WIRELESS INTERNET AND MOBILE COMPUTING

Interoperability and Performance

Yu-Kwong Ricky Kwok
Colorado State University

Vincent K. N. Lau
The Hong Kong University of Science and Technology

IEEE PRESS

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

SAMSUNG 1020

This Page Intentionally Left Blank

WIRELESS INTERNET AND MOBILE COMPUTING



THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

WIRELESS INTERNET AND MOBILE COMPUTING

Interoperability and Performance

Yu-Kwong Ricky Kwok
Colorado State University

Vincent K. N. Lau
The Hong Kong University of Science and Technology

IEEE PRESS

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

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, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Kwok, Yu-Kwong Ricky.
Wireless Internet and mobile computing : interoperability and performance /
Yu Kwong Ricky Kwok and Vincent Lau.
p. cm.
ISBN 978-0-471-67968-4 (cloth)
1. Wireless communication systems. 2. Mobile computing. 3. Internetworking (Telecommunication) I. Lau, Vincent K. N. II. Title.
TK5103.2.K95 2007
621.382—dc22

2007001709

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To our wives and kids:
Fion, Harold, Amber,
Elvina, and Sze-Chun

This Page Intentionally Left Blank

CONTENTS

Preface xvii

Acknowledgments xxix

Acronyms xxxi

PART I ESSENTIALS OF WIRELESS COMMUNICATIONS

1 The Mobile Radio Propagation Channel 1

1.1 Introduction 1
1.2 Large Scale Path-loss 3
1.3 Shadowing Effects 6
1.4 Small Scale Multipath Fading Effects 7
1.5 Practical Considerations 13
1.6 Summary 16
Problems 16

2 Modulation Techniques 19

	2.1	Introduction	19
	2.2	Signal Space and Geometric Representation of Signals	21
	2.3	Modulation Design and Signal Constellations	25
	2.4	Demodulation Design and Optimal Detection in AWGN Channels	28
	2.5	Performance and Tradeoffs	34
	2.6	Practical Illustrations: Digital Modulation	38
	2.7	Summary	44
		Problems	45
3	**Multiuser Communications**		**47**
	3.1	Introduction	47
	3.2	Information Theoretic Overview	49
	3.3	Orthogonal Resource Partitioning	54
	3.4	Non-Orthogonal Resource Partitioning	65
	3.5	Spectral Efficiency and Performance Issues	73
	3.6	Practical Illustrations: GSM, CDMA and Wireless LAN	81
	3.7	Summary	85
		Problems	86
4	**Diversity Techniques**		**87**
	4.1	Introduction	87
	4.2	Effects of Flat Fading on BER Performance	88
	4.3	Effects of Frequency Selective Fading on BER Performance	89
	4.4	Diversity: A Key Technique to Combat Flat Fading Channels	90
	4.5	Equalization*	98
	4.6	Practical Illustration: RAKE Receiver	101
	4.7	Summary	105
		Problems	106
	PART II CELLULAR WIRELESS TECHNOLOGIES		
5	**Overview and Evolution of Cellular Technologies**		**111**
	5.1	Introduction	111
	5.2	Evolution of Cellular Systems	113
	5.3	Technical Challenges to Realize 3G Services	121
	5.4	Summary	131

Problems 132

6 CDMA (IS-95) 133

6.1 Introduction 133
6.2 System Architecture of IS95 134
6.3 Physical Layer and Physical Channels 136
6.4 Call Processing 140
6.5 Power Control 154
6.6 Soft Handover 164
6.7 Summary 173
Problems 174

7 GSM 175

7.1 Introduction 175
7.2 GSM System Architecture 176
7.3 GPRS System Architecture 179
7.4 Radio Interface 184
7.5 Core Network Interface and Services 194
7.6 Summary 216
Problems 217

8 Wideband CDMA and Beyond 219

8.1 Introduction 219
8.2 UMTS Architecture 221
8.3 Packet Switched Connections in UMTS (Rel 99) 243
8.4 Packet Scheduling in HSDPA (Rel 5) 247
8.5 Summary 254
Problems 255

PART III SHORT-RANGE WIRELESS TECHNOLOGIES

9 IEEE 802.11x WLAN Standards 259

9.1 Introduction 259
9.2 Design Goals 260
9.3 IEEE 802 Architecture 260
9.4 IEEE 802.11 MAC Layer 264

9.5	IEEE 802.11 Physical Layers	272
9.6	IEEE 802.11e for QoS Provisioning	277
9.7	Advanced Developments	279
9.8	Practical Illustration: Home Network	282
9.9	Summary	283
	Problems	284
10	**Bluetooth WPAN**	**285**
10.1	Introduction	285
10.2	Design Goals	286
10.3	Bluetooth Protocol Stack	287
10.4	Bluetooth Physical and MAC Layers	292
10.5	Piconets and Scatternets	296
10.6	Performance Issues	303
10.7	Practical Illustration: Sensor Network	304
10.8	Summary	305
	Problems	307
11	**Coexistence Issues**	**309**
11.1	Introduction	309
11.2	The ISM Band Spectrum	310
11.3	Packet Collision	311
11.4	Possible Solutions	313
11.5	IEEE 802.15 TG2	318
11.6	Interference Source Oriented Adaptive Frequency Hopping	320
11.7	Interference Source Oriented Master Delay MAC Scheduling	322
11.8	Performance Issues	322
11.9	Practical Illustration: Bluetooth and Wi-Fi Colocate Transmissions	324
11.10	Summary	327
	Problems	328
12	**Competing Technologies**	**329**
12.1	Introduction	329
12.2	IrDA	331
12.3	HomeRF	338

12.4 HIPERLAN 343
12.5 Practical Illustrations—Implementation of HIPERLAN 352
12.6 Summary 354
Problems 355

PART IV PROTOCOL ADAPTATIONS FOR WIRELESS NETWORKING

13 Mobile IP 359

13.1 Introduction 359
13.2 Advertisement Mechanisms 360
13.3 Registration 363
13.4 Tunneling Approaches 368
13.5 Route Optimizations 369
13.6 Practical Illustrations—Hierarchical Mobility Management 372
13.7 Summary 373
Problems 373

14 IPv6 377

14.1 Introduction 377
14.2 Design Goals 378
14.3 Mobility Support 387
14.4 Home Agents Discovery 389
14.5 Practical Illustrations—IPv6 Based VTHD Network 390
14.6 Summary 394
Problems 394

15 Wireless Application Protocol (WAP) 395

15.1 Introduction 395
15.2 WAP Service Model 396
15.3 WAP System Architecture 400
15.4 WAP Protocol Stack 403
15.5 WAP Profiles and Caching 407
15.6 Practical Illustrations—Location-Aware Advertising System 410
15.7 Summary 412
Problems 412

16 TCP over Wireless 413

16.1 Introduction 413
16.2 TCP Congestion and Error Control 414
16.3 Deficiencies and Overview of Solutions 418
16.4 Link Layer Approaches 419
16.5 Split Connection Approaches 422
16.6 End-to-End Approaches 425
16.7 Practical Illustrations—Wireless WAN 427
16.8 Summary 429
Problems 430

PART V WIRELESS RESOURCES MANAGEMENT

17 Wireless Packet Scheduling 433

17.1 Introduction 433
17.2 The Scheduling Problem 434
17.3 System Model 437
17.4 Fairness Notions 444
17.5 Fair Queueing Approaches 449
17.6 Practical Illustrations—HSDPA 453
17.7 Summary 453
Problems 454

18 Power Management 455

18.1 Introduction 455
18.2 Characterization of Power Consumption 456
18.3 Power Conservation Schemes 458
18.4 Performance Issues 468
18.5 Practical Illustrations—The pcMAC Scheme 470
18.6 Summary 472
Problems 473

19 Ad Hoc Routing 475

19.1 Introduction 475
19.2 Application Scenarios 477
19.3 System Model 478
19.4 Performance Metrics 480
19.5 Routing Protocols 482

19.6 Practical Illustrations—WLAN Based Ad Hoc Routing 497
19.7 Summary 499
Problems 500

20 Wireless Data Caching 501

20.1 Introduction 501
20.2 Application Scenarios 502
20.3 System Architecture 504
20.4 Performance Metrics 506
20.5 Data Caching Approaches 507
20.6 Practical Illustrations—Proactive Key Caching 517
20.7 Summary 518
Problems 518

21 Security Issues 519

21.1 Introduction 519
21.2 Security Considerations 521
21.3 Cellular Wireless Security 522
21.4 Short-Range Wireless Security 529
21.5 Practical Illustrations—Wireless LAN Security Auditing 538
21.6 Summary 540
Problems 541

PART VI MOBILE COMPUTING APPLICATIONS

22 VoIP on Wireless 545

22.1 Introduction 545
22.2 iGSM VoIP Approach 546
22.3 iGSM System Architecture 547
22.4 iGSM Call Processing 548
22.5 Practical Illustrations—Problems and Solutions for VoIP over IEEE 802.11 556
22.6 Summary 560
Problems 560

23 Wireless Video 561

23.1 Introduction 561
23.2 Real-Time Service Goals 562
23.3 System Constraints 563
23.4 Error Concealment Techniques 563
23.5 Joint Source Channel Coding Techniques 573
23.6 Practical Illustrations—Rate Adaptation for MPEG-4 Video 574
23.7 Summary 575
Problems 577

24 Wireless File Systems 579

24.1 Introduction 579
24.2 File System Service Model 580
24.3 General Principles for the Design of a Distributed File System 582
24.4 Replication Services and Mechanisms 582
24.5 Disconnected Operations and Caching 584
24.6 Weakly Connected Operations 588
24.7 Practical Illustrations—Mobile Distributed Database 590
24.8 Summary 595
Problems 595

25 Location Dependent Services 597

25.1 Introduction 597
25.2 Mobile Location Service Model 598
25.3 Spatial Analysis Techniques 600
25.4 Mobile Positioning Techniques 604
25.5 Practical Illustrations—The Cricket Location Support System 608
25.6 Summary 613
Problems 613

26 Trust Bootstrapping in Wireless Sensor Networks 615

26.1 Introduction 615
26.2 Key Predistribution 619
26.3 Key Predistribution with Deployment Knowledge 625
26.4 Key Establishment 630
26.5 Discussions and Future Work 642
26.6 Summary 643

Problems 643

27 Peer-to-Peer Computing over Wireless 645

27.1 Introduction 645
27.2 A Taxonomy of P2P Systems 646
27.3 Wired P2P Systems 647
27.4 Topology Control in Wireless Ad Hoc Networks 651
27.5 Media Objects Replication Techniques 654
27.6 Practical Illustrations—Localized Topology Control 655
27.7 Summary 661
Problems 661

28 Incentives in Peer-to-Peer Computing 663

28.1 Introduction 663
28.2 Incentive Issues in P2P Systems over the Internet 667
28.3 Incentive Issues in Wireless P2P Systems 688
28.4 Discussion and Future Work 693
28.5 Summary 695
Problems 695

Appendix A: Optimality of Minimum Distance Decoder 697

References 701

Topic Index 725

This Page Intentionally Left Blank

PREFACE

The Theme:

The all-mighty Internet has extended its reach to the wireless realm. With this exciting development, we have an ambitious goal in this book—to help the reader to build up sound technical understanding of what is going on in a large-scale networking system as depicted graphically in Figure 0.1. Using a bottom-up approach, we would like the reader to understand how it is feasible, for instance, for a cellular device user (top left corner in the figure) to communicate, via the all-purpose TCP/IP protocols, with a wireless notebook computer user (bottom left corner in the figure), traversing all the way through a base station in a cellular wireless network (e.g., GSM, CDMA), a public switched telephone network (PSTN), the Internet, an intranet, a local area network (LAN), and a wireless LAN access point. In traveling through this long path, the information bits are processed by numerous disparate communication technologies (in slightly more technical terms, processed by many different protocol stacks and wireless air interfaces). We also describe the technologies involved in infrastructureless (i.e., ad hoc) wireless networks (the bottom right corner in the figure), which are widely envisioned to be the most popular form of mobile computing infrastructure, pertaining to many interesting applications: most notably, wireless network games. Our focus is on how these distinctive technologies can work together

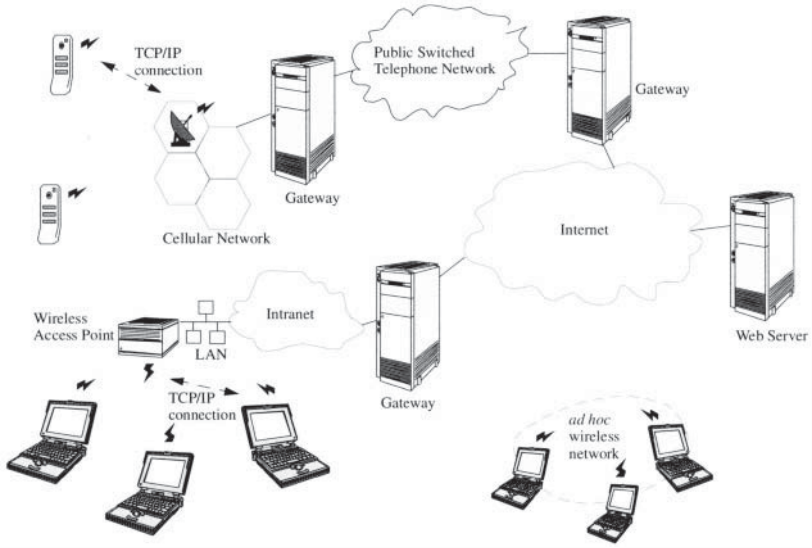


Figure 0.1 The Internet extends into the wireless realm.

through various adaptation methods. Specifically, throughout this book we use two major technical challenges to delineate and motivate our discussion—*interoperability* and *performance*.

We believe that interoperability is the very first concern in integrating so many different components to realize a seamless communication path for users with disparate wireless devices. Indeed, many interesting features of intelligent communication technologies emerge out of the need to address the interoperability issues. Performance is another major challenge in a resource-limited wireless environment. Combined, these two challenges have baffled researchers and engineers for years, and there are many technical imperfections yet to be tackled.

Cutting vertically through the protocol stacks at various junctions in the distributed wireless networking environment shown in Figure 0.1, we divide the book into six parts:

I. Essentials of Wireless Communications: In this part we provide the reader with a solid foundation about wireless communications in a concise, yet complete manner. We do not aim to repeat all the minute details that can be found in classical digital communications theory books, but we explicate the overall theory in a unified manner.

II. Cellular Wireless Technologies: In this part we describe in an evolutionary manner the developments of the various practical cellular wireless technologies: the last wireless mile of the Internet, highlighting the ingenious adaptation schemes

designed for satisfying the interoperability requirement while enhancing performance continuously from one generation to another.

III. Short-Range Wireless Technologies: In this part we describe and explain the various short-range wireless technologies: the last 100 m of the Internet, including Bluetooth, wireless LAN, and infrared technologies. Unlike the cellular counterpart, which is driven by voice communication-oriented application requirements, short-range wireless technologies are driven by computing oriented application requirements, and as such, entail a number of different system-level issues. In particular, the coexistence issue between the two most popular technologies, Bluetooth and IEEE 802.11b WLAN, is very acute, as it critically affects the performance perceived by users.

IV. Protocol Adaptations for Wireless Networking: In this part we describe the various ingenious adaptation techniques for our all-purpose HTTP/TCP/IP protocol suite, which is a vital component in many parts of the distributed wireless networking environment, as shown in Figure 0.1. We start with the IP-Mobile IP and IPv6. We then describe the infamous wireless application protocol (WAP) and its design rationales, which suggest a complete reworking of the TCP/IP stack. We move on to the TCP adaptations, which are important in providing stable application-level performance.

V. Wireless Resources Management: In this part we discuss management of the scarce resources in a wireless environment for different layers: physical layer, multiple access control layer, network layer, and application layer. Specifically, we describe power management, packet scheduling policies, routing, data caching, and security issues. From the interoperability and performance angles, the interplay between the wireless resources management schemes and the various protocol adaptation technologies is also our focus throughout this part.

VI. Mobile Computing Applications: In this part we move on to the application level in that we describe the various adaptation techniques, interoperability issues, and performance aspects in realizing mobile computing services, including voice, video, and file system abstraction. To achieve efficient anytime, anywhere wireless information services, location-dependent applications are envisioned as the core underlying delivery mode. Finally, we also look into the future by discussing the ultimate form of truly peer-to-peer wireless ad hoc wireless networking and wireless sensor networking.

Blending Practice and Theory:

Throughout the book we use practical real-life case studies, derived from our own research and development experience in both the industry and academia, to serve as motivating examples for helping the reader to gain in-depth understanding of the technologies taught. These examples are designed to bring out the dynamics of the

underlying systems. For example, in addition to describing various subsystems (e.g., soft handover, power control, etc.) in cellular systems, we illustrate the interactions of various subsystems involved in several user application scenarios in a dynamic manner.

The technologies we describe in the book have evolved from several areas: digital communications, information theory, signal processing, wireless networking, Internet protocol advancements, and mobile computing applications. We firmly believe that these technological areas will synergize with each other, and the ultimate goal is to achieve optimized performance at the user level. For instance, while video processing techniques have traditionally been the outcome of signal processing research, recent advancements are the fruits of research projects that take into account wireless networking and physical layer design results.

We strongly feel that a comprehensive book, developed via the interoperability and performance perspective, and cutting through the various protocol layers—from the physical layer to the application layer—is in pressing need by both researchers and practitioners in the field. Indeed, we have conducted various lectures on the subject for senior undergraduate and master-level students in the past few years. It is difficult to find a comprehensive textbook and reference book on the market that covers both the scope and depth of the subject. This motivated us to develop a book that will allow people to better understand the engineering details, followed by a firm grasp of development trends at all the protocol layers which are interdependent, and then to form an opinion as to when and how the technologies can be applied in various settings, be they commercial or educational.

Our target audience of this book includes senior undergraduate students, graduate students, and engineers in the field. The prerequisite knowledge required includes basic communication engineering concepts and basic understanding of internetworking protocols. We expect our readers to walk through the technological descriptions and most important, the practical illustrative examples, and then map them to their own academic and/or practical experience. We would describe the book as being on an advanced level, but not a research-oriented specialist treatment.

Road-Map:

- **Part I: Essentials of Wireless Communications**
  - **Chapter 1: The Mobile Radio Propagation Channel** This chapter aims at providing the reader with a solid understanding of the properties of the wireless transmission medium. Such a background will be useful for the reader to understand the various adaptation and optimization approaches devised to enhance system performance. To strengthen the reader's understanding, we provide a detailed description of the wireless channel propagation characteristics in the popular GSM (narrowband), CDMA (wideband), and IEEE 802.11a/b/g (wideband) environments.
  - **Chapter 2: Modulation Techniques** The core of this chapter is the evolution of different digital modulation techniques, which are critical

in achieving high bandwidth efficiency. The basic trade-off of bit rate, power/bandwidth, and quality of communication is illustrated both qualitatively and quantitatively. Design considerations and issues with respect to various channel models described in Chapter 1 (narrowband and wideband) are illustrated. To allow the reader to appreciate the intelligent design of popular modulation methods, we use the GSM, CDMA, and IEEE 802.11a/b/g as examples to illustrate the concepts.

- **Chapter 3: Multiuser Communications** While Chapter 2 deals with point-to-point communications, we extend the framework in this chapter to consider multiuser communications. The chapter begins with an information-theoretical introduction on multiuser communications. Next, we introduce various commonly adopted resource partitioning techniques, together with a quantitative comparison of their strengths and weaknesses. We then illustrate the practical design choices in the GSM, CDMA, and IEEE 802.11b systems.
- **Chapter 4: Diversity Techniques** Diversity is becoming an important problem dimension that can be exploited to further enhance system capacity and performance. In this chapter we discuss various diversity parameters that are considered as practicable performance enhancement approaches. We describe the practical implementation methodology in realizing the various diversity techniques (e.g., RAKE receiver, open-loop transmit diversity, closed-loop transmit diversity). To give the reader a solid understanding of the efficacy of the diversity concept, we use the UMTS and IEEE 802.11a systems as real-life examples.

- **Part II: Cellular Wireless Technologies**

  - **Chapter 5: Overview and Evolution of Cellular Technologies** In this chapter we aim to provide a comprehensive review of the rationale, theory, and practical issues of modern cellular wireless communications technologies. Motivated by the efficient frequency reuse features, cellular networks are designed and engineered to be highly cost-effective but also entail a number of system issues such as mobility management and cochannel interference. We give a brief overview of the evolution of cellular systems from 1G to 3G+ and compare the paths in Europe, the United States, and Japan. We discuss the technical challenges in realizing 3G services. Finally, we describe briefly how GPRS/EDGE/3G1x/UMTS overcomes these technical challenges.
  - **Chapter 6: CDMA (IS-95)** Based on the theoretical foundation given in Chapter 4, we describe in detail the design goals and system architecture of a practical CDMA (IS-95) cellular infrastructure. To prepare for the 3G topics in Chapter 10, we present the details of various subsystems [such

as the air interface, the radio resource management (soft handoff, power control), and mobility management]. To highlight the dynamic interactions of various subsystems, we describe several real-life call-processing scenarios.

- **Chapter 7: GSM** GSM is by far the most successful cellular communication technologies, driven perhaps not only by technical superiority but by economic forces as well. In a retrospective angle, we describe its original design goals, followed by a detailed presentation of its practical architecture. Besides describing the design of various subsystems (such as the air interface, the protocol architecture, radio resource management, mobility management), we shall explain the design based on the theories described in Part I. To help the reader appreciate the efficacy of the GSM system, we describe the interactions of various subsystems in a dynamic way through the use of detailed real-life call-processing examples.
- **Chapter 8: Wideband CDMA and Beyond** To cope with bandwidth-hungry applications (e.g., multimedia messaging services), there is no choice but to let the cellular network systems march into the broadband era. Here, interoperability is a prime concern, as operators have to provide a seamless migration of the huge user population and install base to next-generation systems. Thus, there are a plethora of engineering chores that must be tackled with high efficiency in the actual implementations of the so-called 3G systems. In this chapter we describe the most common modes of IMT2000 umbrella: the CDMA2000 and UMTS systems. The focus is on packet-switched services. High-data-rate services are then described in detail. Some advanced topics such as macroscopic scheduling and microscopic scheduling, are also discussed. Besides describing various subsystems, we also explain the design philosophy behind the standards. Finally, we use a number of practical wireless Internet access scenarios to illustrate the salient features of 3G systems.

- **Part III: Short-Range Wireless Technologies**
  - **Chapter 9: IEEE 802.11x WLAN Standards** Due to its local cost (commoditized technologies and free wireless spectrum), IEEE 802.11x-based wireless LAN systems have proliferated. Based on the mature MAC design principles in the original IEEE 802 architecture, the IEEE 802.11x standards employ several different physical layer designs (FHSS, DSSS, OFDM, IrDA, etc.), providing a hierarchy of data rates. We describe these system issues in detail and provide a mobile computing scenario based on IEEE 802.11b LANs, using both the point coordination function (PCF) and distributed coordination function (DCF).
  - **Chapter 10: Bluetooth WPAN** Bluetooth attracted high-level media attention when it was introduced around the beginning of the new millen-

nium. Also based on the free ISM frequency band, Bluetooth turned out to be much delayed in its deployment for various engineering and economic reasons. Nevertheless, Bluetooth's compact design—the TDD TDMA master-slave communication model—has eventually made it pervasive in a wide range of practical gadgets. We discuss the design rationale of the many interesting features of Bluetooth, including the piconet and scatternet modes, the latter yet to be realized. We provide several users' application scenarios to let the reader know the unique characteristics of Bluetooth.

- **Chapter 11: Coexistence Issues** The free ISM frequency band is now occupied by a crowd of short-range wireless technologies, most notably Bluetooth and IEEE 802.11b. The consequence is that when devices of different technologies come into each other's range, interference inevitably occurs, rendering significant performance degradation unavoidable. Thus, various commercial efforts and IEEE task forces have been dispatched to investigate possible solutions of the coexistence problem. We first illustrate the collision problem quantitatively and then describe various suggested solutions and their rationales. Among the many suggestions, adaptive frequency hopping (e.g., to be used in Bluetooth systems) seems to be most practicable. We provide a solid description of such technologies, followed by a real-life illustrative example.
- **Chapter 12: Competing Technologies** Short-range wireless communication has traditionally been the territory of IrDA. We describe the functionality and features of IrDA-based systems and its applications. Recent competitors, other than Bluetooth and IEEE 802.11x, include several commercial efforts, such as HomeRF and HIPERLAN. We also discuss in detail the strengths and weaknesses of these new technologies. We then illustrate the usefulness of IrDA using a few real-life application examples.

- **Part IV: Protocol Adaptations for Wireless Networking**

  - **Chapter 13: Mobile IP** In this chapter we begin our discussion of higher-layer protocols. Specifically, we motivate the reader regarding the need to devise intelligent adaptation strategies to make the classical wireline protocols (e.g., IP, TCP, HTTP) work in a wireless environment to satisfy the interoperability requirement. We describe various ingenious designs in making a mobile version of IP interoperable with the wireline version. We use several practical scenarios to illustrate the functionality of Mobile-IP.
  - **Chapter 14: Ipv6** IPv6, designed to tackle a number of problems in wireline Internet, is also very important in the wireless environment. Indeed,

IPv6 is designed deliberately to have a number of features that make IP independent of the underlying physical medium. We describe such features in detail and compare them against those discussed in Chapter 15. We use similar practical situations to illustrate similarities and differences in relation to the two different IP adaptation approaches.

- **Chapter 15: WAP** The Wireless Application Protocol also caught intense public attention when it was first published. Designed for seamless integration of wireline Internet and cellular wireless networks, WAP possesses a number of special features to optimize the transportation of heavy weight Internet contents while satisfying the interoperability requirements. We use a real-life scenario that is found in a commercial environment to illustrate the usefulness of WAP as well as to show the weakness of WAP, thereby letting the reader know why it has not become as popular as it was initially touted to be.
- **Chapter 16: TCP over Wireless** TCP has a number of salient features, making it the hard core of many Internet services for over three decades. Among these features, congestion control has been receiving great attention in the wireline Internet research community, and as such, many intelligent new techniques are devised to make the congestion control performance better. However, such techniques are not suitable in a wireless environment, in which packet reception errors (e.g., missing packets within a TCP window) may be due not only to congestion but to channel error as well. We describe such phenomena in detail to motivate the reader about the need to design new techniques for the adaptation of TCP in a wireless environment. We describe several more mature approaches using numerous examples. We then use several practical scenarios to illustrate how these methods are being used in a real-life situation.

- **Part V: Wireless Resources Management**

  - **Chapter 17: Wireless Packet Scheduling** This chapter kicks off our discussion of resources management, which is of utmost importance in a wireless computing environment, wherein resources (power and bandwidth) are desperately deficient compared with the connected wireline counterpart. We talk about packet scheduling in this chapter by walking through several popular scheduling approaches, designed based on the fundamental fairness notions. We discuss in detail the performance evaluation criteria for scheduling techniques. We then use the scheduling approach deployed in practical cdma2000 systems to show the reader that a good scheduler can indeed enhance the utilization and performance of a wireless infrastructure.
  - **Chapter 18: Power Management** Power is always a limiting factor in a wireless environment. We describe quantitatively how we quantify

power consumption. We then discuss general principles in conserving power. We describe the performance issues related to these power-saving strategies. Using the popular IEEE 802.11b standard as an example, we illustrate how far we can go in saving power in a mobile computing system.

- **Chapter 19: Ad Hoc Routing** Finding a stable and good (in the sense that the bandwidth is higher) route in an infrastructureless network (i.e., ad hoc network) is critical for two devices that are out of each other's range to communicate. Leveraging the legacy techniques designed in the 1980s for packet radio networks, many new algorithms are devised for efficient routing in ad hoc networks. We describe in detail the system model, performance metrics, and protocol features. We then use an IEEE 802.11b-based ad hoc environment to illustrate the use of such ad hoc routing protocols.
- **Chapter 20: Wireless Data Caching** Contrary to an infrastructureless environment, in today's mobile computing systems, there usually are centralized servers to provide wireless access and information services to the tetherless client devices. Among the many system services provided by the server, the cooperative caching (cooperation between the clients and the server) model is one of the most important features. Specifically, the server, usually acting as a proxy, will periodically broadcast invalidation reports to alert clients that some data cached in the clients may be obsolete. We discuss in detail the application scenarios, system architecture, and performance issues of the caching problem. We then describe several interesting approaches to tackling this problem. We use an IEEE 802.11b-based infrastructure to illustrate the practicality of the caching techniques.
- **Chapter 21: Wireless Security** Wireless security is an important but largely immature technical area. In this chapter we aim to provide a general overview of the issues involved and the current approaches used. We divide our discussion into cellular long-range wireless security and short-range wireless security. We use an IEEE 802.11b network to illustrate the use of security enhancement techniques and the difficult problems faced by security protocol designers.

- **Part VI: Mobile Computing Applications**

  - **Chapter 22: VoIP on Wireless** This chapter begins the discussion of various higher-level mobile computing issues. We talk about an alternative approach in providing voice services over wireless-based VoIP. We discuss about the advantages and weaknesses of this approach, through a detailed description of the use of VoIP over wireless. We then describe

the various engineering efforts required to fulfil the performance and interoperability requirements. We use an IEEE 802.11b-based system to illustrate the actual practice of VoIP over wireless.

- **Chapter 23: Wireless Video** Video is beyond doubt a very important media in the use of wireless infrastructure. Due to the heavy resource requirements of video, sophisticated approaches are devised to carry video efficiently over a wireless link. Most notably, many joint source channel coding techniques have been suggested in the research community. We provide a unified overview of such techniques. We use an IEEE 802.11b-based system to highlight the difficulties involved and the solutions designed for handling video data.
- **Chapter 24: Wireless File Systems** A file system is a diehard abstraction that any meaningful information system must efficiently provide. A wireless mobile computing environment is no exception. However, the frequent disconnection problem and the time-varying channel quality make providing such an abstraction rather difficult. We discuss in detail several effective techniques in tackling the various file system problems, such as consistency, integrity, and security. We use the infamous Coda file system as an illustrative example.
- **Chapter 25: Location-Dependent Services** It is widely conceived that providing suitable services in a suitable environment at a suitable time is the most desirable to a wireless device user, due to the needs for conserving power, simplifying user input, and maximizing system utilization. However, providing truly seamless location-dependent services entails a number of difficult engineering issues, including location management, intelligent information pushing, and efficient remote query processing. We discuss these issues in detail and then we use a practical real-life example to illustrate the functionality of location-dependent advertising.
- **Chapter 26: Trust Bootstrapping in Wireless Sensor Networks** We discuss a recently hot research topic about wireless sensor networks in this chapter—the establishment of mutual trust among sensors. Without such mutual trust, it is infeasible for the wireless sensors to carry out secure data exchange. However, a major constraint is that usually we have no control about the network topology of a wireless sensor network. Consequently, we usually do not know the neighboring relationship among sensors so as to set up encryption keys in an a priori manner. We summarize representative major research findings in this interesting topic.
- **Chapter 27: Peer-to-Peer (P2P) Computing over Wireless Networks** In this chapter we aim to help the reader to look ahead into the future about what mobile computing will look like. We provide our personal opinions based on our own research as well as on various radical ideas from the research community. In particular, we describe in detail to the

reader what we believe is the ultimate form of wireless mobile information processing—selfish wireless computing.

- **Chapter 28: Incentives in Peer-to-Peer Computing** In this chapter we follow up Chapter 27 by discussing a fundamental requirement of P2P computing—to provide adequate incentives for autonomous and possibly selfish peers to cooperate. Several types of incentive schemes have been proposed recently by world-renowned researchers. The reader should find this chapter's concepts and techniques interesting, yet intriguing.

RICKY AND VINCENT

Hong Kong

This Page Intentionally Left Blank

ACKNOWLEDGMENTS

The authors would like to thank our families for their love and support in this project. We sincerely thank Mr. Tyrone Kwok and Mr. Carson Hung (both are Ph.D. students of Y.-K. Kwok) for their invaluable help in compiling the materials and illustrations in the book. In particular, Tyrone has provided various criticisms and comments on various parts of the book. We also thank Mr. Ray Lau (a close friend of Y.-K. Kwok), who made important contributions in the design of the book cover. The authors also thank all the people who have helped in the preparation of this book in one way or the other. Thanks are due to Ms. Rachel Witmer and Ms. Angioline Loredo for their professional advice and assistance. Indeed, we owe a deep sense of gratitude to our friends and families, especially now that this project has been completed.

Y.-K.

This Page Intentionally Left Blank

ACRONYMS

ACK	Acknowledgment
ACL	Asynchronous Connectionless Link
AFH	Adaptive Frequency Hopping
AOA	Angle of Arrival
AODV	Ad Hoc On-Demand Distance Vector
ARQ	Automatic Repeat Request
AVSG	Accessible Volume Storage Group
AWGN	Additive White Gaussian Noise
BPSK	Binary Phase-Shift Keying
CCK	Complementary Code Keying
CDMA	Code-Division Multiple Access
CRC	Cyclic Redundancy Check
CSMA/CA	Carrier-Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier-Sense Multiple Access with Collision Detection

CTS	Clear to Send
D-CDMA	Deterministic CDMA
DCF	Distribution Coordination Function
DHT	Distributed Hash Table
DMT	Discrete Multitone
DSDV	Destination-Sequenced Distance-Vector
DSL	Digital Subscriber Line
DSR	Dynamic Source Routing
DSSS	Direct Sequence Spread Spectrum
FDMA	Frequency-Division Multiple Access
FHSS	Frequency-Hopping Spread Spectrum
GGSN	Gateway GPRS Support Node
GMSK	Gaussian Minimum Shift Keying
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HLR	Home Location Register
IFS	Interframe Space
IMSI	International Mobile Subscriber Identity
ISI	Intersymbol Interference
ISM	Industrial, Scientific, Medical
IV	Initialization Vector
L2CAP	Logical Link Control and Adaptation Protocol
LMP	Link Manager Protocol
LMU	Location Measurement Unit
MAC	Multiple Access Control
MDMS	Master Delay MAC Scheduling
MSC	Mobile Switching Center
NAP	Network Access Point
NAT	Network Address Translation
OBEX	Object Exchange
OFDM	Orthogonal Frequency-Division Multiplexing

OMA	Open Mobile Alliance
OVSF	Orthogonal Variable Spreading Factor
P2P	Peer-to-Peer
PCF	Point Coordination Function
PCS	Personal Communication System
PPM	Pulse Position Modulation
PPP	Point-to-Point Protocol
PSMM	Pilot Strength Measurement Message
PSTN	Public Switched Telephone Network
PSNR	Peak Signal-to-Noise Ratio
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase-Shift Keying
R-CDMA	Random CDMA
RLP	Radio Link Protocol
RREP	Route Reply
RREQ	Route Request
RSSI	Received Signal Strength Indication
RTS	Request to Send
SACK	Selective ACK
SCO	Synchronous Connection-Oriented
SDMA	Space-Division Multiple Access
SDP	Service Discovery Protocol
SGSN	Serving GPRS Support Node
SIM	Subscriber Identification Module
SMLC	Serving Mobile Location Center
SRES	Signed Response
TCP	Transport Control Protocol
TKIP	Temporal Key Integrity Protocol
TLS	Transport Layer Security
TMSI	Temporary Mobile Subscriber Identity
TODA	Time Difference of Arrival

UDP	User Datagram Protocol
VLR	Visitor Location Register
WAE	Wireless Application Environment
WAP	Wireless Application Protocol
WDP	Wireless Datagram Protocol
WEP	Wired Equivalent Privacy
WLAN	Wireless Local Area Network
WML	Wireless Markup Language
WPA	Wi-Fi Protected Access
WSP	Wireless Session Protocol
WTA	Wireless Telephony Application
WTAI	Wireless Telephony Application Interface
WTLS	Wireless Transport Layer Security
WTP	Wireless Transaction Protocol
XML	Extensible Markup Language

PART I

ESSENTIALS OF WIRELESS COMMUNICATIONS

This Page Intentionally Left Blank

CHAPTER 1

THE MOBILE RADIO PROPAGATION CHANNEL

The beaten path is the safest. (*Via trita est tutissima.*)

—Latin proverb

1.1 INTRODUCTION

A communication system consists of a transmitter and a receiver connected by a *channel*. The channel is a *black box* representation of the actual physical medium where a signal at the channel input will produce a corresponding channel output after suffering from various *distortions*. Examples of a physical medium include telephone wire, coaxial cable, radio frequency (RF), and optical fiber. Different meda have different signal propagation characteristics and require different designs. In this chapter we focus on understanding the signal propagation characteristics of wireless channels.

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau

Copyright © 2007 John Wiley & Sons, Inc.

The simplest wireless channel is the *additive white Gaussian Noise* (AWGN) channel where the output signal from the channel is given by:

$$y(t) = x(t) + z(t)$$

where $x(t)$ is the channel input and $z(t)$ is the white Gaussian channel noise. In this channel, the received signal composes of an undistorted transmit signal contaminated with channel noise. In Chapter 2, we discuss the design and performance of basic digital communication systems using AWGN channel as a simple example. Note that AWGN channel is quite an accurate model for deep space communications between earth station and geostationary satellites or spacecrafts.

On the other hand, for terrestrial wireless communications, the channel model is much more complicated due to the time-varying nature as well as the multipath nature of the wireless channel [271]. For instance, multipath refers to the situation that there are more than one propagation path between a transmitter and a receiver. The received signals from the multipaths are superimposed with each other. Depending on the phase difference (or path difference) of the multipaths, the superposition may be *constructive* or *destructive*. When the superposition is constructive (path difference around $n\lambda$ for some integer n where λ is the carrier wavelength), the received signal will be strong. On the other hand, when the superposition is destructive (path difference around $n\lambda/2$ for some integer n), the received signal will be quite weak. This is illustrated in Figure 1.1

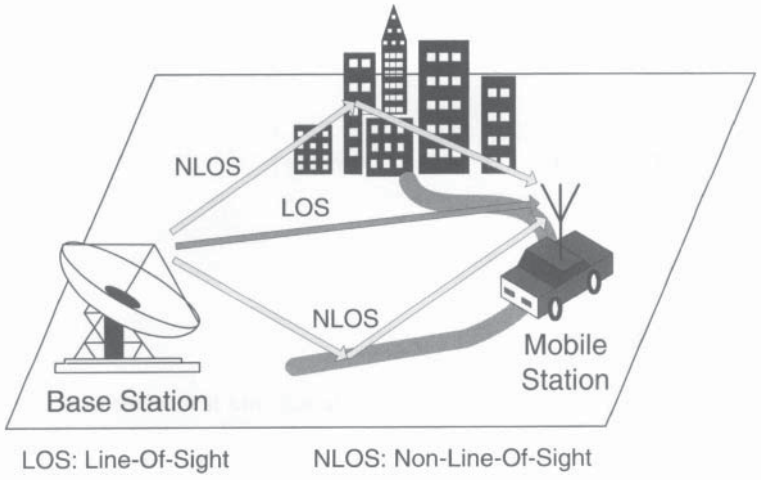


Figure 1.1 Illustration of multipath in terrestrial wireless communications.

Typically, the signal envelop as a result of multipath superposition can fluctuate over 30dB or more over a span of a few λ in distance. Suppose the carrier frequency

is 3GHz, the wavelength λ is around 10cm and the large fluctuation of signal strength (35 dB) can be observed with a span of ~30cm which is quite short.

On the other hand, time variation of the wireless channel results from the mobility of the transmitter, the receiver or the environment (e.g., scatterer). Hence, the number of multipaths, the strength of multpaths as well as the delays of the multipaths can be time varying.

To capture different time scale of the distortions introduced by the wireless channels, we introduce a *three-level model*. The first level model is called the *large scale path loss model*. The path loss model focuses on the study of the long-term or large scale variations on the average received signal strength due to the variation of distance from the transmitter and the receiver. The second level model is called the *shadowing model* which focuses on the study of the medium term variation on the medium-term average (over hundreds of λ) received signal strength due to local changes of terrain features or man-made obstacles (such as blockage or coverage shadows). The third level model is called the *microscopic fading model* which focuses on the study of the short-term received signal strength (over a few λ) due to constructive and destructive interference of the multipath. Figure 1.2 illustrates the three level models of wireless channels.

The path loss model is usually used for system planning, coverage analysis and link budget [271]. On the other hand, the shadowing model is usually used for power control analysis, second order interference analysis as well as more detailed coverage and link budget analysis [152]. Finally, the microscopic fading model is usually used for the detail design of the physical layer transmitter and receiver such as coding, modulation, interleaving, etc. [320]. In this chapter, we elaborate the three level models using a qualitative approach in the following sections.

1.2 LARGE SCALE PATH-LOSS

Path loss model focuses on the study of the large scale variation on the average received signal strength due to variation in distance between the transmitter and the receiver [203]. The simplest path loss model is the *free space path loss model* where the averaged received signal strength is inversely proportional to the square of the distance between the transmitter and the receiver. For instance, the received signal power P_r is given by:

$$P_r = \frac{P_t G_t G_r c^2}{16\pi^2 d^2 f^2}$$

where c is the speed of light, P_t is the transmitted power, G_t and G_r are the transmit and receive antenna gains, d is the distance separation between the transmitter and the receiver and f is the carrier frequency of the transmitted signal. In dB form, the received signal power is given by:

$$P_r(dB) = P_t(dB) + K - 20\log_{10} d - 20\log_{10} f.$$

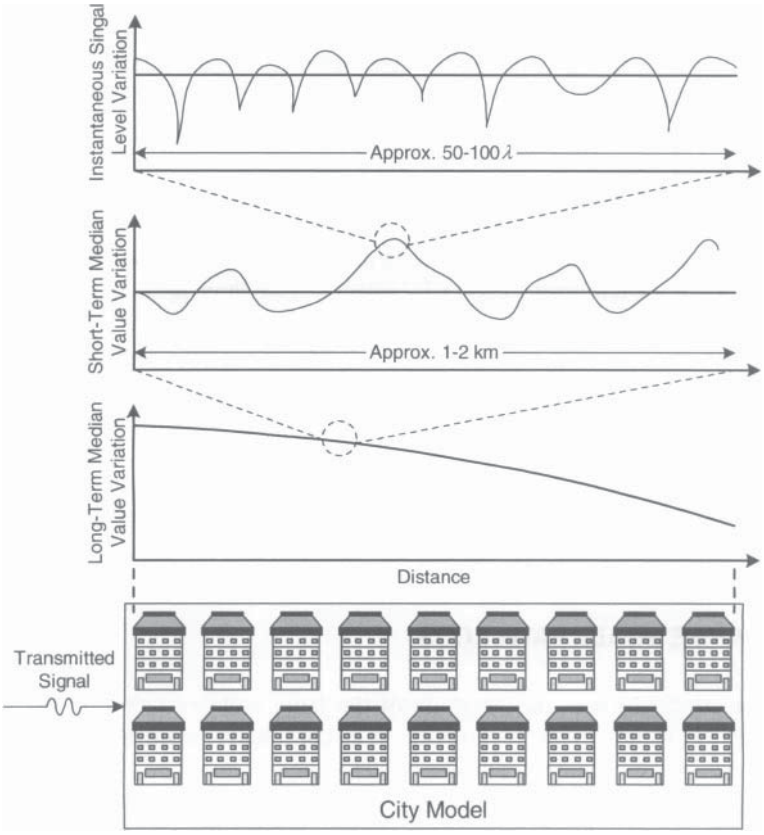


Figure 1.2 Three level channel models for wireless systems.

Define the path loss as:

$$PL(dB) = P_t(dB) - P_r(dB).$$

Hence, the free space path loss equation[23] is given by:

$$\begin{aligned} PL(dB) &= -K + 20\log_{10} d + 20\log_{10} f \\ &= PL(d_0) + 20\log_{10}\frac{d}{d_0} \end{aligned} \tag{1.1}$$

for some reference distance d_0. The path loss equation indicates how fast the received signal strength drops with respect to the change in distance. For example, the received signal strength is reduced by 4 times if we double the distance between transmitter and the receiver or double the carrier frequency. The inverse square law is a direct result of the *Maxwell wave equation* [23] in free space and can be used to model the signal loss due to *line-of-sight* (LOS) propagation mode where the transmitter and receiver can *see* each other.

The free space path loss equation is not accurate to model the path loss in terrestrial wireless communications with multipaths from the transmitter to the receiver. In general, the path loss model can be represented by the equation:

$$PL(dB) = PL(d_0) + 10n\log_{10}(d/d_0)$$

where n is the *path loss exponent* which depends on the environment. Table 1.1 illustrates the values of path loss exponents under various environment.

Table 1.1 Path loss exponents.

Environment	Path Loss Exponent
In building line-of-sight	1.6–1.8
Free space	2
Obstructed in factories	2–3
Urban area cellular	2.7–3.5
Shadowed urban cellular	3–5
Obstructed in building	4–6

The higher the path loss exponent is, the faster the signal strength drops with respect to the increase in distance. For example, a commonly assumed path loss exponent in non line-of-sight environment is $n = 4$. In this case, doubling the distance separation between the transmitter and the receiver will result in 16 times reduction in received signal strength.

In some more complicated propagation environments such as irregular terrain and cities, there is no simple analytical path loss model. Empirical models based on

extensive channel measurements are used to model the path loss versus distance in such complex environments. Examples are the Okumura model [261], Hata model and COST 231 extension to the Hata model [316] for cellular systems simulations and link budget analysis.

1.3 SHADOWING EFFECTS

In the shadowing model, we are interested to study the medium term variation in received signal strength when the distance between the transmitter and receiver is fixed. For example, image a mobile station is circling around a base station with radius r. As the mobile moves, one expects some medium term fluctuations in the received signal strength but the variation is not due to the path loss component because the distance between the transmitter and the receiver is not changed. This signal variation is due to the variations in terrain profile such as variation in the blockage due to trees, buildings, hills, etc. This effect is called shadowing. Consider a signal undergoing multiple reflections (each with an attenuation factor a_i) and multiple diffractions (each with an attenuation factor b_i) as illustrated in Figure 1.3

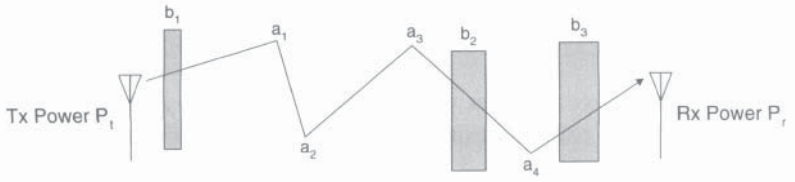


Figure 1.3 Shadowing model—variations in path loss predication due to buildings, trees, hills, etc.

The received signal strength is given by:

$$P_r = P_t \Pi_{i=1}^{N_r} a_i \Pi_{i=1}^{N_d} b_i \tag{1.2}$$

where N_r and N_t denote the number of obstacles with reflecting and diffracting the signal respectively. Expressing the received power in dB, we have:

$$\begin{aligned} P_r(dB) &= P_t(dB) + \sum_{i=1}^{N_r} a_i(dB) + \sum_{j=1}^{N_d} b_i(dB) \\ &= P_t(dB) + \sum_{i} \alpha_i(dB) \end{aligned}$$

where α_i is the attenuation coefficient due to reflections or diffractions. Each of the term $\alpha_i(dB)$ represents a random and statistically independent attenuation. As the number of reflectors and diffractors increase, by central limit theorem, the sum

$S(dB) = \sum_i \alpha_i(dB)$ approaches a Gaussian random variable. Hence, the received signal power can be expressed as:

$$\begin{aligned} P_r(dB) &= S(dB) + P_t(dB) \\ &= \mu_S(dB) + X(dB) + P_t(dB) \end{aligned} \tag{1.3}$$

where $S(dB)\,\mathcal{N}(\mu_S, \sigma^2)$ and $X(dB)\,\mathcal{N}(0, \sigma^2)$. The mean shadowing μ_S is usually included into the path loss model and that is why the path loss exponent can be larger than 2. Hence, without loss of generality, we move the term $\mu_S(dB)$ to the path loss model and consider only the shadowing effect $X(dB)$. Expressing the received power in linear scale, we have:

$$P_r = P_t 10^{X/10} = A_s P_t$$

where $A_s = 10^{X/10}$ is the power attenuation due to shadowing and is modeled by the *log-normal distribution* with standard derivation σ (in dB).

Combining the path loss model and the shadowing model, the overall path loss is given by:

$$PL(dB) = PL_{av}(dB) + X(dB)$$

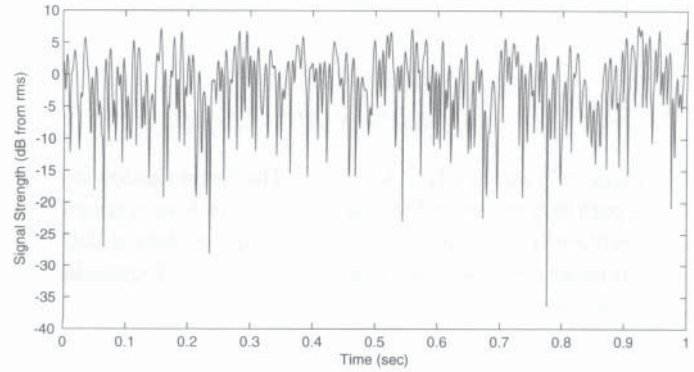
where PL_{av} is the path loss component obtained from the large scale path loss model and $X(dB)$ is the shadowing component which is modeled as a zero-mean Guassian random variable with standard derivation σ in dB.

1.4 SMALL SCALE MULTIPATH FADING EFFECTS

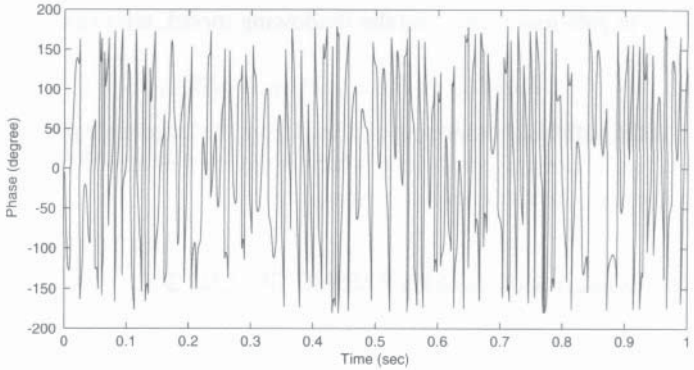
The third component of variation of received signal strength is called the small scale fading or *microscopic fading*. In the small scale fading component, the signal strength fluctuates over 30dB within a very short time scale of the order of milli-second as illustrated in Figure 1.4. There are two independent dimensions in microscopic fading, namely the *multipath dimension* and the *time-varying dimension*. They are elaborated below.

Figure 1.5 illustrates the concept of multipath dimension and time varying dimension of microscopic fading. For example, a narrow pulse transmitted at $t = t_0$ results in one resolvable received pulse. The received pulse in general has a different shape compared with the transmitted pulse because of multipaths propagation. At $t = t_1$, the same transmit pulse results in two resolvable pulses with the first pulse stronger than the second pulse. At $t = t_2$, the same transmit pulse results in two resolvable pulses with the first pulse weaker than the second pulse. Note that the multipath dimension and the time-varying dimension are two independent dimensions. We quantify the two dimensions below.

Multipath dimension: To quantify the multipath dimension in microscopic fading, we can either look at the *delay spread* or *coherence bandwidth*. Delay spread



(a) Rayleigh fading envelope



(b) Rayleigh fading phase

Figure 1.4 Illustration of the envelope and phase variations in microscopic fading.

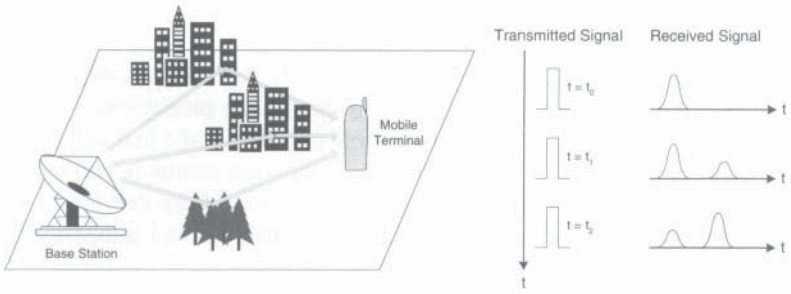


Figure 1.5 Illustration of the multipath and time-varying dimensions in microscopic fading.

is defined as the range of multipath components with *significant* power when an impulse is transmitted as illustrated in Figure 1.6 In Figure 1.6(a), a narrow pulse (in time) at the transmitter results in a spread of energy across the delay dimension. The graph of received power versus delay is called the *power-delay profile* and is used to measure the delay spread experimentally. Typically, the delay spread in enclosed indoor environment is smaller than the outdoor environment because the range of delays in the multipaths are more contained in the indoor environment.

On the other hand, to quantify the multipath dimension [167], an equivalent parameter, namely the *coherence bandwidth*, can be used. Figure 1.6(b) illustrates the concept of *coherence bandwidth*. Suppose at the transmitter, we transmit two single tone signals (impulse in frequency domain) with unit power and frequencies f_1 and f_2 respectively. At the receiver, the two single tones will be received with received powers α_1 and α_2 respectively. α_1 and α_2 represent the random channel attenuation (or fading) experienced by the two single-tone signals. It is found that if $|f_1 - f_2| > B_c$, the attenuation α_1 and α_2 are uncorrelated. Otherwise, the attenuation α_1 and α_2 are correlated. Hence, the minimum separation of frequency such that uncorrelated fading is resulted is called the *coherence bandwidth*, B_c. It is found that for 0.5 correlation coefficient between α_1 and α_2, the coherence bandwidth is related to the delay spread σ_τ by:

$$B_c \approx \frac{1}{\sigma_\tau}$$

Typical values of the coherence bandwidth for indoor and outdoor environment are 1MHz and 100kHz respectively. Note that delay spread and coherence bandwidth are two sides of the same coin to characterize the multipath dimension of microscopic fading.

Time Varying dimension: To characterize the time variation dimension, we can have either the *Doppler spread* or *coherence time*. Doppler spread [167] refers to the spread of frequency introduced by the fading channel when we transmit a single tone signal (narrow pulse in frequency domain) as illustrated in Figure 1.7(a). The spread in frequency is due to the mobility between the transmitter and the receiver or the mobility of the surrounding obstacles. The maximum Doppler spread is given by:

$$f_d = \frac{v}{\lambda}$$

where λ is the wavelength of the signal and v is the maximum speed between the mobile and the base station. Typical values of Doppler spread for pedestrian users (5km/hr) and vehicular users (100km/hr) at carrier frequency of 2.4 GHz are given by 14 Hz and 300 Hz respectively.

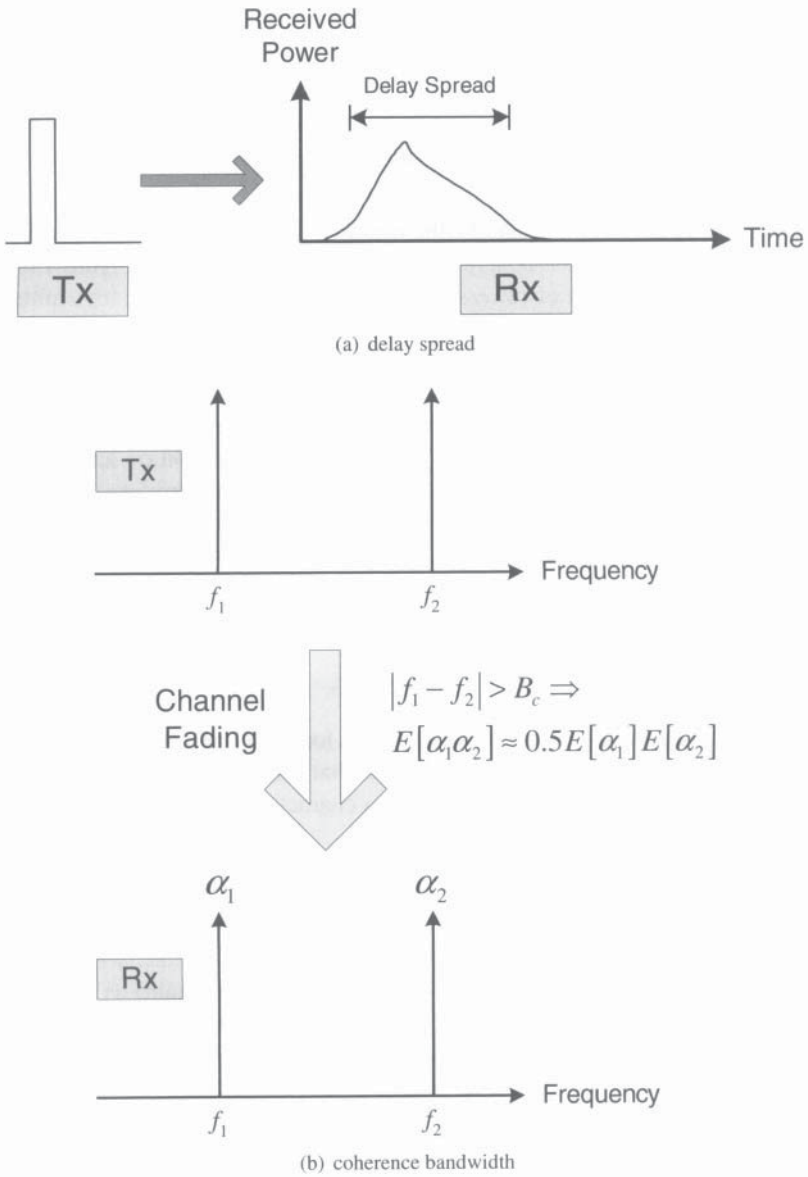


Figure 1.6 Illustration of the delay spread and coherence bandwidth.

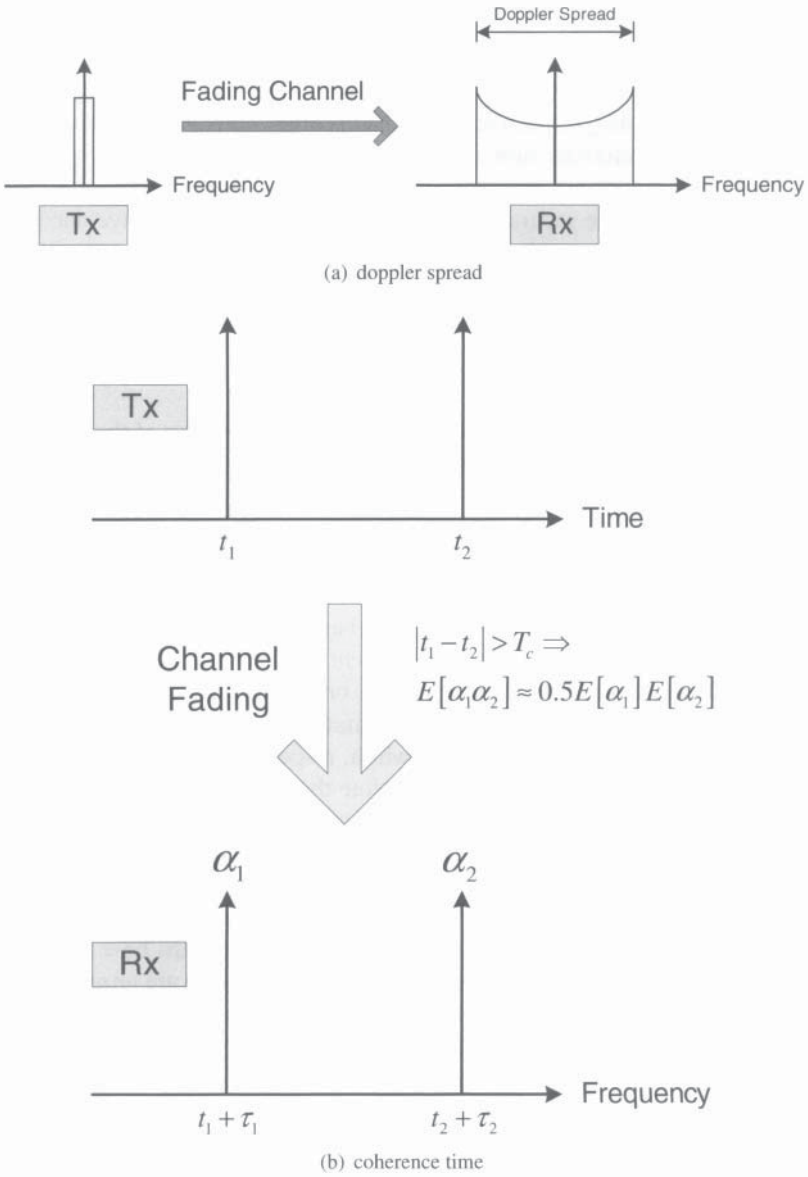


Figure 1.7 Illustration of the doppler spread and coherence time.

Similarly, an equivalent parameter to quantify the time variation dimension of microscopic fading is given by the *coherence time*. Figure 1.7(b) illustrates the concept of coherence time. Suppose we transmit two narrow pulses in time domain at times $t = t_1$ and $t = t_2$ with unit energy. After channel fading, the corresponding signals will be received at $t = t_1 + \tau_1$ and $t = t_2 + \tau_2$ with energies α_1 and α_2 respectively. Similar to coherence bandwidth, if we separate the two transmissions sufficiently far in time, ($|t_1 - t_2| > T_c$), the channel fading α_1 and α_2 are uncorrelated. Hence, the coherence time is defined as the minimum time separation at the transmitter in order to have uncorrelated fading at the receiver. It is found that for 0.5 correlation coefficient between α_1 and α_2, the coherence time T_c is related to the Doppler spread f_d by:

$$T_c \approx \frac{1}{f_d}$$

Note that Doppler spread and coherence time are two sides of the same coin to characterize the time variation dimension of microscopic fading.

1.4.1 Flat Fading vs Frequency Selective Fading

Based on the delay spread and the transmitted symbol duration, (or equivalently, the coherence bandwidth and the transmitted bandwidth), the multipath fading can be classified as *frequency flat fading* or *frequency selective fading* [287]. A fading channel is said to be *frequency flat* if $\sigma_\tau < T_s$ or equivalently, $B_c > W_{tx}$ where T_s and W_{tx} are the symbol duration and the transmitted signal bandwidth, σ_τ and B_c are the delay spread and the coherence bandwidth, respectively. Otherwise, the fading channel is said to be *frequency selective*. Note that whether the transmitted signal will experience frequency flat fading or frequency selective fading depends on both the *channel parameters* (such as σ_τ, B_c) as well as the *transmitted signal parameters* (such as T_s and W_{tx}).

When a transmitted signal experience frequency flat fading, the received signal consists of superposition of multipath signals but the multipaths have delays much smaller than the symbol duration T_s and hence, the multipaths are *unresolvable*. As a result of the unresolvable multipath, the received pulse shape will be different from the transmitted pulse shape but the received pulse appears as one single pulse at the receiver. The received signal is modeled as:

$$y(t) = h(t)x(t - \tau) + z(t)$$

where $h(t)$ is the time-varying channel fading which is modeled as complex Gaussian random variable with zero mean and unit variance, $z(t)$ is the complex AWGN noise and $x(t)$ is the low-pass equivalent complex transmitted signal.

On the other hand, when a transmitted signal experience frequency selective fading, the received signal consists of multipaths with delay » symbol duration (T_s). Hence,

the received signal consists of pulses at *resolvable delays* (integral multiples of T_s). In general, the number of resolvable multipaths is given by:

$$L_p = \lceil \frac{\sigma_\tau}{T_s} \rceil = \lceil \frac{W_{tx}}{B_c} \rceil$$

The received signal for frequency selective fading channel is modeled as:

$$y(t) = \sum_{i=1}^{L_p} h_i(t)x(t - iT_s) + z(t)$$

where $h_i(t)$ is the time varying channel fading for the i-th resolvable multipaths which is modeled as complex Gaussian random variable with zero mean and unit variance.

1.4.2 Fast Fading vs Slow Fading

Similarly, based on the Doppler spread and the transmitted bandwidth (or equivalently the coherence time and the symbol / frame duration), the fading channel is classified as *fast fading* or *slow fading*. A fading channel is classified as *fast fading* if the symbol duration T_s is larger than the coherence time T_c. Otherwise, the fading channel is called *slow fading*. Note that whether a transmitted signal experiences fast fading or slow fading depends on both the channel parameters (such as the Doppler spread or coherence time) and the transmitted signal parameters (such as the transmitted bandwidth or the symbol duration). Hence, for slow fading channels, the channel fading coefficients $h_i(t)$ remains to be quasi-static within a symbol duration T_s.

Note that sometimes, we define slow fading to be the case when T_c (coherence time) is larger than T_f (frame duration) instead of symbol duration.

1.5 PRACTICAL CONSIDERATIONS

In this section, we briefly discuss the implications of path loss, shadowing and microscopic fading components on system design. In fact, the design and performance of communication systems depends heavily on the underlying channels. For instance, path loss exponent determines how fast the received signal strength attenuates with respect to distance separation between the transmitter and receiver. For a point to point digital link, higher path loss exponents results in faster signal attenuation and therefore is undesirable. This is because for the same transmitter and receiver design, a higher path loss exponent results in shorter communication range (for the same transmit power) or higher required transmit power (for the same distance) to overcome the higher path loss. This holds in general for *noise-limited systems*. However, if we consider multicell systems such as cellular networks like GSM, or CDMA systems, higher path loss exponent results in more confined interference (as the interference

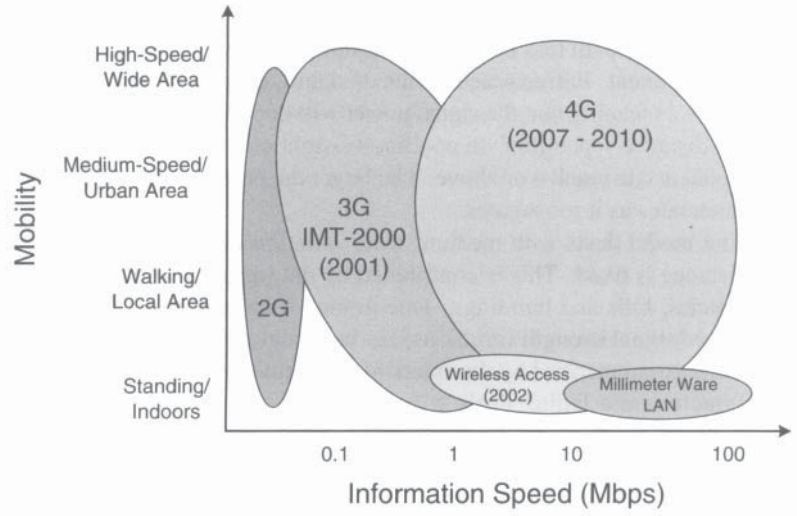
signal cannot propagate very far) and this translates into higher system capacity because more aggressive resource reuse can be realized. In general, higher path loss exponent is desirable for *interference-limited systems*. We will elaborate the cellular system designs in Chapter 3.

Shadowing introduces randomness in the coverage of cellular systems. For instance, if the standard derivation of the shadowing component is large, the average received signal power at the cell edge will have large fluctuations. To achieve a certain Quality of Service (QoS) such as 90% probability that the received signal strength at the cell edge is above a target threshold, higher power or shorter cell radius is needed to allow for some shadowing margins to satisfy the QoS target. Hence, shadowing with larger standard derivation is undesirable.

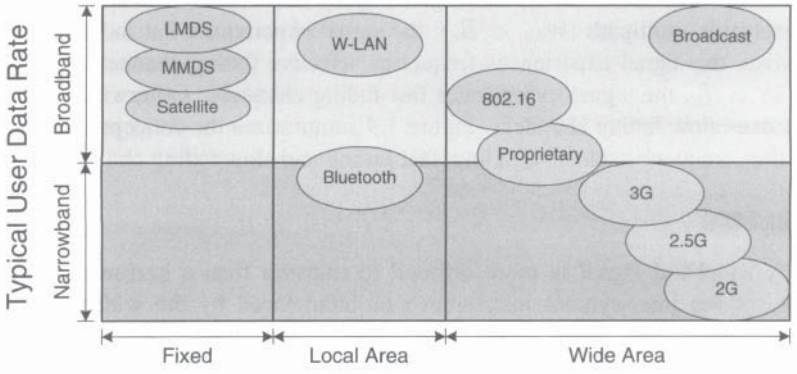
Finally, the effects of microscopic fading have high impact on the physical layer design of communication systems. For instance, by common sense, it is more challenging to setup a wireless link to transmit high quality video then audio signals. This is because for the same environment, transmitting a video signal generally involves a higher transmit bandwidth, W_{tx}, relative to an audio signal. From Section 1.4.1, it is very likely that the video signal will experience frequency selective fading while the audio signal will experience frequency flat fading. As we elaborate in Chapter 4, frequency selective fading will introduce *intersymbol interference* (ISI) and this induces irreducible error floor. Hence, complex equalization at the receiver is needed. On the other hand, for audio signal, since $W_{tx} << B_c$, the signal will experience flat fading only and no equalization at the receiver is needed. Hence, this results in more simple design. Another common-sense example is that it is easier to setup a wireless link for indoor environment than an outdoor environment to transmit the same signal. This again can be analyzed by the frequency selective or frequency flat fading channels. For the same transmit signal, the indoor environment in general has a larger coherence bandwidth and hence, the number of resolvable multipaths in indoor environment will be small. On the other hand, the coherence bandwidth for outdoor environment in general will be small and this results in larger number of resolvable multipaths (frequency selective fading). Hence, more complicated designs are needed in the latter case.

In general, it is imperative to understand the target operating environment in order to optimize the transmitter and receiver design of the communication systems. There is no such thing as a *universal communication systems* that is effective for all communication environments. For examples, both wireless LAN systems and 3G networks are designed to offer high bit rate. For IEEE 802.11g system, the highest bit rate is 54Mbps. For UMTS systems (Rel 99), the highest bit rate is just 2Mbps. Obviously, the cost of the UMTS system is much higher than the wireless LAN systems. Hence, one might query about why we still require UMTS system when wireless LAN systems cost only less than US$100. Figure 1.8 illustrates a more comprehensive comparison between Wi-Fi and UMTS systems. We do not just look at the maximum bit rate offered by both systems but we also have to compare the target environment of operations. For instance, while Wi-Fi systems offer higher bit rate,

they are designed to operate at pedestrian mobility and indoor coverage. For 3G/4G cellular systems, although their maximum bit rate is lower than Wi-Fi systems, they are designed to operate at very high mobility and macroscopic coverage.



(a) mobility



(b) coverage

Figure 1.8 A comparison between WiFi systems and 3G cellular systems.

1.6 SUMMARY

In this chapter, we have elaborated on the modeling of wireless channels. We have introduced the 3 levels of channel models, namely the *large scale path loss model*, the *medium scale shadowing model* as well as the *microscopic scale fading model*. The path loss model deals with the variation of received signal strength with respect to variation of distance between the transmitter and the receiver and is focused on the time scale of seconds. The path loss between a transmitter and a receiver is characterized by a path loss exponent. In free space or line of sight propagation condition, the path loss exponent is 2 meaning that the signal power will be reduced by 4 times for every double in the distance separation. In non line-of-sight propagation environment, the path loss exponent can reach 4 or above. The larger the path loss exponent, the faster the signal attenuates as it propagates.

Shadowing model deals with medium scale variation of received signal strength when the distance is fixed. This is contributed by the variation in the terrain profile such as obstacles, hills and buildings. Due to law of large number, the shadowing effects (received signal strength variations) can be modeled by *log-normal shadowing* component and parameterized by the standard derivation $\sigma(dB)$.

Finally, microscopic fading deals with small scale variation of received signal strength due to constructive and destructive multipath superpositions. The time scale of interests can be of millisecond order. Microscopic fading can be parameterized by the *delay spread* (*coherence bandwidth*) for the multipath dimension as well as the *Doppler spread* (*coherence time*) for the time variation dimension. Note that the multipath dimension and the time variation dimension are two *independent* dimensions. The number of resolvable multipaths is given by $L_p = \lceil W_{tx}/B_c \rceil$. When there is only one resolvable multipath ($W_{tx} < B_c$), the signal experiences flat fading channels. Otherwise, the signal experiences frequency selective fading channels. Similarly, when $T_s > T_c$, the signal experiences fast fading channels. Otherwise, the signal experiences slow fading channels. Figure 1.9 summarizes the concept of frequency flat fading, frequency selective fading, fast fading and slow fading channels.

PROBLEMS

1.1 A wideband signal is more difficult to transmit than a narrowband signal because of the inter-symbol interference problem faced by the wideband signal. [True/False]

1.2 In flat fading channels, there is no multipath effect at all as there is only a single resolvable echo at the receiver. [True/False]

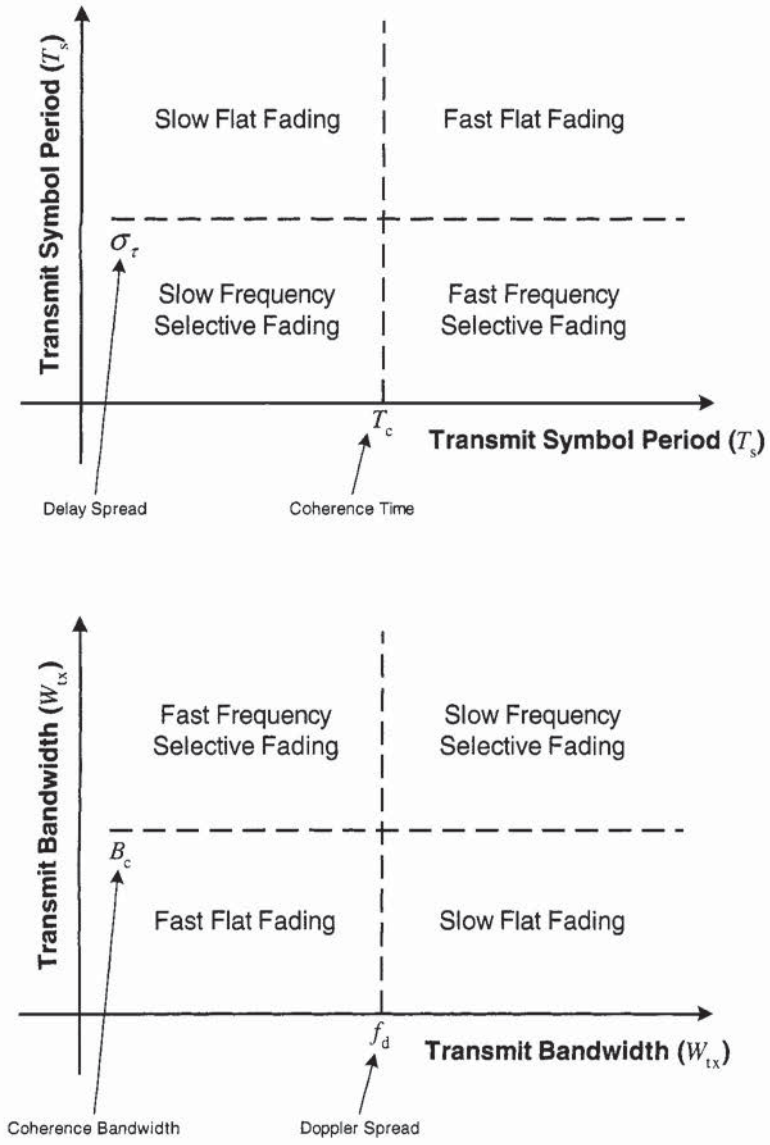


Figure 1.9 Summary of classifications of fading channels.

This Page Intentionally Left Blank

CHAPTER 2

MODULATION TECHNIQUES

The whole earth was of one language, and of one speech.

—Genesis XI, 1, c. 700 B.C.

2.1 INTRODUCTION

In this chapter, we introduce a generic model for digital modulation and demodulation design based on the concept of signal space. We focus on the point-to-point communications where there is a single transmitter, a single receiver, and a channel connecting the transmitter and receiver, as illustrated in Figure 2.1

Performance:

In Chapter 1, we have discussed various channel models for wireless media. In general, the design of the transmitter and the receiver has to match the characteristics of the channels. There does not exist a *universal communication system* that is optimized in all different operating environments. In this chapter, we elaborate on the design of digital modulator and digital demodulator for various wireless channels.

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau

Copyright © 2007 John Wiley & Sons, Inc.



Figure 2.1 Point-to-point communication link.

To study the digital communication design in Figure 2.1, we discuss and quantify the tradeoff between the performance dimensions and the cost dimensions. Specifically, the performance dimension of a digital communication link refers to *bit rate* and *bit error rate* (or packet error rate). Bit rate, which is defined as the number of bits delivered to the receiver per second (bits / sec), is a measure of how fast information bits are transferred in the communication link. Obviously, from a user perspective, we would like the bit rate to be as high as possible. Bit error rate (BER), on the other hand, is a measure of the *reliability* of the communication link. BER is defined as the number of bits in error over the total number of bits transmitted. From a user perspective, we would like the BER to be as low as possible.

On the other hand, the cost dimension of a digital communication link refers to the *transmit power* and *channel bandwidth*. Obviously, a communication link that requires a higher power or larger channel bandwidth will be most costly to operate. Hence, a very important focus of this chapter is to understand and study the detail dynamics behind the tradeoff of the *performance* and *cost* parameters. The tradeoff relationship is summarized in Figure 2.2

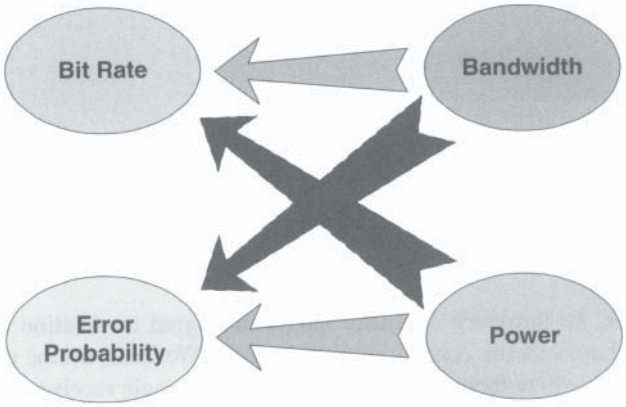


Figure 2.2 Tradeoffs of the performance and cost in digital communication systems.

We first introduce the concept of signal space[287] and how to represent time domain signals using vectors in the signal space[147] in Section 2.2. We then discuss modulation techniques in Section 2.3. Next, we consider the design and performance of detection techniques over a simple *additive white Gaussian noise* channel (AWGN)

in Section 2.4. Finally, we give some practical illustrations on the application of digital modulations in real-life systems in Section 2.6. We conclude with a brief summary of main points in Section 2.7.

2.2 SIGNAL SPACE AND GEOMETRIC REPRESENTATION OF SIGNALS

While the physical transmitted signals and received signals are all in continuous time domain, it is quite difficult to gain proper design insights by working in the original forms. For example, in digital communication receiver, the primary task is to deduce what has been transmitted at the transmitter given a noise observation (or received signal) $y(t)$. It is not very obvious on what should be the optimal structure to extract the information bits carried by the signal. However, as we illustrate, if we represent the time domain signals as points or *vectors* on a *signal space*, the design of the digital receiver will be much easier. Before we discuss the detail operation of the digital modulator and digital receiver, we try to introduce the concept of *signal space* and how to represent continuous time signals as *points* or *vectors* in the signal space.

If we want to discuss the coordinate of a point, say $(1, 2)$, in Euclidean space, we have to first establish what is the *frame of reference* or equivalently, what is the x-axis and y-axis. In other words, before knowing the coordination reference, it is meaningless to discuss the coordinate of a point. Similarly, when we say we want to represent a time domain signal $x(t)$ by a point $\vec{x}$ in the signal space, we mean expressing the signal $x(t)$ into a coordinate $\vec{x} = (x_1, \ldots, x_D)$ where D is the dimension of the signal space. However, for similar reasons, before we can discuss the coordinate of a point, we have to first establish the *frame of reference* or *axis* of the signal space.

In fact, signal space and Euclidean space are two specific examples of vector space. Recall that a vector space is a collection of objects that satisfies a number of *axioms*[258]. Hence, signal space is simple a vector space where the *objects* or the vectors are time domain signals. We define the *dot-product* or *inner product* between two signals $x(t)$ and $y(t)$ (with duration T_s) in the signal space as:

$$< x(t), y(t) >= \int_0^{T_s} x(t)y^*(t)dt. \tag{2.1}$$

To specify a $D-$dimensional signal space, we need D orthonormal basis functions $\{\phi_1(t), \ldots, \phi_D(t)\}$ where

$$\langle \phi_i(t), \phi_j(t) \rangle = \begin{cases} 1 & \text{if} i = j \\ 0 & \text{otherwise} \end{cases}$$

Note that the D *basis functions* in signal space play the same role as the D *axis* in Euclidean space.

Suppose we would like to represent any M time domain signals $\{x_1(t), \ldots, x_M(t)\}$ with finite symbol duration T_s into M vectors $\{\vec{x}_1, \ldots, \vec{x}_M\}$ in a signal space. We have to go through the following two steps.

Step 1. Find out the basis functions: The first step is to find out the basis functions or frame of reference of the underlying signal space that contains the M signals. It can be shown that given any M time domain signals $\{x_1(t), \ldots, x_M(t)\}$ with finite duration T_s, there exists a D dimensional signal space with $D \leq M$ that can contain these M signals. The corresponding orthonormal basis $\{\phi_1(t), \ldots, \phi_D(t)\}$ for the D-dimensional signal space can be obtained by the *Gram-Schmidt Orthogonal Procedure*[283] as illustrated below.

(i): The first basis function $\phi_1(t)$ is given by:

$$\phi_1(t) = x_1(t)/\sqrt{E_1}$$

where $E_1 = \int_0^{T_s} |x_1(t)|^2 dt$.

(ii): The second basis function $\phi_2(t)$ is obtained by first computing the projection of $x_2(t)$ onto $\phi_1(t)$, which is:

$$c_{1,2} = \int_0^{T_s} x_2(t)\phi_1^*(t)dt.$$

Then $c_{1,2}\phi_1(t)$ is subtracted from $x_2(t)$ to yield

$$f_2(t) = x_2(t) - c_{1,2}\phi_1(t).$$

Note that $f_2(t)$ is orthogonal to $\phi_1(t)$ because $< f_2(t), \phi_1(t) >= 0$. Normalizing $f_2(t)$ to unit energy signal, we have:

$$\phi_2(t) = f_2(t)/\sqrt{E_2}$$

where $E_2 = \int_0^{T_s} |f_2(t)|^2 dt$.

(iii): In general, the k-th orthogonal basis $\phi_k(t)$ is given by:

$$\phi_k(t) = f_k(t)/\sqrt{E_k}$$

where

$$f_k(t) = x_k(t) - \sum_{i=1}^{k-1} c_{i,k}\phi_i(t)$$

and

$$c_{i,k} =< x_k(t), \phi_i(t) >= \int_0^{T_s} x_k(t)\phi_i^*(t)dt.$$

The orthogonalization process continues until all the M signals $\{x_1(t), \ldots, x_M(t)\}$ have been exhausted and $D \leq M$ orthonormal basis functions have been constructed.

Step 2. Find out the coordinates of the M points: After the D orthonormal basis $\{\phi_1(t), \ldots, \phi_D(t)\}$ have been found from Step 1, we can determine the coordinates of the M vectors. For instance, let $\vec{x}_m = (x_{m1}, x_{m2}, \ldots, x_{mD})$ be the coordinates of the $m-$th point. The j-th coordinate of the m-th point is given by:

$$x_{mj} =< x_m(t), \phi_j(t) >= \int_0^{T_s} x_m(t)\phi_j^*(t) \tag{2.2}$$

Example 1 (Geometric Representation of 4 time domain signals) *Let $x_1(t)$,$x_2(t)$, $x_3(t)$,$x_4(t)$ be 4 time domain signals with duration $T_s = T$ as illustrated in Figure 2.3*

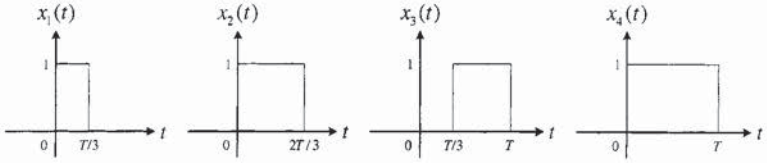


Figure 2.3 Illustration of the time domain waveforms of the 4 signals in example 1.

Following the approach in step 1, we requires a 3-dimensional signal space to contain all the 4 signals. The basis functions of the signal space $\{\phi_1(t), \phi_2(t), \phi_3(t)\}$ are illustrated in Figure 2.4

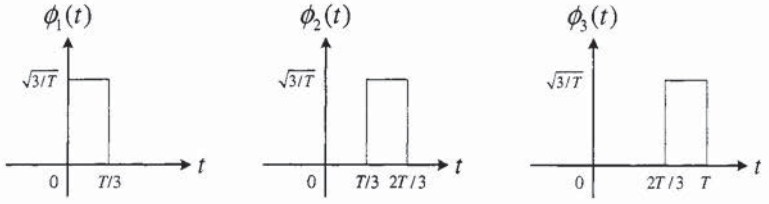


Figure 2.4 Basis functions of the signal space in example 1.

Following the step 2, the coordinates of the 4 time domain signals are given by $\vec{x}_1 = (\sqrt{T/3}, 0, 0)$, $\vec{x}_2 = (\sqrt{T/3}, \sqrt{T/3}, 0)$, $\vec{x}_3 = (0, \sqrt{T/3}, \sqrt{T/3})$ and $\vec{x}_4 = (\sqrt{T/3}, \sqrt{T/3}, \sqrt{T/3})$ as illustrated in Figure 2.5

Hence, given any M time domain signals, we can find a signal space of at most M dimension to represent these M signals as M points. Notation-wise, we denote such process as

$$x_m(t) \longrightarrow \vec{x}_m = (x_{m1}, \ldots, x_{mD}).$$

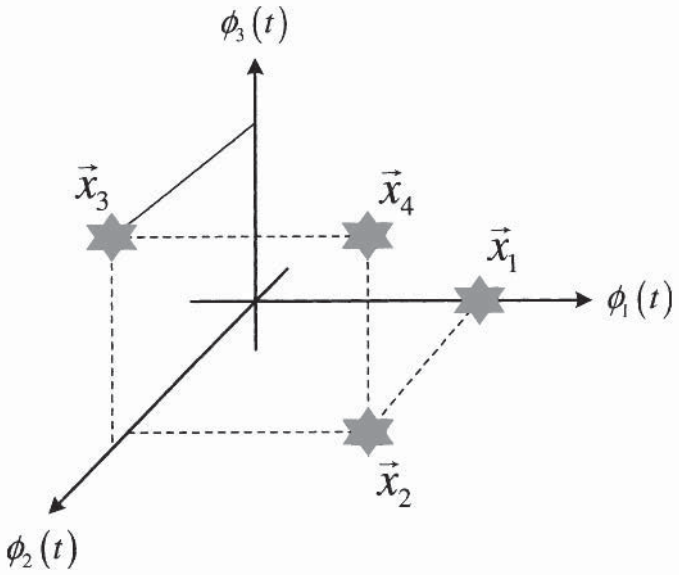


Figure 2.5 Geometric representation of the 4 signals in example 1.

One point to emphasis is that nomatter how we describe the M signals (either in time domain or in geometric domain), the physical nature of the signal is not changed. For example, the energy of a signal is defined in time domain as:

$$E = \int_0^{T_s} |x(t)|^2 dt \tag{2.3}$$

However, by the definition of the *dot-product* in the signal space in (2.1), we have

$$< x(t), x(t) >= \int_0^{T_s} |x(t)|^2 dt = E = \|\vec{x}\|^2. \tag{2.4}$$

In other words, the energy of a signal $x(t)$ is also given by the square of the length of the vector $\vec{x}$. A signal point further away from the origin implies higher energy. Similarly, the power of a signal in time domain is defined as:

$$P = \frac{1}{T_s} \int_0^{T_s} |x(t)|^2 dt \tag{2.5}$$

Using the definition of dot-product, we have the power of the signal also given by:

$$P = \frac{1}{T_s} \|\vec{x}\|^2. \tag{2.6}$$

2.3 MODULATION DESIGN AND SIGNAL CONSTELLATIONS

In Section 2.2, we have introduced the concept of representing time domain signals as *points* or *vectors* on a signal space. We apply this *geometric representation* of signals to model the digital communication link design. In this section, we elaborate on the digital transmitter (or modulator). A digital modulator[296] can be modeled by a *black box* with information bits $\{b_1, b_2, \ldots\}$ as inputs and modulation symbols $s(t)$ as output. The digital modulator is parameterized by:

Symbol Duration T_s: The symbol duration T_s specifies how often the modulator produces a *modulation symbol*. The *baud rate* or *symbol rate* of a digital modulator is defined as the number of modulation symbols per second, which is given by:

$$R_{baud} = \frac{1}{T_s} \tag{2.7}$$

The baud rate or symbol rate is a very important parameter because it determines the required channel bandwidth of a digital modulator.

Signal Set $\mathcal{S}$: The signal set of a digital modulator is a collection of M time domain signals given by:

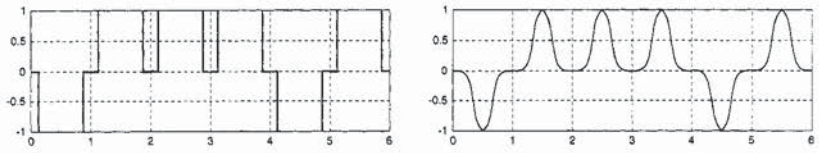
$$\mathcal{S}_M = \{s_1(t), \ldots, s_M(t)\} \tag{2.8}$$

for some integer M. Each of the signal element in the signal set $\mathcal{S}$ is labeled with a unique bit pattern which will be explained below.

The operation of a $M - ary$ digital modulator is very simple. For simplicity, we consider the *memoryless* modulator where the encoding of information bits is done on a *symbol-by-symbol* basis. At the n-th symbol duration, the digital modulator accepts $\log_2(M)$ bits as input. Depending on the bit pattern of the input bits, one of the M modulation signals (with exactly the same label as the input bit pattern) in the signal set $\mathcal{S}_M$ is chosen as the modulation symbol output. For example, when $M = 2$ (binary modulator), the modulator accepts one single bit on every symbol duration. The signal set $\mathcal{S}_2 = \{s_1(t), s_2(t)\}$ has the bit label as $s_1(t)$ associated with bit value = 0 and $s_2(t)$ associated with bit value = 1. When the input bit is 0, $s_1(t)$ will be produced at the modulator output during the current symbol duration and vice versa. After *modulation process* is completed for the current symbol duration, we proceed to the next symbol duration and fetch a new information bit for modulation. Figures 2.6(a) and 2.6(b) illustrate the output waveforms of two binary modulators with different signal sets. In both binary modulators, the symbol durations are $T_s = 1$. In the first binary modulator, the signal set is given by $\mathcal{S}_2(A) = \{s_1^{(A)}(t), s_2^{(A)}(t)\}$ which is illustrated in Figure 2.7(a). In the second binary modulator, the signal set is given by $\mathcal{S}_2(B) = \{s_1^{(B)}(t), s_2^{(B)}(t)\}$ which is illustrated in Figure 2.7(b).

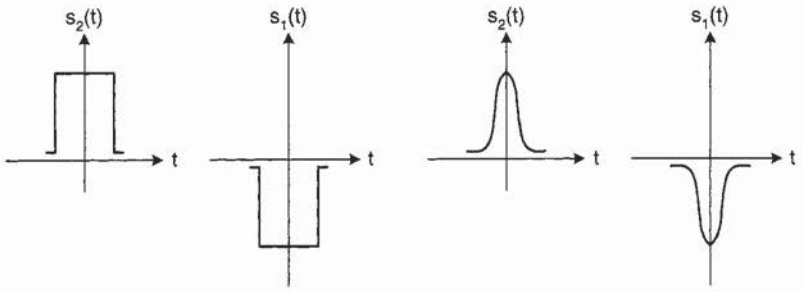
Note that for binary modulators, the bit rate R_b is given by:

$$R_b = \frac{1}{T_b} = R_{baud} \tag{2.9}$$



(a) binary modulator A (b) binary modulator B

Figure 2.6 Illustration of Binary Modulations.



(a) binary modulator A (b) binary modulator B

Figure 2.7 Illustration of the Signal Sets of the Binary Modulators A and B.

where T_b denotes the bit duration. In other words, the bit rate is the same as baud rate for binary modulators. Hence, it is obvious that increasing the channel bandwidth W_{tx} can result in an increase of information bit rate as indicated in the tradeoff diagram in Figure 2.2

On the other hand, for 4-ary modulator ($M = 4$), the signal set $\mathcal{S}_4 = \{s_1(t),\ldots,s_4(t)\}$ consists of 4 distinct signal elements which are labeled as $\{00, 01, 10, 11\}$. During each symbol duration, 2 bits of information is fetched and modulated. Depending on the specific bit pattern observed, the signal with the corresponding label will be produced at the output. Similarly, we can extend the operation of a general M-ary modulator by the *black box* model. For $M-$ary modulator, since one symbol can be considered as carrying $\log_2 M$ information bits, the bit rate is given by:

$$R_b = \frac{1}{T_b} = \log_2 M \frac{1}{T_s}. \tag{2.10}$$

On the other hand, the transmission bandwidth of a digital modulator with signal set $\mathcal{S}$ is in general given by [287]:

$$W_{tx} \propto \text{bandwidth} \left[\frac{1}{M} \sum_{m=1}^{M} |S_m(f)|^2 \right] \tag{2.11}$$

where $|S_m(f)|$ is the amplitude spectrum of the m-th signal $s_m(t)$ in the signal set and W_{tx} denotes the required channel bandwidth. For example, if the signal set $\mathcal{S}$ consists of *baseband signals* (all with pulse duration T_s), the required transmission bandwidth (one-sided lowpass bandwidth) is given by:

$$W_{tx} = (1 + \alpha)\frac{1}{2T_s} \tag{2.12}$$

where $\alpha \geq 0$ is a constant determined by the *pulse shape* of the modulation symbol. If the signal set $\mathcal{S}$ consists of *bandpass signals* (all with pulse duration T_s), the required transmission bandwidth (one-sided bandpass bandwidth) is given by:

$$W_{tx} = (1 + \alpha)\frac{1}{T_s} \tag{2.13}$$

Finally, the average transmit power of the modulator (assuming equiprobable input information bit) is given by:

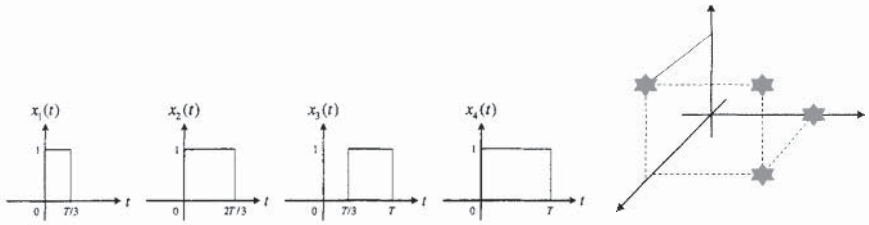
$$P_{tx} = \frac{1}{MT_s}\|\vec{s}_m\|^2 = \frac{1}{MT_s}\int_0^{T_s} |s_m(t)|^2 dt.$$

Hence, intuitively, the mean square distance of the constellation points from the origin determines the average transmit power of the digital modulator.

From Equation (2.10), it seems that given any fixed channel bandwidth W_{tx} (and hence a fixed baud rate or symbol duration T_s), one can increase the bit rate to an

arbitrarily high value by increasing M. The analogy is that modulation symbols are like mini-bus. The bandwidth or baudrate determines the *frequency* of mini-buses (number of mini bus per second). However, the M determines the number of *seats* in a mini bus. Hence, given a fixed mini bus frequency, one can still increase the bit rate (number of passengers per second) by increasing the number of seats in a mini bus ($\log_2 M$). This is obviously too good to be true as otherwise, bandwidth will be worthless. We will revisit this point when we discuss the operation of the demodulator to get a complete tradeoff picture.

Finally, as we have mentioned, to specify the behavior and operation of a modulator, we just need to specify the symbol duration and the signal set. For the signal set, we can choose to represent it using time domain approach (draw out the waveforms of the M signal elements). Equivalently, we can represent the signal set using *geometric domain* approach. In that case, the M signal elements are represented as M points on the signal space $\mathcal{S}_M = \{\vec{s}_1, \ldots, \vec{s}_M\}$. If we draw out the M points on the signal space, this is called the *constellation diagram* of the digital modulator. Figure 2.8 illustrates the time domain (signal waveforms) and geometric domain (signal constellation) of a quaternary modulator.



(a) time domain (b) geometric domain

Figure 2.8 Illustration of the time domain and geometric domain representations of the signal set.

Example 2 (Signal Constellation of BPSK and BASK) *Consider two binary modulators, namely the Binary Phase Shift Keying (BPSK) and the Binary Amplitude Shift Keying (BASK) modulators. The signal outputs and the signal constellations are illustrated in Figures 2.9 and 2.10*

2.4 DEMODULATION DESIGN AND OPTIMAL DETECTION IN AWGN CHANNELS

In Section 2.3, we focused on the modelling of the digital modulator. In order to have a complete picture of the tradeoff in Figure 2.2, we discuss the design of the digital demodulator in this section. The primary task of the digital demodulator is

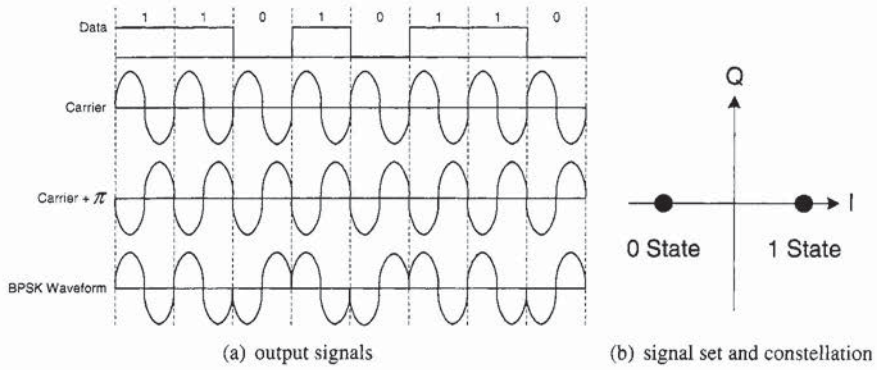


Figure 2.9 Illustration of BPSK modulation and the constellation.

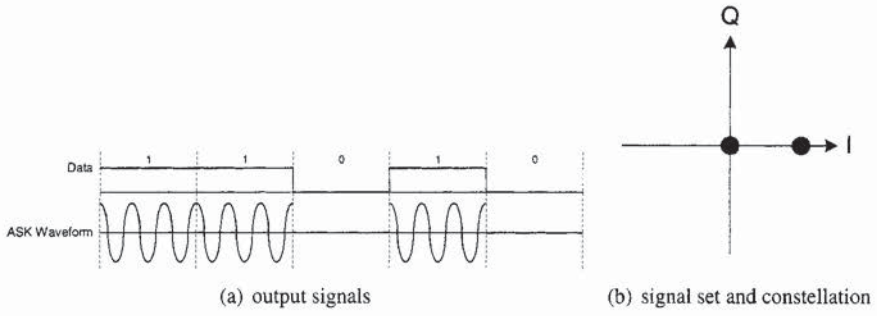


Figure 2.10 Illustration of BASK modulation and the constellation.

to detect what was transmitted at the digital modulator based on noisy or distorted observations $y(t)$. Depending on the specific media, the received signal $y(t)$ will be distorted in different ways as discussed in Chapter 1. In the following, we consider a simple example of Additive White Gaussian Noise (AWGN) channel. The received signal $y(t)$ is given by:

$$y(t) = x(t) + z(t)$$

where $x(t)$ is the transmitted signal and $z(t)$ is the white Gaussian noise. Figure 2.11 illustrates a segment of the received signal for a binary modulator after transmission over the AWGN channel.

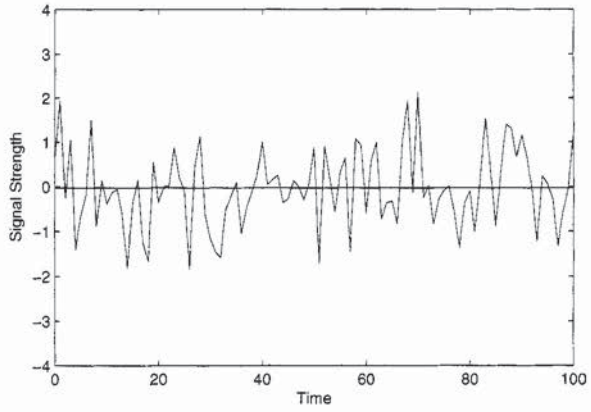


Figure 2.11 Illustration of received signal $y(t)$ of a binary modulator after transmission over the AWGN channel.

From the time domain perspective, there is little insight on what should the demodulator do in order to extract the information bits from the noisy observation $y(t)$. Moreover, it is difficult to deduce what should be the *optimal* detection rule for the demodulator. In the followings, we illustrate the demodulation process using geometric domain perspective.

Assume that symbol synchronization[1] has been achieved before passing the received signal to the demodulator. Using the concept of signal space, the received signal during the n-th symbol duration can be expressed as:

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{z}_n$$

where $\mathbf{x}_n$ is the transmitted signal vector with respect to the signal space $\mathcal{X}$ that defines the constellation of the modulator, $\mathbf{z}_n$ is the noise vector of the channel

[1]Symbol synchronization refers to the timing synchronization with respect to the symbol boundary. In other words, after achieving symbol synchronization, the demodulator is able to partition the received signal into multiple symbols, each with duration T_s.

noise $z(t)$ projected[2] onto the signal space $\mathcal{X}$ and $\mathbf{y}_n$ is the received signal vector of $y(t)$ projected onto the signal space $\mathcal{X}$. For Gaussian noise, the noise vector $\mathbf{z}_n$ is a random vector with uniformly distributed orientation and Rayleigh distributed length. Since the square of the length of the signal vector represents signal energy, hence, the stronger the channel noise is, the longer the average length of the noise vector will be. Without loss of generality, consider a quaternary modulator with signal set $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_4\}$. Assume that $\mathbf{x}_n = \mathbf{s}_1$ is the actual transmitted signal. If the channel is noiseless, the received signal $\mathbf{y}_n$ must always be given by $\mathbf{s}_1$. However, due to the random nature of the noise vector $\mathbf{z}_n$, the received signal $\mathbf{y}_n$ will fall within a *noise cloud* about the actual transmitted signal $\mathbf{s}_1$. This is illustrated in Figure 2.12.

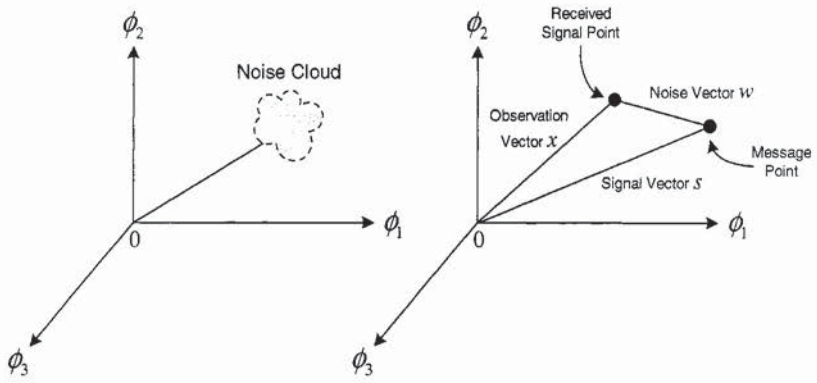


Figure 2.12 Illustration of noise cloud on the received signal.

Suppose now, the received signal $\mathbf{y}_n$ is at the position as illustrated by Figure 2.13 Note that due to quartenary modulator, the demodulator realizes that only one of the four possible positions, $\{\mathbf{s}_1, \ldots, \mathbf{s}_4\}$, could have been transmitted. Given that the current received vector is $\mathbf{y}_n$, a simple and intuitive scheme is to pick the decoded signal $\hat{\mathbf{x}}$ as the point that is *closet* to the received vector $\mathbf{y}_n$. In the example illustrated in Figure 2.12, the demodulator should pick the point $\mathbf{s}_1$ as the detected symbol $\hat{\mathbf{x}}_n$. This detection scheme is called the *minimum distance detection*.

In general, the minimum distance detector for a M-ary modulator is summarized below.

Minimum distance detector:

[2]Strictly speaking, the vector $\mathbf{z}_n$ and $\mathbf{y}_n$ are not exact geometric representations of $z(t)$ and $y(t)$ because an infinite dimension signal space is required to represent $z(t)$ and $y(t)$. $\mathbf{z}_n$ and $\mathbf{y}_n$ are vectors with finite dimension on a signal sub-space $\mathcal{X}$ that contains the transmitted constellation only. Since the null space of $\mathcal{X}$ contains no useful information on the transmitted bits, it is sufficient to consider the projection of the noise and received signals onto $\mathcal{X}$ only.

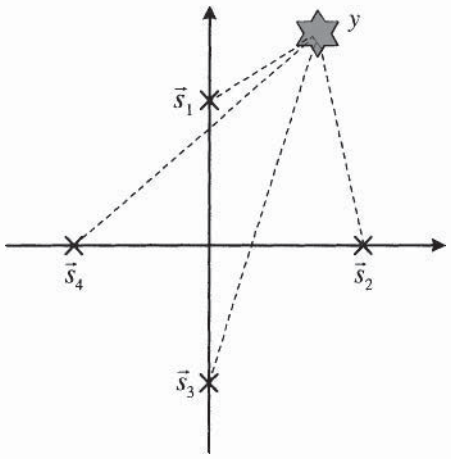


Figure 2.13 An example of the received signal $\mathbf{y}_n$ on a two-dimensional signal space.

Step 1: Convert the time domain received signal $y(t)$ to the received vector $\mathbf{y}_n$. As mentioned in the previous section, suppose the constellation signal space $\mathcal{X}$ (with D dimension) has basis $\{\phi_1(t), \ldots, \phi_D(t)\}$. The j-th coordinate of the received vector $\mathbf{y}_n$ is given by:

$$\mathbf{y}_n(j) = < y(t), \phi_j(t) > = \int_{(n-1)T_s}^{T_s} y(t)\phi_j^*(t)dt \tag{2.14}$$

for $j = \{1, 2, \ldots, D\}$.

Step 2: Determine the M distances. Given the received vector $\mathbf{y}_n$ as determined in step 1, we measure M distances between the received vector and the M possible *hypothesis points* $\{\mathbf{s}_1, \ldots, \mathbf{s}_M\}$. That is,

$$\begin{aligned} d(\mathbf{y}_n, \mathbf{s}_m)^2 &= \|\mathbf{y}_n - \mathbf{s}_m\|^2 = < y(t) - s_m(t), y(t) - s_m(t) > \\ &= \int_{(n-1)T_s}^{nT_s} |y(t) - s_m(t)|^2 dt \end{aligned} \tag{2.15}$$

Step 3: Determine the minimum distance. The detected symbol $\hat{\mathbf{x}}_n$ is given by:

$$\hat{\mathbf{x}}_n = \mathbf{s}_{m^*} \tag{2.16}$$

where

$$m^* = \arg\min_m d(\mathbf{y}_n, \mathbf{s}_m)^2 \tag{2.17}$$

The three steps are illustrated in Figure 2.14 Note that the first step can be implemented by a correlator or a *matched filter*. The minimum distance detector is simple

and easy to understand intuitively. Moreover, it can be shown (see Appendix A) that it is the optimal detector (best possible detector) for Gaussian noise.

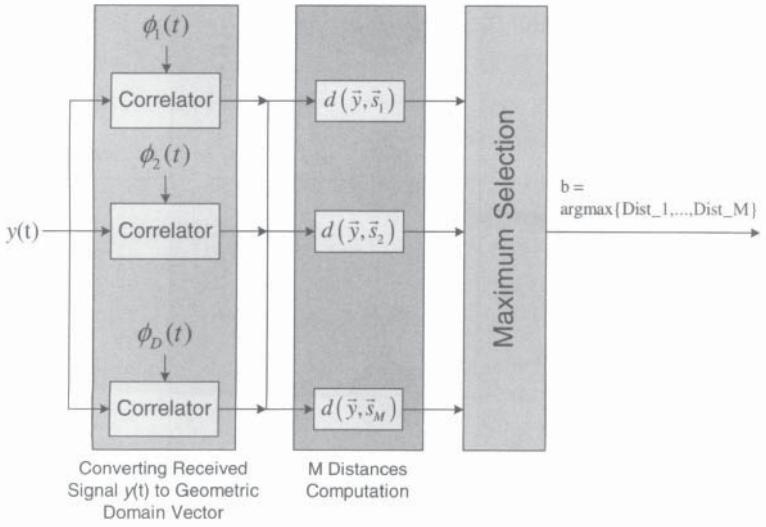


Figure 2.14 A block diagram of the minimum distance detector.

In fact, the 3-stage minimum distance detector can be further simplified. For instance, Equation (2.15) can be expressed as:

$$\begin{aligned} d(y(t), s_m(t))^2 &= \|\mathbf{y}\|^2 + \|\mathbf{s}_m\|^2 - 2 < \mathbf{y}, \mathbf{s}_m > \\ &= \|\mathbf{y}\|^2 + E_m - 2\int_{(n-1)T_s}^{nT_s} y(t)s_m^*(t)dt \end{aligned} \tag{2.18}$$

where $E_m = \|\mathbf{s}_m\|^2$ is the energy of the m-th signal element $s_m(t)$. Since the key of the minimum distance detector is to select the signal point that gives the minimum distance and $\|\mathbf{y}\|^2$ is independent of m, hence, the minimum distance detection rule is equivalent to:

$$m^* = \arg\max_m \int_{(n-1)T_s}^{nT_s} y(t)s_m^*(t)dt - E_m/2 \tag{2.19}$$

Hence, the minimum distance detector structure can be simplified as illustrated in Figure 2.15

Note that the first stage consists of M correlators or matched filters between the received signal $y(t)$ and $\{s_1(t), \ldots, s_M(t)\}$.

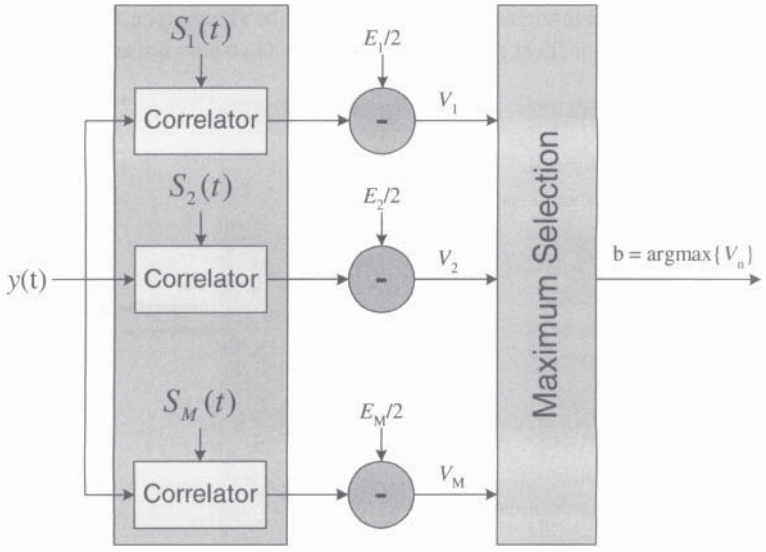


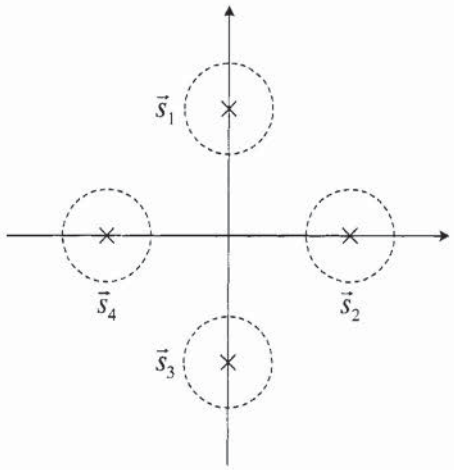
Figure 2.15 A simplified structure of the minimum distance detector.

2.5 PERFORMANCE AND TRADEOFFS

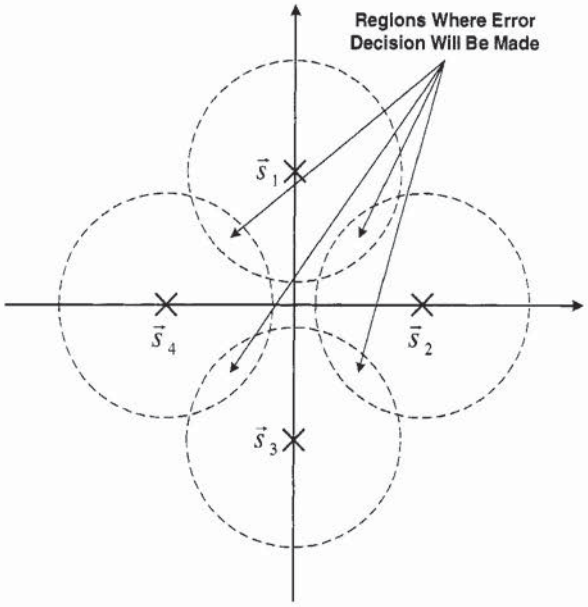
In the previous sections, we have introduced the concept of geometric representation of signals (signal space), the geometric models of digital modulators and digital demodulators. In this section, we complete the tradeoff big picture as stipulated in Figure 2.2

To understand the tradeoffs, we have to first understand the mechanism that a demodulator makes error on the detection process. Suppose the optimal demodulator, namely the *minimum distance detector*, is used and assume that $\vec{s}_1$ is actually transmitted. Figure 2.16(a) illustrates the situation when the channel noise power is small. As illustrated, the received signals $\vec{y}_n$ will fall within the noise cloud most of the time and since the noise cloud is small, the demodulator (minimum distance detector) will make the right decisions most of the time.

On the other hand, Figure 2.16(b) illustrates the situation when the channel noise power is large. When the received vector $\vec{y}_n$ falls in the area A, the minimum distance detector will pick $\vec{s}_1$ as the detected output and no error is made. However, when the received vector $\vec{y}_n$ falls in the area B, the minimum distance detector will pick $\vec{s}_2$ as the detected symbol and this results in decoding error. In general, there is a noise cloud around each constellation point at the receiver side and the minimum distance detector will make errors whenever the noise clouds between adjacent constellation points overlaps as illustrated in Figure 2.16(b). The probability of making detection error (symbol error probability) is given by the area of overlapping between constellation



(a) small noise



(b) large noise

Figure 2.16 Illustration of detection errors in minimum distance detector.

points. Obviously, the symbol error probability is limited by the minimum distance between any two constellation points in the signal constellation.

$$P_e \approx Q\left(\sqrt{\frac{d_{min}^2}{\eta_0}}\right) \tag{2.20}$$

where $d_{min} = \min_{i,j} \|\vec{s}_i - \vec{s}_j\|$, η_0 is the noise power spectral density and $Q(x)$ is the Gaussian Q-function. Hence, to reduce the error probability P_e, we have to maximum the minimum separations between constellation points.

Next, we apply the concept of noise cloud to complete the tradeoff picture. This is elaborated below.

Increasing transmit power can reduce error probability. Recall that P_e is related to the minimum distance between any two constellation points, d_{min}. Increasing the transmit power is equivalent to pulling all the constellation points further away from the origin. In this way, the minimum separation between any two constellation points will be increased and therefore, the error probability will be decreased without decreasing the bit rate.

Increasing transmit power can indirectly increase bit rate. On the other hand, one method to increase the bit rate without increasing the bandwidth requirement is to increase the modulation level M. For example, the bit rate can be doubled by using 16QAM constellation instead of QPSK constellation. However, doing so will increase the density of the constellation points unless the transmit power is increased. In that case, more constellation points can be squeezed into the signal space (higher bit rate) and at the same time, maintaining a similar *minimum distance separation between constellation points* (similar error probability) at the expense of higher transmit power.

Increasing bandwidth can increase bit rate. Recall that bandwidth is directly proportional to the baud rate or symbol rate of a digital communication link. Hence, increasing the bandwidth will directly increase the baud rate. Hence, the bit rate, which is given by $\log_2 M \times$ baud rate, will be increased.

Increasing bandwidth can indirectly reduce error probability. Finally, increasing bandwidth (while keeping bit rate unchanged) can increase the *dimension* of the signal space. For instance, say the dimension of the signal space is increased from 2 to 4 due to an increase in bandwidth. For the same number of signal constellation points and the same average transmit power (mean square distance of the points from the origin), the inter-constellation point spacing will be increased if the dimension of the signal space is increased. This results in a decrease of error probability without increasing power (at the expense of higher bandwidth).

The qualitative model of digital modulator and demodulator enable the design to compare the performance of different digital modulator designs without going through complicated math calculations. This is illustrated by a few examples in the following.

Example 3 (Design of Quarternary Modulator) *Consider a quarternary modulator with 4 points on the 2 dimensional signal space. What is the optimal transmit constellation given a fixed power constraint?*

Since the energy of a signal point $\vec{s}_i$ is given $\|\vec{s}_i\|^2$, the average power of the quarternary modulator with points $\{\vec{s}_1, \ldots, \vec{s}_4\}$ is given by:

$$P_{av} = \sum_{i=1}^{4} \frac{\|\vec{s}_i\|^2}{4T_s}$$

Suppose the four points have different lengths *to the origin. To keep the average power P_{av} constant, if we pull any two points $\{\vec{s}_1, \vec{s}_2\}$ away from the origin (so as to increase the separation of the two constellation points), the remaining two points $\{\vec{s}_3, \vec{s}_4\}$ have to move closer to the origin. Since the error probability is limited by the minimum distance between any two constellation points, the bottleneck of the constellation will be on $\vec{s}_3$ and $\vec{s}_4$. Hence, by symmetry, the four points should be equi-distance from the origin (or lie on a circle). Again by symmetry, the four points should be equi-distance from each other (i.e., lie on the four corners of a square) as illustrated in Figure 2.17*

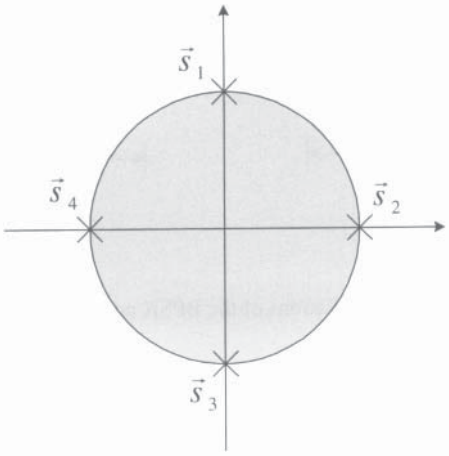


Figure 2.17 Optimal constellation of quarternary modulator on a 2-D signal space.

Example 4 (Comparison of ASK and BPSK modulators) *Consider a ASK and BPSK modulators as illustrated in Figure 2.9 Which of the two binary modulator designs is more energy efficient?*

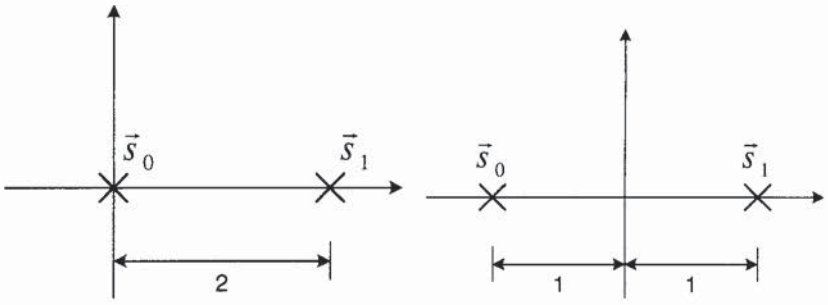
From the time domain waveforms, it is hard to say which of the two binary modulators has a better performance. Consider the signal constellation of the two binary modulators. Figure 2.18 illustrates the signal constellations of the BASK and BPSK modulators. Both designs have two constellation points in the signal set but the BASK design has one signal point at the origin. On the other hand, the BPSK design has the two signal points equi-distance from the origin with one on the positive x-axis and the other on the negative x-axis. Let the BPSK and BASK have the same error probability. This means that the distance separation between the two constellation points in both cases are the same (normalize to 2). For BASK, the average power is given by

$$P_{av}(BASK) = \frac{1}{2T_s}\left(\|\vec{s}_0\|^2 + \|\vec{s}_1\|^2\right) = \frac{2}{T_s}$$

On the other hand, the average power of the BPSK design is given by:

$$P_{av}(BPSK) = \frac{1}{2T_s}\left(\|\vec{s}_0\|^2 + \|\vec{s}_1\|^2\right) = \frac{1}{T_s}$$

Hence, we have $P_{av}(BASK) = 2P_{av}(BPSK)$. That is, the BASK design requires 3dB more power than the BPSK design for the same error probability.



(a) BASK (b) BPSK

Figure 2.18 Constellations of the BPSK and BASK modulators.

2.6 PRACTICAL ILLUSTRATIONS: DIGITAL MODULATION

In this section, we illustrate how digital modulations are applied in practical systems. From the previous sections, the power and bandwidth are two important parameters to characterize the *resource space*. Given the transmit power P_{tx} and bandwidth resource W, there is a maximum limit on the supported bit rate at a given error probability. For instance, there are two extreme operating regions in a digital communication system. They are namely systems with *high bandwidth efficiency* and

systems with *high bandwidth expansion*. When a system operates in the *high bandwidth efficiency* region, the system is bandwidth limited (rather than power limited). In this case, the digital modulation design relies on $M-$ary modulation scheme to increase the bit rate (at the expense of higher power). Examples are GSM, GPRS and EDGE cellular systems[141]. On the other hand, when a system operates in the *high bandwidth expansion* region, the system is power limited (rather than bandwidth limited). In this case, the digital modulation design relies on the expansion of signal space dimensions in order to reduce the error probability. Examples are satellite communications, deep space communications and spread spectrum systems. We elaborate how various practical systems operate in these two different regions.

High bandwidth efficiency. Consider the IS54 (Digital AMPS) system, which is the US digital cellular system. In IS54, the digitalized speech is protected using convolutional encoding and modulated using Offset-Quadrature Phase Shift Keying (O-QPSK) modulation. In regular QPSK system, the signal constellations is given by $\{\pi/4, -\pi/4, 3\pi/4, -3\pi/4\}$. Hence, the four time domain signals can be expressed as:

$$\begin{aligned} s(t) &= A\cos(\omega_c t + \theta) \\ &= \frac{A}{\sqrt{2}} d_I(t)\cos(\omega_c t) + \frac{A}{\sqrt{2}} d_Q(t)\sin(\omega_c t) \end{aligned} \tag{2.21}$$

where $d_I(t)$ is given by:

$$d_I(t) = \cos(\theta_n) = \{\pm\frac{1}{\sqrt{2}}\}$$

and $d_Q(t)$ is given by:

$$d_Q(t) = \sin(\theta_n) = \{\pm\frac{1}{\sqrt{2}}\}.$$

Hence, the QPSK modulator can be regarded as an independent I-Q modulator with the information bit streams $\{d_0, d_1, d_2, \ldots\}$ split into two independent bit streams $\{d_0, d_2, \ldots\}$ driving the I-th modulator $d_I(t)$ and $\{d_1, d_3, \ldots\}$ driving the Q-th modulator $d_Q(t)$. Figure 2.19(a) illustrates the QPSK modulation from the I-Q perspective.

Figure 2.19(b) illustrates the QPSK time domain waveform. The carrier phase change once every $2T_b$ duration where T_b is the bit duration. If from one bit interval to the next, neither the I or the Q bit stream changes sign, then the carrier phase remains unchanged. If either the I or the Q component changes sign, a phase change of $\pi/2$ occurs. However, if both the I and Q components change signs, a phase change of π occurs. For practical consideration, the QPSK signal with a phase change of π will go through zero momentarily after

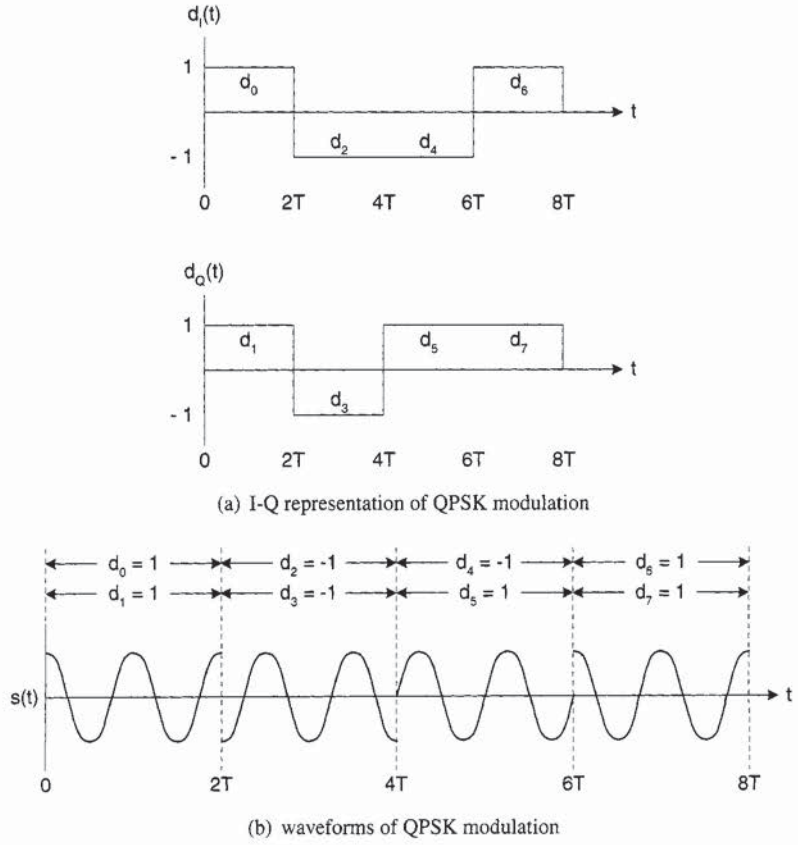
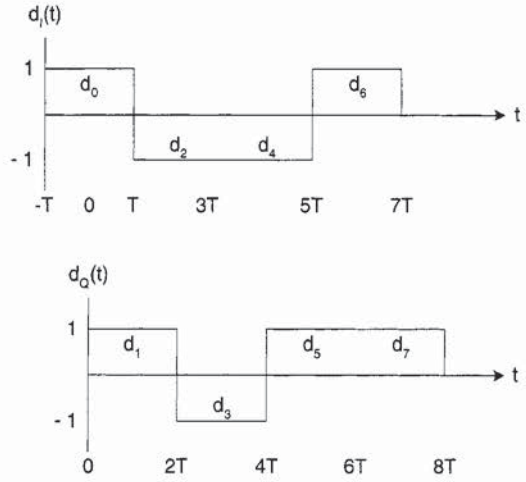


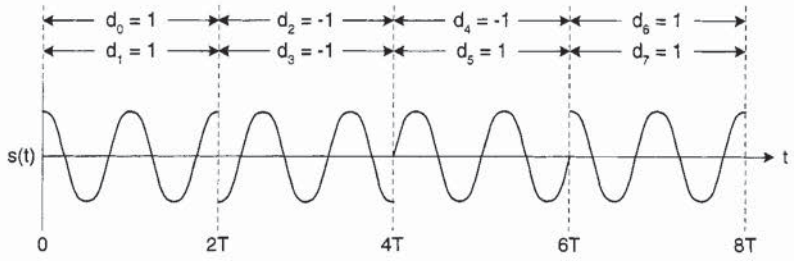
Figure 2.19 Illustration of regular QPSK modulation.

pulse filtering. This results in non-constant envelop in the QPSK waveform or in general, a higher Peak-to-Average (PAR) ratio[299]. Signals with high PAR is notorious of *spectral regrowth* in the side lobes after non-linear process (such as the class B power amplifier). Hence, we need to operate the power amplifier with a higher *backoff factor* to avoid the saturation of the power amplifier during some peaks of the envelop. Therefore, the regular QPSK is not desirable from a spectral regrowth perspective.

Offset QPSK (O-QPSK) is designed to overcome the high PAR problem of the QPSK signal. For instance, the I and Q streams of information $d_I(t)$ and $d_Q(t)$ are offset by $T_b/2$ in time as illustrated in Figure 2.20(a).



(a) I-Q representation of O-QPSK modulation



(b) waveforms of O-QPSK modulation

Figure 2.20 Illustration of Offset-QPSK modulation in IS54 systems.

Thus, the possible range of phase transition becomes 0 to $\pi/2$. The phase change of π is eliminated by the $T_b/2$ time offset. While the amplitude fluctu-

ations still occur in O-QPSK waveform, it fluctuates over a smaller amplitude range because the amplitude will not pass through a zero point anymore. Hence, the O-QPSK signal has a smaller PAR[306] and therefore, smaller power back-off is required at the power amplifier to avoid spectral regrowth problem. This enhances the efficiency of the power amplifier and therefore, the O-QPSK was adopted in the IS-54 standard. Note that the O-QPSK modulation has the same error performance and bit rate performance as the regular QPSK modulation. The only difference is the O-QPSK systems will have a lower PAR and this favors the robustness towards the non-linear spectral regrowth effects.

Similarly, the modulation system of GSM system, which is the second generation cellular system, follows a similar consideration. Speech signal is first digitalized using speech compression technique (vocoder) to produce a digital stream of source data at a rate of 13 kbps. The source data is protected by convolutional encoder and modulated using Gaussian Minimum Shift Keying (GMSK) and transmitted over the air to the receiver. The GMSK is derived from the O-QPSK design by replacing the rectangular pulse in the I and Q waveforms with a *Gaussian pulse*. The time domain equation of a Gaussian pulse is given by:

$$g(t) = \frac{1}{2T_b}\left[Q\left(2\pi B\frac{t - T_b/2}{\sqrt{\log_e 2}}\right) - Q\left(2\pi B\frac{t + T_b/2}{\sqrt{\log_e 2}}\right)\right] \qquad (2.22)$$

where $BT_b \in [0, \infty)$ and $Q(t)$ is the Q-function defined as:

$$Q(t) = \int_t^\infty \frac{1}{\sqrt{2}} \exp - \left(-x^2/2\right) dx$$

and B is the low-pass bandwidth of the Gaussian pulse. Figure 2.21 illustrates the Gaussian pulse with $BT_b = 0.5$.

We not go into the details of GMSK modulation here and the interested reader could refer to [306] for a more quantitative treatment. Similar to O-QPSK, the design of GMSK is to reduce the spectral regrowth as a result of non-linear processing in the transmitter chain.

Consider one more example in the cellular system. EDGE (Enhanced Data rate for GSM System) system is regarded as a 2.5G technology designed to enhance the data rate of the GSM / GPRS air interface. It increases the maximum data rate from 100kbps of GPRS system to around 384kb/s in EDGE systems. The first order enhancements in the EDGE system is the application of higher order modulation (BPSK and 8PSK) as well as the use of *adaptive modulation*. For instance, as we have discussed above, the 8PSK modulation offers a higher bit rate at the expense of higher required power for same error performance. The BPSK modulation offers a more robust performance (error probability) at the expense of lower bit rate. It is found that the received Signal-to-Noise ratio

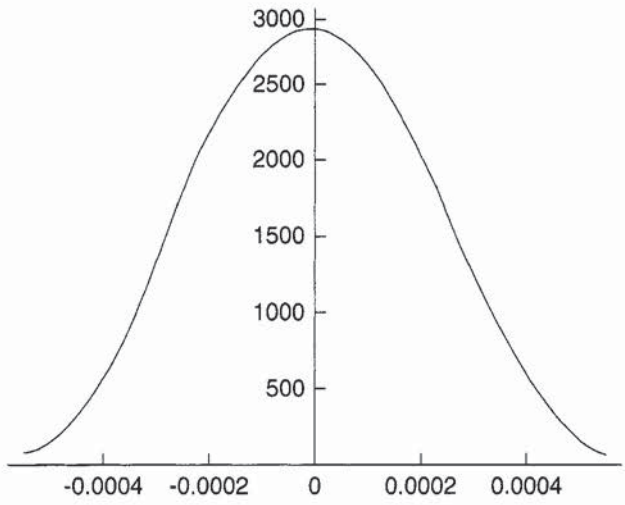


Figure 2.21 Illustration of Gaussian pulse shape in GMSK modulation with $BT_b = 0.5$.

(SNR) of mobiles is a random variable with a certain mean value. Sometime, the instantaneous received SNR of a mobile is high and sometimes, it is small. If the transmitter knows about the instantaneous received SNR, it can adaptively change the modulation constellations according to the SNR level. Say if the current SNR is high, the transmitter should transmit information using 8PSK modulation to boost the throughput. On the other hand, when the current SNR is low, the transmitter should transmit using BPSK modulation for robust error performance at the expense of bit rate. In EDGE, the transmitter can adaptively select the modulation level on a *block level* (a block consists of several frames). This adaptive modulation approach substantially enhances the average bit rate of the mobile users.

High bandwidth expansion. On the other hand, some practical systems are designed to operate in high bandwidth expansion mode. For example, in CDMA systems such as IS-95, CDMA2000 or W-CDMA systems, the modulation design is based on high-bandwidth expansion approach. For instance, in IS95 systems, *time-orthogonal modulation* is used in the reversed link (mobile to base station). Six information bits $\{d_0, d_1, \ldots, d_5\}$ are grouped together to select one out of 64 *modulation symbols* $\{\vec{s}_1, \ldots, \vec{s}_{64}\}$. Hence, it is a 64-ary modulation but the constellation (with 64 points) are positioned in a 64-dimensional signal space. Moreover, the 64 modulation symbols are *orthogonal* to each other. That is:

$$< \vec{s}_i, \vec{s}_j >= 0$$

where $< x, y >$ denotes the vector dot-product. Hence, the modulation scheme is also called 64-ary orthogonal modulation. The 64-dimensional signal space are derived from the *time-spreading operation* which will be discussed in detail in Chapter 3. The design philosophy in IS95 orthogonal modulation is to expand the required bandwidth to trade for robustness in error performance in the presence of strong interference (multi-user interference, intersymbol interference as well as multi-cell interference). This robustness is also called the *interference suppression capability* in CDMA systems. In other words, it is operating in the tradeoff direction of using extra bandwidth to reduce the error probability.

2.7 SUMMARY

In this chapter, we have elaborated the fundamental of point-to-point digital communication system design. We have elaborated a *black box model* for the digital modulator and the digital demodulator. For instance, a digital modulator can be characterized by (i) the symbol duration, and (ii) the signal set $\mathcal{S} = \{s_1(t), \ldots, s_M(t)\}$. The required bandwidth is given by $BW = \alpha \frac{1}{T_s}$. The transmitted bit rate is given by $R_b = \log_2(M)\frac{1}{T_s}$. The signal set can be described by either the time domain approach or *geometric domain approach* (vectors in a signal space). Using the signal space concept, the operations of digital demodulator can be easily understood and the optimal digital demodulator with respect to the AWGN channel is the *minimum distance decoding rule*. The decoded symbol is chosen to be the one of the M possible points $\{\vec{s}_1, \ldots, \vec{s}_M\}$ that is closet to the received vector $\vec{y}$. The error probability is a function of the minimum distance between any two constellation points.

After the introduction of digital modulation and digital demodulation, we have elaborated on the big picture of system tradeoffs between the bandwidth, power, bit rate and error probability. For instance, using higher bandwidth can result in higher bit rate for obvious reasons. On the other hand, using a higher bandwidth can also result in lower error probability indirectly due to the expansion of the dimension of the signal space. Using a higher transmit power results in smaller error probability directly because signal constellations are separated further. On the other hand, using a higher transmit power can also result in a higher bit rate indirectly because it allows the use of higher level modulation (M) over the same dimension signal space. Higher power allows a similar separation between the constellation points after more constellation points are squeezed into the signal space.

Finally, we have illustrated the use of digital modulation in some commercial systems such as the digital AMPS, GSM, EDGE as well as CDMA systems. The digital AMPS, GSM, EDGE systems are designed to operate in high bandwidth efficiency manner where a higher power is required to tradeoff for higher bit rate. On the other hand, the CDMA systems are designed based on high bandwidth expansion modes

where a higher bandwidth expansion is required to tradeoff for robustness against interference.

PROBLEMS

2.1 Suppose there is a scheme that makes use of a special traffic light to send digital information. There are 3 lights (with different colors: red, yellow and green). The traffic light changes pattern once every 10ms. Suppose we want to build a "modulator" that transmits a "symbol" based on the different light combinations (each light can be on or off but the three lights cannot be "on" simultaneously) and the modulation process is symbol-by-symbol (encoding across one 10ms symbol only). What are the baud rate and the bit rate of the system if we consider a basic symbol with duration 10ms?

2.2 One way to increase the information bit rate through such a setting is to encode information bits based on two "symbols" together. (i.e., a new "super-symbol" of duration 20ms is formed by two consecutive symbols and information bits are carried by the new "super-symbol"). What is the resulting bit rate?

2.3 What is the bit rate when grouping N symbols at a time during the modulation? How about the asymptotic bit rate as $N \to \infty$?

This Page Intentionally Left Blank

CHAPTER 3

MULTIUSER COMMUNICATIONS

Men, like all other animals, naturally multiply in proportion to the means of their subsistence.
—Adam Smith: *The Wealth of Nations*, I, 1776

3.1 INTRODUCTION

In Chapter 2, we have introduced a generic model for digital modulation and demodulation design based on the concept of signal space. We have also discussed various channel coding and decoding techniques. All these focused on the point-to-point communications where there is a single transmitter and a single receiver. However, in many situations, communications may involve multiple users. In this chapter, we focus on both the theoretical and application aspects of techniques for multi-user communications.

We consider a system with single base station and multiple mobile users. In general, there are two directions of multi-user communications, namely the uplink direction and the downlink direction. Uplink refers to the direction from mobile users

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

to the base station as illustrated in Figure 3.1(a). This is a *many-to-one* communication scenario and is called the *multi-access problem*.

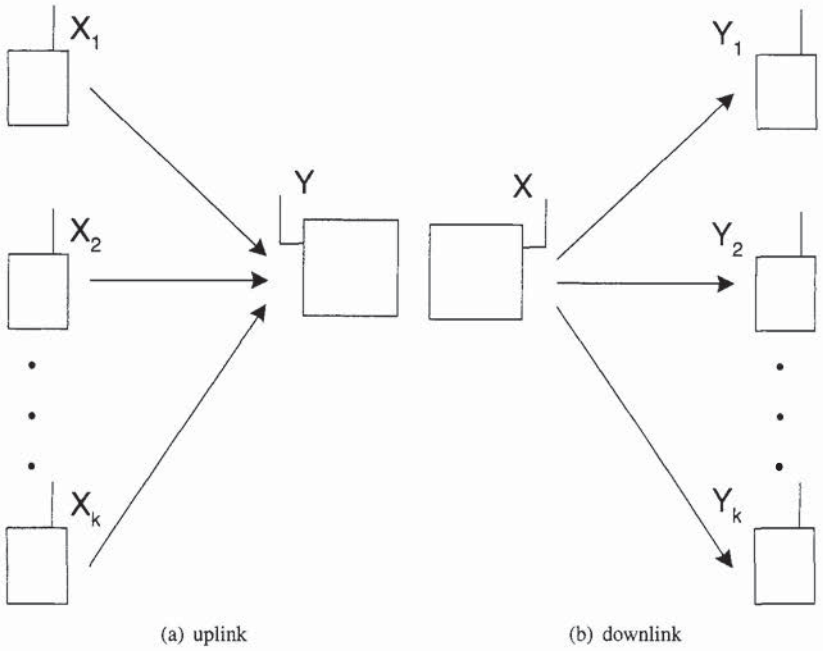


Figure 3.1 Illustration of two directions of multi-user communications.

On the other hand, downlink refers to the communication direction from the base station to the mobile users as illustrated in Figure 3.1(b). This is a *one-to-many* communication scenario and is called the *broadcast problem*. The two problems were treated separately for the past few decades but lately, it is shown that the two problems are dual of each other. In Section 3.2, we give an overview on the information theoretical framework for the two problems. In Section 3.3, we discuss various commonly used techniques to partition resources in both the uplink and downlink directions in an orthogonal manner. For example, commonly used techniques such as FDMA (Frequency Division Multiple Access), TDMA (Time Division Multiple Access), Deterministic CDMA (Code Division Multiple Access) belong to this category. In Section 3.4, we discuss various commonly used techniques for resource partitioning in a non-orthogonal manner. For example, random CDMA as well as space division multiple access belong to this category. In Section 3.5, we compare the pros and cons of various technologies in single cell and multi-cell scenarios. In Section 3.6, we describe how these technologies are applied in practical systems such

as GSM, CDMA and IEEE 802.11b systems. Finally, we give a summary of the main points and concepts in Section 3.7.

3.2 INFORMATION THEORETIC OVERVIEW

In the context of point to point communications, there are two approaches to attack the design problem. One approach, namely on the *signal processing approach*, focused on the specific design of modulation and error control coding schemes. Obviously, different specific design will have different physical layer performance (as elaborated in Chapter 2) in terms of supported bit rate, required transmit power, required bandwidth and resulting error probability. A natural question to ask is *"could we do a better physical layer design"*?. The signal processing approach could not offer a complete answer to this question.

Another approach, namely on the *information theoretical approach*, focused on characterizing the *channel capacity* [78]. Specifically, the theoretical framework isolates the best performance from the specific implementation details of modulation and error control coding. It gives the performance bound achieved by the most optimal design of error control coding and modulation schemes. For example, given a fixed transmit power and a fixed bandwidth, the maximum data rate [78] that could be transmitted *reliably*[1] over the *Additive White Gaussian Noise* (AWGN) channel is given by:

$$C = W \log_2 \left(1 + \frac{P}{\eta_0 W}\right) \tag{3.1}$$

where W is the channel bandwidth, P is the received power and η_0 is the channel noise density. In other words, there is no channel coding and modulation scheme that could *reliably* deliver a data rate larger than C and C is called the *channel capacity* or *Shannon's capacity*. This approach has the beauty that it gives a complete answer to the previous question because it presents the optimal solution. Unfortunately, this approach gives no indication on the structure of the optimal channel coding design. Therefore, there are still plenty of different researchers adopting these two approaches to address the point-to-point communication link design.

Similarly, there are two approaches to address the multi-user communication system design, namely the *signal processing approach* and the *information-theoretical approach*. We introduce major information-theoretical results with respect to multi-user communications in this section. However, unlike the point-to-point situation where the channel is characterized by a scalar called *channel capacity*, the multi-user

[1]Reliable transmission of information bit to the receiver refers to the asymptotic situation that the error probability could be made arbitrarily small at arbitrarily large block length.

$$\lim_{N \longrightarrow \infty} P_e = 0$$

channel is characterized by a $K-$dimensional region $\mathcal{R}$ where K is the number of users involved in communications.

3.2.1 Multi-Access Problem

3.2.1.1 General Problem Formulation As illustrated in Figure 3.1(a), let $X_k \in \mathcal{C}$ be the transmitted symbol out of user k in the uplink and $Y \in \mathcal{C}$ be the received symbol at the base station where $\mathcal{C}$ represent a set of complex number. The multiaccess channel is in general characterized by a probabilistic function (or *channel transition probability*) $p(\mathbf{Y} = \mathbf{y}|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_K = \mathbf{x}_K)$ where $\mathbf{Y} = [Y(1), \ldots, Y(n), \ldots, Y(N)]$ is the block of received symbols, $\mathbf{X}_k = [X_k(1), \ldots, X_k(N)]$ is the block of transmitted symbols from user k and n is the time index. If the channel transition probability could be expressed as

$$\begin{aligned} p(\mathbf{Y} &= \mathbf{y}|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_K = \mathbf{x}_K) \\ &= \Pi_{n=1}^{N} p(Y(n) = y(n)|X_1(n) = x_1(n), \ldots, X_K(n) = x_K(n)) \end{aligned} \quad (3.2)$$

the multiaccess channel is called *memoryless*. Otherwise, the mutliaccess channel is called *memory*. For simplicity, we assume a simple multiaccess channel, namely the multiaccess channel with Additive White Gaussian Noise. The AWGN channel is a memoryless multiaccess channel and the received signal is given by:

$$Y = \sum_{k=1}^{K} X_k + Z \quad (3.3)$$

where Z is the white Gaussian noise with variance σ_z^2. Putting the channel model into the form of *channel transition probability*, we have:

$$\begin{aligned} p(Y(n) &= y(n)|X_1(n) = x_1(n), \ldots, X_K(n) = x_K(n)) \\ &= \frac{1}{\sqrt{2\pi}\sigma_z} \exp\left(-\frac{|y(n) - \sum_k x_k(n)|^2}{2\sigma_z^2}\right) \end{aligned} \quad (3.4)$$

We first have a few definitions.

Definition 1 *A* multiaccess coding scheme *with block length N for K users is a collection of codebooks $\mathcal{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_K\}$, where the codebook for the $k-$th user is a mapping from the message set $\mathcal{M}_k$ to the block of transmitted symbols $\mathbf{X}$. That is: $\mathcal{B}_k : \mathcal{M}_k \longrightarrow \mathcal{C}^N$ where $\mathcal{M}_k$ is a set of positive integers with cardinality $|\mathcal{M}_k|$.*

Figure 3.2(a) illustrates the concept of 2-user multi-access coding scheme as an example.

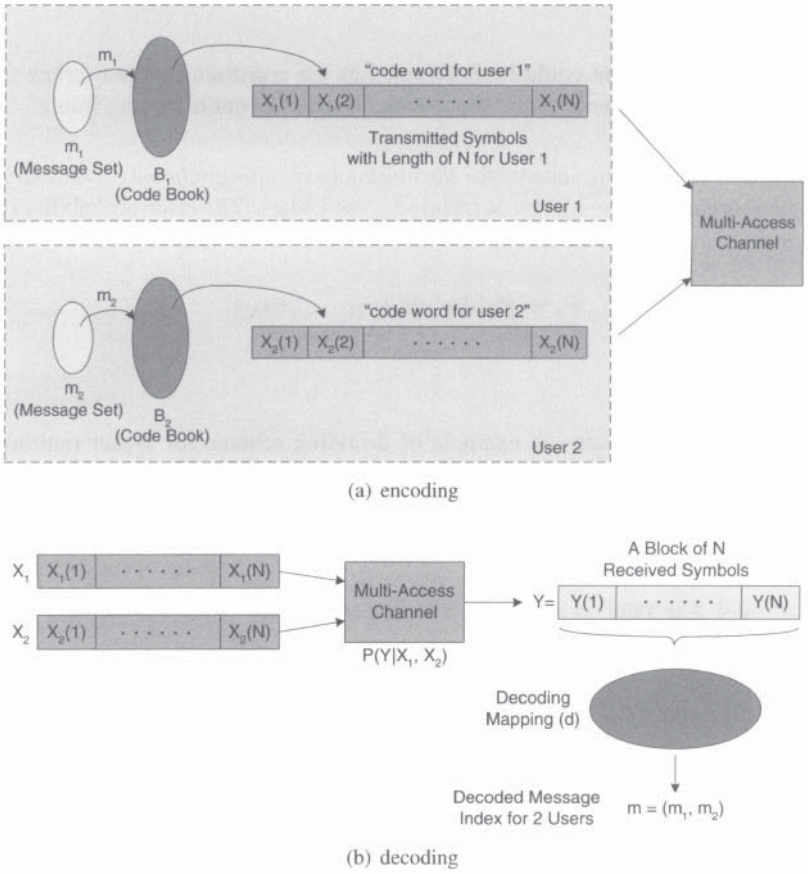


Figure 3.2 Illustration of two-user multiaccess encoding and decoding scheme.

Definition 2 *The multiaccess coding scheme is said to satisfy* power constraints, $(P_1, \ldots, P_K)$, *if for each* $k \in [1, K]$ *and each* $m_k \in \mathcal{M}_k$, *we have*

$$\frac{1}{N} \sum_{n=1}^{N} |X_k(n)|^2 \leq P_k$$

The power constraint could be interpreted as the constrain on the average transmitted power out of user k where the average is over the entire coding frame.

Definition 3 *A* decoding scheme *for the multiaccess coding scheme is characterized by a mapping* $d : \mathcal{C}^N \longrightarrow [\mathcal{M}_1] \times [\mathcal{M}_2] \times \ldots \times [\mathcal{M}_K]$. *The* error probability *of the decoding scheme is given by:*

$$P_e = \Pr\left[d(\mathbf{y}) \neq (m_1, \ldots, m_K)\right]$$

Figure 3.2(b) illustrates an example of decoding scheme for 2-user multiaccess code. In other words, the task of the decoder at the base station is to separate the *embedded information* from all the K users based on the observation $\mathbf{Y}$ over the frame of N symbols.

Definition 4 *The* rate vector *of the multiaccess coding scheme is defined by* $\mathbf{r} = (r_1, \ldots, r_K)$ *where:*

$$r_k = \frac{1}{N} \log_2(|\mathcal{M}_k|)$$

The multiaccess code is called achievable *at a rate vector* $\mathbf{R}$ *if there exists a codebook* $\mathcal{B}$ *such that* $\lim_{N \to \infty} P_e = 0$ *and* $\lim_{N \to \infty} \mathbf{r}(\mathcal{B}) \geq \mathbf{R}$. *Note that* $(r_1, r_2) \geq (R_1, R_2)$ *if and only if* $r_1 \geq R_1$ *and* $r_2 \geq R_2$.

Definition 5 *The* capacity region *of a multiaccess channel with* K *users satisfying the power constraints* $\{P_1, \ldots, P_K\}$ *is defined as the* convex closure *of all achievable rate vectors.*

With the above definitions, the problem of multiaccess communication is to find out the *capacity region*[78] given a multiaccess channel model (or the channel transition probability) and power constraints. Therefore, the *channel capacity* in the context of point-to-point communication link is simply a special case of the *capacity region* for multiaccess problem.

3.2.1.2 General Solution to the Multiaccess Problem The general solution of the multiaccess problem [78] is summarized by the following theorem.

Theorem 1 *The* capacity region *of a multi-access channel is given by the region defined by the following inequalities.*

$$\mathbf{r}(S) \leq I\left(Y; (X_k)_{k\in S} | (X_j)_{j\notin S}\right) \quad \forall S \subset [1, \ldots, K]$$

for some input distributions $p(X_1)p(X_2)\ldots(X_K)$ *where* $\mathbf{r}(S) = \sum_{k\in S} r_k$ *and* $I(X;Y|B)$ *denotes the conditional mutual information between* X *and* Y *given* B.

In other words, there exists *achievable* multiaccess codes if the rate vector lies within the above region. On the other hand, if a rate vector lies outside the above region, all multiaccess codes will be *non-achievable*. The proof of this theorem is omitted as this is outside the scope of the book.

For example, consider a 2-user multiaccess system with AWGN channel. Similar to the point-to-point situation, the *capacity achieving input distribution* in this case is also Gaussian. Hence, the capacity region is given by:

$$r_1 \leq I(X_1; Y|X_2) = \log_2\left(1 + \frac{P_1}{\sigma_z^2}\right) \tag{3.5}$$

$$r_2 \leq I(X_2; Y|X_1) = \log_2\left(1 + \frac{P_2}{\sigma_z^2}\right) \tag{3.6}$$

$$r_1 + r_2 \leq I(X_1, X_2; Y) = \log_2\left(1 + \frac{P_1 + P_2}{\sigma_z^2}\right) \tag{3.7}$$

The 2-user AWGN capacity region is illustrated in Figure 3.3

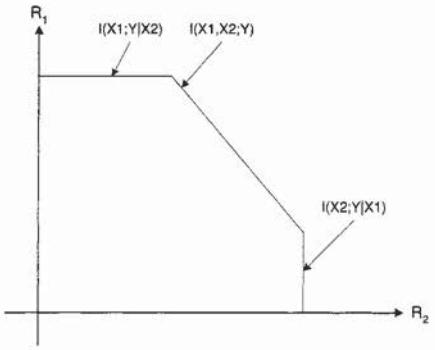


Figure 3.3 2-user multiaccess capacity region.

All the points could be realized by *successive interference cancellation* and/or *time sharing* citeVerdu1998. For instance, point A could be realized by decoding

information for user 2 first, subtract X_2 from Y and then decode information for user 1. Point B could be realized by decoding user 1 followed by user 2. Point C could be realized by time sharing between the 2 decoding order.

3.2.2 Broadcast Problem

Broadcast problem[77] refers to the communication directions from the base station to the K mobile users.

3.3 ORTHOGONAL RESOURCE PARTITIONING

While Section 3.2 highlights the main results in multiuser communication theory, the theory itself simply gives the optimal performance bound. It does not provide information on how to realize the expected performance. In practice, the optimal schemes as suggested by theories have high computational complexity[2] and are usually infeasible. Yet, the theoretical results are important in the sense that they could give us a reference on the best achievable results. Based on the optimal performance reference, we could compare between various practical designs and evaluate how far they are from the optimal schemes.

In the next few sections, we address the multiuser communication problem based on the second approach, *signal processing approach.* We introduce various schemes for multiuser communications commonly used in practice due to their simplicity and compare their relative performance in single cell and multi cell scenarios. In this section, we focus on the *orthogonal schemes* such as FDMA, TDMA and deterministic CDMA. In the next section, we discuss on the *non-orthogonal schemes* such as random CDMA and SDMA. In Section 3.5, we compare their performance.

Assuming AWGN channel, the received signal $Y(t)$ (uplink direction) or the transmitted signal (downlink direction) at the base station is given by:

$$Y(t) = \sum_{k=1}^{K} X_k(t) + Z(t) \tag{3.8}$$

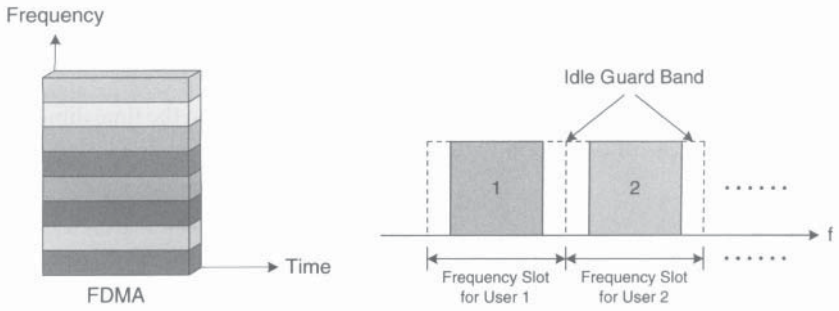
where $Z(t)$ is the AWGN channel noise, $X_k(t)$ is the transmitted signal from (or to) the $k-$th user. In the uplink direction, the task of the *multiuser detector* is to separate information related to the K users. In the downlink direction, the task of the *multiuser transmitter* is to make sure user k receives the signal specific to it ($X_k(t)$) without severe degradation. In other words, K *user channels* have to be formed between the base station and the K users. In general, some means of *user coordination* is needed in order to reduce the complexity of base station and mobile station processing. One

[2]For example, the optimal multiaccess detection algorithm is shown to have exponential order of complexity with respect to number of users.

obvious and simple approach is to partition the resource space based on *orthogonality principle*. That is, K orthogonal user channels are formed as a result of the resource partitioning schemes. By orthogonal, we mean that the *processed signal* for user k_1 do not interfere with the *processed signal* for user k_2. We have 3 popular orthogonal multiuser communication schemes, namely the frequency division multiple access (FDMA), time division multiple access (TDMA) and *deterministic* time division multiple access (CDMA). The FDMA scheme partitions the K user channels based on the frequency dimension of the signal space. The TDMA scheme partitions the K user channels based on the time dimension of the signal space. Finally, deterministic CDMA partitions the user channels based on the *code domain* of the signal space. These three schemes are applicable to both the multiaccess problem and the broadcast problem and they will be explained in details below.

3.3.1 Frequency Division Multiple Access (FDMA)

FDMA partitions the communication resource in the *frequency dimension* as illustrated in Figure 3.4(a).



(a) partitioning of K user channels for FDMA

(b) guard band of FDMA

Figure 3.4 Principle of FDMA.

Specifically, the allocated spectrum with bandwidth W is divided into K frequency slots to form K user channels. Each channel has a bandwidth of W/N. Different mobiles are assigned to transmit (or receive) at different frequency slots and hence, there is no interference of signals between different users. The K channels are therefore *orthogonal channels*. Users could transmit (or receive) at the assigned frequency slot(s) over the entire time frame.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Let the modulation throughput be 1 bit per symbol and the required data rate of individual

user be 25 kbps. Find out how many users could be supported based on the FDMA scheme.

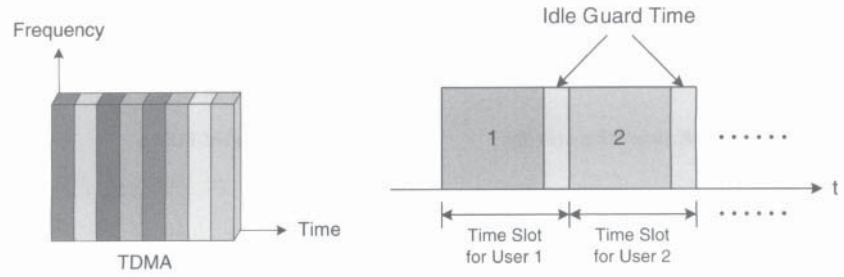
Solution With binary modulation, the required bandwidth per user to support an average bit rate of 25kbps is 25kHz. Hence, the number of frequency slots available over a spectrum of 15MHz is given by:

$$N_{FDMA} = 15M/25k = 600.$$

For the uplink multiaccess communication, signal transmitted from user k ($X_k(t)$) could be extracted from the received signal using a bandpass filter tuned to the frequency range of the assigned channel. For the downlink communications, signal to user k ($X_k(t)$) is modulated to the assigned frequency slot. However, in practice, *perfect bandpass filtering* is not realizable. Practical filters have finite slope in the frequency response. To facilitate simple processing at the base station, *frequency guard band* is maintained between channels as illustrated in Figure 3.4(b). Another design consideration is that the number of bandpass filters or RF units required at the base station is equal to the number of mobile users in the system. Since these components are analog circuits, they are usually quite bulky compared with digital components.

3.3.2 Time Division Multiple Access (TDMA)

On the other hand, TDMA partition the K user channels based on the time dimension of the signal space as illustrated in Figure 3.5(a).



(a) partitioning of K user channels for TDMA

(b) guard time of TDMA

Figure 3.5 Principle of TDMA.

Specifically, all the K user channels occupy the same entire bandwidth W. However, each user takes turn to use the bandwidth in a coordinated manner. Different mobiles are assigned to transmit (or receive) at different *time slots* and therefore, there

is no interference of signals between different users. The resulting K user channels are therefore orthogonal channels.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Let the modulation throughput be 1 bit per symbol and the required average data rate of individual user be 25 kbps. Find out how many users could be supported based on the TDMA scheme.

Solution: Let T_p be the TDMA frame period and T_k be the time slot duration assigned to user k. When user k transmits in the assigned time slot, the peak bit rate is given by 15Mbps (because the modulation throughput is 1 bit per symbol). However, user k does not transmit during the rest of the time on T_p. Hence, the average bit rate of user k is given by:

$$R_b = 15M \times (T_k/T_p) = 25k$$

On the other hand, the number of time slots available for assignment is given by T_p/T_k. Hence, the capacity of TDMA is given by:

$$N_{TDMA} = T_p/T_k = 15M/25k = 600$$

For the TDMA scheme to work, all the K users have to be *synchronized* so that they have a common timing reference. However, in practice, it is not possible to have all the K users having exactly the same timing reference as the base station due to the variation in propagation delays of the K users. To allow certain tolerance as a result of user distribution, *guard time* is used between time slots as illustrated in Figure 3.5(b). During the guard time interval, no signal is transmitted and therefore, it does not interfere with other channels even if the user timing is slightly offset from the base station timing. In fact, the guard time imposes a limit on the maximum cell radius. Another characteristics of TDMA is the burstiness of physical data rate per user. A user will transmit at the *peak rate* (using the entire bandwidth) during the assigned time slot and remain silence at other time slots. This presents two potential problems. Firstly, it is more difficult to transmit the wideband signal (at the allocated time slot) because multipath fading will be seen by the transmitted signal. Complex channel equalization is therefore required at the receiving end to reduce the *inter-symbol-interference* (ISI). Secondly, the rapid switching on and off of the power amplifier at the TDMA transmitter will generate high level of *electromagnetic interference*. On the other hand, one advantage of TDMA scheme is the requirement on a single RF transceiver at the base station regardless of the number of TDMA channels. Therefore, in general, operators seldom partition the system resource based on TDMA entirely.

3.3.3 Deterministic Code Division Multiple Access (D-CDMA)

While FDMA and TDMA partition the system resource based on the frequency dimension and time dimension of the signal space, CDMA [339] partition the system resource based on the *code-dimension* of the signal space. In other words, all the K

user signals share the entire bandwidth and the entire time duration but differentiate among themselves based on different *code channels*. This extra code dimension is introduced through the *spreading process* as illustrated in Figure 3.6(a).

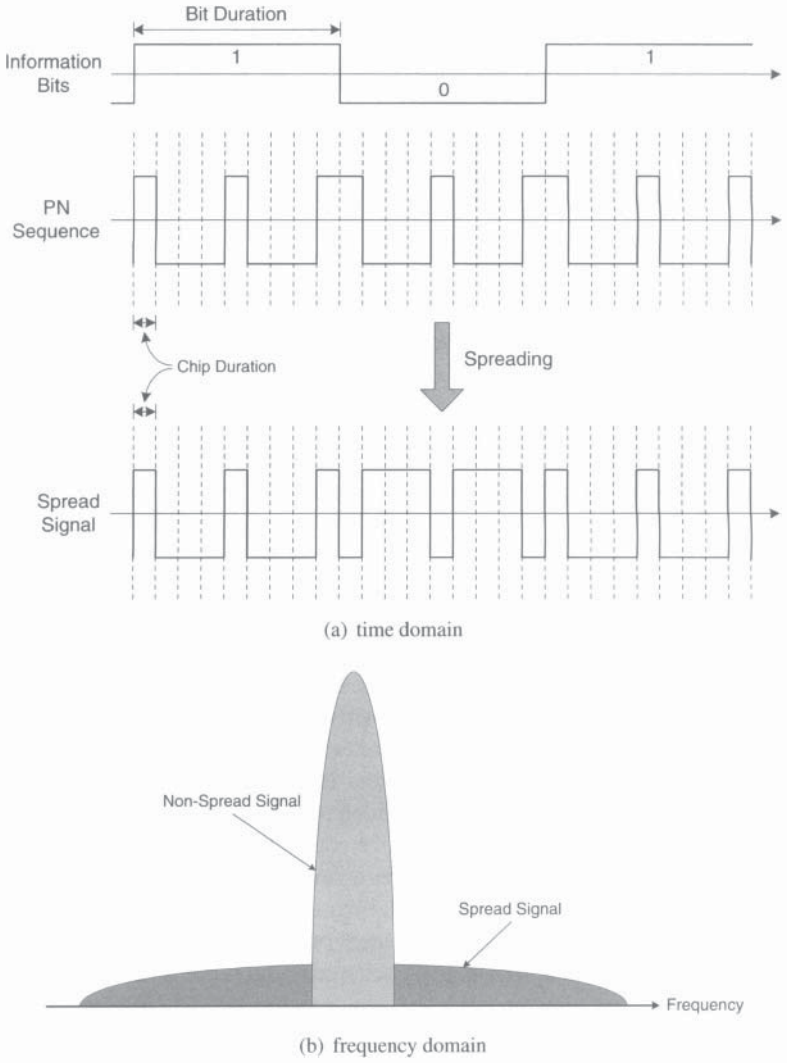


Figure 3.6 Principle of CDMA spreading.

Specifically, consider the transmit processing of the signal for user k. Individual information bit (or modulation symbol) is *spread* with a code sequence $c_k(t) \in$

$\{+1, -1\}$ to form the transmit signal. The code sequence is a time sequence with the smallest time unit of transition called a *chip*. The time duration of a chip is called *chip duration* T_c. Hence, effective, a modulation symbol is *chopped* into many tiny pieces by the code sequence. The ratio between the modulation symbol duration T_s and the chip duration T_c is called the *spreading factor* (SF). That is:

$$SF = \frac{T_s}{T_c} \tag{3.9}$$

As a result of the spreading process, the transmit bandwidth is increased by SF times. That is:

$$W_{tx} = W_s \times SF \tag{3.10}$$

where W_s is the signal bandwidth without spreading. Figure 3.6(b) illustrates the spreading process in frequency domain.

Each of the K transmit signals from (or to) the K users are spread by the code sequence $\{c_1(t), \ldots, c_K(t)\}$ respectively. Note that these K signals all share the same time and same bandwidth and they are separated based on the distinct code sequence only. For the uplink direction, the received signal at the base station is given by:

$$Y(t) = \sum_{k=1}^{K} S_k(t)c_k(t) + Z(t)$$

where $S_k(t)$ is the information signal for the $k-$th user, $c_k(t)$ is the spreading code sequence for the $k-$th user and $Z(t)$ is the channel noise. Since the transmitted signal all share the same time domain and frequency domain, we have to exploit on the characteristics of code sequence $c_k(t)$ in order to separate the K signals. Specifically, we apply *despreading* processing at the receiving end. To extract signal for user k, we multiply the received signal $Y(t)$ with the corresponding code sequence $c_k(t)$ as illustrated in Figure 3.7(a).

The despread signal is given by:

$$W_k(t) = Y(t) \times c_k(t) = \underbrace{S_k(t)}_{\substack{\text{Signal} \\ \text{Component}}} + \underbrace{\sum_{j \neq k} S_j(t)c_j(t)c_k(t)}_{\text{MultiuserInterference}} + \underbrace{Z(t)c_k(t)}_{\substack{\text{Channel} \\ \text{Noise}}} \tag{3.11}$$

due to the fact that $c_k(t) \times c_k(t) = 1$. Hence, we could see that the desired signal component for user k, $S_k(t)$, is *constructed* after the despreading operation. However, in general, in addition to the channel noise component, there is another *multiuser interference* component. If the code sequence $\{c_1(t), \ldots, c_K(t)\}$ are chosen in such a way that the multiuser interference is zero, the CDMA scheme is called *deterministic CDMA* and similar to FDMA or TDMA, the K user channels become orthogonal.

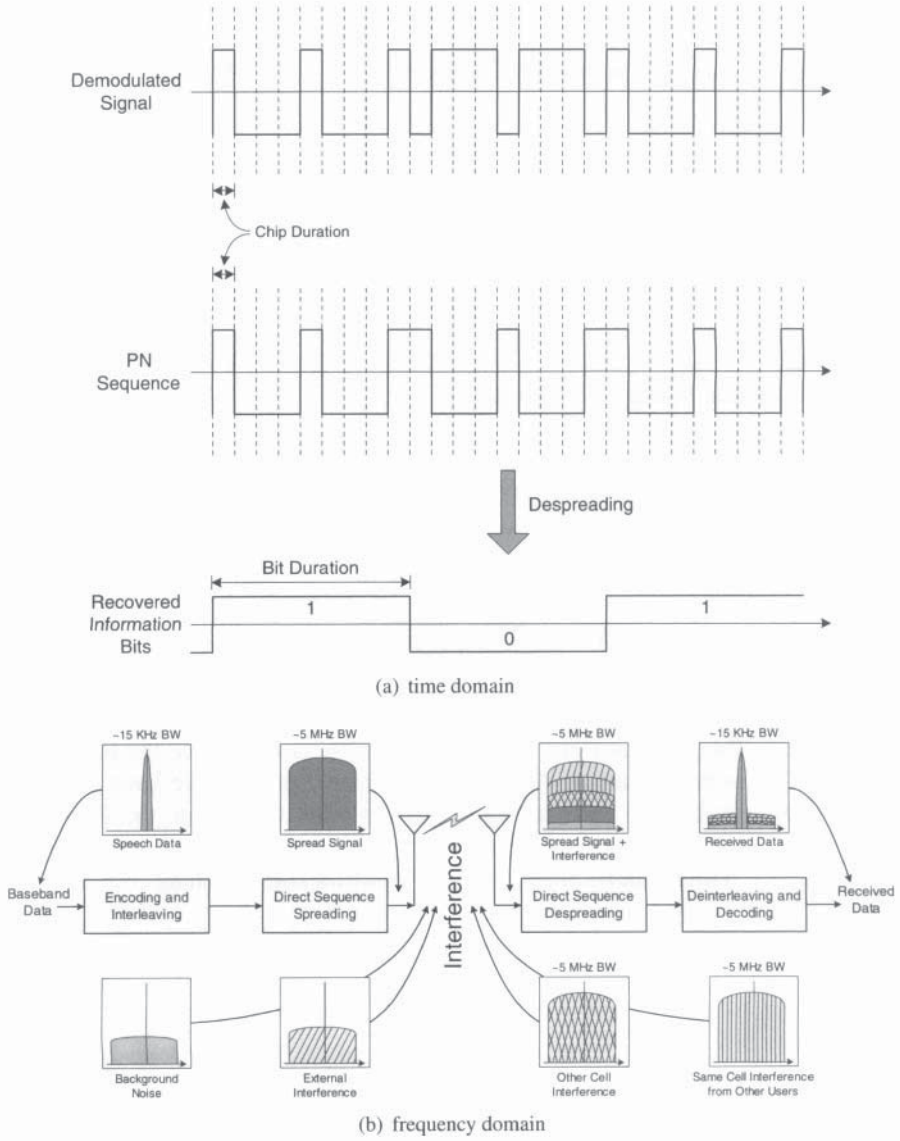


Figure 3.7 Principle of CDMA despreading.

We derive the condition on code sequence design in order to achieve deterministic CDMA. If the code sequence (periodic with period T_s/T_c) is expressed in discrete time domain $c_k(t) = \sum_{n=0}^{T_s/T_c-1} c_{k,n}u(t-nT_c)$ where $c_{k,n} \in \{+1,-1\}$ and $u(t)$ is a rectangular unit pulse with duration T_c, the received signal is expressed as:

$$W_k(t) = S_k(t) + \sum_{j\neq k} S_j(t) \sum_{n=0}^{T_s/T_c-1} c_{k,n}c_{j,n}u(t-nT_c) + Z(t) \sum_{n=0}^{T_s/T_c-1} c_{k,n}u(t-nT_c) \tag{3.12}$$

Theorem 2 *If $T_c \ll T_s$ and the code sequence is chosen such that*

$$R_{k,j} = \frac{1}{SF} \sum_{n=0}^{SF-1} c_{k,n}c_{j,n} = 0 \tag{3.13}$$

where $SF = \frac{T_s}{T_b}$ is the spreading factor, *the interference component in (3.12) becomes* 0 *and the resulting CDMA scheme achieves* deterministic CDMA.

Proof: Converting $W_k(t)$ to a signal vector $\mathbf{W}_k = (W_{k,1}, \ldots, W_{k,D})$ in a $D-$dimensional signal space with basis $\{\phi_1(t), \ldots, \phi_D(t)\}$, we have:

$$\mathbf{W}_k = (< W_k(t), \phi_1(t) >, \ldots, < W_k(t), \phi_D(t) >)$$

where $< W_k(t), \phi_d(t) >$ denotes the *dot-product* of the signal space and is given by:

$$\begin{aligned} < W_k(t), \phi_d(t) > &= \int_0^{T_s} W_k(t)\phi_d^*(t)dt \\ &= \underbrace{\int_0^{T_s} S_k(t)\phi_d^*(t)dt}_{\text{DesiredSignal}} + \underbrace{\sum_{j\neq k} \sum_{n=0}^{T_s/T_c-1} c_{k,n}c_{j,n} \int_0^{T_s} S_j(t)\phi_d^*(t)u(t-nT_c)dt}_{\text{Multiuserinterference}} \\ &+ \underbrace{\sum_{n=0}^{T_s/T_c-1} c_{k,n} \int_0^{T_s} Z(t)\phi_d^*(t)u(t-nT_c)dt}_{\text{Noise}} \end{aligned}$$

In general, since $T_c \ll T_s$, we have $f(n) = \int_0^{T_s} S_j(t)\phi_d^*(t)u(t-nT_c)dt = \int_{nT_c}^{(n+1)T_c} S_j(t)\phi_d^*(t)u(t-nT_c)dt$ is *approximately* independent of n. Hence, the multi-user interference term could be expressed as $f\sum_{j\neq k} c_{k,n}c_{j,n}$. If $R_{k,j}(0) = 0$, the multi-user interference term vanishes.

Hence, the despread signal vector of user k is given by $\mathbf{W}_k = \mathbf{S}_k + \mathbf{Z}_k$ where $\mathbf{S}_k$ is the desired signal vector and $\mathbf{Z}_k$ is the channel noise vector.

Similar to TDMA and FDMA, the capacity of the resource partitioning is given by the number of available orthogonal channels. In deterministic CDMA, the number of available channels is equal to the number of code sequences with spreading factor $SF = T_s/T_c$ that satisfy the *orthogonality requirement* in (3.13). In other words, the capacity of deterministic CDMA system is *code limited*. Since the code dimension is created by the spreading operation, the degree of freedom of the code dimension is equal to the spreading factor $SF = T_s/T_c$ and therefore, the number orthogonal code sequences available in the system is given by the spreading factor SF. Furthermore, complete orthogonality of code sequence requires perfect synchronization between $c_k(t)$ and $c_j(t)$. In general, it is easier to achieve this condition in the downlink direction. This is because signals for the K users originates from the same point (the base station). As long as the transmitted codes are synchronized at the transmitting point, the same timing relationship could be maintained at the mobile and therefore, deterministic CDMA is feasible in the downlink direction. On the other hand, it is more difficult to achieve this synchronous state between code sequence in the uplink direction. This is because the code sequence $c_k(t)$ and $c_j(t)$ originates from user k and user j respectively and they are in general at different locations and uncoordinated. Even if the two codes start at the same system time, this timing relationship could not be maintained at the receiving point (the base station) because they have different propagation delays. Hence, deterministic CDMA is not applied in the uplink direction.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Let the modulation throughput be 1 bit per symbol and the required average data rate of individual user be 25 kbps. Find out how many users could be supported based on the deterministic CDMA scheme.

Solution: With binary modulation, the maximum chip rate based on a bandwidth of 15 MHz is 15 Mcps. Hence, the spreading factor of the D-CDMA system is $SF = T_s/T_c = 15M/25k = 600$. Hence, there are 600 code channels in D-CDMA system.

Note that the maximum number of spreading code is limited by the *spreading factor* $SF = T_s/T_c$. Given a fixed bit rate (and hence a fixed bit duration T_b), the symbol duration $T_s = \log_2(M)T_b$ is a function of the modulation level M. For example, $T_s(QPSK) = 2T_b$ and $T_s(16QAM) = 4T_b$. Hence, the number of orthogonal channels could be expanded by using higher order modulation. The tradeoff is possibly higher required bit-energy to noise ratio as a result of higher order modulation.

D-CDMA shares the same advantages as the TDMA system in the sense that single RF transceiver is needed regardless the number of users in the system because all the K user channels share the same RF channel. On the other hand, since despreading operation require re-construction of the desired signal component by multiplying the corresponding code sequence to the received signal, accurate timing synchronization is required so that the despreading sequence $c_k(t)$ is time aligned with the spreading sequence $c_k(t)$ embedded in the received signal.

Similar to TDMA, the transmitted signal has high bandwidth due to the spreading operation. Hence, multipath fading will be present and in general, expensive channel equalization is needed at the receive to alleviate the inter-symbol interference (ISI). However, due to property of code sequence, a simple equalization structure exists for CDMA and this is called the *RAKE receiver*.

Finally, in terms of practical implementation, the implementation bottleneck of D-CDMA is usually on the maximum supportable chip rate $1/T_c$. In particular, the D-CDMA receiver front end operates at a speed of the chip rate order and there is a natural limitation on how fast the integrated circuit could operate. Therefore, in general, operators seldom partition the channel based entirely on D-CDMA.

3.3.4 Hybrid Design of FDMA, TDMA and D-CDMA

As introduced in the previous sections, FDMA, TDMA and Deterministic CDMA are all orthogonal resource partitioning schemes. As illustrated in the examples, the number of channels as a result of the 3 partitioning methods are the same for single cell system. (We elaborate the comparison more in Section 3.5) and this is in general always the case. Therefore, all the 3 schemes are equally effective theoretically. However, each of the schemes have their own implementation limitations. For example, FDMA system suffers a disadvantage of requiring one RF element per user. TDMA system suffers from a disadvantage of bursty transmission with very high peak rate. CDMA system suffers from a disadvantage of high speed signal processing at the chip rate level. In general, system resource is partitioned by either hybrid FDMA/TDMA or hybrid FDMA/CDMA schemes and they are introduced below through two examples.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Let the modulation throughput be 1 bit per symbol and the required average data rate of individual user be 25 kbps. Assume that the channel coherence bandwidth is 200kHz. Suggest how the operator should partition the system resource based on FDMA and TDMA so that no equalizer is needed at the receiver.

Solution: Figure 3.8 illustrates the hybrid partitioning scheme.

Since the channel coherence bandwidth is 200kHz, we constraint the maximum bandwidth of signal transmission to be 200kHz so that expensive equalization is not required at the receiving side. Hence, the 15 MHz spectrum is first partitioned into $15M/200k = 75$ carriers based on FDMA as illustrated in the figure. Each carrier is then further partitioned into N_t channels using TDMA. Similar to the example described earlier, the number of time slots (channels) per carrier is given by:

$$N_t = 200k/25k = 8.$$

The total number of channels in the system is given by $N_t N_f = 600$ which is the same as partitioning based on FDMA or TDMA alone. Using the hybrid scheme, the number of RF transceivers required at the base station is reduced to 75 instead of 600 (for pure FDMA). On the other hand, the transmitted bandwidth of user signal

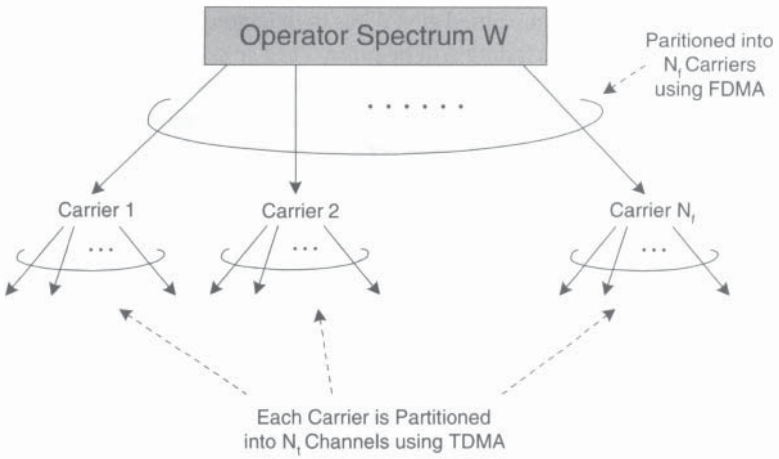


Figure 3.8 Hybrid FDMA and TDMA.

is reduced to $200k$Hz (instead of 15MHz as in pure TDMA) and therefore, avoiding expensive channel equalization.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Let the modulation throughput be 1 bit per symbol and the required average data rate of individual user be 25 kbps. Assume that maximum supportable chip rate is 1.5Mcps. Suggest how the operator should partition the system resource based on FDMA and D-CDMA so that no equalizer is needed at the receiver.

Solution: Figure 3.9 illustrates the hybrid partitioning scheme.

Since the maximum supportable chip rate is 1.5 Mcps and the modulation throughput is 1 bits per symbol, the D-CDMA signal requires 1.5MHz transmission bandwidth. Hence, the 15 MHz spectrum is first partitioned into $15M/1.5M = 10$ carriers based on FDMA as illustrated in the figure. Each carrier is then further partitioned into N_c channels using D-CDMA. The maximum spreading factor per carrier becomes $T_s/T_c = 1.5M/25k = 60$ and hence, the number of code channels per carrier is given by:

$$N_c = 60.$$

The total number of channels in the system is given by $N_c N_f = 600$ which is the same as partitioning based on FDMA or D-CDMA alone. Using the hybrid scheme, the number of RF transceivers required at the base station is reduced to 75 instead of 600 (for pure FDMA). On the other hand, the maximum chip rate is reduced to 1.5Mcps (instead of 15Mcps as in pure D-CDMA).

In both of the hybrid schemes, FDMA is used to partition the spectrum into a few *carriers* of smaller bandwidth. Each of the carrier is further partitioned based on either TDMA or D-CDMA schemes to a number of *user channels*. In this way,

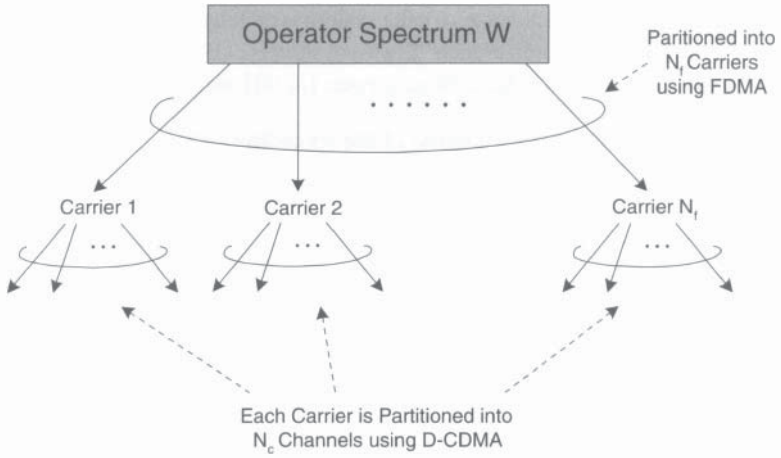


Figure 3.9 Hybrid FDMA and D-CDMA.

we could limit the number of RF units required due to FDMA and at the same time limit the peak rate transmission (for TDMA) or maximum supportable chip rate (for D-CDMA) of individual carriers. The total number of channels available after the hybrid partitioning are the same.

3.4 NON-ORTHOGONAL RESOURCE PARTITIONING

In the previous section, we introduce a few common resource partitioning schemes based on the orthogonality principle. The resulting channels are orthogonal to each other, meaning that users on different channels do not interfere with each other. However, as we could see in Section 3.5, the orthogonal resource partitioning schemes are in fact not the optimal schemes. They are chosen because of low implementation complexity. In this section, we introduce a few other resource partitioning scheme that are *non-orthogonal*. In other words, users on different channels do interfere with each other but in a controlled manner.

3.4.1 Random Code Division Multiple Access (R-CDMA)

Similar to deterministic CDMA, each user is assigned a spreading code sequence $c_k(t) = \sum_n c_{k,n} u(t - nT_c)$ where $c_{k,n} \in \{+1, -1\}$ and $u(t - nT_c)$ is a unit pulse with interval $t \in [nT_c, (n + 1)T_c]$. If we choose the code sequence such that the orthogonality condition in (3.13) is satisfied, the system is limited by the number of codes. On the other hand, if we relax the orthogonality condition slightly to tolerate

small cross-correlation (instead of zero), the system is no longer limited by the number of codes but limited by the interference. Such a system is called *random CDMA* [339].

Specifically, the criteria of the code sequence $\{c_k(t)\}$ satisfies:

Auto-correlation: The auto-correlation of the spreading code:

$$R_{k,k}(\tau) = \frac{1}{T_s}\int_0^{T_s} c_k(t)c_k(t+\tau)dt$$

has sharp fall off to a small *residual value* at $\tau \geq T_c$. This is illustrated in Figure 3.10

Time-averaged Cross-correlation: The cross-correlation of the spreading code:

$$R_{k,j}(\tau) = \frac{1}{T_s}\int_0^{T_s} c_k(t)c_j(t+\tau)dt$$

has small value for all τ.

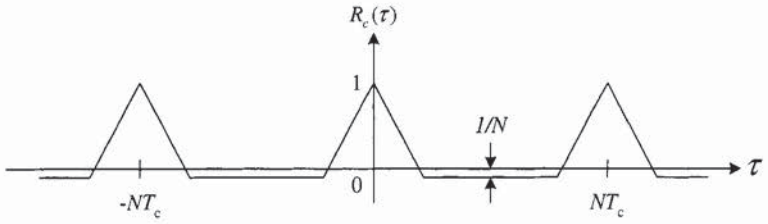


Figure 3.10 Requirement of autocorrelation property of spreading code sequence.

The transmitted signal of user k is given by $S_k(t)c_k(t)$ and the received signal as a result of K users transmission through the AWGN channel is given by:

$$Y(t) = \sum_{k=1}^{K} S_k(t)c_k(t) + Z(t)$$

There are several ways to extract information signal for user k. One simple receiving technique, which has been widely applied in today's CDMA systems, is called *isolated decoding*. Specifically, signals due to other users are treated as interference and the desired information signal of user k is extracted through despreading with respect to $c_k(t)$. That is:

$$W_k(t) = Y(t)c_k(t) = \underbrace{S_k(t)}_{\substack{\text{Signal}\\ \text{Component}}} + \underbrace{\sum_{j\neq k} S_j(t)c_j(t)c_k(t)}_{\text{MultiuserInterference}} + \underbrace{Z(t)c_k(t)}_{\substack{\text{Channel}\\ \text{Noise}}}$$

Expressing $W_k(t)$ as a vector in a signal space in a similar way as in the previous section, we have:

$$\mathbf{W}_k = \mathbf{S}_k + \sum_{j \neq k} \sum_n c_{j,n} c_{k,n} \mathbf{V}_{j,n} + \mathbf{Z}_k \tag{3.14}$$

where $\mathbf{S}_k$ is the information signal vector of user k, $\mathbf{Z}_k$ is the noise vector projected onto the signal space and $\mathbf{V}_{j,n}(d) = \int_{nT_c}^{(n+1)T_c} S_j(t)\phi_d^*(t)dt$ represents the $d-$th component of the multi-user interference vector contributed by user j. Unlike deterministic CDMA, the *multi-user interference* term $\sum_{j \neq k} \sum_n c_{j,n} c_{k,n} \mathbf{V}_{j,n}$ is non-zero and is a random process. In general, the power of the interference component depends on the specific choice of code sequences. Detail description of the code design[288] and performance is outside the scope of the book. To capture the performance of random CDMA based on isolated detection, we employ a statistical approach.

Assume that the code sequences for the K users $\{c_1(t), \ldots, c_K(t)\}$ are *random sequences* given by:

$$c_k(t) = \sum_n c_{k,n} u(t - nT_c) \tag{3.15}$$

where $c_{k,n}$ is an identical and independent distributed (i.i.d.) random variable drawn from $\{+1, -1\}$ with equal probability. Hence, we have $\mathcal{E}[c_{k,n}] = 0$, $\mathcal{E}[c_{k,n}^2] = 1$, $\mathcal{E}[c_{k,n}c_{k,m}] = 0$ and $\mathcal{E}[c_{k,n}c_{j,m}] = 0$.

Since the received signal vector $\mathbf{W}_k$ in (3.14) is a *sufficient statistic* for the information signal $\mathbf{S}_k$, the maximal likelihood (ML) detection rule on $\mathbf{S}_k$ is given by:

$$\text{argmax}_{\mathbf{S}_k} [p(\mathbf{W}_k|\mathbf{S}_k)] \tag{3.16}$$

For sufficiently large number of users, the multi-user interference becomes Gaussian. Hence, together with the Gaussian channel noise vector, the decision vector $\mathbf{W}_k$ is a Gaussian vector. To evaluate the performance of the ML detection, we have to evaluate the mean and co-variance of the received signal vector $\mathbf{W}_k$.

Conditional Mean of $\mathbf{W}_k$:

Given $\mathbf{S}_k$, the mean of $\mathbf{W}_k$ is given by:

$$\mathcal{E}[\mathbf{W}_k|\mathbf{S}_k] = \mathbf{S}_k$$

The co-variance of the Noise Term:

For AWGN channel, the noise vector $\mathbf{Z}_k$ is a zero-mean Gaussian random vector with covariance matrix given by $\mathcal{E}[\mathbf{Z}_k\mathbf{Z}_k^*] = \eta_0\mathbf{I}_D$ where η_0 is the noise spectral density of $Z(t)$.

The co-variance of the Multiuser Interference Term:

When the number of users K is sufficiently large, the interference term $\mathbf{V}_I = \sum_j \sum_n c_{j,b} c_{k,n} \mathbf{V}_{j,n}$ is approximately Gaussian. We have the mean given by:

$$\mathcal{E}[\mathbf{V}_I|\mathbf{S}_k] = \mathbf{0} \tag{3.17}$$

because $\mathcal{E}[R_{j,k}(mT_c)] = 0$.

and the covariance matrix given by:

$$\begin{aligned}
&\mathcal{E}[\mathbf{V}_I \mathbf{V}_I^* | \mathbf{S}_k] \\
&= \sum_{j \neq k} \sum_n \mathcal{E}[|c_{k,n}|^2] \mathcal{E}[|c_{j,n}|^2] \left| \int_{nT_c}^{(n+1)T_c} S_j(t) \phi_d^*(t) dt \right|^2 \mathbf{I} \\
&\leq \sum_{j \neq k} \sum_n \int_{nT_c}^{(n+1)T_c} |S_j(t)|^2 dt \int_{nT_c}^{(n+1)T_c} |\phi_d^*(t)|^2 dt \mathbf{I} \\
&= \frac{T_c}{T_s} \sum_{j \neq k} \int_0^{T_s} |S_j(t)|^2 dt \mathbf{I} \\
&= \sum_{j \neq k} E_s(j) \mathbf{I}
\end{aligned}$$

where $E_s(j) = \int_0^{T_s} |S_j(t)|2dt$ is the *symbol energy* due to user j and the inequality is due to Cauchy-Swartz inequality.

Hence, the log-likelihood detection rule in (3.16) is equivalent to the *minimum distance decoding*:

$$\text{argmax}_{\mathbf{S}_k} [\|\mathbf{W}_k - \mathbf{S}_k\|^2] \tag{3.18}$$

Let $E_b(k) = P_k T_b$ be the bit energy of user k and $E_s(j) = P_j T_s$ be the symbol energy of user j. The error performance of such a decoder is usually expressed as a function of *bit-energy to variance ratio* $\gamma = \frac{E_b}{\sigma^2}$ where σ^2 is the total variance of the channel noise and multiuser interference. Hence, γ of random CDMA is given by:

$$\gamma_k = \frac{E_b}{\eta_0 + \frac{T_c}{T_s} \sum_{j \neq k} E_s(j)} = \frac{P_k}{\eta_0 W + \sum_{j \neq k} P_j} \frac{T_b}{T_c} \tag{3.19}$$

where $W = \frac{1}{T_c}$ and P_j is the transmitted power of user j. From this equation, we could see that the γ is given by the SNIR (*signal-to-noise plus interference* power ratio of the $k-$th user) multiplied by a factor $\frac{T_b}{T_c}$. This factor is called the *processing gain* of random CDMA.

$$PG = \frac{T_b}{T_c} \tag{3.20}$$

Hence, the key characteristic of random CDMA is its ability to suppress interference power by a factor of PG. Hence, signals from multiple users could be transmitting using the same time and frequency dimensions. The *associated interference* power as a result of transmitting in the same frequency and the same time is reduced by the action of spreading and despreading. Note that the interference is *suppressed* but not eliminated. Let us consider an example to illustrate the operation of random CDMA.

Example: In a 3-user uplink system, signals from user 1, 2, 3 are received at power levels of 20dBm, 18 dBm and 23dBm respectively at the base station. All users use BPSK modulation and transmit at a bit rate of 30 kbps. Random CDMA is applied on the uplink with a chip rate of 3M cps. That is: user k is spread with a code sequence $c_k(t)$ according to equation (3.15). Find out the BER performance of the 2nd user where the BER performance of BPSK modulation in AWGN channel is given by $Q\left(\sqrt{(\frac{2E_b}{\eta_0})}\right)$.

Solution: The processing gain is given by $PG = \frac{T_b}{T_c} = W/R_b = 100$. The transmit bandwidth = 3MHz and hence, the noise power (in linear scale) is given by $\eta_0 W = 10^{-13.9}$. The received power levels for the 3 users (in linear scale) are 0.1, 0.063 and 0.2 respectively. For the 2nd user, the bit energy to noise plus interference density (γ_2) is given by:

$$\gamma_2 = 100\frac{P_2}{\eta_0 W + P_1 + P_3} = 21$$

Hence, the BER performance of the user 2 is given by $Q(\sqrt{(2\gamma_2)}) \approx 10^{-8}$.

Unlike deterministic CDMA where the maximum number of channels is limited by the number of available orthogonal code sequence, the random CDMA system is *interference limited*. In other words, since we relaxed the requirement on the cross-correlation of different code sequence to be non-zero, the number of code sequence satisfying this mild criteria is abundant. Therefore, the system is no longer *code limited*. However, channels become *non-orthogonal* and every time a new user is added to the system, the overall interference level to all other users increases. Eventually, the system capacity is limited by the maximal tolerable interference level. This point is illustrated by an example below.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Assume that binary modulation is used and the required average data rate of individual user be 25 kbps. Find out how many channels could be partitioned based on random CDMA scheme alone if the maximal *tolerable* E_b/σ is given by 7dB (assuming mobiles are transmitting at infinity power).

Solution: Given a bandwidth of 15MHz and a bit rate of 25kbps, the processing gain is given by: $PG = T_b/T_c = W/R_b = 600$. From (3.19), the asymptotic number of channels as $P_j \to \infty$ is given by:

$$N_{R-CDMA,\infty} = 1 + \frac{600}{\gamma_{\text{target}}} = 1 + \frac{600}{5} = 121.$$

Note that PG is the operational parameter of R-CDMA. Furthermore, *spreading factor* and *processing gain* are two *different* parameters for deterministic CDMA and random CDMA respectively. The spreading factor governs the maximum number of orthogonal spreading code sequence. On the other hand, the *processing gain* determines the ability to *suppress* multiuser interference. When binary modulation is

used, $T_s = T_b$ and therefore, the two parameters are equivalent. On the other hand, when $M-$ary modulation is used, the two parameters is different. For instance, given a fixed bit rate (and therefore a fixed T_b), number of orthogonal channels supported by deterministic CDMA could be increased by using higher order modulations (e.g., 16QAM) at the expense of higher required γ. Hence, there is a natural tradeoff between the number of orthogonal channels and the required γ for deterministic CDMA using higher order modulation.

However, the processing gain of random CDMA (and therefore, the interference suppression capability) is independent of the modulation level. In other words, it is always sub-optimal to apply higher level of modulation in random CDMA. That is why most commercial systems operating at random CDMA have either BPSK or QPSK modulation but not 8PSK, 16QAM. This point is better illustrated using the following example.

Example: Suppose a spectrum of 15 MHz is allocated to a mobile operator. Possible modulation schemes include BPSK, QPSK, 16QAM modulations. The required average data rate of individual user be 25 kbps. At a given target BER of 10^{-3}, the required $\gamma_{dB} = 10\log_{10}(E_b/\sigma^2)$ for the 3 modulation schemes are 7 dB, 7 dB, 17 dB respectively. Which modulation is the best choice for R-CDMA system if we assume that all users transmit at infinite power?

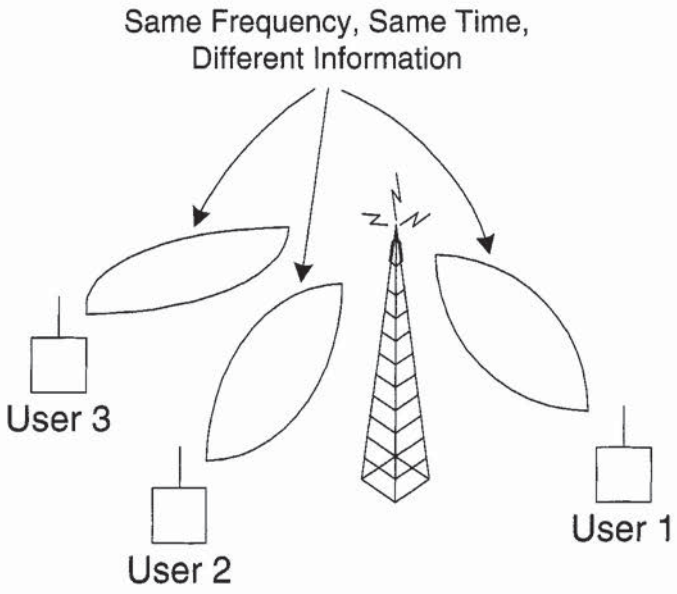
Solution: The *processing gain* of R-CDMA system is given by $PG = T_b/T_c = W/R_b = 15M/25k = 600$. The *asymptotic number of channels* available in R-CDMA with such processing gain is given by:

$$N_{R-CDMA,\infty} = 1 + \frac{600}{\gamma_{\text{target}}}$$

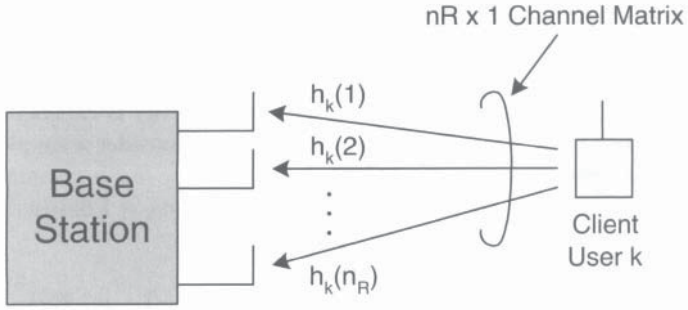
Obviously, using 16 QAM is a sub-optimal choice because the required γ_{dB} is the largest. On the other hand, it is equally effective to use either QPSK or BPSK in R-CDMA system because they require the same bit-energy to noise plus interference density ratio.

3.4.2 Space Division Multiple Access (SDMA)

Another new techniques of non-orthogonal resource partitioning is to utilize the spatial dimension through multiple antenna technologies. Specifically, SDMA serves different users by utilizing the fact that each users are located in different "positions" and are therefore *separable*. Hence, different users could transmit signals based on the same frequency and the same time duration but via different *spatial channels*. For line-of-sight propagation environment, one could visualize the process as forming different narrow beams to focus the transmission to individual users situated in different spatial positions as illustrated in Figure 3.11(a). However, the concept of beams may not apply in a more general environment, namely the *scattering environment* where there is no line of sight propagation. In this case, a more general concept, namely *spatial filtering*, is more appropriate.



(a) illustration of SDMA using concepts of beam-forming



(b) channel matrix

Figure 3.11 Illustration of SDMA using concepts of beam-forming.

We illustrate the concept using uplink as an example. Assume that the base station has n_R antennas and the client station has single antenna. Let $S_k(t)$ be the transmitted signal from user k and $\mathbf{Y}(t)$ be the $n_R \times 1$ dimensional received signal vector at the base station. Each client user is linked up to the base station through an $n_R \times 1$ dimension complex channel matrix, $\mathbf{h}_k$ as illustrated in Figure 3.11(b). The channel matrix is in general a complex random process and the distribution depends on the propagation environment. In one extreme case when we have complete line of sight, $\mathbf{h}_k$ becomes a deterministic quantity whose amplitude is 1 and phase value depends on the relative position between the client user and the base station. In another extreme when we have complete scattering environment, the channel matrix is Rayleigh distributed in amplitude and uniformly distributed in phase. In some intermediate cases, the channel matrix is a mixture between the two ends and it could be characterized by Rician distribution. In any case, the channel matrix is normalized such that

$$\mathcal{E}[\mathbf{h}_k \mathbf{h}_k^*] = \mathbf{I}_{n_R}$$

The received signal vector is given by:

$$\mathbf{Y}(t) = \mathbf{h}_k S_k(t) + \sum_{j \neq k} \mathbf{h}_j S_j(t) + Z(t) \tag{3.21}$$

where $\mathbf{h}_k \neq \mathbf{h}_j$ for all $k \neq j$. The first term represents the desired signal and the middle term represents the multi-user interference. The last term represents the channel noise. Similar to R-CDMA, all signals are transmitted at the same frequency range and the same time but have different *spatial signature* $\mathbf{h}_k$ and hence, they are separable. Hence, there is interference between the user signals and therefore, the SDMA is a non-orthogonal resource partitioning scheme.

There are various ways to separate signals from the K users. The optimal scheme is given by the joint-detection and successive interference scheme as suggested by information theory. However, this implementation complexity is exponential in K and is not feasible in practice. Similar to R-CDMA, we consider a simple and effective method to separate signals from the K users, namely *linear spatial filtering*. Specifically, to extract information from user k, we apply a $n_R \times 1$ dimension *linear complex weight* $\mathbf{w}_k$ to the received signal vector. That is:

$$X_k(t) = \mathbf{w}_k^* \mathbf{Y}(t) = \mathbf{w}_k^* \mathbf{h}_k S_k(t) + \sum_{j \neq k} \mathbf{w}_k^* \mathbf{h}_j S_j(t) + \mathbf{w}_k^* Z(t)$$

The problem becomes how to choose an appropriate weight $\mathbf{w}_k$. In general, there could be several design objectives, such as *interference nulling* and *minimum mean-square error*. For interference nulling, the problem is: Choose weight $\mathbf{w}_k$ such that

$$\mathbf{w}_k^* \mathbf{h}_k = 1.$$

$$\mathbf{w}_k^* \mathbf{h}_j = 0 \quad \forall j \neq k$$

For minimum mean-square error (MMSE), the optimization problem becomes: Choose weight $\mathbf{w}_k$ such that the Mean-Square-Error (MSE) is minimized.

$$\text{argmin}_{\mathbf{w}_k} \left(\mathcal{E} \left[|\mathbf{w}_k^* \mathbf{Y}(t) - S_k(t)|^2 \right] \right)$$

For interference nulling, we are solving a system of linear equation and in order to have solution, we have:

$$K \leq n_R \tag{3.22}$$

Hence, the maximum number of *spatial channel* as a result of interference nulling spatial filtering is the number of antennas at the base station (n_R). Note that interference nulling gives n_R orthogonal spatial channels at the expense of enhanced noise power.

On the other hand, for MMSE spatial filtering, the optimal weight is given by:

$$\mathbf{w}_k = \mathbf{\Phi}^{-1} \mathbf{p}_k$$

where $\mathbf{\Phi} = \mathcal{E}[\mathbf{Y}(t)\mathbf{Y}^*(t)]$ and $\mathbf{p}_k = \mathcal{E}[\mathbf{Y}(t)S_k^*(t)]$. If we look at the MMSE filtered output for user k, it consists of the desired signal for user k, the multi-user interference (*suppressed* by MMSE filtering) and the residual channel noise. Therefore, the spatial channels are not *orthogonal* with each other and they suffer from multi-user interference but with a suppressed power due to spatial filtering. Similar to R-CDMA, the number of spatial channels available in the system is *interference limited* and depends on the maximal tolerable interference level.

3.5 SPECTRAL EFFICIENCY AND PERFORMANCE ISSUES

In the previous sections, we have introduced an information theoretical framework for multi-user communications. We have also introduced various common resource partitioning schemes, resulting in both orthogonal and non-orthogonal channels. In this section, we give a performance comparison of the various schemes in both single cell and multi-cell scenarios.

3.5.1 Single Cell Comparison

We first assume that the system consists of a single cell and K mobile users. In this scenario, a very important multiuser performance metric is the system capacity given by $\sum_k R_k$ (R_k) is the data rate of user k. On the other hand, when all users have the same data rate (such as voice applications), the system capacity could be expressed as the number of users (or number of channels) that could be supported. As we illustrate below, all the orthogonal resource partitioning schemes (namely FDMA, TDMA and D-CDMA) are *equally effective*. On the other hand, the performance of non-orthogonal resource partitioning schemes depends heavily on the decoding methods employed. For linear complexity signal separation as we have illustrated in the

previous section, the performance is *interference limited* and is in general worst than that of the orthogonal schemes. However, when more sophisticated interference cancellation schemes are employed (such as multi-user joint detection), the performance of non-orthogonal schemes are actually better than that of the orthogonal schemes.

Performance of FDMA System:

Let W be the spectrum available for resource partitioning, and R_b be the required average bit rate per user. The transmission bandwidth per user is given by:

$$W_k = R_b / \log_2(M)$$

where M is the modulation level in terms of bits per modulation symbol. Since the bit rate per user is the same for all users, the number of channels partitioned by FDMA is given by:

$$N_{FDMA} = W/W_k = (W/R_b)\log_2(M)$$

Performance of TDMA System:

Let T_p be the frame duration of a TDMA frame and T_k be the width of the time slot assigned to user k. The peak bit rate of user k during the assigned time slot is given by $R_p = W\log_2(M)$. Hence, the average bit rate per user k is given by:

$$R_b = P_p(T_k/T_p) = W\log_2(M)T_k/T_p$$

The total number of time slots available is thus given by:

$$N_{TDMA} = T_p/T_k = (W/R_b)\log_2(M)$$

Performance of D-CDMA Systems:

The modulation symbol duration is given by $T_s = T_b\log_2(M) = \log_2(M)/R_b$. With a bandwidth of W, the chip duration $T_c = 1/W$. Hence, the *spreading factor* of D-CDMA is given by:

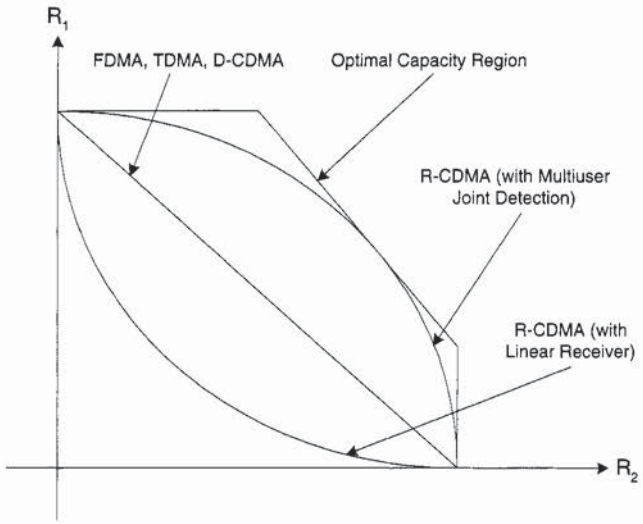
$$SF = T_s/T_c = \log_2(M)(W/R_b)$$

Hence, the total number of channels available for D-CDMA is given by:

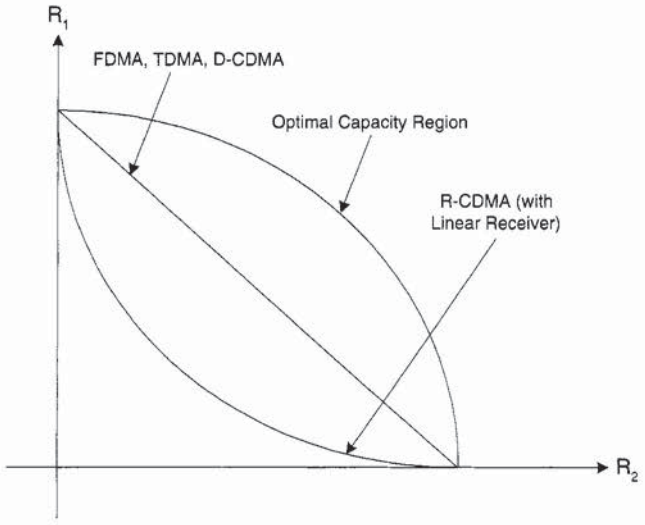
$$N_{D-CDMA} = \log_2(M)(W/R_b)$$

As could be seen from the above results, we observe that $N_{FDMA} = N_{TDMA} = N_{D-CDMA}$. In other words, they are equally effective in terms of system capacity. Furthermore, observe that the system capacity could be enhanced by increasing the modulation level M. Yet, the price to pay is increased signal to noise ratio to maintain the BER target.

Comparing the orthogonal resource partitioning schemes with the information theoretical (optimal) results, we could see that the orthogonal schemes are in fact, sub-optimal. This is illustrated in the *capacity region* of the multiple access (uplink) and broadcast (downlink) problems in Figures 3.12(a) and 3.12(b) respectively.



(a) multiples access region



(b) broadcast region

Figure 3.12 Comparison of FDMA, TDMA, D-CDMA and R-CDMA with respect to the information theoretical capacity region.

Performance of R-CDMA:

For R-CDMA with *isolated decoding*, the system capacity is *interference limited* instead of code limited. In other words, as more and more users are added to the system, the background interference level grows and individual existing users have to further increase their transmit powers to offset the rise in background interference (so that the target SNR could be maintained.). There reaches a point when no more users could be added and every existing users are transmitting at infinite power levels. When this happens, we said the capacity of the R-CDMA system reaches the *pole capacity*. This concept is illustrated below.

Consider the reverse link as a simple illustration of R-CDMA. As illustrated in the previous section, the received bit-energy-to noise plus interference level of user k, γ_k, is given by:

$$\gamma_k = \frac{P_k}{\eta_0 W + \sum_{j \neq k} P_j} \frac{T_b}{T_c} \tag{3.23}$$

where P_j is the received signal power from user j. Assume that to maintain a target BER for user k, the required γ_k has to be at least γ_{target}. This translates into the requirement:

$$\frac{P_k}{\eta_0 W + \sum_{j \neq k} P_j} \frac{T_b}{T_c} \geq \gamma_{\text{target}}$$

For voice applications, the source rate is quite steady and the real time speech frames could not be queued. Therefore, by symmetry, the optimal operating point (with respect to the number of users that could be accommodated) is when all users signal powers are equal and given by P. In this case, we have:

$$\frac{P}{\eta_0 W + (K-1)P} \frac{T_b}{T_c} \geq \gamma_{\text{target}}$$

Therefore, the capacity of R-CDMA is given by:

$$N_{R-CDMA} \leq \left(\frac{T_b/T_c}{\gamma_{\text{target}}} - \frac{\eta_0 W}{P} + 1 \right)$$

Figure 3.13 illustrates the required received power P versus the number of users in the system N_{R-CDMA}. Observe that in the *lightly loaded* range, as more users are added to the R-CDMA system, there is a proportional increase in the received signal power P to compensate for the rise in background interference. However, as the system gets more and more *heavily loaded*, the corresponding increase in the received signal power required rise in a dis-proportional manner and *diminishing return* is observed.

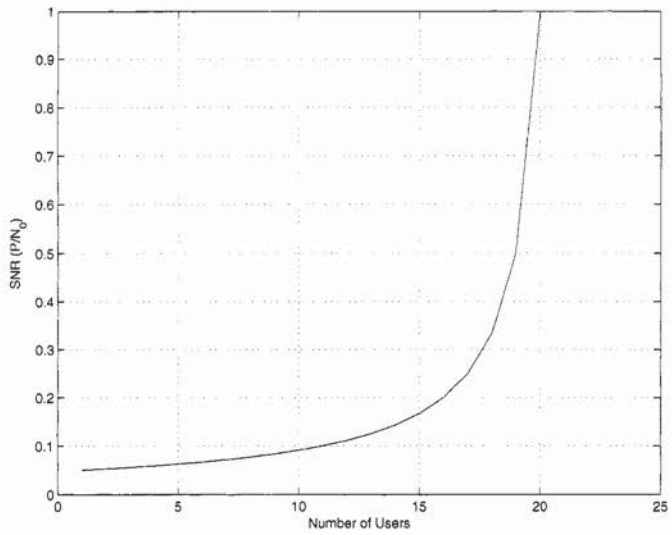


Figure 3.13 Received power versus the system capacity in R-CDMA (isolated decoding). $T_b/T_c = 100$, target $\gamma = 5$.

In the limit when $P \rightarrow \infty$, we have the *pole capacity*[3] given by:

$$N_{R-CDMA}(\infty) = 1 + \frac{T_b/T_c}{\gamma_{\text{target}}} = 1 + \frac{W/R_b}{\gamma_{\text{target}}} \tag{3.24}$$

In general, the required γ_{target} is a function of the modulation level and coding schemes employed. In most cases, the required γ_{target} would be of the order 7 dB for a BER of 10^{-3}. Hence, we could easily see that the pole capacity N_{R-CDMA} is less than the system capacity for FDMA, TDMA or D-CDMA. The performance of R-CDMA (with isolated decoding) with respect to the information theoretical performance is illustrated in Figure 3.12(a). On the other hand, when *multiuser joint-detection* is used, the performance of R-CDMA is substantially improved. This is because with joint detection, multi-user interference could be cancelled between successive decoding stages. Yet, the implementation complexity of such receiver is in general very high and it is outside the scope of the book to discuss the topics on multiuser detection.

Another point to make is on the relationship between the R-CDMA capacity N_{R-CDMA} and the modulation level. Unlike the orthogonal schemes (FDMA, TDMA, D-CDMA) where the number of channels is increased by increasing the

[3]It is called the pole capacity because to achieve $P \rightarrow \infty$, one way is to have all the users transmitting at finite power but situated arbitrarily close to the base station antenna. Therefore, it is called the pole capacity.

modulation level M, the system capacity of R-CDMA depends solely on the required γ. In general, using higher modulation level (for bandwidth efficient modulation schemes such as 8PSK, 16QAM, 64QAM) requires higher γ to maintain the same BER, which is detrimental to the system capacity. Therefore, we seldom use anything beyond QPSK modulation for R-CDMA systems.

3.5.2 Multi-Cell Comparison

The previous section compares the system performance of various resource partitioning schemes in single cell environment. Here, we extend the comparison to consider multi-cell environment[203] where the service area is covered by multiple cells. Although FDMA/TDMA/D-CDMA achieves better performance compared to R-CDMA (isolated detection) in single cell environment, we illustrate that R-CDMA (isolated detection) performs better than the orthogonal resource partitioning schemes in multi-cell environment.

Resource Reuse in Orthogonal Resource Partitioning

Without loss of generality, assume that we have N orthogonal channels in the system. These N orthogonal channels could be partitioned using FDMA, TDMA, D-CDMA or hybrids of these schemes but we do not try to differentiate the detail partitioning here. In multi-cell scenarios, we have to assign the N orthogonal channels to the multi-cells. On one hand, we want to increase the *channel reuse efficiency* so that the overall system capacity could be multiplied based on the N channels. On the other hand, we have to make sure that *co-channel interference*[4] as a result of *channel reuse* is controlled at an acceptable level. There are many ways to do that. A simple and commonly used scheme is called *fixed channel assignment* (FCA).

With FCA, the N orthogonal channels are partitioned into N_c *clusters*. These N_c clusters are assigned to the multi-cells according to a reuse rule that the distance separation between the *co-cluster* cells are maximized. The corresponding system capacity is therefore given by:

$$N_{orthogonal} = (N/N_c) \times B \tag{3.25}$$

where B is the total number of cells in the system. For hexagonal cell shape, the condition on N_c for tessellating the cells is that $N_c = i^2 + ij + j^2$ for all non-negative integers i, j. Figure 3.14 illustrates various clusterization patterns.

In one extreme case, we could have $N_c = 1$ where this is equivalent to *complete reuse*. In this case, the system capacity from (3.25) will be maximized. However, the side effect is the co-channel interference. In this case, since co-channel cells are adjacent to each other, the interference level is very high and therefore, the quality of the communication link is very bad.

[4]Co-channel interference is the interference between multi-users as a result of re-using a channel in different cells.

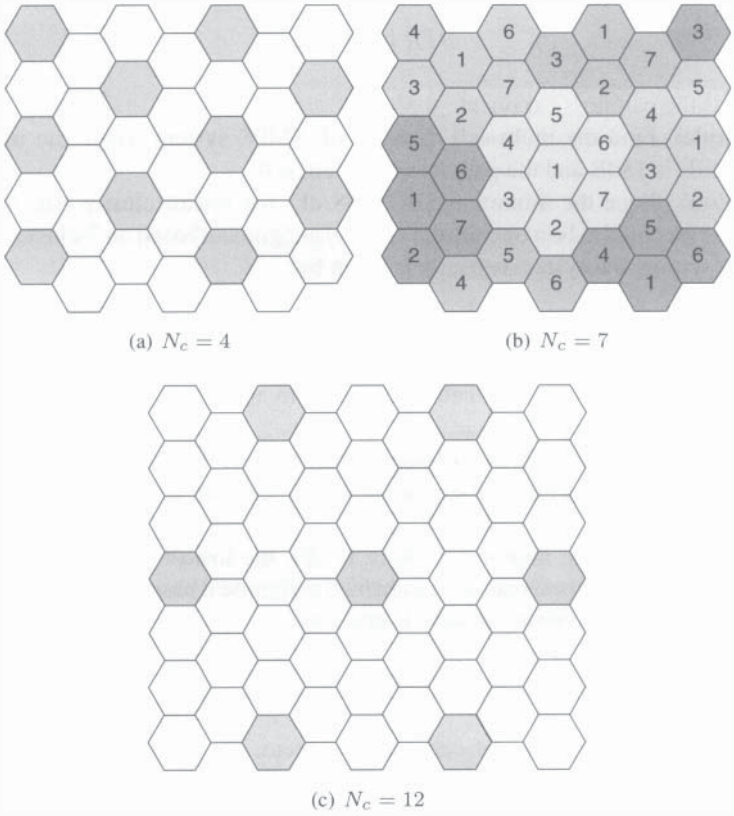


Figure 3.14 Illustration of several cluster reuse patterns in FCA.

In another extreme case, we could have a very large N_c and this is equivalent to *no reuse*. In this case, the system capacity is simply equal to N but the communication quality is very good because there is no co-channel users.

Obviously, these two extremes are not desirable operating point and the desirable operating point should be a minimum cluster size N_c where the co-channel interference is just below the maximal tolerable level. In fact, it can be easily verified that the signal to interference level (SIR) as a result of channel reuse is given by:

$$SIR \approx \sqrt{3N_c}^{\alpha}$$

where α is the path loss exponent.

Example: Find the multi-cell capacity of AMPS system where the minimum required SIR is 18dB and the path loss exponent is 4.

Solution: Since the minimum SIR is 18 dB, the lowest cluster size $N_c = 7$. Therefore, we could at best operate the channel assignment based on 7-cluster pattern. With this setting, the system capacity is given by:

$$N = (N_{AMPS}/7)B \tag{3.26}$$

where B is the total number of cells in the system and N_{AMPS} is the total number of channels in AMPS system.

Example: Find the multi-cell capacity of Digital-AMPS system where the minimum required SIR (as a result of digital transmission, error correction and sectorization) is 12dB.

Solution: Since the minimum SIR is 12 dB, the lowest cluster size $N_c = 4$. Therefore, we could at best operate the channel assignment based on 4-cluster pattern. With this setting, the system capacity is given by:

$$N = (N_{DAMPS}/4)B = (W/R_b)(B/4) \tag{3.27}$$

where B is the total number of cells in the system, N_{DAMPS} is the total number of channels in AMPS system, W is the bandwidth of the operator spectrum and R_b is the required bit rate per user. Note that due to digital speech compression, $N_{DAMPS} = 3N_{AMPS}$ for the same available spectrum. Hence, the capacity gain of 2G digital systems with respect to 1G analog systems are contributed by two factors, namely *higher number of channels as a result of speech compression* and *better reuse efficiency as a result of smaller cluster size*.

Since received signal strength is inversely proportional to d^{α} where $\alpha \approx 4$ in terrestrial propagation environment, the co-channel interference as a result of reuse is reduced by the path loss effect. In a way, the increased path loss exponent in terrestrial propagation helps the channel reuse in multi-cell because signals are more *contained* within their own cells.

Multi-cell Performance of R-CDMA

For R-CDMA, the extension to multicell scenario is relatively easy. This is because the intra-cell channels are non-orthogonal already and there is not much difference between intra-cell interference and inter-cell interference. To account for the additional intercell interference, we have

$$I_{oc} = \beta I_{sc} = \beta(K-1)P\mu$$

where I_{oc} denotes the other cell interference power, I_{sc} denotes the same cell interference power and μ denotes the voice activity factor of users. Typical values of β and μ are $[0.4, 0.7]$ (depending on the path loss exponent) and 0.5 respectively. Hence, the received bit-energy to noise plus interference ratio of user k γ_k is given by:

$$\gamma_k = \frac{P}{\eta_0 W + (1+\beta)\mu(K-1)P}\frac{T_b}{T_c} \geq \gamma_{\text{target}}$$

The pole capacity of a cell is given by $N_{R-CDMA,cell} \approx 1 + \frac{T_b/T_c}{(1+\beta)\gamma_{\text{target}}}$. Since all the adjacent cells use the same frequency and the system is not code limited, the total system capacity in multi-cell scenario is multiplied by B and is given by:

$$N_{R-CDMA,sys} = B\left(1 + \frac{W/R_b}{(1+\beta)\mu\gamma_{\text{target}}}\right) \tag{3.28}$$

where B is the total number of cells in the system.

Example: Find the multi-cell capacity of IS 95 system where the minimum required SIR (as a result of digital transmission, error correction and sectorization) is 7dB. The voice activity factor μ is 0.5 and the co-channel interference factor β is 0.4.

Solution: Since the minimum SIR is 7 dB (which is approximately 5 in linear scale), the system *pole capacity* is given by:

$$N_{IS95,sys} \approx B\left(1 + \frac{W/R_b}{3.5}\right)$$

Observe that R-CDMA multi-cell capacity is larger than the multi-cell capacity of the orthogonal resource partitioning schemes (FDMA/TDMA/D-CDMA) because of the complete channel reuse. Figure 3.15 illustrates the system capacity comparison of the orthogonal and non-orthogonal partitioning schemes. It is shown that the larger the number of cells in the system, the greater the capacity difference is.

3.6 PRACTICAL ILLUSTRATIONS: GSM, CDMA AND WIRELESS LAN

In this section, we explain how different practical systems apply the resource partitioning techniques to enable multi-user communications. We look at GSM (2G cellular system), CDMA (2G cellular system) and Wireless LAN (short-range wireless system) [129, 274] as examples.

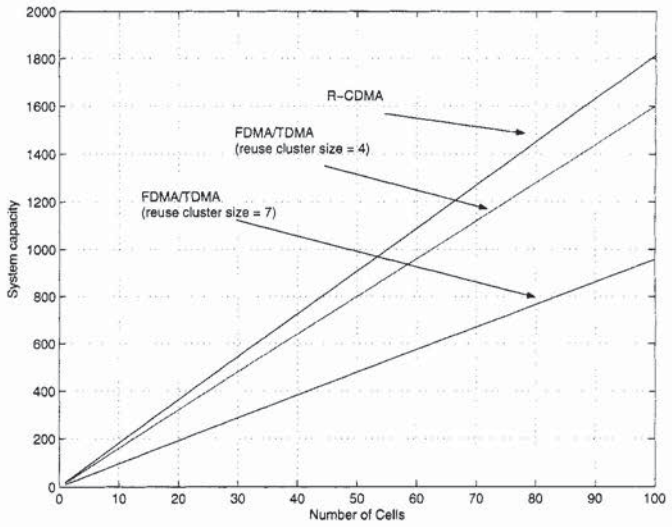


Figure 3.15 Comparison of multicell system capacity between the orthogonal schemes (FDMA,TDMA,D-TDMA) and the non-orthogonal schemes (R-CDMA). $W = 1.5$ MHz, $R_b = 25$kbps.

3.6.1 Multiuser resource partitioning in GSM

GSM is one of the major second generation cellular standards with over 70% worldwide penetration based on TDMA. We leave it to Part II for a detail elaboration of the GSM architecture and design but in this section, we focus on the multi-access aspects in GSM system[141]. In GSM, the spectrum bandwidth is partitioned into radio channels based on hybrid FDMA / TDMA approach. The radio spectrum is first partitioned into a number of *carriers* based on FDMA where each carrier has a bandwidth of 200kHz. Each 200kHz carrier is further partitioned into 8 *time slots* based on TDMA. Voice traffic and control traffic are delivered to and from the mobiles via the timeslots. For example, a GSM operator may be allocated a paired spectrum of 15MHz on the uplink and downlink respectively. The 15MHz spectrum is first partitioned into 75 *carriers* (each carrier has a bandwidth of 200kHz) based on FDMA. Each of the 200kHz carrier is partitioned into 8 timeslots based on TDMA [143]. Hence, the total number of *full duplex channels* available to the operator is given by $75 \times 8 = 600$. The choice of 200kHz carrier bandwidth is a compromise between two factors. For instance, since the end-user channel is partitioned based on TDMA, the user on the assigned timeslot will be transmitting over the entire carrier bandwidth. If the carrier bandwidth is too large (say 2MHz instead of 200kHz), the number of resolvable multipaths as seen by the GSM mobile will be quite high because the typical outdoor coherence bandwidth is of the order of 100kHz. Hence,

complex equalization is required at the GSM mobile which is highly undesirable. We elaborate more on the effect of multi-path fading in Chapter 4. On the other hand, if the bandwidth of the carrier is too small (say 50kHz instead of 200kHz), the number of carriers derived from the 15MHz spectrum will be 300 instead of 75. Hence, a base station (fully equipped with all the 600 channels) will require 300 pairs of RF units which is very bulky and expensive. Therefore, the choice of 200kHz carrier bandwidth is a compromise between the equalization complexity at the GSM mobile and the base station complexity.

In fact, the 600 channels available to the GSM operator in the example is a small number. Imagine the operator has to serve the entire city with only 600 channels from the 15MHz paired spectrum. This is highly insufficient and obviously, multi-cell spatial reuse has to be exploited in order to scale up the total number of channels in the GSM system as the number of base stations increases. Strictly speaking, since a user channel in GSM is labeled by both the timeslot number and the carrier number, two user channels in GSM systems are regarded as *co-channel* users if the two user channels have the same time slot and carrier number coordinate. This is because if either the time slot number or the carrier number of two user channels differs, the two user channels will not be interfering with each other. For example, two user channels with (time slot number, carrier number) coordinate = $(1, 3)$ and $(1, 4)$ are not co-channel users. Two user channels with coordinate = $(1, 3)$ and $(2, 3)$ are also not co-channel user. Hence, strictly speaking, the *user channels* on adjacent cells can be separated based on either the time slot dimension as well as the carrier dimension. However, in GSM systems (FDD), the base stations are only *coarsely synchronized* in time with each other. Furthermore, due to the uncertainty of propagation delay between mobiles in different cells, it is technically too complicated if we try to separate adjacent cell users based on time slots. In GSM systems, the adjacent cells are assigned with different carriers in a similar way as traditional frequency planning (hence, channels from adjacent cells are separated based on the carrier dimension only). The carrier is reused as long as co-channel cells are sufficiently separated. For example, if base station 1 and base station 2 are next to each other, they will be assigned two different carriers. Within each of the assigned carriers, the base station can utilize all the 8 time slots without worrying about mutual interference. Due to the digital transmission and the error correction in the speech frame, the required SIR in GSM system is around 13dB. Together with sectorization, the frequency reuse factor can be reduced from $K = 7$ to $K = 4$. This represents substantial system capacity gain compared to the 1st generation systems.

3.6.2 Multiuser resource partitioning in IS95

Similarly, we leave to Part II for a more detail elaboration of the system architecture and operations of IS95 (another example of 2G digital cellular systems based on CDMA technology). In this section, we just focus on the multi-user aspects of IS95 systems. In IS95 systems, user channels are partitioned from the allocated

spectrum based on hybrid FDMA/CDMA. The allocated spectrum is first partitioned into 1.5MHz *carrier*. Each of the 1.5MHz carrier is further partitioned into user channels based on CDMA. For example, suppose the operator has been allocated 15MHz of paired spectrum. We have $15M/1.5M = 10$ full duplex CDMA carriers after FDMA partitioning. Each of the 1.5MHz CDMA carrier is further partitioned into a number of code channels based on CDMA. Similar to the GSM system, the choice of 1.5MHz is a compromise between CDMA receiver complexity and the efficiency of CDMA. For instance, if the CDMA carrier bandwidth is too high, the chip rate will be high and this will induce higher complexity in the CDMA receiver processing (chip synchronization, searching and despreading). On the other hand, if the CDMA carrier bandwidth is too low, the spreading factor and the processing gain will be too low and this reduces the efficiency of CDMA to suppress various interference.

Unlike the GSM example where we can precisely specify the total number of full duplex channels from the 15MHz pair spectrum, we cannot precisely calculate the total number of channels here. This is because in IS95 downlink, D-CDMA (with spreading factor of 64) is used to partition the 1.5MHz carrier into orthogonal code channels. Hence, there are a total of $64 \times 10 = 640$ downlink channels. However, on the uplink, R-CDMA (with processing gain of 64) is used to partition the 1.5MHz carrier into *non-orthogonal* code channels. As we have introduced, the number of channels R-CDMA can support is not code limited but interference limited. Hence, we have *soft capacity* instead of the *hard capacity*. Nonetheless, in IS95, the R-CDMA capacity is around 14-16 concurrent users per 1.5MHz CDMA carrier.

Similarly, the total number of channels have to be able to scale in the multicell network as we increase the number of base stations. As we have illustrated, CDMA systems allowed complete frequency reuse due to the inherit interference suppression capability. Adjacent cells in IS95 networks are assigned with the same CDMA carriers and the co-channel interference will be suppressed by the inherit processing gain of 64. When the capacity of the 1.5MHz CDMA carrier has been used up in some hot spots, additional CDMA carrier can be added. In other words, each CDMA carrier can be viewed as a *layer* and multiple carriers are *overlaid* with each other in the IS95 multicell system.

3.6.3 Multiuser resource partitioning in Wireless LAN

In cellular systems such as GSM or IS95 systems, the channels are partitioned and coordinated at the base station. This refers to the centralized approach. On the other hand, wireless LAN (Wi-Fi) systems [129] are designed based on a *distributed* manner. There is no centralized controller (similar to the base station) in a wireless LAN network. The key design challenge in Wi-Fi systems is robust operations under distributed and uncoordinated situations. In Part III, we elaborate more on the architecture and the operation of Wi-Fi systems. In this section, we just focus on the multi-user resource aspects in Wi-Fi.

Wi-Fi system is designed to operate in ISM band (2.4GHz or 5GHz) which is an unlicensed spectrum. For instance, the ISM band in 2.4GHz is partitioned into 11 *channels*. However, each Wi-Fi transmission has a bandwidth of 20MHz which occupies 5 consecutive ISM channels. Hence, in the 11 ISM channels, there can only be three non-overlapping Wi-Fi channels, namely channel 1, channel 6 and channel 11. The key to Wi-Fi systems concerning multi-user communications is based on the *carrier sense multiple access with collision avoidance* (CSMA/CA). For instance, before a Wi-Fi user can transmit a packet, it has to listen to the channel and make sure there is no existing transmissions. In other words, in normal operations, there will be no concurrent transmissions among multiple Wi-Fi users operating in the same channel. Occasionally, when there is concurrent transmissions among Wi-Fi users, this results in *collision* and none of the packets will get through. Hence, in the steady state, the Wi-Fi users *coordinate* among themselves in a *distributed* manner to time share the channel dynamically. At any time instance, only one Wi-Fi user is allowed to use the channel.

Notice that Wi-Fi system such as 802.11b utilize spread spectrum to transmit the packet (using Barker code). The spreading factor of the Barker code is 11. Unlike IS95 systems, all the Wi-Fi users in the systems are assigned the same and unique Barker code. Hence, the Barker code is not used to separate other Wi-Fi users for simultaneous transmissions. The Barker code (spread spectrum) is used for interference suppression with respect to other ISM users.

On the multi-cell aspects, there is no need for coordination in Wi-Fi systems as well. For example, multiple access points in Wi-Fi systems can be configured to operate in the same frequency. The *co-channel interference* between adjacent access points is automatically taken care by the CSMA/CA protocol. If the two access points are within the *sensing range*, the Wi-Fi users associated with the two access points will time-share the channel dynamically using the CSMA/CA. On the other hand, if the two access points are separated sufficiently far apart that they are outside the *sensing range*, the Wi-Fi users associated with the two access points will simultaneously use the same channel without affecting each other. Hence, in Wi-Fi systems, the design of sensing range is critical to the scalability of the network capacity. For instance, if the sensing range is too large, the capacity scalability will be very poor and the total throughput supported by the entire network will not scale effectively as we increase the number of access points. If the sensing range is too small, there will be increased *hidden node* populations and this will dramatically reduce the network throughput due to collisions.

3.7 SUMMARY

In this chapter, we address the problem of multiuser communications using both the *information theoretical* approach and *practical design* approach. In the information theoretical context, we state the multiuser communication problem as the *multiaccess*

problem (uplink) and *broadcast problem* (downlink) and offer a general solution for AWGN channel. In practical design approach, we have reviewed and compared the performance of various *orthogonal resource partitioning* schemes (such as FDMA, TDMA, D-CDMA) as well as *non-orthogonal resource partitioning* schemes (such as R-CDMA and SDMA). While orthogonal schemes offer better performance than non-orthogonal scheme (with isolated detection) in single cell scenario, the non-orthogonal schemes give better performance in multi-cell scenario due to the ability to have complete resource reuse. Practical illustrations on the use multiuser communication technologies in GSM, CDMA, and wireless LAN are given in the final section.

PROBLEMS

3.1 Suppose a digital voice user has a source bit rate of 8kbps (full duplex). Assuming QPSK is used as the common modulation scheme and the baud rate is equal to the required channel bandwidth. Consider a single cell system with 1.6 MHz bandwidth allocated for uplink and 1.6MHz for downlink respectively. Calculate the capacity (in turns of number of concurrent active voice users) achievable by the FDMA. [Assume that coherent phase reference is available at the receiver].

3.2 Using the same setting as in Question 1, what about the capacity of the single-cell TDMA system?

3.3 What about the forward link capacity of the single-cell CDMA system Hint: The number of distinct orthogonal spreading codes available is equal to the number of chips per modulation symbol? Which technology is more effective in terms of forward link capacity in a single cell system? Will there be any difference in multiple cell system?

CHAPTER 4

DIVERSITY TECHNIQUES

Diversity—The division of classes among a certain population.
—Wikipedia, the free encyclopedia

4.1 INTRODUCTION

Communication system consists of a transmitter and a receiver connected by a *channel*. In Chapter 1, we have elaborated the common channel models to capture the major propagation characteristics of mobile radio channels. In Chapter 2, we have discussed the operation and tradeoff picture of digital modulator and demodulator. However, we have used AWGN as a simple example when we discuss the performance of digital communication systems. In this chapter, we shall further elaborate and focus on the detrimental effects (and the possible solutions) of channel fading on digital communication systems.

We shall first start with flat fading channels and frequency selective fading channels and discuss their effects on the BER performance of the physical layer. Next, we shall

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

discuss the corresponding solutions to the detrimental effects, namely the *diversity techniques* and the *equalization effects*. We shall elaborate various means of obtaining diversity, namely the *time diversity*, *frequency diversity* and *spatial diversity* [289]. Following these, we shall illustrate the practical application of diversity techniques, namely the RAKE receiver [67], in CDMA systems (such as IS95, UMTS, 3G1X). Finally, we shall conclude with a brief summary of main points.

4.2 EFFECTS OF FLAT FADING ON BER PERFORMANCE

Recall in Chapter 1 that when the transmit signal bandwidth is less than the coherence bandwidth (narrowband transmission), the transmit signal experiences flat fading channels. The low-pass equivalent received signal through slow flat fading channels is given by:

$$y(t) = \alpha(t)x(t-\tau) + z(t)$$

where $x(t)$ is the transmitted signal, $\alpha(t)$ is the complex fading coefficient, τ is the propagation delay and $z(t)$ is the complex Gaussian channel noise with zero mean and power spectral density N_0. In the AWGN channel, the transmitted signal is affected by the channel noise only. However, in flat fading channels, the transmitted signal is *distorted* by the fading coefficient $\alpha(t)$ as well as the channel noise. Both the channel noise and the fading coefficient $\alpha(t)$ are random processes. Since we consider slow fading channels, $\alpha(t) \approx \alpha$ within a modulation symbol duration. The amplitude $|\alpha|$ represents the random attenuation introduced by flat fading. The phase $\angle\alpha$ represents the random phase shift introduced by flat fading. Since the complex fading coefficient α is Gaussian distributed with zero-mean and unit variance $\mathcal{E}[|\alpha|^2] = 1$, the amplitude $|\alpha|$ is Rayleigh distributed and the phase $\angle\alpha$ is uniformly distributed over $[0, 2\pi]$.

For simplicity, consider BPSK modulation where the signal set is given by $\mathcal{S} = \{s_1(t), s_2(t)\}$. Conditioned on α, the flat fading channel is the same as AWGN channel. As illustrated in Chapter 2, the conditional error probability of BPSK modulation in flat fading channels (conditioned on α) is determined by the Euclidean distance between the 2 constellation points $\|\alpha\vec{s}_1 - \alpha\vec{s}_2\|^2$ and is given by:

$$P_e(\alpha) = Q\sqrt{\left(\frac{\|\alpha\vec{s}_1 - \alpha\vec{s}_2\|^2}{2N_0}\right)} = Q\sqrt{\left(\frac{2|\alpha|^2 E_s}{N_0}\right)} \tag{4.1}$$

where $E_s = \int_0^{T_s} |s_i(t)|^2 dt$.

Hence, the average error probability of the BPSK modulation (averaged over the random fading coefficient α) is given by [287]:

$$P_e = \mathcal{E}[P_e(\alpha)] = Q\sqrt{\left(\frac{2\gamma E_s}{N_0}\right)} f_\gamma(\gamma)d\gamma \approx \frac{1}{4E_s/N_0} \tag{4.2}$$

where $\gamma = |\alpha|^2$, $f_\gamma(\gamma) = \exp(-\gamma)$ is the pdf of γ and the approximation holds for large SNR E_s/N_0. Consider plotting the BER (in log-scale) versus SNR (in dB), we have

$$10\log_{10} P_e \approx -10\log_{10}(E_s/N_0) + K.$$

Hence, the BER curve (in log scale) of BPSK modulation in flat fading channels has a slope of -1 at large SNR.

Figure 4.1 illustrates the effect of flat fading on the BER vs SNR curve of BPSK modulation. The error probability drops rapidly with respect to SNR for AWGN channel. However, for flat fading channels, the slope of the BER curve *flattened*. Hence, at a given target BER (say 10^{-3}), the required SNR is 40 dB which is 30 dB larger than that required for AWGN channel[287]. As illustrated in the figure, due to the flat fading coefficient α, the required SNR has to be increased substantially to maintain a certain target error probability. For point-to-point communications, the penalty of flat fading is a reduced communication range (for the same transmit power) or extra transmit power (for the same communication range). For multi-user communications (such as R-CDMA), the penalty of the flat fading is translated into a reduction of system capacity.

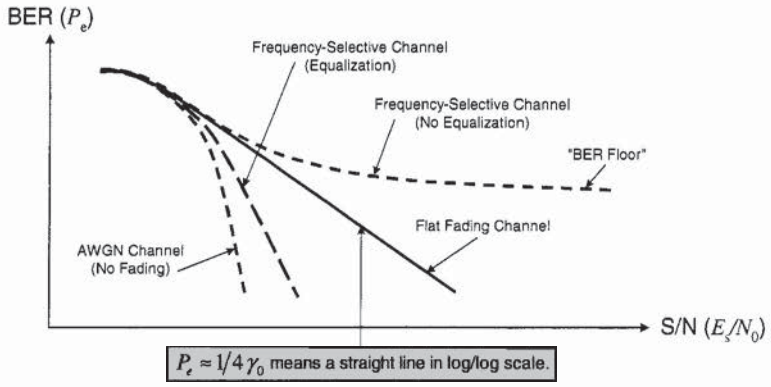


Figure 4.1 BER curves of BPSK in AWGN and flat fading channels.

4.3 EFFECTS OF FREQUENCY SELECTIVE FADING ON BER PERFORMANCE

When the transmit signal bandwidth is larger than the channel coherence bandwidth, the multipaths become *resolvable* and the transmit signal experiences frequency selective fading channels. The lowpass equivalent received signal $y(t)$, after passing through frequency selective fading channels, is given by:

$$y(t) = \sum_{l=0}^{L_p-1} \alpha_l(t)x(t - lT_s) + z(t)$$

where $\alpha_l(t)$ is the complex random channel fading coefficient of the $l-$th resolvable path, T_s is the symbol duration, $z(t)$ is the complex Gaussian channel noise with zero mean and power spectral density N_0, and $L_p = \lceil W_{tx}/B_c \rceil$. Due to the resolvable multipaths, the *influence* of a transmitted symbol at the n-th symbol duration spans over multiple symbol durations $(n - L_p + 1, n - L_p + 2, \ldots, n)$. In other words, the received symbols are mixed up with each other and this results in *intersymbol interference* (ISI). This is illustrated in Figure 4.2 The larger the number of resolvable multipaths (L_p), the more severe the ISI will be.

The effects of frequency selective fading channels is therefore more severe than the flattening of the slope of the BER curves. In addition, due to the ISI, the BER curves will exhibit *error floor* as illustrated in Figure 4.2 This can be explained by the following. The BER can be seen as a function of SINR (Signal to Noise + Interference Power).

$$\gamma = \frac{S}{N + I}$$

where S is the received signal power, N is the noise power, and I is the ISI power. At the $n-$th symbol duration, the ISI is contributed by all the past echos of the past symbols. Hence, the ISI power (conditioned on the fading coefficients) is given by:

$$I = S \sum_{l=1}^{L_p-1} |\alpha_l|^2$$

When the SNR (S/N) is small, $N > I$ and therefore, the BER decreases with increasing SNR. However, at large SNR, we have $I >> S$ and hence the SINR γ is given by:

$$\gamma \to S/I = \frac{1}{\sum_{l=1}^{L_p-1} |\alpha_l|^2} = \gamma_0$$

Hence, γ approaches a constant γ_0 and therefore, the BER does not decrease further with increasing SNR. Therefore, irreducible error floor[219] is resulted. Note that the error floor level depends on the residual SIR γ_0. Hence, when L_p is large, the error floor will be high.

4.4 DIVERSITY: A KEY TECHNIQUE TO COMBAT FLAT FADING CHANNELS

In this section, we shall elaborate a common and effective technique to combat the *flattening* of the slope of the BER curve as a result of flat fading. In Chapter 2, we

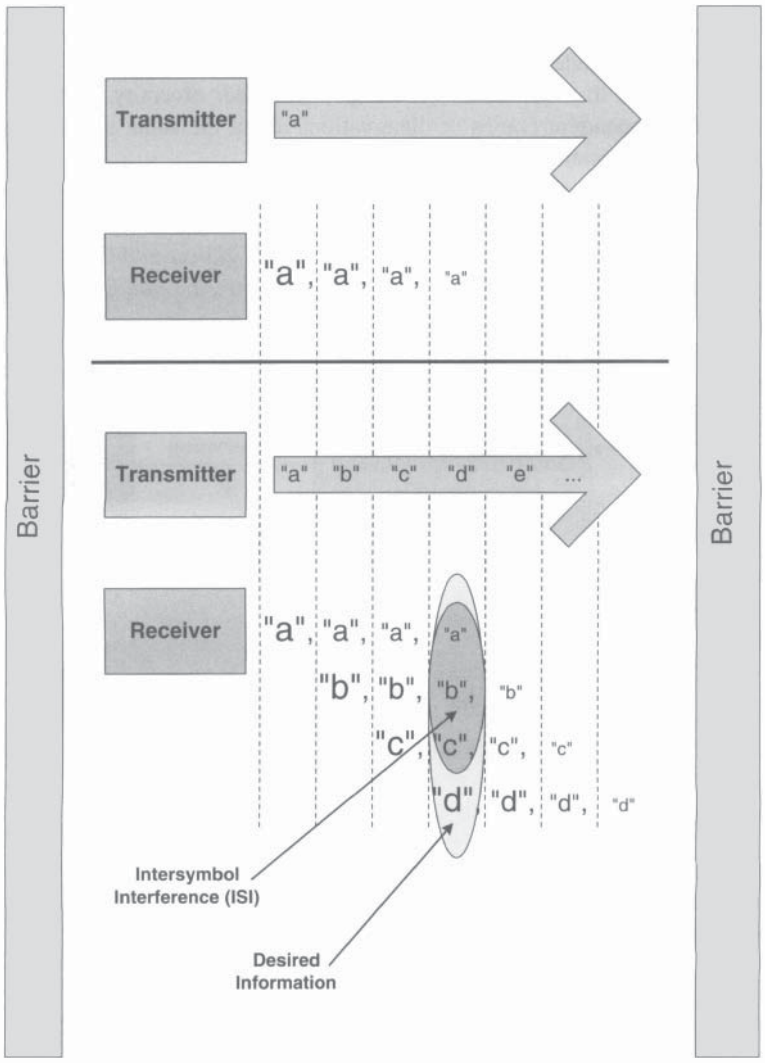


Figure 4.2 Illustration of intersymbol interference (ISI). The transmitter (transmitting letter "a") results in multiple echos at the receiver. Hence, when the transmitter transmits a sequence of letters "a,b,c,...", ISI will be resulted at the receiver where the desired information (letter 'd') is interfered by the past echos of letters "c", "b", "a".

have elaborated on the minimum distance detection design. For a given received symbol, the "best" guess on what has been transmitted is the one that is *closet* to the *observation*. Normally, the detection is made based on a single observation. However, if L *independent observations* about the same information are available at the receiver, the detection can be made based on all the L observations. In that case, the receiver is said to exploit L-th order diversity. With L-th order diversity, the probability that all the L independent copies of observations about the same information fade simultaneously is dramatically reduced.

There are in general three common ways of making use of L independent observations about the same information at the receiver. They are referred to *diversity combining techniques*, namely *selection diversity*, *equal gain combining* and *maximal ratio combining*. These methods, as illustrated in Figure 4.3, are elaborated in the following subsections.

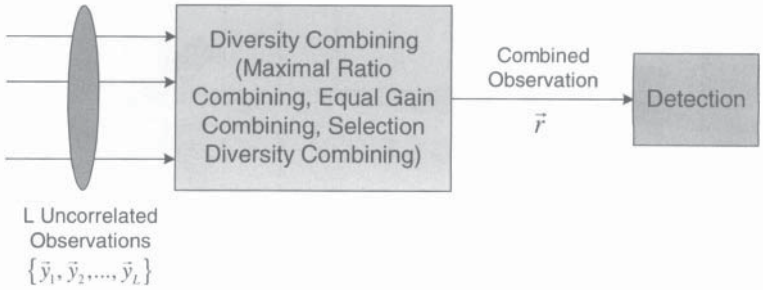


Figure 4.3 Diversity Combining.

4.4.1 Diversity Combining

For simplicity, consider a binary modulator with signal set $\mathcal{S} = \{\vec{s}_1, \vec{s}_2\}$ transmit over flat fading channels with L independent observations. The received vectors from the l-th observation is given by:

$$\vec{y}_l = \alpha_l \vec{x} + \vec{z}_l$$

for $l = [1, \ldots, L]$ where $\vec{z}_l$ is the complex Gaussian noise vector with zero mean and covariance matrix $\mathcal{E}[\vec{z}_l \vec{z}_l^*] = N_0 \mathbf{I}$. We shall consider three possible diversity combining rules below.

Maximal Ratio Combining: Since $\{\vec{y}_1, \ldots, \vec{y}_L\}$ contains the same information $\vec{x}$, the optimal detection should be based on the entire observations $\{\vec{y}_1, \ldots, \vec{y}_L\}$. The ML detection can be formulated as:

$$\vec{s} = \arg \max_{\vec{x} \in \mathcal{S}} p(\vec{y}_1, \ldots, \vec{y}_L | \vec{x}, \alpha_1, \ldots, \alpha_L) \tag{4.3}$$

Since the L observations are independent, $\alpha_1, \ldots, \alpha_L$ and $\vec{z}_1, \ldots, \vec{z}_L$ are independent with each other. Hence, the joint density $p(\vec{y}_1, \ldots, \vec{y}_L | \vec{x}, \alpha_1, \ldots, \alpha_L)$ can be expressed as:

$$\begin{aligned} p(\vec{y}_1, \ldots, \vec{y}_L | \vec{x}, \alpha_1, \ldots, \alpha_L) &= \Pi_{l=1}^{L} p(\vec{y}_l | \vec{x}, \alpha_l) \\ &= K \exp\{-\left(\frac{\sum_{l=1}^{L} \|\vec{y}_l - \alpha_l \vec{x}\|^2}{2\sigma_z^2}\right)\} \end{aligned} \quad (4.4)$$

for some constant K. Hence, the ML detection rule in (4.3) can be simplified as:

$$\begin{aligned} \vec{s} &= \arg\min_{\vec{x} \in \mathcal{S}} \sum_{l=1}^{L} \|\vec{y}_l - \alpha_l \vec{x}\|^2 \\ &= \arg\min_{\vec{x} \in \mathcal{S}} - \sum_{l} < \vec{y}_l, \alpha_l \vec{x} > - \sum_{l} < \alpha_l \vec{x}, \vec{y}_l > + \sum_{l} |\alpha_l|^2 \|\vec{x}\|^2 \\ &= \arg\min_{\vec{x} \in \mathcal{S}} - < \sum_{l} \alpha_l^* \vec{y}_l, \vec{x} > - < \vec{x}, \sum_{l} \alpha_l^* \vec{y}_l > + \sum_{l} |\alpha_l|^2 \|\vec{x}\|^2 \\ &= \arg\min_{\vec{x} \in \mathcal{S}} - < \vec{r}_{MRC}, \vec{x} > - < \vec{x}, \vec{r}_{MRC} > + \sum_{l} |\alpha_l|^2 \|\vec{x}\|^2 \end{aligned} \quad (4.5)$$

where

$$\vec{r}_{MRC} = \sum_{l=1}^{L} \alpha_l^* \vec{y}_l \quad (4.6)$$

Since the ML detection metric in 4.5 is a function of the *combined observation* $\vec{r}_{MRC}$, the observation $\vec{r}_{MRC}$ is called the *sufficient statistics* with respect to the information $\vec{x}$. In other words, there is no loss of information to build a detector based on the combined observation $\vec{r}_{MRC}$ relative to a detector based on the L independent observations $\{\vec{y}_1, \ldots, \vec{y}_L\}$. The combined observation $\vec{r}_{MRC}$ in (4.6) is called the *maximal ratio combined* observation as illustrated in Figure 4.4 and it is the *optimal* way of combining the L independent observations. However, to achieve maximal ratio combining, the receiver has to estimate the fading coefficients $\{\alpha_1, \ldots, \alpha_L\}$.

Next, we shall look at the BER performance of BPSK modulation with L-order diversity and maximal ratio combining. The observation obtained by maximal ratio combining $\vec{r}_{MRC}$ is given by:

$$\vec{r}_{MRC} = \left(\sum_{l=1}^{L} |\alpha_l|^2\right) \vec{x} + \sum_{l=1}^{L} \alpha_l \vec{z}_l$$

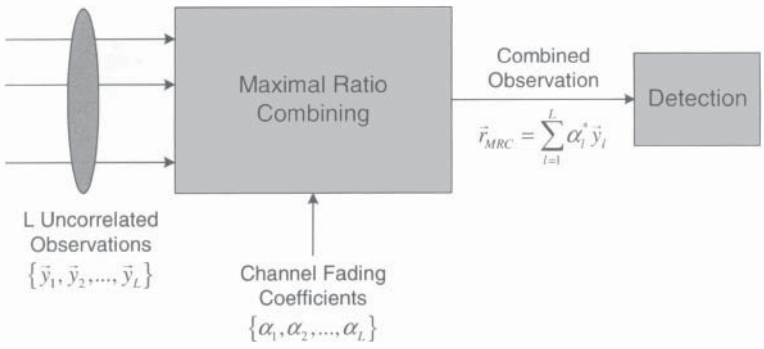


Figure 4.4 Illustration of Maximal Ratio Combining (MRC).

Conditioned on the fading coefficients $\{\alpha_1, \ldots, \alpha_L\}$, the BER is given by:

$$P_e(\alpha_1, \ldots, \alpha_L) = Q\left(\sqrt{\frac{2\gamma E_s}{N_0}}\right)$$

where $\gamma = \sum_{l=1}^{L} |\alpha_l|^2$. Comparing with the flat fading performance in (4.2), the maximal ratio combining alters the distribution of the term γ from *negative exponential distribution* to *central χ-square distribution* with $2L$ degrees of freedom. The average BER expression (average over the fading coefficients $\alpha_1, \ldots, \alpha_L$) is given by:

$$P_e = \mathcal{E}[P_e(\alpha_1, \ldots, \alpha_L)] \approx \binom{2L-1}{L} \frac{1}{(4E_s/N_0)^L}$$

for large SNR. Hence, the slope of the BER versus SNR curve (in log scale) is given by $-L$. The L-th order diversity increases the slope of the BER curve from -1 to $-L$ (steepens the BER curves).

Equal Gain Combining: For equal gain combining, the combined observation $\vec{r}_{EGC}$ is given by:

$$\vec{r}_{EGC} = \sum_{l=1}^{L} \exp(-\angle\alpha_l)\vec{y}_l = \vec{x}\left(\sum_{l=1}^{L} |\alpha_l|\right) + \sum_{l=1}^{L} \exp(-\angle\alpha_l)\vec{z}_l$$

This is illustrated in Figure 4.5

Hence, conditioned on the fading coefficients $\alpha_1, \ldots, \alpha_L$, the conditional BER is given by:

$$P_e(\alpha_1, \ldots, \alpha_L) = Q\left(\sqrt{\frac{2\gamma E_s}{2N_0}}\right)$$

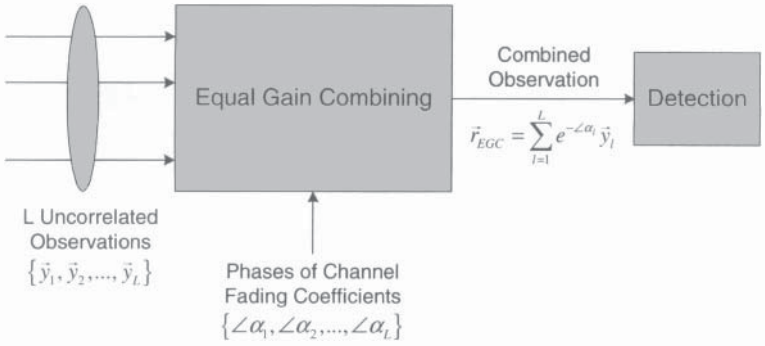


Figure 4.5 Illustration of Equal Gain Combining (EGC).

where $\gamma = (\sum_l |\alpha_l|)^2$. Hence, the average BER (average over fading coefficients $\alpha_1, \ldots, \alpha_L$) is given by:

$$P_e(EGC) = \mathcal{E}[P_e(\alpha_1, \ldots, \alpha_L)] \approx \frac{K_{EGC}}{(4E_s/N_0)^L}$$

for some constant K_{EGC} and large SNR. Hence, the slope of the average BER versus SNR curve is $-L$ with a shift of $10\log_{10}(K_{EGC}/4^L)$ dB. For equal gain combining, we need to know the phase of α_l but not the amplitude of α_l.

Selection Diversity Combining: Consider a third diversity combining technique, namely the selection diversity combining. Specifically, the combined observation $\vec{r}_{SDC}$ is given by:

$$\vec{r}_{SDC} = \vec{y}_{l^*} \tag{4.7}$$

where

$$l^* = \arg\max_l \{|\alpha_1|, \ldots, |\alpha_L|\}$$

This is illustrated in Figure 4.6

In other words, the combined observation is obtained by selecting the *best* observation from the L independent observations $\{\vec{y}_1, \ldots, \vec{y}_L\}$. Similarly, the conditioned BER is given by:

$$P_e(\alpha_1, \ldots, \alpha_L) = Q\left(\sqrt{\frac{2\gamma E_s}{2N_0}}\right)$$

where $\gamma = \max_l \beta_l$ and $\beta_l = |\alpha_l|^2$. The cdf of γ is given by $F_\gamma(\gamma) = [F_\beta(\gamma)]^L$. Hence, the pdf of γ is given by:

$$f_\gamma(\gamma) = \frac{dF_\gamma}{d\gamma} = L(1 - \exp(-\gamma))^{L-1} \exp(-\gamma)$$

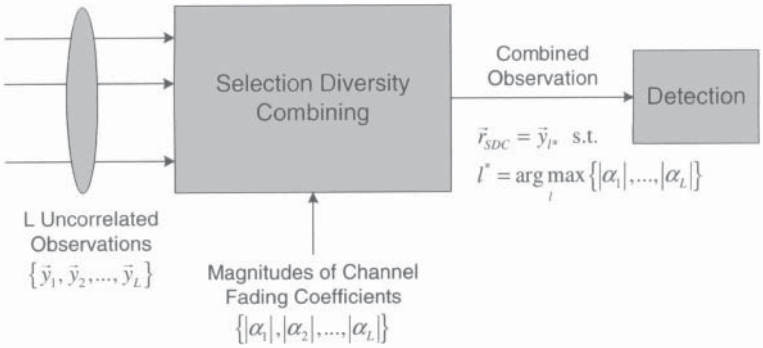


Figure 4.6 Illustration of Selection Diversity Combining (SDC).

The average BER (average over fading coefficients $\alpha_1, \ldots, \alpha_L$) is given by:

$$P_e(SDC) = \mathcal{E}[P_e(\alpha_1, \ldots, \alpha_L)] \approx \frac{K_{SDC}}{(4E_s/N_0)^L}$$

for some constant K_{SDC} and large SNR. Hence, with L-th order selection diversity, the slope of the average BER versus SNR curve is $-L$. For selection diversity combining, we do not need to know the phase of α_l and hence, this method is suitable for receivers without accurate channel estimation or equivalently, it is robust to channel estimation errors in $\alpha_1, \ldots, \alpha_L$.

In all the above combining methods with respect to the L-th order diversity, the slope of the average BER versus SNR curves are steepen from -1 to $-L$. The difference between the three combining schemes in BER performance is just the horizontal shifts as illustrated in Figure 4.7 As illustrated, the slope of the BER curves approach that of the AWGN curve as $L \to \infty$. The maximal ratio combining achieves the best BER performance, followed by the equal gain combining and the selection diversity.

4.4.2 Realization of Diversity

The key of diversity is to obtain L independent received observations *about the same information*. In general, there are three possible ways of realizing L independent received observations, namely the *time diversity* [351], *frequency diversity* [196] and the *spatial diversity* [87]. They are elaborated briefly below.

For time diversity, the L independent observations are obtained from multiple transmissions on the time dimension (separated sufficiently far in time). For example, if the same information symbol is transmitted twice in time domain, we obtain two observations (in time domain). As we have illustrated in Chapter 1, the channel fading experienced by the two observations separated in time by Δt is uncorrelated if $|\Delta t| > T_c$ where T_c is the coherence time of the fading channel. Hence, to realize

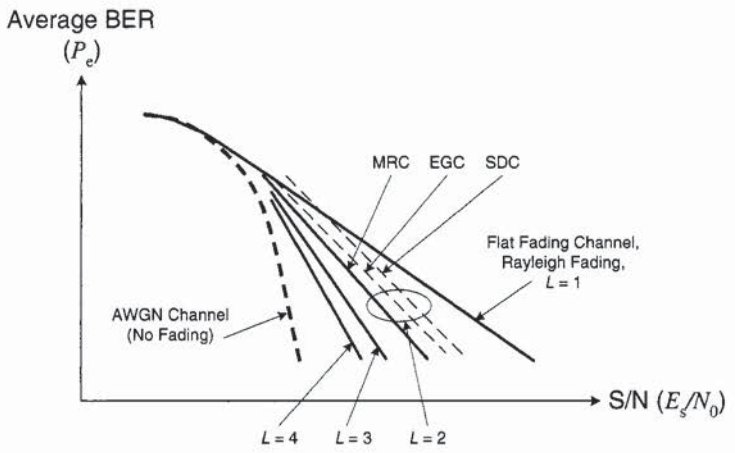


Figure 4.7 Diversity Combining.

L-th order diversity based on the time-domain, we must separate the L copies of observations by more than LT_c. In practice, this is realized by error correction coding over time domain and fast fading (channel fading coefficients are independent between modulation symbols). For slow fading channels where the channel fading coefficient remains quasi-static for an encoding frame, time diversity cannot be realized.

For frequency diversity, the L independent observations are obtained by transmission of the same information symbol at different frequencies. To achieve independent observations, the range of frequencies must be separated by more than LB_c where B_c is the coherence bandwidth of the fading channel. Hence, frequency diversity can be realized for wideband transmission over frequency selective fading channels. For narrowband transmission, the channel fading is quasi-static within the transmission bandwidth and therefore, no frequency diversity can be realized.

Finally, for spatial diversity, the L independent observations are obtained by either transmission of the same information symbol using different antennas at the transmitter or by receiving L observations using different antennas at the receiver. The former case is referred to the *transmit diversity* and the latter case is referred to the *receive diversity*. For independent observations, the antennas must be separated sufficiently far apart. For instance, the separation between any two antennas must be greater than D_c (coherence distance). Since at the base station side, the antennas are elevated at high altitude and hence, the angle spread as seen by the transmitter antennas is smaller compared with that as seen by the mobile antennas. Therefore, the coherence distance D_c at the base station side is larger than that at the mobile side. In general, D_c at the base station is of the order of 10λ whereas D_c at the mobile station is of the order of $\lambda/2$.

4.5 EQUALIZATION*

In Section 4.3, we have illustrated the effects of frequency selective fading channels (multiple resolvable multipaths) on the BER of the physical layer. Due to the inter-symbol interference (ISI), there will be two penalties on the BER vs SNR curve, namely the *flattening of BER curve* and the *irreducible error floor*. While the solution with respect to the former degradation is diversity, the fundamental solution[1] to the latter degradation is *equalization* [213, 287]. A channel equalizer [109] is a signal processing unit at the receiver side which is designed to estimate and cancel the ISI on the received signal. We shall briefly outline the principle of equalization below.

4.5.1 Maximal Likelihood Sequence Estimator (MLSE)

For simplicity, consider a digital modulator where the signal set $\mathcal{S}$ can be characterized by a two dimensional signal constellation such as QPSK, 16QAM, etc. The received signal after passing through a frequency selective fading channel with L_p resolvable multipaths is given by:

$$y_n = \sum_{l=0}^{L_p-1} h_l x_{n-l} + z_n$$

where $x_n \in \mathcal{S}$ is the transmitted modulation symbol, h_l is the complex Gaussian random variable (characterizing the l-th resolvable multipath fading coefficient) and z_n is the complex Gaussian additive channel noise. The optimal equalization process can be formulated as a *maximal likelihood sequence estimator* (MLSE). Consider a span of $N+1$ observations $\mathbf{y}_0^N = \{y_0, \ldots, y_N\}$. The MLSE equalization is to deduce the most likely transmitted information sequence $\mathbf{x}_0^N = \{x_0, \ldots, x_N\}$ based on the observations. This is formulated mathematically as:

$$\begin{aligned} \hat{\mathbf{x}_0^N} &= \arg\max_{\mathbf{x}_0^N} \log p(\mathbf{y}_0^N | \mathbf{x}_0^N, \mathbf{h}_0^{L_p-1}) \\ &= \arg\max_{\mathbf{x}_0^N} \sum_{n=0}^{N} \log p(y_n | \mathbf{x}_{n-L_p+1}^n, \mathbf{h}_0^{L_p-1}) \end{aligned} \tag{4.8}$$

where the second equality is due to the fact that the current observation y_n depends on the past L_p transmitted modulation symbols only. Define the system state $S_n =$

[1] Strictly speaking, diversity with maximal ratio combining may alleviate the error floor as well. This is because with MRC, the overall SINR is enhanced by the independent observations as well. Hence, the error floor will be lowered. On the other hand, if the number of resolvable multipath L_p is large, the effect of selection diversity on the error floor is insignificant because the overall average SINR is virtually not affected.

$(x_{n-L_p+1}, \ldots, x_{n-1})$ and the optimal log-likelihood function as

$$F^*(m, 0, S_m) = \max_{\mathbf{x}_0^m} \sum_{n=0}^{m} \log p(y_n | S_n, x_n, \mathbf{h}_0^{L_p-1}) \tag{4.9}$$

The MLSE optimization problem can be expressed into the following recursive form:

$$\begin{aligned} F^*(m, 0, S_m) &= \max_{x_m} \Big(\log p(y_m | S_m, x_m, \mathbf{h}_0^{L_p-1}) \\ &\quad + \max \mathbf{x}_0^{m-1} \sum_{n=0}^{m-1} \log p(y_n | S_n, x_n, \mathbf{h}_0^{L_p-1}) \Big) \\ &= \max_{x_m} \Big(\log p(y_m | S_m, x_m, \mathbf{h}_0^{L_p-1}) \\ &\quad + F^*(m-1, 0, (S_m / x_m)) \Big) \end{aligned} \tag{4.10}$$

for $m = 0, 1, \ldots$ and $F^*(-1, 0, S) = 0$. Hence, the solution of the MLSE optimization problem can be divided and conquered based on *dynamic programming* approach. More precisely, the MLSE algorithm is identical to the conventional *Viterbi decoding algorithm* for convolutional code with a branch metric given by $\log p(y_m | S_m, x_m, \mathbf{h}_0^{L_p-1})$.

Note that the MLSE equalizer requires the knowledge of the frequency selective fading coefficients $\{h_0, \ldots, h_{L_p-1}\}$. This can be estimated from the pilot channels or pilot symbols. Furthermore, since there are M^{L_p} state sequence S_n in the trellis, the complexity of the MLSE equalizer is of exponential order in L_p. The performance analysis of the MLSE equalizer is quite involved and we shall just quote the results here. The conditional symbol error probability of $M-$ary modulator (conditioned on the channel fading coefficients $\{h_0, \ldots, h_{L_p-1}\}$) is given by:

$$P_e(h_0, \ldots, h_{L_p-1}) \leq KQ\left(\sqrt{\frac{6}{M^2-1} E_s/N_0 \delta_{min}^2(\mathbf{h}_0^{L_p-1})} \right)$$

for some constant K where $\delta_{min}^2(\mathbf{h}_0^{L_p-1})$ is given by:

$$\delta_{min}(\mathbf{h}_0^{L_p-1})^2 = \sum_i \sum_j \alpha_{ij} h_i h_j^*$$

for some constants α_{ij}. The term δ_{min}^2 represents the SNR loss due to the ISI. The average error probability is given by $\mathcal{E}[P_e(h_0, \ldots, h_{L_p-1})]$ where the expectation is taken over all possible realizations of $h_0, \ldots, h_{L_p-1}$. We observe that after the optimal equalization, there is no error floor anymore. In addition, the slope of the BER curve is steepened due to the inherit frequency diversity due to frequency selective fading coefficients $h_0, \ldots, h_{L_p-1}$. This is illustrated in Figure 4.8 and the steepening of the BER curve can be understood by noting that δ_{min}^2 consists of the terms $\sum_i \alpha_{ii} |h_i|^2$ which is equivalent to the MRC error probability.

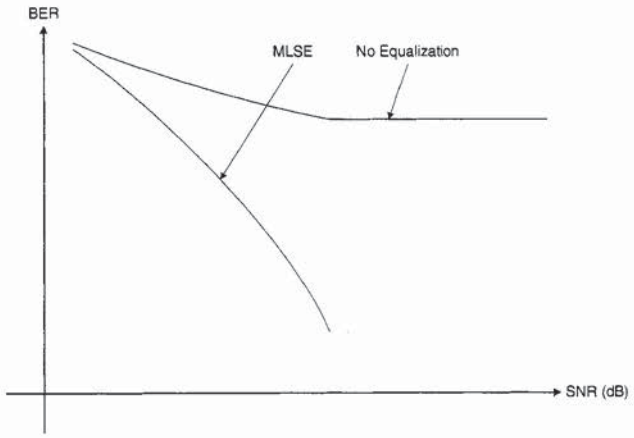


Figure 4.8 BER performance of MLSE equalizer in frequency selective fading channels.

4.5.2 Linear Equalization

While the MLSE equalization achieves the optimal performance, the complexity is of exponential order in L_p and hence, would be too complicated for moderate to large L_p. In this section, we consider a sub-optimal equalization algorithm[132], namely the minimum mean-square error (MMSE) equalization[210], that has linear order of complexity.

Figure 4.9 illustrates a generic block diagram of a linear equalizer for frequency selective fading channels. The received observations $\{y_n, y_{n-1}, \ldots, y_{n-2L}\}$ are processed by a *linear filter* with coefficients $\{w_0, \ldots, w_{2L}\}$ (where $2L$ is the length of impulse response of the linear filter) to produce a processed output r_n for detection. In other words, the processed output is given by:

$$r_n = \mathbf{w}^* \mathbf{y} \tag{4.11}$$

where $\mathbf{w} = [w_0, \ldots, w_{2L}]^*$, $\mathbf{y}_n = [y_n, \ldots, y_{n-2L}]^*$ and $(.)^*$ denotes conjugate transpose. The detection (demodulation) is performed on the processed output r_n to obtain $\hat{x}_n$.

Define the mean square error as:

$$\mathcal{E}[|e_n|^2] = \mathcal{E}[|r_n - x_n|^2]$$

The MMSE weight $\mathbf{w}$ is given by:

$$\mathbf{w}_{opt} = \arg\min_{\mathbf{w}} \mathcal{E}[|r_n - x_n|^2] \tag{4.12}$$

Substituting (4.11) into (4.12), the MMSE weight is given by:

$$\mathbf{w}_{opt} = \left(\mathcal{E}[\mathbf{y}_n \mathbf{y}_n^*]\right)^{-1} \mathcal{E}[\mathbf{y}_n x_n]$$

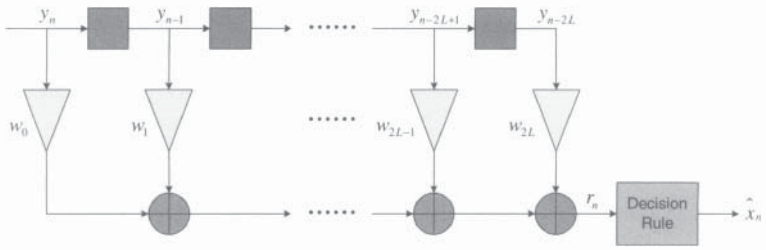


Figure 4.9 Block diagram of linear equalizer.

where

$$(\mathcal{E}[\mathbf{y}_n \mathbf{y}_n^*])_{ij} = \begin{array}{ll} x_{i-j} + N_0\delta_{ij} & |i-j| \le L_p \\ 0 & \text{otherwise} \end{array}$$

and

$$(\mathcal{E}[\mathbf{y}_n x_n])_i = \sigma_x^2 h_i^* \quad \forall i \in [0, L_p]$$

where $\sigma_x^2 = \mathcal{E}[|x_n|^2]$.

While the MMSE equalizer has linear processing complexity with respect to L_p, there is residual ISI at the filtered outputs r_n. Hence, there will be residual error floor in the BER vs. SNR curve.

4.6 PRACTICAL ILLUSTRATION: RAKE RECEIVER

In this section, we shall illustrate how the RAKE receiver in spread spectrum systems (such as CDMA IS95 and 3G) combat frequency selective fading channels. The RAKE receiver can be considered as an equalizer with linear complexity with respect to L_p. Furthermore, the RAKE receiver allows L_p independent observations about the same information to be combined before detection on the information symbol is done. Hence, it allows L_p-order frequency diversity. The operation of RAKE receiver is explained below.

Figure 4.10 illustrates the processing of RAKE receiver in spread spectrum systems. For simplicity, consider a frequency selective fading channel with two multipaths. The complex channel fading coefficients of the two paths are $\{h_1, h_2\}$ and the corresponding delays are $\{\tau_1, \tau_2\}$. At the transmitter, we assume the information signal $s(t)$ (with symbol duration T_s) is spread by the sequence $c(t)$ (with chip duration T_c). The received signal is given by:

$$y(t) = \underbrace{h_1 s(t-\tau_1)c(t-\tau_1)}_{\text{path 1}} + \underbrace{h_2 s(t-\tau_2)c(t-\tau_2)}_{\text{path 2}} + z(t)$$

To extract the information for detection, the receiver has to perform despreading. Suppose the receiver knows the spreading sequence $c(t)$ and the path delays τ_1, τ_2,

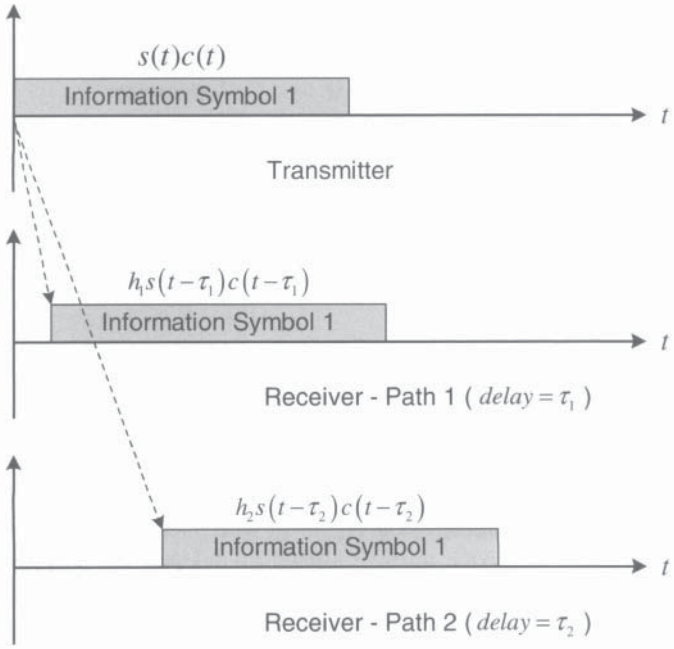


Figure 4.10 Illustration of the RAKE receiver processing.

there are two choices of despreading, namely despreading with respect to path 1 ($c(t-\tau_1)$) or despreading with respect to path 2 ($c(t-\tau_2)$). Consider the former case, the output after despreading is given by:

$$\begin{aligned} r_1(t) &= < c(t-\tau_1), y(t) >= h_1 s(t-\tau_1) < c(t-\tau_1), c(t-\tau_2) > \\ &\quad + h_2 < c(t-\tau_1), s(t-\tau_2)c(t-\tau_2) > + < c(t-\tau_1), z(t) > \\ &= \underbrace{h_1 s(t-\tau_1)}_{\text{Desired signal}} + \underbrace{h_2 s(t-\tau_2) < c(t-\tau_1), c(t-\tau_2) >}_{\text{ISI}} \\ &\quad + \underbrace{\tilde{z}(t)}_{\text{channel noise}} \end{aligned} \tag{4.13}$$

where $< c(t-\tau_1), y(t) >= \frac{1}{T_s}\int_{\tau_1}^{\tau_1+T_s} c(t-\tau_1)y^*(t)dt$, $< c(t-\tau_1), c(t-\tau_1) >= 1$ and the equivalent noise $\tilde{z}(t)$ has the same power as the original noise $z(t)$. Note that the second term represents the ISI term but the ISI power is weighted by the autocorrelation $< c(t-\tau_1), c(t-\tau_2) >= R_c(\tau_1-\tau_2)$. It can be shown that the autocorrelation of a properly designed pseudo-random noise (PN) sequence is upper bounded by $R_c(\tau) \approx \sqrt{fracT_sT_c} = \sqrt{N_{SF}}$. Hence, the ISI power is suppressed by the spreading factor N_{SF} and is given by:

$$ISI_1 = S/N_{SF}$$

The SINR of the decision variable based on $r_1(t)$ is given by:

$$SINR(1) = \frac{S}{N + S/N_{SF}}$$

Compare with the system without spread spectrum, the error floor of the BER vs SNR curve will be lowered because the SIR is increased by N_{SF} times due to the processing gain of spread spectrum. Hence, the RAKE receiver acts as a simple equalizer with respect to frequency selective fading channels. Figure 4.11 illustrates the BER curve of the spread spectrum system based on despreading with respect to the first path.

Similarly, if the receiver despread with respect to the second path $c(t-\tau_2)$, the despread signal is given by:

$$\begin{aligned} r_2(t) &= < c(t-\tau_2), y(t) >== \underbrace{h_2 s(t-\tau_2)}_{\text{Desired signal}} \\ &\quad + \underbrace{h_1 s(t-\tau_1) < c(t-\tau_2), c(t-\tau_1) >}_{\text{ISI}} + \underbrace{\tilde{z}(t)}_{\text{channel noise}} \end{aligned} \tag{4.14}$$

Hence, the residual ISI power is given by S/N_{SF}. Observe that while the error floor is reduced due to the processing gain, the slope of the BER curve is still flattened.

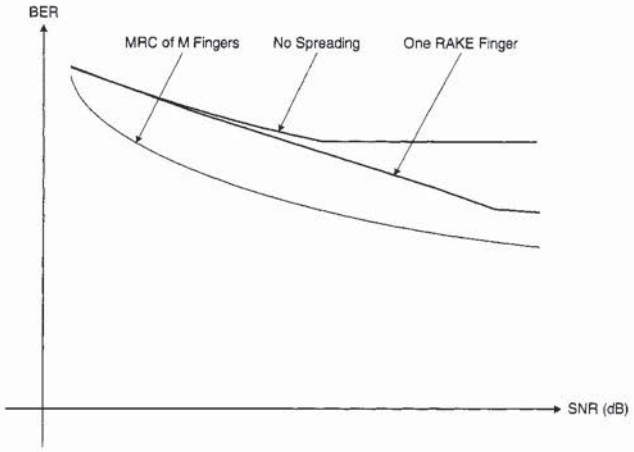


Figure 4.11 BER versus SNR of the RAKE receiver in frequency selective fading channels.

However, since $r_1(t)$ and $r_2(t)$ contain independent observations about the information signal, we can better exploit the inherit frequency diversity by combining the two observations before detecting the signal. In this way, we achieve two-order frequency diversity. The BER curve of the MRC combined output based on $r_1(t)$ and $r_2(t)$ is illustrated in Figure 4.11 Due to the two-order diversity, the slope of the BER curve is increased. Hence, the RAKE receiver kills two birds with one stone.

Figure 4.12 illustrates the block diagram of a typical RAKE receiver. The received signal $y(t)$ is feed into multiple fingers where the $m-$th finger is responsible for despreading with respect to the $m-$th resolvable multipath. The outputs from the M fingers $r_1(t), \ldots, r_M(t)$ are fed into diversity combining block where detection of the information is made on the diversity-combined signal. In general, when the frequency selective fading channel has L_p resolvable multipaths, the RAKE receiver can achieve L_p-th order frequency diversity. Furthermore, since the maximum number of RAKE fingers required is given by L_p, the complexity of the RAKE receiver is linear with respect to L_p

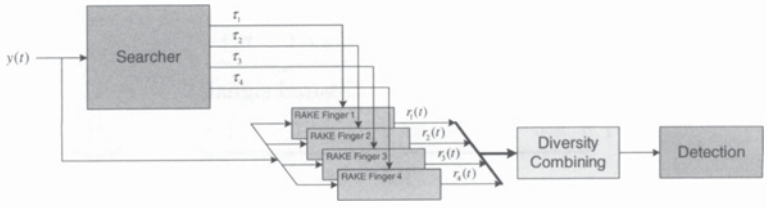


Figure 4.12 Block diagram of the RAKE receiver with $M = 4$ fingers.

In the RAKE processing, we have assumed that the RAKE receiver knows about the path delays $\tau_1, \ldots, \tau_M$. In general, the path delays are estimated by a block called the *searcher*. The searcher consists of a correlator which correlates the received signal with respect to a local spreading sequence $c(t)$. In spread spectrum systems, there is usually a pilot channel where the data symbol consists of all-ones or all-zeros sequence. Hence, the received signal from the pilot channel is given by:

$$y_p(t) = h_1 c(t - \tau_1) + h_2 c(t - \tau_2) + \ldots + h_{L_p} c(t - \tau_{L_p}) + z(t)$$

where $c(t)$ is the spreading sequence used in the pilot channel. Based on the pilot channel, the searcher obtains correlation with respect to the spreading sequence at various delays, i.e.,

$$r(\tau) = < y_p(t), c(t - \tau) > = h_1 R_c(\tau - \tau_1) + \ldots + h_{L_p} R_c(\tau - \tau_{L_p}) + \tilde{z}(\tau)$$

Hence, the correlation output value $r(\tau)$ will hit large values only when $\tau \approx \tau_m$ for $m \in [1, \ldots, L_p]$. Figure 4.13 illustrates a typical searcher output versus the delays. The paths with the correlation values exceeding some predefined thresholds are used to configure the RAKE fingers to perform despreading and RAKE processing.

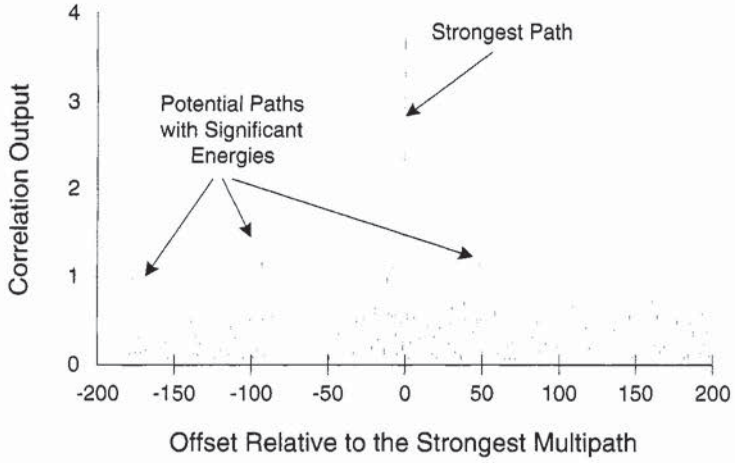


Figure 4.13 Illustration of searcher results—correlation values versus path delays.

4.7 SUMMARY

In this chapter, we have elaborated the effects of fading channels on the physical layer performance of a digital communication link. For instance, flat fading channels introduces the flattening of the BER vs SNR curve where the slope is reduced to -1.

Hence, for the same target BER, more power is required for digital communication over flat fading channels. In multi-user systems, the SNR loss is translated into capacity loss.

For frequency selective fading channels, there are two penalties on the BER curves. In addition to the flattening of the BER curves, we have an irreducible error floor. This error floor is caused by the intersymbol interference (ISI) in the frequency selective fading channels. The problem of ISI cannot be resolved by simply transmitting at a higher power because the ISI power will also increase with the transmit signal power.

A common solution for combating flat fading is diversity. Diversity refers to the technique of detecting information based on multiple independent observations with respect to the same information symbol. With L-th order diversity, the slope of the BER vs SNR curve becomes $-L$ and hence, in the limit of very large diversity order, the slope of the BER vs SNR curve approaches that of the AWGN performance. There are three common techniques for diversity combining, namely the maximal ratio combining, equal gain combining as well as selection diversity. The MRC requires accurate knowledge of the channel fading coefficients from the L independent observations but achieves the best performance. The EGC requires just the phases of the channel fading coefficients and achieves slightly inferior performance relative to MRC. Finally, the SDC requires knowledge of the amplitude of channel fading coefficients and achieve the worst performance of the three combining schemes. However, in all the three schemes, the slope of the BER curve is given by $-L$. The difference in the three schemes are just the *offsets in SNR* in the BER vs SNR curves. Diversity can be realized in time domain (through error correction coding over fast fading channels) or frequency domain (through wideband transmission over frequency selective fading channels) or in spatial domain (through multiple antenna transmission / receptions).

The fundamental solution for frequency selective fading channels is more than diversity. On the contrary, channel equalization at the receiver is needed to cancel the effects of the ISI and hence, alleviating the irreducible error floor. The optimal equalizer is called the MLSE which has exponential order of complexity with respect to the number of resolvable multipaths L_p. For moderate to large L_p, the MLSE becomes impractical in terms of implementation.

Finally, we illustrate the use of RAKE receiver to achieve frequency diversity as well as combating the ISI in CDMA systems. Due to the interference suppression capability of spread spectrum, the ISI power is reduced by the spreading factor N_{SF}. On the other hand, by despreading with respect to different path delays, each RAKE finger produces one independent observation about the information signal. Hence, by combining the L observations from the L RAKE fingers, we can achieve L-th order frequency diversity. In general, the maximum diversity order we can achieve in RAKE receiver is given by $L = L_p$ (number of resolvable multipaths). The complexity of the RAKE receiver is only linear with respect to the number of resolvable paths L_p.

PROBLEMS

4.1 Power control is used to overcome the inter-symbol interference problem in frequency selective fading channel [True/False].

4.2 Transmitting the same information at different time instances, separated smaller than the coherence time of the fading channel, is an effective means to combat frequency flat fading [True/False].

This Page Intentionally Left Blank

PART II

CELLULAR WIRELESS TECHNOLOGIES

This Page Intentionally Left Blank

CHAPTER 5

OVERVIEW AND EVOLUTION OF CELLULAR TECHNOLOGIES

Every cell comes from a cell. (*Omnis cellula e cellula.*)
—Rudolf Virchow: *Cellular-Pathologie*, I, 1858

5.1 INTRODUCTION

In this chapter, we shall apply the theories we have introduced in Part I and review the design of cellular systems. A cellular system [203] is a collection of entities. Figure 5.1 illustrates a generic architecture of modern cellular systems. A cellular system consists of three basic elements, namely the *mobile station* (mobile phone)[MS], the *base station* [BS], as well as the *mobile switching center* [MSC].

The mobile stations are connected to the base station over the *radio interface* while the base stations are connected to the mobile switching center (MSC) through fixed line transmission systems such as T1, E1 or ATM. The service area is covered by multiple base stations where the coverage of each base station is restricted to a limited range only. Mobile stations cannot communicate directly with each other and

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

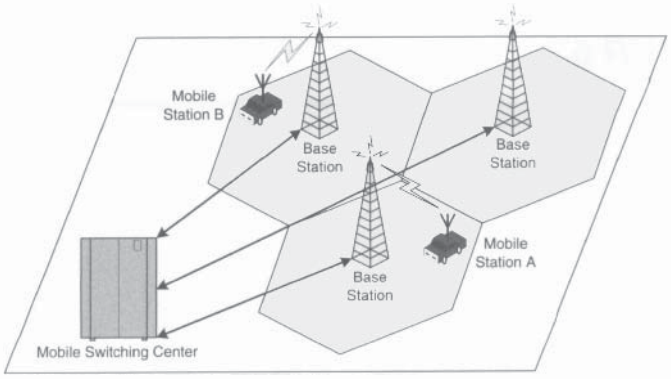


Figure 5.1 A generic architecture of cellular system.

all communications have to go through the base station and the switch. For example, suppose mobile station A establishes a connection with mobile station B. The end to end connection path between A and B consists of two radio segments [from mobile A to BS and from BS to mobile B] and two fixed-line segments [from BS to MSC and from MSC to BS].

In general, the bottleneck of the cellular system (say in terms of the maximum number of users that can be supported or system capacity) is on the radio interface. Let us say the capacity of the fixed line segment is used up. We can easily increase the capacity by installing an additional line. On the other hand, this is not possible in radio or wireless communications. This is because the cellular system has to operate in a license spectrum where the radio signal can only be transmitted in a given range of frequencies. Once the bandwidth in the allocated spectrum is used up, the radio capacity will be used up as well. Hence, various technologies such as FDMA, TDMA, CDMA or SDMA have been developed to utilize the limited spectrum more efficiently. Furthermore, the technologies of the radio interface form an important part in defining the standards of 1G, 2G, 3G as well as 4G systems. In Part I, we have reviewed basic communication theories from point-to-point design to multi-user and multi-cell designs. In Part II (starting from this chapter), we shall elaborate more on the specific designs of industrial standards on cellular systems. We shall first discuss the motivations of the evolution of the cellular technologies from 1G to 3G in the following sections.

5.2 EVOLUTION OF CELLULAR SYSTEMS

In this section, we shall outline the evolution of various cellular systems from 1st generation systems to the current 3rd generation systems. Figure 5.2 illustrates the timeline of the evolution of various cellular systems.

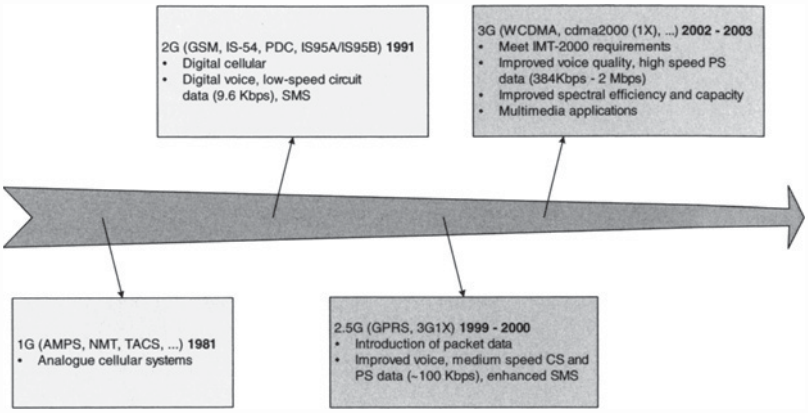


Figure 5.2 Timeline of the evolution of some common cellular systems.

The first generation (1G) cellular systems [202] are also called *analog cellular systems* because the voice is delivered using FM. Commercial deployment started in around 1981. Common examples are the Advanced Mobile Phone Systems (AMPS), the TACS as well as the NMT systems. The AMPS is the US standard and the TACS is the UK standard. Moreover, the 1G systems support voice communications only and the different standards are incompatible with each other. Hence, no roaming across different 1G systems is possible. Channels are partitioned based on FDMA approach. As we have illustrated in Chapter 3, the frequency reuse factor in 1G systems are quite high ($K = 7$) due to the high SIR requirement for FM transmission (SIR 18dB). Hence, the voice capacity (number of voice channels in the system) is in general a primary concern in the 1G systems because the multi-cell spectral efficiency of the radio interface is quite low.

On the other hand, the second generation (2G) cellular systems [324] are called the *digital cellular systems* because the voice transmission is based on digital communication technologies. The speech signal is first digitalized and compressed into digital bit streams by speech vocoder. Hence, the speech channel is *less bulky* compared with the 1G systems. That means for the same channel bandwidth, 2G systems can support more speech users compared to the 1G systems. Hence, the total number of voice channels that can be partitioned from the allocated spectrum is increased and this contributes to the capacity gain in 2G systems. Moreover, the digitalized speech is error protected and transmitted over the air using digital modulation techniques

as introduced in Chapter 1. Due to the digital transmission techniques, the required SIR is generally lower (SIR ~13dB or lower) and this allows a more aggressive frequency reuse. Hence, the voice capacity in 2G systems scales better as we increase the number of base stations compared with 1G systems. Common examples of 2G systems are GSM [324], IS54 (Digital AMPS) and IS95 (CDMA) [127]. 2G systems can support not only the voice service but also the digital data service.

In 2.5G systems [141], such as GPRS, EDGE and 3G1X, a new wireless data service, namely the *packet switched service*, is introduced. In 1G and 2G systems, the primary application is voice service which is supported by *circuit-switched connection*. By circuit switched connection, a dedicated channel is setup for the user during the entire communication session. The user has to pay for the *air time* for the radio resource assigned to it. This connection mode is well suited for voice communication because the source bit rate for voice applications is quite steady and time invariant. On the other hand, in packet switched connection, the dedicated radio resource assigned to the user will be dynamically adjusted based on the traffic requirement. The radio resource may be released to the system if the application source has been idle for some time. The resource will be re-established again when there is a surge of application packets to deliver. Hence, this mode of connection is well suited to accommodate bursty data applications such as video streaming, web browsing or email.

Finally, 3G systems, such as UMTS[152] and CDMA2000 [127] are based on wideband CDMA technology to offer high speed wireless internet access[232] in addition to the conventional voice service. They support a peak bit rate of 2Mbps with Quality of Service (QoS), improved spectral efficiency and multimedia services.

In general, a cellular system has to go through the following typical life cycle, namely the *initial deployment*, the *growth stage* and the *mature/saturated stage* [203]. Suppose an operator just bid for a licensed spectrum from a regulatory body in a city. The operator is about the setup the cellular system. At this stage, the bottleneck of the system is network coverage rather than network capacity. This is because the number of subscribers will not be too high during the initial stage. The primary concerns of the potential subscribers will be the coverage or the size of service area. Hence, the primary strategy of the operator is to maximize the network coverage with the smallest possible investments. This refers to the initial deployment stage where the operator will install a minimum number of base stations to provide a large network coverage. The operator would like to minimize the number of base station because it is quite expensive to setup a new base station. The cost include the hardware cost of the base station as well as the real estate cost to install the base station. In other words, the coverage of the base stations will be large and we called that *macro-cells* as illustrated in Figure 5.3

As the business of the operator evolves, more and more subscribers may join the network and this shall increase the *traffic loading* of the cellular network. For instance, in some densely populated areas, capacity hotspots may be developed where the total number of subscribers covered by a base station may far exceed the number of channels available in that base station. In that case, the network operator can

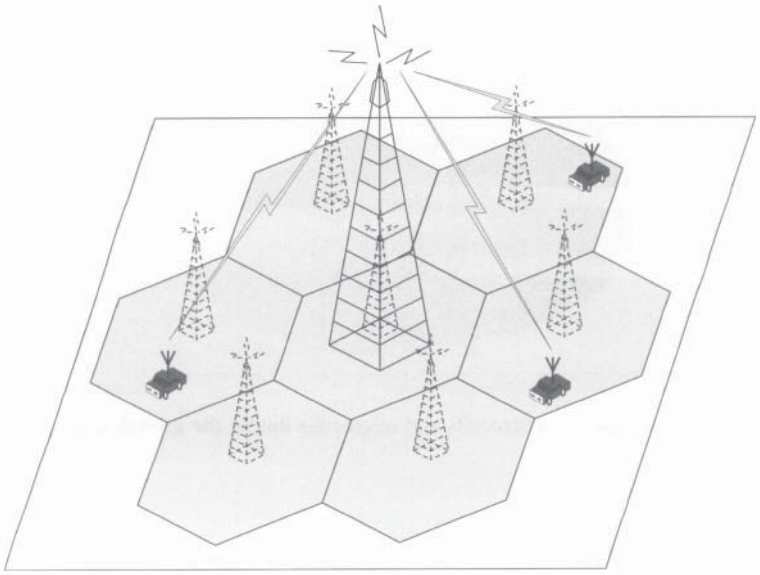


Figure 5.3 Illustration of macrocells during initial deployment stage of a cellular network.

alleviate the hotspot problem by *cell splitting*. Note that cell splitting is a system engineering approach which does not involve technology change or upgrade. By cell splitting, the original macrocell is split into multiple smaller cells. Each cell has a smaller coverage area compared to the original macrocell. With a smaller coverage area, the total number of subscribers served by the base station will be smaller and hence, the traffic loading with respect to the smaller size cell is alleviated. With the same number of channels per cell, the capacity hotspot problems can be resolved. On the other hand, cell splitting requires additional investment on the new base stations and the associated real estate costs. Hence, in the growth stage, the cellular network will have cells of mixed sizes, ranging from macrocell for rural coverage to microcells for higher capacity in urban areas as illustrated in Figure 5.4

Finally, as the cells are split into smaller and smaller cells (such as microcells or picocells), there is a limit on how small a cell can be. Hence, when this limit is hit, the operator cannot further increase the capacity by cell splitting only. In this case, the network is said to reach the matured or saturated stage. The operator has to acquire more licensed spectrum (which is usually difficult) or upgrade the technology to utilize the limited spectrum more effectively.

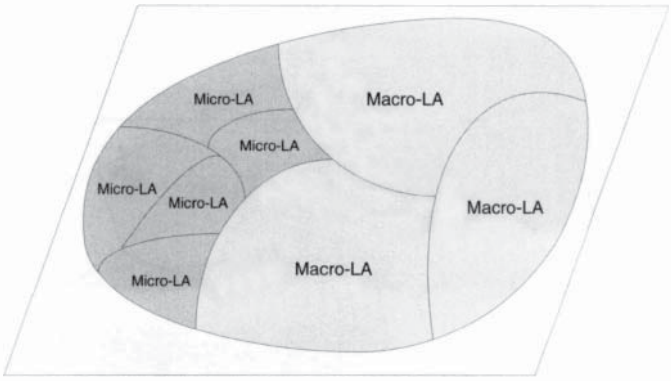


Figure 5.4 Illustration of macrocells and microcells during the growth stage of a cellular network.

5.2.1 Evolution from 1G to 2G

First generation cellular systems are also called analog cellular systems because the voice channels are provisioned through analog modulations. For example, in both the AMPS and TACS systems, voice signals are modulated based on FM with a transmission bandwidth of 30kHz and 25kHz respectively. Signals are separated by FDMA. The total number of channels available to an operator is limited. Furthermore, due to the high SIR requirement for FM transmissions (SIR = 18 dB), the frequency reuse cluster cannot be too small ($K = 7$). Hence, the efficiency of frequency reuse is limited. Due to these two factors, the capacity and the spectral efficiency of the 1G systems are quite limited. 1G operators quickly run out of capacity after cell splitting and reached the mature stage. Hence, there is a major driving force to upgrade the technology to enhance the spectral efficiency in order to meet the ever increasing demand for mobile voice communications.

There are several common second generation cellular standards in the world. For example, we have IS54 (Digital AMPS), IS95 (CDMA), GSM, as well as PDC (Personal Digital Cellular) systems [144] Second generation cellular systems are designed based on digital communication technologies. For instance, the speech signals are digitalized and compressed to 8kbps or 13.3kbps bit streams. The digital bit streams are protected by error correction coding and transmitted over the air using digital modulation methods as introduced in earlier chapters. The digital voice channels are partitioned from the spectrum based on hybrid FDMA/TDMA or hybrid FDMA/CDMA technologies. For example, in IS54 (digital AMPS), the same 30kHz bandwidth carrier is partitioned into three timeslots and hence, it can support 3 digital voice channels compared with one analog voice channel in AMPS system. Hence, due to the digital speech compression, the total number of voice channels available to the operator in D-AMPS is increased by three times compared to the AMPS system. Moreover, due

to the digital modulation and error correction coding, the required SIR is reduced to less than 13 dB. Hence, together with sectorization, a more aggressive frequency reuse factor (e.g., $K = 4$) can be employed and this substantially increases the capacity scaling as more base stations are installed. Hence, these two factors contribute to the substantial increase in the overall network spectral efficiency of 2G cellular systems.

Another major driving force for the evolution of cellular systems from 1G systems to 2G systems is the ability to roam between different countries. There are various incompatible 1G systems such as the AMPS in the US and TACS in the UK. Many of the 1G systems are incompatible with each other and hence, mobile users of one system cannot roam in countries using other 1G systems. In the evolution from 1G to 2G, the US and the European countries took on different approaches. In the US, there is a widespread AMPS deployment over the country. Since US has a large domestic market and there is no issue of roaming across US because there is only one standard (AMPS) deployed, the US took an *evolutionary approach* in upgrading from 1G to 2G systems. Specifically, the 2G systems in the US are designed to be *backward compatible* with the AMPS systems. For example, IS54 or Digital AMPS system is designed to be backward compatible with AMPS systems. The channel bandwidth (30kHz) in D-AMPS is the same as that in AMPS system. In D-AMPS, the voice channels are further partitioned from the 30kHz carrier using TDMA. *Inter-system handoff* is supported when the D-AMPS mobile moves into AMPS coverage. Similar, IS95 (CDMA) system is also designed to be backward compatible with AMPS. Hence, in the US, the 2G digital cellular systems deployed co-exists with the 1G AMPS systems. In fact, a large percentage of rural areas in the US are still covered by the AMPS system.

On the other hand, the European countries took a different approach, namely the *revolutionary approach*, in upgrading from 1G to 2G systems. The reason behind is due to the incompatible standards in the 1G systems among European countries. Hence, during the upgrade to the 2G systems, the European countries have to scrapped the deployed 1G systems and replaced with a common standard 2G systems, namely the GSM systems. Hence, there is no such thing as a dual mode GSM / TACS handset. The GSM system also does not support inter-system handoff (say from GSM to TACS). In fact, GSM systems have around 474 networks in 190 countries worldwide and capture more than 70% of the cellular market. It has a rapid growth in subscribers reaching around 900 million subscribers in 2003. Lately, there are some GSM deployments in some hot spots of the US.

New spectrum has been allocated worldwide to support 2G cellular systems. There are *cellular spectrum* and *PCS (Personal Communication System) spectrum* allocated by the telecom regulators. In the US, the cellular spectrum is located at around 800MHz and the PCS spectrum is located at around 1.9GHz. In most of the other places (such as Europe, China, HK, Taiwan), the cellular spectrum is located at around at around 900MHz and the PCS spectrum is located at around 1.8GHz. Hence, a *tri-band* GSM mobile (covering the cellular spectrum 900MHz, the PCS spectrum

1.8GHz and the US PCS spectrum 1.9GHz) is required to support roaming from HK to the US. Note that the term *PCS* or *Cellular* refer to the spectrum service definitions and is usually technology independent. The operators are free to choose the appropriate technologies (such as GSM, CDMA, D-AMPS, PHS, etc.) to apply on the PCS or cellular spectrum respectively. The cellular spectrum was first allocated for the introduction of 2G systems. Later on, telecom regulators introduced the PCS spectrum to enhance competitions among the cellular industry. In fact, in some cases, the same operator has the license to operate in both the cellular spectrum and the PCS spectrum. To save cost, the operator may choose the same technology to deploy in these two spectrum. For example, we have GSM for cellular spectrum and GSM (also called DCS-1800) for the PCS spectrum.

In addition to the conventional voice service, the 2G cellular systems also support limited wireless data service such as low bit rate circuit switched service (9.6kbps or 14.4kbps transparent or non-transparent bearers) or fax service. Yet, the application of wireless data in 2G systems is limited due to the low bit rate and poor performance.

5.2.2 Evolution from 2G to 2.5G and 3G Systems

While the major driving force of moving from 1G to 2G systems is voice capacity and spectral efficiency, the major driving force for 2G to 3G is on the demand for high quality *wireless internet access*[62]. Similar to the 2G systems, the regulators worldwide has introduced new spectrum for the 3G systems. Figure 5.5 illustrates the worldwide spectrum allocations. Paired spectrum at 1.9GHz (Uplink) and 2.1GHz (Downlink) have been allocated for 3G applications.

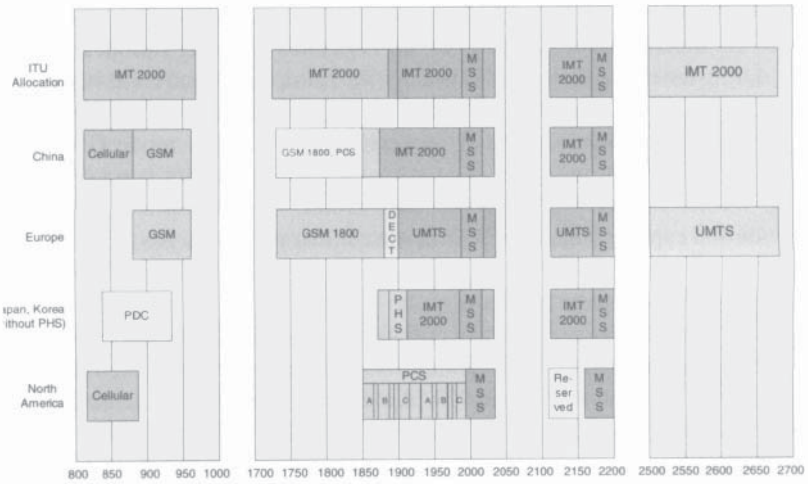


Figure 5.5 Illustration of spectrum allocation for 3G service.

As illustrated in the figure, the IMT2000 defined 3G spectrum is unused in most countries except the US. In the US, all the spectrum around the IMT2000 spectrum has been deployed for other services. Hence, around year 2000, people generally believe that US will be lagging behind the European countries in terms of 3G deployment. In fact, before the internet bubble burst around year 2000, many European operators have spent tens of billions of dollars to bid for the 3G spectrum. At that time, the business plans and financial plans of 3G operators worked fine to justify the expensive license fees. On the other hand, the US operators did not have the chance to bid for 3G spectrum because the spectrum is simply not available. Very soon after the internet bubble burst, the high technology industry collapsed and the financial forecasts / business plans of the 3G operators that sounded reasonable before became unrealistic after the burst. The demand for wireless data access and the *killer applications* became questionable. Hence, most of the 3G operators that spent a lot of money to bid for the license were in financial troubles. In the toughest times, some operators did not even have the budget to acquire 3G base stations and 3G infrastructure. Many operators have to form joint venture to share the equipment costs and the operating costs. Hence, the deployments of 3G systems were delayed for many years until around 2004. In the US, the situation is totally different. The PCS and cellular operators in the US somehow escaped the bidding of 3G licenses and hence, they are relatively *wealthy* after the internet bubble burst. They were ready to invest in the infrastructure to try out wireless internet access using *2.5G technologies* such as 3G1X systems. Hence, the deployment of the wireless data service in the US turned out to be earlier than the rest of the world.

In the new 3G spectrum allocated, there are new service definitions associated with it. One of the most important difference from the 2G systems is the introduction of packet switched connections. Conventionally, circuit switched connection is set up to support voice service. In circuit switched connection, the same dedicated radio resource is assigned to the user throughout the entire communication session. For voice service with *constant bit rate* (CBR) vocoder, the source bit rate is quite steady and it is very effective to support the CBR voice source using circuit switched connection. On the other hand, for data applications, the source bit rate is quite bursty and it is very ineffective to support the bursty application using circuit switched connection as illustrated in Figure 5.6 As illustrated in the figure, if the bit rate of the circuit switched connection is at the average source bit rate, then there will be overflow during the peak source rates. The overflow packets cannot be buffered due to the delay constraint and will be dropped at the transmitter. On the other hand, if the bit rate of the circuit switched connection is set at the peak rate, then most of the time, the connection will be under-utilized. In packet switched connection, the dedicated resource will be dynamically allocated according to the demand (buffer status). When the source demand drops, the dedicated radio resource will be released to the system. Hence, packet switched connection is more suitable to support bursty applications. This efficient use of scarce radio resources means that large numbers of users can potentially share the same bandwidth and be served from a single cell.

The actual number of users supported depends on the application being used and how much data is being transferred.

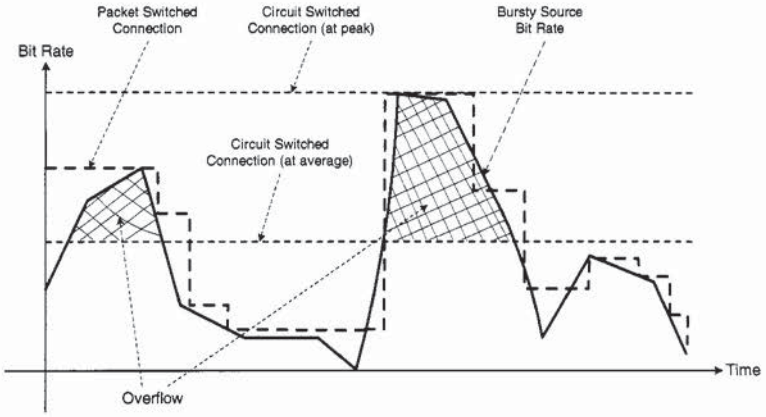


Figure 5.6 Illustration of circuit switched and packet switched connections to support bursty data applications.

The following briefly summarizes the new 3G service definitions.

Circuit Switched Voice Service: Similar to the conventional 2G systems, the 3G systems also support circuit switched voice services. Due to the circuit switched nature, the users are charged by *air time*.

High Speed Circuit Switched Data Service: High speed circuit switched data service up to 64kbps is supported in 3G system. Similarly, the users are charged by the *air time*.

High Speed Packet Switched Data Service: The 3G systems introduced a new data service, namely the *packet switched data service*. The radio resource assigned to support the packet switched data service is adaptive with respect to the traffic volume. Hence, when there is no active packets to transmit in the application, the dedicated radio resource will be released to the system to be shared by other active users. Hence, this mode of data service is very efficient with respect to the bursty source applications. The users will be charged by packet volumes instead of purely air time. Depending on the environment and the mobility of the users, the 3G systems support peak bit rates of 14.4kbps to 2Mbps. Table 5.1 illustrates the bit rates supported under different situations.

The 3G systems also support asymmetric traffic where the downlink bit rate can be different from the uplink bit rate. In 3G systems, there are four QoS classes defined, namely the *conversational class*, the *streaming class*, the *interactive class* and the *background class*. Table 5.2 illustrates typical configurations of the QoS class and the connection modes in various common applications.

Table 5.1 Illustration of peak bit rates of 3G systems in different scenarios.

Environment	Mobility	Peak Bit Rate (kbps)
Indoor	Pedestrian	2M
Outdoor	Vehicular	144
Outdoor	Pedestrian	384

Table 5.2 Illustration of services and applications mappings to 3G connections.

Type	Rate (kbps)	Connection Type	Mode	QoS Class
Voice	8–16	Circuit	Symmetric	N/A
Voice	8–16	Packet	Symmetric	Conversation
Web Browsing	128–2000	Packet	Asymmetric	Interactive
Video Streaming	64–384	Packet	Asymmetric	Streaming
Email, Fax, SMS	9.6–64	Packet	Asymmetric	Background

While 3G systems are designed with respect to the new 3G spectrum and the new 3G service definitions, there are some technology enhancements over the existing 2G systems that operates on the existing 2G PCS spectrum only. These new technologies are called *2.5G systems*. Examples are GPRS, EDGE and 3G1X systems. The 2.5G systems offer packet switched connections and offer an advantage for existing 2G operators to try out the market for wireless data access before fully upgrading the systems to 3G technologies. Figure 5.7 illustrates the evolution paths from 2G systems (GSM and IS95) to the 2.5G and 3G systems. According to Ovum, there will be 36.3 million users of EDGE technology at the end of 2005. Due to the very small incremental cost of including EDGE capability in GSM network deployment, virtually all new GSM infrastructure deployments are also EDGE capable and nearly all new mid- to high-level GSM devices also include EDGE radio technology.

5.3 TECHNICAL CHALLENGES TO REALIZE 3G SERVICES

In the previous sections, we have discussed the evolution of the cellular systems from 1G to 3G. In this section, we shall discuss briefly on the technical challenges involved in realizing the 3G services. In particular, there are three major aspects of technical challenge. They are briefly summarized in the following subsections. We shall outline using GPRS, EDGE and UMTS as examples on how these systems overcome the corresponding technical challenges.

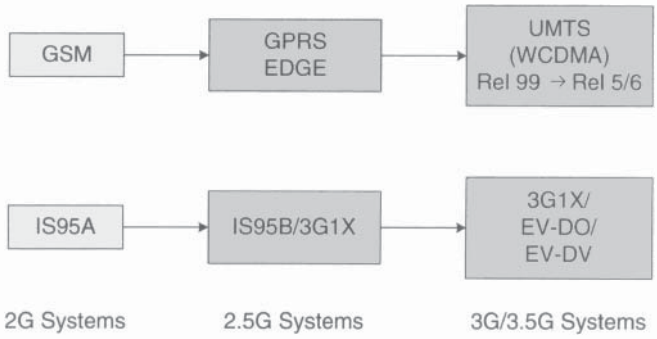


Figure 5.7 Evolution path of 2G systems (GSM, IS95) to 2.5G and 3G systems.

5.3.1 Transmission over the Air

In 3G systems, the physical bit rates of the radio connections are much higher than 2G systems. Hence, one important challenge is to design the radio interface to support such high bit rate requirement. As we have mentioned in earlier chapters, the radio interface is usually the bottleneck in the overall network capacity due to the limited radio spectrum. In other words, higher spectral efficiency is needed to support such higher bit rates given the limited bandwidth.

In GPRS, the air interface is very similar to GSM systems and higher bit rate is supported by *time slot bundling*. Specifically, a GPRS frame has eight time slots and by assigning multiple time slots to a mobile, the bit rate can be increased. In GPRS systems, the specification allows a maximum of assigning 8 time slots to a mobile. However, in practice, the number of time slot assigned to a mobile is limited by the capability of the GPRS mobile. In GPRS mobile, there is a specification of $N + 1$. This means that the GPRS mobile supports N downlink time slots and 1 uplink time slot. For example, most of the GPRS mobiles nowadays support $4 + 1$, meaning that it can support 4 time slots in the downlink. On each time slot, GPRS supports four frame formats (CS1, CS2, CS3 and CS4) with data rate ranging from 9.6kbps to about 21.5kbps. The CS1 format has a data bit rate of 9.6kbps and has the highest level of redundancy for error protection. The CS4 format has a data rate of around 21.5kbps and has no error protection at all. Hence, the theoretical maximum GPRS data rate is given by $8 \times 21.5kbps = 172kbps$. Achieving the theoretical maximum GPRS data transmission speed of 172 kbps would require a single user taking over all eight timeslots without any error protection. Clearly, it is unlikely that a network operator will allow all timeslots to be used by a single GPRS user. Additionally, the initial GPRS terminals are expected be severely limited- supporting only one, two, three or four timeslots. The bandwidth available to a GPRS user will therefore be severely limited. As such, the theoretical maximum GPRS speeds should be checked against the reality of constraints in the networks and terminals. Relatively high mobile data

speeds may not be available to individual mobile users until Enhanced Data rates for GSM Evolution (EDGE) or Universal Mobile Telephone System (3GSM) are introduced.

Further enhancements to GSM networks are provided by Enhanced Data rates for GSM Evolution (EDGE) technology. EDGE provides up to three times the data capacity of GPRS. Basically, EDGE only introduce a new modulation technique (8PSK) and new channel coding that can be used to transmit both packet-switched and circuit-switched voice and data services. EDGE is therefore an add-on to the GPRS system and cannot work alone. Table 5.3 compares the air interface performance of GPRS and EDGE systems. Although the symbol duration of GPRS and EDGE systems are the same, the modulation bit rates are different. EDGE can transmit three times as many bits as GPRS during the same period of time and this is the main reason for the higher bit rates in EDGE system. Using EDGE, operators can handle three times more subscribers than GPRS; triple their data rate per subscriber, or add extra capacity to their voice communications. EDGE uses the same TDMA (Time Division Multiple Access) frame structure, logic channel and 200kHz carrier bandwidth as today's GSM networks, which allows it to be overlaid directly onto an existing GSM network.

Table 5.3 Comparison of GPRS and EDGE air interface specifications.

	GPRS	EDGE
Modulation	GMSK	8-PSK/GMSK
Symbol rate	270 ksym/s	270 ksym/s
Modulation bit rate	270 kb/s	810 kb/s
Radio data rate per time slot	22.8 kb/s	69.2 kb/s
User data rate per time slot	20 kb/s (CS4)	59.2 kb/s (MCS9)
User data rate (8 time slots)	160 kb/s (182.4 kbs/s)	473.6 kb/s (553.6 kb/s)

The EDGE system supports nine channel coding and modulation formats with bit rates ranging from 8.4kbps (MCS1) to 59.2kbps (MCS9). Both GPRS CS1 to CS4 and EDGE MCS1 to MCS4 use GMSK modulation (binary modulation) with slightly different throughput. Figure 5.8 illustrates the modulation and coding formats in GPRS and EDGE systems.

The spectral efficiency of the EDGE system is enhanced compared with GPRS system through *link adaptation*. For instance, the received SIR of mobile stations are time varying. For example, the average received SIR of a mobile maybe 15dB. However, occasionally, the instantaneous received SIR may be higher than the mean (say 20dB). At some other times, the instantaneous received SIR may be lower (say 10dB). Figure 5.9 illustrates the histogram of the received SIR of a mobile.

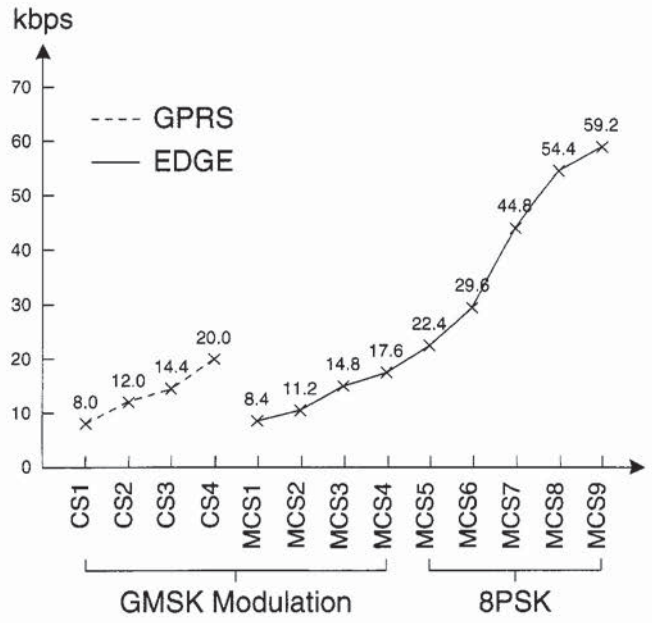


Figure 5.8 Illustration of GPRS and EDGE modulation and coding formats.

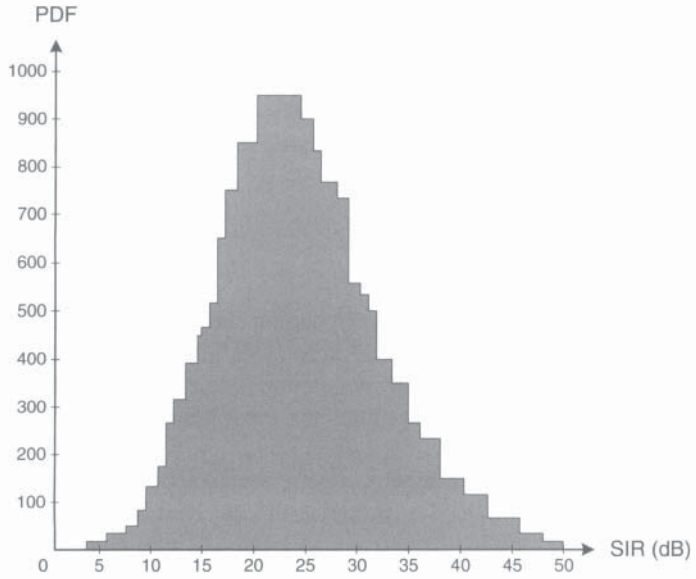


Figure 5.9 Histogram of the received SIR at the mobile.

As we have discussed in earlier chapters, 8PSK modulation offers a higher bit rate at the same bandwidth but a higher received SNR is required to maintain a similar error probability. Hence, by link adaptation, the transmitter (base station in the downlink or mobile station in the uplink) estimates the instantaneous rx SIR and selects the most appropriate modulation and coding formats. If the received SIR is high, the transmitter will selects higher MCS formats which results in higher instantaneous throughputs. On the other hand, if the received SIR is low, the transmitter will select more robust MCS formats for better error protection (resulting in lower throughputs). The modulation and coding formats can be changed for each radio blocks (4 bursts) but a change is usually triggered by new radio channel quality estimates. Hence, the practical link adaptation rate is determined by the measurement interval.

While GPRS and EDGE systems offer an average bit rate of 100kbps and 384kbps respectively, the UMTS system offers a peak bit rate of 2Mbps using wideband CDMA technology. The UMTS system achieves this high bit rate using *variable spreading factor* (VSF) and *multi-code transmissions*. In UMTS, the frequency bandwidth is 5MHz with a chip rate of 3.84Mcps and QPSK modulation. The spreading factor in the downlink changes from 512 (lowest bit rate)[1] to 4 (highest bit rate). The spreading factor in the uplink changes from 256 to 4. Hence, the bit rates of dedicated physical channels can be adjusted by adjusting the spreading factor of the channelization codes. One challenge of supporting variable spreading factor in UMTS is to maintain orthogonality among the codes. This is because for the same chip duration, channelization codes with different spreading factors will have different lengths. This is resolved in UMTS based on *Orthogonal Variable Spreading Factor* (OVSF) as illustrated in Figure 5.10 To maintain code orthogonality, all the codes belonging to the same sub-tree of the assigned code(s) should not be assigned to any users. For example, if $C_{2,1}$ has been assigned, the codes $C_{4,1}$ and $C_{4,2}$ cannot be assigned to any other users.

Due to the OVSF channelization codes, one issue in UMTS is the limitation of the number of orthogonal codes. For instance, if we assign spreading factor of 256 to all users in UMTS, we can support 256 orthogonal downlink channels using D-CDMA. However, if there is high bit rate data user assigned with code channel of spreading factor 4, this single data user will consume one quarter of the orthogonal codes space (with spreading factor 256). Hence, only 3×64 users (with SF=256) can be assigned together with this high bit rate data user (with SF=4). In other words, the UMTS system can easily become code limited if a few high bit rate users are assigned with low spreading factor channelization codes.

In UMTS uplink, a single code channel (with spreading factor 4) can support a bit rate of 400kpbs - 500kbps with coding. Similar to time slot bundling in GPRS, the bit rate of a user can be further increased by assigning multiple codes to a single user. In multicode transmission, the spreading factor of the codes is the same. In UMTS,

[1]In fact, spreading factor of 512 is seldom used in the downlink due to timing synchronization between Node-Bs involved in soft handover.

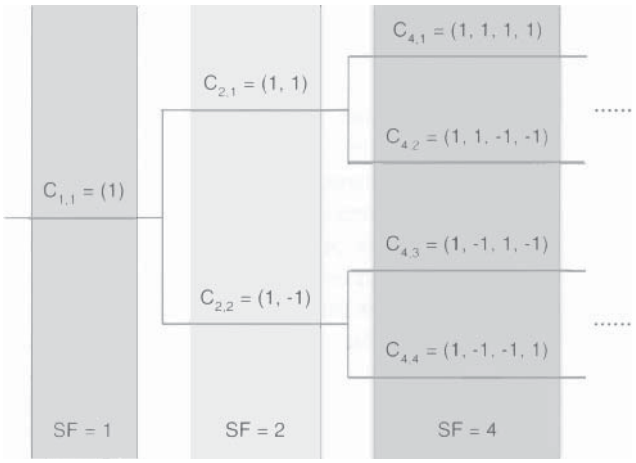


Figure 5.10 Illustration of Orthogonal Variable Spreading Factor (OVSF) in UMTS.

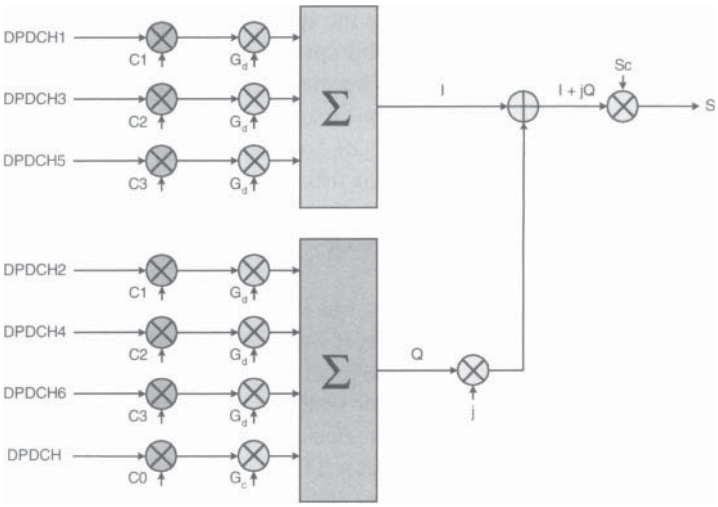


Figure 5.11 Illustration of multicode transmission in uplink of UMTS.

a maximum of six orthogonal codes can be assigned to a user in the uplink, with a total bit rate of around 2.4Mbps. Figure 5.11 and Table 5.4 illustrate, respectively, the multi-code structure as well as the supported data rates in the UMTS uplink.

Figure 5.12 and Table 5.5 illustrate, respectively, the multi-code structure and the supported data rates in the UMTS downlink. A maximum of 3 multi-codes (each

Table 5.4 Supported data rates in uplink of UMTS.

DPDCH Spreading Factor	DPDCH Channel Bit Rate (kbps)	Maximum User Data Rate with 1/2-Rate Coding (Approx., kbps)
256	15	7.5
128	30	15
64	60	30
32	120	60
16	240	120
8	480	240
4	960	480
4, with 6 parallel codes	5740	2.3M

with spreading factor 4) is allowed in downlink resulting in a data channel of bit rate 2.3Mbps.

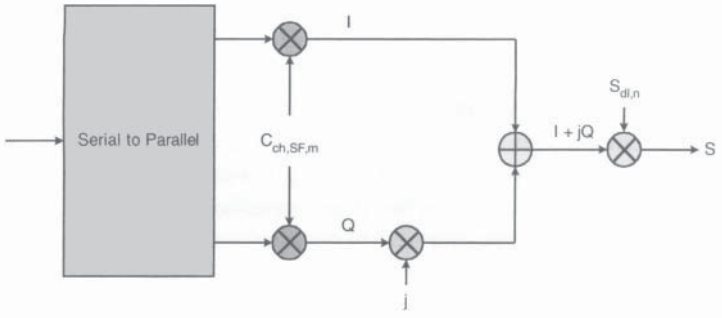


Figure 5.12 Illustration of multicode transmission in downlink of UMTS.

5.3.2 Quality-of-Service (QoS)

In 2G systems, the primary application is voice communications and is supported through circuit switched connections. Hence, there is no issue of maintaining Quality of Service (QoS) in the cellular systems[322]. However, to support wireless data applications in 3G systems, the QoS is an important issue. For instance, there are different types of data and multimedia applications with very different QoS requirement[135]. For example, email applications may be delay insensitive but error sensitive. On the other hand, video streaming may be delay sensitive but could tolerate a higher packet error rate. Hence, it is a challenging task to mix the data

Table 5.5 Supported data rates in downlink of UMTS.

Spreading Factor	Channel Symbol Rate (kbps)	Channel Bit Rate (kbps)	DPDCH Channel Bit Rate Range (kbps)	Maximum User Data Rate with 1/2-Rate Coding (Approx., kbps)
512	7.5	15	3–6	1–3
256	15	30	12–24	6–12
128	30	60	42–51	20–24
64	60	120	90	45
32	120	240	210	105
16	240	480	432	215
8	480	960	912	456
4	960	1920	1872	936
4, with 3 parallel codes	2880	5760	5616	2.3M

and multimedia applications with different QoS requirements efficiently in a cellular system (in particular the air interface QoS differentiation).

In 2.5G systems such as GPRS and EDGE, a descent bit rate is supported in the air interface. Moreover, packet switched data connection is supported for bursty applications. In fact, there were times when people questioned about the demand of 3G systems when GPRS or EDGE systems can deliver packet switched data connections with bit rates of 100kbps or 384kbps. One of the major difference between GPRS/EDGE and 3G systems (besides the physical bit rate) is in fact on the support for QoS differentiation in the air interface. In 2.5G systems, there is no QoS differentiation in the air interface and different users running different applications are treated with the same priority in the allocation radio resource (best effort delivery). This is probably sufficient for typical wireless data applications such as emails. However, for video streaming or interactive service, QoS differentiation is needed and therefore, in 3G systems, four QoS classes, namely the *conversational class*, *the interactive class*, the *streaming class* as well as the *background class*, have been defined. These classes are characterized by maximum bit rates, maximum packet size, transfer delay and traffic handling priority. Table 5.6 illustrates the typical parameter settings for the four QoS classes.

The end-to-end QoS support in UMTS is realized by *QoS provisioning* in both the radio interface as well as the fixed line infrastructure and *QoS signaling*[243]. QoS provisioning refers to the enforcement of QoS in the radio interface and vari-

Table 5.6 QoS parameters for the four QoS classes supported by UMTS Bearer (from 3GPP TS 23.107).

Traffic Class	Conversational Class	Streaming Class	Interactive Class	Background Class
Max. Bit Rate (kbps)	< 2048	< 2048	< 2048 – overhead	< 2048 – overhead
Delivery Order	Yes/No	Yes/No	Yes/No	Yes/No
Max. SDU Size (octets)	≤ 1500 or 1502	≤ 1500 or 1502	≤ 1500 or 1502	≤ 1500 or 1502
Erroneous SDUs Delivery	Yes/No/-	Yes/No/-	Yes/No/-	Yes/No/-
Residual BER	5×10^{-2},10^{-2}, 5×10^{-3},10^{-3}, 10^{-4},10^{-5}, 10^{-6}	5×10^{-2},10^{-2}, 5×10^{-3},10^{-3}, 10^{-4},10^{-5}, 10^{-6}	4×10^{-3}, 10^{-5}, 6×10^{-8}	4×10^{-3}, 10^{-5}, 6×10^{-8}
SDU Error Ratio	10^{-2},7×10^{-3}, 10^{-3},10^{-4}, 10^{-5}	10^{-1},10^{-2}, 7×10^{-3},10^{-3}, 10^{-4},10^{-5}	10^{-3},10^{-4}, 10^{-6}	10^{-3},10^{-4}, 10^{-6}
Transfer Delay (ms)	100 – max. value	100 – max. value		
Guaranteed Bit Rate (kbps)	< 2048	< 2048		
Traffic Handling Priority			1, 2, 3	
Allocation/ Retention Priority	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3

ous interfaces. QoS signaling refers to informing all the network entities about the QoS requirement of the UMTS bearer. Detail description will be elaborated in later chapters.

5.3.3 Infrastructure

For 2G networks, the primary application is circuit switched voice service. The target public network for circuit switched voice service is basically the *public switched*

telephone network (PSTN), which is designed based on circuit switched architecture. Hence, the infrastructure (or core network) in 2G systems is designed based on circuit switched architecture and the interfacing between cellular core network and PSTN is straight-forward (via SS7 in the signalling path and via PCM-64 in the payload path).

However, in 3G networks, the mixed voice and data applications complicated the design of the core network. For instance, the target public network of circuit switched voice service is PSTN and this favors a circuit-switched core network in the cellular system. On the other hand, the target public network of wireless data network is the public internet, which is packet switched. Hence, this favors a packet switched architecture in the core network of the cellular systems. Hence, the challenge in designing the core network is to accommodate both PSTN and internet interfacing for both voice and wireless data applications.

In 2.5G systems such as GPRS and EDGE, the core network design is based on the *overlay approach*. Specifically, a new packet switched layer is added on top of the existing circuit switched layer in the core network as illustrated in Figure 5.13 The voice path involves the mobile station (MS), base station (BS), base station controller (BSC) and the mobile switching center (MSC), which is then interfaced to the PSTN via the gateway MSC. The packet switched data path involves the MS, the BS, the BSC and the SGSN, which is then interfaced to the internet via the GGSN. The GGSN and SGSN are similar to regular IP routers except with GPRS mobility management and GPRS session management capability.

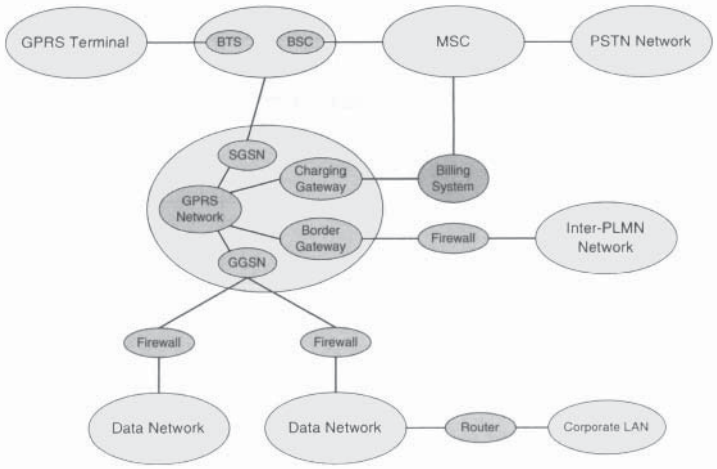


Figure 5.13 Overlay approach in GPRS core network to support circuit switched voice and packet switched data applications.

The advantage of the overlay approach is simple and it does not introduce any impact onto the existing circuit-switched core network (MSC). However, the overlay approach suffers from a disadvantage of higher operation and management cost be-

cause the operator has to maintain two core networks of very different nature (circuit switched core vs packet switched core).

In the earlier releases of 3G systems (such as UMTS Rel 99), the core network design is also based on the overlay approach with MSC nodes responsible for regular circuit switched voice service and SGSN/GGSN nodes responsible for the new packet switched data services. However, the latest trend (such as in UMTS Rel 6) is to integrate the core network into a single packet-switched architecture to support both PSTN and public internet[54]. The interface to the public internet is via GGSN which is pretty straight-forward as GGSN is also an IP router. On the other hand, the circuit switched voice is delivered to the core network in the form of VoIP packets. The interface to the PSTN is via VoIP gateway which converts the VoIP packets and SIP signalling into standard PCM-64 and SS7 signalling respectively.

5.4 SUMMARY

In this chapter, we have elaborated on the evolution of cellular systems from 1G to 3G systems. The primary driving force behind the evolution of 1G systems to 2G systems is voice capacity and international roaming. For instance, 1G systems are designed based on FM transmissions and have quite poor spectral efficiency. Hence, the voice capacity quickly saturated even after cell splitting is employed. Furthermore, the 1G systems are incompatible with each other and no inter-system roaming can be supported. In moving from 1G to 2G systems, the US and the Europe took different approaches. Evolutionary approach was taken in the US where 2G systems were designed to coexist with the existing AMPS operation. However, revolutionary approach was taken in the European countries where all the existing 1G systems were scrapped. GSM, IS54 (D-AMPS) and IS95 (CDMA) are typical examples of 2G systems where the radio interface is designed based on digital transmission technologies. The channels are partitioned based on hybrid FDMA/TDMA and hybrid FDMA/CDMA respectively. Due to the speech compression, more channels are available to the operator. Due to more robust digital transmission, more aggressive frequency reuse can be employed which increase the capacity scalability in multi-cell systems. Hence, the overall network spectral efficiency in 2G systems has been increased substantially compared to 1G systems.

In moving from 2G to 3G systems, the major driving force behind is no longer voice capacity but the demand for high speed wireless and multimedia data services. To support the bursty wireless data applications, packet switched connection is introduced. The dedicated radio resource associated with a data connection is dynamically adjusted according to the demand. Hence, the user can be logically *online* without having to pay for the air time because no dedicated radio resource has been attached to the logical connection. In 3G systems, the maximum peak bit rate supported is about 2Mbps.

Between 2G systems and 3G systems, a number of 2.5G technologies such as the GPRS and EDGE systems are introduced. GPRS increases the bit rate by time slot bundling whereas the EDGE system further increase the bit rate by higher modulation (8PSK) and link adaptation. In UMTS, higher bit rate is supported by variable spreading factor and multi-code transmissions.

Finally, we have elaborated on the three major technical challenges to realize the 3G services. For instance, we have the transmission challenge in the radio interface to support higher bit rate. We have the QoS challenge to support QoS differentiation and end-to-end QoS. We also have challenge of the infrastructure design to support both PSTN and internet interfacing for both circuit switched voice and packet switched data applications.

PROBLEMS

5.1 Second generation wireless systems cannot support digital data communications. [True/False]

5.2 UMTS system allows QoS differentiation in the air interface [True/False].

CHAPTER 6

CDMA (IS-95)

A secret may be sometimes best kept by keeping the secret of its being a secret.
—Henry Taylor: The Statement, 1836

6.1 INTRODUCTION

In this chapter, we shall elaborate the system architecture and the system operations of IS95 CDMA network as one of the 2G system example. We shall focus on the air interface of the IS95 and the circuit switched operations. Specifically, we shall first look at how physical channels are partitioned in the uplink and downlink based on *deterministic CDMA* and *random CDMA*. We shall then look at the call processing procedure in circuit-switched connections. We will look at the processing behind (such as cell search, physical synchronization) as soon as mobile station turns on the power. We will also look at the background process when the mobile station is in idle state (able to see the operator's logo). Next, we will explore the channel access procedure and explain why there is a "SEND" button in mobile stations (but

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau

Copyright © 2007 John Wiley & Sons, Inc.

not in fixed line telephones). Finally, we will elaborate the call origination and call termination processing between the mobile station and the infrastructure.

The next topic we will explore in this chapter is the power control. As explained in earlier chapters, power control is very important to CDMA systems due to near far problem. In fact, quite sophisticated power control procedures have been defined in the reverse link of IS95. The objective is much more than saving battery power but to maintain a reasonable system capacity. We will first look at the open loop power control and explain why the procedure is designed the way it is. We will then look at the inner loop power control and outer loop power control procedures. Finally, we will illustrate the operation with an example.

After power control, the next important topic in CDMA is soft handover. Soft handover is a unique feature in CDMA systems which is not found in GSM or IS54 systems. We shall first elaborate the reason how CDMA systems facilitate soft handover. We will then explore how the mobile station physically measure the pilot signal strength of the home cell and neighboring cell. Next, we will go over the mobile station processing for soft handover, followed by the corresponding infrastructure processing. Finally, we will illustrate the soft handover procedures using an example.

6.2 SYSTEM ARCHITECTURE OF IS95

The IS95 CDMA system can be viewed as a collection of network entities and the associated interfaces. Reference interface is defined as a reference point or interconnections between adjacent network elements. The reference interface is usually specified in terms of a set of procedures and associated signaling that define the operational responsibility of the connected network entities. Figure 6.1 illustrates the standard reference model or system architecture of IS95 system[127]. The IS95 system consists of three parts, namely the mobile station, the radio access network (RAN) as well as the core network (CN). The mobile station subsystem consists of CDMA mobiles (MS). The radio access network consists of base stations (BS) and base station controller (BSC). The core network consists of mobile switching center (MSC), interworking function (IWF), visiting location register (VLR) and home location register (HLR). Note that the IS95 interfaces are fully standardized or open interfaces, which allows the operators to deploy a network consisting of components from different manufacturers. For example, the Um interface is defined for the radio connection between the mobile and the base station (and is specified in the IS95 specification). The A-bis interface is defined for interconnection between base station and base station controller. The A interface[226] is defined for the interconnection between base station controller and the MSC (and is specified in the IS41 specification). In this chapter, we shall primarily focus on the IS95 radio interface.

The protocol architecture of the IS95 radio interface is illustrated in Figure 6.2 There is a user plane and control plane in the protocol stack. The user plane consists of user application data (such as circuit switched voice and circuit switched data).

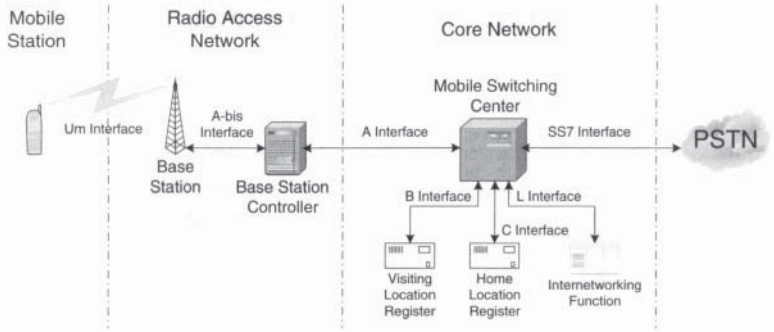


Figure 6.1 System architecture of IS95 CDMA systems.

The control plane supports a variety of functions such as the broadcasting of system information, radio resource management, mobility management and call processing. We will elaborate the functions of the control plane in later sections.

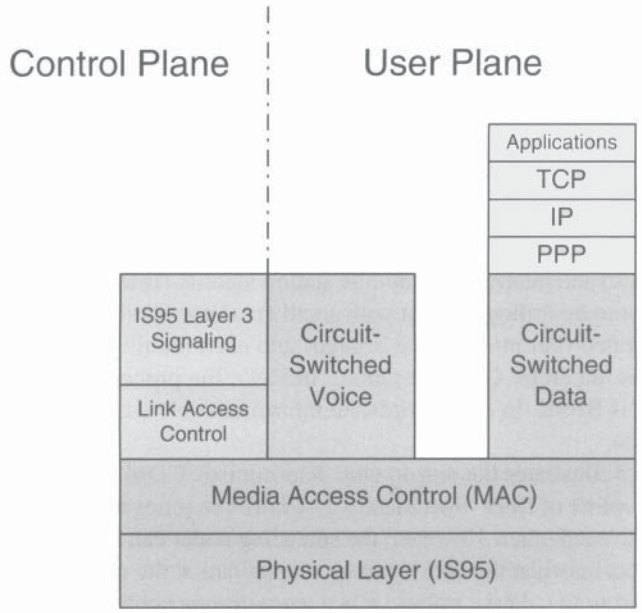


Figure 6.2 Protocol architecture of the IS95 radio interface.

6.3 PHYSICAL LAYER AND PHYSICAL CHANNELS

The physical layer of the IS95 radio interface represents the lowest layer of the IS95 protocol stack and defines the mobile station and base station interoperability procedure over the air interface. The physical layer specifies the physical channel baseband processing such as error correction encoding, interleaving, modulation and spreading (CDMA processing) at the transmitter side and the reverse operations (such as despreading, demodulation, deinterleaving and error correction decoding) at the receiver side. The physical layer also defines the modulation pulse shape, transmit power spectral mask and the radio frequency band over which IS95 can be operated. In this chapter, we shall primarily focus on the baseband processing (specifically CDMA processing) in the IS95 physical layer.

The mobile station and the base station communicates by means of physical channels, which are partitioned from a paired spectrum based on CDMA technology. By paired spectrum, we mean that a dedicated range of frequencies are defined for the uplink (mobile station to base station) and the downlink (base station to mobile station) respectively. The uplink and downlink are separated by *frequency division duplexing* (FDD). The corresponding frequency separation between the uplink channel and the downlink channel is 45MHz in the cellular spectrum (800MHz) and 80MHz in the PCS spectrum (1.9GHz). We shall elaborate on the channel partitioning and the channel structure in the uplink and downlink of the IS95 system.

6.3.1 Channelization in the Uplink

In the reverse link or uplink of IS95 (mobile to base station), the channels are partitioned based on random CDMA (R-CDMA) techniques. Specifically, all the uplink channels are separated by a spreading sequence called the *long code*. The long code is a unique pseudo-random (PN) sequence derived from the user equipment serial number (ESN) and international mobile station identity (IMSI). These long codes are designed not to be orthogonal but with small cross correlation. Hence, the system is not code limited (but interference limited) and each mobile user can have a unique long code in the entire CDMA system. In IS95, the processing gain is 64 and the modulation is BPSK. In other words, an information bit is chopped into 64 chips by the long code.

Figure 6.3 illustrates the reason why deterministic CDMA (D-CDMA) cannot be used in the uplink of IS95. Specifically, D-CDMA requires the spreading codes to be orthogonal to each other. However, the spreading codes can be orthogonal only if the timing offsets between the codes can be maintained at the receiver (synchronous or quasi-synchronous). In the uplink, it is a *many-to-one* configuration where multiple transmitters (at different locations) are communicating to the base station. Since the mobile stations are at random locations, the received signals from the mobile stations suffer from random propagation delays. Hence, the relative timing offsets between the spreading codes at the base station receiver are random and cannot be guaranteed

to be completely synchronized. Hence, the orthogonality of the spreading codes cannot be guaranteed and R-CDMA has to be used to separate the channels.

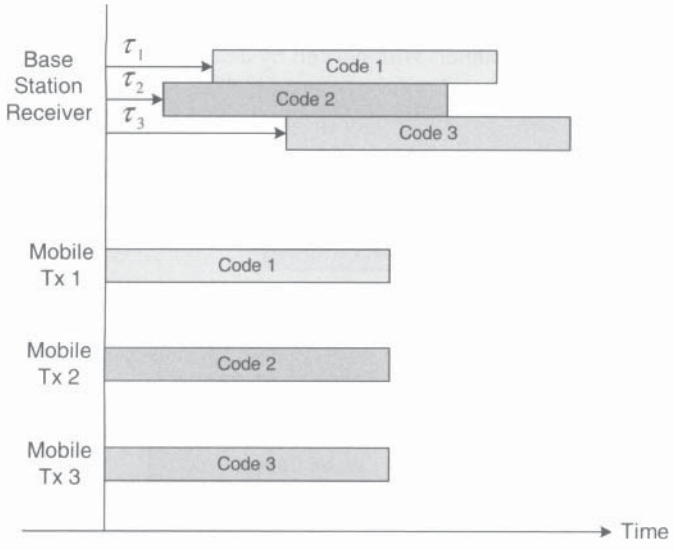


Figure 6.3 Illustration of the random timing offsets between the spreading codes in the uplink of IS95 systems.

While R-CDMA partitions the uplink spectrum into a number of uplink physical channels, the physical channels can be divided into common channels and dedicated channels. Common channels are also called overhead channels because they are shared by all mobiles in the base station. These common channels are always setup in the base station no matter how many active users are there in the base station. Dedicated channels, on the other hand, are assigned to one and only one mobile at a time. Hence, these dedicated channels are revenue generating because once assigned to a mobile station, the user has to start paying for the air time. In IS95, the common channel in the reverse link is called the *random access channel*. The dedicated channel in the reverse link is called the *dedicated traffic channels*. The random access channel is spread by a specific long code (which is defined in the specification and known to the system). The dedicated traffic channels are spread by the unique long codes derived from the mobile user ESN and MSN.

6.3.2 Channelization in the Downlink

In the downlink or forward link of IS95 (base station to mobile stations), the channels are partitioned based on hybrid deterministic CDMA (D-CDMA) and random CDMA (R-CDMA). Specifically, downlink channels are first spread by the *Walsh codes* with spreading factor $SF = 64$. Walsh codes are special spreading sequences which

are mutually orthogonal (zero cross correlation) when they are synchronous (at zero offset). In the downlink, since it is a one-to-many configuration, the Walsh codes between the different downlink channels remain to be synchronous with each other at the mobile receiver as illustrated in Figure 6.4 Hence, the Walsh code spreading is used to separate the downlink channels within a cell by *deterministic CDMA* (D-CDMA).

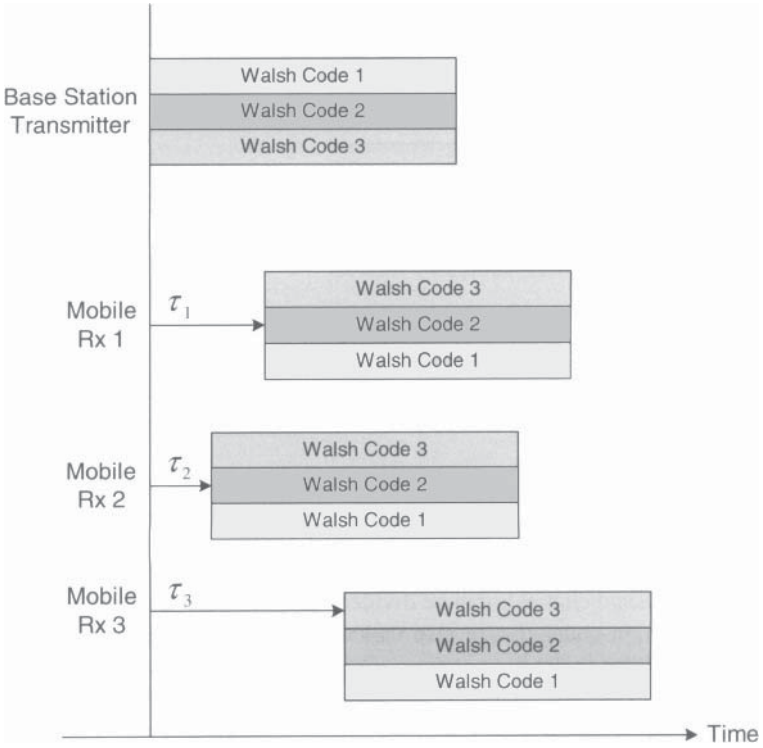


Figure 6.4 **Illustration** of the random timing offsets between the spreading codes in the downlink of IS95 systems.

In addition to the Walsh code spreading, IS95 downlink channels are further spread by a short PN sequence. The short PN sequence is a unique PN sequence per cell. All the downlink channels of the same cell are spread by the same PN sequence associated with the cell. The short PN sequence is used to mitigate inter-cell interference between different base stations in the system through random CDMA (R-CDMA). Hence, the downlink channels are separated by hybrid D-CDMA (same cell channels) and R-CDMA (multi-cell channels). In IS95, all the base stations are synchronous with each other (i.e., they have the same $t = 0$ time reference.). There is a common master short PN sequence defined in the system which is the same short PN sequence used in all the base stations. Each base station has a parameter called *PN Offset* (PNO). The

actual short PN sequence transmitted from a base station is derived from the common master PN sequence except with an offset in the starting phase given by the PNO as illustrated in Figure 6.5 The advantage of this arrangement is simpler cell search procedure (as will be elaborated in the next section). However, the disadvantage is the requirement to accurately synchronize all the base stations in IS95 systems (up to the chip intervals). In practice, such accurate timing synchronization can be achieved using GPS on each base station. Yet, in some densely populated urban areas where the GPS reception is poor, the synchronization between base station will be quite challenging.

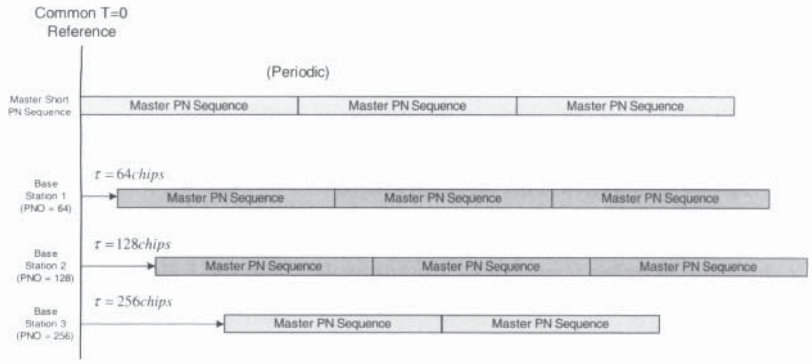


Figure 6.5 Illustration of the common time reference in all base stations and the PN offsets (PNO) in IS95 systems.

Similarly, downlink channels are partitioned into common channels and dedicated channels [190]. Common downlink channels in IS95 include the *pilot channel*, *SYNC channel* and *PAGING channel(s)*. The pilot channel is spread by Walsh code 0 and contains no data (all 1 sequence). It is used for timing acquisition, tracking, channel estimation and soft handover measurement. We shall elaborate on the usage of the pilot channels in later sections. The SYNC channel (spread by Walsh code 32) is a broadcast channel used to convey the *SYNC channel message*. The SYNC channel message contains various important system parameters such as the *system ID* (SID), *network ID* (NID), base station ID (PNO), system time, long code state, paging channel data rate, etc. The paging channel is spread by Walsh code 1 - 7 which carries the *system parameter message*, *access parameter message*, *neighbor list message*, *CDMA channel list message* or *page message*. We shall elaborate the important parameters contained in these system messages in later sections.

On the other hand, the dedicated downlink channels are traffic channels (spread by the remaining Walsh codes) to be assigned to dedicated users. Theoretically, in IS95 systems, there are at most 64 downlink channels per cell. However, in practice, since the dedicated traffic channels are assigned in pairs (uplink and downlink) and the uplink of the IS95 system is interference limited (supporting about 15-16 channels

per sector). Hence, the actual number of voice channels is only around 15-16 per sector.

6.4 CALL PROCESSING

In this section, we shall focus on the call processing procedure in IS95 system. Call processing here refers to the procedure and handshaking between the mobile phone and the infrastructure (radio access network) from the time a mobile phone is turned on to the time that a user press the "SEND" button and establish a traffic channel.

A mobile phone (after power on) can be in one of the four substates, namely the *Initialization State*, the *Mobile Idle State*, the *Mobile Access State* as well as the *Traffic Channel State* [190]. This is illustrated in Figure 6.6 We shall elaborate the processing in each of the substates in the following subsections.

6.4.1 Initialization and Cell Search State

When mobile station is turned on, it will enter into the mobile initialization state. The mobile initialization state can be further broken into the following substates as illustrated in Figure 6.7

System Determination Substate: When the mobile station is powered on, it will first determine which system to use according to user selected preference. As we have discussed in earlier chapters, the IS95 system in the US is designed to co-exist with the AMPS system. Hence, the IS95 specification allows the mobile (dual mode) to acquire onto the AMPS system or the IS95 system. User can specify to acquire CDMA system (IS95) only or CDMA system (IS95) preferred.

In IS95, the channels are partitioned based on hybrid FDMA/CDMA where the cellular spectrum (or PCS spectrum) is first partitioned into a number of 1.5MHz carrier. For each 1.5MHz CDMA carrier, the physical channels are partitioned based on R-CDMA and R-CDMA/D-CDMA as we have explained in the previous section. While the number of CDMA carriers assigned to an operator may be limited, the IS95 mobile should be able to scan all the CDMA carriers in the cellular spectrum (or PCS spectrum).

The specific algorithm for system determination is not specified in the IS95 specification. However, most mobile phone vendors make use of a *preferred roaming list*. The roaming list is a list of CDMA carriers with priority that the mobile phone would attempt to acquire the pilot channel. The preferred roaming list is used in such a way that the mobile station, upon power up, tries to acquire the home system (the system which it is subscribed). Each entry of the roaming list contains the system identifier (SID) and the network identifier (NID) of the system to acquire, the block of CDMA carrier numbers, and the

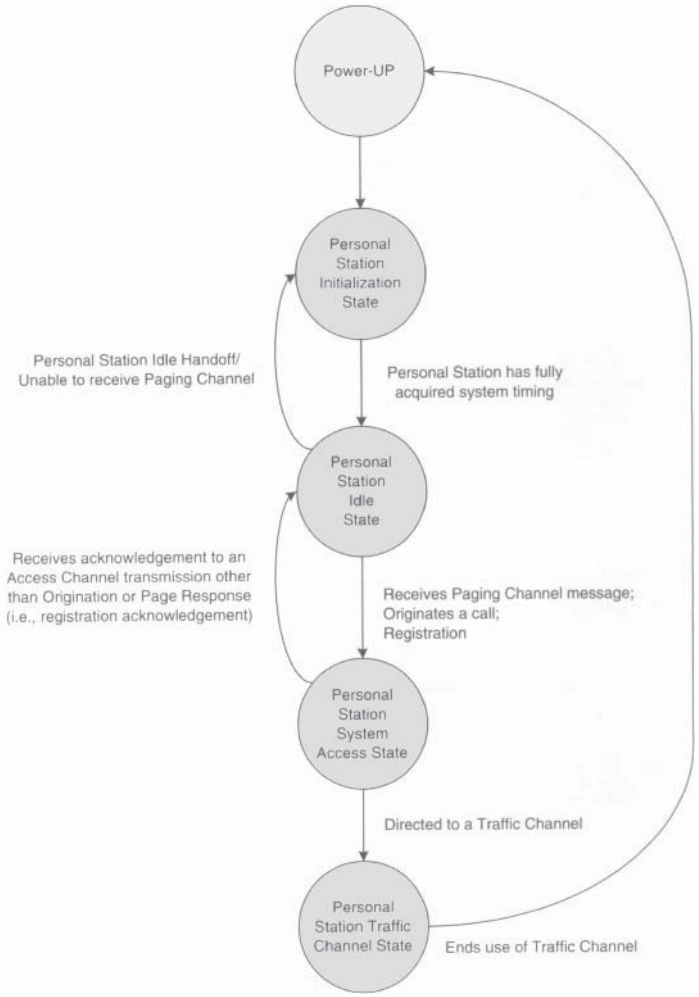


Figure 6.6 Illustration of the four system states when a mobile phone is turned on in IS95 systems.

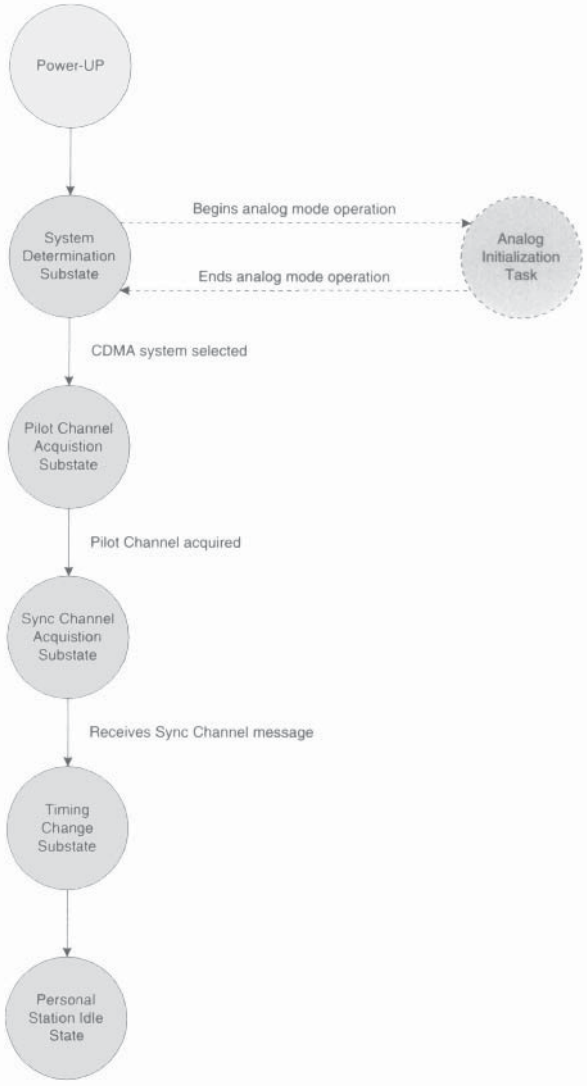


Figure 6.7 Illustration of the mobile initialization states when a mobile phone is turned on in IS95 systems.

corresponding roaming flag. The roaming flag, if turned on, makes the roaming indicator of the mobile station's display flash once service is provided on this entry. The algorithm also accounts for the previously visited systems to speed up acquisition. Some vendors may also maintain a dynamic table listing the CDMA carriers (channels) on which the service was most recently provided.

The system determination procedure builds an ordered scanning list. The top entries are from the most recently visited CDMA carriers, followed by those in the preferred roaming list. Once the scan list is built, the mobile station sequentially attempts to acquire each channel in the scan list using the procedures described below.

Pilot Channel Acquisition Substate: Say the mobile phone attempts to acquire to CDMA system. The mobile will enter into the *pilot channel acquisition substate* and it will attempt to acquire the downlink pilot channel on the selected CDMA carrier. The acquisition to the pilot channel is very important and is the first step before the mobile can decode any channels in the system. As we have discussed in Chapter 3, a CDMA receiver decode the information through the despreading process. The despreading process requires knowledge of the spreading sequences used to spread the information at the transmitter as well as the received timing (or received phase) of the spreading sequence. In IS95 downlink, channels are first spread by the Walsh code and followed by the short PN code. Both the Walsh code (Walsh code 0 for pilot channel) and the short PN code used in spreading the target downlink channel are known to the mobile station. Hence, the only unknown to the mobile is the starting phase (PNO) and the received chip timing. Hence, the objective of pilot acquisition is to estimate the unknown parameters.

Figure 6.8 illustrates the process of pilot acquisition. The mobile station has a local copy of the same short PN sequence and correlates with the received signal. Since the unknown parameter at the mobile is the timing offset and the mobile has no prior information about the timing offset, the mobile has to perform exhaustive search over all the possible offsets. When the current local PN sequence has a non-zero offset with the received PN phase, the result of the correlation will be quite small. However, when the local PN sequence is *in-phase* with the received PN sequence, the correlation value will be large. Hence, the estimated offset (or PN phase) is given by the delay position at which a large correlation value is observed. Once the mobile state estimated the correct PN phase offset, it can use the same PN phase value to despread the other downlink channels (with the correct Walsh codes) because all the downlink channels have the same PN phase offset. Note that due to multipath and signals from adjacent base stations, there might be more than one correlation peaks. The value of the correlation peak is equal to the received path energy from the corresponding base station. Since the PNO between base stations are separated by at least 64 chips, the peaks from one base station (due to multipath) will be clustered

together relative to the peaks from different base stations. The mobile station will pick the peaks with the highest received energy and therefore, camping to the strongest base station.

In any case, if the mobile fails to acquire to any pilot channel in any CDMA carrier within 15 sec, it will return to the system determination substate to try to acquire the pilot channel on a different CDMA carrier.

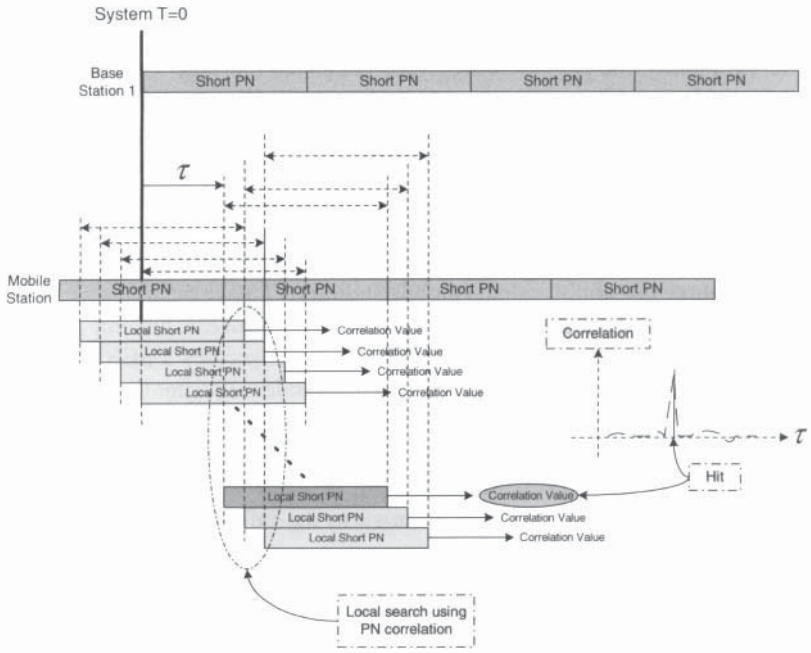
Sync Channel Acquisition Substate: After the mobile has acquired to the pilot channel of a base station, the mobile station attempts to decode for the SYNC channel using Walsh code 32 and the PN code offset (obtained from the pilot acquisition search) to obtain the system timing information. For instance, while the mobile has searched on the appropriate local PN phase offset to acquire the pilot channel, the mobile station has no idea on where is the system $T = 0$ reference. By decoding the SYNC channel message, the mobile station obtains the current PNO of the base station. Hence, the mobile station can align the local $T = 0$ reference with the system $T = 0$ reference. Furthermore, in the SYNC message, the mobile station obtains the operator and system ID (SID and NID) and display the operator logo or roaming status. The mobile also synchronizes the local long code generator with the long code state of the base station.

Timing Change Substate: After the mobile station obtains the timing related information from the SYNC message, it will update the local timing and align with the system timing accordingly. From this point onwards, the mobile and the base stations in the IS95 system have a common $T = 0$ reference.

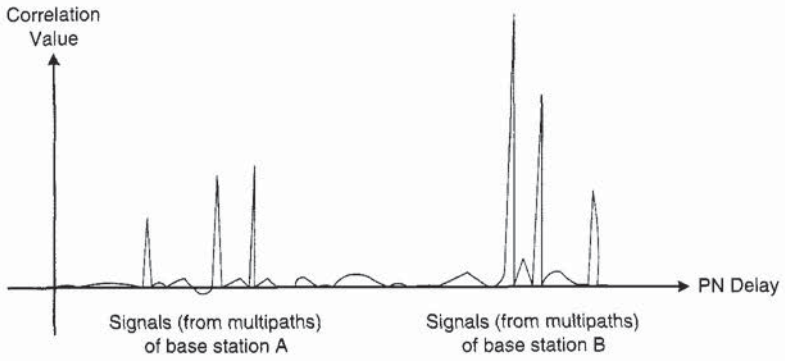
After the mobile station has completed the initialization process, the mobile station enters the idle state and is ready to originate or receive calls.

6.4.2 Idle State

During the idle state, the mobile station periodically monitors the messages in the paging channel as well as the pilot strength of the home cell and the neighboring cells. In particular, the mobile station monitors the paging channel for the *System Parameter Message*, the *Access Parameter Message*, the *Neighbor List Message*, the *CDMA Channel List Message* as well as the *Page Message*. All the messages (except the page message) are broadcast to all mobiles and contain important system parameters for the mobile to participate in call processing in the base station. For example, the system parameter message contains the search window for pilot, power control parameters as well as pilot thresholds in soft handover. The access parameter message contains the important parameters related to the random access procedure (such as the number of access channels, initial access power, access channel timing information and authentication parameters). The neighbor list message contains the



(a) pilot acquisition process



(b) correlation

Figure 6.8 Illustration of the pilot acquisition process in IS95 systems.

list of PNOs in the neighbor base stations and this is very important to speed up pilot strength measurement of neighbor cells during soft handoff. The CDMA channel list message contains a list of active 1.5MHz CDMA carriers in the current cell. These parameters will be elaborated in later sections when we discuss the corresponding procedures. Finally, the paging message is a dedicated message to a mobile user. It is used to notify a mobile station about call request (mobile terminated call).

When a mobile station is in idle state, it will continuously monitor the pilot strength from neighbor cells and perform idle handoff to the new (and stronger) base station when the home cell signal becomes too weak. Note that while IS95 supports soft handover, the idle handover is not soft. The handover decision during idle state is entirely made and initiated by the mobile and it is transparent to the infrastructure.

There are three possibilities that a mobile station will leave the idle state and enter the access state. They are namely, the *call origination process*, the *call termination process* and the *location update process*. When a user presses a SEND button, the call origination process will be triggered and the mobile will enter the access state to communicate this request to the infrastructure. When there is a mobile terminated call, the concerned base station will page the target mobile via the paging channel. The mobile station upon receiving the paging message will enter the access state and communicate the *page response message* to the infrastructure. Finally, when a mobile station moves to a new *location area* (a new cell with new location area identity (LAI)), the mobile station enters into the access state to send the location update message to the infrastructure.

6.4.3 Access State

In the system access state, the mobile station exchanges signaling message with the base station in response to messages requiring acknowledge (page message), to register with the infrastructure (location update message), or call origination. Suppose the user triggers a mobile originated call request by pressing the SEND button in the mobile phone. The mobile station attempts to send the "Call Origination Request" message to the infrastructure. However, at this stage, the only uplink channel available to the mobile station is the random access channel (there is no dedicated channel assigned to the mobile yet). Hence, the mobile station will leave the idle state and move to the channel access state, trying to utilize the random access channel (uplink) to convey the call origination request message to the infrastructure. In other words, a random access procedure is initiated by the mobile to send the call origination message. Figure 6.9 illustrates the random access procedure in IS95.

The entire random access attempts are grouped into *random access probe sequence*. Within each random access probe sequence, there are multiple *random access probes*. During the first random access attempt (the first random access probe in the first random access probe sequence), the mobile station transmit the random access message to the base station via the random access channel (spread by a long code assigned for access channel) at a transmit power given by *Initial Power*. The initial power is

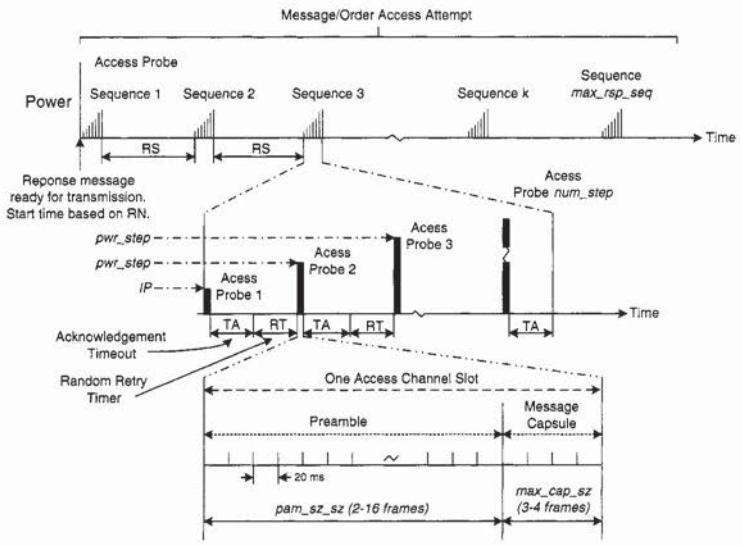


Figure 6.9 Illustration of the random access procedure in IS95 systems.

computed based on the following equation

$$P_i = nom_pwr + init_pwr - 76 - RSSI$$

where $RSSI$ denotes the received signal strength indicator of the downlink signals from the base station and $nom_pwr, init_pwr$ are access parameters defined in the access parameter message. If the mobile station does not receive an acknowledgement from the base station after a timeout period TA (which is also defined in the access parameter message), the mobile station will retransmit the access packet after a random waiting time RT. In the second attempt, the mobile station will transmit at a slight higher power defined by pwr_step (which is another access parameter defined in the access parameter message). The process repeats until the mobile station receives the acknowledgement or the number of access attempts exceed num_step (another access parameter). In the former case, the random access is successful and the mobile enters the traffic conversation state. In the latter case, the mobile station will backoff with a random time RS and starts another access probe sequence. The process repeats until the number of access probe sequence exceeds max_resp_seq (another access parameter message). If so, the mobile station will move to system determination state and acquire the pilot channel on another CDMA carrier (channel) once more.

Observe that there are two possible reasons that the base station does not send back the acknowledgement to the mobile station. The first reason is because the mobile access probe is not transmitted at a sufficiently high power. Hence, the random access probes has an increasing power between repeated attempts. The setting of

initial power is a tradeoff between the capacity loss due to random access[1] and the access delay because it takes a while for the transmit power of access probes to ramp up. The second reason that the base station may miss the access probe is *collision*. For instance, the random access channel is a common uplink channel shared by all the mobiles in the system. Hence, there are some cases when more than one mobile stations attempt to send access probes at similar time. In this case, the access probes may have been transmitted at sufficiently high power but due to collision, the access attempts between mobiles collide and nothing is received at the base station due to mutual interference. Hence, a random backoff timer (RT) between access probes and a random backoff timer (RS) between access probe sequence are introduced to randomize the chance of access collisions in the system. This also explains why there is a SEND button in mobile phones but not in fixed line telephone. In fixed line telephone, the local loop (telephone wire) is a dedicated channel not being shared with other fixed line telephones. However, in cellular systems such as IS95, the uplink access channel is a common channel shared by all the mobiles in the base station. Without a SEND button, the *channel occupancy time* of each access attempt (containing origination request with called telephone number) will be quite long and this increases the chance of collisions between access attempts of mobiles.

6.4.4 Traffic Conversation State

After the mobile station successfully transmit the signaling message (call origination or page response) to the base station via the uplink random access channel, the mobile will enter the traffic conversation state and try to setup dedicated traffic channel (both uplink and downlink)[2]. Figure 6.10 illustrates the three substates in the traffic conversation state.

Traffic Channel Initialization Substate: In this substate, the base station setup a downlink dedicated traffic channel (by assigning an unused Walsh code) and sends null frames at the eighth rate. The base station then notifies the mobile of the assigned Walsh code using a *Channel Assignment Message*. Note that the channel assignment message is sent to the mobile over the paging channel because at this time, there is no dedicated channels established between the base station and mobile station. The mobile station has to verify that it can properly demodulate and decode the downlink frames before starting transmission on the reverse traffic channel. In IS95, the mobile station has to receive two consecutive good frames (passed CRC check) on the dedicated traffic channel first before it can declare the downlink channel is of sufficient quality. The

[1] The background interference floor due to random access will be increased if mobiles transmit the access probe at higher than the required power.

[2] Note that if the mobile station enters the access state due to location update (or registration), it will return to idle state after successfully transmitted the location update message to the base station in the access state.

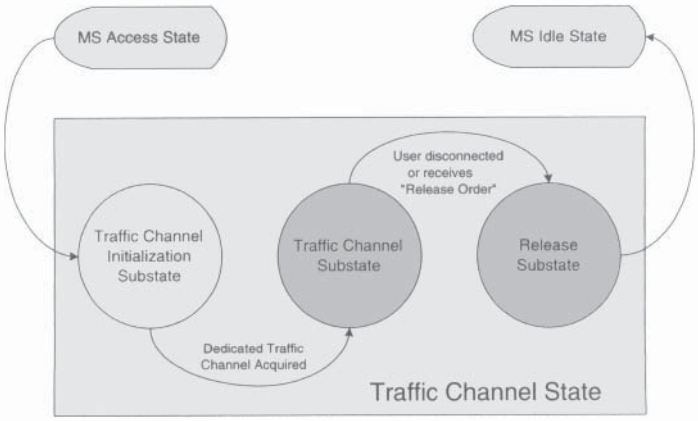


Figure 6.10 Mobile station substates in traffic channel state of IS95 systems.

mobile station will abort the call and re-enter the idle state if no two consecutive good frames are received within 1 second. Say the downlink traffic channel is good, the mobile station can start transmission on the uplink traffic channel. At this point, the mobile station will transmit uplink traffic channel preamble (with data part gated off). Note that the mobile station does not need inform the base station of the spreading code used to spread the uplink traffic channel because the uplink spreading code (long code) is derived from the ESN and IMSI of the mobile and the base station can derive the same long code from the ESN and IMSI of the mobile (which were communicated to the base station via the Call Origination Request Message or Page Response Message). Meanwhile, the base station attempts to acquire the reverse traffic channel preamble and send an acknowledgement message to the mobile station over the downlink traffic channel. This indicates to the mobile that it can start transmitting null data frames on the uplink traffic channel. At this point, both the uplink and downlink traffic channels have been setup properly and the mobile station enters the traffic channel substate.

Traffic Channel Substate: In this substate, the mobile station and the base station transmits signaling and data traffic on the forward and reverse dedicated traffic channel setup during the traffic channel initialization substate. Before the mobile and base station can exchange regular traffic frames, they have to negotiate the service options such as multiplex option, radio configuration and data rates. After the negotiation is completed, speech frames are exchanged between the mobile and the base station over the established dedicated traffic channel continuously once every 20ms. During this state, signaling messages are exchanged between the mobile and base station over the dedicated traffic channel using *dim-and-burst* or *blank and burst* signaling. Dim-and-burst

signaling refers to carrying the signaling message in a 20ms traffic frame by gating off part of the speech data. Blank-and-burst signaling refers to carrying the signaling message in a 20ms traffic frame by gating off all of the speech data.

During the traffic channel substate, both the mobile station and base station perform supervision of the forward and reverse traffic channels, respectively, to detect loss of connection due to radio link failure. A radio link failure may occur if the mobile moves outside the coverage area of the access network. On the mobile side, the mobile station must disable its transmitter if it receives 12 consecutive traffic channel bad frames (CRC failed). The mobile station re-enables its transmitter only when receiving two consecutive good frames. In addition to disabling its transmission, the mobile station must decide when to release the connection based on *fade timer*. The timer is enabled when the mobile station first enables its transmitter upon entering the traffic channel state. The fade timer is reset for 5 seconds whenever the mobile station receives two consecutive good frames on the traffic channel. If the timer expires, the mobile station declares a loss of traffic channel and release the connection.

On the base station side, supervision is implementation dependent and is typically performed by monitoring the quality of the reverse traffic channel after frame selection (at the base station controller). The base station may also use a fade timer. For example, when the reverse traffic channel frames are in error consecutively for a period equal to the base station's fade timer, the base station declares loss of traffic channel and releases the connection. Usually, the fade timer at the base station is set to a value larger than that at the mobile station.

Upon the user disconnects the call or the mobile station receives *Release Order* from the base station, the mobile station enters the release substate and releases the dedicated traffic channel.

Release Substate: When a mobile station enters into the release substate, all the dedicated traffic channels are released and the radio resource is returned to the system. The mobile station moves to the idle state.

6.4.5 Example: Circuit Switched Connection for Voice

Figure 6.11 illustrates an example of call setup processing for mobile originated voice call. Steps (1) to (6) correspond to the procedures described in the *traffic channel initialization substate*. Steps (7) and (8) correspond to the procedures described in the *traffic channel state* where user is engaged in speech conversation. Step (9) corresponds to the *release substate* where the mobile receives a *release order* from the network and release the dedicated traffic channel. Note that the messages such as Call Origination Request and Release Order are terminated and generated from the core network. For mobile terminated calls, the procedures are similar except that the Call Origination Request message is replaced with Page Response message.

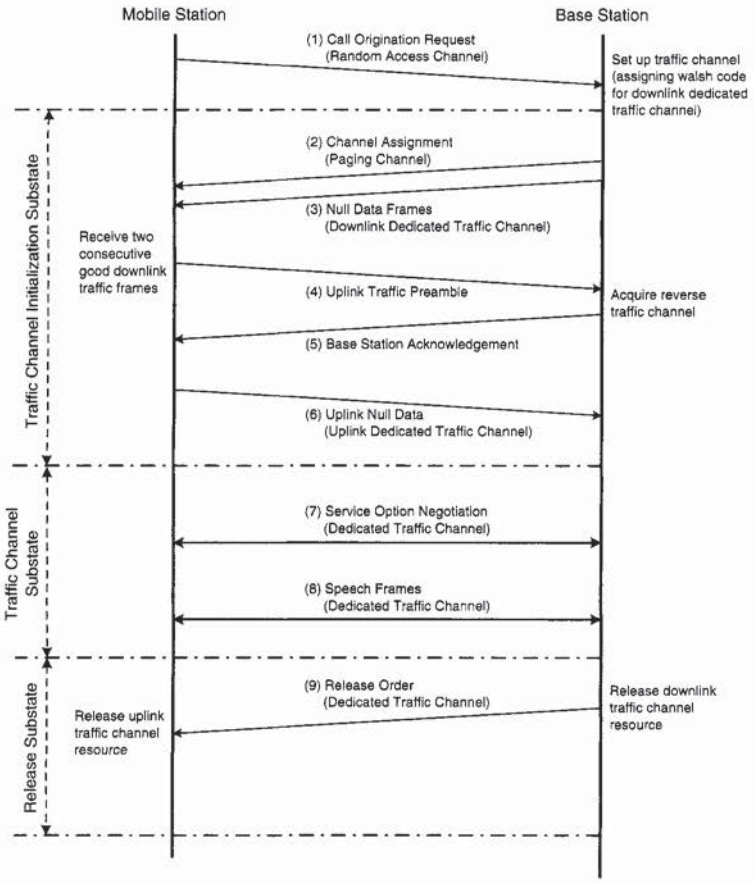


Figure 6.11 Call setup procedure for mobile originated call. In this example, the mobile initiated the call request and the network initiated the call release.

Figure 6.12 illustrates the end-to-end block diagram showing all the entities involved in the speech path for a mobile to fixed line call example. The voice signal is digitalized and compressed at the mobile station and the compressed speech frames are delivered to the base station over the air. The base station acts as a relay which delivers the compressed speech frames to the MSC where it terminates with a corresponding speech decoder. At the output of the speech decoder, the speech frames becomes standard PCM 64kbps speech data and interfaces directly to the PSTN via the gateway MSC. The speech decoder is usually located at the MSC because this could save the bandwidth occupied by a voice channel in the fixed line transmission systems between the base station and the MSC.

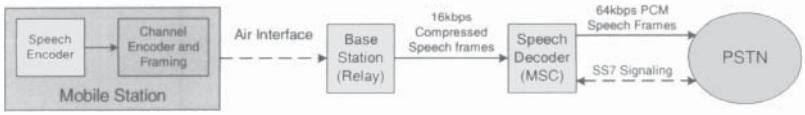


Figure 6.12 Illustration of all the network entities involved in the speech path for a mobile to fixed line call.

6.4.6 Example: Circuit Switched Connection for Data

The call setup procedure for circuit-switched data connection is similar to the procedure illustrated in Figure 6.11 except that the mobile has to indicate to the infrastructure that the circuit switched connection is for data service. This is because the infrastructure has to prepare different entity to terminate the data call. Figure 6.13 illustrates the end-to-end block diagram showing all the entities involved in the digital data path for a mobile to fixed line call example. The digital data (e.g., a file) from the laptop is transferred to the mobile via the infrared interface. The digital data is error protected and fit into the physical frames, which are delivered over the air to the base station with a bit rate of 9.6kbps or 14.4kbps. The base station acts as a relay which delivers the digital data to the InterWorking Function (IWF) (located at MSC). The IWF is a pool of modem gateways which interface the IS95 digital data with common modem standards (such as V.90). The V.90 side of the IWF handshakes directly with the V.90 modems in the ISP via the PSTN to establish the physical connection. Hence, the IWF can be regarded as a layer 1 gateway between the IS95 data frames and the V.90 modem standard. After the physical layers between the mobile and the IWF and between the IWF and the ISP modem are established, the PPP layer will handshake with the PPP layer in the ISP to activate the IP address in the mobile. Since then, the IP address in the mobile will be visible to the external internet and the applications (such as TCP) can work in the regular way as in the fixed line situation.

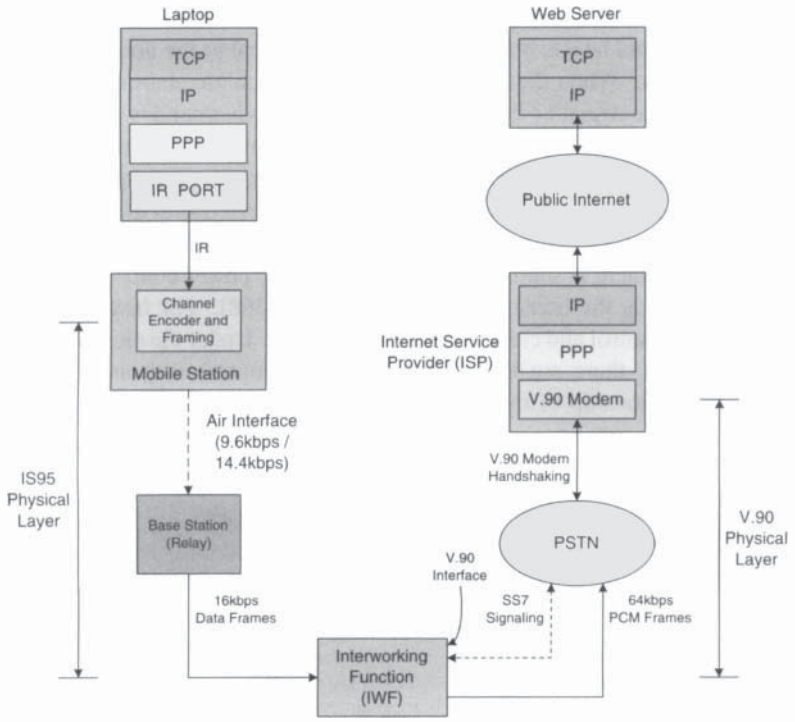
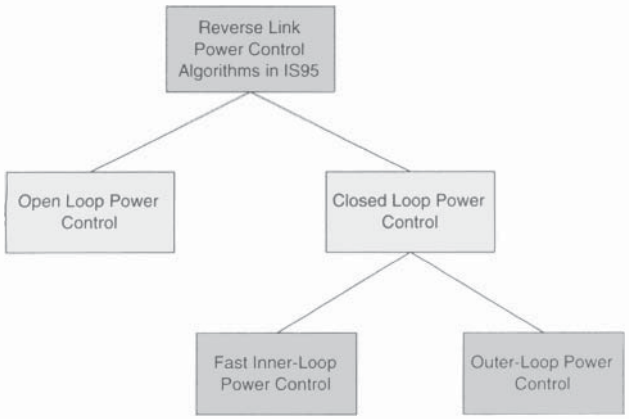


Figure 6.13 *Illustration of all the network entities involved in the circuit switched data path for a mobile to fixed line data call.*

6.5 POWER CONTROL

In this section, we shall focus on the power control in IS95 systems. In GSM or FDMA systems, power control is used primarily to save battery power of mobiles as well as reducing the interference between co-channel cells in the systems. The speed of the power control is therefore relatively slow because the channels for intracell users are orthogonal. However, in IS95 systems, the requirement for power control[131, 315] is more than saving battery power or reducing intercell interference between co-channel cells but rather to control the interference between intracell users as well. As we have explained in Chapter 4, the capacity of random CDMA systems are limited by the interference floor and hence, power control is very crucial to the normal functioning of CDMA systems. When the transmit power of the mobile station is too low, the voice quality will be degraded due to high FER. On the other hand, when the transmit power of the mobile is too high, there will be excessive interference to the other users (because the channels in R-CDMA are not orthogonal). Hence, the mobile station should be controlled to transmit only the minimum required power as any extra power will contribute to the rise of interference floor and consequently, reducing the system capacity. In IS95, there is sophisticated mechanism for power control in the reverse link (i.e., controlling the transmit power of mobiles)[368]. For instance, there are open loop power control and closed loop power control. Under the closed loop power control framework, there are fast inner-loop power control and outer-loop power control. The relationship of the power control algorithms in IS95 is illustrated in Figure 6.14



The open loop reverse link power control refers to the automatic adjustment of mobile station transmit power without any *feedback* from the base station. On the other hand, the closed loop power control mechanism refers to the adjustment of

mobile station transmit power by explicit feedback of the base station. The open loop power control and the closed loop power control algorithms are designed to achieve an ideal operating condition of *equal received power at the base station from all the mobile stations*[86]. Obviously, this ideal operating condition can never be achieved perfectly. The errors in power control will be translated into *residual fluctuations* in the base station received power. The better the power control algorithms are, the smaller the dynamic range of the residual fluctuations of the base station received power are as illustrated in Figure 6.15 These power control mechanisms are elaborated in detail in the following subsections.

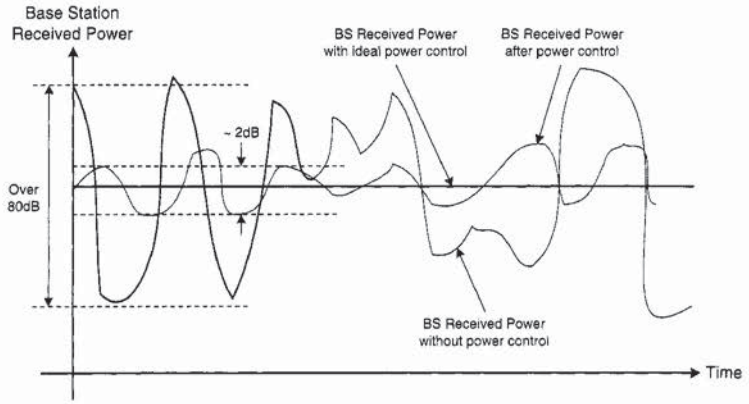


Figure 6.15 Performance of power control algorithms in CDMA systems.

6.5.1 Open Loop Power Control

The received power at the base station from the mobile stations fluctuates with time based on three factors, namely the path loss, shadowing and the microscopic fading. These three effects together account for over 80dB fluctuation of received signal strength from the mobiles as illustrated in Figure 6.15 In other words, the transmit power of the mobile station has to be adjusted over a dynamic range of more than 80dB in a relatively short time. This poses a great challenge to the closed loop power control design. Hence, in IS95 systems, there is an autonomous open-loop power control algorithm at the mobile station (without any explicit feedback from the base station) designed to reduce the dynamic range of the base station received power fluctuations from mobiles. Hence, after the open loop power control, the residual dynamic range requirement of the closed loop power control is substantially reduced. Note that the objective of the open loop power control is not on the accurate control of mobile transmit power but simply to reduce the dynamic range of the base station

received power fluctuations so that the closed loop power control algorithms will have an easier job to fine tune the transmit power of the mobiles to the right target level.

We shall discuss how the open loop power control can reduce the dynamic range of the base station received power fluctuation without explicit feedbacks from the base station. The 80dB dynamic range is contributed by the path loss, shadowing and the microscopic fading. While the microscopic fading (due to multipath) may not be symmetrical between the uplink and the downlink, the first two factors (path loss and shadowing) are basically symmetrical between the uplink and downlink directions. In other words, when the path loss or shadowing from the base station to the mobile is large, it is very likely that the corresponding path loss or shadowing from the mobile station to the base station is large as well. Hence, the mobile station can have a first order estimate of the path loss and shadowing by measuring the received signal power from the base station. For instance, if the received power from the base station is large, it is likely that the mobile is close to the base station (path loss and shadowing are small) and the mobile station should transmit with a smaller power. On the other hand, if the received power from the base station is small, the mobile is likely to be far away from the base station and the mobile station should transmit at a larger power. Note that this logic works only if the transmit power from the base station is constant. In IS95, the pilot channel (Walsh channel 0) is always transmitted at constant power and could serve as a good reference for open loop power control. However, for implementation simplicity, the IS95 utilizes the total received signal strength (RSSI) from the base station for open loop power control. The transmit power of the mobile station is given by:

$$Pwr_{ms-tx}(dBm) = -76 - \overline{RSSI}(dBm)$$

The required response time for open loop power control is from 100ms to 300ms. Figure 6.16 illustrates the variation of mobile transmit power due to open loop power control.

Obviously, since the microscopic fading component is not symmetrical between the uplink and downlink, the open loop power control can never be perfect in adjusting the transmit power of the mobile. The fine tuning of the mobile station transmit power is the task for the closed loop power control.

6.5.2 Inner Loop Power Control

The inner loop power control algorithm in IS95 is the core of the power control algorithm. Basically, the base station measures the instantaneous received SIR and generates a 1-bit power control command (PCB) once per 1.25ms to control the mobile station transmit power. Hence, there are 800 PCBs per second transmitted from the base station to the mobile station and the speed of the inner loop power control is 800Hz. Figure 6.17 illustrates the block diagram of the inner loop power control in IS95 systems. There are two inputs, namely the *measured SIR*[86] and the *target SIR*, and one output, namely the *power control command*, in the inner loop power control

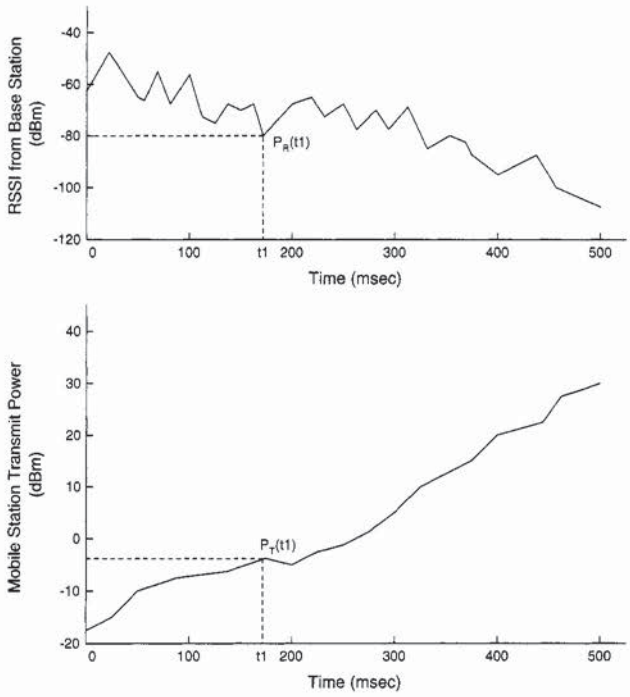


Figure 6.16 Example of open loop power control in IS95.

algorithm[359]. The base station measures the instantaneous received SIR from the uplink frames and compares with the target SIR. If the instantaneous SIR exceeds the target SIR, the mobile station is probably transmitting at excessive power and hence, a negative power control command (bit = 0) will be generated and the mobile station is asked to decrease the transmit power by a power control step size Δ. On the other hand, if the instantaneous SIR is less than the target SIR, a positive power control command (bit = 1) will be generated and the mobile station is asked to increase the transmit power by a power control step size Δ. Note that the 1-bit power control commands are sent to the mobile station in the downlink frame. Since a frame lasts for 20ms, there are 16 power control commands embedded in each downlink frame. To allow fast decoding of the power control commands in the mobiles, the power control commands are not protected by error correction codes. Furthermore, there is a dedicated power control loop per dedicated traffic channel.

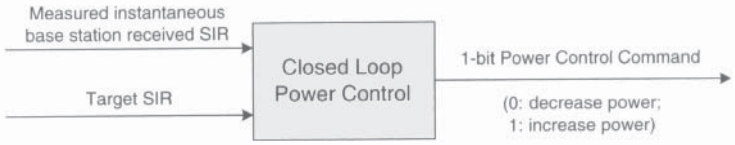


Figure 6.17 Block diagram of closed-loop power control in IS95.

Figure 6.18 illustrates an example of mobile transmit power adjustment as a result of inner loop power control in IS95 systems. During the first 20ms, the instantaneous base station received SIR exceeds the target SIR and hence, the mobile station transmit power is reduced by the power control step size Δ. When the instantaneous base station received SIR come close to the target SIR, the received SIR oscillates up and down about the target SIR and this is the *steady state* of the inner loop power control [189]. Note that due to the 1-bit power control action, the received SIR in steady state will not be a constant. This is a well-known characteristic of the *bang-bang control loop*. The target SIR may be adjusted to a new value once every 20ms frame. The power control commands are generated accordingly by comparing the instantaneous received SIR with the new target SIR. Note that choice of the power control step size Δ involves a tradeoff between the *steady state error* and the *response time*. For large Δ, the response time of the inner loop will be faster (i.e., the instantaneous received SIR will be brought to the steady state about the target SIR faster). However, since the received SIR fluctuates up and down about the target SIR in the steady state, large Δ will result in large steady state error of the received SIR, which is undesirable. In IS95, Δ is chosen to be 0.25dB–0.5dB. Furthermore, for a fixed Δ, a longer response time will be needed to bring the received SIR to the target level if the initial SIR offset from the target is large. This helps to justify the needs of open loop power control to reduce the dynamic range of the SIR offset.

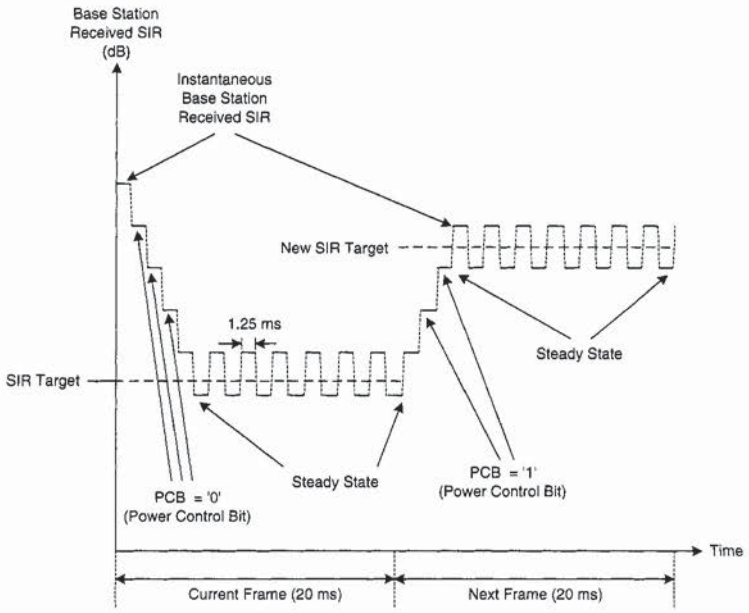


Figure 6.18 Example of mobile transmit power adjustment of inner-loop power control in IS95.

Note that after the inner loop power control, the base station received SIR will not be exactly at the target SIR value but will fluctuate with some residual variance as illustrated in Figure 6.15 The residual variance is contributed by the power control step size Δ as well as the delay in the power control commands and the errors in the power control commands. In particular, the residual variance of the received SIR at the base station increases with the mobile speed because the inner loop power control fails to track the microscopic fast fading at high mobility. To ensure small delay in the generation of power control commands, the inner loop power control algorithm resides at the base station.

6.5.3 Outer Loop Power Control

The inner loop power control algorithm is responsible to bring the received SIR at the base station to the target SIR level. However, the received SIR is not something that a user concerns. In fact, users only concern with the voice quality which is related directly to the frame error rate (FER). Hence, from the user's QoS perspective, the FER is the ultimate target rather than the SIR.

The translation from SIR to FER is usually obtained from the FER versus SIR curve of the physical layer. Given a target FER (e.g., 10^{-1}), we can read the target SIR from the physical layer curve. However, the mapping from the target FER to the target SIR is not always straight-forward and is time varying as well. For instance, different mobile speed will result in different coherence time and this will affect the FER curves of the physical layer. Hence, for the same target FER, the required target SIR is not a constant and changes with the mobility and the channel environment as the mobile moves around. In IS95, the outer-loop power control is designed to ensure an automatic adjustment of the target SIR in order to achieve the overall target FER. Figure 6.19 illustrates the conceptual block diagram of the outer-loop power control in IS95. The inputs and the output of the outer-loop algorithm are *target FER*, *instantaneous measured FER* and the *target SIR* respectively. The target SIR is updated once per 20ms. Basically, the target SIR is increased if the current FER exceeds the target FER and vice versa. However, the difficulty here lies in the ability to measure the FER with small delay. For example, to measure a FER of 10^{-2}, we need at least several hundred frames. Hence, brute force measurement of FER will incur unacceptable delay and will be useless. As a matter of fact, different vendors will have different implementations of the outer loop power control algorithm to get around this problem[185]. We shall elaborate one simple algorithm below.

Figure 6.20 illustrates an example of outerloop power control algorithm in IS95 systems. At the base station, the SIR target setpoint is reduced by a small amount ($down_{frr}$) for each consecutive frame until uplink frame error is detected. Note that frame error can be detected by CRC check in each frame. When uplink frame error is detected, the target SIR is increased by a large amount (up_{frr}). Within each 20ms where the target SIR setpoint remains constant, the inner loop power control will bring the actual instantaneous received SIR to the target SIR setpoint based on the

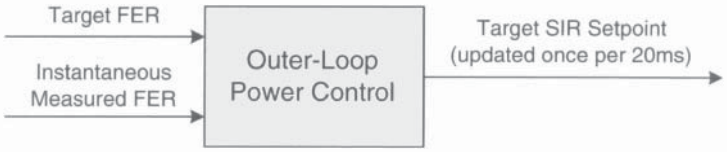


Figure 6.19 Conceptual block diagram of outer-loop power control in IS95.

1-bit power control commands. During steady state of the outer loop power control, there will be 2 uplink frame errors per cycle of $down_{frr}/up_{frr}$ (the factor of 2 is to account for the 1 frame latency in the outer loop). Hence, we have

$$FER_{\text{target}} = 2\frac{up_{frr}}{down_{frr}}$$

Since the response of the outerloop is just 50Hz and to incorporate the potential macrodiversity of soft-handover into the power control (will be elaborated in the next section), the outer loop power control algorithm resides in the base station controller.

6.5.4 Example

Figure 6.21 illustrates a summary of the reverse link power control in IS95, integrating the open loop and closed loop power control actions into a single example. The mobile station is initially in the clear with the base station and hence, receiving large RSSI from the base station. The corresponding mobile transmit power is small due to the action of open loop power control. At the same time, the inner loop power control reaches the steady state where the base station received SIR fluctuates up and down about the target SIR. When the mobile moves into a shadow, the received RSSI from the base station drops suddenly and this brings an immediate increase in the mobile transmit power. At the same time, the base station instantaneous received SIR drops below the target and the inner loop power control generates consecutive "+" commands to increase the mobile station transmit power. Furthermore, due to the frame error (when the mobile enters into the shadow), the target SIR setpoint is increased by up_{frr} due to the action of outer loop power control. As the mobile transmit power gradually increases due to the actions of inner loop power control, the base station instantaneous received SIR reaches the new target SIR level shortly and maintains at the steady state. Later on, when the mobile station leaves the shadow and enter into the clear again, the base station instantaneous received SIR suddenly exceeds the target SIR and the mobile station transmit power is decreased gradually through the actions of inner loop power control as well as the open loop power control. Shortly, the instantaneous base station received SIR reaches the steady state of the new target SIR again.

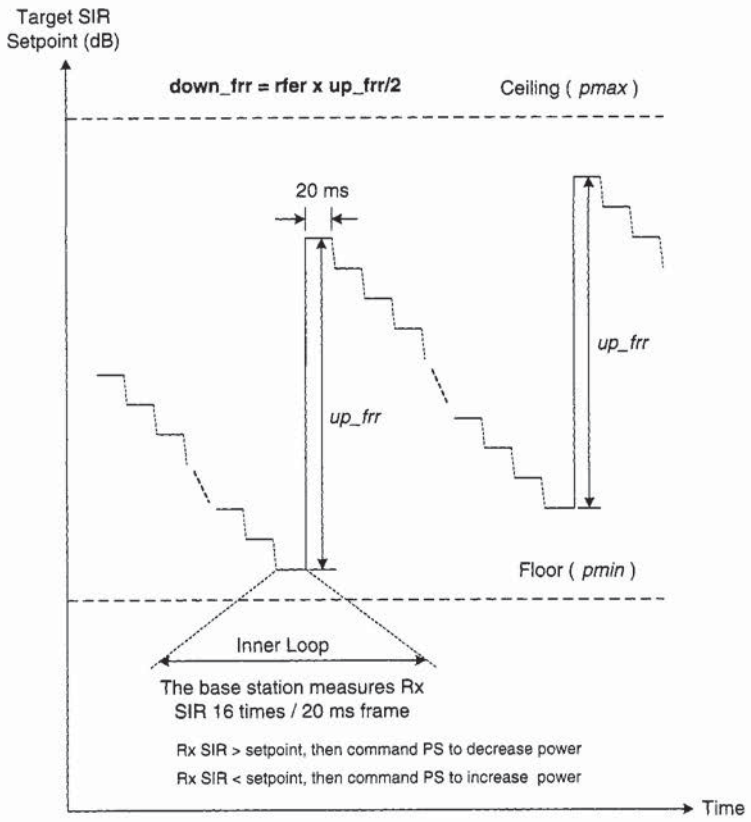


Figure 6.20 Example of outer-loop power control algorithm in IS95.

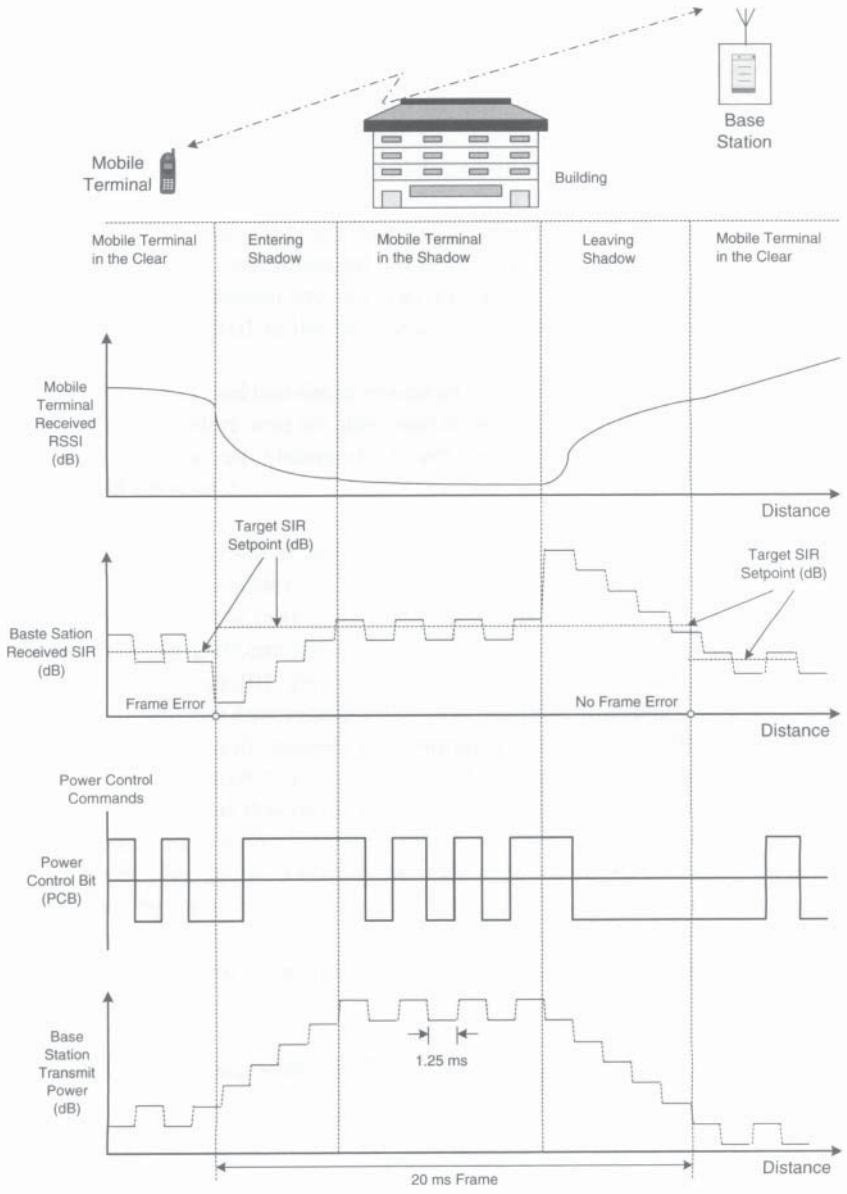


Figure 6.21 Integrated example of open loop and closed loop power control algorithms in IS95.

6.6 SOFT HANDOVER

In this section, we shall discuss the soft handover procedure in IS95. Handover is a generic process of transferring calls when a mobile moves out of the coverage of one cell and enters the coverage of another cell in the cellular system. In a handover process, the system must determines (i) the need for handover, (ii) the new cell that should continue to serve the call. To facilitate such knowledge, the mobile station as well as the cellular infrastructure have to coordinate the exchange of information. In IS95 systems, the mobile station assists the infrastructure by providing measurement information about the surrounding cells to the infrastructure. This refers to *mobile assisted handover*. In fact, mobile station plays a very important role to provide an accurate and timely picture of handover needs as well as handover direction to the cellular infrastructure.

After the infrastructure acquires the handover needs and handover direction information, the infrastructure have to coordinate with the new cells and mobile station to *execute* the handover decision. Note that while mobile station offers timely measurement data to the infrastructure, the handover decision is done at the infrastructure (base station controller). There are in general two common ways of handover execution, namely the *hard handover* and the *soft handover*. In hard handover, the existing connection (between the mobile and the home cell) is released before the new connection (between the mobile and the new cell) is established. In other words, we have a *break-before-make* situation. Examples of hard handover includes GSM and DAMPS systems. On the other hand, in soft-handover[110], the mobile station may establish the new connection (with the new cell) together with the existing connection (with the home cell) simultaneously during the transition. In other words, we have a *make-before-break* situation and therefore, a more reliable handover is possible with soft handover. Existing systems that can benefit from soft handover include IS95, CDMA2000 as well as UMTS [340]. Note that soft handover in CDMA systems refers to the dedicated traffic channel when the mobile is in conversational state. On the other hand, when the mobile is in idle state, no soft-handover will be supported in IS95[3]. In fact, the design of CDMA systems is well-suited for soft-handover and the detailed processing will be elaborated in the following subsections.

6.6.1 Pilot Strength Measurement and Soft Handover Combining

Before we elaborate on the soft handover algorithms in IS95 systems, we shall zoom into the physical layer and discuss how mobile station measures the signal strength of the environment. We shall also elaborate on why soft handover is well-suited for CDMA systems but not for other systems (such as GSM).

[3]Only hard handover will be supported when mobile is in idle state and this refers to *idle handover*. In idle handover, the decision is made entirely at the mobile station without any coordination from the infrastructure.

In IS95 systems, the mobile station has to measure the signal strength of the home cell as well as surrounding cells frequently so that the infrastructure will not miss any opportunities or needs for handover. However, how does the IS95 mobile physically measures the signal strength? Recall that every base station in IS95 system has a common downlink pilot channel (transmitted at constant power) spread by a master PN sequence at an offset PNO. Hence, the pilot strength measurements (measured at the mobile) from the home cell as well as from the surround cells are good indications for handover decision. Hence, it is important to understand how the mobile measures the pilot strength. When the mobile station powers up, it has to synchronize the local PN sequence with the pilot channel of the home cell. The delay position of the peaks in the autocorrelation measurement of the PN sequence at the mobile indicates the delay positions with synchronized PN timing. On the other hand, the magnitude of the peaks indicates the strength of the pilot and is proportional to the pilot strength. At the time of power up, the mobile station has no prior information about which is the home cell and hence, the mobile has to exhaustively search over all the delay offsets to synchronize with the strongest pilot. However, when the mobile is in conversation state, the pilot strength measurement is time critical as timely measurement is crucial to the appropriate execution of handover. To facilitate fast pilot strength measurement of the home cell as well as the surrounding cells, the base station periodically broadcasts a *neighbor list message* in paging channel. In the neighbor list message, there is a list of the PNO of the surrounding cells (with respect to the current home cell). Mobile station, upon decoding the neighbor list message, will have prior information about the target PNO of the pilot channel it is going to measure. For example, if the PNO of a neighbor cell B is 128 chips, the mobile station can shorten the time to measure the pilot strength of cell B by moving the local PN phase at the mobile to a delay position of $128 \pm \Delta$ where Δ is a measurement window to account for the delay uncertainty of the received pilot signal from base station B due to propagation delay and delay spread. Hence, instead of exhaustive search over all the possible PNO space, the mobile only needs to search over a small window Δ about the target PNO (e.g., PNO=128 chips) to measure the pilot strength from base station B. Note that the pilot strength is given by the sum of the peaks of the autocorrelation measurement at the mobile about the target PNO as illustrated in Figure 6.22 Typical delay in pilot strength measurements is of the order of 100ms in IS95 systems.

The next question is why soft handover is well-suited for CDMA systems but not say GSM systems. To understand this, we shall first discuss the general requirement of soft handover on the physical layer processing. During soft-handover, the mobile has to detect information signals from more than one base stations simultaneously. This poses a first requirement on the RF processing at the mobile. For example, in GSM systems, the mobile has to equip with two parallel RF units in order to support simultaneous detection of signals from two cells because the signals from two neighbor cells will be transmitted at different carriers (with different frequencies). On the other hand, for CDMA systems, only one RF unit is needed at the mobile to detect signals from more than one base stations simultaneously because the signals from neighbor

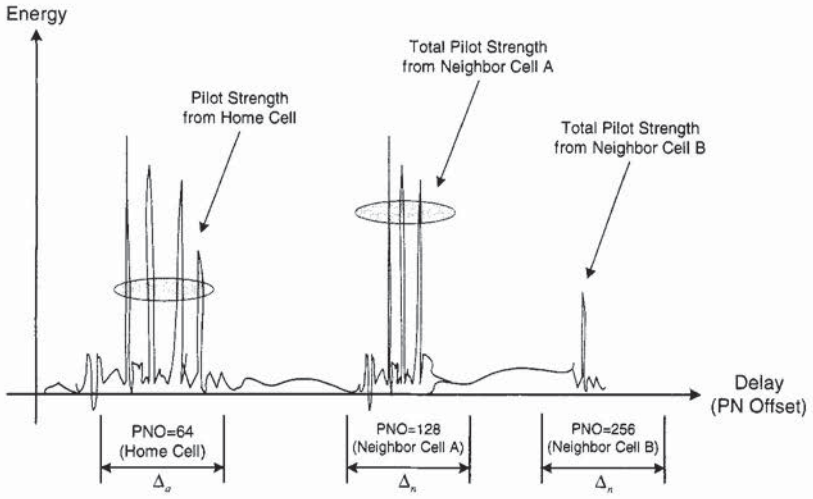


Figure 6.22 An example of pilot strength measurement for a neighbor cell at the mobile.

cells are transmitted using the same carrier frequency (due to the interference suppression capability of CDMA). Hence, CDMA systems post a distinctive advantage in terms of low cost consideration. A second requirement to support soft-handover is the baseband processing involved to combine multiple signals (containing same information) from different cells. For instance, consider a 2-way soft-handover scenario where a mobile is detecting information from cell A and cell B simultaneously. The transmitted signals from cell A and cell B contains the same information and hence, they serve as "independent observations" about the same information at the mobile. However, in order to be able to detect the information as well as exploiting the *macro-diversity* as a result of the two independent observations, the baseband has to align and combine the two observations properly, which is not trivial. For GSM systems, two baseband units will be required to achieve advantage of soft-handover, which is costly. On the other hand, for CDMA systems, the RAKE receiver is well suited for separating and combining these independent observations from cell A and cell B. This is because there is virtually no difference in the processing complexity for the RAKE receiver to track multipath signals from cell A only versus the RAKE receiver to track signals from cell A and cell B simultaneously. For example, if there are four RAKE fingers in the mobile, 5 significant multipaths from cell A and 3 significant multipaths from cell B, two of the RAKE fingers can be assigned to detect 2 strongest multipaths from cell A and 2 strongest multipaths from cell B during soft handover. Hence, due to the complete frequency reuse and the RAKE receiver architecture, there is virtually no increase in mobile complexity to support soft-handover in CDMA systems.

6.6.2 Mobile Soft Handover Processing

In IS95 systems, the soft handover procedures require actions from both the mobile station and infrastructure. For instance, mobile station has to measure the pilot strengths of the home cells as well as the neighbor cells frequently. When mobile spots any strong pilots from the neighbor, it has to report to the infrastructure at once in order not to miss any handover needs. The infrastructure will make a decision about whether to engage in soft handover or not based on the reported pilot strength as well as the target cell loading. Hence, whether the call can be maintained as the mobile station moves depends on how fast the mobile can measure the pilot strength from the neighbors. As we have explained in the above section, certain processing power (searching over the uncertainty window Δ) is needed for mobile to measure the strength of one pilot in the IS95 systems. Hence, the mobile station has to *prioritize* the measurement of pilots. This is elaborated below.

In mobile station, the pilots in the system are divided into the following four sets, namely the *active set*, the *candidate set*, the *neighbor set* and the *remaining set*. They are elaborated as follows.

Active Set: The active set contains all pilots (PNO) of the base stations that are directly engaged in carrying the payload traffic to and from the mobile. In other words, this is the set of all pilot offsets (PNO) corresponding to the active base stations with respect to the mobile. Note that when a mobile station is in conversational state, there must be at least one member in the active set. When there are more than one members in the active set, the mobile station is said to be engaged in *soft handover*. In IS95, the maximum entry of the active set is 6, meaning that there are at most 6 soft handover legs in any call.

Candidate Set: The candidate set contains pilots (PNO) that are not yet in active set but have sufficient signal strength to carry the mobile payload connection (successful detection of the payload frames). These are good candidates to be *promoted* into active sets and they are awaiting the decision from the infrastructure to be added into active set. In IS95, the maximum entry of the candidate set is 5.

Neighbor Set: The neighbor set contains pilots (PNO) that are not in the active set and candidate set but they are pilots from base stations that have potentials to be promoted to candidate set. In IS95, the maximum entry is 20.

Remaining Set: The remaining set contains some lower priority pilots (pilots that are not included in active set, neighbor set or candidate set).

The mobile station shall scan the pilot strength according to the priorities defined in the four pilot sets. Figure 6.23 illustrates a flow chart for the pilot measurement in the mobile station side. The mobile station first measures the signal strength of all pilots in the active set. Next, the mobile station measures the pilot strength in the neighbor

set (trying to find strong pilots). If there is any pilot with strength greater than t_{add}, the pilot is regarded as *strong pilot* and is promoted to the candidate set by the mobile. At the same time, a signaling message called the *Pilot Strength Measurement Message* (PSMM) will be sent to the base station (via blank-and-burst or dim-and-burst over the dedicated traffic channel established). On the other hand, if none of the pilots in the neighbor set exceeds t_{add}, the mobile station then measures the pilot strength of the remaining set to look for strong pilots (pilot strength exceeds t_{add}). Similarly, if any strong pilot(s) is found, the pilot(s) will be promoted to candidate set and the corresponding PSMM message will be sent to the base station. This corresponds to *pilot addition event*. Note that one single PSMM message can contain pilot strength measurement of multiple strong candidate pilots to reduce the signalling overhead.

Next, the pilot strength of all the candidate pilots are compared with the active pilot strength. If any of them exceeds the weakest active pilot by t_{comp}, PSMM message will be sent to the base station so that the system may replace (or swap) the weakest active pilot with a stronger candidate. This refers to *pilot swapping event*. Finally, the mobile station compares the signal strength of all active pilots against a parameter t_{drop}. If there is any active pilot(s) with signal strength below t_{drop}, a timer is started at the mobile station. If the condition (the signal strength of that active pilot below t_{drop}) lasts for longer than t_{tdrop}, the mobile station sends the PSMM message indicating the weak active pilot. This corresponds to the *pilot drop event*.

Different mobile vendors may have different ways of implementing the pilot measurement and prioritization for pilots within a set. Note that a mobile station can decide on its own to move the pilots between the candidate set, neighbor set and remaining set. However, any movement of pilots into or outside the active set requires the approval of the infrastructure and is not decided by the mobile station alone.

6.6.3 Infrastructure Soft Handover Processing

The soft handover processing in the infrastructure resides at the base station controller. Hence, the base station just acts as a relay to deliver the handover signalling message to and from the base station controller and the mobile. At the base station controller, the handoff evaluation starts when the PSMM message from a mobile station is received. The base station controller forms a list of candidate pilots (from the PSMM message) that do not appear in the active list and rank the pilots by signal strength. The base station controller then checks the available resource for soft handover. For example, if the PSMM message contains a candidate pilot B and the resource at base station B is available to support the soft handover request, the base station controller will first inform the base station B to set up a full duplex (uplink and downlink) radio channel as well as a channel in the fixed line transmission between the base station controller and base station B. When the setups in base station B is completed, the base station controller will deliver a *Handover Direction Message* (HDM) to the mobile station via the home cell (e.g., base station A). In the HDM, the pilot (PNO) corresponds to the new active set configuration is delivered. The mobile station upon

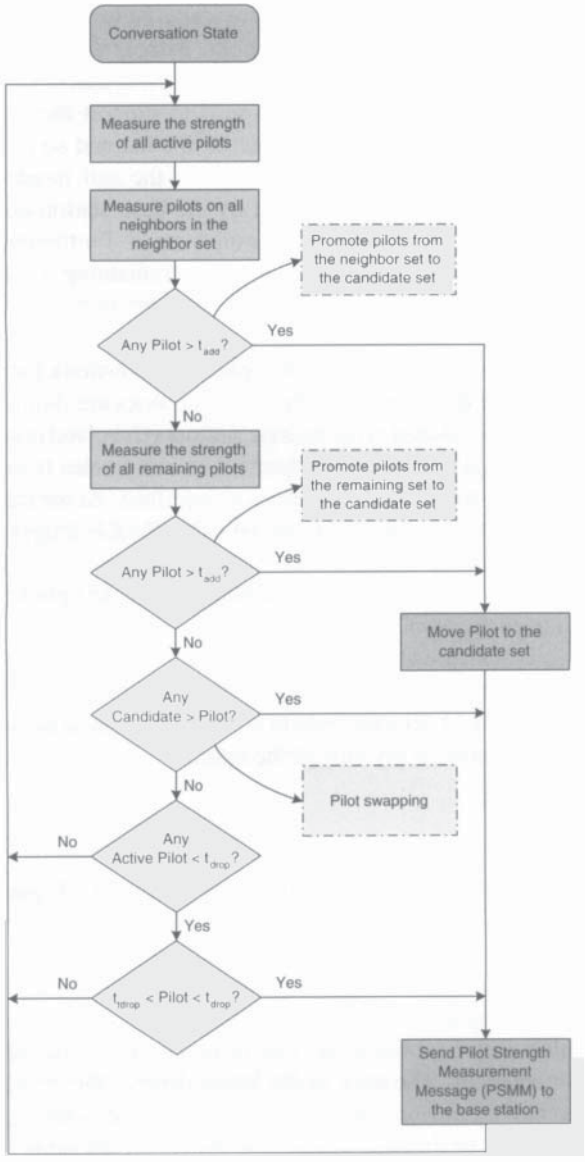


Figure 6.23 A flow chart of pilot measurement processing in the mobile station of IS95 systems.

receiving the HDM should reply with a *Handoff Completion Message* (HCM), which acts as an acknowledgement to complete the soft handoff transaction. Hence, one soft handover transaction starts with PSMM message, followed by HDM and HCM. There are three possible soft handoff transactions, namely the *pilot add transaction*, *pilot drop transaction* and the *pilot swap transaction*. Note that in the above example, if there is no resource available in base station B to support the soft handover, the base station will look at the second candidate in this list and so on. If none of the candidate base stations has the resource to support the soft handover, the PSMM message sent by the mobile will be ignored by the base station controller and the active set configuration of the mobile remains unchanged. Furthermore, note that the concept of active set, candidate set, neighbor set and remaining set applies to mobile station side only. At the base station controller, there are only the active list as well as the candidate list.

Figure 6.24 illustrates the data path of the uplink and downlink for mobile involved in soft handover. In the downlink path, the speech frames are duplicated at the base station controller and distributed to all the base station(s) involved in soft handover (all the base stations in the active set). The identical speech frames from the active base stations are delivered to the mobile over the radio interface. At the mobile, the signals from different active base stations are received using RAKE fingers as illustrated in Figure 6.22 and combined by one of the three diversity combining (selection diversity or equal gain combining or maximal ratio combining). In the uplink path, the speech frames delivered from the mobile are received respectively by the active base stations. The speech frame is detected at each of the base station independently. The detected frames from the active base stations are delivered to the base station controller where one out of N frames will be selected (where N is the number of active base stations). Hence, there is selection diversity only at the uplink.

6.6.4 Example

We shall complete the soft handoff description with an example. Figure 6.25 illustrates a hypothetical trace of pilot strength measurement by a mobile station. In this example, we have two base stations (A and B) in the system. At $t = 0$, the mobile station is in conversation state with pilot A in the active set (i.e., the home cell of the mobile at $t = 0$ is A.). The trace is the pilot strength of pilot from base station B. Hence, at $t = 0$, the mobile's active set has pilot A as member and the mobile's neighbor set has pilot B as the member. The trace in the figure denotes the pilot strength of B as measured by the mobile station. At $t = (1)$, the pilot B exceeds t_{add} and hence, a PSMM message is sent to the base station controller. At the same time, the pilot B is promoted from the neighbor set to the candidate set. At $t = (2)$, the base station controller confirms the addition of pilot B into active set by sending the HDM (with pilot A and pilot B). The mobile station acknowledges the HDM by sending HCM at $t = (3)$, completing the soft handover-pilot addition event. At the same time, pilot B is promoted into the active set (i.e., active set of the mobile consists of pilot A and

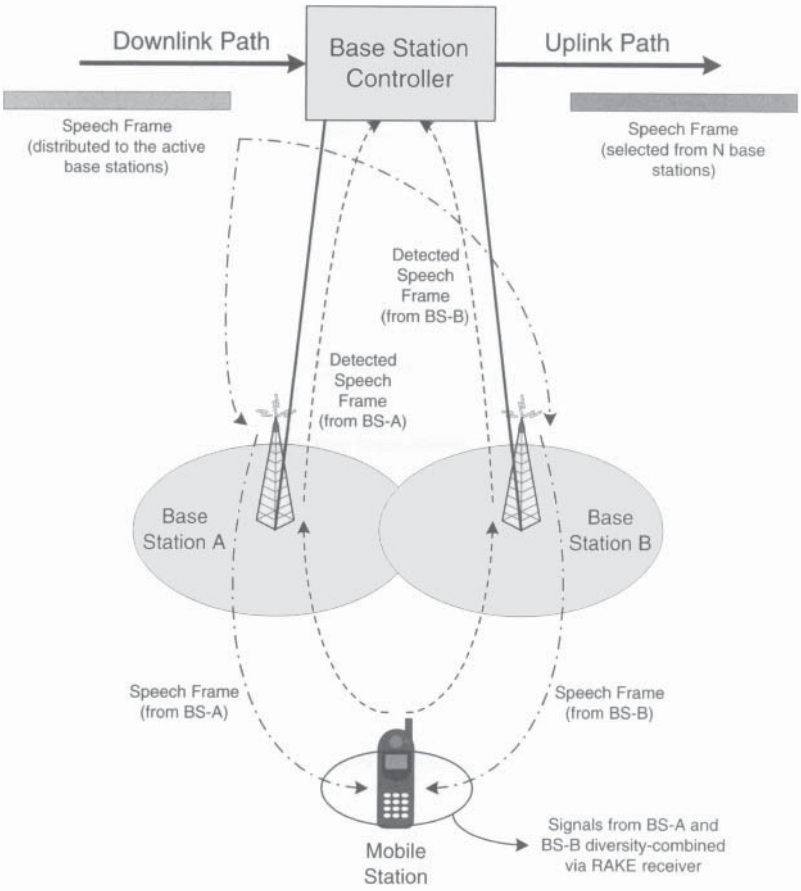


Figure 6.24 Uplink and downlink data path between mobile and base stations engaged in soft handover in IS95 systems.

pilot B). Since then, there are more than one members in the mobile's active set and the mobile is said to be engaged in *two-way soft handover* (from $t = (3)$ to $t = (7)$).

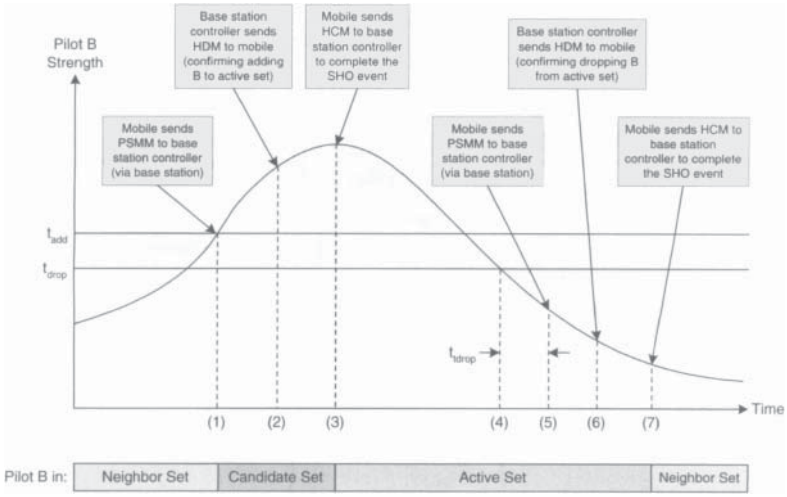


Figure 6.25 A flow chart of pilot measurement processing in the mobile station of IS95 systems.

At $t = (4)$, the pilot strength of B drops below t_{drop} and the mobile starts a timer. At this time, no message is sent to the base station controller and pilot B still remains in the active set. At $t = (5)$, the timer exceeds t_{tdrop} and hence, a PSMM message is sent to the base station controller, suggesting to drop off pilot B in the active set. Similarly, the base station controller sends the HDM at $t = (6)$ (with pilot A only) and the mobile station acknowledges with HCM at $t = (7)$. At the same time, pilot B is demoted from the active set to the neighbor set at the mobile and the active set contains pilot A only. After that, the mobile is not in soft handover anymore because there is only one member in the active set. This corresponds to a pilot drop event. Observe that the system is designed in such a way that pilot dropping is more careful compared with pilot addition. For example, to drop a pilot from an active set, the mobile has to make sure the signal strength is below t_{drop} for longer than t_{tdrop}. This is because the penalty of falsely dropping a pilot can be large (e.g., call drop).

Before we conclude this section, we shall elaborate the system consideration when soft handover and power control are considered together. In fact, power control or soft handover on its own perspective is relatively straight-forward. However, complications occur when we look at these two algorithms together. For instance, consider a 2-way soft handover example where a mobile station is engaged in 2-way soft handover with base station A and B respectively. There will be two closed-loop power control loops running between the base station A, base station B and the mobile. For instance, base station A will send power control commands to control the transmit

power of the mobile based on the measured uplink SIR at A. Similarly, base station B will send power control commands to control the transmit power of the mobile based on the measured uplink SIR at B. Since the two base stations do not coordinate when generating the power control commands, there is a chance when base station A sends a power control command (0) and base station B sends a power control command (1) to the mobile. In IS95, the mobile will reduce the transmit power as long as there is a power control command of 0 from any of the active base stations. This is because after the frame selection at the base station controller, only one good frame is needed. Hence, the mobile station only needs to maintain the transmit power in such a way that at least one base station can detect the uplink frame.

6.7 SUMMARY

In this chapter, we have elaborated the system operation of IS95 systems. For instance, we have elaborated the multi-access configuration in the uplink and downlink of IS95 systems. In the uplink, mobile signals are separated by R-CDMA using a unique long code. In the downlink, intracell users are separated by D-CDMA using orthogonal Walsh codes. On the other hand, intercell users are separated by R-CDMA using short PN sequence (which is unique per base station). We have looked at the call processing procedure to establish a circuit-switched dedicated connection. There are four call processing states in a mobile station, namely the *initialization state*, *the idle state*, *the access state* and the *conversation state*. In the initialization state, the mobile station has to acquire the pilot channel by exhaustive search over the PNO space. In the idle state, the mobile station has to monitor the broadcast messages in various downlink common channels (such as SYNC channel and PAGING channel). When mobile originates a call request or when mobile receives a dedicated paging message or when mobile performs location updating, the mobile enters the access state and using the common uplink access channel, transmit the appropriate signaling messages to the base station based on the random access procedure. In the conversation state, dedicated traffic channels in the uplink and downlink are established to carry the payload.

Next, we have looked at the power control operation in IS95 systems. There is an open loop procedure and closed loop procedure to control the transmit power of mobile users. In the open loop procedure, mobile station adjusts the transmit power by measuring the RSSI from the base station downlink signals. The objective of the open loop power control is not to accurately adjust the mobile transmit power but to reduce the dynamic range of error power so that the closed loop procedure will have an easier job to bring the mobile transmit power to the target level. In the closed-loop power control procedure, there are inner loop power control and outer-loop power control. The inner loop procedure resides at the base station and transmit power control commands once per 1.25ms to control the mobile transmit power up or down by Δ dB so that the measured uplink SIR is similar to the target SIR. On the other

hand, the outer-loop procedure adjusts the SIR target on a frame-by-frame basis based on a target FER.

Finally, we have elaborated on the soft handoff processing at both the mobile station and base stations in IS95 systems. Soft handoff refers to the handoff arrangement where the mobile establishes a new connection with the new base station before releasing the old connection (make-before-break). Soft handover is a unique feature in CDMA systems such as IS95, CDMA2000 and UMTS due to the low-cost mobile receiver to receive and combine signals from multiple base stations. We have elaborated the physical procedure to measure the pilot strength of the home cell and neighbor cells using CDMA structure. At the mobile station, pilots are partitioned into active set, candidate set, neighbor set and remaining set. Members of the active set are the base stations that are directly engaged in carrying payload of the user. When there are more than one members in the active set, the mobile station is engaged in soft handover. Pilot strength from the various sets are measured frequently at the mobile. If there are any strong pilots on the neighbor set or remaining set, the mobile station reports to the base station controller using PSMM . The base station controller replies with a handoff direction message (HDM) and the mobile station completes the transaction using HCM. There are three possible handoff transactions, namely the *pilot addition event*, the *pilot dropping event* and the *pilot swapping event*.

PROBLEMS

6.1 The outer loop power control is done at the base station of IS95 systems [True/False].

6.2 In IS95 system, the mobile station will combine messages (diversity combining) from the paging channels of the neighboring base stations to enhance the reliability of receiving the paging message [True/False].

CHAPTER 7

GSM

It takes all sorts to make a world.

—English Proverb, traced to the XVII century

7.1 INTRODUCTION

Global System for Mobile Communications (GSM) is a digital wireless network standard designed by standization committees of the European telecommunication operators and vendors [298, 239]. The objective is to design a high capacity cellular system to operate in the 900MHz cellular band with service portability, quality of service and security, high spectral efficiency as well as low cost design. The GSM system offers a common set of compatible circuit-switched services to all mobile users across European and the rest of the world. More than 70% of the mobile population in the world are GSM-based system. Hence, GSM represents a very important and successful deployment of digital cellular system and is one typical example of 2G cellular system. Around year 2000, the GSM system was evolved to GPRS system to

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

offer high speed packet switched data service. This refers to the so-called 2.5G system where packet-switched connection is the key new element. After GPRS, we have the EDGE system which further enhances the radio interface of the GPRS system to offer higher bit rate data services. Eventually, the GSM system will be evolved into UMTS system (today's 3G networks) to offer high quality wireless packet data access with quality of services. In this chapter, we shall provide an overview of the GSM and GPRS system. We first introduce the overall GSM/GPRS system architecture. Next, we shall look at the radio interface design. We will look at how logical channels are partitioned based on the hybrid FDMA/TDMA multi-access scheme. We will also look at the core network design in GSM/GPRS system and discuss some GSM-based services such as mobility management and international roaming. Moreover, we shall illustrate with practical illustration of GSM-based services such as the end-to-end call setup scenarios for circuit switched and packet switched connections. Finally, we shall conclude with a brief summary.

7.2 GSM SYSTEM ARCHITECTURE

We shall first elaborate on the GSM system architecture, followed by the GPRS evolution. Figure 7.1 illustrates the system architecture of a GSM network [324]. In this architecture, a mobile station (MS) communicates with a base station system (BSS) over the radio interface. The BSS is connected to the core network via a mobile switching center (MSC) using the *A-interface*. In the BSS, a number of base stations are connected to a base station controller (BSC). In the core network subsystem, the MSC connects with each other to switch calls between mobiles as well as interworking with external PSTN network.

7.2.1 Mobile Station

Unlike the IS95 system architecture in Chapter 6, the MS consists of two important parts [259]: the *subscriber identity module* (SIM) and the *mobile equipment* (ME). The SIM can be realized as a smart card or a smaller-sized "plug-in SIM". The SIM is protected by a personal identity number (PIN) between four to eight digits in length. The PIN is initially loaded by the network operator at the subscription time. Depending on the mobile implementation, this PIN can be deactivated or changed by the user. To use the mobile phone, the user may be asked to enter the PIN.

More importantly, the SIM contains very important subscription data (such as the service profile and the keys for encryption and authentication into the GSM network). The SIM also includes a list of abbreviated dialing, short messages received as well as the contact list. Parts of the SIM information can be modified by the subscriber either by using the keypad of the mobile or by a personal computer. However, some parts of the SIM information (such as the subscription data and the keys) cannot be modified by the user. In GSM, the SIM card can be updated over the air through

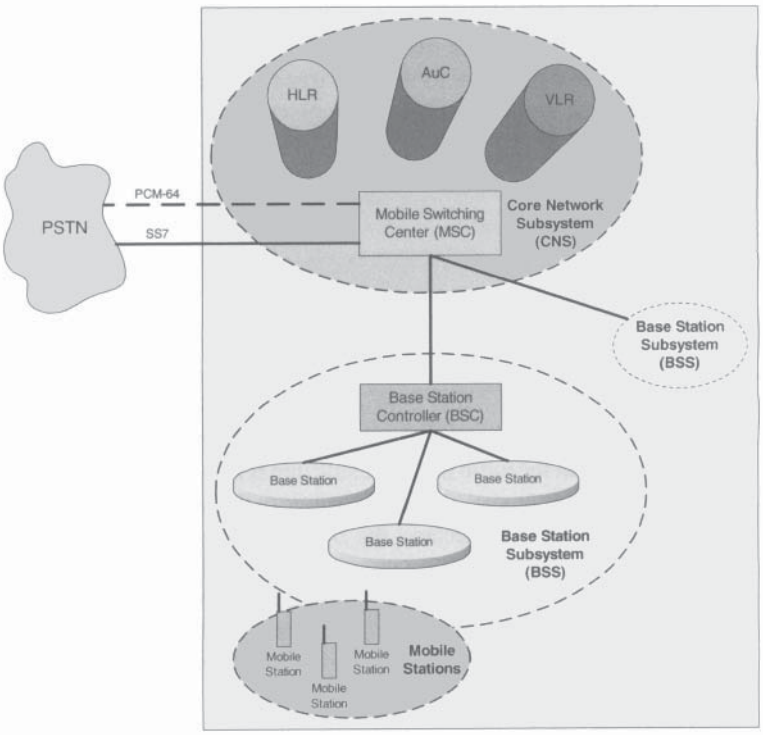


Figure 7.1 System architecture of GSM system.

the SIM toolkit as well. The network operator may remotely upgrade a mobile by sending codes through the SMS. These SMS messages are issued from a SimCard server and are received by mobiles equipped with SIM-toolkit capability.

There are two types of mobile station defined in GSM system, namely the *type 1* and *type 2* mobiles. Type 1 mobiles are not required to transmit and receive at the same time. Almost all GSM phones are of type 1 for simpler design. This means that although they operate in FDD mode with frequency offset of 45MHz between the uplink and downlink in the 900MHz cellular band or 95MHz in the 1900MHz PCS band, they are not operating in full duplex in strict sense. This is because they do not transmit or receive at the same time. On the other hand, type 2 mobiles are required to transmit and receive at the same time. Hence, they operate in full duplex mode and they usually require a duplexer in the RF circuit to separate the mobile transmit signal from jamming the mobile receive path.

7.2.2 Base Station Subsystem

The base station subsystem consists of multiple base stations connected with a base station controller (BSC) by fixed line transmission systems such as T1, E1 or ATM. The base station (BTS) interfaces the mobile station with the base station controller via the radio interface. On the radio interface side, the BTS contains radio transceivers as well as baseband digital channel elements specific to the radio interface of the GSM specification. On the fixed line interface, the BTS contains modems for T1/E1/ATM connections with the BSC. The fixed line interface is also called the *A-bis interface*. In fact, most of the signalling messages are terminated into the BSC or the core network rather than the BTS. Hence, the BTS simply acts as a physical relay for signals between the mobile and the BSC. An important part of the BTS is the transcoder and rate adapter unit (TRAU) that carries out GSM-specific speech encoding (downlink) and decoding (uplink) and rate adaptation in data transmission. However, in some implementations, the TRAU may be located in the MSC so that one voice channel occupies only 16kbps of the T1 time slot (instead of 64kbps) in the fixed line transmission link between the BTS, BSC and MSC.

On the other hand, the BSC is responsible for the switching functions and signaling protocol terminations (related to the radio resources) and is in turn connected to the MSC in the core network subsystem. The BSC supports radio channel allocations, handover management as well as other radio resource related functions (such as call admission). In GSM systems, multiple BTS can be connected to one BSC but one BSC can be connected to a BTS only. If there is only one BTS and one BSC in the base station subsystem, the BTS and the BSC may be integrated without the A-bis interface.

7.2.3 Core Network Subsystem

The core network subsystem [217] supports the call switching functions, subscriber profiles and mobility management (such as location update and paging). The basic switching function resides in MSC, which is a circuit-switched switching fabric similar to the PSTN switch. However, the software protocols in the MSC supports all GSM-based signalling protocols such as the GSM-MAP and mobility management protocols (MM). Multiple MSCs can be interconnected together in a hierarchical manner using fixed line transmission systems. There is a gateway MSC which interfaces with external public PSTN via PCM-64 in the speech payload path and SS7 in the signaling path. The MSC also communicates with other network entities in the core network subsystem such as the Visiting Location Register (VLR), Home Location Register (HLR) and the Authentication Center (AuC) using the GSM-MAP protocol (based on SS7). The VLR and HLR are very important elements (forming a distributed database) to keep track of the location of the mobile stations as well as supporting international roaming. When an MS moves from the home system to a visited system, its location is registered at the VLR of the visited system. The VLR informs the HLR of the mobile's home system of its current location. The AuC is used in the security data management for the authentication of the subscribers. The AuC may also be colocated with the HLR.

When a call is established between the mobile and the PSTN fixed line telephone, the entities involved in carrying the speech payload are the mobile, BTS, BSC, MSC and the gateway MSC. The entities involved in the signaling are the VLR, HLR, AuC (if outgoing call), MSC and gateway MSC. In GSM systems, since only the circuit-switched connection (for voice or data) is supported, the core network is designed based on circuit-switched architecture as well similar to the PSTN network.

7.3 GPRS SYSTEM ARCHITECTURE

In early 2000, there is strong demand for wireless data applications due to the bloom in the fixed line internet. While the existing GSM system supports wireless data connection, the bit rate and efficiency are very low. While 3G networks were still remote at that time, General Packet Radio Service (GPRS) serves as an interim solution to offer high quality, high data rate and low cost wireless data access based on the GSM network. This refers to the 2.5G solution that enables existing operators to *upgrade* the GSM network to offer wireless packet data access over the existing cellular and PCS spectrum. GPRS is designed based on an *overlay approach* that reuses the existing GSM architecture. For instance, in GPRS systems, both the traditional circuit-switched connection and the new packet-switched data connection are supported [325]. However, as explained in Chapter 5, the challenge in the infrastructure design lies in the interfacing between two different types of public networks (PSTN and internet). For instance, the PSTN is a circuit-switched network while the public internet is a packet-switched network. In GPRS system, the circuit-switched connec-

tion is provided through the core network subsystem (circuit-switched oriented). On the other hand, the packet-switched connection is provided through a new layer of network entity in GPRS (packet-switched oriented). This is illustrated in Figure 7.2

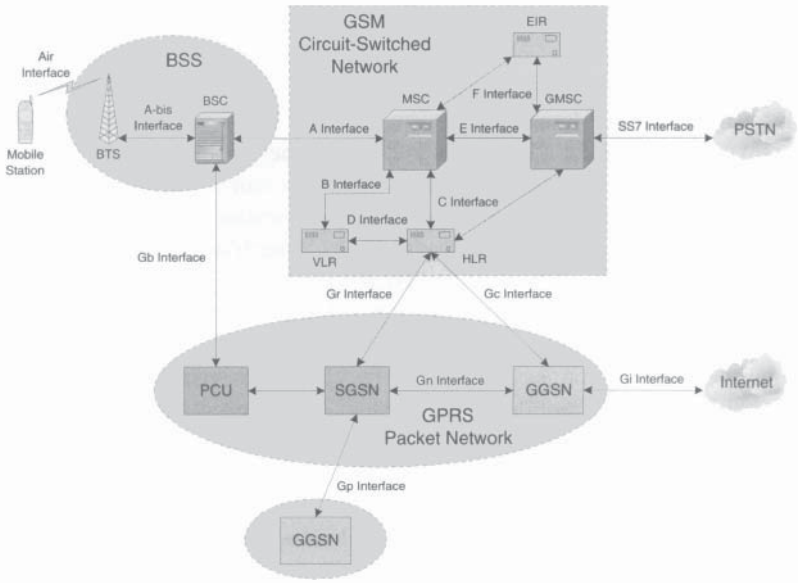


Figure 7.2 System architecture of GPRS system.

In GPRS system, the mobile, BSS, MSC, VLR and the HLR in the existing GSM system are modified. For example, the HLR is enhanced with GPRS subscriber information. In addition, a new packet switched layer is introduced for the core network in the GPRS system. The new packet-switched layer consists of two new entities, namely the serving GPRS support node (SGSN) and the gateway GPRS support node (GGSN). The SGSN and GGSN are IP routers which interconnects with each other via an IP-based GPRS backbone. The GGSN provides interworking with external public IP networks. The GGSN and SGSN are regular IP routers (in fixed line internet) except that they support mobility management protocols specific to the GPRS networks.

The BSS and the SGSN are connected via the Gb interface using ATM or frame relay. Within the same GPRS network, SGSNs and GGSNs are connected through the Gn interface. When SGSN and GGSN are in different GPRS networks, they are interconnected via the Gp interface. The GGSN connects to the public internet via the Gi interface. The HLR connects to the GGSN and the SGSN via the Gc and Gr interfaces respectively. Both the Gc and Gr interfaces follow the GSM-MAP protocol. The mobile, BTS, BSC, SGSN and GGSN are involved directly to carry

the packet-switched data payload. The SGSN, GGSN, VLR and HLR are involved in the signaling for packet-switched connections.

7.3.1 GPRS Mobile

A GPRS mobile consists of a mobile terminal (MT) and terminal equipment (TE). An MT communicates with the BSS over the air based on GSM/GPRS radio interface specification. The MT is equipped with software for GPRS functionality in order to handshake and establish signaling and payload connections with the core network (SGSN/GGSN). We shall illustrate the protocol architecture of the GPRS system in later sections. A TE can be a computer attached to the MT. On the other hand, modern GPRS devices such as palm pilot or powerPC are equipped with GPRS modem. In that case, the MT and the TE are integrated into a single physical unit. Existing GSM mobiles do not support GPRS. For example, GPRS utilizes time slot bundling to increase the downlink bit rate. In GSM, mobile can only utilize a single time slot to deliver speech payload. Furthermore, there are four new encoding modes in GPRS which are not supported in regular GSM mobiles. Figure 7.3 illustrates the growth of GPRS enabled handsets over the past few years. As of 2002, 96 operators in 45 countries have GPRS systems in operation. GPRS enabled handset grows from a scant of 10 million in 2001 to over 280 million in 2005.

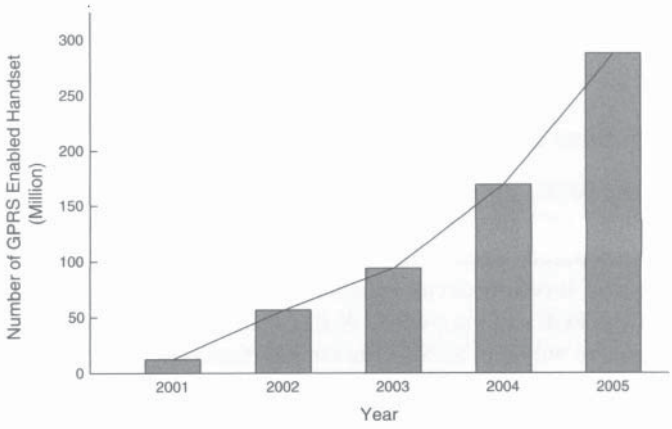


Figure 7.3 Growth of GPRS enabled handset.

Three operation modes were defined in GPRS, namely the *class A*, *class B* and *class C* mobiles. The class A mode of operation allows simultaneous circuit-switched and packet-switched services using different time slots within the GSM TDMA frame structure. Hence, class A mobile allows user to be in a GSM call or a GPRS data session and then monitor the applicable paging channel of the other service and react to it appropriately. Class B mobile supports GSM or GPRS sessions but not

simultaneously. Once a user is in a GSM call or a GPRS data session, it is not required to monitor the paging channel for the other service. Hence, it is not required to respond to a page message if the mobile is paged for incoming calls from the network. Class C mobile supports only GPRS or GSM data services. These types of mobiles are typically modems. No provision is provided for the mobile to react to pages from either mode. By default, GSM circuit switched data mobiles are Mobile Station Class C mobiles.

The mobile stations that access GPRS services may or may not contain GPRS-aware subscriber identity modules (SIMs). An MS maintains mobility management (MM) and Packet Data Protocol (PDP) contexts to support GPRS mobility management. Some of the MM context fields stored in a GPRS aware SIM include international mobile subscriber identity (IMSI), which uniquely identifies the MS. In addition, the SIM stores the packet temporary mobile subscriber identity (P-TMSI), which is the GPRS equivalent of the TMSI in GSM. The SIM stores the routing area identity, which is a subset of the location area in GSM. If the SIM is not GPRS aware, these fields are stored in the mobile equipment. For data routing purposes, the MS maintains a PDP context which includes the PDP type (X.25, IP, PPP), PDP address, PDP state (ACTIVE/INACTIVE) as well as the QoS profile. We shall elaborate the handshaking and signaling involved in packet data session in later section.

Figure 7.4 illustrates the protocol stack in a GPRS mobile (packet data service only). We shall elaborate the details of the protocol layers and as well as the handshaking and signaling involved in packet data session in later section.

7.3.2 GPRS Base Station Subsystem

To accommodate GPRS, the base transceiver station (BTS) and the base station controller (BSC) in the BSS are modified. A new component, namely the packet control unit (PCU) is introduced. The BTS is modified to support new GPRS channel coding schemes. The BSC forwards circuit-switched calls to the MSC, and packet-switched data (through the PCU) to the SGSN. A BSC can connect to only one SGSN. A BSC can connect to only one SGSN. The Gb interface is implemented to accommodate functions such as paging and mobility management for GPRS. The BSS should also manage GPRS-related radio resources such as allocation of packet data traffic channels in cells.

The PCU can be viewed as an equivalent of the transcoder and rate adaptor unit (TRAU) for the packet switched service. The PCU is either colocated with the BTS or remotely located in the BSC or the SGSN. Most vendors follow the remote PCU option so that no hardware modifications to the BTS/BSC are required. In the remote option, the existing A-bis interface between the BTS and the BSC is reused and the GPRS data/signaling messages are transferred in modified TRAU frames with a fixed length of 320 bits (20ms). The PCU is responsible for medium access control (MAC) and radio link control layer functions, such as packet segmentation and reassembly,

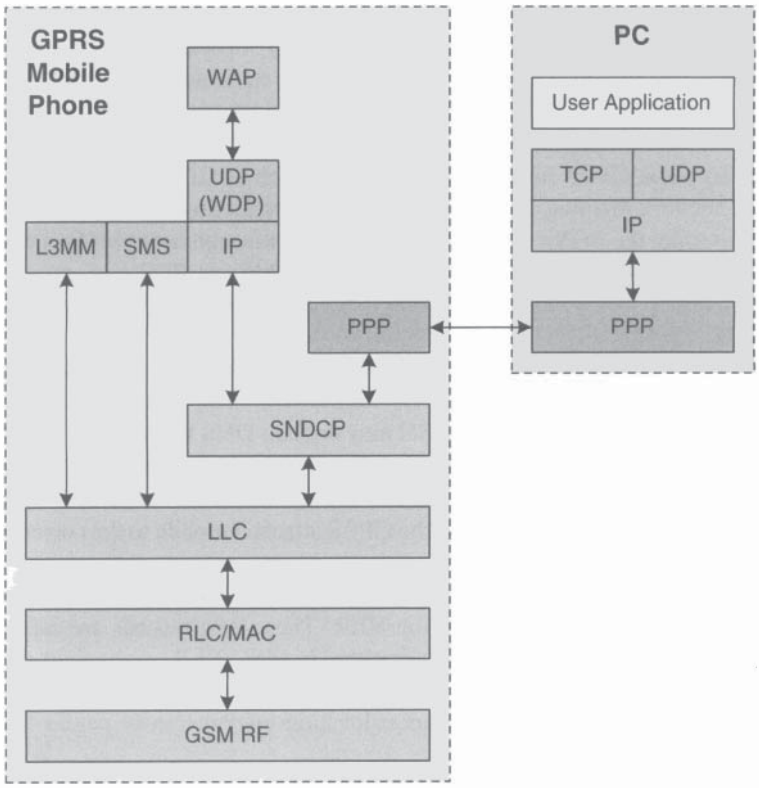


Figure 7.4 Illustration of GPRS protocol stack.

packet scheduling and ARQ as well as radio resource management (such as power control, congestion control and broadcast control information).

7.3.3 GPRS Support Nodes, HLR and VLR

There are two kinds of GPRS support nodes, namely the serving GPRS support node (SGSN) and the gateway GPRS support node (GGSN). The functionality of the SGSN and GGSN may be combined into a single physical node or distributed in separate nodes. A GSN is typically implemented using multi-processor architecture with hardware redundancy, hot swapping and robust software platform that support uninterrupted operation. The role of the SGSN is equivalent to the MSC/VLR in the current GSM network. It is involved in carry the packet data payload as well as signaling in a GPRS packet data session. SGSN connects the BSS to GGSN, which offers encryption, GPRS mobility management (such as GPRS routing area update and inter-network roaming), charging and statistics collection. To provide services to a GPRS mobile, the SGSN establishes an mobility management (MM) context that contains mobility and security information for the MS. At PDP context activation, the SGSN establishes a PDP context that is used to route data between the MS and the GGSN.

GGSN is primarily provisioned by a router, which supports traditional gateway functionality such as publishing subscriber addresses, mapping addresses as well as routing and tunneling packets. A GGSN may contain DNS functions to map routing area identifiers with serving SGSNs and DHCP functions to allocate dynamic IP address to the GPRS mobiles. The GGSN also maintains an activated PDP context for tunneling the incoming packets of the GPRS attached mobile to the corresponding SGSN.

To accommodate GPRS subscription and GPRS routing area, new fields in the GPRS MS record are introduced in the HLR. These new records are assessed by SGSN and GGSN using IMSI as the index key. In MSC/VLR, a new field called the SGSN number is added to indicate the SGSN currently serving the GPRS mobile. The MSC/VLR may contact SGSN to request location information or paging for voice calls.

7.4 RADIO INTERFACE

In GSM/GPRS system, the spectrum resource is partitioned into channels based on hybrid FDMA and TDMA. The uplink and downlink signals in GSM are separated in frequency (namely the frequency division duplexing (FDD)). For instance, the 900MHz cellular spectrum for the GSM downlink and uplink signals are 935–960 MHz and 890–915 MHz, respectively. The 1800 MHz PCS spectrum for the GSM (DCS-1800) downlink and uplink signals are 1805–1880 MHz and the 1710–1785 MHz respectively. In the cellular (900MHz) spectrum, the frequency band (both

uplink and downlink) are first partitioned into 124 pairs of frequency duplex channels (carrier) with 200kHz carrier bandwidth. Each of the 200kHz carrier is further partitioned into 8 time slots based on TDMA. The time slot is the basic channel to be allocated to mobile users. Such a hybrid arrangement is to strike a balance between the number of radio transceivers needed at the base station and the equalization complexity of the mobile. This point has been explained in Chapter 3. In addition to the frequency separation of 45MHz for cellular spectrum (900MHz) and 95MHz for PCS spectrum (1800MHz), the uplink and downlink bursts of class 1 mobiles are further separated by 3 time slots. Hence, this removes the need for class 1 mobile to transmit and receive simultaneously and therfore, simplifies the design of the mobile. Figure 7.5 illustrates the timing schedule of uplink and downlink bursts in GSM for class 1 mobile. For instance, the mobile receives a downlink burst from the base station, re-tune to the uplink frequency and transmit the uplink burst 3 time slots later.

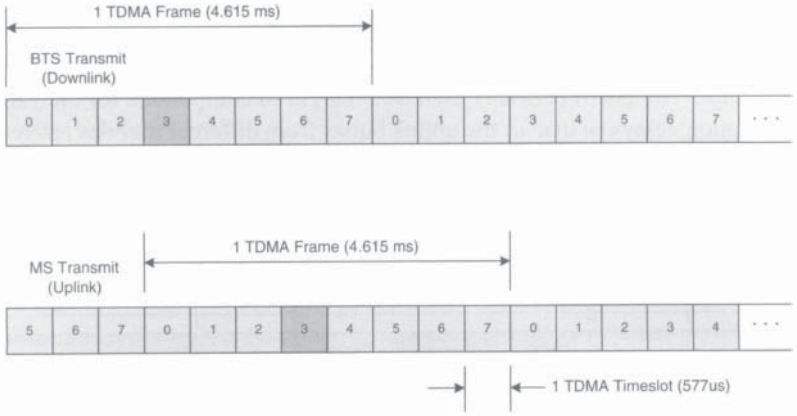


Figure 7.5 Timing schedule of uplink and downlink burst in GSM-FDD for class 1 mobiles.

Note that the propagation path loss of lower frequency band is in general smaller than that in higher frequency band. Hence, to save the mobile power, the uplink is usually assigned to the lower frequency band (890–915 MHz) while the downlink is assigned to the higher frequency band (935–960 MHz). To further save mobile battery power, GSM supports discontinuous transmission and discontinuous reception. Discontinuous transmission exploits the silence gaps between human conversation. The mobile station turns on the transmitter only when there is voice packet and turns off the transmitter during the silence gaps. In discontinuous reception, the mobile station can turn off the receiver during the time slots of the other users and the mobile only need to listen to its allocated time slot for paging. The 8 time slots per 200kHz carrier are the physical channels used to carry user traffic. However, to facilitate the call processing and controls, the time slot (physical channels) are first partitioned

into a number of logical channels. We shall elaborate the important logical channels associated with GSM and GPRS systems in the following subsections.

7.4.1 GSM Radio Interface

As we have introduced, each GSM carrier (200kHz) is partitioned into 8 time slots based on TDMA and the data is transmitted in the form of *bursts* and the bursts are designed to fit within the time slots. In GSM, five different burst types [143] have been defined. The five burst types are namely the *normal burst* (NB), *frequency correction burst* (FB), *synchronization burst* (SB), *access burst* (AB) and the *dummy burst* (DB). Figure 7.6 illustrates the normal burst structure. The normal burst is the most commonly used bursts in GSM to carry user payload data. It contains 116 bits of information (user payload), 26 bit training sequence and a guard period of 8.25 bits. The training sequence is to allow the mobile station to estimate the channel fading response and to perform simple equalization (to alleviate the effect of frequency selective fading). For 200kHz carrier, the number of resolvable echos experienced by the carrier will be limited and hence, the equalization complexity is feasible in GSM mobile. The guard period is used as a margin to compensate for the variation of propagation delays so that all the uplink bursts from different mobiles arrive at the base station without *colliding* with each other. The length of the guard period will limit the maximum coverage of the GSM BTS.



Figure 7.6 Illustrate of GSM burst.

The frequency correction burst is used by the mobile to detect a special carrier which is transmitted by all the BTS in a GSM network. This carrier is called the broadcast channel (BCCH) carrier which delivers important system parameters related to the call processing. The synchronization burst (SB) is used by the mobile to synchronize to the BTS transmission timing to within a quarter-bit resolution based on the extended 64 bit training sequence. The access burst is used by the mobile to access the system (e.g., call request, page response) and it is the first uplink burst that a BTS will have to receive and demodulate from any particular mobile. Similar to the synchronization burst, the training sequence in the AB is extended to ease the demodulation process in the base station. Furthermore, the access burst is much shorter than the other burst and this results in a large guard period of 68.25 bits. This guard period is used to compensate for the propagation delay between the MS and the BTS. Such a large guard period is needed for AB because during access, there is no

timing advance mechanism yet and hence, a large guard period is needed to cover the worst case propagation delay. For instance, 68.25 bit period is equivalent to 252us which allows the MS to be up to 38km from the BTS before the received uplink AB will spill into the next time slot. In other words, 38km is the maximum cell radius in GSM. Finally, the dummy burst is similar to the normal burst in that it has the same structure and uses the same training sequences. However, the information bits in the dummy burst are set to a predefined sequence in the DB. The DB is used to fill inactive time slots on the broadcast channel carrier, which is transmitted continuously at constant power.

The radio spectrum is partitioned into 200kHz carriers and then time slots. Hence, a combination of carrier number and time slot number identifies a *physical channel* uniquely in GSM system. However, not all physical channels can be assigned to users to carry user payload. The physical channels have to be mapped to a number of logical channels in the GSM system. There are two types of logical channels, namely the *traffic channels* (TCH) and the *control channels* (CCH). TCHs are intended to carry user information (e.g., speech packets or data packets) while CCHs are used to carry control or signaling information. The traffic and control channels are further elaborated below.

7.4.1.1 Traffic Channels in GSM There are two kinds of TCHs defined in GSM, namely the *full-rate TCH* (TCH/F) and the *half rate TCH* (TCH/H). The TCH/F provides transmission speed of 13kbps for speech or 9.6kbps, 4.8kbps or 2.4kbps for data. A full rate TCH will occupy a complete physical channel (i.e., one time slot in each 4.615ms TDMA frame) on one uplink and one downlink carrier). The half rate TCH allows speech transmission at around 7kbps and data at primary user rates of 4.8 kbps and 2.4kbps. The half-rate channel uses one time slot in every other TDMA frame on average and hence, each physical channel (a time slot and a carrier) can support two half-rate TCH. The half rate channel is primarily intended to support the GSM half-rate speech coder, which produces digitalized speech at a reduced source rate of around 7kbps. TCH/F and TCH/H always use normal bursts.

Figure 7.7 illustrates the GSM frame and hyperframe structure. In GSM, one hyperframe consists of 2048 superframes. One superframe contains 51 *multi-frames* and each multi-frame contains 26 TDMA frames. The BTS and all the mobiles in the GSM system are time-synchronized from the hyper-frame level to the time slot level. In other words, a time slot in a carrier is *labeled* by a *time slot number*, a *TDMA frame number*, a *multi-frame number*, a *super-frame number* and a *hyper-frame number*. All the mobiles and the BTS have the same "notion" (or time reference) of the time slot number, TDMA frame number, Multi-frame number, super-frame number as well as the hyper-frame number. In fact, each of the 2715648 time slots in a hyperframe has a unique number and this is used in the ciphering (security) and frequency hopping algorithms.

Figure 7.8(a) illustrates the mapping of physical channels (physical time slots) into logical channel (TCH/F). The time slots in the first 12 TDMA frames (0-11) are used

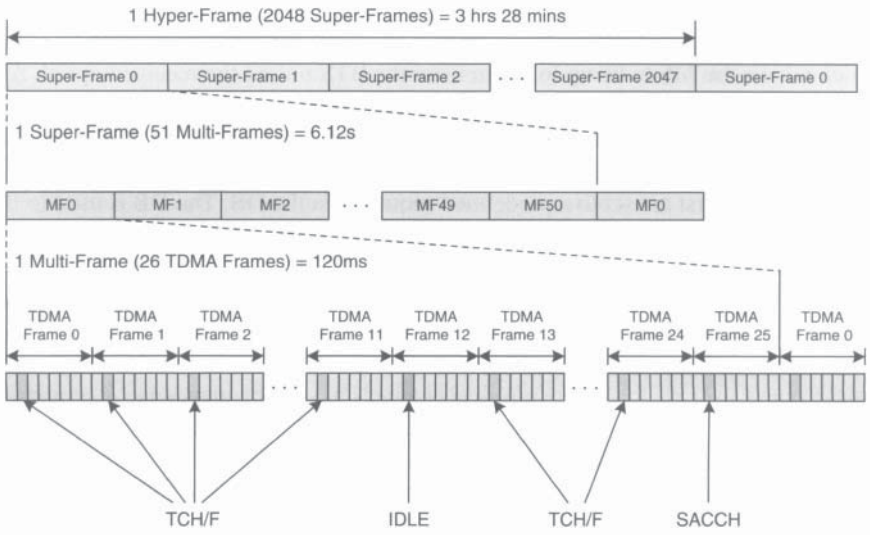


Figure 7.7 Illustration of GSM frame structure.

for TCH/F. The time slots in the 13-th TDMA frame (12) are not used for transmission and are labeled as IDLE. The next 12 TDMA frames (13-24) are used by the TCH/F. The time slots in the 26-th TDMA frame (25) are used by the slow associated control channel (SACCH), which is a control logical channel to be elaborated in the next subsection. The organization of the TCH/H is illustrated in Figure 7.8(b) with two half rate traffic channels, TCH0 and TCH1. The two half rate traffic channels share a time slot in every TDMA frame. Similarly, the time slots in TDMA frames 0-11 and TDMA frames 13 - 24 are used to carry TCH0 and TCH1. The time slots in TDMA frame 12 are used for SACCH corresponding to TCH0. The time slots in TDMA frame 25 are used for SACCH corresponding to TCH1.

7.4.1.2 Control Channels in GSM The control channels in GSM are intended to carry signaling information. Specifically, there are three types of CCHs defined in GSM, namely the *common control channels* (CCCH), the *dedicated control channels* (DCCH) and the *broadcast channels* (BCH). The control channels are needed to carry the signaling or control information between the mobiles and the BTS or to provide system information related to radio resource management and call processing.

The common control channels are shared by all mobiles in the BTS, which include the *paging channel* (PCH), the *access grant channel* (AGCH) and the *random access channel* (RACH). They are elaborated below.

Random Access Channel (RACH): The RACH is an uplink control channel used by the mobiles for initial access such as call origination, paging response or

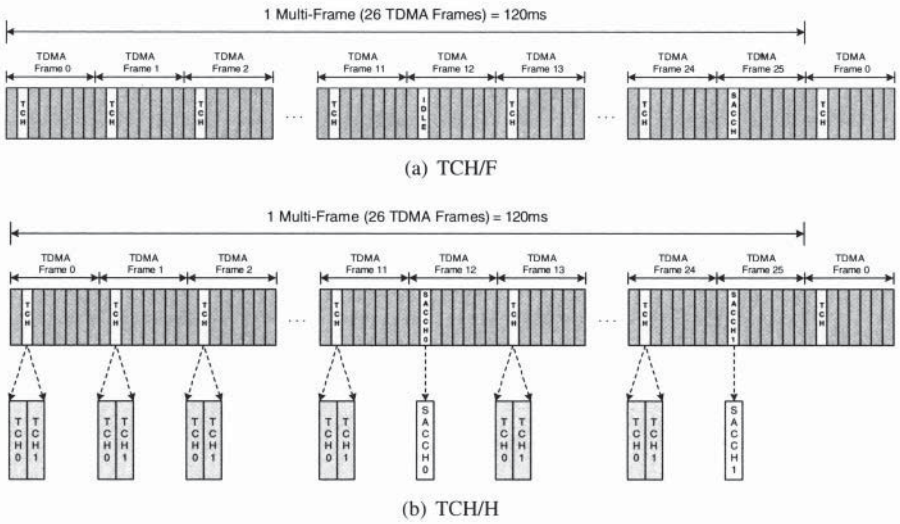
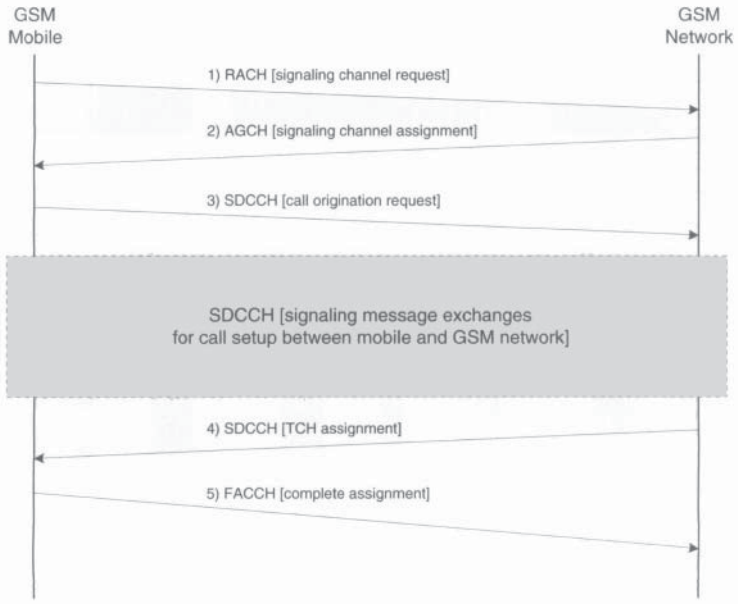


Figure 7.8 Illustration of the mapping from physical channel to logical traffic channels in GSM.

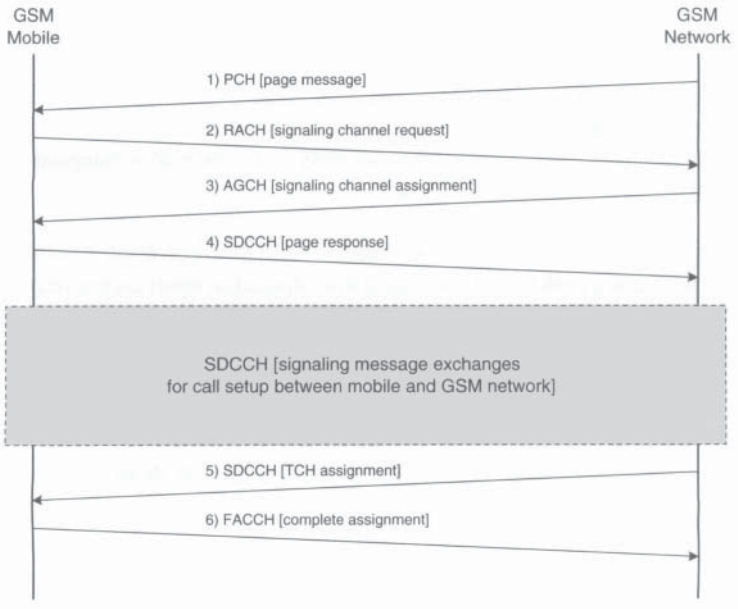
location update. Figure 7.9 illustrates the handshaking of a typical GSM call origination and call termination. The call request message is sent by the mobile via the RACH. Since all mobiles share the same RACH for initial access, there is potential collisions between access requests and hence, slotted Aloha (similar to that described in random access of IS95 in Chapter 6) is adopted in RACH to resolve potential access collision.

Access Grant Channel (AGCH): The AGCH is also a downlink control channel used by the network to indicate radio link allocation upon access requests from mobiles. For example, as illustrated in Figure 7.9(a), the BTS assigns a low-data rate SDCCH (introduced below) to carry the signaling messages and notify the mobile using the AGCH.

The dedicated control channels are control channels dedicated to a mobile in GSM. There are four types of dedicated control channels defined in GSM, namely the *stand-alone dedicated control channel* (SDCCH), the *slow associated dedicated control channel* (SACCH), the *fast associated dedicated control channel* (FACCH) and the



(a) GSM call origination



(b) GSM call termination

Figure 7.9 Illustration of the handshaking in GSM call *origination and call termination.*

cell broadcast channel (CBCH). The CBCH is used on the downlink only while the SDCCH, SACCH and FACCH are used in both downlink and uplink. They are elaborated as below.

Standalone Dedicated Control Channel (SDCCH): The SDCCH is a dedicated channel setup to carry signaling between a mobile and the BTS or for short message delivery. In fact, the SDCCH is similar to the TCH in the sense that they are both used to provide a dedicated connection between the mobile and the BTS. In some situations (such as location update and call setup), signaling information must flow between the mobile and the BTS. This may be accomplished by setting up the TCH directly. However, this may waste the radio resource because the data transfer requirement of the signaling message is far less than that of speech transmission. Hence, the SDCCH is defined to be a low-date rate dedicated control channel which has only 1/8 of the capacity of a TCH. For example, as illustrated in Figures 7.9(a) and 7.9(b), a SDCCH is first setup in both call origination and call termination scenarios to carry the call origination request message (or page response message) from the mobile to the BTS. The TCH assignment is also delivered from the BTS to the mobile via the SDCCH.

Slow Associated Dedicated Control Channel (SACCH): The SACCH (as illustrated in Figure 7.8) is associated with either a TCH or an SDCCH. The SACCH is used for non-urgent control messages such as the transmission of power control and time alignment information over the downlink, and measurement reports from the mobiles over the uplink. As a result, the power control in GSM (delivered in SACCH) is a slow closed-loop control (as compared with IS95 closed loop control) because the power control commands are delivered once per 120ms.

Fast Associated Dedicated Control Channel (FACCH): The FACCH is used for time-critical control messages such as call-establishment progress, authentication of mobiles, or handover signaling. The FACCH makes use of the TCH during a call and hence, there is a loss of user data because the FACCH steals the bandwidth of the TCH. This is similar to the blank-and-burst and dim-and-burst signaling in IS95 introduced in Chapter 6.

Cell Broadcast Channel (CBCH): Finally, the CBCH carries only the short message service cell broadcast messages, which use the same time slot as SDCCH.

The broadcast channels (BCH) are used by the BTS to broadcast system information to the mobiles in its coverage area. There are three types of BCHs, namely the *frequency correction channel* (FCCH), the *synchronization channel*(SCH) and the *broadcast control channel* (BCCH). The FCCH and the SCH carry information from the BTS to the mobile to allow the mobiles to acquire and achieve time synchronization (from time slot level to hyperframe level) with the GSM system. The BCCH