McGraw-Hill TELECOM
PROFESSIONAL

W-CDMA and cdma2000 for 3G MOBILE NETWORKS

- IMT-2000
- Linking GSM and W-CDMA systems
- Designing the 3G Mobile Communications Network

M. R. KARIM • MOHSEN SARRAF

TEAMFLY

W-CDMA and cdma2000 for 3G Mobile Networks

M.R. Karim
and
M. Sarraf

McGraw-Hill
New York Chicago San Francisco Lisbon
London Madrid Mexico City Milan New Delhi
San Juan Seoul Singapore Sydney Toronto

McGraw-Hill

A Division of The McGraw-Hill Companies

Copyright © 2002 by M.R. Karim and Lucent Technologies, Inc.0-07 All rights reserved. Manufactured in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

0-07-140956-4

The material in this eBook also appears in the print version of this title:0-07-138513-4.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please contact George Hoare, Special Sales, at george_hoare@mcgraw-hill.com or (212) 904-4069.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. ("McGraw-Hill") and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill's prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED "AS IS". McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

DOI: 10.1036/0071409564

To our families

Rahima, Razi, and Nayeem
—MRK

Maryam, Artin, and Shawhin
—MS

ABOUT THE AUTHORS

M. R. Karim, formerly a Distinguished Member of Technical Staff of Bell Laboratories, was a member of the original team that developed the world's first cellular system. He has published in the areas of mobile communications and packet switching, and is author of the book *ATM Technology and Services Delivery* (Prentice Hall, 1999).

Mohsen Sarraf received his Ph.D. degree in 1986 from the University of Southern California. He joined Bell Laboratories in 1987 where he has been involved in various aspects of communications systems. He has worked on wireless systems from design and implementation to project leadership during the last ten years. Currently he is the Director of Advanced Multimedia Communications Department of Bell Labs.

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

For more information about this book, click here.

CONTENTS

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

TEAMFLY

PREFACE

At the time we were working on *third-generation* (3G) wireless systems at Lucent Technologies, we realized that there were not many books available on this topic. ITU-R had defined four 3G systems, and published a set of standards in 1999. In most cases, our only sources of information were these standards, which were necessarily quite elaborate and were not available as a single document. The purpose of this book is to fill that void and provide a comprehensive description of 3G systems. The standards specify air interfaces based upon both *wideband CDMA* (W-CDMA) and *wideband TDMA*. However, since W-CDMA is the preferred interface, we have chosen to deal with W-CDMA and more specifically cdma2000 and UMTS FDD. Technologies used in 3G and necessary background material required to understand and, in some instances, develop a 3G system are presented. The treatment of topics is neither too detailed nor too brief, and our expectation is that a wide spectrum of readers—systems engineers, engineering managers, people who are new in this area but want to understand the system, and even designers—will find the book useful.

The book is organized as follows. We begin by tracing, in Chapter 1, the evolution of mobile telephony from analog systems (that is, *Advanced Mobile Phone Service* [AMPS]) through the *second generation* (2G) systems of the nineties and leading up to 3G systems. Included in this chapter is an overview of 3G capabilities, features, and requirements.

Knowledge of the propagation characteristics of a mobile radio channel is essential to the understanding and design of a cellular system. As such, an overview of this topic is presented in Chapter 2.

Chapter 3 describes the basic principles of *wideband CDMA* and deals with various topics that, in essence, provide the physical layer functionalities of a 3G system.

cdmaOne and cdma2000 are the subject matter of Chapter 4. Because cdma2000 is an evolution of cdmaOne, uses the same core network standards (that is, IS-41) as cdmaOne, and may coexist with this system, we begin with a synopsis of cdmaOne and follow it up with a description of cdma2000.

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

Chapter 5 is devoted to GSM and *General Packet Radio Service* (GPRS). The reasons we have included these two systems are the following: Both GSM and UMTS share the same core network and use the same *Mobile Application Part* (MAP) protocol of Signaling System 7. Similarly, the packet mode data services in UMTS and the associated network entities and protocols have been harmonized with those of GPRS. Thus, even though there are significant differences in the air interface standards of *UMTS Terrestrial Radio Access Network* (UTRAN) and GSM, a description of GSM and GPRS may be helpful to the reader in this context.

UMTS is described in Chapter 6, where, among other things, we discuss the protocols of different layers, synchronization schemes, power controls, and handover procedures.

Since packet mode data is an important aspect of 3G, existing core networks, which are built around a circuit-switch fabric, work in conjunction with routers and gateways to provide packet mode data services. In fact, because of high volume data transfer requirements in next generation systems, the core network is evolving to an all-IP architecture. Chapter 7 describes the evolution of mobile communication networks.

Chapter 8 touches briefly on call controls and mobility management in wireless networks. To help the reader understand this topic better, a brief description of protocol stacks at various interface points is also included.

Chapter 9 deals with the *quality of service* (QoS) concepts as they relate to 3G, provides the reader with a basic understanding of the subject, and discusses the need for implementing a flexible resource management scheme in the network that will provide mobile stations with an end-to-end QoS across all-IP networks.

Network planning and design issues, such as spectrum requirements, link budget calculation, frequency planning, and cellular growth, are presented in Chapter 10.

We conclude the book with our reflections, in Chapter 11, on what may come about beyond 3G, discuss the driving force behind the evolution of the *fourth-generation* (4G) system, and mention some technologies that might play a key role in the development of 4G.

ACKNOWLEDGMENTS

The authors would like to thank Reed Fisher who read almost the entire manuscript, and gave us valuable comments. Special thanks go to Ken Smolik who gladly reviewed much of the material and offered suggestions that have greatly enhanced the quality of the book. Thanks are also due to Nikil Jayant, Victor Lawrence, and an anonymous reviewer for going over a few chapters and giving us their comments. We are grateful to Marjorie Spencer for inviting us to write this book and for her continued interest in this endeavor. Finally, we would like to express our most sincere gratitude to our families because without their constant support and encouragement, we could not have undertaken this work and completed it on time.

M. R. KARIM
M. SARRAF
MARCH 2002

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

This page intentionally left blank.

FOREWORD

Throughout history and across boundaries, people have been engaged in a constant quest for information. What they have learned is that information is one of the most valuable and enabling commodities in the world. Those who have it become more powerful, and those who can access it faster than others gain an extra edge. For this reason, people are constantly in search of means to generate, archive, access, and transfer information as quickly as possible. This quest for obtaining and transferring information has made people innovate in many dimensions. It has made them create new words, new means of recording information, new means of interpreting information, and, above all, new means of transmitting information. In the latter area, over the past several thousand years we have observed the use of smoke signals and the creation and evolution of languages, mail systems, messenger services, the telegraph, wireless broadcast, telephony, wireless telephony, and now e-mail and wireless messages. Among important parameters in this quest are the amount, the type, the speed, the security, and the ease of access of the underlying information to be transferred.

As with many other scientific and technological quests, the advances in communications have come in cycles of slower progress in the beginning until a critical mass has been achieved, followed by a leap and the continuation of the cycle. Eventually, these leaps will take the technology to the point where the underlying service (be it agricultural, medical, engineering, scientific, or another type of service) will become inexpensive and reliable enough to make it economically viable for mass production, resulting in a big jump in quality of life. We are fortunate to live at a point in history that allows us to observe the many technological advances in information transfer taking place right in front of our eyes. Never before have we been able to transfer information of most types (text, image, sound) fast and securely enough for real-time applications, from anywhere to anywhere with portable gadgets light and small enough to fit easily in our pockets. Only a couple of decades ago, this achievement would have been relegated to science fiction writers and movie producers. The aforementioned scientific and technological leaps, however, have swiftly moved the achievement from imagination to implementation. To make implementations cheap and, at the same time, ubiquitous, those involved in bringing this technology to the public have created the Third Generation Wireless Telephony standards, commonly referred to as 3G.

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

Those who produce and implement 3G solutions will provide the public with great social and economic benefits. Learning about the basics of the technologies and methods upon which 3G solutions are based is the first step in this important task, and this book is an excellent vehicle to accomplish that step. With a depth that is just right for graduate students, engineers who are developing the systems, and others who want to grasp the breadth of the subject, it describes the most important issues in the design of the overall 3G system. (It is also suitable for business managers, product managers, sales and marketing, attorneys, and others who need to gain general knowledge of the subject.) At the same time, it easily accommodates the more advanced readers, who can use it to pinpoint the important issues in the field and follow up on them in the more advanced literature cited in its references. Some of the issues discussed in this book are the challenges of the wireless channel, the evolution of the older technologies to the current ones, the basics of the Code Division Multiple Access (CDMA) technology, systems planning, and the architecture of the systems and their evolution. All are presented in a highly readable manner, providing a great all-inclusive source for learning and references on the subject of 3G wireless technology.

I hope every reader enjoys and takes advantage of this book, as I did.

Victor B. Lawrence
Vice President
Advanced Communications Technology
Bell Laboratories—Lucent Technologies

CHAPTER 1

Introduction

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

Early Systems

The earliest recorded instance of radio service to moving vehicles, such as ships, trains, and automobiles, was an experimental system in 1919 that provided two-way radio communication among coastal steamers between Boston and Baltimore [2], [3]. For the next 12 years or so, considerable improvement was made to radio communications technology to provide an effective high-seas mobile radio service. For land-based users, however, the earliest mobile phone service was in 1933, although research laboratories started experimenting with it much earlier. This system used a 35 MHz frequency band and was available only to police and fire departments. There were only 10 channels in the system with a 40 kHz spacing. It was a manual system where channel assignment and dialing were performed by the telephone operator. Because the mobile could not receive and transmit information simultaneously, the user had to "push to talk." There was no roaming feature available. In other words, users would receive service only in their registered home areas and would be denied the service if they moved to different serving areas.

Subsequently, in 1946, the FCC granted some spectrum on the 150 MHz band for an improved mobile telephone service; that year, following this spectrum allocation, the first commercial service was introduced in St. Louis, Missouri, and by the end of the same year, services were available to 25 other U.S. cities. These earlier systems were manual in that all calls were handled by a telephone operator. Because of the heavy demand for this service, the FCC allocated six more channels around 150 MHz and 12 new channels around 450 MHz in 1956. This is the first time that a 450 MHz system was used for commercial service [1].

An improved version of the mobile telephone service was introduced in 1964. Known as the MJ, this system operated at 150 MHz and had 11 channels. Initially, the channel spacing was 120 kHz, but with the advancement of *radio frequency* (RF) circuit technology, this spacing was reduced to 30 kHz with a peak frequency deviation of 5 kHz. Each mobile serving area consisted of a single, fixed-tuned FM transmitter, which was located centrally at a high enough elevation so that it could serve all mobiles in the serving area with a high

probability. The RF power output of a transmitter was 50 to 250 W, while with the antenna gain, the radiated power at the antenna was usually in the range of 500 W. A number of FM receivers were placed at different points in the serving area to receive the signal from all vehicles. These transmitters and receivers were then connected to a control terminal in a local switch. Roaming features were now provided. However, because the complete routing information was not available to the local switch, a land-originated call to a roaming mobile had to be completed manually by telephone operators. The mobile unit could scan all available channels, lock onto an idle one, and then start dialing. Signaling was done using low-frequency audio tones. The maximum range between a serving transmitter and a mobile unit was about 25 miles. To provide satisfactory operation, frequencies could be reused but only at distances of 75 miles or more.

To meet the growing demand from customers, the FCC opened up another spectrum in the 450 MHz band. This system, which was introduced in 1969, was known as the MK system and had 12 channels with a frequency spacing of 25 kHz. Like its predecessor, it supported automatic dialing and operator-assisted roaming.

These early systems provided three types of mobile telephone service:

- *Complete Mobile Telephone Service* (MTS) for voice communication to land-mobile users assisted with mobile telephone operators where necessary.
- *Automatic Dispatch Service* (ADS) was used between one or more dispatchers and a fleet of mobile units. This service supported only one two-way conversation at a time between a dispatcher and a mobile unit. Conference calls between a dispatcher and multiple mobile units were not possible.
- One-way paging.

The spectrum allocated by the FCC for these early systems was usually quite small compared to the relatively large number of contending users. Also, because of the limitations of the hardware technology, the frequencies could not be reused at distances any closer than 75 miles or so. Thus, naturally, as the demand grew, users experienced high probability of call blockage. To overcome this

fundamental problem, the FCC set aside a bandwidth of 75 MHz in the 850 MHz range and asked common carriers to submit their proposals for a *high-capacity mobile telecommunication system* (HCMTS) [1]. In response, the Bell System submitted comprehensive details of one such system based on the cellular concept that had been under development in Bell Laboratories since 1947 [4]. Finally, in 1974, the FCC ruled that 40 MHz of the original 75 MHz spectrum could be used by common carriers to provide advanced mobile telephone service, and the remaining 30 MHz was reserved for private services. In 1975, the Illinois Bell Telephone Company filed a petition to the FCC asking for permission to build and test a cellular system. The permission was granted in 1977. Consequently, in 1978, a development system that was built in Bell Laboratories during 1972 to 1977 was installed in Chicago to verify the system concept and design issues. This phase of the trial, known as the *Equipment Test*, involved only 100 mobile units. A follow-up test phase, known as the *Service Test*, was launched in the following years using about 2,000 mobile units that were designed by outside vendors[1] according to Bell Laboratories specifications.

TEAM-DLI

The Cellular System

The FCC allocated a bandwidth of 20 MHz—from 870 to 890 MHz—in the forward direction (that is, from base station transceivers to mobile stations) and another band of 20 MHz—from 825 MHz to 845 MHz—in the reverse direction.[2] These frequency bands were divided into a number of channels, each with a bandwidth of 30 kHz. The operating frequencies of these channels are shown in Figure 1-1.

The idea behind a cellular system is simple [22]. Because the spacing between adjacent channels is 30 kHz, there are altogether 666 channels in either direction. Of these channels, a few are set aside for

[1]They were Motorola and E. F. Johnson of the United States and Oki of Japan.

[2]Because a different frequency band is used for transmission in each direction, the system is said to operate in a *frequency division duplex* (FDD) mode.

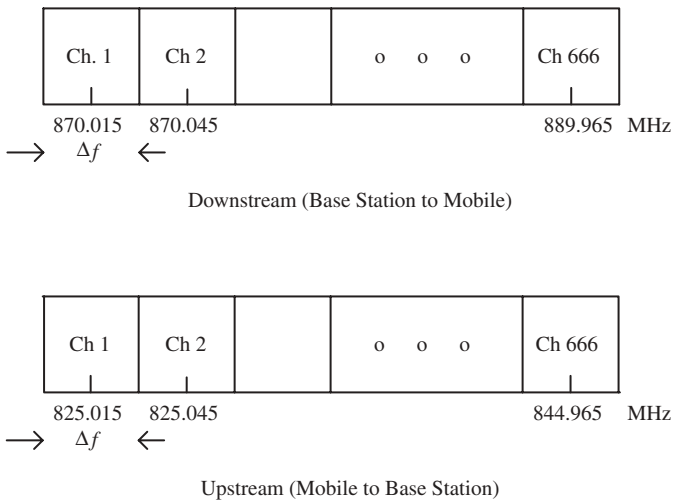


Figure 1-1 Spectrum allocation and channel assignment in advanced mobile phone service

access and control purposes, while the rest are used as voice channels to provide two-way voice communications. Because each user is assigned a different channel operating at a different frequency, the system is called *frequency division multiple access* (FDMA).

In the simplest case, the desired serving area is partitioned into a number of hexagonal cells of equal size. A base station may be located at the center of each hexagonal cell and provide coverage on the entire cell using an omnidirectional antenna. Alternatively, a base station may be located at each alternate corner of a cell and cover each of the three 120-degree sectors of the cell using a directional antenna. The actual radius of each cell depends upon a number of parameters, one of which is the traffic density. The available voice channels are divided into seven sets[3] in such a way that the

[3]Here the available channels have been divided into seven sets, assuming a cluster of seven cells. As will be shown shortly, the channels could have been divided differently, leading to a different cluster size. For example, clusters of 3, 4, 9, 12, and so on could be used. However, the cluster size of seven has some advantages. They will be discussed later.

separation between any two neighboring channels in any set is as large as possible so that the adjacent channel interference becomes minimum. Each channel set is assigned to one of a cluster of seven cells as depicted in Figure 1-2 and reused in other cells outside the cluster over and over again as shown in Figure 1-3, where each cell is identified by the number of the channel set being used in that cell. Cells that use the same channel set are called *co-channel cells*. In this example, the cluster consists of seven cells. Thus, the co-channel reuse ratio is 7.

To see how the channel sets should be reused, refer once more to Figure 1-3. From the center of a cell, say, cell 2, we go across two cells along vector OA as indicated by $i = 2$ and then one cell along vector AB as indicated by $j = 1$. The cell where we finally land is the co-channel cell with channel set 2. Clearly, for any given cell, there are exactly six co-channel cells. In a general case, for any value of i and j, the distance D between any two neighboring co-channel cells is given by

$$D = \sqrt{i^2 + ij + j^2} \tag{1-1}$$

In terms of the radius of a cell R

$$D/R = \sqrt{3(i^2 + ij + j^2)} \tag{1-2}$$

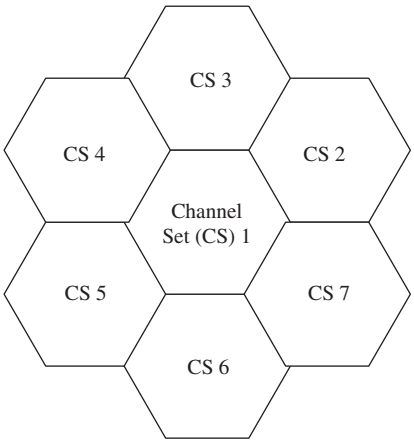


Figure 1-2
Available channels assigned to a cluster of seven cells

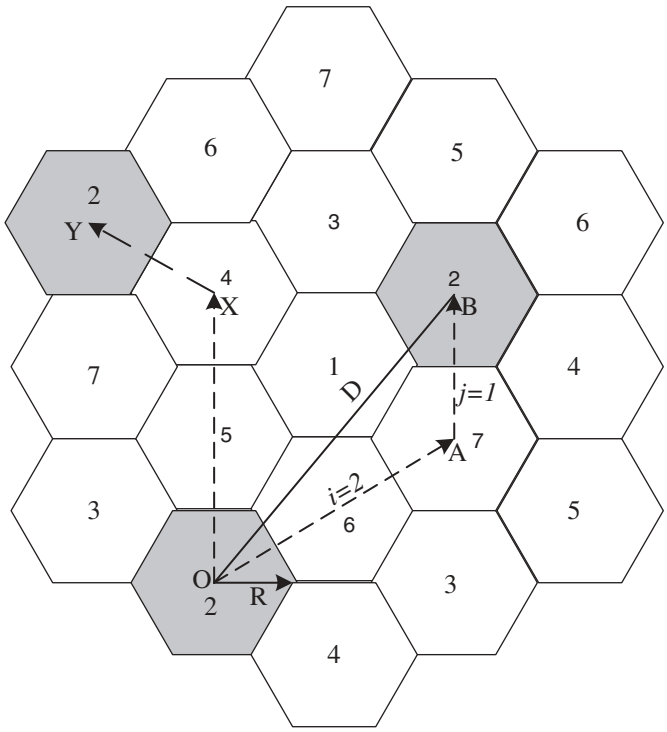


Figure 1-3
Channel reuse in a cellular system

It can be further shown [14] that the co-channel reuse ratio N is given by[4]

$$N = i^2 + ij + j^2 \tag{1-3}$$

Thus, substituting equation 1-3 in equation 1-2, we have

$$D/R = \sqrt{3N} \tag{1-4}$$

With $i = 2$ and $j = 1$, $N = 7$, which is the co-channel reuse ratio that is being used here, and D/R = 4.6. A few other permissible values of N are $N = 3$ with $i = 1$ and $j = 1$, $N = 12$ with $i = 2$ and $j = 2$, $N = 19$ with $i = 3$ and $j = 2$, and so on.

[4]Because the area is proportional to D^2, it is intuitively obvious that N, which is the number of cells in a cluster, would be given by equation 1-3.

The interference experienced by a mobile station from its neighboring co-channel cells is called *co-channel interference*. The signal-to-co-channel interference at a mobile station depends upon the co-channel reuse ratio and the path loss characteristics of the RF signal.[5]

The mobile phone system that was developed in Bell Laboratories using the cellular concept was called *Advanced Mobile Phone Service* (AMPS). It was first commercially deployed by Ameritech in Chicago in 1983. This system, which was subsequently standardized as TIA-553, was based on essentially the same technical specifications and design principles as the development system of the trial phase and used the 40 MHz spectrum allocation.

Later in 1989, the FCC allocated another 10 MHz band. Thus, a total bandwidth of 50 MHz was now available for cellular systems. The spectrum allocation is shown in Figure 1-4. The B bands consisting of subbands B and B′ were provided for use by wire-line service providers such as AT&T, MCI, Verizon, and so on. The A bands consisting of A, A′ and A″ were opened to nonwireline service providers. With a channel spacing of 30 kHz, the number of channels available in either direction is 833. System features are summarized in Table 1-1. The parameters of the table will be discussed later.

---

[5]Assume that the received signal strength varies inversely as the nth power of the distance, that is, $S = k/d^n$, where k is a constant and d is the distance. If the mobile is at the edge of its serving cell, the interference to the mobile due to a co-channel cell at a distance D from the mobile is given by $I = k/D^n$. Because there are six co-channel cells, the *signal-to-interference ratio* (SIR) at the mobile is given by $S/I = (\frac{k}{R^n})/(\frac{6k}{D^n}) = \frac{1}{6}(\frac{D}{R})^n = \frac{1}{6}(3N)^{n/2}$. Here, all base stations are assumed to have the same transmitter power level and antenna gain, among other things. The exponent n depends on the terrain and environmental clutter and may vary from 2 to 5. Assuming $n = 3.5$ and $N = 7$ for a cluster of seven cells, $S/I = 34.33$ or 15.36 dB. The previous expression for the SIR shows that the larger the value of N, the greater the SIR. However, a disadvantage of a large value of N is that now, for a given spectrum allocation, each channel set has fewer channels. As a result, the capacity of a cell (that is, the number of active calls per cell) is diminished. In most cases of spectrum allocation, $N = 7$ gives a fairly good SIR.

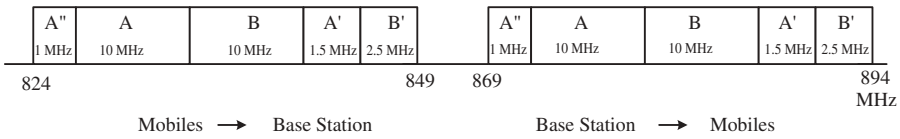


Figure 1-4
The spectrum allocation by the FCC for commercial deployment of cellular systems

Table 1-1
The cellular system features

Multiple Access Scheme	FDMA with FDD
Channel Bandwidth	30 kHz
No. of Users per Channel	1
Speech Signal	Analog, band-limited to 300–3000 Hz
Data Rate	10 kb/s (only for control)
Modulation	FM for speech, FSK for data
Frequency Deviation	12 kHz for speech and 8 kHz for data
User Data Transfer Capability	None

TDMA System

IS-54 and IS-136

In Cellular System TIA-553, where each user is allocated one physical channel with a bandwidth of 30 kHz, about 21 channels are reserved for access and paging, and the remaining 811 are reserved for voice channels. Thus, in a cellular system with a co-channel reuse ratio of 7, each cell has 116 channels (that is, 811/3) or about 39 channels (that is, 116/3) per sector.

Because of the tremendous demand for cellular services, it was soon found that there was a need to increase the capacity of an existing system, or alternatively build new systems with higher capacity. The capacity of an installed system could be increased with the

standard procedure of cell splitting, which indeed was used in many systems.[6] The other way of doing it is to use a *time division multiple access* (TDMA) scheme, where data from multiple users is time-division multiplexed using a number of time slots and sent out over a physical channel.[7] Because each time slot used may be assigned to a different user, in essence, the capacity is increased in the same proportion. Based on this concept, TIA/EIA developed a TDMA standard, called *IS-54* [5], and systems designed to these specifications were introduced in this country in 1993. This standard was eventually superceded by a newer version called *TIA/EIA IS-136.1* and *IS-136.2* [6]. In these specifications, a TDMA frame is 40 ms long and consists of 6 time slots, each 6.67 ms (see Figure 1-5). A full-rate traffic channel contains two equally spaced time slots. For example, it may use time slots 1 and 4, 2 and 5, or 3 and 6, thus in essence assigning each user to two slots. As a result, a 30 kHz wide physical channel that was previously used for a single user can now accommodate 3, thus increasing the capacity threefold. The capacity can be further increased, if necessary, by using lower-bit-rate speech encoders and assigning each user to a single slot instead of two. This method was developed in laboratories but was never commercially deployed.

Clearly, to be able to accommodate multiple users in the same bandwidth, it is necessary to use low-bit-rate coding of speech. The

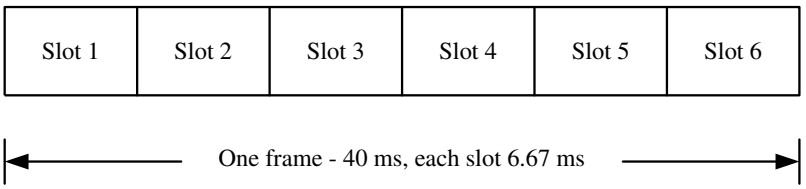


Figure 1-5
A TDMA frame in IS-136

[6]Cell splitting will be discussed later in the book in more detail. For the time being, however, it is sufficient to say that cells may be split by installing a new cell site midway between two existing cells, thus increasing the density of cells and, consequently, the capacity of the system by a factor of four.

[7]Because each channel operates at a different frequency, this scheme is actually a combination of TDMA and FDMA.

System features are summarized in Table 1-2. Obviously, in these digital systems, it is possible to multiplex user data with digitally encoded speech, thus opening up the possibility of providing data services (both basic and enhanced, such as mobile access to the Internet), which would be outside the realm of the older analog systems.

GSM

In Europe, cellular mobile telephony was first introduced in Sweden, Norway, Finland, and Denmark in 1981. These were all analog systems operating at 450 and 900 MHz bands. Over the next few years, many large service providers, such as *Nordic Mobile Telephone* (NMT) and *Total Access Communications Systems* (TACS), installed similar systems in almost every other country of Western Europe. One of the problems with these systems was that they were incompatible with each other and thus did not permit

Table 1-2

The IS-136 system features

Multiple Access Scheme	TDMA
Spectrum Allocation	824–849 MHz uplink
869–894 MHz Downlink	
Channel Bandwidth	30 kHz
Modulation Data Rate on an RF Channel	48.6 kb/s
Modulation	π/4-Shifted DQPSK
No. of Users per Channel	3 for full-rate speech and 6 for half-rate. There are 6 time slots/frame.
Digital Coding of Speech	*Vector Sum Excited Linear Predictive* (VSELP) coder at 7.95 kb/s with 159 bits per 20 ms frame
Channel Coding	Combination of 7-bit CRC and convolutional coding of rate 1/2
User Data Transfer Capability	Limited capability, such as short messages on a *dedicated control channel* (DCCH)

inter-system or international roaming. To overcome this problem, a new standard called *Global System for Mobile Communications* (GSM) was developed in 1990 for next-generation digital cellular mobile communications in Europe. Systems based on this standard were first deployed in 18 European countries in 1991. By the end of 1993, it was adopted in nine more countries of Europe, as well as Australia, Hong Kong, much of Asia, South America, and now the United States.

GSM, like IS-54 and IS-136, combines FDMA and TDMA access schemes and uses 2 frequency bands around 900 MHz [7]. As shown in Figure 1-6, the first band, dedicated to the reverse link, operates at 890 to 915 MHz and the second at 935 to 960 MHz on the forward link. Each physical channel has a bandwidth of 200 kHz and consists of 8 time slots, each assigned to an individual user. Among the features supported by the system are the following:

- Voice, call forwarding, call screening, and call hold.
- Facsimile.
- *Short messaging service* (SMS).
- Circuit-switched data services at rates up to 12 kb/s. The system can support a maximum of 76.8 kb/s data rate by bundling 8 transport channels.
- International roaming.
- Interoperability with ISDN.
- Discontinuous transmissions of mobile stations, thus leading to increased battery life.

In addition to the standard voice telephony, call forwarding, and call screening, the system supports transmission of digital data in the range of 0.3 to 9.6 kb/s transparently using the normal channel coding procedure of the system as well as nontransparently using

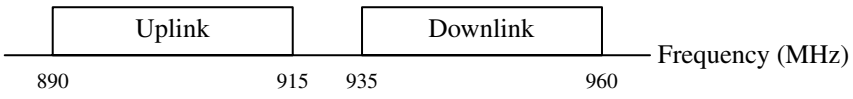


Figure 1-6
The spectrum allocation in the GSM system

Table 1-3 GSM system features

Multiple Access Scheme	TDMA
Spectrum Allocation	890–915 MHz uplink
935–960 MHz Downlink	
Channel Bandwidth	200 kHz
Modulation Data Rate on an RF Channel	270.8333 kb/s
Modulation	0.3 GMSK
No. of Users per Channel	8 for full-rate speech
Digital Coding of Speech	*Regular Pulse Excitation with Long-Term Predictor* (RPE-LTP) at 13 kb/s for full-rate coding
Channel Coding	Combination of block coding and convolutional coding
User Data Transfer Capability	Circuit-switched data up to 12 kb/s and SMS

special coding procedures as required by a user interface. In GSM, a TDMA frame is 4.615 ms long and consists of 8 time slots, each assigned to a user. Thus, the effective bandwidth per user is only 25 kHz (that is, 200/8). Other features include SMS, in which alphanumeric texts of limited lengths are transmitted by base stations along with the regular voice traffic and such supplementary ISDN services as caller identification, call diversion, and so on. The system features are summarized in Table 1-3.

cdmaOne (Based on IS-95-A and IS-95-B)

Concurrently, the application of spread spectrum technology to a mobile communication system was being explored. The feasibility of such a system based on *code division multiple access* (CDMA) scheme was demonstrated in 1998. According to this scheme, each

user is assigned a unique *pseudonoise* (PN) code whose clock rate (that is, the chip rate) is generally much higher than the user data rate. The PN code modulates the user data and the resulting output phase-modulates a carrier. The available spectrum is divided into a number of channels, each with a much higher bandwidth—1.25 MHz—compared to the TDMA systems that were previously discussed. However, the same carrier can now be used in all cells, adjacent or otherwise, and not just in those cells that are outside a cluster as in the cellular or GSM system. In other words, the co-channel reuse ratio is 1. It was found that CDMA systems can provide much larger capacity, more efficient utilization of the spectrum, better speech quality using low-bit-rate linear predictive coders, more robust communication of data services employing efficient channel coding, and much larger bandwidth per channel, thus leading to the possibility, for the first time, of truly multimedia services in wireless networks. With more efficient and dynamic power controls and novel transmission algorithms, transmitter power requirements for the base station or even the mobile station can be minimized. Thus, the handsets could be smaller and more compact in design, resulting in increased battery life. Furthermore, handoff strategies used in a CDMA system provide for a better coverage and lead to an improvement in the system performance. Because of these potential benefits that the system may eventually offer to the end users, this new technology is fast becoming popular. In the United States, there is now the CDMA system called *cdmaOne* based on TIA/EIA specifications IS-95A and IS-95B at Cellular 850 and PCS 1800 MHz bands [8]. Table 1-4 lists the basic features of this system.

Table 1-4 *cdmaOne system features*

Multiple Access Scheme	CDMA, FDD
Spectrum Allocation	Cellular CDMA: 824–849 MHz uplink and 869–894 MHz downlink PCS CDMA: 1850–1910 MHz uplink 1930–1990 MHz downlink

Table 1-4 cont. cdmaOne system features

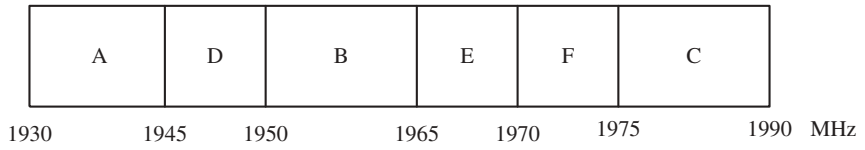
Channel Bandwidth	1.23 MHz
Chip Rate	1.2288 Mb/s
Modulation (for Digital Data)	QPSK and OQPSK[8]
Speech Coding	*Code Excited Linear Predictive Coder* (CELP)—1.2, 2.4, 4.8, 9.6 kb/s for Cellular IS-95 and CELP—14.4 kb/s for PCS IS-95
Number of Users per Channel	~16
User Data Transfer Capability	Packet data at 9.6 and 14.4 kb/s. In IS-95B, higher data rates may also be supported in steps of 8 kb/s.

Personal Communications System

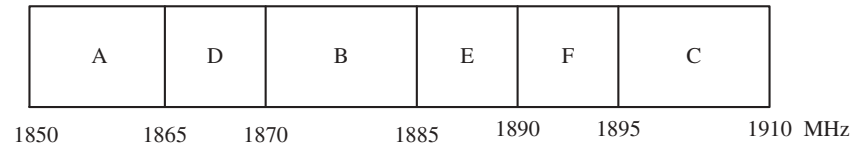
At about the same time, the concept of global *personal communications services* (PCS)—wireless access to *anybody, anywhere, anytime*—emerged. The FCC allocated another spectrum block in the 1.8 to 2.0 GHz band for providing PCS in this country. As shown in Figure 1-7, the spectrum consists of 6 bands with a total of 60 MHz width in either direction with a guard space of 20 MHz in between. Frequency bands A, B, and C are each 15 MHz wide in either direction. Bands D, E, and F have a bandwidth of 5 MHz each. The spectrum in either direction is divided into CDMA channels with a spacing of 50 kHz. Thus, there are altogether 1,200 channels.

A number of technical ad hoc groups were formed by the *Joint Technical Committee* (JTC) to provide enough technical detail to build PCS equipment with different access technologies (that is, IS-95-based CDMA, IS-136 TDMA, GSM, TDMA for *Digital European Cordless Telephone* [DECT] and *wideband CDMA* [W-CDMA]). The standards developed for PCS in North America appear in References [15]–[19].

[8]OQPSK is *offset quadrature phase shift keying*.



Downlink - from Base Station to Mobiles



Uplink - from Mobiles to Base Station

Figure 1-7
The spectrum allocated by the FCC for PCS

Third-Generation (3G) Wireless Technology

As mentioned earlier, the first-generation mobile telecommunication systems to be introduced in the 1980s were analog. These systems, which are still in service, do not have any user data transport capability. To provide data services in these analog systems, a new platform—say, *Cellular Digital Packet Data* (CDPD)—has to be overlaid on the cellular system. However, even this arrangement supports only slow-speed data. The second-generation systems—IS-136, cdmaOne, and GSM—are digital and have data transport capabilities but only to a limited extent. For example, GSM supports SMSs and user data at rates only up to 9.6 kb/s. With IS-95B, it is possible to provide data rates in the range of 64 to 115 kb/s in increments of 8 kb/s over a 1.25 MHz channel. In 1997, to provide for packet mode data services in GSM 2G1, ETSI defined a new standard called *General Packet Radio Service* (GPRS), whereby a single time slot may be shared by multiple users for transferring packet mode data [9], [10]. In GPRS, each slot can handle up to 20 kb/s. Because each user may be allocated up to 8 slots, data rates up to about 160 kb/s per user are possible.

To support high-speed data rates and, more importantly, to be able to provide for multimedia services, the *International Telecommunications Union-Radio Communication Sector* (ITU-R) undertook the task of defining a set of recommendations for *International Mobile Telecommunication in the year 2000* (IMT-2000). Reference [21] gives a historical background on the standardization activities that resulted in the development of many different proposals for 3G radio interfaces and eventually culminated in the selection of a few basic technologies. Briefly, research organizations, equipment manufacturers, and service providers from many different countries of the world started working on different aspects of 3G mobile communications. They developed algorithms and air interfaces, performed simulation, built prototypes, and conducted field tests to verify their validity. 3G partnership projects were established to coordinate the technical activities of various groups and help work out their details. Based on their work, a number of regional standards bodies began to develop the relevant standards. They were

- *Telecommunications Industry Association* (TIA) and T1P1 in the United States. Here, two proposals emerged. One, from TIA TR45.5, is cdma2000. Based upon the direct sequence spread spectrum technology, cdma2000 works in the FDD mode, operates with one or more carriers, and is backwards compatible with the 2G system cdmaOne.

  The other proposal, from TR45.3, is UWC-136, which is wideband TDMA based on recommendations from the *Universal Wireless Communications Consortium* (UWCC) that developed the TDMA standard IS-136 for the United States.
- The *Association of Radio Industries and Business* (ARIB) of Japan. Initially, this organization made a number of proposals, based on W-CDMA, TDMA, and even *Orthogonal Frequency Division Multiplexing* (OFDM) schemes. At the end of the process, however, it submitted only one proposal to ITU that is based on W-CDMA.
- *European Telecommunications Standards Institute* (ETSI)/*Special Mobile Group* (SMG). Here also a number of proposals were initially studied. Eventually, only two proposals were submitted. One is *Universal Mobile Telecommunications*

System (UMTS) W-CDMA FDD, which was actually harmonized with the ARIB proposal. The other is UMTS, based on *time-division, code-division multiple access* (TD-CDMA) principles 17 and operates in the TDD mode.[9]

- *Telecommunications Technology Association* (TTA) of South Korea. Here two proposals were developed—one of them was similar to cdma2000, and the other was similar to the ETSI/ARIB proposal.

Thus, eventually, there were only 4 systems for 3G mobile communications—cdma2000, UWC-136, W-CDMA UMTS FDD, and W-CDMA UMTS TDD. Recommendations on these systems were published by ITU-R as a harmonized standard with four modes in 1999. cdma2000 is required to comply with EIA/TIA IS-41 and W-CDMA UMTS with GSM MAP intersystem networking standards. ITU-R also stipulated that IMT-2000 might provide for other modes as necessary in support of systems that may be developed from time to time around the world with new spectrum allocation.

3G Requirements

3G systems are required to operate in many different radio environments, such as indoor or outdoor, urban, suburban, or rural. The end users may be fixed or moving at various speeds. For example, services may involve:

- Stationary users or pedestrian (0 to 10 km/h)
- Ordinary vehicular applications up to 100 km/h
- High-speed vehicular applications up to 500 km/h
- Aeronautical applications up to 1500 km/h
- Satellites up to 27000 km/h.

[9]In TDD, the same carrier frequency is used in either direction. Information is transmitted in frames, each consisting of a number of time slots, some of which are used for uplink transmissions and the rest for downlink.

The infrastructure used to deliver 3G services may be either terrestrial or satellite based. The information types may include speech, audio, data, text, image, and video [11]. Radio interfaces must be designed to provide voiceband data and variable bit rate services to end users. Both circuit and packet mode data must be supported. The data rates may be

- 144 kb/s or more in vehicular operations[10]
- At least 384 kb/s for pedestrians
- About 2.048 Mb/s for indoor or low-range outdoor applications

Many different cell sizes are permissible in 3G. For example, they could be

- Large or so-called megacells more than 35 km in radius
- Macrocells with a radius of 1 to 35 km
- Indoor or outdoor microcells with a radius of up to 1 km
- Indoor or outdoor picocells with a radius of less than 50 m

3G networks must interoperate with legacy networks, such as a *Public Switched Telephone Network* (PSTN) or *Integrated Services Digital Network* (ISDN) [12], as well as packet-switched public data networks, for example, the Internet.

Some user applications may require bandwidth on demand and a guaranteed *quality of service* (QoS) from networks. Thus, the core network should be capable of reserving resources based on user requests and making sure that all users get the requested quality. 3G standards call for efficient utilization of the spectrum and, in some cases, phased introduction of these services [13]. For example, the data rate supported may be only 144 kb/s in the first phase, 384 kb/s in the second phase, and 2.048 Mb/s in the final phase, all phases being backwards compatible. The goal here is to provide 3G services to users regardless of their locations, in both rural and urban areas, and to support both national and international roaming in a seamless manner. Mobile stations should be able to interwork

[10]It is understood that the system must also be capable of supporting lower data rates (such as 14.4 kb/s, 64 kb/s, and so on) as well.

with different multimedia terminal types that may be used on the fixed side and also connect to other mobile users over satellite links if necessary. Additional services, such as user identity, global position identification, and so on, may also be offered to a customer as available options [12]. Mobile stations could be in different sizes. For example, they could be as small as a pocket radio or large enough to require mounting in a vehicle, and should be able to operate satisfactorily in extreme weather conditions. Open interfaces should be used wherever possible. The service quality to be provided to mobile users is intended to be comparable to that available from a PSTN or an ISDN and should be maintained even when there is more than one service provider in a given serving area [13]. The received speech quality at a mobile station should be equivalent to 32 kb/s *adaptive differential pulse code modulation* (ADPCM). Services should be provided to each user with an acceptable degree of privacy and security that would be at least as good as or better than what is currently available over a PSTN. Finally, the 3G networks should be synergistic with the architecture of the future network.

The 3G standards envisage different types of user traffic. For example, it may be

- Constant bit rate traffic, such as speech, high-quality audio, video telephony, full-motion video, and so on, which are sensitive to delays and, more importantly, delay variations.
- Real-time variable bit rate traffic, such as variable bit-rate encoded audio, interactive MPEG video, and so on. This type of traffic requires variable bandwidths and is also sensitive to delays and delay variations.
- Non-real-time variable bit rate traffic, such as interactive and large file transfers, that can tolerate delays or delay variations.

Some possible applications that appear commercially attractive are

- Conversational voice, video phone and video conferencing, interactive games, and two-way process control, and telemetry information.
- High-speed Internet access applications, such as web browsing, e-mail, data transfer to or from a server (such as a database

download for later analysis), transaction services (that is, e-commerce), and so on.

- Audio streaming, one-way video, still images, large-volume data transfers, and telemetering information for monitoring purposes at an operations and maintenance center.
- Entertainment-quality audio.
- Inquiries/reservation (such as, plane ticket ordering and so on.

Evolution to 3G Systems

One of the goals of 3G standards is to enable the graceful evolution of the current, 2G wireless networks, using as much of the existing infrastructure as possible. The evolution path to 3G is shown in Figure 1-8.

cdma2000 is actually an evolution of cdmaOne. As indicated before, it is a direct sequence spread spectrum system, may use one or more carriers, and operates in the FDD mode. In a multicarrier

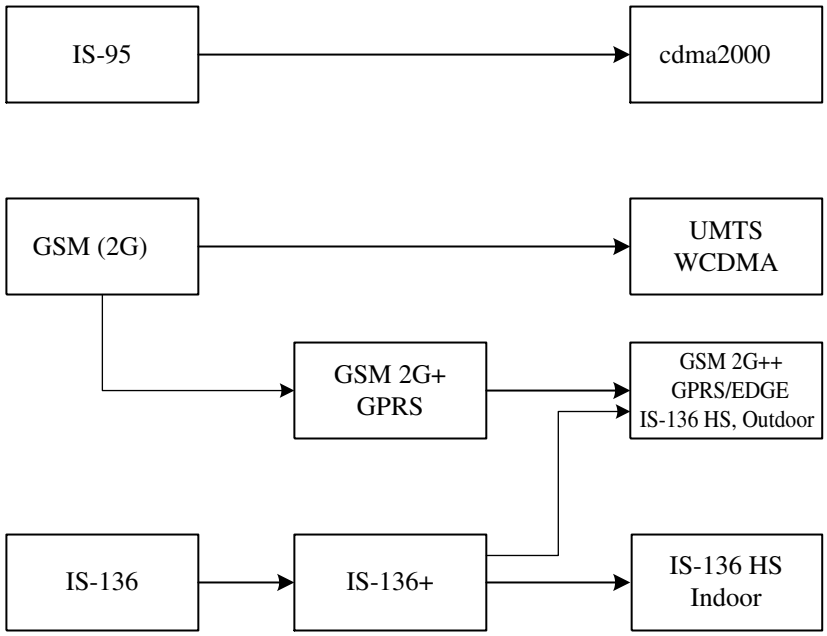


Figure 1-8 The evolution path to 3G systems

system with N carriers (N = 1, 2, or 3), each individual carrier usually has a bandwidth of 1.25 MHz. However, for N = 3, the total bandwidth required is 5 MHz, including the necessary guard bands. To provide for high-speed data services, say, up to 2 Mb/s, a single carrier may have a nominal bandwidth of 5 MHz[11] with a chip rate of 3.6864 Mc/s (that is, 3 × 1.2288 Mc/s). Commercial viability may require the cdma2000 technology to be introduced in different phases. For example, phase 1 may use a single carrier that will support data rates up to 144 kb/s. In phase 2, two more carriers may be added to provide still higher data rates.

Standards have been designed to harmonize core networks of UMTS with those of GSM. Similarly, packet mode data services of UMTS have been harmonized with GPRS, which is a service capability of GSM 2G1. W-CDMA, which is the radio interface of the *UMTS Terrestrial Radio Access* (UTRA), uses a direct sequence spread spectrum on a 5 MHz bandwidth and operates in both FDD and TDD modes.

The TDMA version of the 3G system for use in North America is known as *UWC-136*. As shown in Figure 1-8, its evolution takes place in three phases: IS-1361, IS-136 HS Outdoor/Vehicular, and IS-136 HS Indoor. The first phase, IS-1361, provides voice and up to 64 kb/s data. The per-channel bandwidth is still the same (that is, 30 kHz) as for IS-136. However, to support higher data rates, 8-PSK modulation is used instead of the usual QPSK. The second phase provides data rates up to 384 kb/s for outdoor/vehicular operations, using high-level modulation and a bandwidth of 200 kHz per channel. It should be mentioned here that ETSI has defined a standard called *Enhanced Data Rates for GSM Evolution* (EDGE) to support IP-based services in GSM at rates up to 384 kb/s [20], [21]. IS-136 HS for outdoor/vehicular applications is designed to use this standard in the access network. In the third stage, IS-136 HS Indoor, end users may have a data rate of up to 2 Mb/s with a bandwidth of 1.6 MHz. The spectrum allocation for UWC-136 is the same as for cdma2000.

The system features of UMTS and cdma2000 are summarized in Table 1-5.

[11]Or, if necessary, the bandwidth of a single carrier may be some multiple of 5 MHz.

Table 1-5

System features of UMTS and cdma2000

	W-CDMA (UTRA)	cdma2000
Multiple Access Mode	FDD, TDD	FDD
Spectrum Allocation	FDD mode 1920–1980 MHz uplink, 2110–2170 MHz downlink TDD mode 1900–1920 MHz 2010–2025 MHz	1850–1910 MHz uplink 1930–1990 MHz downlink
Channel Bandwidth	5 MHz	1.25 × N MHz. Initially, *N* may be 1, 2, or 3, but later could be 6, 9, or 12.
Chip Rate	3.84 Mc/s	1.2288 × *N* Mc/s
Frame Structure	10 ms	20 ms
Modulation (for Digital Data)	QPSK	QPSK
Speech Coding	*Adaptive Multirate* (AMR) coding	AMR
User Data Transfer Capability	Circuit mode—up to 144 kb/s, 384 kb/s, and 2.048 Mb/s; packet mode data at least 144 kb/s, 384 kb/s, and 2048 kb/s	144, 384, and 2048 kb/s
3G Network Interface	GSM MAP (evolved version)	ANSI-41 (evolved version)

Summary

This chapter has briefly traced the evolution of mobile communications. A chronology of the important developments is presented in Table 1-6. The first version of cellular telephony to be commercially deployed in the 1980s consisted of analog systems, where frequency modulation is used for analog voice and FSK for signaling and control data. The bandwidth of each channel allocated to an individual

user is 30 kHz. These systems, which had no user data transport capability, were later followed by TDMA systems, where a channel is divided into a number of synchronized slots, each allocated to a single user. The TDMA systems installed in United States are based on standards IS-54 and IS-136, use a channel spacing of 30 kHz, and

Table 1-6

Chronology of important developments in mobile communications

1946	First domestic public land mobile service introduced in St. Louis. The system operated at 150 MHz and had only three channels.
1956	First use of a 450 MHz system. Users had to use a push-to-talk button and always needed operator assistance.
1964	First automatic system, called MJ. It operated at 150 MHz and could select channels automatically. However, roaming was operator-assisted.
1969	First MK system. Like the MJ system, it was automatic, but worked at 450 MHz bands.
1970	FCC sets aside 75 MHz for high-capacity mobile telecommunication systems.
1974	FCC grants common carriers 40 MHz for development of cellular systems.
1978	First cellular system called AMPS was introduced in Chicago on a trial basis.
1981	Cellular systems deployed in Europe.
1983	First commercial deployment of cellular system in Chicago. It is an analog system and does not have a user data transport capability. Analog systems around 450 and 900 MHz band were also introduced in many countries of Europe during 1981–90.
1989	FCC grants another 10 MHz bandwidth for cellular systems, thus giving a total of 50 MHz.
1991	GSM introduced in Europe and other countries of the world.
1993	TDMA system called IS-54 introduced in the United States. SMS available in GSM.
1995	CDMA cellular and PCS technology introduced in the United States.
1997	ETSI publishes GPRS standard.
1999	Standards for 3G wireless services published.

TEAMFLY

provide six slots per frame, eventually tripling the capacity compared to the older analog system. GSM, which is used in much of Europe and many other countries of the world, is also based on the TDMA technology, where each channel has a bandwidth of 200 kHz, and each frame consists of six slots. A distinctive feature of these systems is their support of SMS and circuit-switched user data. An enhanced data service called GPRS is also now available in GSM. CDMA systems, which use direct sequence spread spectrum technology, have been deployed in this country since 1995. Standards for 3G wireless services were published in 1999. Support for high-speed data at rates from 144 kb/s for urban and suburban outdoor environments to 2,048 Mb/s for indoor or low-range outdoor environments is one of the most important features of 3G. Because of the many advantages that it offers, the CDMA technology forms the basis of 3G systems.

References

[1] W.R. Young, "Advanced Mobile Phone Service: Introduction, Background, and Objectives," *Bell Syst. Tech. J.,* Vol. 58, No. 1, January 1979, pp. 1–14.

[2] E.F. O'Neill (ed.), *A History of Engineering and Science in the Bell System.* Indianapolis, Indiana: AT&T Bell Laboratories, 1985, pp. 401–418.

[3] R.F. Rey (ed.), *Engineering and Operations in the Bell System.* Murray Hill, New Jersey: 1984, pp. 516–525.

[4] *High Capacity Mobile Telephone System.* Technical Report Prepared by Bell Laboratories for submission to the FCC, December 1971.

[5] EIA Standard IS-54-B, "Cellular System Dual-Mode Mobile Station—Base Station Compatibility Standard," 1992.

[6] EIA Interim Standard IS-136.2, "800 MHz TDMA—Radio Interface—Mobile Station—Base Station Compatibility—Traffic Channels and FSK Control Channels," 1994.

[7] GSM Specifications 2.01, Version 4.2.0, Issued by ETSI, January 1993. Also, ETSI/GSM Specifications 2.01, "Principles of Telecommunications Services," January 1993.

[8] EIA Interim Standard IS-95, "Mobile Station—Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System," 1998.

[9] GSM Specifications 3.60, Version 6.4.1, "General Packet Radio Service (GPRS); Service Description, Stage 2," 1997.

[10] GSM Specifications 4.60, Version 7.2.0, "General Packet Radio Service (GPRS); Mobile Radio-Base Station Interface, Radio Link Control/Medium Access Control (RLC/MAC) Protocol," 1998.

[11] Recommendations ITU-R M.1034-1, "International Mobile Telecommunications-2000 (IMT-2000)," 1997.

[12] Recommendations ITU-R M.816-1, "Framework for Services Supported on International Mobile Telecommunications-2000 (IMT-2000)," 1997.

[13] Recommendations ITU-R M.687-2, "International Mobile Telecommunications-2000 (IMT-2000)," 1997

[14] V.H. MacDonald, "The Cellular Concept," *Bell Syst. Tech. J.,* Vol. 58, No. 1, January 1979, pp. 15–41.

[15] TR-45.4, Microcellular/PCS.

[16] TR-46, Mobile and Personal Communications 1800.

[17] TR-46.1, Services and Reference Model.

[18] TR-46.2, Network Interfaces.

[19] TR-46.3, Air Interfaces.

[20] E. Dahlman, et al., "UMTS/IMT-2000 Based on Wideband CDMA," *IEEE Commun. Mag.*, September 1998, pp. 70–80.

[21] T. Ojanpera, et al., "An Overview of Air Interface Multiple Access for IMT-2000/UMTS," *IEEE Commun. Mag.,* September 1998, pp. 82–95.

[22] EIA/TIA-553 Cellular System Mobile Station—Land Station Compatibility Specification.

CHAPTER 2

Propagation Characteristics of a Mobile Radio Channel

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

Knowledge of the propagation characteristics of a mobile radio channel is essential to the understanding and design of a cellular system. For example, an appropriate propagation model is required when estimating the link budget or designing a rake receiver for a wideband *Code Division Multiple Access* CDMA system.

There are two types of variations of a mobile radio signal. First, the average value of the signal at any point depends on its distance from the transmitter, the carrier frequency, the type of antennas used, antenna heights, atmospheric conditions, and so on, and it may also vary because of shadowing caused by terrain and clutter such as hills, buildings, and other obstacles. This type of signal variation, which is observable over relatively long distances, say, a few tens or hundreds of wavelengths of the *radio frequency* (RF) carrier, has a log normal distribution and is classified in the literature as a large-scale variation.

The second type of variation is due to multipath reflections. In urban or dense urban areas, there may not be any direct line-of-sight path between a mobile and a base station antenna. Instead, the signal may arrive at a mobile station over a number of different paths after being reflected from tall buildings, towers, and so on. Because the signal received over each path has a random amplitude and phase, the instantaneous value of the composite signal is found to vary randomly about a local mean. A fade is said to occur when the signal falls below its mean level. These fades, which occur roughly at intervals of one-half of a wavelength, may sometimes be quite severe. In fact, fades as deep as 25 dB or more below the local mean are not uncommon. Consequently, a moving vehicle experiences a rapidly fluctuating signal. The rate at which the received signal crosses the fades depends upon the mobile velocity, the RF carrier wavelength, and the depth of the fades. There are other effects due to the motion of the vehicle. For example, if a vehicle moves with a fixed velocity, the power spectrum of the received signal is not constant any more, but varies within a narrow band of frequencies around the carrier. Second, because the in-phase and quadrature components of the fading signal are inherently time varying, the frequency of the received FM signal varies randomly—this is known as *random FM*. Generally, the deeper the fades, the higher its frequency deviation. In fact, this deviation may be much higher than the Doppler shift.

The purpose of this chapter is to summarize the propagation characteristics of a mobile radio channel. We begin with large-scale variations of the signal and consider the effect of terrain and clutter that usually characterize an urban area. Signal variations as a function of the distance, carrier frequency, and antenna heights, as well as the propagation characteristics of suburban and rural areas, will be discussed. Because there is no straightforward relationship between the signal and these factors, path loss models are presented that are based upon empirical relations. The next section deals with short-term variations of the signal resulting from multipath reflections, their effects, coherence bandwidth, and power delay profiles. The chapter concludes with a simulation model of a mobile radio channel in terms of a small number of resolvable paths, each associated with an attenuation and delay that characterize the environment in which the mobile station is operating.

Large-Scale Variations

Signal Variations in Free Space

Consider an ideal, lossless antenna that radiates power equally in all directions. Such an antenna is called *isotropic*. If its input power is P_t, the power density (that is, power per unit area) at a distance r is given by

$$p_i(r) = \frac{P_t}{4\pi r^2} \tag{2-1a}$$

assuming that the medium is the free space and that there is no clutter or environmental obstruction.

For a directional antenna, the power density depends upon the direction. If the direction is such that $p_d(r)$ is the maximum value of the power density, then the antenna gain A with respect to an isotropic antenna is defined as

$$A = p_d(r)/p_i(r) \tag{2-1b}$$

Thus, combining equations 2-1(a) and 2-1(b), $p_d(r)$ is given by

$$p_d(r) = \frac{AP_t}{4\pi r^2} \tag{2-1c}$$

When expressed in dB by taking its logarithm with respect to base 10, the antenna gain is taken to be

$$G_{dBi} = 10 \log(A) \tag{2-1d}$$

In this context, the term *effective isotropic radiated power* (EIRP) of a directional antenna is useful. It is defined as the input power of an isotropic antenna such that the two antennas have identical power densities. In other words, if the directional antenna has an input power P_t and gain A as defined in 2-1(b), then

$$EIRP = AP_t \tag{2-1e}$$

The power P_r received by an antenna depends on the antenna size, that is, the antenna aperture, which in turn is directly proportional to the antenna gain and square of the wavelength. More specifically, using equation 1(c), P_r is given by

$$P_r(r) = \frac{A_t P_t}{4\pi r^2}\left(\frac{A_r \lambda^2}{4\pi}\right) \tag{2-1f}$$

where A_t and A_r are, respectively, the transmitting and receiving antenna gains with regard to an isotropic antenna, and λ is the wavelength of the signal frequency. The term within the parentheses is the effective aperture of the receiving antenna.

There are many other factors that affect the signal attenuation. For example, rain, snow, and other similar atmospheric conditions increase the attenuation. Furthermore, the higher the frequency, the greater the attenuation. The attenuation due to a rainfall rate of 1 mm/hour at 10 GHz is about 0.01 dB/km, whereas it increases to about 5 dB/km for a rainfall rate of 100 mm/hour. Similarly, the attenuation due to a rainfall rate of 1mm/hour at 20 GHz is 0.1dB/km and about 1 dB/km at 100 GHz.

Variations in Urban Areas Due to Terrain and Clutter

In equation 2-1(f), it is assumed that the transmission takes place over the free space and that the received signal is composed of only direct rays between the two antennas. Because in most environments, there are buildings, towers, trees, and hills along the propagation path, there may not be any direct line-of-sight path, and so the signal received at an antenna may not have any direct waves. Instead, it may consist of only reflected rays or possibly a combination of both direct and reflected waves as shown in Figure 2-1.[1]

The propagation characteristics of the mobile radio signal have been extensively studied by a number of authors: [1], [2], and [18]–[20]. For example, Young [18] measured the mobile radio signal in New York at 150, 450, 900, and 3,700 MHz. Okumura et al. [2] measured the signal strength received by a mobile antenna in and around Tokyo in the frequency band from 200 MHz to 1,920 MHz using different base station and mobile antenna heights. Black and Reudink [19] studied the mobile radio signal characteristics at 800 MHz in Philadelphia. Measurements by these and other authors indicate that the signal strength received by a mobile would depend

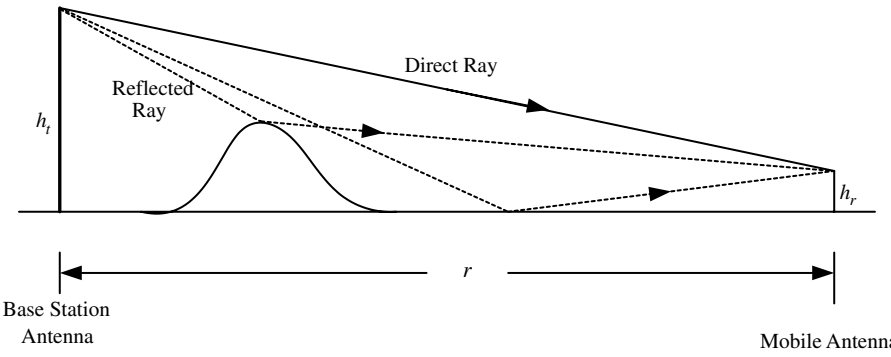


Figure 2-1
Signal propagation between a base station and a mobile

[1]An electromagnetic wave can penetrate an object, entering it at one angle and exiting it at another or bend around an object (such as a hill) due to diffraction. As such, the signal received by a mobile may also include the refracted and diffracted rays.

not only on the transmitter power, the separation distance between the mobile and the base station, carrier frequencies, and antenna heights as discussed previously, but also on the terrain features; environmental clutter such as buildings, tall structures, trees, lakes, or other bodies of water; the width of the streets traversed by the mobiles; the angle at which the signal is incident at the receiving antenna; and the direction in which the vehicles travel with respect to the signal propagation. The terrain may be smooth or quasi-smooth with small undulations, say, on the average of 20 m or so, or it could be quite irregular such as rolling hills, sloping terrain, a mountain range, or an isolated mountain. Sometimes the signal path might include large water bodies such as a sea or a lake. Based on the environmental clutter, a serving area could be urban or dense urban, featuring built-up areas with tall buildings. Similarly, there may be suburban areas with buildings not as tall or dense and rural areas that have very few obstacles except for trees and hills.

The next few subsections describe the effects of the distance, frequency, antenna heights, and other parameters on the received signal for an urban environment. In this description, we have used the results of reference [2] because the general trends in signal variations as shown in [2] are valid for most cities of similar types.

Effect of Distance Figure 2-2 shows signal variations at various distances from the transmitter and at two different frequencies. The values are given relative to the signal level in the free space at a distance of 1 km from the transmitter. The terrain is considered to be quasi-smooth where the average height of the surface undulations is 20 m or less. Also shown in the figure is the signal level in the free space as computed from equation 2-1(a) and 2-1(b).

First of all, notice that the free space signal decreases by 6 dB per octave or 20 dB per decade. Secondly, the difference between the actual signal strength measured in an urban area and the free space signal sharply increases at a distance of about 25 km.[2] It is shown in Reference [2] that for any given frequency, the signal level varies with the distance according to the following empirical relation:

[2]This difference is due to the environmental clutter in the densely built city of Tokyo where Okumura et al. took the measurements.

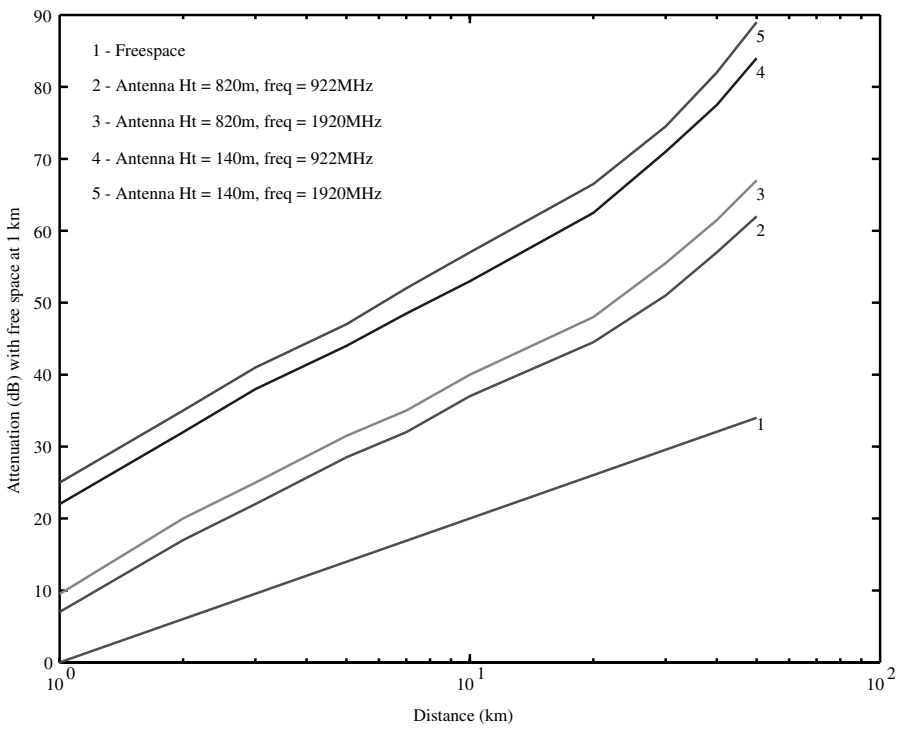


Figure 2-2
The relative signal strength in an urban area as a function of the distance from the base station for two different frequencies

$$P_r = k/r^n \tag{2-2}$$

where k is a constant. In the previous expression, the exponent n is not constant, but varies with the distance itself as well as the antenna heights. For example, with base station antenna heights of 20 to 200 m, the value n for a typical urban area may be in the range of 1.5 to 3.5.

Effect of Frequency The received signal level is also seen to be a function of the frequency, decreasing as the frequency increases. As shown in Figure 2-2, the median value of this decrease varies from about 4.0 dB at distances of 3 km to approximately 7 dB at a distance of about 50 km. In fact, the signal level appears to vary with the frequency according to the relation

$$P_r = k/f^n \tag{2-3}$$

Table 2-1

Values of exponent n in the expression for the received signal as a function of the frequency

Distance from transmitter (km)	Value of n in equation (2-3) at 500–1,000 MHz band	Value of n in equation (2-3) at 1,000–2,000 MHz band
1–20	0.35–0.42	0.5–0.6
20–100	0.42–0.66	0.6–0.8

where k is a constant. Table 2-1 lists approximate values of n for different distances and frequency bands.

Effect of Antenna Heights The received signal level increases with base station antenna heights. See Figure 2-2. This increase in the signal also depends on the distance between the mobile station and base station antennas. For distances of up to 10 km, the signal level increases by about 6 dB/octave. At longer distances, the signal increases by 9 dB/octave if the base station antenna is higher than 200 m or so, but by only 6 dB/octave if the antenna heights are lower. This trend in the signal variation as a function of the base station antenna height is almost independent of the frequency.

The signal level also depends on the mobile antenna height. For example, if the height is increased from 1.5 m to 3 m, the increase in the signal level is about 3 dB [2]. However, this increase is virtually constant at all distances from the base station.

Effect of Other Parameters Other factors that affect the signal attenuation include irregular terrain such as rolling hills, isolated mountains, mixed land-sea paths, tunnels, foliage, bodies of water, and so on, and orientation of the street traversed by a vehicle with respect to a radius from the base station. Although there have been some experimental studies of these parameters, there is no sufficient data to make any conclusive statement about their effects. Okumura et al. suggested some correction factors that can be used to predict the signal level in rolling, hilly terrain. Generally, the signal level decreases as the average terrain undulation height averaged over a

few kilometers increases. For a more detailed description, the interested reader is referred to [1], [2].

Signal Variations in Suburban and Rural Areas

So far, we have only discussed signal variations in urban areas. Because the effect of the environmental clutter in suburban or rural areas is not as severe, the average signal level in these areas is comparatively better. This improvement in the signal levels increases with frequencies, but does not appear to depend on the distance between base stations and mobile terminals or on the antenna heights. Compared to an urban area, the average signal level at 920 MHz is higher by about 10 dB in a suburban area and by 29 dB in rural areas. If the frequency is 1,920 MHz, these improvements are,

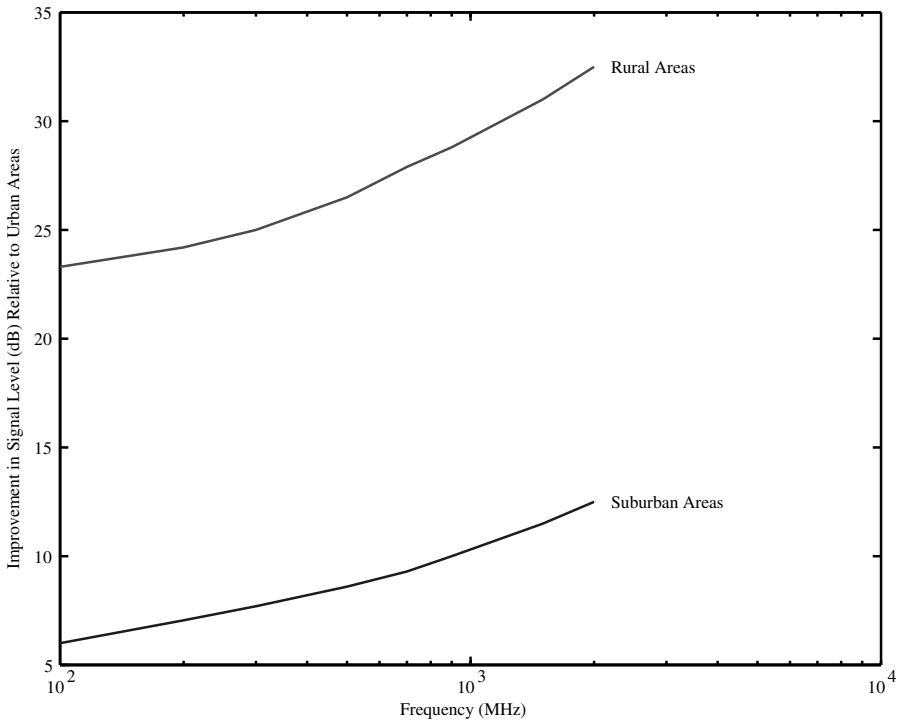


Figure 2-3
Improvement in the signal level in suburban and rural areas over urban areas

respectively, about 12 dB and 32.5 dB. Okumura et. al [2] have suggested using some prediction curves to compute this signal improvement that is statistically valid for most suburban and rural areas. These curves are shown in Figure 2-3.

Variation of the Local Mean Signal Level

It is evident from the previous discussions that even if such factors as the distance from the base station, the antenna heights, frequency, and so on were to remain the same, the local mean signal level[3], because of the environmental clutter, would vary randomly. In fact, this variation is found to have a log-normal distribution.[4] No general characterization can be made of its standard deviation. For the city of New York, it is about 8 dB at a distance of about 2 km from the base station and increases to 12 dB at points farther away from the transmitter. In other cities, it may either decrease with the distance or may not vary with the distance at all but instead may depend only upon the frequency.

In general, then, the received signal at a distance r from the transmitter may be given by

$$P_r(r) = \frac{kA_tA_r}{r^n}P_t \tag{2-4}$$

where k is a constant that depends on the transmitter and receiver antenna heights. The exponent n depends on the environment. Values of n for a few environments are given in Table 2-2.

---

[3]When we talk about the local mean signal, it is understood that variations of the received signal due to fading have been removed by averaging the received signal over a distance of about 10 to 20 m.

[4]The density function of a log-normal variable is given by the following expression:

$$f(x) = \begin{cases} \dfrac{1}{x\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \text{ if } x > 0 \\ 0 \text{ otherwise} \end{cases}$$

where μ is the average value of the variable x and σ^2 is the variance. In other words, the mean signal level when expressed in decibels, has a normal distribution.

Table 2-2

Values of exponent n that determine RF signal attenuation in different environments

Transmission Environment	Values of n in equation (2-4)
Outdoor urban and dense urban areas	3.0–4.0
Indoor urban and dense urban areas	4.5–6.0
Rural areas	2.0–3.0

If P_t is in watts, then taking the logarithm of expression (4), the received power P_r in dB is given by

$$P_r = 10 \log(k) + 10 \log P_t + 10 \log A_t + 10 \log A_r - 10n \log(r)$$

If P_t is in dB, and G_t and G_r are respectively the transmitter and receiver antenna gains in dB, then the previous expression may be rewritten as

$$P_r(\text{dB}) = a + P_t\,(\text{dB}) + G_t + G_r - 10n \log(r)$$

where a is a constant. The received signal power P_r (in dB) may also be expressed in terms of the signal P_{r_0} (in dB) at a reference distance, say, r_0:

$$P_r = P_{r_0} - 10n \log(r/r_0)$$

The reference distance r_0 may be taken as 1 km for cells of an average size.[5] In this case,

$$P_r = P_{r_0} - 10n \log(r) \qquad (2\text{-}5)$$

Notice that if n is assumed to be 4, the average signal at a distance of 10 km is 40 dB below the signal at 1 km.[6] Like the exponent n, sig-

[5]For microcells, it is about 100 m or less. For picocells, it could be a few meters.

[6]For this value of n, if the distance increases by a factor of 10, the signal decreases by 40 dB. In other words, the signal falls at a rate of 40 dB per decade.

nal level P_{r_0} at a reference distance r_0 also depends upon the environment. For rural areas, this value is higher than for an urban environment. As an example, assume that 1w (that is, 30 dBm) is transmitted from a base station antenna. Then typical signal variations as a function of the distance are shown in Figure 2-4.

The actual signal level measured at any point differs from the calculated value using equation 2-1(f). This difference, referred to in the literature as an excess path loss, is also a log-normal distribution. Actual measurements in New York and New Jersey show that for urban areas, the excess path loss has a standard deviation of about 8 to 12 dB for locations about 1 mile from the base station. As we will

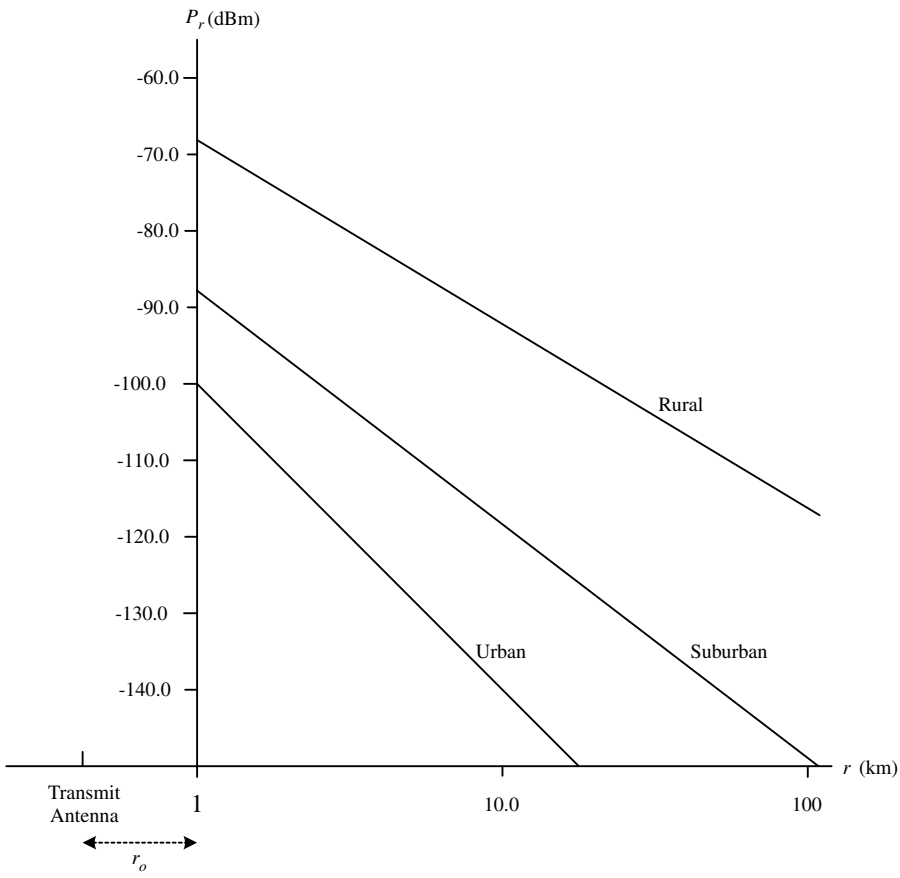


Figure 2-4
An example of variations of the received signal with distance for urban, suburban, and rural areas

see later, these uncertainties in signal levels are dealt with in practice by providing appropriate margins when designing a cellular system.[7]

Propagation Model

As mentioned earlier, the path loss at any point depends on a number of factors, the principal among them being the environmental clutter, the distance from the transmitter, the frequency, the base station antenna height, and, to a much lesser extent, the mobile antenna height. This dependence is usually so complex that it is very difficult to describe it with exact mathematical expressions. However, a number of propagation models based on empirical formulas are available that can be used to estimate the path loss, and consequently, the signal distribution, when designing a cellular network. Using these results, one can then determine the cell size and the number of base stations necessary to provide satisfactory coverage in a serving area.

References [22] to [24] discuss these propagation models in detail. A simple model whose validity appears to be borne out by theoretical studies and practical measurements expresses the path loss at a distance r with respect to the path loss at a distance r_0:

$$P_L(r) = P_L(r_0) + 10n \log(r/r_0) \qquad \text{(2-5a)}$$

If the reference point r_0 is 1 km away from the transmitter antenna, this expression is reduced to

$$P_L(r) = P_L(r_0) + 10n \log(r)$$

where r is in kilometers. The path loss is plotted in Figure 2-5.

[7]For example, when estimating the link budget, a log-normal fade margin of about 10 dB is included to provide a certain level of coverage for 8 dB log-normal standard deviation.

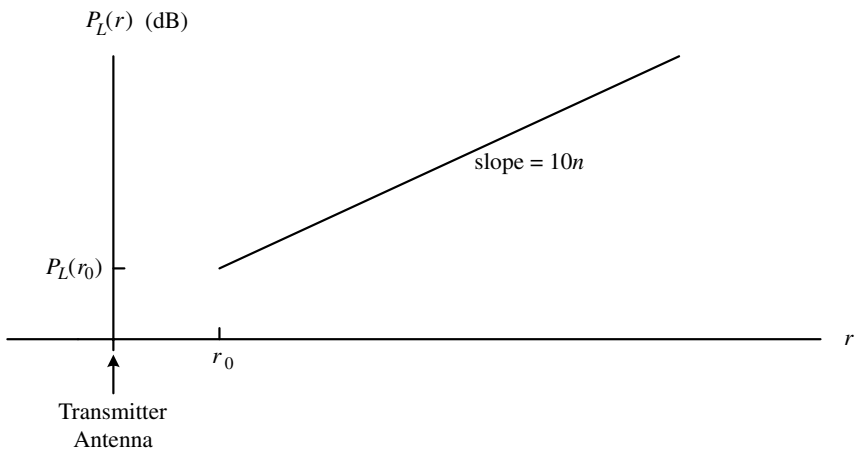


Figure 2-5
A simple path loss model

A similar model, called the *Hata-Okumura model*, is based on actual field strength measurements of Okumura that were previously discussed. As in equation (2-5a), the path loss at any point according to this model is given by

$$P_L = a + b \log r \tag{2-5b}$$

where r is the distance of the point in kilometers from the transmitter, P_L is the path loss, and a and b are constants. These constants depend on terrain characteristics, carrier frequencies, and antenna heights. For example, if the base station antenna height is 50 m and the mobile antenna height 1.5 m, the model gives the following path loss at 900 MHz for a typical urban area:

$$P_L = 123.33 + 33.77 \log r \text{ dB}, r \geq 1 \text{ (km)}, f_c = 900 \text{ MHz} \tag{2-5c}$$

Notice that the path loss at 1 km from the transmitter is 123.33 dB.

Similarly, the path loss for the same antenna heights at 1,900 MHz is given by

$$P_L = 131.82 + 33.77 \log r \text{ dB}, r \geq 1 \text{ (km)}, f_c = 1900 \text{ MHz} \tag{2-5d}$$

The path loss in suburban and open areas is less than in urban areas. For example, at 1,950 MHz, this improvement in path loss is about 12 dB for suburban and 32 dB for open areas.

Short-term Variations of the Signal

As described before, in urban and dense urban areas, there is very often no direct line-of-sight path between a mobile and a base station. In these instances, the signal is composed of a large number of reflected rays because of scattering and reflections from buildings and obstructions. As a result, over short distances, say, of the order of a few wavelengths, the average signal level received at any point remains virtually constant, but its instantaneous value (that is, the envelope of the RF signal) varies randomly about the mean level with a Rayleigh distribution, while its phase is uniformly distributed between 0 and 2π. Because in those cases that are of interest to us, the received signal e at a mobile antenna may consist of a number of randomly varying components; we can represent it as

$$e = x_1 \cos\omega_c t + x_2 \sin\omega_c t \tag{2-6}$$

where x_1 and x_2 are two independent Gaussian random variables with zero mean and equal variance, say, E^2_{rms}, and ω_c is the carrier frequency [21]. The amplitude of e is given by the random variable $z = \sqrt{x_1^2 + x_2^2}$. The variable z, defined this way, may be shown to have Rayleigh distribution [25] with the probability density function given by

$$f(z) = \frac{2z}{E_{rms}^2} e^{-\frac{z^2}{E^2_{rms}}}, z \geq 0 \tag{2-7}$$

Figure 2-6 shows the amplitude variations of a Rayleigh fading signal. As the mobile moves through this signal pattern, the amplitude of the received signal varies, going alternately through the maxima and minima. When the amplitude falls below a given level with respect to its average value, we say that the mobile has gone into a fade.

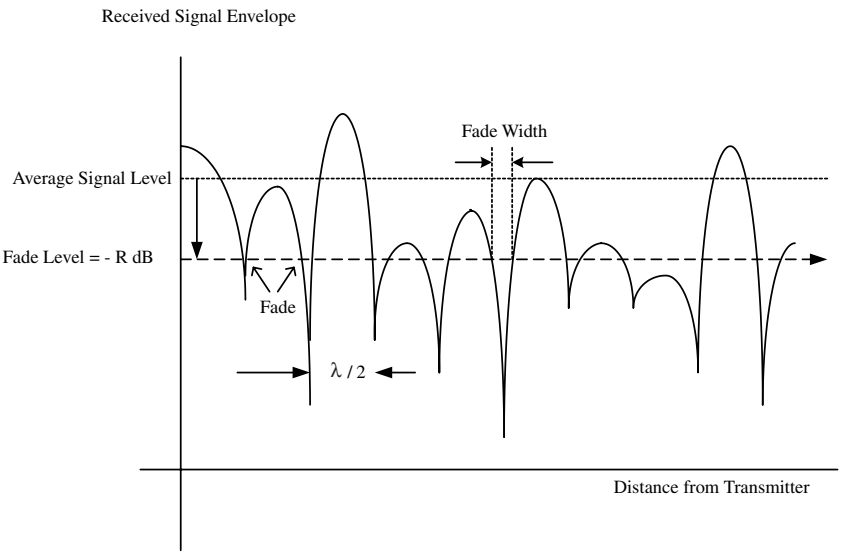


Figure 2-6
A Rayleigh fading signal. The figure shows fades at a signal level R with respect to the average signal. The signal minima occur approximately at one-half the wavelength.

The rate, N_R, at which the instantaneous value of the received signal goes below the level $z = E$ is called the level crossing rate and is given by [21]

$$N_R = \sqrt{2\pi}\, f_d \frac{E}{E_{rms}} e^{-(E/E_{rms})^2} \tag{2-8}$$

where $f_d = \nu/\lambda$ is the Doppler shift due to a mobile velocity ν and carrier wavelength λ.[8]

The average duration of a fade at level $z = E$ is given by [21]:

$$w = \frac{1}{\sqrt{2\pi}\, f_d} \frac{E_{rms}}{E} (e^{(E/E_{rms})^2} - 1) \tag{2-9}$$

Obviously, the level crossing rate, which is the same as the number of fades per second, and the fade duration depends on the fade level, among many other parameters. Table 2-3 shows the number of fades per second at 850 MHz. The average fade duration for the same carrier is given in Table 2-4. Notice that the deeper the fade

[8]The fade level is usually specified in dB. For example, there may be a −10 dB fade. In this case, $20\log(E/E_{rms}) = -10$, or $E/E_{rms} = 0.316228$.

Table 2-3

The number of fades/second at 850 MHz

Vehicle Speed (km/h)	−10 dB fades	−15 dB fades
32	18	10.8
112	64	38.4

Table 2-4

The average fade duration (in ms) at 850 MHz

Vehicle Speed (km/h)	−10 dB fade	−15 dB fade
32	5.31	2.88
112	1.49	0.81

level (that is, the larger the absolute value of R in Figure 2-6) with respect to the average value of the signal, the fewer the number of fades per second and the shorter the fade duration.

Assuming that an unmodulated carrier at frequency ω_c is transmitted, if a mobile moves with a constant velocity v, the power spectrum of the received carrier is no longer confined to ω_c, but is distributed over a frequency band $|\omega - \omega_c| \leq \omega_d$, where, as before,

$$\omega_d = \frac{2\pi V}{\lambda} \cos \alpha \tag{2-10}$$

is the Doppler shift in radians.[9] Here, α is the angle between the direction of the incident signal and the direction of the vehicle motion as shown in Figure 2-7. The power spectral density $S(\omega)$ of the received signal envelope is given by

$$S(\omega) = \frac{E_{rms}^2}{\omega_d} \left[1 - \left(\frac{\omega - \omega_c}{\omega_d} \right)^2 \right]^{-0.5} \tag{2-11}$$

[9]In other words, ω_d is the maximum apparent change in the frequency of the received signal due to the Doppler effect.

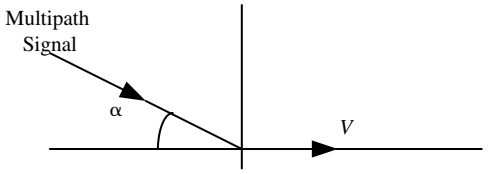


Figure 2-7
The direction of the incoming signal with respect to the vehicle velocity

for $|\omega - \omega_c| \leq \omega_d$ and is depicted in Figure 2-8. Figure 2-8(a) shows the spectrum of an unmodulated carrier transmitted by a base station. Here, all the energy is concentrated in a single frequency ω_c. When this signal is received by a mobile station that is travelling at a velocity ν, the energy is no longer concentrated in the carrier frequency alone but instead distributed over a bandwidth $2\omega_d$ around ω_c as in Figure 2-8(b).[10]

The above analysis assumes that the antenna used is omnidirectional, and that the signal is arriving at the mobile antenna at all

TEAMFLY

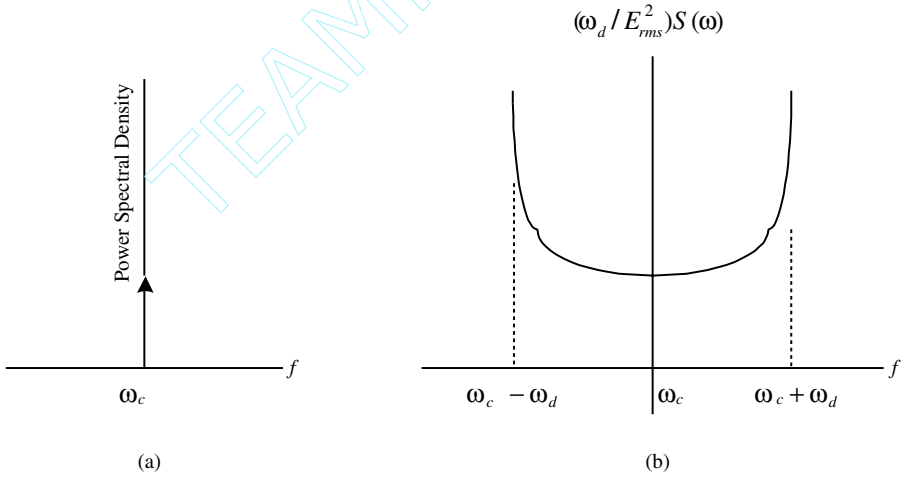


Figure 2-8
Effect of mobility on an RF signal. (a) The power spectrum of an unmodulated carrier transmitted by a base station. (b) The power spectrum of the signal envelope received at the antenna of a moving vehicle.

[10]In practical terms, this means that when an unmodulated carrier is being transmitted, the output at the baseband of a stationary receiver is a low-level thermal noise, which, for an average carrier-to-noise ratio, would be inaudible. If the receiver now moves at a velocity ν, the output at the baseband will be an audible noise with noise power concentrated in a narrow band of frequencies that depends on the velocity and the carrier frequency.

angles with equal probability. As indicated before, E_{rms} is the RMS value of the signal envelope. Equations 2-6 and 2-7 show that random fluctuations of the received signal envelope depend on the vehicle velocity and the carrier frequency and may in some cases be 30 dB or more below the average value of the signal.

In suburban areas, the signal may, sometimes have a direct path and a few indirect paths as well, although possibly not as many as in an urban area. In rural areas, there is often a direct path, and depending on the terrain, the signal may also come over one or two reflected paths. When there is a direct line-of-sight component in addition to one or more reflected rays, the channel is said to be a Rician fading channel.[11]

Effect of Short-term Variations

In a digital system, short-term variations of the RF signal result in burst errors in the received data stream. When the signal goes into a fade, the data bits in the faded portion of the signal are in error with a probability that usually depends upon the fade level. To understand the severity of the burst errors, consider a narrow-band system (such as TDMA based on IS-136 or GSM). Assume that the carrier frequency is 850 MHz and that the mobile velocity is 32 km/h. In this case, the signal falls 15 dB below its local mean about 11 times a second, and each time remains below that level about 3 ms (see Tables 2-3 and 2-4). Hence, the signal is in a fade approximately three percent of the time.[12] Thus, for a −15 dB fade, assuming that the probability of a faded bit being in error is 0.5, the burst errors during a fade will result in a bit error rate of 0.015. Because wideband CDMA systems are inherently less susceptible to fades, the burst error rates due to the vehicle motion are less severe in

[11]The signal with Rician distribution is given by

$$e = \mathrm{Re}[(x_0 + x_1 + jx_2)e^{-j\omega_c t}]$$

where x_0 is a constant line-of-sight component, and x_1 and x_2 are two independent Gaussian random processes as in equation 2-6.

[12]That is, during each second, the signal is in a fade $3 \times 11 = 33$ ms.

these systems.[13] Also, because a mobile radio channel is time varying, and because the signal at a mobile antenna is subjected to a Doppler shift that varies with the mobile velocity, the signal frequency as perceived at a mobile station goes through a random variation with time. This phenomenon, which is known as *random FM*, manifests itself as additional noise in the baseband, which in a digital system may cause the bit error rate to increase even more.

Coherence Bandwidth and Power Delay Profiles

To understand coherence bandwidth, assume for simplicity that signals are being transmitted at two different frequencies over a mobile radio channel. If the difference between the two frequencies is small, the two signals fade in the same way. In other words, their short-term variations (with respect to time) have identical statistical properties and consequently are correlated to each other. As the separation between the frequencies is increased, they begin to fade differently, and if the separation is wide enough, the signal variations become statistically independent. The separation between the two frequencies that the signal variations are correlated below is called the *coherence bandwidth*. When the channel bandwidth is less than the coherence bandwidth, the resulting fading is called *flat fading*. If the channel bandwidth is any greater, it leads to a frequency-selective fading.

Suppose that a base station transmits a single pulse on a coherent channel as shown in Figure 2-9. As we have seen earlier, this results in a number of reflected rays, each of which travels to a mobile station along a different path. Because of attenuation, only a finite number of rays have sufficient energy to be useful at a mobile station. In Figure 2-9(b), only three reflected waves are shown to have any significant amount of energy, the last two being delayed with respect to the first by, respectively, τ_1 and τ_2. The variation of the sig-

[13]See Chapter 3 "Principles of Wideband CDMA."

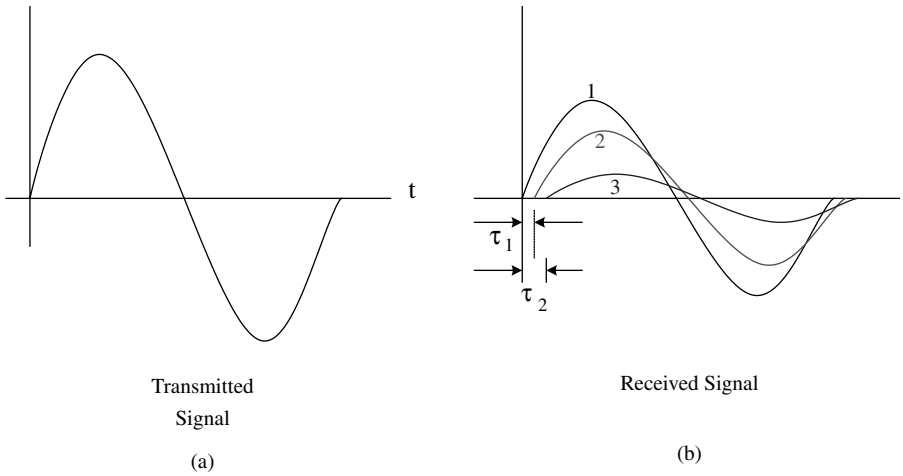


Figure 2-9
Signal propagation on a coherent channel.
(a) Transmitted signal that consists of a single pulse.
(b) The received signal consists of three waves that have any significant amount of energy.

nal power as a function of the delays associated with the reflected waves is called the *power delay profile*. The maximum delay over which the signal energy is non-negligible is called the *delay spread*. In Figure 2-9(b), the delay spread is τ_2.

Because the statistical properties that characterize the signal variations of the reflected waves in Figure 2-9(b) are the same, they cannot be combined in a diversity receiver to cancel or reduce fading. On the other hand, in a wideband CDMA system, where the channel bandwidth (1.25 to 5 MHz) is usually much higher than the coherence bandwidth of most urban areas (about 100 to 200 kHz), each of the reflected waves that have any significant energy, with delays that are multiples of the chip periods (that is, 0.2713 to 0.8138μs), can be coherently combined in a rake receiver to provide diversity.

The delay spread and the coherence bandwidth are obtained by measuring the power delay profile. Because the autocorrelation function of a received signal is a measure of its energy content, the delay profile may also be looked upon as an auto-correlation function [5].

The delay profile in any region depends upon its terrain and environmental clutter [4], [5], [7]–[9]. Reference [1] provides detailed mathematical treatment of the delay spread, signal envelope correlation, and phase correlation as a function of the separation frequency.

Cox and Leck [4], [5] have studied the delay spreads, excess delays,[14] and distributions of delay spreads in selected areas of New York City and have found that at 910 MHz, the maximum delay spread is about 3.5 μs. The delay spread exceeds 3 μs in 5 percent of the areas where measurements were taken, 2.5 μs in about 10 percent of the areas, and 1.2 μs in 50 percent of the region. Determination of the cumulative distribution from these values indicates that the delay spread approximates a log-normal distribution with a mean of 1.3 μs and a standard deviation of 0.6 μs. The excess delay, on the other hand, is found to have a mean of 1.1 μs, standard deviation of 0.9 μs, and a maximum value of about 4 μs, and exceeds 2 μs in about 10 percent of the areas observed. Evidently, the delay spread is smaller in suburban and rural areas and possibly the smallest in those areas that have a direct line-of-sight path as in a satellite communications system [17]. Table 2-5 gives average delay spreads for typical urban, suburban, rural, and in-building communications channels.

Reference [5] gives measured power delay profiles of a few places in New York City. These profiles are generally not smooth functions of the delay variations. However, to further explain our ideas, some simple, idealized power delay profiles are shown in Figure 2-10. In Figure 2-10(a), the maximum delay that might be considered in the design of a rake receiver is 4 μs. However, most of the signal energy is concentrated in multipaths that have delays of 2.0 μs or less.

Table 2-5

Delay spreads for typical urban, suburban, rural, and in-building channels

Environment Class	Average Delay Spread (μs)
Urban area	2.5
Suburban area	0.5–1.5
Rural/open area	0.1–0.25
In-building	$<$0.1

[14]The excess delay is the average delay spread (obtained by taking the first moment of the power delay profile) less the first arrival delay.

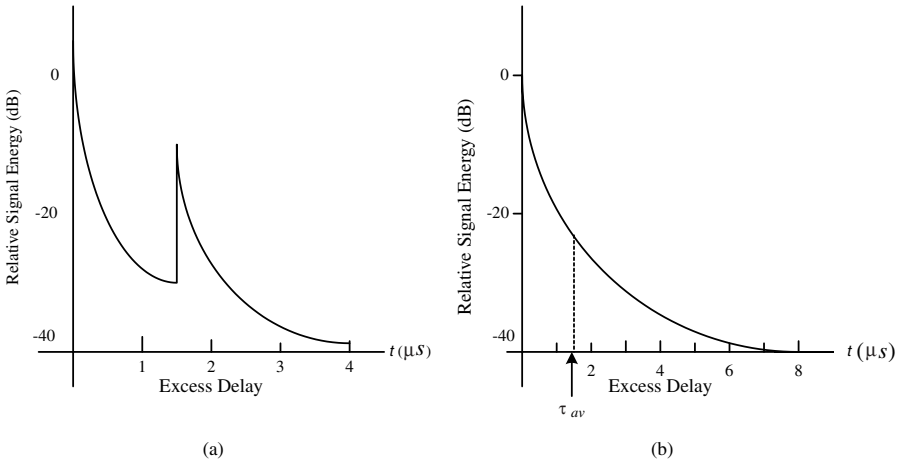


Figure 2-10
Two idealized power delay profiles showing the signal energy as a function of the delay relative to a direct ray

A number of empirical formulas based upon field measurements have been used to express the coherence bandwidth [5], [23]. As an example, the following expression defines this bandwidth $B_{coherence}$ in terms of the average delay:

$$B_{coherence} = 0.16/\tau_{av}$$

where τ_{av} is the average value of the delay spread. If Figure 2-10 (b) is the power delay profile of a certain urban area, then because $\tau_{av} = 1.5\ \mu$s, the coherence bandwidth is

$$B_{coherence} = 0.16/\tau_{av} = 106.67 \text{ kHz}$$

Simulation Model of a Mobile Radio Channel

In many instances, it is useful to simulate multipath signals in the laboratory because in the absence of such simulation, one would have to actually build the system, test it in the field to see if it actually works satisfactorily in a mobile environment, adjust the design parameters if necessary, retest, and repeat the process until the desired performance is achieved.

A number of procedures for simulating a mobile radio channel have been suggested by various authors [11]–[17]. The procedure presented here is based on the fact that the power delay profile of a mobile radio environment may be resolved into a small number of multipaths that have significant energy [3].

A block diagram of the simulation model is shown in Figure 2-11. The signal on each path i is attenuated by $e^{-\alpha_i}$ (where α_i is the attenuation in dB) and delayed by τ_i second. The first branch represents a direct line-of-sight path between the transmitter and receiver antennas. The delayed signal from each of the other paths is passed through a Doppler shaping filter. The frequency response of this filter, which is shown in Figure 2-12, is designed to approximate the theoretical spectrum of the received signal envelope of Figure 2-8(b) [11]. Notice that the filter response increases to a peak value at the Doppler frequency $\omega_d/2\pi$ [11] and then falls off at 18 dB/octave. Signals from all branches are summed together and then combined with additive white Gaussian noise. The resulting output simulates the signal received over a mobile radio channel. The number of paths, the attenuation, and delay of each are obtained from the power delay profile.

Reference [3] discusses five mobile radio channels corresponding to the following environments: urban fast, urban slow, rural, terrain-obstructed, and satellite mobile channels. In most urban areas, there is no direct path, and as such their simulation excludes the first

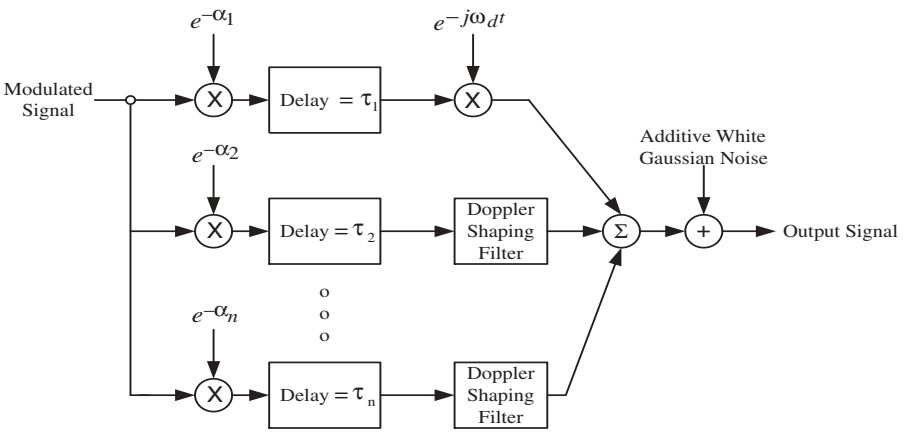


Figure 2-11
A simulation model of the Rayleigh fading channel. The Doppler shaping filter has a Doppler spread of $\omega_d/2\pi$ Hz where $\omega_d = 2\pi V/\lambda$. Here V is the vehicle speed, and λ is wavelength of the carrier frequency.

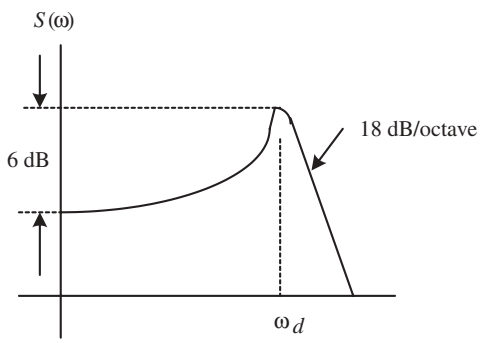


Figure 2-12
The frequency response of the Doppler shaping filter

branch of Figure 2-11. The number of paths to be included depends on the environment.[15] For example, nine paths, as shown in Table 2-6, have been used in simulating a mobile radio channel in a typical urban area for *Digital Audio Radio* (DAR) systems. For the representative power delay profiles of other environments, see Reference [3].

Table 2-6
The power delay profile for an urban area used in testing DAR systems

Path #	Attenuation (dB)	Delay (μs)
1	2	0
2	0	0.25
3	3	0.5
4	4	0.75
5	2	1.0
6	0	1.25
7	3	2.0
8	5	2.5
9	10	3.0

[15]A satellite channel usually consists of a line-of-sight ray, but may include one or two multipaths as well.

Summary

In this chapter, we have described the propagation characteristics of a mobile radio channel. There are two types of variations experienced by an RF signal—the large scale variations observed over relatively long distances, say, a few tens or hundreds of wavelengths of the carrier, and short-term variations that result from multipath reflections and occur over very short distances, such as a fraction of the wavelength. Factors responsible for large-scale variations include clutter and environmental obstructions, distance from the transmitter, the carrier frequency, antenna heights, and so on. Signal variations as functions of these parameters are described in terms of field measurements and empirical relations. The mean signal level at any point obtained by averaging it over a small area around that point varies randomly with a log-normal distribution. Simple models are presented that can be used to compute the path loss of a mobile radio channel. Short-term variations have a Rayleigh distribution and may cause the instantaneous value of the signal envelope to vary by as much as 30 to 40 dB about the local mean. The effect of these variations, the coherence bandwidth, and power delay profiles are discussed. We conclude the chapter with a simulation model of a mobile radio channel. The model consists of a small number of resolvable paths, each associated with an attenuation and delay that characterize the particular propagation environment—urban, suburban, or rural—that is being simulated.

References

[1] W.C. Jakes, Jr. (Ed.), *Microwave Mobile Communications*. New York: John Wiley, 1974.

[2] Y. Okumura, et al., "Field Strength and Its Variability in VHF and UHF Land-Mobile Radio Service," *Review of the Electrical Communication Laboratory*, Vol. 16, No. 9–10, September–October, 1968, pp. 825–873.

[3] L. Thibault, G. Soulodre, and T. Grusec, "EIA/NRSC DAR Systems Subjective Tests Part II: Transmission Impairments,"

IEEE Trans. Broadcasting, Vol. 43, No. 4, December 1997, pp. 353–369.

[4] D.C. Cox and R.P. Leck, "Distributions of Multipath Delay Spread and Average Excess Delay for 910-MHz Urban Mobile Radio Paths," *IEEE Trans. Ant. & Prop*., Vol. AP-23, No. 2, March 1975, pp. 206–213.

[5] D.C. Cox and R.P. Leck, "Correlation Bandwidth and Delay Spread Multipath Propagation Statistics for 910-MHz Urban Mobile Radio Paths," *IEEE Trans. Comm*, Vol. COM-23, No. 11, November 1975, pp. 1271–1280.

[6] R.D.J. van Nee, et al., "Rician Fading Land-Mobile Satellite Channel," *IEEE J. Sel. Areas Comm.,* Vol. 10, No. 2, February 1992, pp. 350–357.

[7] A. Afrashteh, et al., "Performance of a TDM/TDMA Portable Radio Link for Interference, Noise, and Delay Spread Impairments," *IEEE Trans. Veh. Tech*., Vol. 43, No. 1, February 1994, pp. 1–7.

[8] M.J. Feuerstein, et al., "Path Loss, Delay Spread, and Outage Models as Functions of Antenna Height for Microcellular System Design," *IEEE Trans. Veh. Tech*., Vol. 43, No. 3, August 1994, pp. 487–497.

[9] E.S. Sousa, et al., "Delay Spread Measurements for the Digital Cellular Channel in Toronto," *IEEE Trans. Veh. Tech.,* Vol. 43, No. 4, November 1994, pp. 837–847.

[10] N.L.B. Chan, "Multipath Propagation Effects on a CDMA Cellular System," *IEEE Trans. Veh. Tech.,* Vol. 43, No. 4, November 1994, pp. 848–855.

[11] G.A. Arredondo, et al., "A Multipath Fading Simulator for Mobile Radio," *IEEE Trans. Comm.,* Vol. COM-21, No. 11, November 1973, pp. 1325–1328.

[12] J.I. Smith, "A Computer Generated Multipath Fading Simulation for Mobile Radio," *IEEE Trans. Veh. Tech.,* Vol. VT-24, No. 3, August 1975, pp. 39–40.

[13] M.R. Karim, "Packet Communications on a Mobile Radio Channel," *AT&T Tech. J*., Vol. 65, Issue 3, May/June 1986, pp. 12–20.

[14] C. Loo, et al., "Computer Models for Fading Channels with Applications to Digital Transmission," *IEEE Trans. Veh. Tech.*, Vol. 40, No. 4, November 1991, pp. 700–707.

[15] C. Loo, "A Statistical Model for a Land Mobile Satellite Link," *IEEE Trans. Veh. Tech.*, Vol. VT-34, No. 3, August 1985, pp. 122–127.

[16] G.E. Corazza, et al., "A Statistical Model for Land Mobile Satellite Channels and Its Applications to Nongeostationary Orbit Systems," *IEEE Trans. Veh. Tech.,* Vol. 43, No. 3, August 1994, pp. 738–741.

[17] E. Lutz, et al., "The Land Mobile Satellite Communication Channel—Recording, Statistics, and Channel Model," *IEEE Trans. Veh. Tech.*, Vol. 40, No. 2, May 1991, pp. 375–385.

[18] W.R. Young, Jr., "Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 MC," *Bell System Tech.* J. 31, November 1952, pg. 1068.

[19] D.M. Black and D.O. Reudink, "Some Characteristics of Radio Propagation at 800 MHz in the Philadelphia Area," *IEEE Trans. Veh. Tech.*, Vol. 21, May 1972, pp. 45–51.

[20] D.O. Reudink, "Comparison of Radio Transmission at X-Band Frequencies in Suburban and Urban Areas," *IEEE Trans. Ant. Prop.*, AP-20, July 1972, pg. 470.

[21] W.C. Jakes, Jr., "A Comparison of Specific Space Diversity Techniques for Reduction of Fast Fading in UHF Mobile Radio Systems," *IEEE Trans. Veh. Tech.*, Vol. VT-20, No. 4, November 1971, pp. 81–92.

[22] M. Hata, "Empirical Formula for Propagation Loss in Land Mobile Radio Services," *IEEE Trans. Veh. Tech.*, Vol. 29, May 1980.

[23] K. Garg, *IS-95 and cdma2000.* New Jersey: Prentice Hall, 2000, pp. 232–243.

[24] J. Lee and L. Miller, *CDMA Systems Engineering Handbook.* Artech House, 1998.

[25] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1965.

TEAMFLY

CHAPTER 3

Principles of Wideband CDMA (W-CDMA)

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

IMT-2000 has defined four 3G systems, only one of which, namely, UWC-136 is based upon the *time division multiple access* (TDMA) scheme, while the other three—*Universal Mobile Telecommunications System* (UMTS) W-CDMA *Frequency Division Duplex* (FDD), UMTS W-CDMA *Time Division Duplex* (TDD), and cdma2000—use *direct-sequence code division multiple access* (DS-CDMA).[1] In fact, CDMA appears to be the preferred technology for wireless communications because of the many advantages it offers including, for example, the multipath diversity and soft handoff [1], [31]. The purpose of this chapter is to provide the reader with a basic understanding of the principles of wideband CDMA technology. The chapter begins with an overview of the various access technologies. It next considers a wideband CDMA system architecture and presents some technical detail of its various physical layer components.

Multiple Access Schemes

UMTS W-CDMA FDD is a direct-sequence CDMA system with a nominal bandwidth of 5 MHz [16], [19], [20]. The second system, UMTS W-CDMA TDD, also uses CDMA with a bandwidth of 5 MHz, but now the frequency band is time shared in both directions—one half of the time, it is used for transmission in the forward direction and the other half of the time in the reverse direction. cdma2000 is a multicarrier, direct-sequence CDMA FDD system. Like cdmaOne, its first phase is expected to use a single carrier with a bandwidth of 1.25 MHz. In the second phase, it may have as many as three carriers. In this system, even though each carrier has a nominal bandwidth of 1.25 MHz, the total bandwidth required is 5 MHz. The fourth system supported by IMT-2000 is based upon a TDMA scheme whereby each physical channel is divided into a number of fixed, synchronized time slots. Each user is assigned one or more slots and is permitted to transmit its information only during its

[1]To be more accurate, however, W-CDMA UMTS TDD uses a combination of TDMA and CDMA schemes.

allocated period. These multiple access techniques are not new in communication systems. In fact, spread spectrum systems, of which the direct sequence spread spectrum using code division multiple access technique is one variation, have found its application in military systems for a long time. A number of different multiple access schemes have been used in conventional and wireless local area networks [2]–[4], [30].

Historically, the most common multiple access techniques are: *Frequency Division Multiple Access* (FDMA), TDMA, and CDMA.

FDMA

In this scheme, the available spectrum is divided into a number of smaller bands of equal bandwidths, each of which is assigned to a different user for the duration of a call. A guard band must be provided between any two adjacent bands, or *channels* as they are called, for satisfactory operation of the system. This is depicted in Figure 3-1(a). Figure 3-1(b) shows a transmitter using the FDMA

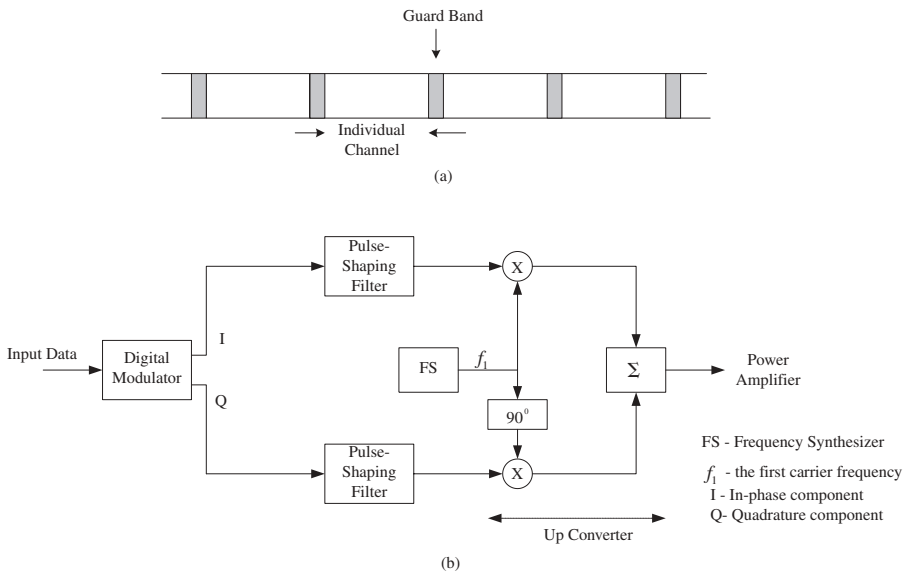


Figure 3-1
An FDMA system: (a) The available bandwidth is divided into a number of smaller bands with guard bands between them. (b) Functional block diagram of a generic transmitter in an FDMA system.

technique. An FDMA system would actually consist of many of these transmitters, each using a different carrier frequency. Pulse-shaping filtering of the incoming data is usually done at the base band. An example of this access mechanism is cellular system TIA-553, where each channel has a bandwidth of 30 kHz. Separate channels are used on uplinks and downlinks.

TDMA

In TDMA [5], information from multiple users is sent out in fixed-size frames, each consisting of a number of equal time slots. Any given user may be assigned one or more of these slots and is allowed to transmit only during the allocated slot(s) in each frame using the entire bandwidth of the channel.[2] A guard period must be provided between any two adjacent time slots so that transmissions from different users do not overlap. The access scheme is illustrated in Figure 3-2.

Digital systems, such as the North American standard IS-54, its newer version IS-136, and *Global System for Mobile Communications* (GSM), are all based on TDMA and FDD. In IS-54 and IS-136, each voice channel has a bandwidth of 30 kHz and consists of six time slots. For full-rate speech, three users use these six slots in a certain way. For example, user 1 uses time slots 1 and 4, user 2 time slots 2 and 5, and user 3 time slots 3 and 6. For half-rate speech, one user is assigned to each of the six time slots. Because the same bandwidth is now being time shared among three and possibly six users, the system capacity increases three to six times.

In GSM, on the other hand, each physical channel has a bandwidth of 200 kHz. Each frame consists of eight time slots. As such, the system can support up to eight users per channel with full-rate speech. Each TDMA frame is 4.615 ms. Each slot is 0.577 ms, of which the guard period is 0.03462 ms. In these TDMA systems, physical channels use different frequencies as in FDMA. Thus, in essence, they combine TDMA and FDMA schemes.

[2]This bandwidth is not to be confused with the available spectrum, which is divided into a number of separate channels for use in the TDMA system.

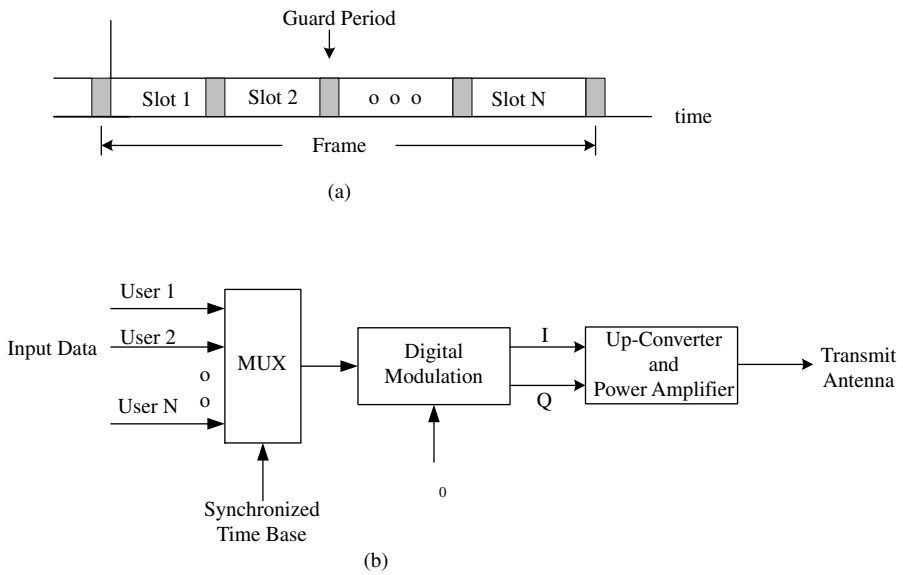


Figure 3-2
A TDMA system: (a) A TDMA frame (b) A TDMA transmitter

Spread Spectrum Multiple Access

As we have just seen, each user in FDMA or TDMA is allocated a fraction of the available bandwidth. In the spread spectrum multiple access scheme, on the other hand, all users can transmit simultaneously on the entire available bandwidth using a pseudorandom code that is unique for each user. These codes, which are also known as *pseudonoise* (PN) codes, are random sequences generated by means of a multistage shift register, where some selected outputs are added modulo 2 and fed back to the input of the shift register. As will be shown later in this chapter, the code sequences repeat themselves after a finite, although usually quite long, period and behave as random functions for all practical purposes. The receiver separates the different users by correlating the received signal with these codes.

Many spread spectrum techniques are currently available. For example, there are *direct sequence spread spectrum* (DSSS), *frequency-hopping spread spectrum* (FHSS), *time-hopping spread spectrum* (THSS), and hybrid techniques, which are combinations of the

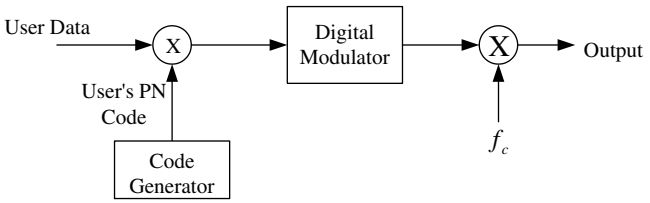


Figure 3-3 Principles of a DSSS

first three. Because UMTS W-CDMA and cdma2000 use direct sequence spread spectrum, only this access scheme will be explained in this section.

In a direct-sequence spread spectrum system [6], [7], also known as the CDMA, each user is assigned a unique PN code. Its data stream is first spread out by that PN code, and then modulates the carrier frequency. The clock rate of the spreading code is known as the *chip rate*. This principle is illustrated in Figure 3-3.

The chip rate of the PN code is usually much higher than the user data rate. The ratio of the chip rate to the data rate is called the *spreading factor*. In UMTS, the spreading factor varies from 4 to 256.

CDMA Technology

Direct-Spread CDMA Principles

As will be seen later, PN codes have some unique properties. One of them is that any physical channel or user application, when spread by a PN code at the transmitter, can be uniquely identified at the receiver by multiplying the received baseband signal with a phase-coherent copy of that PN code. To illustrate how a CDMA receiver can detect the signal from a desired user in the presence of signals received from other users in a CDMA system, consider Figure 3-4 (a), which shows the block diagram of an overly simplified CDMA receiver. Suppose that the receiver wants to detect the data stream

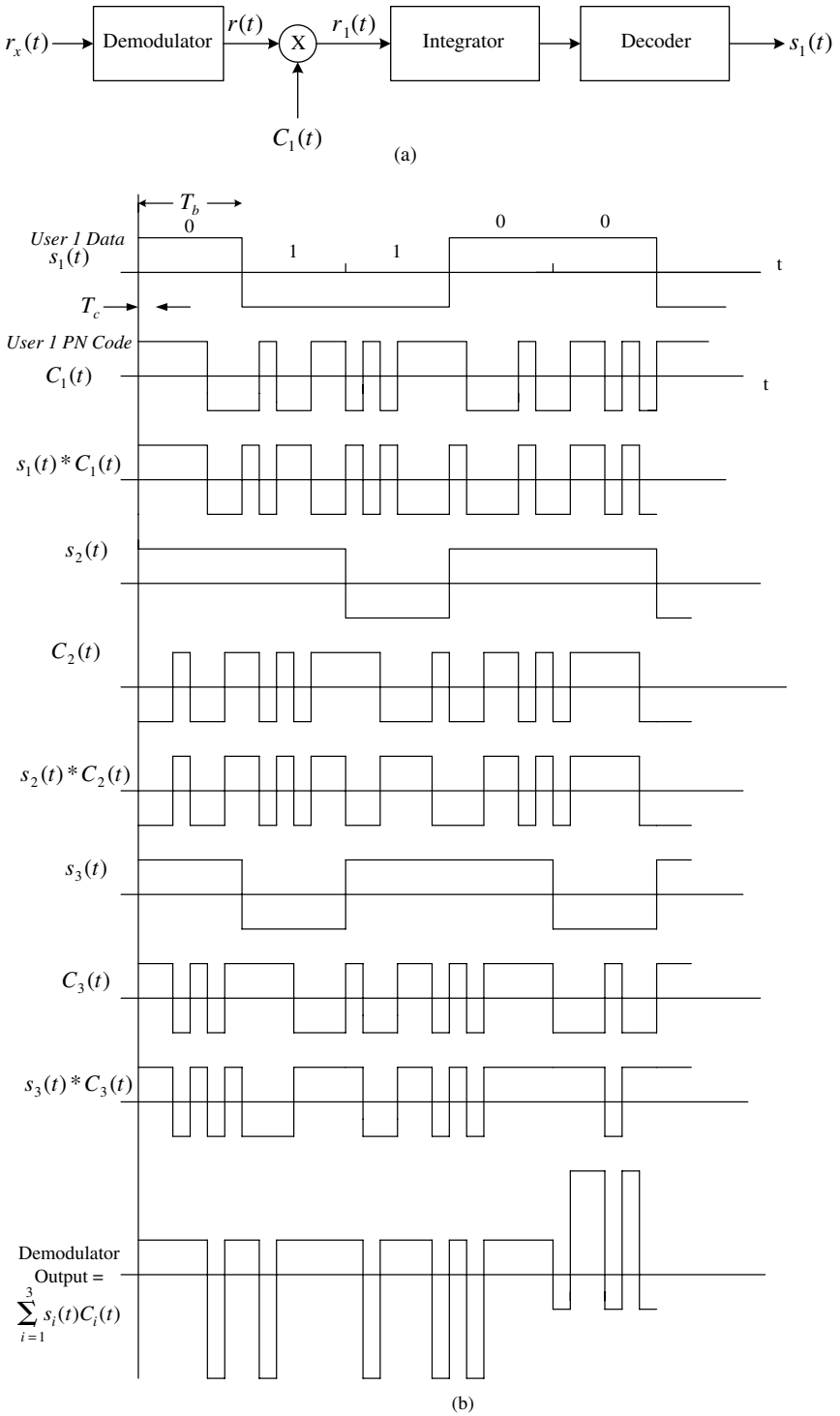
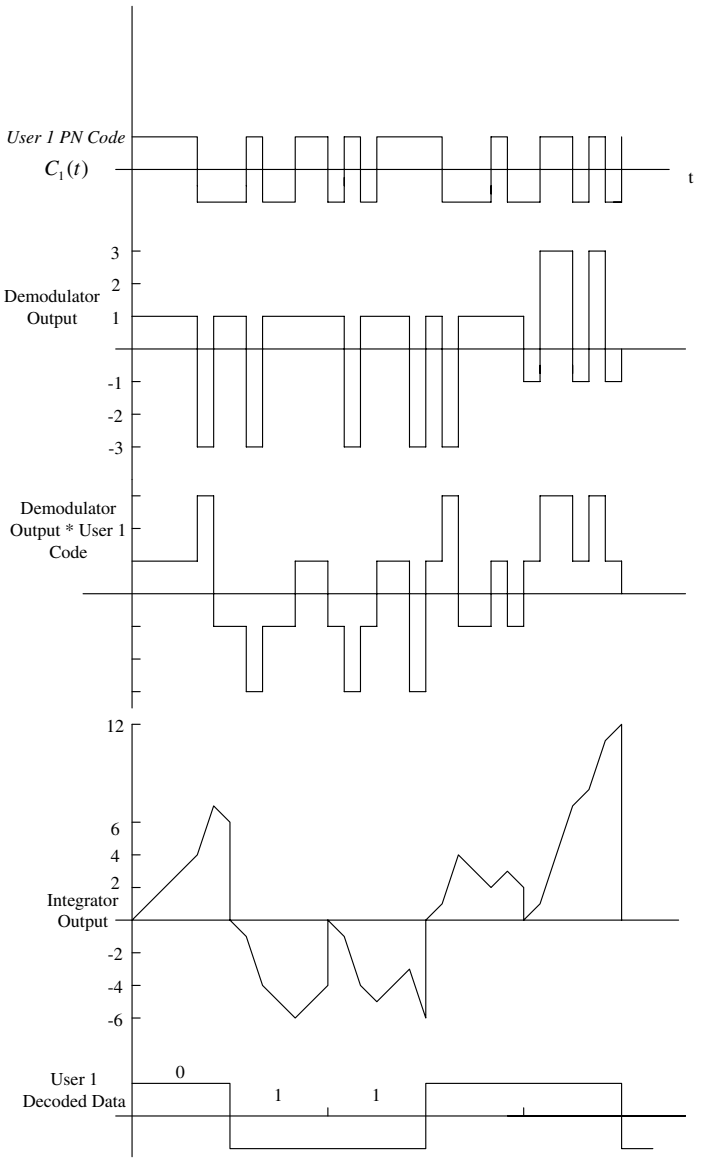


Figure 3-4
A CDMA system: (a) A simplified CDMA receiver. The received signal at the input of the demodulator is composed of signals from multiple users. The data stream from user 1 is being detected in this figure. (b) Diagrams showing how the signal arriving at a receiver is composed of the transmitted signals from multiple users. (c) Diagrams illustrating how the signal from a desired user can be detected. The decoder reads the integrator output at the end of each symbol period, and if it is positive, it takes the data to be a binary zero. If it is negative, the data is decoded to be a binary one. Notice that after the decoder has read the integrator output, the integrator must (cont.)

Figure 3-4 cont. be reset (that is, its output must be dumped) so that the process can recommence at the start of the next symbol period.



from user 1. The received signal from multiple users is first demodulated. The output of the demodulator, which is a baseband signal, is multiplied by the PN code assigned to user 1. The resulting output is applied to the input of an integrator where it is integrated over each symbol period. The decoder reads the output of the integrator and

decodes it into binary data, following certain rules. The result is the recovered data from user 1.

To see that this indeed is the case, assume that the data stream from any user is represented by $s_i(t)$ and its associated PN code by $C_i(t)$. The output at the transmitter after spreading is $\nu_i(t) = s_i(t)*C_i(t)$. Notice that in $s_i(t)$ or $C_i(t)$, the signal level is either $+1$ or -1, with $+1$ representing a binary 0 and -1 a binary 1. If the noise introduced by the channel is negligible, the demodulated signal at the baseband is given by

$$r(t) = \sum_{i=1}^{N} s_i(t)*C_i(t)$$

where N is the number of users in the system. If $r(t)$ is now multiplied by a copy of the PN code $C_1(t)$ of user 1, the resulting output is given by

$$\begin{aligned} r_1(t) &= C_1(t)*r(t) = C_1(t)*\sum_{i=1}^{N} s_i(t)*C_i(t) \\ &= s_1(t)*C_1(t)*C_1(t) + s_2(t)*C_2(t)*C_1(t) + s_3(t)*C_3(t)*C_1(t) + \ \ldots \end{aligned}$$

Because the cross-correlation between $C_1(t)$ and $C_2(t)$ is very small, the second term appears as noise so that when it is integrated over a symbol period, the output of the integrator due to this term is virtually zero. The same is true of the third and following terms. However, the output of the integrator due to the first term, when averaged over a symbol period, is $s_1(t)$ because

$$C_1(t)*C_1(t) = 1$$

These ideas are illustrated in the time diagrams of Figure 3-4(b) and (c).

Capacity of a CDMA System

Consider a single cell CDMA system where a number of mobiles are simultaneously transmitting at the same frequency. Here, each mobile is assigned a unique PN code sequence. Let

P = carrier power,
E_b = Energy per bit
B_c = spread spectrum signal bandwidth

f_{data} = information bit rate

I = power due to interference

N_o = noise power per bit

Then,

$$E_b = P/f_{data}$$

$$E_b/N_0 = \frac{P}{N_0 f_{data}} \tag{3-1}$$

$$N_0 = I/B_c$$

So,

$$E_b/N_o = \frac{P}{I}\frac{B_c}{f_{data}} = \frac{P}{I} \times G_p$$

Here G_p is the RF bandwidth divided by the information bit rate. In the CDMA system being discussed here, the signal is *Quadrature Phase Shift Key* (QPSK)-modulated, where the RF bandwidth is approximately equal to the chip rate. In other words, if f_{chip} is the chip rate, then the RF bandwidth $B_c = f_{chip}$, and in that case $G_p = f_{chip}/f_{data}$ is called the *process gain*. For a given bit error rate, E_b/N_0 is fixed. Consequently, the greater the process gain, the larger the allowable interference (that is, I/P) for that bit error rate.

If there are N transmitters, all transmitting at the same power and using the same chip rate, then

$$I = (N - 1)P$$

So, using equation (3-1),

$$I/P = \frac{(N-1)P}{P} = N - 1 = \frac{G_p}{E_b/N_o}$$

Or,

$$N = 1 + \frac{G_p}{E_b/N_o} \approx \frac{G_p}{E_b/N_o} \tag{3-2}$$

for large values of N.

Notice that for a fixed bit error rate (that is, a fixed value of E_b/N_0), the greater the process gain, the larger the capacity N of the system. Similarly, with a fixed process gain, the capacity increases if the value of E_b/N_0 required to provide a satisfactory operation decreases.

The capacity given by the previous equation is achieved only under ideal conditions. In actual practice, it may be significantly less for a number of reasons. For example, the capacity will decrease if the power control is not perfect. Similarly, in a multicell system, where each cell operates at the same frequency, transmissions in other cells may cause the interference to be increased by 60–85 percent.

Because the system is interference limited, the capacity of the system can be increased by reducing the interference. There are a number of ways of doing this. First, the interference due to other users can be reduced by replacing an omnidirectional antenna with a directional one. For example, a 3-sector antenna would increase the capacity by a factor of about 2–3.

Second, human conversation is characterized by talk bursts followed by silence periods. If the transmitter is turned off during these silence periods, the interference to other transmitters will decrease, and consequently, the overall system capacity will increase. Thus, actual capacity may be given by

$$N = 1 + \frac{G_p}{E_b/N_o}\frac{\alpha}{(1+\beta)v} \tag{3-3}$$

where α is the correction factor due to imperfect power control, β is the effect of co-channel interference from other cells in a multicell system, and ν is the voice activity factor. Table 3-1 gives some typical values of these parameters.

As an example, suppose that $\alpha = 1$ (that is, perfect power control), $\nu = 0.4$, $\beta = 0.85$ for a 3-sector cell, data rate = 9.6 kb/s for an 8 kb/s vocoder, and chip rate = 1.2288 Mc/s. The required $E_b/N_0 = 7$dB. So, the value of $E_b/N_0 = 10^{0.7} = 5.01$. $G_p = 1.2288 \times 10^6/9600 = 128$. So, $N = 1 + (128/5.01)(1/1.85)(1/0.4) = 35$. This capacity is also known as the *sectorized pole capacity*.

Notice that the capacity can be increased by simply reducing E_b/N_0, but that would result in increased bit error rates for all users. On the other hand, it is possible to minimize E_b/N_0 without

Table 3-1

Typical values of parameters that affect the system capacity

Parameter	Average Values
Power control correction factor, α	0.5–1.0
Voice activity factor, ν	0.4–0.6
Effect of co-channel interference from other cells in the system, β	0.5–0.9. A typical value for a 3-sector cell is 0.85. For an omnidirectional antenna, it is 0.6.

necessarily running the risk of increasing the bit error rate. One way to do this is to select an appropriate modulation technique. For example, if the desired bit error rate is 10^{-5}, the required E_b/N_0 is 12.6 dB with *Binary Frequency Shift Keying* (BFSK), whereas it is only 9.6 dB for *Binary Phase Shift Keying* (BPSK) or QPSK using coherent detection [9].

Because the bit error rate increases as the signal-to-interference ratio is minimized, it is necessary to use an error-correcting code. The coding that is normally used in CDMA and W-CDMA systems is convolutional coding where it is possible to achieve a coding gain of 4–6 dB with hard decision sequential and soft decision Viterbi decoding. Thus, the capacity of a CDMA system can be increased by using channel coding.[3]

It is interesting to know the minimum *signal-to-noise ratio* (SNR) that one can possibly use. The maximum attainable data rate $R_{\max}$ on a channel with infinite bandwidth in the presence of Gaussian noise is given by Shannon's channel capacity theorem:

$$R_{\max} = \frac{P}{N_0 \ln 2} \tag{3-4}$$

Comparing equations (3-1) and (3-4), we see that the minimum SNR is $\ln 2 = 0.693$, that is, $10 \log(0.693)$ or -1.6 dB. The maximum data rate is determined not only by this SNR but also the transmitter power.

[3]This is true even though the use of a channel code leads to a somewhat increased channel rate.

3G Radio Transmitter Functions

To understand the technology used in the implementation of the physical layer functions of a typical W-CDMA system, consider Figure 3-5, which shows a simplified block diagram of the transmit functions of a multicarrier cdma2000 base transceiver station.

The incoming data stream (destined to each user) is encoded into a CRC code and passed through a convolutional encoder of

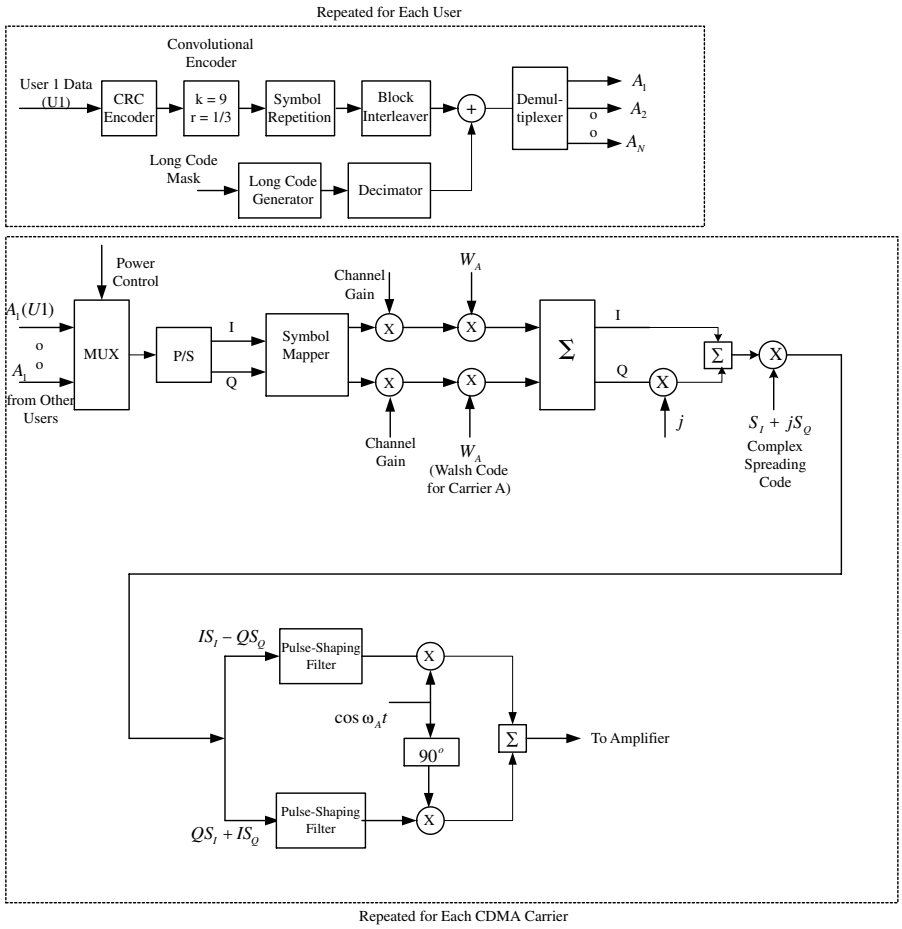


Figure 3-5
A simplified block diagram of the transmit functions of a base transceiver station of a multicarrier cdma2000 system

constraint length 9 and rate 1/3.[4] Depending upon the data rate, the output of this coder may have to be repeated a few times in the block marked *Symbol Repetition* so that the resulting output matches the physical channel rate.

The output of the symbol repetition block is applied to an interleaver that spreads out in time adjacent bits of the input stream to provide protection against burst errors. A long PN code that is unique for each user scrambles the output of the interleaver and applies the scrambled sequence to a demultiplexer where it is broken into N subsequences, where N is the number of CDMA carriers. Each of these subsequences is transmitted over a separate CDMA carrier as shown in the lower half of the diagram. In other words, data is transmitted in parallel over multiple CDMA carriers. The chip rate used is $N \times 1.2288$ Mc/s. The value of N may be 1, 3, 6, 9, or 12. However, standards currently specify $N = 1$ and 3 only.

Subsequences (shown as A_1, A_2, . . .) are multiplexed with the power control bits, converted into a parallel form, and then split into I and Q streams corresponding to the *in-phase* (I) and *quadrature* (Q) component of the transmitted signal. Each bit of the I and Q stream is mapped into a BPSK symbol: a zero into $+1$ and a one into -1. The symbols of the I and Q branches are multiplied by a gain factor and spread by a Walsh code, say, W_A, which is different for each CDMA carrier. As described later, these codes are constructed with the elements of a row of an orthogonal matrix, whose entries are either $+1$ or -1 so that when channels are spread with different Walsh codes, they become mutually orthogonal.

The I and Q symbols after Walsh spreading are added in quadrature to form complex symbols that are again spread (that is, multiplied) by a complex PN code $S_I + jS_Q$, where S_I and S_Q are, respectively, the cell-specific I-channel and Q-channel pilot PN sequence. The I and Q components of the output from complex spreading are passed through a pulse-shaping filter and modulate the desired CDMA carrier shown as ω_A in Figure 3-5.

[4]In IS-95, the convolutional code used has a constraint length 9 and rate 1/2.

Basic ideas behind the system components, such as a channel encoder, an interleaver, PN code sequences, and so on, are presented in the following section.

Speech Encoding

Speech encoders used in different mobile communications systems are listed in Table 3-2. UMTS uses *Adaptive Multirate* (AMR) coding of speech based on the principles of *Algebraic Code Excited Linear Prediction* (ACELP) [43]–[46]. Eight encoded bit rates are supported: 12.2 (GSM enhanced full rate), 10.2, 7.95, 7.40, 6.70, 5.90, 5.15, and 4.75 kb/s.

ACELP coders belong to the vocoder class of encoders that, unlike a waveform quantizer, model the vocal tract as a time-varying digital filter such that when it is excited with an appropriate input, the output is a desired speech signal. The filter coefficients are determined

Table 3-2

Speech encoders used in mobile communications systems

Mobile Communications System	Speech Coding Algorithm
IS-54/IS-136	*Vector-Sum Excited Linear Predictive Coding* (VSELP). The bit rate is 7.95 kb/s. The coder operates on 20-ms frames. Its output consists of 159 bits per frame.
European Telecommunications Standards Institute (ETSI)/GSM Standard 06.xx	13 kb/s *Regular Pulse Excitation with Long Term Predictor* (RPE-LTP). Every 20 ms, the speech encoder generates 260 bits.
cdmaOne based on IS-95A and IS-95B	*Code Excited Linear Predictive Coding* (CELP). The bit rate may be 9.6, 4.8, 2.4, or 1.2 kb/s.
UMTS	AMR based upon ACELP. There are eight possible bit rates varying from 4.75 kb/s to 12.2 kb/s. At 12.2 kb/s, the output of the coder is 244 bits for every 20-ms frame.

by analyzing the speech input. The input to the filter is selected from two code books that contain excitation signals. The output of the synthesis filter is compared with the incoming speech, and the difference signal is minimized by choosing the best excitation signals from the code books. The resulting pitch, gain, and code indices together with the filter coefficients (rather than the speech samples themselves) are transmitted to the far end. The receiving end, which also maintains two similar code books, decodes these indices to derive an excitation signal and feeds it into a *linear predictor* (LP) synthesis filter that is constructed using the received filter coefficients.

Figure 3-6(a) shows the block diagram of the encoder. It consists of two code books—a fixed or algebraic code book containing a small number of predefined excitation vectors (that is, nonzero pulses) and an adaptive code book containing excitation vectors, which are adapted during every 5 ms subframe of the incoming speech signal.[5]

The speech signal at a mobile station is sampled at 8000 Hz, each sample consisting of 13 bits. Samples are read into the encoder in 20 ms frames and are analyzed to derive the short-term coefficients of an LP filter of order 10.[6] Because speech is encoded in *public switched telephone networks* (PSTNs) using the 8-bit A-law (or μ-law) companding schemes, the speech signal received from a PSTN is first converted into 13-bit uniform *pulse code modulation* (PCM), and then applied to the input of an ACELP encoder.

Excitation vectors from the code books are multiplied by gains, summed together, and then fed into the LP synthesis filter. For each 5 ms subframe, the code books are searched for vectors to minimize the mean squared error using a perceptual weighting filter where various formant frequencies are weighted according to their contribution to the perceptual quality of the speech. Because each of the codec parameters—the pitch, gain, code indices, and filter coefficients—affects the speech quality differently, they are encoded in different error-protecting codes before they are transmitted over the channel. At 12.2 kb/s, a 20 ms frame at the output of the encoder

[5]The pitch, gain, and code are together known as a vector.

[6]In other words, each frame contains 160 samples.

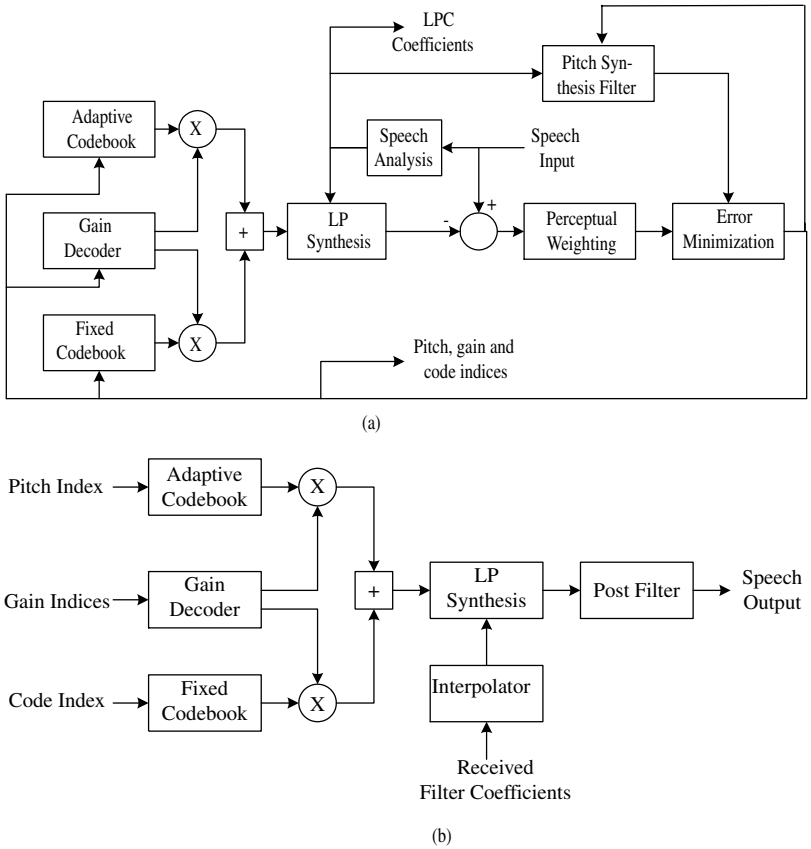


Figure 3-6 Codecs used in UMTS. The coding scheme is based on ACELP: (a) An ACELP encoder (b) An ACELP decoder at the receiving end.

includes 38 bits of LP parameters, 30 bits of pitch delay, 16 bits of pitch gain, 140 bits of algebraic codes, and 20 bits of code book gain.

Figure 3-6(b) shows the block diagram of the ACELP decoder at the receiving end. The received filter coefficients are interpolated every subframe before synthesizing speech at the receiver.

Channel Coding

Convolutional Encoder

In W-CDMA, user data and voice are encoded into convolutional codes before they are transmitted. These codes are different from

block codes not only in the way they are generated at the transmitter but also in the way they are usually decoded at the receiver. Consider, for example, an *(n,k)* block code of length *n* with *k* message bits and *n-k* check bits. When decoding this block code, it is necessary to first retrieve a block of *n* bits from a synchronous frame, and then apply the decoding procedures to that block. A block code is memoryless because decoding any given code block does not require the knowledge of any previous blocks. In convolutional codes, on the other hand, there is no concept of a block; decoding can be done with any finite or semi-infinite sequence of input data [8]–[11]. However, when encoding the incoming data stream, only a finite number of the past data bits will determine the output. Thus, in both the encoding and decoding of these codes, it is necessary to save some past data bits in memory. In this section, we shall define this code and describe the commonly used decoding procedure.

A convolutional encoder is shown in Figure 3-7(a). It consists of three one-stage shift registers; the first registers holds the current bit, and the last two contain the past two bits of the incoming data. Input data is shifted one bit at a time. For each input bit, there are two bits of the output that are obtained by adding modulo 2 outputs of certain registers. Given the input stream 0110011 . . . , the two output streams are respectively 0100110 . . ., and 0010101 . . . . Because each bit of the input stream generates two bits of the output, the code is rate ½. The constraint length of the coder is equal to the length of the registers. Here, the constraint length is 3. Notice that in our example, a 2-bit shift register would have been sufficient because the first bit is the current bit of the incoming data stream.

In general, there could be any number of registers, say *m*, each of length *k* bits as shown in Figure 3-7(b). The input is shifted k bits at a time. As there are n output bits for each set of k bits at the input, the code is of rate $^{k}/_{n}$ and constraint length mk. In Figure 3-7(a), k = 1, n = 2 and m = 3.

The two generating functions of the code of Figure 3-7(a) are

$$g_1(x) = 1 + x + x^2, \text{ and}$$

$$g_2(x) = x + x^2$$

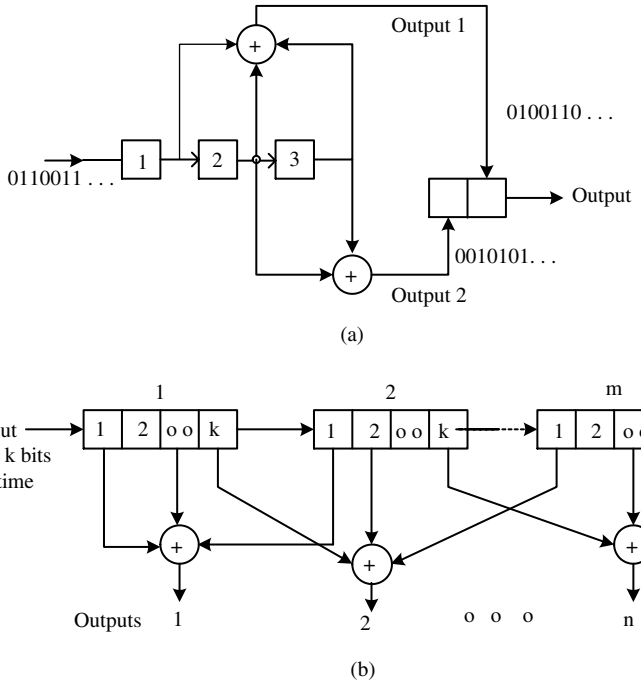


Figure 3-7
Convolutional encoders: (a) An encoder of rate $^1/_2$ and constraint length 3. (b) A more general encoder structure. The code is of rate $^k/_n$ and constraint length mk.

The encoder output for any input can be determined from a sequential state machine. The states of the machine are given by the contents of the last k-1 stages of the shift register that save the past k-1 inputs. In our figure, because only two bits of the shift register save the past data, the state machine has four states. The output at any instant is determined by the current input and the state of the state machine at that instant. Figure 3-8 shows the states of the encoder of Figure 3-7a. Beside each path is a number of the type a/bc, which means that the input that causes the state transition is a and the resulting outputs are b and c.

To explain this figure, assume that at t = 0, the encoder is in state 00. If the first bit of the incoming data is 0, it is shifted into bit position 2 of the shift register at the end of the bit period, and so the state of the machine still remains at 00. If, instead, the input bit is 1,

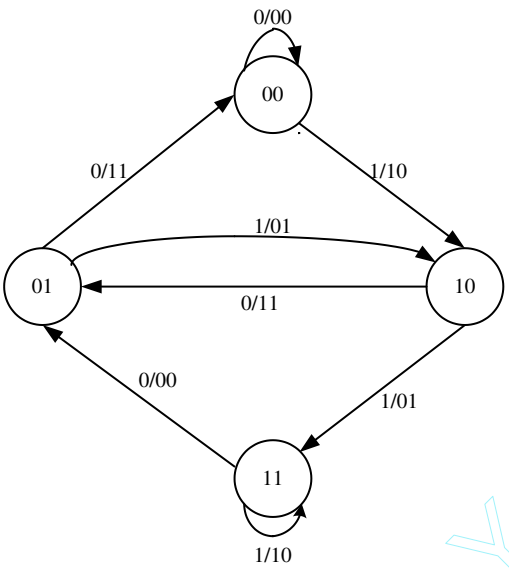


Figure 3-8
The state machine representation of the convolutional encoder of Figure 3-7

this bit moves into bit position 2 of the register, and so the state at the end of this bit period is 10. Similarly, the reader can easily trace the transitions at other states.

A limitation of the previous state machine is that it does not explicitly show how the states change with respect to time. They have to be derived by tracing the diagram for a given sequence of input. We can overcome this limitation and show the state transitions as a function of time by means of a trellis diagram, which has a tree-like structure, where diverging branches remerge at a node at some later time. This is shown in Figure 3-9. The number of nodes is equal to the number of states. The states at different instants of time are indicated on the vertical axis. Each line or trellis represents a state change at some instant $(n - 1)T$ to the next instant nT. Branches indicating state changes caused by an incoming bit 0 are shown by heavy lines, and those due to an incoming bit 1 by light lines. Alongside each line is a 2-bit number that represents the output of the encoder.

Referring to Figure 3-9, the initial state of the encoder at t = 0 is 00. If now the input to the encoder is 0, the output is 00, and the new

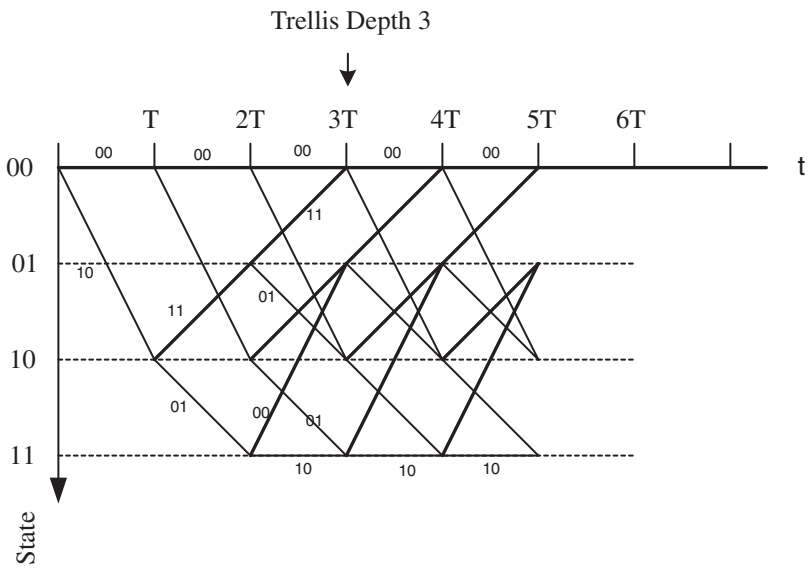


Figure 3-9
The trellis diagram for the encoder of Figure 3-7

state is the same as before, namely, 00. If, instead, the input is 1, you can see that the new state will be 10, and the corresponding outputs will be 1 and 0, denoted as 10 alongside the trellis. With the machine in this state at instant T, assume that another input bit comes in. If this input is 0, the new state is 01, and the resulting output is 11. If the input is 1 instead, the state transitions to 11 with an output 01 as shown in the figure. Similarly, if the state at 2T is 01, it transitions to 00 if the input is 0 or 10 if the input is 1. In this way, one can see state changes and resulting outputs at different instants of time.

There are a few things to notice in Figure 3-9. First, the trellis diagram has a repetitive structure that manifests itself after t = 3T. In general, this repetition period is equal to the constraint length of the encoder. At t = 3T (that is, at trellis depth 3) and beyond, every node has two incoming and two outgoing branches. This assumes that there are three registers and that the incoming data is being shifted only one bit at a time. In this case, the paths that diverge at a node remerge after two consecutive identical bits at the input.

Decoding Convolutional Codes

Conceptually, a convolutional code can be decoded in the following manner. For a given number of information bits, say, n at the source, we could compute the likelihood of the transmitted code sequence corresponding to each path of the trellis and choose the most likely one as the desired sequence. Because the number of all possible patterns containing n information bits is 2^n, this approach is not very practical for large values of n.

A procedure that overcomes this difficulty and is widely used is known as the *Viterbi algorithm* or the *maximum likelihood decoding* [8], [9], [11]. It uses sequential decoding at each sampling instant and is based on the fact that if at any instant t_k, there is a sequence of k information bits for which the receiver performance is optimum, then that sequence will be the first k information bits of a sequence that optimize the performance at any later instant $t_l > t_k$. Appendix A provides a brief description of the Viterbi algorithm. Interested readers are referred to references [9], [11] for a more detailed description of the algorithm.

Punctured Codes

The encoding procedure discussed in the previous paragraphs using a single register with the incoming message shifted one bit at a time is suitable for codes of rates 1/2, 1/3, 1/4, and so on. To generate a code of rate, say, 2/3, we could begin with a code of rate 1/2, and then remove one out of every four bits from the output. Because this would leave three bits of output for two bits at the input, the resulting code is rate 2/3. In this case, the code is said to be punctured. The Viterbi algorithm described in this chapter may also be used to decode punctured codes with slight modification. See, for example, reference [29].

Channel Encoders for UMTS

In UMTS, the *Media Access Control* (MAC) layer data carried by different transport channels are multiplexed, segmented if necessary, and then encoded into either a convolutional code or a turbo code,

depending upon the type of transport channels. The convolutional codes used are of rate ½ or ⅓. Because their constraint length is 9, they can be implemented with an 8-bit shift register. The implementation of the coder of rate ½ is shown in Figure 3-10. Notice that the generator polynomials of the two outputs may be represented by two octal numbers 561 (that is, 101 110 001 binary) and 753 (that is, 111 101 011 binary).

The coder of rate ⅓ is presented in Figure 3-11. The octal representations of the three outputs are 557, 663, and 711.

The turbo coder has a functional block diagram of Figure 3-12(a). It consists of an interleaver and two encoders. The two encoders, however, are identical and are built around a 3-bit shift register as

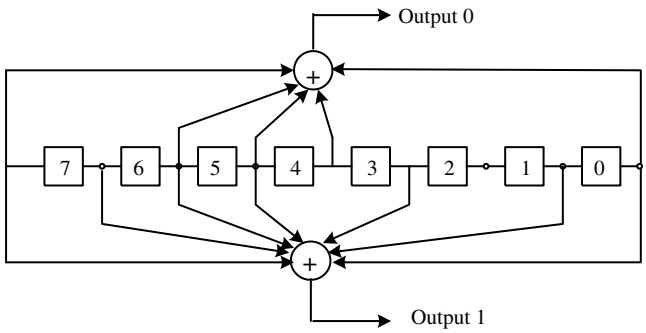


Figure 3-10
A convolutional encoder of rate ½ for UMTS

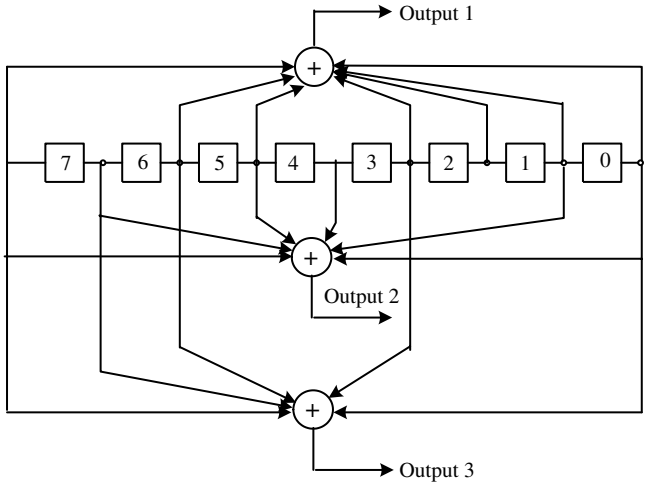


Figure 3-11
A convolutional encoder of rate ⅓ for UMTS

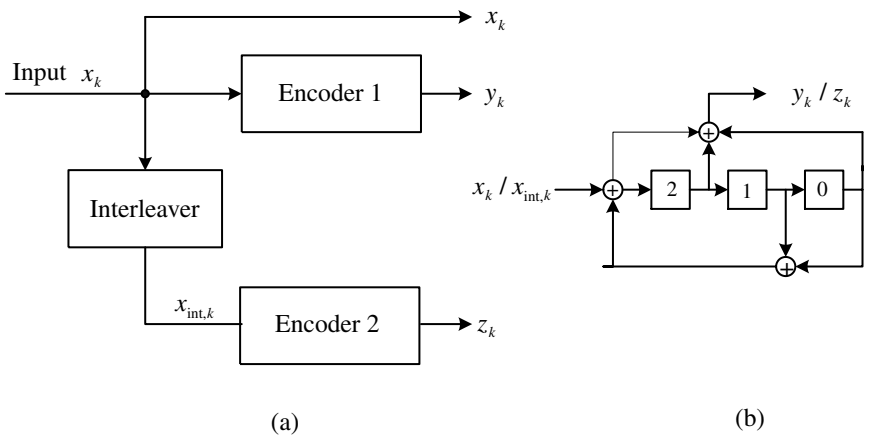


Figure 3-12
Turbo coder used in UMTS.
(a) Block diagram of a turbo coder.
(b) Encoders of Figure (a).

shown in Figure 3-12(b). The input to the first encoder is the incoming data sequence x_k. As for the other encoder, x_k is first interleaved. The output of the interleaver is then applied to the second encoder. Thus, for each input bit x_k, there are three bits at the output: x_k, y_k, and z_k.

In IS-95, *base stations* (BSs) use a convolutional code of constraint length 9 and rate ½. The code on the reverse link of IS-95 is rate ⅓, also with a constraint length k = 9. The generator polynomials of these coders are exactly the same as for the UMTS.

Interleavers

The purpose of an interleaver is to spread out adjacent bits of an incoming data stream so as to minimize the effect of burst errors. Interleaving is best explained with an example. Assume that an incoming data frame containing 600 bits is represented as (a1, a2, . . . , a599, a600). As the data comes in, it is written row by row into a 20 × 30 matrix so that bit a1 goes into row 1, column 1, bit a2 goes into row 1, column 2, and so on, as shown in Figure 3-13. If the matrix is now read out column by column, the output data sequence is a1, a31, a61, . . . , a541, a571, a2, a32, . . . , a542, a572, . . . , a30, a60, . . . , a570, a600. Thus, any two adjacent bits in the input

Figure 3-13
A block interleaver

a1	a2	o o o	a30
a31	a32	o o o	a60
o o	o o	o o o o o o	o o
a571	a572	o o o	a600

sequence are interleaved by 29 bits in the output stream. The interleaver span is the total number of bits in the block which are interleaved, or an equivalent time period. For example, in this case, the span is 600 bits.

In UMTS, interleaving is done in two steps. First, each transport block, after channel coding, is interleaved in the first interleaver. Here, the interleaver matrix (that is, Figure 3-13) may have 1, 2, 4, or 8 columns, depending upon the *Transmission Time Interval* (TTI) (that is, 10, 20, 40, or 80 ms). When there are 4 or 8 columns, they are first permuted in a certain order before reading out the matrix. For example, if there are 4 columns, columns 2 and 3 are interchanged.

After the first interleaver, transport channels are mapped to physical channels. The data associated with each physical channel is passed through the second interleaver, where the number of columns is fixed at 30. Clearly, the number of rows depends on the number of bits in a frame. If the number of incoming data bits is not an integral multiple of 30, the last row is padded with dummy bits. The columns of the matrix are permuted in a specific order and then read out from left to right. The dummy bits are removed from the output before it is spread by a channelization code.

Modulation

The basic idea of modulation is to vary some characteristic property of a *radio frequency* (RF) signal, such as its amplitude, frequency, or

phase, by the digital signal to be transmitted. The incoming serial data is first mapped into a sequence of symbols. The number of bits per symbol may vary depending on the type of modulation. The simplest modulation scheme is BPSK. Here each symbol consists of an incoming bit, and the phase of the carrier is changed by 0 degrees if the symbol is a zero and by 180 degrees if the symbol is a one. In QPSK, each pair of incoming data bits constitutes a symbol. The phase of the carrier is altered according to the symbol. For example, in one possibility, the carrier phase is advanced by 0 degrees if the symbol is (0,0), by 90 degrees if the symbol is (0,1), by 180 degrees if the symbol is (1,1), and by 270 degrees if the symbol is (1,0). In digital cellular systems, QPSK is used. This is usually implemented by dividing the incoming data into two sequences—one with the odd bits and the other with the even bits. Each of these sequences then BPSK-modulates the carrier [12]. The block diagram of a QPSK modulator is shown in Figure 3-14. See Appendix B for more detail.

Demodulation of a Phase Modulated Signal

There are two types of demodulation—coherent and differential. With coherent detection, the receiver needs to know the absolute phase angle of the transmitter carrier. In differential detection, on the other hand, it is not necessary to know this absolute phase angle. All that is required here is the relative phase angle of the carrier during successive symbol periods. In this case, however, the receiver

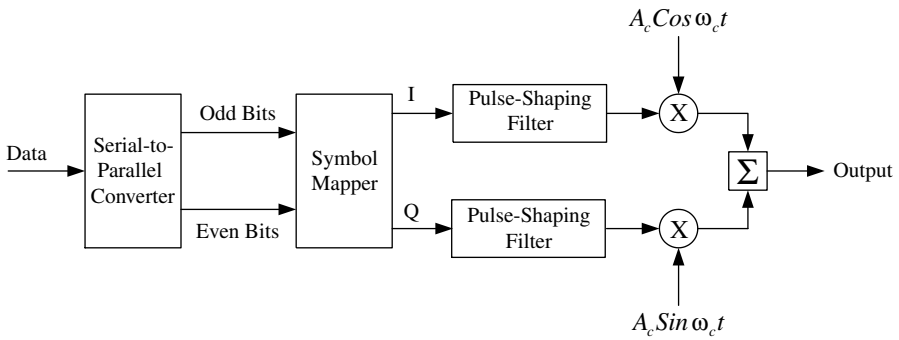


Figure 3-14
Functional block diagram of a QPSK modulator used in digital cellular systems

performance degrades to some extent. For example, this degradation is about 2 dB for BPSK or QPSK.[7]

If the channel noise is Gaussian, coherent detection gives the best performance. Under these circumstances, BPSK and QPSK have almost identical bit error rate performance. This is shown in Figure 3-15 [13]. However, the bandwidth required in QPSK is about one half of that for BPSK. To be more specific, if R_b is the data rate, the bandwidth with BPSK using rectangular pulses is $2R_b$. If rectangular pulses are passed through a raised cosine, pulse-shaping filter, almost all of the power spectral density is contained in a bandwidth of $1.5R_b$ around the carrier frequency. With QPSK, the null-to-null bandwidth is R_b when using rectangular pulses and about $0.75R_b$ with a raised cosine filter.

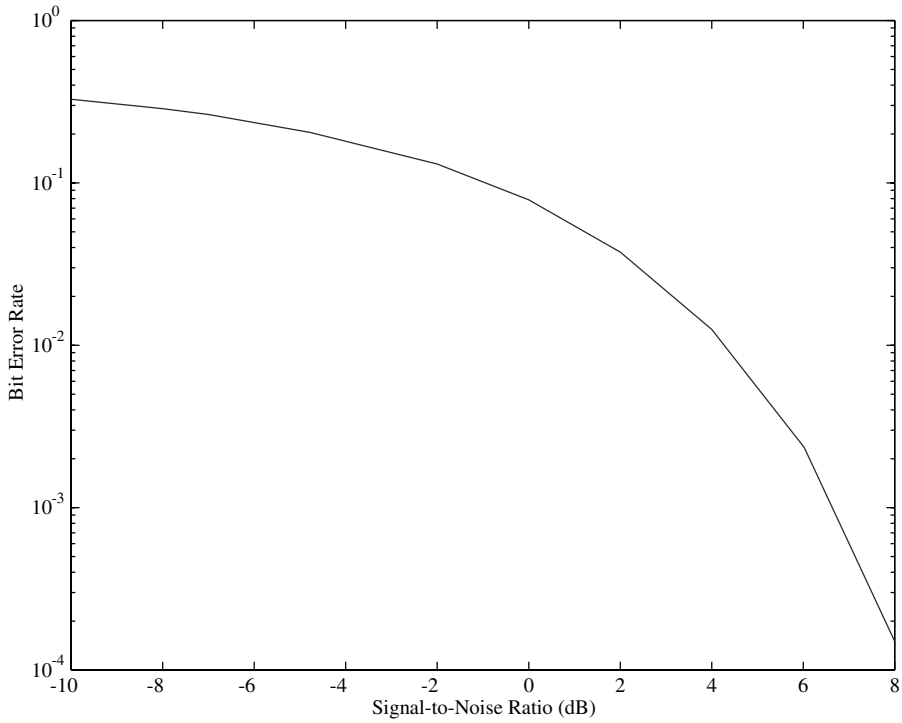


Figure 3-15
The bit error rate performance of BPSK with coherent detection in the presence of Gaussian noise

[7]In other words, the SNR required to achieve the same bit error rate is 2 dB more for the differential detection.

Spreading

In UMTS and cdma2000, signaling and user data is spread twice in succession—first with the channelization codes and later with the scrambling codes. The channelization codes are orthogonal Walsh codes, which are inherently more tolerant of interference caused by multiple users. The scrambling codes, on the other hand, are not necessarily orthogonal and are built from the so-called PN codes.

Walsh Codes

Various physical channels may exist at any time on a radio interface of a 3G system. For example, at a mobile station, there may be one or more dedicated physical data channels, a dedicated physical common control channel, a physical random access channel, and a physical common packet channel. To separate these channels at the receiver, they are spread with Walsh codes at the transmitter. These codes are formed by the rows of an $N \times N$ square matrix, whose entries are either 0 or 1. Usually, $N = 2^n$ where n is an integer. They are orthogonal because if a 0 is mapped to $+1$ and a 1 to -1, then the sum of the term-by-term products of any two rows of this matrix is 0. In other words, if the matrix is assumed to be $[a_{ij}]$, $i, j = 1, \ldots, N$, then $\sum_{k=1}^{N} a_{ik}a_{jk} = N$ for $i = j$, and 0 otherwise.

This matrix, also known as the *Hadamard matrix*, can be generated recursively in the following manner when its dimension N is given by $N = 2^n$ with n an integer:

$$H_{2^0} = H_1 = [1],$$

$$H_{2^n} = \begin{bmatrix} H_{2^{n-1}} & H_{2^{n=1}} \\ H_{2^{n-1}} & -H_{2^{n-1}} \end{bmatrix}$$

Thus, for example,

$$H_{2^0} = H_1 = [1]$$

$$H_{2^1} = H_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$H_{2^2} = H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Channelization codes used in UMTS W-CDMA and cdma2000 are variable-length Walsh codes, also known as *orthogonal variable spreading factor* (OVSF) codes. The spreading factors in UMTS may vary from 4 to 256 chips on uplink channels and from 4 to 512 chips on downlink channels. In cdma2000, Walsh codes used on traffic channels may vary from 4 to 128 chips.

IS-95 uses a set of 64 fixed-length Walsh codes to spread forward physical channels. For example, Walsh code 0 is assigned to the pilot channel, code 32 to the sync channel, codes 1–7 to paging channels, and the rest to the forward traffic channels. In the reverse direction, they are used for orthogonal modulation where every six symbols from the block interleaver output are modulated as one of 64 Walsh codes.

Scrambling Codes

PN codes form the basic building blocks of scrambling codes. These codes are generated by a multibit shift register, where some selected outputs are added modulo 2 and fed back to the input. The underlying theory is well documented in the literature. See, for example, reference [14] for a thorough description of shift register sequences. Reference [15] provides some relevant mathematical background.

Theory of PN Codes To illustrate the principle of PN codes, consider Figure 3-16. The shift register array consists of four single-bit shift registers and is clocked at the chip rate. The outputs of registers 3 and 0 are added in a modulo-2 adder (that is, an exclusive-OR circuit), and then applied to the input of the array.

Assume that the initial states of all stages of the shift register are 1, 1, 1, and 1. Then at instant t = 0, the output of the adder is 0. So when the first clock pulse appears, the states of the 4-bit shift register change to 0, 1, 1, and 1, respectively. Thus, the states with successive clock pulses are

1 1 1 1
0 1 1 1
1 0 1 1
0 1 0 1
1 0 1 0
1 1 0 1
0 1 1 0
0 0 1 1
1 0 0 1
0 1 0 0
0 0 1 0
0 0 0 1
1 0 0 0
1 1 0 0
1 1 1 0
1 1 1 1

TEAMFLY

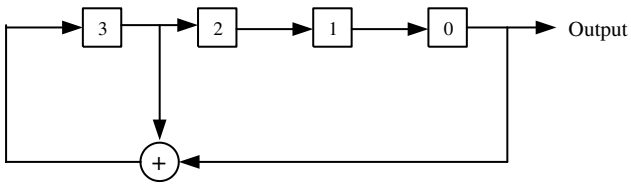


Figure 3-16 A 4-stage shift register to generate a PN code

There is no need to write the output sequence any more because the pattern repeats after every 15 clock pulses. The output may be taken from any one of the four registers. When taken from register 0, the output sequence is 1 1 1 1 0 1 0 1 1 0 0 1 0 0 0 . . . . The same sequence, with some delays, will be generated with any other initial states as long as they are not all 0's.

A few things are to be noticed here. First, the output sequence is periodic, but its bit pattern is random. Thus, it may be termed a *pseudo-random pattern*. In this example, the sequence repeats itself after 15 clock pulses. In other words, its period is 15 or $2^4 - 1$, the maximum that can be achieved with a 4-bit shift register. A sequence with the longest possible period that can be generated with a given shift register is called a *maximal sequence* or a *maximum-length sequence*. The second point to observe is that in each period, there are eight 1's and seven 0's. In other words, the numbers of 1's and 0's are equal within one clock period. Third, this particular output sequence has one run of length 4, one run of length 3, two runs of length 2, and four runs of length 1.

Some shift register sequences, finite as they are, may have all of the following randomness properties [14]:

1. In any given period, the number of 1's is equal to the number of 0's within one bit.
2. Each period of the output sequence contains runs of different lengths. In a random sequence, one half of these runs are of length 1, one fourth of length 2, one eighth of length 4, and so on.
3. The auto-correlation function $C(j)$ of a peroidic sequence $\{b_n\}$ of period N is defined by

$$C(j) = \sum_{i=1}^{N} b_i b_{i+j} \tag{3-5}$$

In other words, it is the sum of the products of the elements of the sequence and its delayed copy. Assuming that each b_i is either -1 or $+1$, the auto-correlation function $C(j)$ of the shift register sequence has only two values:

$$C(j) = \begin{cases} N \text{ if } j = 0 \\ K \text{ if } 0 < \text{j} < \text{N} \end{cases} \tag{3-6}$$

To see the auto-correlation property of the shift register sequence of Figure 3-16, note, first of all, that the auto-correlation function with zero phase shift is $C(0) = +15$. Next, if the output sequence is 1 1 1 1 0 1 0 1 1 0 0 1 0 0 0, then the sequence shifted one bit to the right is 0 1 1 1 1 0 1 0 1 1 0 0 1 0 0. Thus, the auto-correlation function $C(+1) = -1$. Similarly, the auto-correlation function $C(-1)$ with one bit shifted to the left is also -1. In fact, it can be shown that $C(j) = -1, j \neq 0$ as shown in Figure 3-17. This is an important property of a shift register sequence for the following reason: If each user's data is modulated with a unique PN code and transmitted to the remote end, then it can be demodulated at the receiver by correlating the received signal with a local copy of that code. If the period of this sequence $x(t)$ is very large, each user i may be assigned the sequence $x_i(t) = x(t + T_i)$, which is actually the same code but with an offset T_i. In this case, in order to be able to decode each user sequence as unambiguously as possible in the presence of noise, it is necessary that for $j > 0$, the value of $C(j)$ be as small as possible compared to N.

Similarly, the cross-correlation between two sequence x_n and y_n is defined in the following way:

$$R_{xy}(j) = \sum_{i=1}^{N} x_i y_{i+j} \tag{3-7}$$

Given a shift register of length, say, k, not all sequences will have the maximum period of length $2^k - 1$. Consider again the shift reg-

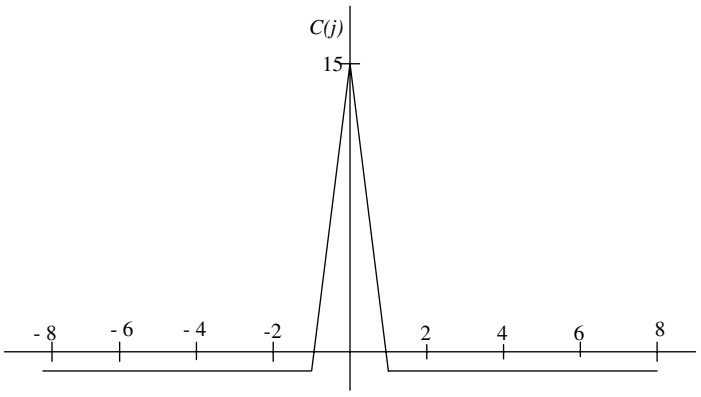


Figure 3-17
The auto-correlation function of a maximal length sequence

ister of Figure 3-16. However, this time, outputs of registers 2 and 0 are added modulo 2, and then applied to the input of the first stage as shown Figure 3-18. Assuming that the initial states of the register array are 1111, the subsequent states are

1 1 1 1
0 1 1 1
0 0 1 1
1 0 0 1
1 1 0 0
1 1 1 0
1 1 1 1

So in this case, the period $n = 6$.

Because the period p depends on the way the feedback signal is constructed, it is necessary to see which registers should be connected to the adder so that the sequence has the previous desirable properties. To this end, assume that the binary sequence $\{b_n\} = b_0, b_1, b_2, b_3, \ldots$ defines the states of register 3 as a function of time. Clearly, the same sequence delayed by one bit also gives the states of register 2, and so on. Thus, the sequence can be used to define the state of the whole shift register array. Using this sequence as coefficients, a polynomial *f(x)* can be defined:

$$f(x) = \sum_i b_i x^i \tag{3-8}$$

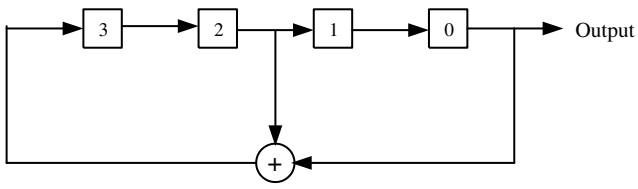


Figure 3-18
Another 4-bit shift register sequence. This is similar to Figure 3-16 except that the feedback is taken from different registers.

In this case, we can determine a function

$$g(x) = 1 + \sum_{i=1}^{k} g_i x^i \tag{3-9}$$

such that $f(x) = \dfrac{1}{g(x)}$, where the division by $g(x)$ is modulo 2. In other words, the shift register sequence $f(x)$ can be generated by constructing a feedback signal by adding (modulo 2) the outputs of the k shift registers according to the function $g(x)$. Function $g(x)$ is called the characteristic polynomial for sequence b_n. Clearly, in the previous definition of $g(x)$, g_i may be 0 for some values of i.

We will now provide some important results without giving any proof:

1. Given a function $g(x)$ as defined previously, the period of a k-stage shift register sequence is the smallest positive integer n for which $g(x)$ divides polynomial $1 + x^n$. The maximum possible period of a k-stage shift register is $2^k - 1$.
2. If the feedback in a k-stage shift register is such that the output sequence has indeed this maximum period $2^k - 1$, then the function $g(x)$ that generates this sequence is irreducible. In other words, $g(x)$ is divisible only by itself (and of course by 1). Reference [14] lists a few irreducible polynomials.

   Notice that this statement merely gives a necessary condition for the sequence to be maximal length. This condition is not sufficient because there are polynomials that are irreducible, but do not yield a maximal length sequence. For example, consider $g(x) = 1 + x + x^2 + x^3 + x^4$. This polynomial is irreducible, but because it divides $1 + x^5$, the period of the sequence is 5 and not $2^4 - 1$.
3. The period of a k-stage shift register sequence is a factor of $2^k - 1$. Thus, if $2^k - 1$. is a prime, then every irreducible polynomial of degree k will generate a maximal length sequence.

To illustrate these ideas with examples, consider the shift register of Figure 3-16. Here, the shift register consists of four stages. Thus

$k = 4$. Because the outputs from registers 3 and 0 are being added and then applied to the input, the characteristic polynomial is

$$g(x) = x^4 + x^3 + 1 \tag{3-10}$$

The period is 15, the maximum achievable with four registers. It can be shown that the smallest positive integer n for which this $g(x)$ divides $1 + x^n$ is indeed 15, thus satisfying statement 1 previously. Furthermore, as required by statement 2, it can be easily shown that $g(x)$ is in fact irreducible because it cannot be factored into polynomials of lower degrees. But because 15 is not a prime, following statement 3, not all irreducible polynomials of degree 4 will generate a maximal length sequence.

Referring to Figure 3-18, the characteristic polynomial for this shift register sequence is

$$g(x) = x^4 + x^2 + 1 \tag{3-11}$$

It can be shown that this polynomial divides $1 + x^6$. Thus, the period of this sequence is $n = 6$, as shown previously. Also, this characteristic polynomial is not irreducible because $1 + x^2 + x^4 = (1 + x + x^2)^2$, and hence, in accordance with statement 2 previously, cannot generate a maximal length sequence.

Scrambling Codes in UMTS The scrambling codes in UMTS are complex valued and may be either long or short. A long code has a length of 38,400 chips (that is, 10 ms) and a short code only 256 chips. As an example, Figure 3-19 shows how a long code for a UMTS uplink channel is generated [12]. This code is constructed with two PN codes, whose characteristic polynomials are $g_1(x) = x^{25} + x^3 + 1$ and $g_2(x) = x^{25} + x^3 + x^2 + x + 1$. They are implemented as sequences PN1 and PN2 using two 25-bit shift registers. PN1 and PN2 are added modulo 2, and the output is mapped to a real-valued function, say, I. Another function Q is derived by simply delaying I by $2^{24} + 16$ chips. Q is multiplied by $\pm j$, where the sign changes every chip period, and then added to I to yield the long code.

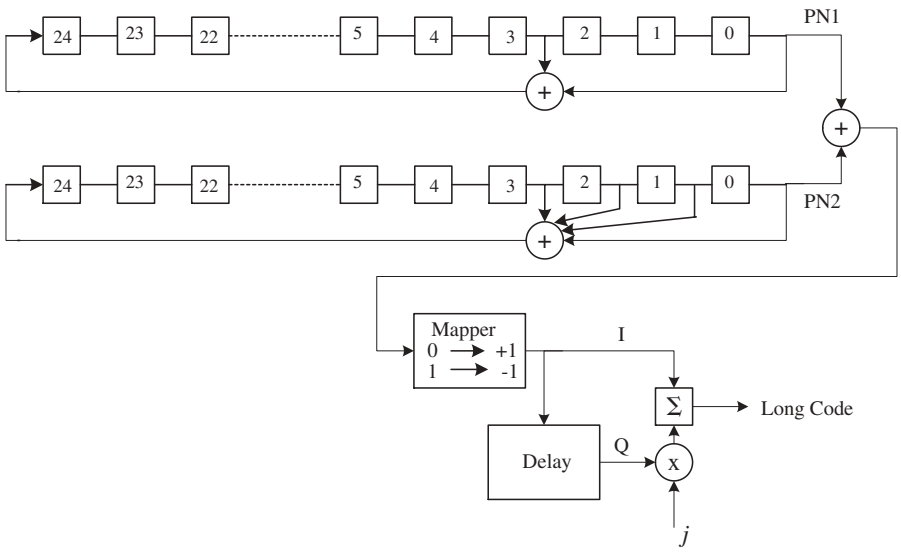


Figure 3-19
A long code generator for a UMTS uplink channel

Receiver

Receiver Structure

A QPSK receiver structure with coherent detection is shown in Figure 3-20. The purpose of the predetection band pass filter is to select the desired RF channel. Because coherent demodulation is being used, it is necessary to generate at the receiver a carrier that has the same or nearly the same frequency and phase as the carrier at the transmitter. The output of the demodulator is low-pass filtered, and then applied to the input of a matched filter. Notice that the output of the demodulator is an analog signal that would be the same as the modulating signal at the transmitter if there were no channel noise and if the carrier recovery were perfect, that is, the frequency and phase of the local carrier were exactly the same as those of the transmitter carrier.

To understand the purpose of the matched filter, recall that QPSK modulation is achieved at the transmitter by dividing the input data

into two streams, each of which BPSK-modulates the carrier. Consequently, we need only consider detecting a BPSK-modulated signal on either of the branches of Figure 3-20.

A BPSK modulated signal with additive white noise may be detected by simply determining whether or not a pulse is present at the receiver input. The probability of error in making this decision can be minimized by using a filter whose impulse response is matched to the (transmitted) pulse. Such a filter is called a *matched filter*. It can be shown that if the transmitted signal is a unit rectangular pulse of duration *T,* then the output of the matched filter is simply the integral of the received signal over the symbol period *T.* At the end of this period, the decision circuit samples the output of the integrator, and based on some algorithm, decides which symbol was transmitted, whereupon the filter must dump its output so that the cycle can commence all over again at the start of the next symbol period. As such, a matched filter is often referred to as an integrate and dump circuit [11].

The output of the matched filter is despread by multiplying it with the I-channel and Q-channel scrambling code. Here, the filter output may be sampled at the chip rate, multiplied by the scrambling code and the samples averaged over a symbol period. This average value may then be applied to the decision circuit.

Hard and Soft Decision

The decision circuit acts on the output of the matched filter to decide whether a 0 or a 1 was transmitted. One could use a simple decision

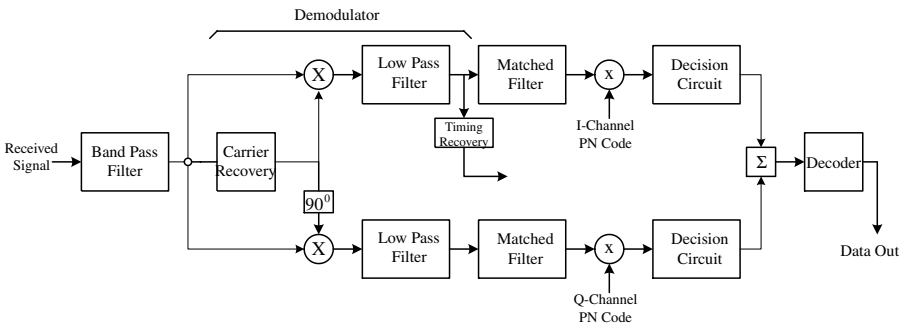


Figure 3-20
A QPSK receiver structure with coherent detection and hard decision decoding

rule and conclude that the demodulated symbol is 0 whenever the output of the matched filter $r_k > 0$ and 1 whenever $r_k \leq 0$. All the incoming symbols could be detected this way, and the detected data sequence applied to the input of a decoder for error detection and correction. When the input bit stream to the decoder has already been decided by the decision circuit, it is called a *hard decision*. This is shown in Figure 3-20, where the I and Q streams are detected by the decision circuits using hard decision rules, multiplexed together, and then applied to the decoder.

Alternatively, the decoder may read the analog samples of the matched filter, and based on certain rules that are designed to optimize the performance of the receiver, determine the received bit sequence. For example, one could quantize the filter output into, say, four levels as shown in Figure 3-21, and use the following decision rules. Assuming that the symbol transmitted is 0, the output of the matched filter is in the range

$0.5\nu < r_k \leq \nu$ with probability, say, 0.9,

$0 , r_k \leq 0.5\nu$ with probability, say, 0.75,

$-0.5\nu < r_k \leq 0$ with probability 0.1, and

$-\nu < r_k \leq -0.5\nu$ with probability 0.05.

A similar set of rules may be used in connection with the transmitted symbol of 1. See Figure 3-22. Thus, given a sequence of output voltages from the filter over a number of symbol periods, the decoder could construct a number of code sequences each associated with a certain probability indicating how likely that code sequence is and then choose the one with the highest probability as the most likely transmitted code sequence. In other words, the decision circuit now becomes part of the decoder as shown in Figure 3-23. In this case, the decoder is said to have used soft decision decoding [11].

Soft decision decoding improves the bit error rate performance of the receiver. Generally, the higher the quantization levels, the greater this improvement. With eight levels, there is a gain of about 1.75 dB in E_b/N_0 compared to the hard decision decoding [9]. If the number of levels is increased beyond eight, the additional improvement is insignificant. For example, with infinitely fine levels, the

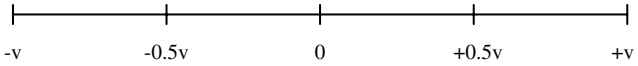


Figure 3-21 Four-level quantization of the input to the decision circuit

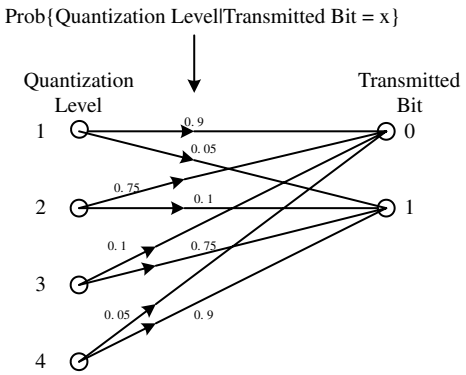


Figure 3-22 Transition probabilities in soft decision decoding with four levels of quantization. As an example, the probability that the matched filter output has a quantization level 1, assuming that the transmitted bit is a 0, is 0.9.

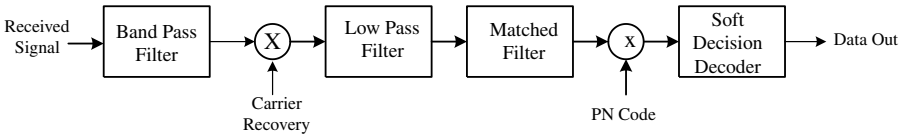


Figure 3-23 Receiver with soft decision decoding

gain in E_b/N_0 is only 2 dB. Thus, for all practical purposes, eight or even four levels are sufficient.

Viterbi Decoding

The Viterbi algorithm for decoding convolutional codes has been studied extensively in the literature [8]–[11]. The algorithm has been shown to be the maximum likelihood and is particularly useful for decoding convolutional codes of small constraint lengths.

Appendix A gives a brief description of this algorithm using soft decision decoding.

Multipath Diversity in CDMA

Wider bandwidths offer some advantages that are not available with narrowband systems. One of them is reduced fading, and the other is concerned with multipath diversity. As described previously, the signal received at a mobile antenna consists of a number of reflected rays, each characterized by a different amount of attenuation and delay. As a vehicle moves around, the attenuation and delay associated with each path also change with time. As a result, the received signal is time varying in nature. If the signal bandwidth is small compared to the coherence bandwidth of the channel, different frequency components of the signal have identical statistics and tend to vary with time in the same manner. The result is that they interfere with each other, causing correlated signal fading. This kind of fading is called *flat fading*. If, on the other hand, the signal bandwidth is large, as in W-CDMA, compared to the coherence bandwidth of the channel, the different frequency components of the signal still interfere with each other, but now they are statistically independent, and therefore, their variations are not correlated with respect to each other. The fading that results from this situation is called *frequency selective fading*. In this case, if one frequency component is below the average level of the signal, another component is likely to be above that level, and so, the net effect would be a reduction in fade [31]. This is shown in Figure 3-24 where the term fade depth is defined as the amount by which the signal falls below its average value with a probability of 0.1. If the channel bandwidth is only 30 kHz as in cellular system TIA-553, the signal goes into a −10 dB fade with probability 0.1. On the other hand, in cdmaOne, where the bandwidth is 1.25 MHz, the signal encounters a −8.75 dB, whereas in W-CDMA with a bandwidth of 5 MHz, the signal experiences a −5.75 dB fade with the same probability. Thus, fading is reduced with increased bandwidth.

When the signal bandwidth is sufficiently large, and multipath components are delayed by more than a chip period, they can be

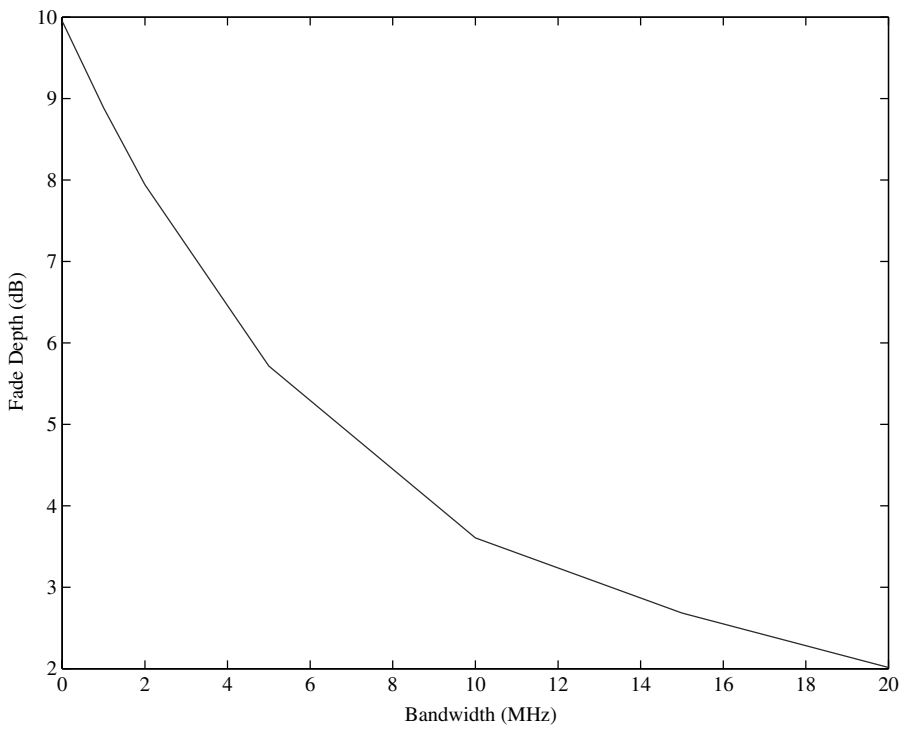


Figure 3-24
The effect of channel bandwidth on fading. Here, fade depth is defined as the amount by which the signal falls below its average value with a probability of 0.1.

treated as multiple copies of the same signal, and may, therefore, be combined constructively in a diversity receiver as discussed in the following section.

Rake Receiver

Studies of power delay profiles of urban and dense urban areas around 900 MHz indicate that most of the energy of the received signal is due to the reflected rays with delays in excess of 0.75 μs. In W-CDMA, because the chip period is quite small (about 0.2604 μs for a chip rate of 3.84 Mc/s) compared to the delay spread, multipath components with delays of more than one chip period may have significant energy. Thus, if the signal bandwidth is greater than the coherence bandwidth of the channel, the various multipath components can be used as different branches of a diversity receiver. This,

in essence, forms the basis of a rake receiver [17], [18], [21], [32], [33]. Each branch of the receiver is called a *finger*.

Because the autocorrelation of a PN code across successive chip periods is very small, that is, $R(n\tau)$ is very small for $n > 0$ compared to its value for $n = 0$, copies of the desired signal can be extracted if their associated delays are multiples of the chip period. The extracted components may then be combined in an optimal manner so that fading is reduced and, consequently, the bit error rate performance of the receiver improved. Notice that these delays, although greater than the chip period, must at the same time be sufficiently small compared to the bit period; otherwise, there may be significant intersymbol interference.

Suppose for illustrative purposes that Figure 3-25 represents the power delay profile of an urban area. Here, two distinct paths are identifiable, one with a delay spread from 0 to τ_1 and the other from τ_1 to τ_2. If the delay spreads are greater than the chip period, the two paths can be resolved and taken advantage of in the receiver. The functional block diagram of a rake receiver is shown in Figure 3-26. After coherent demodulation, the I-channel and Q-channel signals are multiplied by their respective PN codes. The outputs a_I and a_Q are added in quadrature and multiplied by the user-specific PN code. For each branch of the receiver, there is a matched filter, which integrates the input signal over a period that depends on the delay spread of the associated branch. For example, starting at any symbol, say, nT, where n is an integer and T is the symbol period, the first branch is integrated from nT to $(n + 1)T$, the second branch from $nT + \tau_1$ to $(n + 1)T + \tau_1$, and so on. The matched filters are

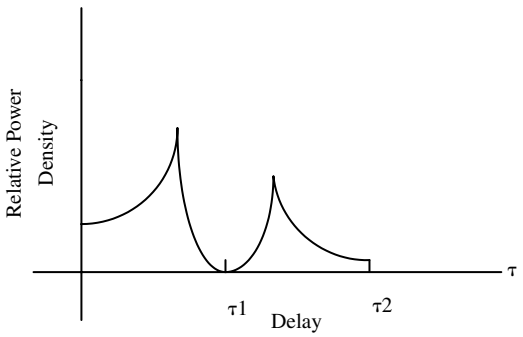


Figure 3-25
An example power delay profile used in a rake receiver

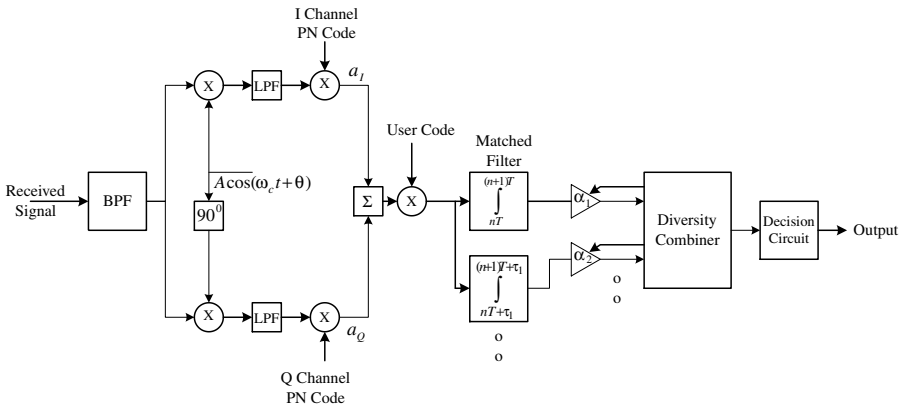


Figure 3-26
Functional block diagram of a rake receiver for W-CDMA

designed to yield the maximum SNR in the presence of Gaussian noise. Outputs of the filter banks are each weighted by gain α_i, summed together in the block marked Diversity Combiner and applied to the decision circuit where symbols are decoded.

The gain coefficients α_i may be either fixed, or dynamically adjusted, as shown in Figure 3-26, using an appropriate diversity-combining algorithm. For example, in an equal-gain combiner, coefficients α_i for all paths are equal and set to a constant value. Alternatively, in a maximal ratio-combining scheme, each gain α_i is adjusted so that it is proportional to the signal-to-noise-power ratio of the associated branch. When the gains are adjusted this way, an optimum SNR value of the receiver is obtained. Because a channel can be modeled as a tapped delay line, it is found that in this case, coefficient α_i for any branch is the complex conjugate of the channel coefficient of the associated path.[8] Because the channel is usually time varying, it is necessary to track and estimate the channel periodically. This can be accomplished either by regularly sending a known pattern from the transmitter (as in IS-136) or by means of an auxiliary pilot, as in W-CDMA.

[8]Based on its impulse response, a channel can be modeled as a tapped delay line. The number of taps depends upon the degree of accuracy required. The power delay profile of this example suggests that two taps are sufficient for this channel.

The maximum amount of multipath delay that can be exploited in a rake receiver is usually limited, and is determined by the power delay profile. As an example, for a city like New York, it lies in the range of 0.25–2.5 μs. Thus, in UMTS W-CDMA, where the chip rate is 3.84 Mc/s, the delay is about 1–10 chips.

Although multipath diversity is a property of all CDMA systems, it is only W-CDMA that provides multipath diversity for small cells (that is, the micro and pico cells). To see this, consider IS-95 where the carrier bandwidth is 1.25 MHz. In this case, because the chip rate is 1.2288 Mc/s and because the delay must be at least one chip long to achieve multipath diversity, the difference in path lengths must be at least 244 meters. On the other hand, for W-CDMA with 5 MHz bandwidth, the chip rate is 3.84 Mc/s, and so this path difference is reduced to 81 meters.

The multipath diversity employed in a rake receiver leads to an improvement in performance. For example, the value of E_b/N_0 required to ensure a bit error rate of 10^{-3} on a fading channel is about 10 dB, assuming BPSK modulation, a 4-branch rake receiver, and equal gain combining. The required E_b/N_0 for the same bit error rate is 14 dB with two branches and about 24 dB with one branch, that is, without any multipath diversity [21]. The maximal ratio combining has the best performance. If most of the signal energy is contained in only one branch, a conventional receiver will perform better than a rake receiver that uses equal gain combining [33] because, in this case, branches with very little signal power will only add to the noise.

Multiuser Detection

Consider the uplink transmissions in UMTS. Here, the user data on various physical channels (such as dedicated physical data channels, dedicated physical control channels, and so on) is first spread with a channelization code, and then scrambled with a user-specific PN code. Because channelization codes are mutually orthogonal and thus more resistant to multiuser interference, the physical channels can be correctly separated at the receiver with a high probability. The

scrambling codes, on the other hand, are generally nonorthogonal. This is not a problem in a synchronous system, such as IS-95, because here, all transmissions are synchronized to a systemwide time reference. Thus, signals from multiple users arrive at the BS with relatively small delays. Consequently, the cross-correlation between the signals is quite small. In contrast, because UMTS W-CDMA is an asynchronous system, these delays are random as shown in Figure 3-27, and may be comparable to the bit period. As a result, the cross-correlation between the received signals from multiple users is no longer negligible and, if ignored, causes significant errors in soft decision decoding.

Besides, very often the power control is not perfect. Even when a mobile is adjusting its transmitter power at 1,500 Hz on command from the BS, this closed-loop power control algorithm does not work well for mobile velocities of 100 km/h or more. Thus, the amplitude of the desired signal may at times be quite small compared to interfering signals. So, the performance of a matched filter followed by a simple decision circuit is not optimum anymore. Multiuser detection attempts to overcome this problem by detecting the desired user signal in the presence of interference from all other users in some optimum way.

A number of multiuser detection algorithms have been suggested [21]. One of them is based on the Viterbi algorithm with soft decision

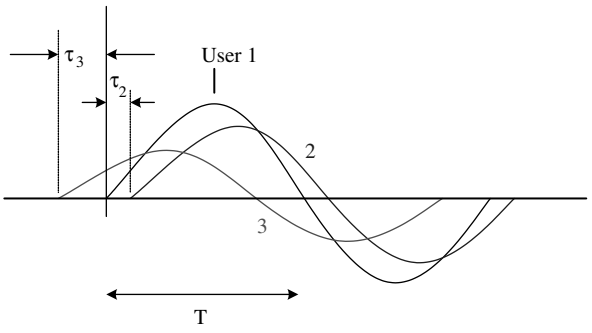


Figure 3-27
Signals received at a BS from multiple users. In an asynchronous system, the time offsets shown as τ_2 and τ_3 with respect to the desired signal from, say, user 1 are significant.

decoding. The ideas here are similar to those discussed in connection with the maximum likelihood decoding of convolutional codes [22], [23]. The received signal, after demodulation, is multiplied by the scrambling code of each user, integrated over a symbol period using a matched filter, and applied to a soft decision decoder. The output of the matched filter corresponding to any desired user depends upon the cross-correlation between the signal from that user and signals from all other users over three consecutive symbol periods. Over a given symbol length, the soft decision decoder considers all combinations of symbols from multiple users, and using a channel model together with the observed outputs of the matched filter, estimates the likelihood of each sequence of symbols. Appendix C presents a brief description of this algorithm. Although the performance of this receiver is optimum, it is not very practical because the number of real-time computations required increases as 2^n, where n is the number of users to be detected. A number of authors have proposed suboptimum receiver structures where these computational requirements are less stringent.

Another technique suggested for multiuser detection involves successive cancellation of interference from the received signal [24]–[28]. Here, the receiver first extracts the strongest signal of all users and subtracts it from the received signal. Next, the second strongest signal is detected from the remaining signal, and subtracted from this latter signal, and so on, until signals from all users have been detected. The idea is illustrated in the block diagram of Figure 3-28. Because the performance of the receiver depends on the accuracy with which the strongest interference is detected in the first stage, reference [24] suggests using a multipath-combining receiver for detecting the strongest interference.[9] The detected data of this user is then passed through a channel model to regenerate a signal, which approximates as closely as possible the received signal from this user. The output of the channel model is subtracted from the received input. The result is used to derive the second strongest signal in the same way. Conventional receivers may be used in the second and subsequent stages.

[9]For this to be possible, it is necessary that the signal bandwidth be much greater than the coherence bandwidth of the channel.

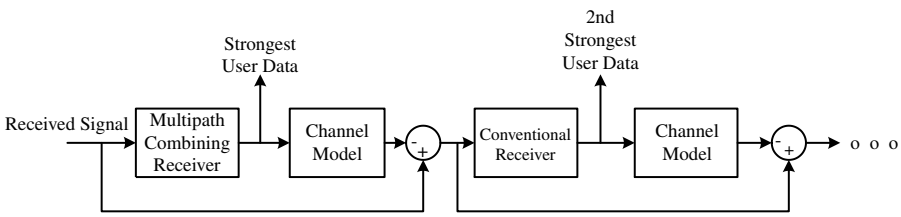


Figure 3-28
Multiuser detection using successive cancellation of interference

Because of the complexity involved, multiuser detection is more amenable to implementation at a BS. Moreover, because a mobile station is only concerned with detecting the signal from a single user, multiuser detection is really not necessary at a mobile station.

In UMTS W-CDMA, both long and short scrambling codes may be used on uplinks. However, short codes are generally more suitable for multiuser detection [41]. Long codes are handled better by the algorithm based on the successive cancellation of interference.

Smart Antennas

In a previous chapter, isotropic and directional antennas were discussed. An *isotropic antenna* is one that radiates energy equally in all directions in any horizontal or vertical plane. Practical antennas, however, are not isotropic. For example, with an omnidirectional antenna, such as a vertically mounted, half-wave dipole, or a short monopole, the signal strength at any given distance from the antenna is distributed equally in all directions in the horizontal plane. In the vertical plane, however, the signal strength at any point depends on its location with respect to the vertical axis. This is shown in Figure 3-29(a). The power density is 0 along the vertical axis and increases as the angle θ increases, attaining a maximum value on a horizontal plane through the antenna such that $\theta = 90$ degrees. As discussed in Chapter 2, the signal strength decreases at points further and further away from the transmitter antenna. An example of an omnidirectional antenna is the antenna at a mobile station or a center-excited BS.

As the name implies, a directional antenna radiates most of its energy only in a certain direction, transmitting the signal in the form of a beam in the direction of the antenna. The radiation pattern for a vertically mounted directional antenna is shown in Figure 3-29 (b). Notice how the signal strength varies even in a horizontal plane. Depending upon the design, the energy in the back lobe is usually very small. Directional antennas are used to provide coverage on highways and in corner-excited, 3-sector cell sites, where each sector has an angular width of 120 degrees. Clearly, there are many advantages of a directional antenna. For example, with a given transmitter power, it extends the coverage area, decreases the probability of the far-near problem that was discussed before, reduces interference to a given mobile due to other active users on the same frequency, and thus increases the system capacity (such as the number of users in a CDMA system).

In 3-sector cells, a sector may be covered by a number of narrow-beam antennas as shown in Figure 3-30. The beams formed by these antennas are fixed, each of which may be used to cover users concentrated in certain directions. In this case, the BS must be able to track each user and switch the beams appropriately as a mobile station moves from the coverage area of one beam to another. A disadvantage of the fixed beam approach is that if the traffic pattern changes from the one for which the beams were originally designed, the system may not operate at the same level of performance.

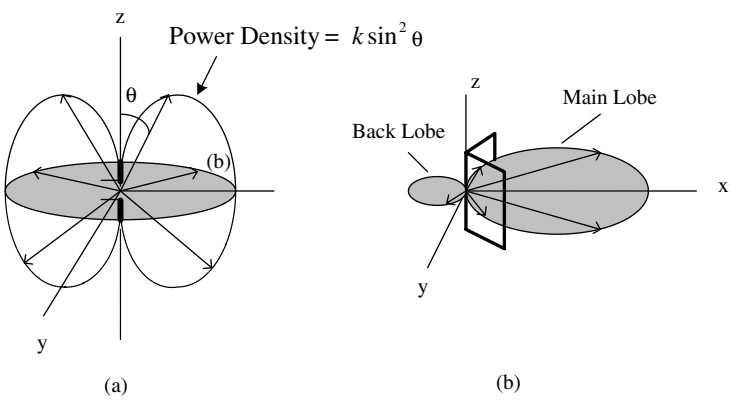


Figure 3-29 Radiation patterns of two antennas: (a) Omnidirectional antenna, (b) Directional antenna

Because each mobile station has a unique physical location, the signal received from each can be processed in real time and separated from the signals of all other users even though they may overlap in the time or frequency domain. Signal processing required to perform this function is called *spatial filtering* or *filtering in the space domain*. This technique is also called by some authors *space-division multiple access* (SDMA) because this enables multiple users to be distinguished even though they may occupy the same frequency or time slot.

Clearly, sectorization of cells with directional antennas and use of fixed beams may be considered as a form of spatial filtering. Another way to implement spatial filtering is to use an adaptive antenna array where the signal received from each element of the array is multiplied by a gain coefficient, called a *weight*, summed together, and then processed using digital signal processing techniques so as to maximize the system performance according to some criteria. The weights are adjusted dynamically using an adaptation algorithm that tries to achieve some design objectives. For example, an objective may be the formation of a beam in a desired direction so that the signal is maximized in that direction and minimized or even reduced to a null in other directions, say, in the direction of co-channel sources. This is called *digital beam forming*. Another objective may be the minimization of bit error rates for users located in a certain geographical area where the error rate would otherwise be excessively high due to clutter or other conditions. The term *smart antennas* refers to both switched beam antennas and adaptive antenna arrays.

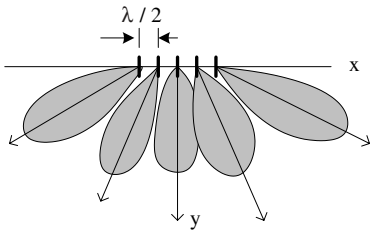


Figure 3-30
Fixed beams formed by narrow-beam antennas

Fundamental to the operation of adaptive antennas is the ability to estimate the angle of arrival of signals from different users and, based on the estimate, steer the beams on downlink channels. The arrival angle is generally quite well defined in rural areas, but not so in microcells or indoors. Because for large cells, the angle of arrival varies much more slowly than the instantaneous fading signal, measurements from mobile stations may also be used in the adaptation algorithm.

The concept and theory of adaptive antennas may be found in References [35], [36]. Various authors have investigated the application of adaptive antennas to mobile communications systems [37]–[39], [42]. Reference [40] discusses the possibility of extending the capacity of an existing cellular system so as to serve areas of high traffic density by using smart antennas. Possible benefits of using smart antennas in 3G systems have been studied under the auspices of the *Technology in Smart Antennas for Universal Advanced Mobile Infrastructure* (TSUNAMI) project in Europe [41], and include the following:

- Extending the range or coverage area in a desired direction with beamforming
- Increasing the system capacity in areas with dense traffic (that is, hot spots)
- Dynamically adjusting the coverage area (say, from 120 to 45 degrees)
- Creating nulls to/from co-channel interferers so as to minimize the co-channel interference
- Tracking individual mobile stations using separate, narrow beams in their direction
- Reducing multipath fading

In this section, we will explain briefly how beam forming is accomplished by adaptive antennas.

Figure 3-31(a) shows a functional block diagram of a system where adaptive antennas are being used to maximize the signal for a given user. Beam forming in a desired direction or creating a null (from co-channel interferers or various multipaths in a TDMA system) as shown in Figure 3-32 is similar in principle. Signals

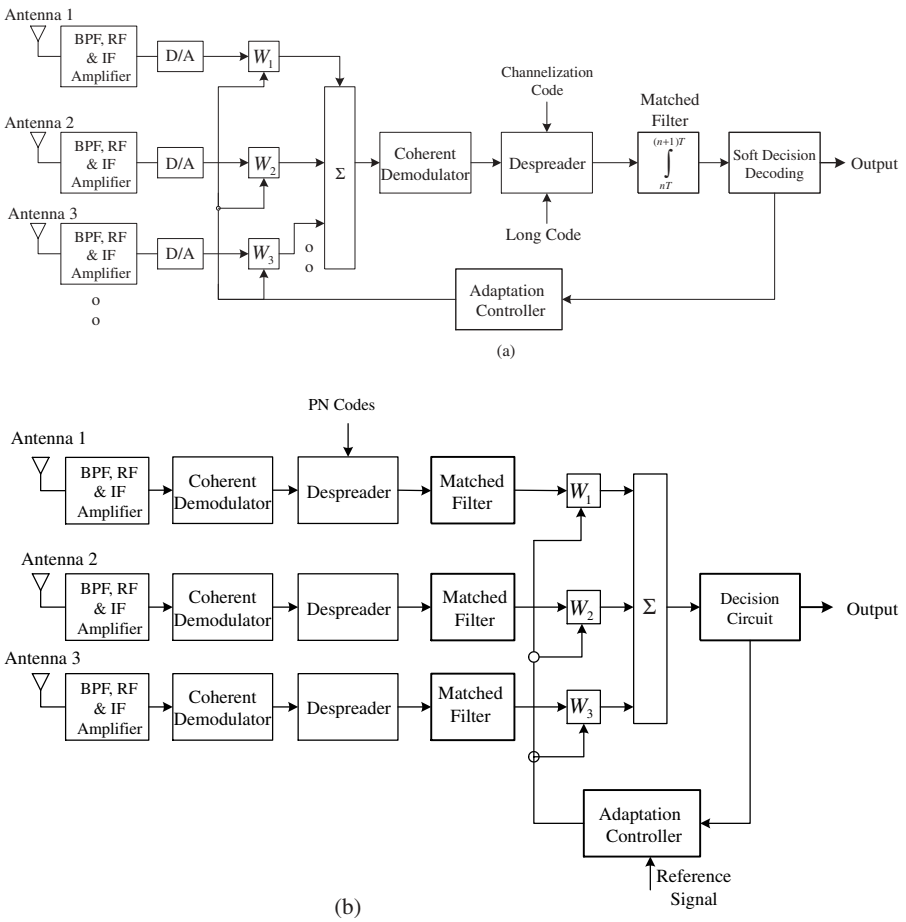


Figure 3-31
A CDMA system using an adaptive antenna array: (a) Beamforming is done at the IF stage. (b) Beamforming is done at the baseband.

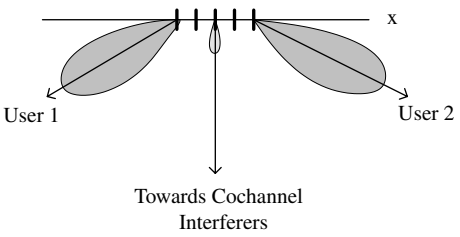


Figure 3-32
Beamforming and steering nulls toward certain directions using adaptive antennas

from various sensors in an antenna array are converted into digital forms, multiplied by weights W_i, summed together, and after coherent demodulation, despread in the usual way using local copies of orthogonal Walsh codes and long user codes. The output

of the matched filter is decoded in a decision circuit. The resulting output is also used by the adaptation controller to adjust the weights so as to maximize the signal-to-interference ratio for the given user in much the same way as a rake receiver, discussed previously.

In this approach, because signals are being weighted and summed at the RF stage, the scheme suffers from the disadvantage that its accuracy is rather limited and that its implementation may become quite complex, particularly when there are many elements in the array. A scheme that performs beamforming at the baseband was shown in Figure 3-31(b). Because signal processing is now being done at the baseband, it is possible to use 16-bit arithmetic compared to a 5- or 6-bit operation that is usual for RF beamforming.

The improvement in performance with adaptive antennas depends upon the antenna type—linear, planar, or circular—the number of elements in the array, and the spacing between adjacent elements. This spacing is usually one half of the carrier wavelength. The improvement in signal-to-interference ratios is about 3 dB with two elements, 6 dB with four elements, 7.75 dB with six elements, and 9 dB with eight elements [42].

Summary

In this chapter we have presented fundamental principles of CDMA and more specifically W-CDMA. The various functional components of a BS transmitter have been discussed in some detail. The receiver structure, soft decision decoding of convolutional codes, methods of multiuser detection at a BS, and smart antennas have been described. In some cases, for the convenience of readers, details have been moved to the following appendices.

Appendix A—Viterbi Decoding of Convolutional Codes

The Viterbi algorithm performs sequential decoding using principles of dynamic programming [9], [11]. The algorithm is based on the fact that if at any instant t_k, there is a sequence of m information bits for which the decoder performance is optimum, then those m bits will be the first m bits of a sequence that optimizes the performance at any later instant $t_l > t_k$. Given a sequence of outputs from the matched filter over a desired observation period, a sequence of bits is chosen at each stage as the most likely transmitted sequence.

To continue with the algorithm, suppose that R is a sequence of samples of the matched filter output (which are analog voltages as mentioned before). At each symbol period, the number of samples read by the decoder equals the number of output bits generated by the encoder for each input bit. That is, for a rate ½ encoder, there are two samples to the input of the decoder at the end of each symbol period. Furthermore, each of these samples is defined by one of the quantization levels R. The maximum likelihood decision theory states that X is the code that was most likely transmitted if the probability of R (assuming X) is maximum, that is, if

$$P(R|X) \text{ is maximum.}$$

To use this algorithm, then, it is first of all necessary to determine the probability of occurrence of each quantization level of the decoder inputs at each symbol period assuming that a transmitted bit is 0. Similarly, the probability of occurrence of each quantization level of the decoder inputs at each symbol period, assuming that a transmitted bit is 1, is determined in the same manner. Because these probabilities will be used at each step for sequential decoding,

it is better to convert them into some suitable numbers that would speed up the computation process. Specifically, suppose that

$$p(r_{kj}|x_{kl}) \tag{A-1}$$

is the probability of occurrence of the *j-th* quantization level of the matched filter output at instant *k,* assuming that the transmitted bit is x_{kl}, where x_{kl} is either a 0 or 1 at any symbol period. As mentioned earlier, because an encoder of rate $1/2$ generates two output bits for each bit of the input, j takes only two values: 1 or 2, and so the probability of a code symbol for a path in the trellis diagram is a product of two terms of type (A-1). In other words,

$$P_k = p(r_{kl}|x_{kl}) \times p(r_{k2}|x_{kl}) \tag{A-2}$$

It is, therefore, convenient to take the logarithm of expression (A-2) and, for ease of computation, transform the result into an integer using an appropriate expression. This value can then be used as a metric for a path. In this way, the branch metrics for all paths of the trellis diagram are computed.

For an encoder with m registers, the number of states for the trellis diagram is 2^{m-1}. For instance, for the diagram of Figure 3-7, $m = 3$, and the number of states is 4. Referring to the trellis diagram of Figure 3-9, the Viterbi algorithm can be summarized in the following way:

1. Starting at state 00 of Figure 3-17 (at depth 3 or beyond), add the metrics of the two paths coming to this state to the previously saved metrics of the two states (namely 00 and 01) from which these two paths have originated.
2. Choose the larger of the two-path metrics computed in step 1 and save it. This becomes the new path metric for this state (that is, state 00) for subsequent use. The branch that gives the larger path metric is called a *survivor path*. Identify this path by adding a 0 to the path history if state 00 has a larger metric. Otherwise, add 1 to the path history. This path history is saved in memory for use in the next step.
3. Repeat steps 1 and 2 for all other states at the same trellis depth.

4. The path with the largest metric gives the desired decoded bit.

Clearly, the number of survivor paths at each iteration is equal to the number of states of the trellis. Eventually, however, at the end of a transmitted sequence, it is necessary to choose only one of these four possible paths corresponding to the most likely transmitted code. This is easily done by adding two 0's ($m - 1$ 0's in a general case) to the end of the information sequence at the encoder input. Because in this case the final survivor path must terminate at state 00, the desired path is the one that ends at this state after the last four encoder output bits have been received and decoded.

Figure 3-33 gives the bit error rate performance of convolutional codes of rate ½ for two values of the constraint length, K = 4 and K = 8 using a quantization level of 8 and assuming Gaussian noise [9]. Referring to Figure 3-15, the value of E_b/N_0 required for a bit

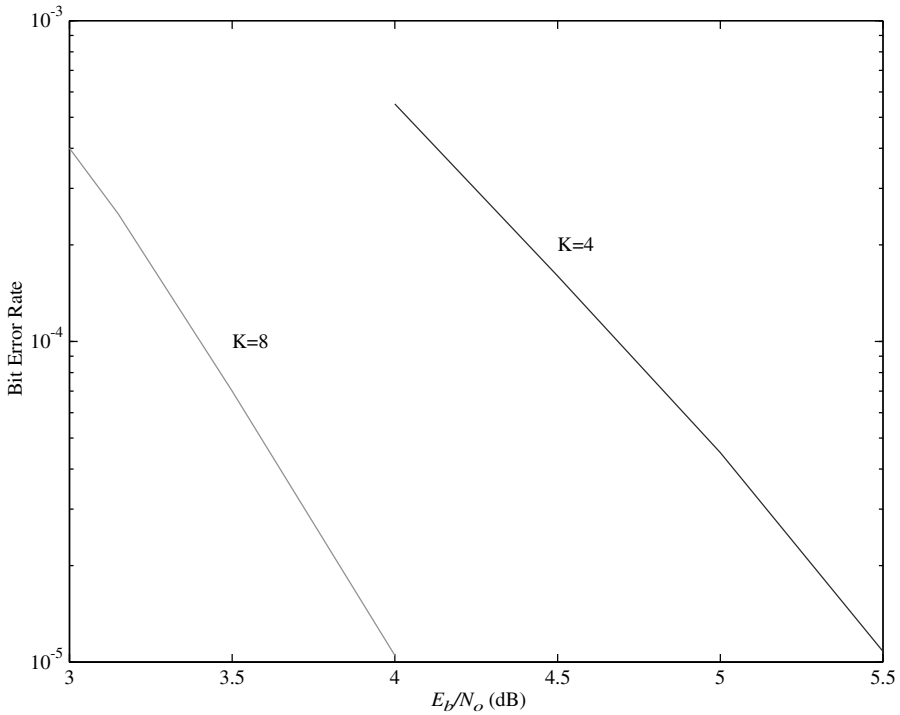


Figure 3-33
Bit error rate of convolutional codes with constraint length K = 4 and K = 8. The quantization level used is 8. [From paper by Heller and Jacobs (1971). © 1971 IEEE]

error rate of 10^{-4} for BPSK without coding is about 8.5 dB, whereas with a convolutional code of rate ½ and constraint length 8, the required value of E_b/N_0 is only 3.3 dB. Thus, the coding gain is about 5.2 dB. Notice, however, that the net information rate with this code is reduced by a factor of 2.

Appendix B—Modulation

QPSK

In digital phase modulation or phase shift keying, as it is called, the phase of the carrier is modulated by the digital data stream. To do this, the incoming serial data is first converted into symbols. The number of bits in a symbol may vary. For example, in BPSK, each incoming bit makes a symbol. In QPSK, each successive pair of bits constitutes a symbol, and so on. In general, if a symbol consists of m bits of digital data, the number of distinct symbols is $N = 2^m$. Each symbol is then transmitted by setting the absolute phase angle of the carrier to an appropriate value between 0 and 2π. More specifically, the absolute phase angle of the carrier corresponding to the n-th symbol is given by

$$\theta_n = \frac{(2n - 1)\pi}{N} \text{ with } n = 1, \ldots, N. \qquad \text{(B-1)}$$

For instance, with QPSK, $N = 4$, and the phase angles are $\pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$. The phase transitions are shown as a constellation in Figure 3-34. The lines connecting the symbol positions indicate how the phase may change with incoming symbols. For example, assume that the present symbol is (0,0). In this case, the phase angle is 45 degrees. If the next symbol is also (0,0), the phase angle remains the same as before. If, instead, it is (1,0), the phase changes to 135 degrees, and so on. Notice how the symbols have been arranged in the constellation diagram. With this arrangement, the most probable errors involve only one bit. For instance, in the pres-

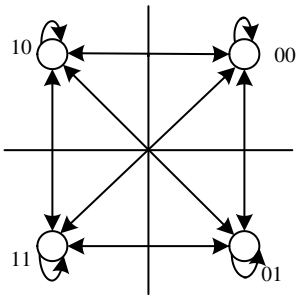


Figure 3-34
The phase transitions of the carrier frequency in QPSK modulation

ence of noise, a transmitted symbol (0,0) might be mistakenly decoded at the receiver as (1,0) or (0,1), and with much lower probability as (1,1).

Offset QPSK (OQPSK)

As mentioned earlier, to perform QPSK modulation, the incoming data is usually split into two streams—the odd bits forming an *in-phase* (I) channel and the even bits forming a *quadrature* (Q) channel. Each stream then modulates the carrier using BPSK. In IS-95, the Q-channel data on reverse channels is delayed by one half of a chip period before modulating the carrier. This is called *offset QPSK* (OQPSK). See Figure 3-35. Phase transitions in OQPSK modulation are shown in Figure 3-36. Because the modulated signals of the I and Q channels undergo phase changes at different instants, the maximum change in the phase angle is only 90 degrees. Thus, even though the output of the wave-shaping filter does not have a constant amplitude all the time, it never goes through 0 (compare Figure 3-34 and Figure 3-36), and is, therefore, more suitable for amplification by a somewhat nonlinear amplifier without producing any spurious side bands.

Differential QPSK (DQPSK)

In the previous definition, each modulating symbol was transmitted using an absolute phase of the carrier. In *differential* DQPSK, an

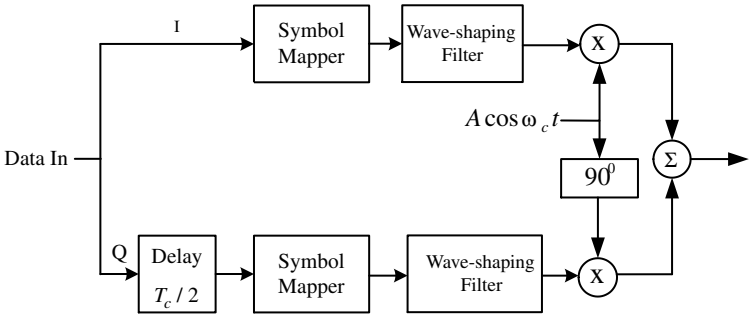


Figure 3-35 OQPSK. This is used on reverse channels in IS-95.

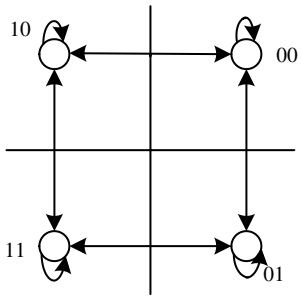


Figure 3-36 Phase transitions in OQPSK modulation

incremental change in the phase instead of an absolute value is used to transmit a symbol. In other words, if θ_{n-1} is the phase of the carrier corresponding to symbol $n - 1$, the phase angle for symbol n is given by

$$\theta_n = \theta_{n-1} + \Delta\theta_n$$

where $\Delta\theta_n = \dfrac{(2n + 1)\pi}{N}$ is the incremental phase change corresponding to the n-th symbol [13]. For example, with $N = 4$,

$$\Delta\theta_n = \begin{cases} \pi/4 \text{ for symbol } (0,0) \\ 3\pi/4 \text{ symbol } (0,1) \\ 5\pi/4 \text{ symbol } (1,0) \\ 7\pi/4 \text{ symbol } (1,1) \end{cases}$$

Notice that in this case, phase changes occur at each symbol period regardless of the incoming data pattern, but not so in Figure 3-34 or 3-36.

Appendix C—Multiuser Detection Using Viterbi Algorithm

In this appendix, we will further expand our ideas behind multiuser detection, and discuss the detection principles based on Viterbi algorithm, using broad, general concepts. For a detailed mathematical analysis of the subject, see references [21]-[24].

Because of its complexity, multiuser detection is more amenable to implementation at a base station rather than a mobile station. First, consider a synchronous CDMA system. Since it uses a system-wide timing reference based on the Global Positioning System (GPS), symbols transmitted by individual mobile stations are synchronous. Thus, even though they undergo variable delays as they arrive at the base station, these delays are usually quite small compared to the symbol period, and therefore, the cross-correlation between scrambling codes assigned to various users is also very small. In this case, with perfect power control, the output of the matched filter corresponding to any user at the end of a symbol period depends only on the signal from that user. If, however, the power control is not perfect, the weaker signals may be swamped by the stronger signals, and as a result the bit error rates for the weaker channels will be high.

In an asynchronous system, on the other hand, as we mentioned previously, time offsets between signals received from multiple users may be comparable to the symbol period. Thus, any symbol of the desired user may overlap with one or more successive symbols from all other users. Because the cross-correlation between scrambling codes is no longer zero, the matched filter output from any given user depends not only on the signal from that user but also on signals received from all other users over a few consecutive symbol periods.

Figure 3-37 shows a channel model describing the signal received at a base station. Here, the user data is mapped by the symbol mapper to a bipolar signal. The resulting data stream, say, $\{s_1(i)\}$ from user 1 is spread out by $y_1(t)$, the PN code sequence for this user. A_1 is the transmitted signal amplitude, ω_c its carrier frequency which is same for all users, ϕ_1 the phase of the carrier, τ_1 the delay and $n_1(t)$ the noise introduced by the channel. Similarly, $\{s_2(i)\}$, $y_2(t)$, A_2, ϕ_2, τ_2

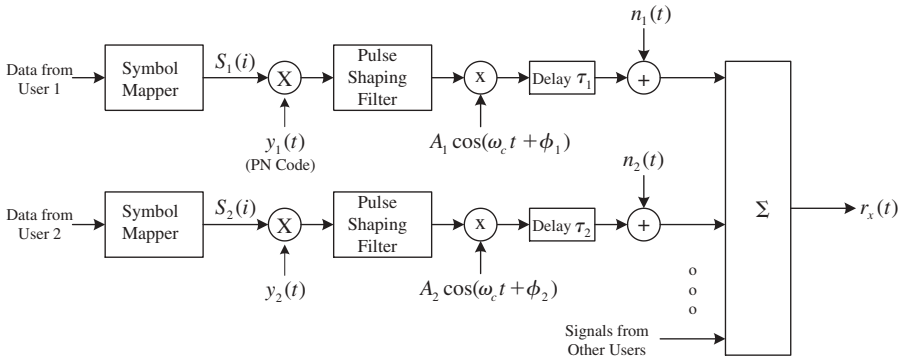


Figure 3-37
Channel model describing the signal received at a base station from multiple users

and $n_2(t)$ are the corresponding parameters for user 2, and so on. The channel noise is assumed to be Gaussian. T is the symbol period.

Figure 3-38 shows the base station receiver that uses matched filters and a soft decision decoder. To detect the signal from any user, say, user 1, the demodulated output of the low pass filter is multiplied by its PN code, that is, $y_1(t)$. The resulting signal $r_{d1}(t)$ is applied to the input of the matched filter, where it is integrated over each symbol period, and the output read into the decoder at the end of each integration cycle.

It may be intuitively clear from Figures 3-37 and 3-38 that the output of the matched filter corresponding to user 1 at the end of the j-th symbol period may be expressed as

$$r_{o1}(jT) = s_1(j) + s_2(j-1)R_{12}(T) + s_2(j)R_{12}(0) + s_2(j+1)R_{12}(-T) + \ \ldots \ + n_1(t) \quad \text{(C-1)}$$

where $n_1(t)$ is the base band noise. The dots in (C-1) indicate that there are similar terms accounting for the interference due to users 3, 4, and so on. Notice in expression (C-1) that the filter output at the end of any symbol period depends on the present bit of this user and three bits of user 2: the present, the previous and the next. The reason for this dependence on three bits is that the mobile radio channel is time-varying, and that based on the relative delays, the signal from user 2 may arrive at the base station either earlier or later with respect to the signal from user 1. Also, the interference due to distant symbols such as $s_2(j \pm 2)$, $s_2(j \pm 3)$, and so on, are ignored because it

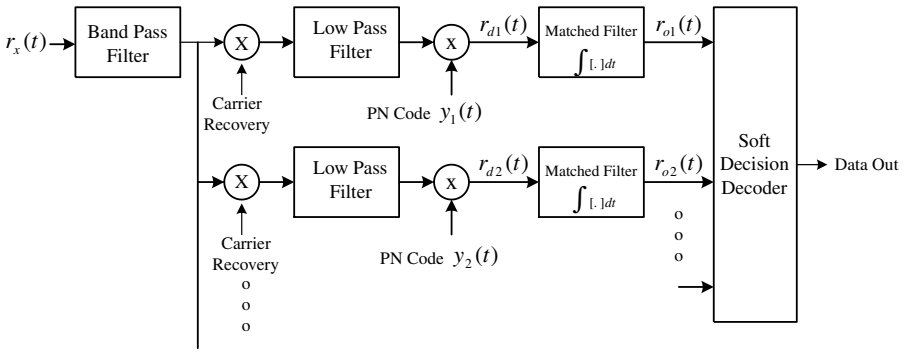


Figure 3-38
The base station receiver model used in a multiuser detection

is assumed that the relative delays between signals from any two users, say, $\tau_i - \tau_j \leq T$. Here, the auto-correlation function of a scrambling code $y_1(t)$ is

$$\frac{1}{T}\int_{jT}^{(j+1)T} y_1(t)y_1(t)dt = 1 \qquad \text{(C-2)}$$

The cross-correlation $R_{12}(.)$ of the two PN codes $y_1(t)$ and $y_2(t)$ is given by

$$R_{12}(jT) = \frac{1}{T}\int_{jT}^{(j=1)T} y_1(t)y_2(t = jT = \tau)dt \qquad \text{(C-3)}$$

Similarly, the output of the matched filter for user 2 is:

$$r_{o2}(jT) = s_2(j) + s_1(j-1)R_{21}(T) + s_1(j)R_{21}(0) + s_1(j+1)R_{21}(-T) + \ \ldots \ + n_2(t) \qquad \text{(C-4)}$$

Expressions (C-1) and (C-4) can be represented by a 2-tap delay line. Figure 3-39 shows this delay line representation assuming only two users.

Symbols can now be decoded using a soft decision decoder. In fact, the Viterbi algorithm can be used to decode them in much the same way as for convolutional codes. Since this algorithm has been previously described, we will simply mention here that the trellis diagram for this two-user model has 16 states corresponding to the current and previous symbols received from each user. The state transitions are caused by next symbols. The metric associated with each branch

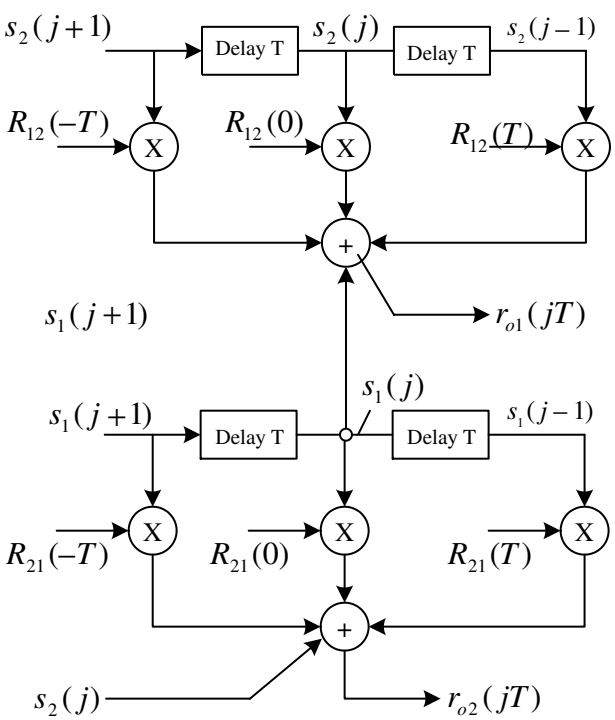


Figure 3-39
Two-user delay line model in multiuser detection. The maximum delay is assumed to be T, the symbol period.

of the trellis is given by values of the cross-correlation functions. However, now, at each state, the survivor path is the one whose metric is closest to the output of the matched filter.

In UMTS, both long and short codes may be used on uplinks. However, short codes are better from the standpoint of multiuser detection because their cross-correlation remains constant over a number of consecutive symbol periods. Because the number of states in the trellis diagrams, and consequently the computational complexity, increase exponentially with the number of users, the procedure is not very useful in practical applications.

References

[1] A.J. Viterbi, "The Evolution of Digital Wireless Technology from Space Exploration to Personal Communication Ser-

vices," *IEEE Trans. Veh. Technol.,* Vol. 43, No. 3, pp. 638–644, August 1994.

[2] N. Abramson, "The Throughput of Packet Broadcasting Channels," *IEEE Trans. Comm*., Vol. COM-25, No.1, pp. 117–128, January 1977.

[3] L. Kleinrock and F.A. Tobagi, "Packet Switching in Radio Channels: Part I—Carrier Sense Multiple Access Modes and Their Throughput-Delay Characteristics," *IEEE Trans. Comm*., Vol. COM-23, No.12, pp. 1400–1416, December 1975.

[4] L. Kleinrock and F.A. Tobagi, "Packet Switching in Radio Channels: Part II—The Hidden Terminal Problem in Carrier Sense Multiple Access Modes and the Busy Tone Solution," *IEEE Trans. Comm*., Vol. COM-23, No.12, pp. 1417–1433, December 1975.

[5] N.D. Wilson, et al., "Packet TDMA Versus Dynamic TDMA for Multiple Access in Integrated Voice/Data PCN," *IEEE JSAC*, Vol. 11, No.6, pp. 870–83, August 1993.

[6] G.L. Turin, "Introduction to Spread Spectrum Antimultipath Techniques and Their Application to Urban Digital Radio," *Proc. IEEE*, Vol. 68, No.3, pp. 328–353, March 1980.

[7] R.L. Pickholz, L.B. Milstein, and D. L. Schilling, "Spread Spectrum for Mobile Communications," *IEEE Trans. Veh. Technol.,* Vol. 40, No.2, pp. 313–322, May 1991.

[8] A.J. Viterbi, "Convolutional Codes and Their Performance in Communications Systems," *IEEE Trans. Comm. Tech.,* Vol. COM-19, No. 5, pp. 751–772, October 1971.

[9] J.A. Heller and I.M. Jacobs, "Viterbi Decoding for Satellite and Space Communication," *IEEE Trans. Comm. Tech.,* Vol. COM-19, No. 5, pp. 835–848, October 1971.

[10] G.D. Forney, "The Viterbi Algorithm," *Proc. IEEE*, pp. 268–278, March 1973.

[11] M.C. Jeruchim, P. Balaban, and K.S. Shanmugan, *Simulation of Communication System.* New York: Plenum Press, 1992.

[12] 3GPP TS 25.213: UMTS; Spreading and Modulation, 2000.

[13] R.W. Lucky, J. Salz, and E.J. Weldon, *Principles of Data Communications.* New York: McGraw-Hill, 1968.

[14] S.W. Golomb, *Shift Register Sequences.* Aegean Park Press, 1982.

[15] S. Lin, *An Introduction to Error-Correcting Codes*. New Jersey: Prentice Hall, 1970.

[16] M.W. Oliphant, "Radio Interfaces Make the Difference in 3G Cellular Systems," *IEEE spectrum*, pp. 53–58, October 2000.

[17] G.J.R. Povey, et al., "A Decision-Directed Spread-Spectrum RAKE Receiver for Fast-Fading Mobile Channel," *IEEE Trans. Veh. Technol.,* Vol. 43, No. 3, pp. 491–502, August 1994.

[18] J.S. Lehnert and M.B. Pursley, "Multipath Diversity Reception of Spread-Spectrum Multiple-Access Communications," *IEEE Trans. Comm*., Vol. COM-35, No. 11, pp. 1189–1198, November 1987.

[19] T. Ojanpera and R. Prasad (Ed.), *Wideband CDMA for Third Generation Mobile Communications.* Boston: Artech House, 1998.

[20] H. Holma and A. Toskala (Ed.), *W-CDMA for UMTS*. New York: John Wiley, 2000.

[21] S. Glisic and B. Vucetic, *Spread Spectrum CDMA Systems for Wireless Communications*. Boston: Artech House, 1997.

[22] S. Verdu, "Minimum Probability of Error for Asynchronous Gaussian Multiple-Access Channels," *IEEE Trans. Inform. Theory*, Vol. IT-32, pp. 85–96, January 1986.

[23] S. Verdu and H.V. Poor, "Abstract Dynamic Programming Models under Commutativity Conditions," *SIAM J. Control Optimization*, Vol. 24, pp. 990–1006, July 1987.

[24] R. Prasad, *CDMA for Wireless Personal Communications*. Boston: Artech House, 1996.

[25] R.H. Kohno, H. Imai, and M. Hatori, "Cancellation Techniques of Co-Channel Interference in Asynchronous Spread Spectrum Multiple Access Systems," *Trans. IECE (Electronics and Communications in Japan),* Vol. 66, pp. 416–423, May 1983.

[26] R.H. Kohno, H. Imai, M. Hatori, and S. Pasupathi, "Combination of an Adaptive Array Antenna and a Canceller of Interference for Direct-Sequence Spread-Spectrum Multiple Access Systems," *IEEE J. Selected Areas Comm.*, Vol. 8, No. 4, pp. 675–682, May 1990.

[27] Y.C. Yoon, R. Kohno, and H. Imai, "Spread-Spectrum Multiple Access System with Co-Channel Interference Cancellation for Multipath Fading Channels," *IEEE J. Selected Areas Comm.*, Vol. 11, No. 7, pp. 1067–1075, May 1992.

[28] P. Patel and J. Holzman, "Analysis of Simple Successive Interference Cancellation Scheme in DS/CDMA-System," *IEEE J. Selected Areas Comm.*, Vol. 12, No.5, pp. 796–807, June 1994.

[29] H. Imai, *Essentials of Error-Control Coding Techniques*. New York: Academic Press, 1990.

[30] N. Abramson, "Multiple Access in Wireless Digital Access Networks," *Proc. IEEE*, Vol. 82, No. 9, pp. 1360–69, September 1994.

[31] D.L. Schilling, "Wireless Communication Going into the 21st Century," *IEEE Trans. Veh. Technol.,* Vol. 43, No. 3, pp. 645–652, August 1994.

[32] U. Grob, A.L. Welti, E. Zollinger, R. Kung, and H. Kaufman, "Micro-Cellular Direct-Sequence Spread Spectrum Radio System Using N-Path Rake Receiver," *IEEE J. Selected Areas Comm.*, Vol. 9, pp. 772–780, June 1990.

[33] N.L.B. Chan, "Multipath Propagation Effects on a CDMA Cellular System," *IEEE Trans. Veh. Technol.,* Vol. 43, No. 4, pp. 848–855, November 1994.

[34] T.S. Rappaport, *Wireless Communications.* New Jersey: Prentice Hall, 1996.

[35] J.C. Liberti and T. S. Rappaport, *Smart Antennas.* New Jersey: Prentice Hall, 1998.

[36] R.T. Compton, *Adaptive Antennas—Concepts and Performance*. New Jersey: Prentice Hall, 1996.

[37] S.C. Swales, et al., "The Performance Enhancement of Multibeam Adaptive Base Station Antennas for Cellular Land

Mobile Radio Systems," *IEEE Trans. Veh. Technol.,* Vol. VT-39, No. 1, pp. 56–67, February 1990.

[38] S. Anderson, et al., "An Adaptive Array for Mobile Communications System," *IEEE Trans. Veh. Technol.,* Vol. VT-40, No. 1, pp. 230–236, February 1991.

[39] J.H. Winters, "Signal Acquisition and Tracking with Adaptive Arrays in the Digital Mobile Radio System IS-54 with Flat Fading," *IEEE Trans. Veh. Technol.,* Vol. VT-42, No. 4, pp. 377–384, November 1993.

[40] J. Kennedy, et al., "Direction Finding and 'Smart Antennas' Using Software Radio Architectures," *IEEE Commun. Mag*., Vol. 33, No.5, pp. 62–68, May 1995.

[41] G. Tsoulos, et al., "Wireless Personal Communications for the 21st Century: European Technological Advances in Adaptive Antennas," *IEEE Commun. Mag.,* Vol. 35, No. 9, pp. 102–109, September 1997.

[42] V.K. Garg, et al., "Application of Adaptive Array Antenna to a TDMA Cellular/PCS System," *IEEE Commun. Mag.,* Vol. 35, No. 10, pp. 148–152, October 1997.

[43] ETSI Draft EN 301 704 Version 7.1.0, Adaptive Multi-Rate Speech Transcoding, 1999.

[44] N.S. Jayant, "High-Quality Coding of Telephone Speech and Wideband Audio," *IEEE Comm. Mag.,* Vol. 28, No. 1, pp. 10–19, January 1990.

[45] P. Vary, et al., "Speech Codec for the European Mobile Radio System," *Proc. ICASSP '88*, pp.227–230, April 1988.

[46] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, Vol. 63, pp. 561–80, April 1975.

CHAPTER 4

cdmaOne and cdma2000

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

As mentioned in Chapter 1, "Introduction," *first-generation* (1G) mobile telecommunication systems in the 1980s were analog, and consisted of cellular system TIA/EIA-553 in the United States operating around 850 MHz, and *Total Access Communication System* (TACS), and *Nordic Mobile Telephone* (NMT) in Europe operating at 450 and 900 MHz bands. The *second-generation* (2G) systems are based on IS-136, IS-95A, IS-95B, and GSM, and have the data transport capability, but only to a limited extent. For example, GSM supports *short messaging services* (SMS) and user data at rates only up to 9.6 kb/s. With IS-95B, it's possible to provide data rates in the range of 64 to 115 kb/s in increments of 8 kb/s over a 1.25 MHz RF bandwidth.

To overcome this limitation and, particularly, to be able to provide multimedia services, the *International Telecommunications Union—Radio Communication Sector* (ITU-R) published in 1999 a set of standards for *third-generation* (3G) wireless systems [1], [2], [5], [7]. These systems include cdma2000, *Universal Mobile Telecommunications System* (UMTS) *Wideband CDMA* (W-CDMA) FDD, UMTS W-CDMA TDD, and *Time Division Multiple Access* (TDMA) system known as *Universal Wireless Communication-136* (UWC-136).

The purpose of this chapter is to describe cdma2000. One of the fundamental requirements of 3G standards is to allow for the graceful evolution of current, 2G wireless networks. In fact, cdma2000 is an evolution of the present North American CDMA system called *cdmaOne*. Thus, we shall begin with a brief description of cdmaOne.

cdmaOne

Spectrum Allocation

Present CDMA systems in the United States, which are known as cdmaOne, are based upon IS-95 standards [6], [7]. The spectrum allocation is shown in Figure 4-1. The allocation is 50 MHz for cellular systems and 120 MHz for *Personal Communications Services* (PCS). The spectrum is divided into a number of bands as shown in the fig-

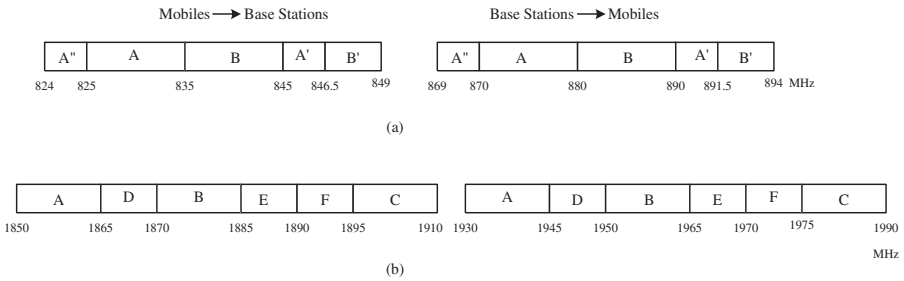


Figure 4-1
The spectrum allocation in IS-95: (a) Cellular (b) PCS

ure. Only those combinations of these bands that are considered as valid by the FCC may be used by a service provider.

The available spectrum may be thought of as consisting of a number of channels with a nominal spacing of 30 kHz for a cellular system and 50 kHz for a PCS. Thus, as an example, there are 1,200 FDD channels in a PCS. Starting from a given channel, a CDMA carrier may include a number of these channels. However, for satisfactory operation, CDMA carriers should be separated by at least 25 channels or 1.25 MHz for PCS. Thus, the nominal bandwidth of a CDMA system is 1.25 MHz.

Physical Channels

In the *uplink*, there are two physical channels—the access channel and traffic channel. A mobile station may use an access channel to send signaling messages such as a call origination request, a page response, a registration request, an order message, and so on. A system may have one or more access channels, each associated with a paging channel. A traffic channel carries user traffic, such as speech or data, and may also be used during a call to send signaling messages, such as a handoff completion or a pilot strength measurement message, report power measurements to the base station, and so on.

In the *downlink*, there are four classes of physical channels—a pilot channel, a sync channel, up to 7 paging channels, and up to 55 traffic channels. In another possible configuration, the sync and paging channels may be replaced by traffic channels.

For each active forward CDMA channel, there is a pilot channel that continuously sends a carrier modulated by an all-zero Walsh code so that mobile stations can synchronize to a base station. This signal can also be used as a reference in coherent demodulation and timing recovery at a mobile station. Similarly, a mobile station can measure the signal strength of a pilot channel, and report the result to the serving base station so that it can decide if a handoff is necessary.

The sync channel transmits information that enables mobile stations within the coverage area of a base station to acquire frame synchronization after achieving the pilot channel synchronization. The information rate on the sync channel is 1,200 b/s. The channel rate, however, is 4.8 kb/s because the synchronization stream is encoded into a convolutional code of rate ½ and then repeated once before being interleaved. The sync channel superframe is 80 ms, while a traffic channel frame is 20 ms.

A paging channel carries system overhead information, such as system parameters, access parameters, a CDMA channel list, a neighbor list, and so on, as well as messages directed to mobile stations such as a general page message, alert, abbreviated alert, release, reorder, and so on. The information rate on a paging channel may be either 9.6 kb/s or 4.8 kb/s. The incoming data is encoded into a covolutional code of rate ½. Encoded symbols are repeated if the data rate is 4.8 kb/s and interleaved before being spread by an orthogonal code.

The purpose of a forward traffic channel is to send the user data as well as signaling messages to a mobile station during a call. The information rate on this channel may be 8.6 kb/s, 4.0 kb/s, 2.0 kb/s, and 0.8 kb/s.

We will now describe the transmit functions of a cdmaOne system.

Reverse Channel Transmit Functions

A functional block diagram of transmit functions on a reverse traffic channel is shown in Figure 4-2. User data streams, which may originate at different rates, are arranged in 20 ms frames. If the data

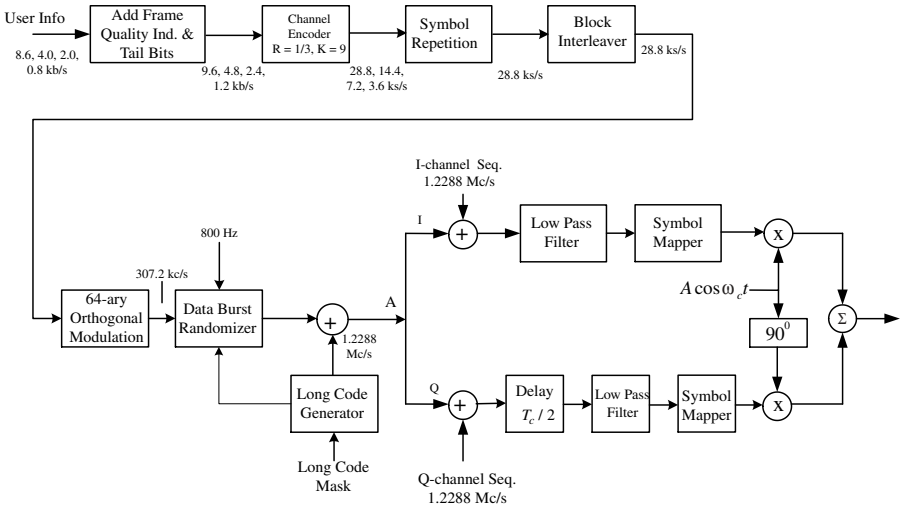


Figure 4-2
Transmit functions of an IS-95 reverse channel

rate is 8.6 or 4.0 kb/s, some frame-quality indicators are added to the incoming frames. To ensure proper operation of the channel encoder, all incoming frames are appended with some tail bits and then passed to a convolutional encoder of rate 1/3 and constraint length 9. Because we would like to have a constant rate at the input of the interleaver regardless of the user data rates, the encoder output for lower data rates (4.0 kb/s and less) is repeated on a bit-by-bit basis in the symbol repetition block so that its output is 28.8 k/s. Thus, the output of the symbol repeater consists of two copies of the code symbols if the user data rate is 4 kb/s, four copies if the incoming data rate is 2.0 kb/s, and so on. This output is applied to a block interleaver of depth 20 ms so that any two adjacent bits at the input are spaced 1.1111 ms apart at the output.

The output of the interleaver is passed through an orthogonal modulator, where each 6-bit code symbol is converted into a 64-bit modulation symbol using a 64 × 64 Walsh matrix. The modulator output feeds into a data burst randomizer that allows only one copy of the code symbols to be transmitted. To this end, each 20 ms frame of the interleaver output is divided into 16 blocks, or *power control groups* as they are called, as shown in Figure 4-3(a). The transmitter is then gated on during only some of these groups, depending upon

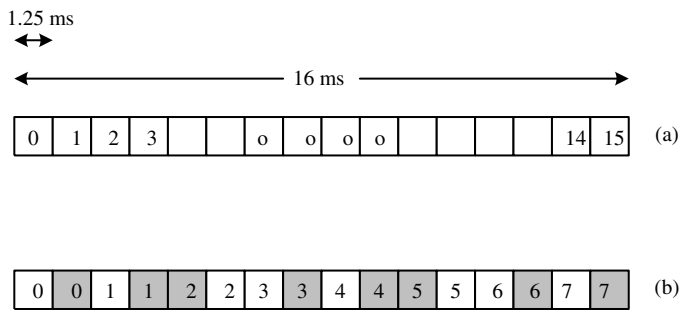


Figure 4-3
Data burst randomization

the user data rate. For example, if the rate is 9.6 kb/s, the transmitter must be gated on all the time. If it is 4.8 kb/s, the transmitter should be activated only half the time, and so on. Figure 4-3(b) shows the power control groups for a data rate of 4.8 kb/s. Notice that the first two groups marked 0 and 0 have identical data streams. Similarly, groups 1 and 1 are identical, and so on. Because only one-half of the code symbols are to be transmitted, in a simple case, the transmitter could be gated on during each alternate group. Instead, the data burst randomizer randomly selects the groups during which the transmitter is gated on, making sure that a given symbol is transmitted exactly once. The random selection is indicated by the shaded groups of Figure 4-3(b).

Referring again to Figure 4-2, the output of the data burst randomizer is spread by a long code, which is derived in the following way. The output of a maximal length shift register sequence with a characteristic polynomial of degree 42 is ANDed with a long code mask that is constructed with the permuted electronic serial number (ESN) of the mobile station. Thus, the long code mask and hence the output of the long code generator are unique for each user. Because the randomizer output is 307.2 kc/s and the chip rate is 1.2288 Mc/s, each bit is spread by a factor of 4.

The resulting output is divided into two sequences, the I and Q sequences, which are spread by a zero-offset, I and Q pilot pseudonoise (PN) sequences of period $2^{15} - 1$ (chips). To minimize the out-of-band energy, the resulting outputs are passed through a low-pass

filter with a nominal bandwidth of about 740 kHz. The filtered outputs are symbol-mapped and then modulate the carrier.

Notice that the Q-channel data, after spreading by the Q-channel pilot PN sequence, is delayed by $T_c/2$ before it is filtered, where T_c is the chip period. This is known as *offset QPSK* (OQPSK) modulation. Because the I channel and Q channel are now delayed by this amount, only one of them undergoes a phase transition at a time. Since, in this case, the maximum phase change that can take place at any time is only ± 90 degrees, the envelope of the modulated signal never goes through zero, and furthermore, varies much less than for QPSK. Consequently, the resulting signal is more suitable for amplification with nonlinear amplifiers.

The transmit functions of the reverse access channel are slightly different. For example, there is only one data rate for this channel, namely, 4.4 kb/s. Thus, the code symbols are repeated just once. No data burst randomizer is used. Each access channel is spread by a long code, which is derived in the same manner as for the reverse traffic channel except for a different long code mask that includes the access channel number, the paging channel number, the base station identification number, and so on.

Forward Channel Functions

cdmaOne uses a system-wide reference time scale that is based upon a global positioning system synchronized with a universal coordinated time. Each base station derives its time base from this reference time scale. The functional block diagram of a base station transmitter is shown in Figure 4-4.

The pilot channel carries an all-zero pattern and is spread by *Walsh function 0* (W0).

The sync channel is used to transmit a synchronizing sequence at 1.2 kb/s. The data is encoded with a convolutional code of rate ½ and constraint length 9. The output of the channel encoder is repeated once so that the resulting bit rate is 4.8 kb/s. It is passed through a block interleaver and then spread out with *Walsh function 32* (W32).

The paging channel data, which may operate at either 9.6 kb/s or 4.8 kb/s, is encoded in a code of rate ½, symbol-repeated, interleaved,

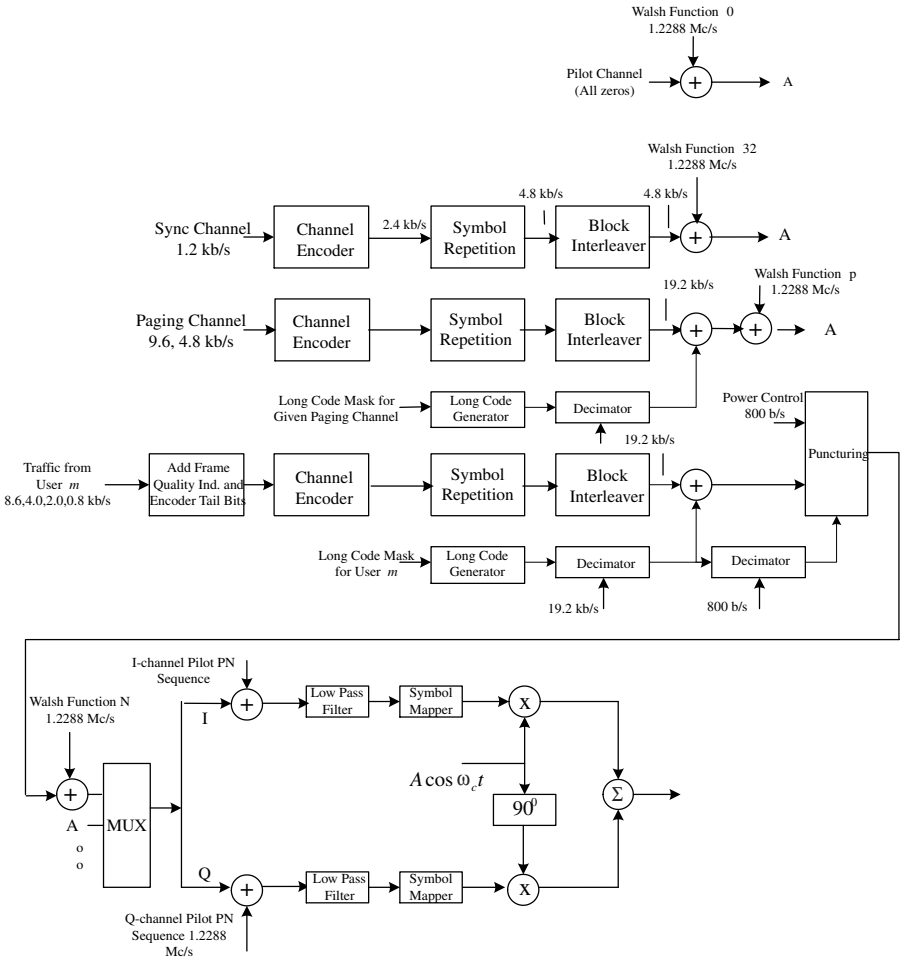


Figure 4-4
Transmit functions on forward channels in IS-95

scrambled, and spread by a Walsh function (W1–W7). The scrambling code is derived by ANDing the output of a 42-bit maximal-length shift register sequence with a 42-bit mask that includes, among other things, a 3-bit paging channel number. The resulting output is passed through a decimator that is clocked at 19.2 kb/s.[1]

[1]Scramblers are generally used to provide encryption and security for the user information and also ensure that there are enough transitions in the channel sequence for timing recovery at the receiver.

Data rates on a forward traffic channel may be 8.6, 4.0, 2.0, or 0.8 kb/s. If it is 8.6 kb/s, 12 bits of frame quality indicators (that is, *Cyclic Redundancy Check* [CRC]) are added to each 20 ms frame by encoding it into a block code using a generator polynomial of degree 12. If the data rate is 4.0 kb/s, 8 bits of CRC are added to each frame by encoding it into a block code with a generator polynomial of degree 8. For other data rates, no such CRCs are added. To reset the encoder, a sequence of all-zero tail bits is appended to each frame. The resulting output is encoded, interleaved, repeated on a symbol-by-symbol basis, and scrambled in the same way as a paging channel. However, in this case, the 42-bit long code mask is constructed with the 32-bit electronic serial number of the particular user.

Once every 1.25 ms, a power control bit with a duration of 2 code symbols is transmitted over a forward traffic channel to indicate to the mobile station whether it should increase or decrease its transmitter power level.[2] If this bit is a 0, the power level should be increased. Otherwise, it is to be decreased. Actually, each power control bit is inserted in the forward traffic channel by replacing, or *puncturing*, as it is called, two adjacent code symbols, which are selected randomly for this purpose.

The I- and Q-channels are spread by two maximal-length pilot PN sequences of period $2^{15} - 1$ (chips) with an offset with respect to a reference PN sequence. This offset, which is unique for each base station, is expressed in terms of the chip rate and is given by $64n$ chips where $0 \leq n \leq 511$. Thus, cdmaOne is a synchronous system where each base station is uniquely identified by the offset index n.

It is worth mentioning here that each forward channel type—pilot, paging, sync, and traffic channels—is separated at a mobile station by means of the Walsh codes. All long codes are the same. However, each paging or forward traffic channel is associated with a unique long code mask. Thus, mobile stations can separate traffic channels by despreading the received signal with a Walsh code (W8–W31, W33–W63) and the user-specific long code.

---

[2]The symbol rate of the interleaver output on a traffic channel is 19,200 symbols/s. Thus, there are 384 code symbols in a 20 ms frame or 24 symbols in each 1.25 ms interval. Each of these intervals is called a *power control group*.

Power Control

Reasons for Power Control One of the most important requirements in a CDMA system is the capability of the *user equipment* (UE) and base transceiver stations to control their transmitter power output. In fact, without this capability, the system will not work effectively and attain its full capacity. To understand the reason, consider the so-called *far-near* problem in a cellular system illustrated in Figure 4-5, where a base station is receiving signals simultaneously from two mobile stations, one close to the base station and the other farther from it near the edge of a cell. Because UE1 is closer than UE2, if both mobile stations are transmitting at the same level, the signal from UE1 will be generally stronger than the signal from UE2. Because all stations in a CDMA cell operate on the same carrier but use only different PN codes, and because the received SIR for UE2 is lower, the detection of this signal at the base station is subject to a higher bit error rate.[3] In fact, depending upon

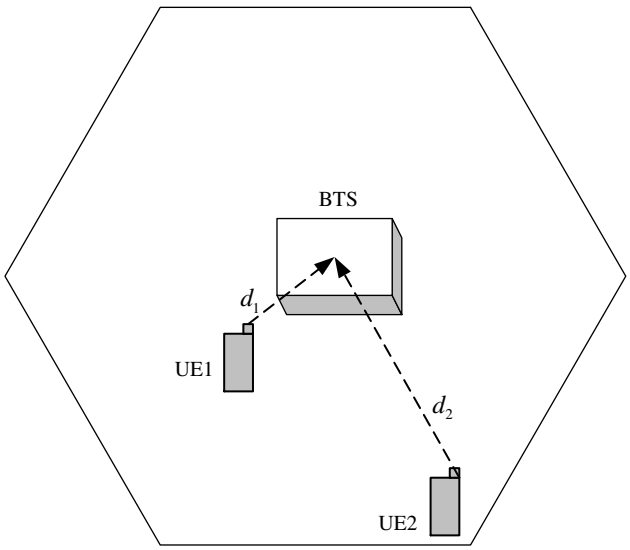


Figure 4-5
The far-near problem in a cellular system

[3]Notice that in a CDMA system, the interfering signal for UE1 is the transmitted signal from UE2 as well as all other stations. Similarly, the interference to UE2 is caused by the transmitted signal from UE1 and all other mobiles.

relative distances d_1 and d_2, the stronger signal may swamp out the weaker signal. Furthermore, because the two signals fade independently, it is possible that the instantaneous value of the signal from UE1 is above its average level while at the same time the signal from UE2 is substantially below its mean, thus making it more difficult for the base station to detect the weaker signal.

To overcome this problem, the base station measures the signal from a mobile station, and if it is above a threshold, it sends a command to that station to reduce its power level so that the signal that the base station receives from that mobile station is equal to signals from other mobile stations. Similarly, if the signal is below a threshold, the user equipment may be asked to increase its transmitter power. However, in this case, there is a possibility of an increased interference to neighboring cells particularly if the mobile station is near the edge of its cell. The power control commands are sent at 800 b/s by puncturing code symbols on a traffic channel once every 1.25 ms (or 16 times per 20 ms frame).

The transmit power control commands can also be used to counteract the effect of fades. Consider, for example, the fading signal of Figure 4-6. The signal is above the local mean during T_1, T_2, and T_3 but below that average at all other times. The bases station can

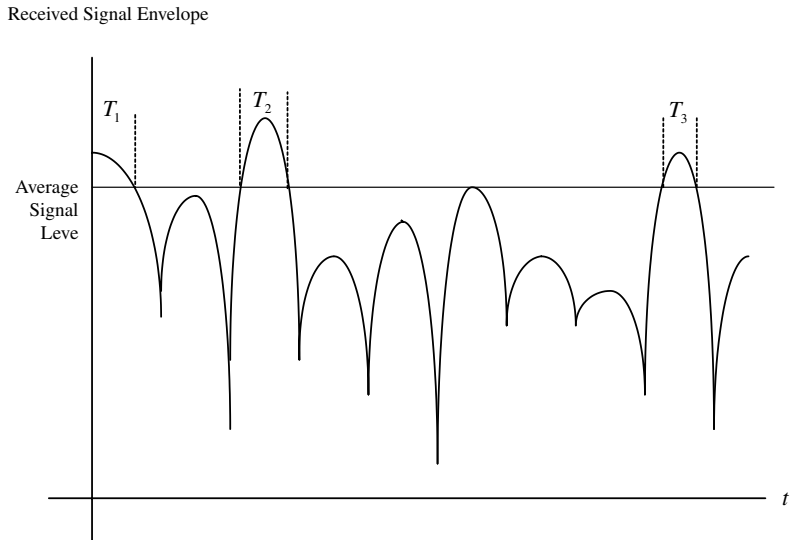


Figure 4-6
A fading signal

remove these variations and receive the signal at a fairly constant level by repeatedly sending transmit power control commands so that the UE reduces the power level during T_1, T_2, and T_3 and increases it at all other times.

Another side benefit of the power control is that mobile stations may now be able to operate at an optimum power level. Thus, CDMA handsets may have a more compact design and longer battery life. It would also result in improved speech quality, particularly in urban and dense urban areas.

Uplink and Downlink Power Control An effective power control is a fundamental requirement for maximizing the capacity of a CDMA system. As we shall see later, the system capacity (that is, the maximum number of mobile stations that can be served in a CDMA system) is maximum when the transmitter power of each mobile station is adjusted such that the received SIR from each mobile is just about enough to maintain the desired quality.

Power control is applied on an uplink (that is, reverse) channel so that the base station receives a satisfactory SIR from each mobile station. To this end, the base station may measure the received signal strength, and if it is outside the limits of satisfactory operation, it orders the mobile station to adjust its transmit power accordingly. This is called a *closed loop power control* because it is based upon direct measurements of the desired signal. Alternatively, we could control the transmitter power on an uplink channel by measuring a different parameter that only gives an indirect indication of the transmitter power. For example, the mobile station could measure the signal received from the base station, estimate the path loss on the forward channel, and accordingly adjust the transmit power of the reverse channel. Clearly, this adjustment works only if the path loss variations on the forward and reverse channels are correlated. This is known as *open loop power control*. Thus, we can use both open and closed loop controls on a reverse channel.

Power control is also used on a downlink channel so that each mobile station receives a satisfactory signal level from the base station. In this case, the algorithms are usually closed loop where each mobile station measures the received signal on the forward channel

and, based upon the measurements, requests the base station to adjust its transmit power.

Handoff in IS-95

As we have indicated elsewhere, when a mobile station moves from the serving area of one base station to the serving area of a second, it should eventually communicate with the second base station because the signal received form this station is stronger. This process of switching communication from one base station to another is called a *handoff*.

There are many different types of handoffs in cellular systems. Consider, for example, an analog *Advanced Mobile Phone System* (AMPS) system or a TDMA system such as IS-136 or GSM. In these systems, each adjacent cell in a cluster is assigned a different set of frequencies. Consequently, if a mobile station moves into an adjacent cell, its transmitter and receiver must begin to operate at new frequencies. This is called a *hard handoff*. In this case, the mobile station must terminate communication with the old base station before starting communication with the new, thereby causing a momentary interruption of the voice or data signal. Similarly, because each sector of a cell in these systems is assigned a different channel set, a hard handover takes place when a mobile station moves from one sector to another.

Things are different in a CDMA system. If two adjacent CDMA systems use two different carrier frequencies, clearly, there will be a hard handoff when a mobile station travels from one system to another. If, however, two adjacent base stations operate on the same CDMA carrier, it may be possible for the mobile station to use traffic channels associated with both base stations at the same time and provide diversity by combining its forward traffic channels, thus reducing the possibility of signal disruptions. In this case, we say that the mobile is in a *soft handoff*.

Even though the mobile is in communication with both stations, only one of them, called the *primary base station*, is responsible for call controls. The other base station is termed a *secondary base*

station. If, at some later time, the signal received from one base station is significantly stronger than the signal from the other base station, the weaker base station may eventually drop out. For example, if based on the pilot strength measurement by the mobile station, the primary base station determines that the secondary base station is weaker, it may request the mobile switching center to drop the secondary base station from the soft handoff. If the primary base station itself is weaker, it indicates to the mobile station that it is going to drop itself from the soft handoff and then transfer control to the target base station. Clearly, a three-way handoff is also possible because mobile stations may be receiving signals with comparable SIRs from three adjoining base stations.

An intracell or intersector handoff is called a *softer handoff*. In this case, if a mobile communicates with two sectors of the same cell, the associated base station can generate an optimum output using a rake receiver.

Handoffs Supported in IS-95 IS-95 supports three types of handoffs:

- *Soft and softer handoff* In this case, all forward traffic channels used by a mobile station operate at the same frequency.
- *Hard handoff* This takes place when the participating base stations use two different CDMA carrier frequencies.
- *CDMA to analog cellular system handoff* This handoff is invoked when a mobile station moves from a CDMA serving area to an adjacent analog system.

Soft Handoff Procedure Recall that in CDMA the pilot channel transmits a continuous, Walsh function W0-modulated carrier, which provides a reference phase that is required for coherent detection and timing recovery at the receiver. Each pilot is identified by a unique offset of the two pilot PN sequences that are used to spread the forward and reverse CDMA channels. Mobile stations monitor the pilots, which have the same CDMA frequency assignment, and use their relative signal strength as a basis for handoff.

The idea behind soft handoffs is rather simple. Each mobile station searches for sufficiently strong pilots that are not associated with any of the forward traffic channels assigned to the particular mobile. If it finds one, it sends the signal strength measurement of that channel to the base station, which can then proceed to assign to this mobile a forward traffic channel associated with that pilot. In this case, the handoff is said to be initiated by the mobile station. Because a base station is capable of measuring signal strengths, it can also trigger a handoff if necessary.

For the purpose of handoff, a mobile station maintains four sets of pilots:

- *Active set* It includes all pilots that are associated with the traffic channels currently being used by this mobile station.
- *Candidate set* It consists of all pilots that are not in the active set but which have been determined by the mobile as sufficiently strong that their associated forward traffic channels can be used by this mobile.
- *Neighbor set* These are pilots that are not included in the previous two sets but are associated with all other base stations in the serving area and thus may be probable candidates for a handoff.
- *Remaining set* Pilots outside of the previous three sets form the remaining set.

The handoff procedure is briefly the following. Refer to Figure 4-7. Base stations specify the size of a search window during which a mobile station should search the pilots. When the signal strength of a pilot in the neighbor set exceeds the pilot detection threshold labeled T_ADD, the mobile station sends the measurement information to the primary base station and moves that pilot to the candidate set.[4] This is shown as time t_1 in the figure.

On receiving the measurement, the base station transmits at t_2 a Handoff Direction message to the mobile station indicating that this pilot should be an active pilot, and also sends a soft handoff request

[4]The signal strength measured is actually E_c/I_0 where E_c is the received pilot energy per chip and I_0 is the total received spectral density due to signal and noise.

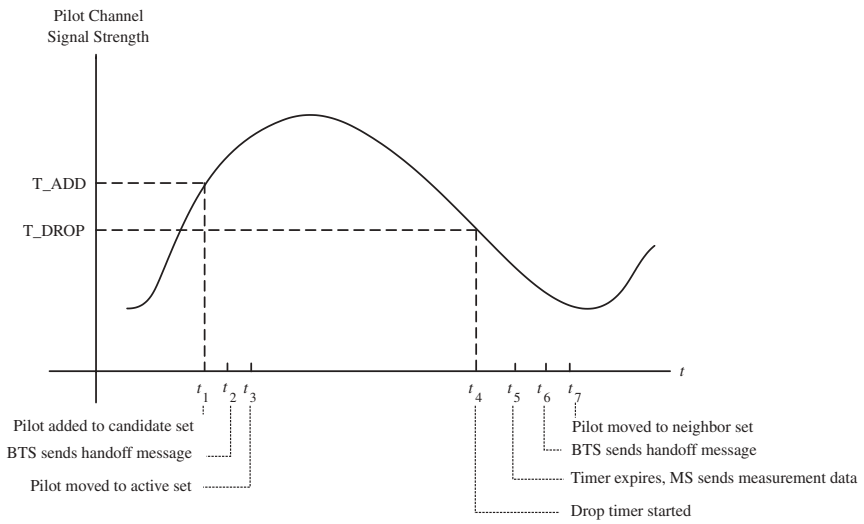


Figure 4-7
Soft handoff in IS-95

message to the base station that uses this particular pilot (that is, the target base station). The target base station may then proceed to join the handoff process and thus exchange necessary messages with the MSC. The mobile station receives the Direction message at t_3, transfers that pilot to the active set, properly updates the candidate set as well, and sends a Handoff Complete message to the primary base station.

From instant t_3, the mobile station continues in the soft handoff state. At instant t_4, when the signal level of an active pilot begins to fall below the pilot drop threshold T_DROP, a timer with a fixed timeout setting is started. If the signal begins to improve back again so that it exceeds T_DROP, the timer is stopped and reset, indicating that this pilot will continue to be active. If, however, the timer expires at instant t_5, and if the signal remains below the threshold for the entire duration from t_4 to t_5 as indicated in the figure, the mobile station sends a pilot strength measurement message to the primary base station.

On receiving the message, say, at instant t_6, the base station sends a Handoff Direction message to the mobile station. Because the forward traffic channel associated with this pilot is no longer usable, the base station sends a release request to the MSC, which forwards

it to the target base station as part of the process to drop it from the soft handoff.

The mobile station receives the Handoff Direction message at time t_7, removes the pilot from the active set, adds it to the neighbor set, and sends a Handoff Complete message to the base station.

CDMA allows for an idle handoff as well. If a mobile station, while in the idle state, detects a pilot channel from another base station to be significantly stronger than the pilot channel of the current base station, it may decide to initiate a handoff.

cdma2000

System Features

Traffic Types Broadly speaking, cdma2000, like all other 3G technologies, is expected to support the following types of traffic. The data rates may vary from 9.6 kb/s to 2 Mb/s:

- Traditional voice and *voice over IP* (VoIP)
- Data services
  - *Packet data* These services are IP-based with the *Transmission Control Protocol* (TCP) or *User Datagram Protocol* (UDP) at the transport layer. Included in this category are the Internet applications, H.323-type multimedia services, and so on.
  - *Circuit-emulated broadband data* Examples of this kind of traffic include fax, asynchronous dial-up access, H.321-based multimedia services where audio, video, data, and control and indication are transmitted using circuit emulation over *Asynchronous Transfer Mode* (ATM), and so on.
  - *SMS*

In addition, there are, of course, signaling services.

3G systems are intended for indoor and outdoor environments, pedestrian or vehicular applications, and fixed environments such as

wireless local loops. Cells sizes may range from a few tens of meters (say, less than 50 m for picocells) to a few tens of kilometers (in excess of 35 km for large cells).

Bandwidth A cdma2000 system may operate at different bandwidths with one or more carriers. In a multicarrier system, adjacent carriers should be separated by at least 1.25 MHz as shown in Figure 4-8(a). In an actual multicarrier system, each individual carrier usually has a bandwidth of 1.25 MHz and is separated from an IS-95 carrier by means of orthogonal codes. However, when three carriers are being used in a multicarrier system, the bandwidth required is 5 MHz. To provide high-speed data services of the type discussed previously, a single channel may have a nominal bandwidth of 5 MHz as indicated in Figure 4-8(b) with a chip rate of 3.6864 Mc/s (that is, 3 × 1.2288 Mc/s).[5] The bandwidth *BW* in Figure 4-8(b), outside of which the power density is negligible, depends

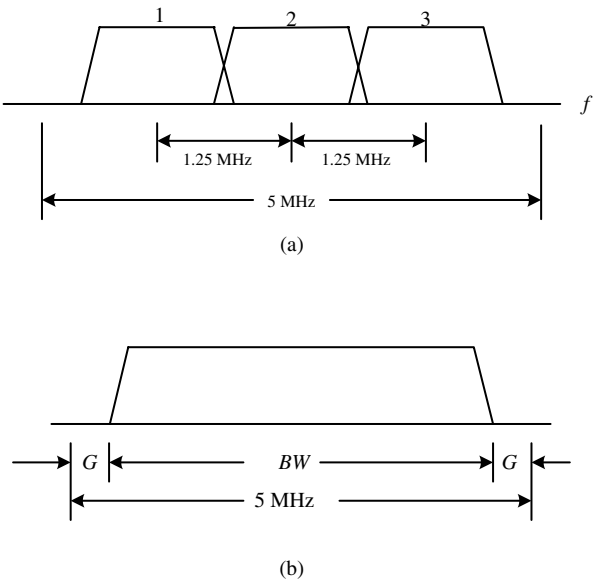


Figure 4-8 Bandwidth requirements in cdma2000

[5]Or, if necessary, the bandwidth of a single channel may be some multiple of 5 MHz.

on the pulse-shaping filter at the baseband.[6] If a raised cosine filter is used, $BW = R_c(1 + \alpha)$, where R_c is the chip rate and α is the roll-off factor. If $\alpha = 0.25$, $BW = 4.6$ MHz, and so the guard band $G =$ 200 kHz. Clearly, an advantage of a wider bandwidth lies in the fact that it provides more resolvable paths that can be used in a multi-path diversity receiver to improve the system performance.

Quality of Service (QoS) At any time, multiple applications may run on a mobile station. A user may request a desired QoS depending on the application, and the network is expected to guarantee the requested quality without any (noticeable) degradation in the QoS contracted by other active users.

Packet Mode Data Services cdma2000 supports packet mode data services [1]. Starting from an initial state, if there is a packet to send, the user attempts to establish the dedicated and common control channels using the multiple-access slotted Aloha scheme.[7] In

---

[6]Recall that the purpose of this filter is to reduce out-of-band energy at the RF stage and minimize the intersymbol interference.

[7]The Aloha system is a wireless computer communication network that was developed in the late 1960s at the University of Hawaii. In this system, multiple user terminals could access a central computer over a radio link using a random access scheme, whereby any terminal could seize the channel at any time and transmit a packet of a fixed length. If there was no contention from other terminals, the central computer would receive the packet error-free, and send an acknowledgment. If a user terminal did not receive the acknowledgment, it would wait for a random period of time, and retransmit the packet. A terminal would repeat this process until it was successful or until it had attempted three times. The radio link operated in the FDD mode, where the two frequencies used were 413.350 MHz and 413.475 MHz. The bandwidth in either direction was 100 kHz. The data rate was 24,000 bauds.

Since this access is purely random, transmissions form two or more terminals may completely or partially overlap, thereby significantly reducing the throughput. In the slotted Aloha scheme, where synchronized time slots are used for transmission purposes, a user can transmit only at the beginning of a slot. Thus, in case of contention, transmissions from multiple users would completely overlap. This approach, therefore, improves the throughput considerably. For a detailed description, see N. Abramson, "The Throughput of Packet Broadcasting Channels," *IEEE Trans. Commun.*, Vol. COM-25, No. 1, pp. 117–128, Jan. 1977.

this scheme, a reference clock is used to create a sequence of time slots of equal duration. When a user has a packet to send, it can begin to transmit, but only at the beginning of a time slot rather than at any arbitrary instant of time. Notice that although users are synchronized via the reference clock, there is some probability that two or more users could begin to transmit at the same time.

When these channels are established, the user may send the packet(s) over the dedicated control channel, and may also request a traffic channel of a desired bandwidth. Once this traffic channel has been assigned, the user transmits the packet(s), maintaining synchronization and power control as necessary, and releasing the traffic channel either immediately following transmission or after a fixed time-out period. If there are no more packets to send, the dedicated control channel is also released after a while, but the network and link layer connections are maintained for a certain length of time so that newly arrived packets, if any, may be sent without any channel setup delays. At the end of that time period, short, infrequent data packets may be sent over a common control channel. The user may either disconnect at this point, continue in this state indefinitely, or reestablish the dedicated control and traffice channels if there are large or frequents packets to send.

Transmit Diversity One of the advantages of W-CDMA is the possibility of transmit diversity. This may be accomplished in two ways. First, with a 5 MHz, direct-spread CDMA system, the user data may be divided into two or more streams, each spread with an orthogonal code, and then transmitted to mobile stations. Because of multipath diversity, the forward channel performance may improve significantly. Second, if it is a multicarrier system, user data streams may be transmitted over different carriers on different antennas (see Figure 3-5).

The Protocol Stack

cdma2000 takes the information—user data and signaling—from the higher layers and adds two lower-layer protocols before trans-

ferring the data over the air interface. This is shown in Figure 4-9. The link layer consists of the *link access control* (LAC) and *media access control* (MAC) layers. The MAC layer is divided into two sublayers: the *physical layer-independent convergence function* (PLICF) and *physical layer-dependent convergence function* (PLDCF) [7], [5].

The various layers and sublayers perform the following functions.

Each traffic type coming from the higher layer has a different QoS requirement in terms of delays, delay variations, and error rates. The function of the LAC is to ensure that various types of traffic are transferred over the air interface according to their QoS requirements. The link layer protocols used for this purpose include an *automatic repeat request* (ARQ) as well as an acknowledged data transfer procedure using *acknowledgment / negative acknowledgment* (ACK/NACK) and sequence numbering for retransmission. The MAC layer also provides a certain degree of transmission reliability. However, when it does not meet the requirements of an application, the LAC may call for an appropriate link layer procedure. Notice that for some traffic, such as circuit-switched voice, the LAC layer function may be null. In other words, associated packets from the higher layers are passed directly to the MAC layer.

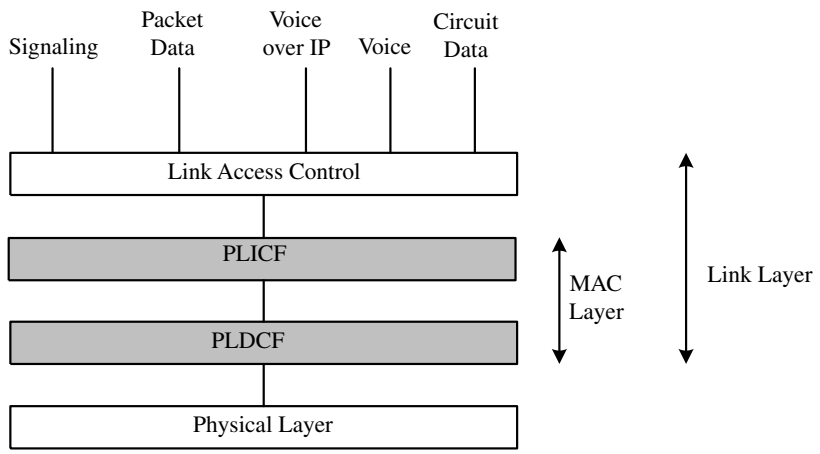


Figure 4-9 The lower layer protocols for cdma2000

A MAC sublayer performs the following functions:

- It controls user access to the physical layer (that is, the medium) by resolving, if necessary, contention among multiple applications from the same user or among multiple users, and scheduling its resources so as to ensure efficient utilization of bandwidth. Resources include buffers, spreading codes, convolutional encoders, and so on.
- User data and signaling information from the upper layers (that is, the LAC layer and the higher layers) are multiplexed, mapped into different physical channels, and delivered to the physical layer on a best-effort basis, providing a basic level of transmission reliability.[8]

The MAC layer is divided into two sublayers:

- Functions that are independent of the physical layer, such as controlling access to the medium so as transmit packets, are performed by the sublayer called PLICF. The user data and control information are passed to the lower sublayer over a set of logical channels, such as a dedicated traffic channel, common traffic channel, dedicated signaling channel, common signaling channel, dedicated MAC channel carrying MAC messages, forward common MAC channel, and reverse common MAC channel.
- The second sublayer is the PLDCF. Functions performed at this sublayer when transmitting over the air interface include multiplexing logical channels coming from PLICF, mapping them into physical channels, assigning proper priorities to each according to its QoS requirement, and delivering them to the physical layer. The best-effort delivery of data services is performed at this layer using a *radio link protocol* (RLP) for streaming-mode user data, and a *radio burst protocol* (RBP) for

[8]In the best-effort service, the user specifies the maximum and minimum data rates. The amount of bandwidth allocated to a user may vary during the life of a call depending on the congestion experienced by the network.

short bursts of user data over a common traffic channel. The RLP uses an ARQ-based retransmission scheme. The corresponding protocols for handling signaling information are the *signaling radio link protocol* (SRLP) and *signaling radio burst protocol* (SRBP).

Physical Channels

Forward Physical Channels As in IS-95, the pilot channel continuously transmits a carrier modulated with an all-zero patttern so that mobile stations can achieve initial cell synchronization. A mobile station may use the received signal as a reference carrier for coherent demodulation, or measure the received signal strength and report the measurement to a base station for handoff purposes.

A common auxiliary pilot channel has been added to cdma2000 so that adaptive antennas can be used for beamforming to extend coverage, increase capacity, and provide higher data rates, among other things. Because beamforming is accomplished by combining signals from different locations in the antenna's aperture in an optimal manner using an adaptation algorithm that requires as accurate a channel estimate as possible, it is necessary that the pilot and data signals travel along the same path to the receiver [3], [4].

A dedicated auxiliary pilot channel is dedicated to a given mobile station (or a group of mobile stations) for the purpose of beam steering using an adaptive antenna array.

A sync channel operates at 1200 b/s, transmitting synchronization messages so that mobile stations in the coverage area of a base station can acquire frame synchronization after cell acquisition. For a single carrier system with a channel bandwidth of 1.25 MHz, the channel encoder used is of rate $1/2$. If the system consists of multiple carriers or a single carrier with a bandwidth of 5 MHz or more, the convolution code used is of rate $1/3$.

The paging channel is used to transmit paging and overhead messages directed to mobile stations in the coverage area of a base station. There are two data rates: 9.6 and 4.8 kb/s. For a single carrier system with a channel bandwidth of 1.25 MHz, the convolutional encoder used is of rate $1/2$. If the system consists of multiple

carriers, or a single carrier with a bandwidth of 5 MHz or more, the encoder used is of rate 1/3.

The quick paging channel has been added so that a base station can send a quick paging message to a mobile station operating in the slotted mode. This message actually consists of a single bit, which is followed by a regular paging message in the slot that has been allocated to the particular mobile.

Next is the broadcast common channel. Instead of combining overhead and paging messages on a paging channel, the system performance can be improved to some extent by separating overhead messages and sending them over this channel.

The common control channel is used to send layer 3 and MAC layer messages to mobile stations at 9.6 kb/s using frame sizes of 5, 10 or 20 ms.

The dedicated control channel is similar to the common control channel, but uses frames that are 5 or 20 ms long.

The fundamental channel is used for lower data rates: 9.6 kb/s and its subrates, grouped as rate set 1, and 14.4 kb/s and its subrates, grouped as rate set 2.[9] This channel is supported in both single-carrier and multicarrier cdma2000 systems. Both 20 ms and 5 ms frames are permissible.

Supplementary channel 1 and 2 are designed for higher data rates. Rates supported are shown in Table 4-1. Frames are usually 20 ms long.

Reverse Physical Channels The reverse pilot channel is similar in concept to the forward pilot channel. Used in conjunction with reverse dedicated channels, it enables a base station to acquire initial time synchronization and recover a phase-coherent carrier for coherent demodulation in a rake receiver. It also includes a power control subchannel, which sends one bit in each 1.25 ms power control group or 16 bits in each 20 ms frame. The base station can use this bit to adjust its power level when necessary.

[9]This is after adding the frame quality indicator bits to incoming frames.

Table 4-1 *Data rates supported on a supplementary channel in cdma2000*

	Rate Set 1	Rate Set 2
Single-carrier cdma2000 with a bandwidth of 1.25 MHz	M × 9.6 kb/s, M = 1, 2, 4, 8, 16, and 32. Uses channel encoder of rate 1/2.	M × 14.4 kb/s, M = 1, 2, 4, 8, and 16. Uses channel encoder of rate 1/2.
Multicarrier cdma2000 where each channel has a bandwidth of 1.25 MHz, or a single-carrier system with a bandwidth of 5 MHz or multiples thereof	M × 9.6 kb/s, M = 1, 2, 4, 8, 16, 32, and 64. Uses channel encoder of rate 1/3.	M × 14.4 kb/s, M = 1, 2, 4, 8, 16, 32, and 64. Uses channel encoder of rate 1/4.

The access channel transmits layer 3 and MAC layer messages from different mobile stations to a base station. Multiple users access this channel using a mechanism that is very similar to the slotted Aloha scheme. The data rate supported is 9.6 kb/s. There may be more than one access channel, each identified by a unique orthogonal code.

The common control channel, like the reverse access channel just described, also carries layer 3 and MAC messages, and is accessed by mobile stations using the same multiple access scheme. Data rates supported include 9.6, 19.2, and 38.4 kb/s.

The dedicated control channel, like the reverse fundamental or supplementary channels, carries user data packets at 9.6 kb/s or 14.4 kb/s in 5 ms or 20 ms frames.

The fundamental channel is similar to the forward fundamental channel. It supports a data rate of 9.6 kb/s and its subrates (4.8, 2.7, and 1.5 kb/s), or 14.4 kb/s and its subrates (7.2, 3.6, and 1.8 kb/s). For these rates, convolutional codes are used. A frame is usually 20 ms long. However, in some cases, a 5 ms frame may also be used. Note that only a fundamental channel supports a 5 ms frame.

Supplementary channel 1 and 2, which are similar to the forward supplementary channels, provide higher data rates: (1) 9.6, 19.2,

38.4, 76.8, and 153.6 kb/s, and (2) 14.4, 28.8, 57.6, 115.2, and 230.4 kb/s. Only 20 ms frames are supported. For these data rates, turbo coding may be used.

Forward Channel Transmit Functions

As an aid to understand the technology used in the implementation of physical layer functions of a typical W-CDMA system, a simplified block diagram of the transmit functions of a multicarrier cdma2000 base transceiver station was presented in Chapter 3, "Principles of Wideband CDMA," (see Figure 3-5 of that chapter). Figure 4-10 shows a similar diagram of the transmit functions of the forward channels of a direct-spread, single-carrier cdma2000 system. For simplicity, only a subset of the forward physical channels is included in this figure. Notice the similarity between cdma2000 and IS-95 (refer to Figure 4-4) forward channel transmit functions. Some of the differences are as follows.

cdma2000 has two traffic channel types—the fundamental and secondary. A number of data rates are supported. Depending upon the data rate, convolutional codes of rate $1/2$, $3/8$, $1/3$, or $1/4$ may be used. Both 10 ms and 5 ms frames are supported.

I- and Q-channel symbols are multiplied by gain factors to provide some additional power control. As in IS-95, cells are separated by different pilot PN sequence offsets.[10] However, now, complex spreading is used by, first, adding the real-valued I and Q sequences in quadrature (so that the result is a complex number) and then multiplying it with a another complex number $S_I + jS_Q$, where S_I and S_Q are, respectively, the I-channel and Q-channel pilot PN sequences. The output of this multiplication is a complex quantity whose in-phase and quadrature components are as shown in the lower part of the figure. With complex spreading, the output of the wave-shaping filter goes through zero only with low probability, thus leading to improved power efficiency.

[10]The period of these sequences is $2^{15} - 1$ chips.

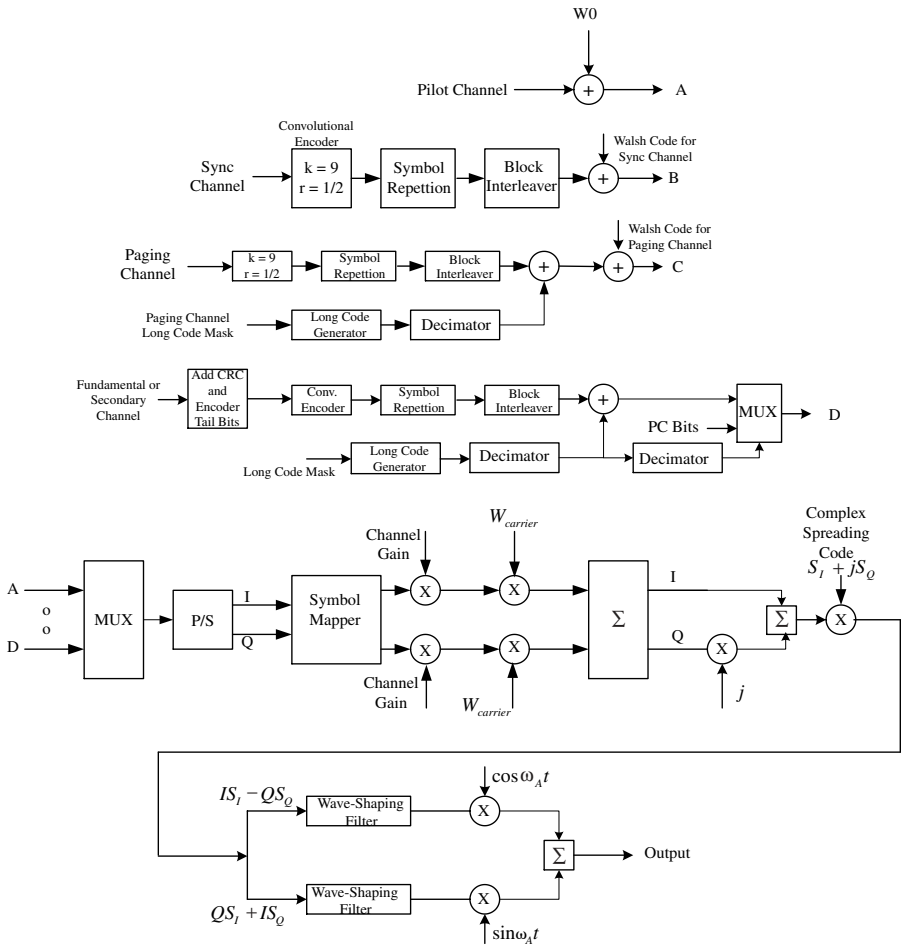


Figure 4-10
The functional block diagram of direct-spread (single-carrier) forward channel transmit functions in cdma2000

Reverse Channel Transmit Functions

The functional block diagram of direct-spread, reverse-channel transmit functions in cdma2000 is shown in Figure 4-11. Consider, first, the fundamental channel. The incoming data on this channel is processed in the usual way. Depending on the user data rate, a variable number of frame-quality indicator bits in the form of CRC are added to a frame. A few tail bits are appended to ensure proper operation of the channel encoder, which may be either a convolutional

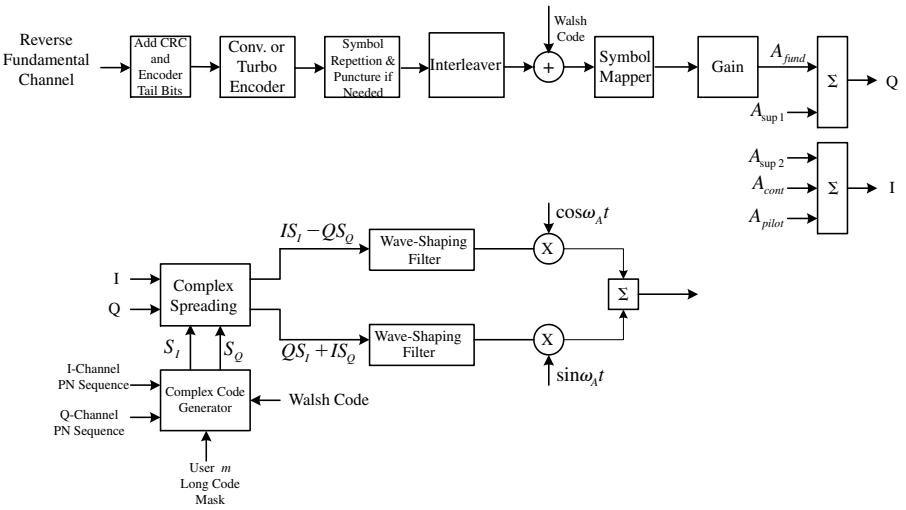


Figure 4-11 The functional block diagram of direct-spread reverse channel transmit functions in cdma2000

coder or a turbo coder. Code symbols are repeated, but depending upon the rates, some of them are also deleted. The output of the interleaver is spread with a Walsh code, mapped into modulation symbols, and multiplied by gain factors, resulting in a signal labeled A_{fund}.

The supplementary channels 1 and 2 and control channels are processed in the same way, although details might vary in some cases. For example, symbol puncturing is not done on a reverse dedicated control channel. Similarly, the reverse pilot channel, which consists of a string of zeros (that is, real values of +1), is treated differently because it is not encoded into a channel code, interleaved in a block interleaver, or multiplied by a Walsh code. However, a power control bit is inserted into the pilot channel for each power control group or 16 power control bits per frame. For simplicity, we have omitted these repetitions and merely indicated the processed outputs of these channels as A_{sub1}, A_{sub2}, A_{cont}, and A_{pilot}.

The fundamental channel and supplementary channel 1 are summed together giving an output Q. Similarly, the remaining channels are summed separately, giving I as the output. Notice that in this case, the I- and Q-channel sequences formed for QPSK modulation are independent of each other because they are derived from different channels and not by splitting the data stream of a given

channel into two sub-streams. The I and Q sequences are spread by a complex code of the type $S_I + jS_Q$, where S_I and S_Q are user-specific because they are obtained from a 42-bit long code mask for the given user, I- and Q-channel pilot PN sequences, and a Walsh code.

Summary

In this chapter, we have described the fundamental aspects of cdma2000, which is one of the systems specified by IMT-2000. Because cdma2000 is an evolved version of the current CDMA system known as cdmaOne, a brief description of this system is also included. The basic features and service capabilities of cdma2000 are discussed. To provide services in cdma2000, a new link layer protocol has been defined that consists of a LAC layer and a MAC layer. The functions performed by the different sublayers are briefly described. This is followed by a description of the physical layer in terms of the physical channels and the forward and reverse channel transmit functions.

The distinctive features of a cdma2000 system may be summarized as follows:

- *Wider bandwidth and higher chip rate* For a direct-spread CDMA system, the nominal bandwidth is 5 MHz. While IS-95B supports data rates in the range of 64 to 115 kb/s, much higher data rates—from 144 kb/s to 2.0 Mb/s—are possible in cdma2000. CDMA in general is inherently resistant to fades. However, the improvement in the bit error rate performance is significantly greater for a 5 MHz system than for 1.25 MHz. Because the chip rate is three times as high as in IS-95, for a given power delay profile, there are many more resolvable paths in direct-spread cdma2000 that can be utilized in a rake receiver. Furthermore, as we discussed before, transmit diversity is a distinct possibility here that will significantly improve the downlink performance.
- *Multicarrier system* cdma2000 may consist of a single, direct-spread, 5 MHz carrier, or multiple carriers, each with a

bandwidth of 1.25 MHz. In a multicarrier system, because each carrier is orthogonally spread, W-CDMA can be overlaid on an existing IS-95 system. Also, a multicarrier system is inherently capable of providing transmit diversity because high-speed user data may be divided into two or more streams and transmitted on multiple carriers over different antennas.

- *Spreading codes* In both IS-95 and cdma2000, the spreading of downlink channels is similar. For example, different cells are separated by means of different offsets of the I- and Q-channel pilot PN sequences. Similarly, traffic channels directed to a given user are spread by user-specific long codes.

  On uplinks, however, there are some differences. In cdma2000, physical channels are separated by Walsh codes, and mobile stations by long codes, whereas in IS-95, long codes are used to separate the access and traffic channels.

- *Variable length Walsh codes* Because a traffic channel of a cdma2000 system is required to support many data rates, it is necessary to use variable-length Walsh codes. This length varies from 4 to 128 chips. On fundamental channels, Walsh codes have a fixed length. But on the secondary channels, as the data rates increase, the code length decreases (which, in essence, reduces the process gain and thus the number of simultaneous users on a CDMA channel).

- *Complex spreading* In cdma2000, complex spreading is used that reduces the amplitude variations of the baseband filter output, thus making the signal more suitable for nonlinear power amplifiers.

- *Additional pilot channels* Many new physical channels have been defined in cdma2000 that have the potential for improving the system performance. For example, in the downlink, there is an auxiliary pilot that may be code-multiplexed to provide beamforming and beam steering with adaptive antenna arrays.

  Similarly, there is a pilot channel in the uplink, which again is code-multiplexed, enabling a base station to recover the carrier for coherent demodulation in a rake receiver.

- *New traffic channels* There are two types of traffic channels: fundamental and supplementary, both of which are code-multiplexed. A fundamental channel is used for lower data rates such as 9.6 and 14.4 kb/s and their subrates. The supplementary channels provide higher data rates. Also, two channel codes are used—convolutional codes on fundamental channels or supplementary channels with a data rate of 14.4 kb/s. At higher data rates on a supplementary channel, turbo codes of constraint length 4 and rate 1/4 are recommended. Fundamental channels support both 20 ms and 5 ms frames, while secondary channels use only 20 ms frames.
- *Packet mode data services* cdma2000 supports a highly flexible packet mode data service. The multiple-access procedure is based upon the slotted Aloha scheme. The physical channels that may be used for this purpose include dedicated traffic channels, dedicated control channels, and common control channels.
- *Quality of service* The support of multimedia services at variable data rates with user-specified QoS is unique to wideband systems.

References

[1] T. Ojanpera and R. Prasad, "An Overview of Air Interface Multiple Access for IMT-2000/UMTS," *IEEE Commun. Mag.*, Vol. 36, No. 9, September 1998, pp. 82–95.

[2] E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, "UMTS/IMT-2000 Based on Wideband CDMA," *IEEE Commun. Mag.*, Vol. 36, No. 9, September 1998, pp. 70–80.

[3] F. Adachi, M. Sawahashi, and H. Suda, "Wideband DS-CDMA for Next Generation Mobile Communications System," *IEEE Commun. Mag.*, Vol. 36, No. 9, September 1998, pp. 56–69.

[4] G. Tsoulos, M. Beach, and J. McGeehan, "Wireless Personal Communications for the 21st Century: European Technological Advances in Adaptive Antennas," *IEEE Commun. Mag.*, Vol. 35, No. 9, September 1998, pp. 102–109.

[5] TIA TR 45.5, "The cdma2000 ITU-RTT Candidate Submission," TR 45-ISD/98.06.02.03, May 15, 1998.

[6] TIA/EIA/IS-95-A: Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System, May 1995.

[7] V.K. Garg, *IS-95 CDMA and cdma2000.* New Jersey: Prentice Hall, 1999.

CHAPTER 5

The GSM System and General Packet Radio Service (GPRS)

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

We mentioned in Chapter 1 that core networks of UMTS are harmonized with GSM. The UMTS core network is also compliant with the *Mobile Application Part* (MAP) protocol of *Signaling System 7* (SS7) that provides signaling between a *Mobile Switching Center* (MSC), the *Visitor Location Registers* (VLR), the *Home Location Register* (HLR), and the *Authentication Center* (AC) in GSM. Similarly, the packet mode data services in UMTS and the associated network entities and protocols have been harmonized with those of GPRS, which is now being offered as an upgrade of GSM. The reader may recall from Chapter 1 that ETSI has also defined another standard called *Enhanced Data Rates for GSM Evolution* (EDGE) to support data rates up to 384 kb/s in GSM networks. The wideband TDMA system IS-136 HS for outdoor/vehicular applications is designed to use this protocol in the access network. Thus, even though there are significant differences in the air interface standards of UTRAN and GSM, a description of GSM and GPRS is appropriate in this context.

GSM was first deployed in a few countries of Europe in 1991. Subsequently, it was adopted in most of Europe, Australia, much of Asia, South America, and the United States. Today, it is the fastest growing technology in many parts of the world and is being continually evolved to provide advanced features, particularly in areas of data communications.

GSM supports voice, circuit-switched data, and short messaging services. The standards work on a packet mode data service in GSM started in 1994, and was completed in 1997. The new system specified by these standards was called GPRS. A number of references are available in the literature that describe the GSM system in great detail. See, for example, [23], [1], and [2]. Reference [9] gives a detailed description of GPRS and discusses its performance based on simulation. GPRS services are described in [11]. An overall description of the GPRS radio interfaces appears in Reference [12]. Details of the radio link control and medium access control protocols are provided in Reference [13]. Our goal in this chapter is to present an overview of GSM and GPRS systems.

GSM System Features

GSM operates in the frequency division duplex mode, using one band for uplinks and a separate one for downlinks. Initially, a 50 MHz bandwidth around 900 MHz was allocated to GSM. The spectrum allocation is shown in Figure 5-1. The 25 MHz spectrum in either direction is divided into 125 physical channels, each with a bandwidth of 200 kHz. To avoid interference with other systems operating at the edges of the GSM spectrum, one of these channels is not used. Later, a second allocation of 150 MHz bandwidth centered around 1800 MHz was set aside for use in systems called *Digital Cellular System* at 1800 MHz (DCS1800).[1]

In GSM, speech is digitally encoded at 13 kb/s using linear predictive coding (also known as *vocoding*). Information is transmitted in frames, each 4.615 ms long and divided into eight equal time slots. Normally, each slot is assigned to a user. Data (such as voice samples) from multiple users is *time-division multiplexed* on a frame and sent out over a physical channel at 270.8333 kb/s. Because each channel operates at a different frequency, the system combines TDMA with *frequency division multiple access* (FDMA). The GSM characteristics are summarized in Table 5-1.

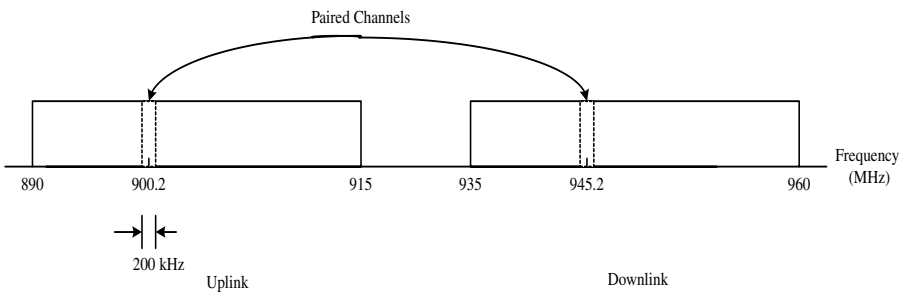


Figure 5-1 Spectrum allocation for GSM

[1]The allocation is 1710 to 1785 MHz for uplinks and 1805 to 1880 MHz for downlinks.

Table 5-1

Summary of GSM system characteristics

Multiple Access Scheme	TDMA/FDMA with FDD
Spectrum allocation	890–915 MHz (uplink), 935–960 MHz (downlink)
Bandwidth of each physical channel	200 kHz
Total number of channels available in either direction	124
Number of users per channel	8
Data rate	270.83333 kb/s, bit period = 3.692 μs
TDMA frame size	4.615 ms
Number of slots per frame	8
Slot duration	0.576923 ms
Modulation	0.3 GMSK
Speech coding	13 kb/s *Regular Pulse Excitation with Long Term Predictor* (RPE-LPT)
Interleaver period	40 ms maximum, using two consecutive 20 ms blocks of data
User data transfer capability	Short messaging service, circuit-switched data, high-speed circuit-switched data, and GPRS for packet data

Among the features and capabilities of GSM are the following:

- *Teleservices* This includes regular telephony via *public switched telephone networks* (PSTN), emergency calling such as police or fire brigade, and voice messaging where a calling party can leave a voice message that can be retrieved later by the called party.
- *Bearer services* These include data services and short messaging services. For data services, the MSC may be connected to a circuit-switched PSTN via a modem or to a public data network via *a packet assembler and disassembler* (PAD). A mobile station may subscribe to circuit-switched data services at all standard rates up to 9.6 kb/s without having to use a modem

and should also be able to access public data networks at 9.6 kb/s, 4.8 kb/s, and 2.4 kb/s.

In GSM, a mobile station is permitted to transmit or receive short messages during both idle and active call states. A message may contain up to 160 alphanumeric characters. Shorter messages with only 93 characters may be broadcast by a base station to all mobiles in a serving area. A subscriber may dictate a message at a service center, which may later be sent to the intended party. If the intended party is not available, the message is saved and later forwarded when the party is available.

- *Supplementary services* These include *private branch exchange* (PBX) features such as call forwarding, call hold, call waiting, call transfer, calling number identification, detailed billing records, three-party conference calls, interoperability with ISDN, and so on.

Other important features, not necessarily in the order of their importance, include international roaming in European countries, secure communication and privacy, use of a *subscriber identity module* (SIM) to distinguish between the identity of a subscriber and that of the mobile station, and *discontinuous transmission* (DTx), that is, turning off the transmitter during silence periods, thus leading to increased battery life.

A new feature that is available with the later version of GSM, known as Phase 2+, is the GPRS, whereby both high-speed and low-speed user data and signaling information may be transferred using packet-switching techniques. In this service, a single time slot may be shared by multiple users for transmitting data in the packet mode, resulting in a throughput of about 12 to 20 kb/s. Thus, if all 8 slots of a frame are used, it is possible to achieve a throughput of about 124 to 171 kb/s.

System Architecture

The GSM system architecture is shown in Figure 5-2. The air interface corresponds to the reference point Um between a *mobile station*

(MS) and a *base transceiver station* (BTS). Each cell is served by a BTS that consists of multiple radio receivers and transmitters. *Base station controllers* (BSC) perform radio control functions such as power control and handoff. Each BSC may connect to one or more BTS over the A-bis interface. There may be one or more base station controllers in a serving area. BTSs and the associated BSCs for a given area are together known as a *base station subsystem* (BSS).

The MSC is responsible for call controls, call routing to and from PSTNs, and switching and controls during a handover process. It connects to a BSS over the A interface. It interfaces with a number of other entities: VLR, HLR, *Equipment Identity Register* (EIR), and *Operations and Maintenance Center* (OMC). It also requires the services of the AC, which is connected to the HLR. A brief description of each of these entities is given in the following paragraphs.

The HLR is a database system of all mobile subscribers who are registered in a *public land mobile network* (PLMN). There may be just one HLR at a central location in a network or many of them distributed throughout the network, but only one of them contains the information about a given subscriber. The VLR contains the database of all mobile subscribers who are visiting this particular serving area. There is a VLR for each serving area controlled by an MSC. Whenever a mobile travels into a foreign serving area, the VLR of the visited system requests the database of that mobile from its HLR and saves it in its memory so that it can continue to serve that mobile as long as it is in this area. At the same time, the MSC of this

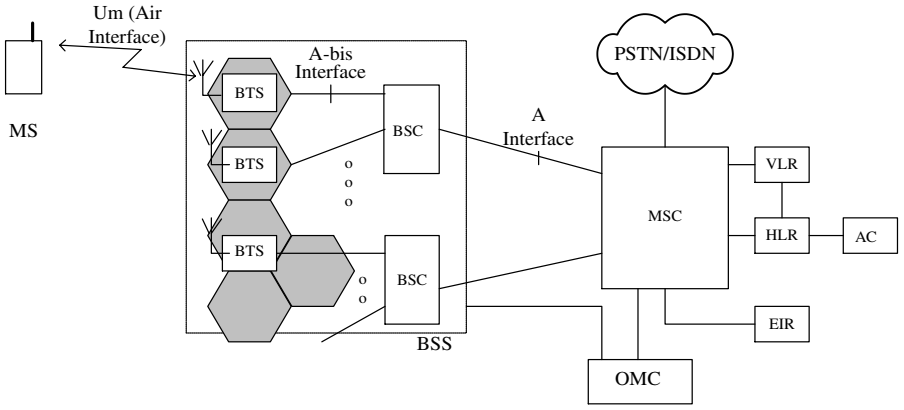


Figure 5-2 GSM system architecture

foreign serving area informs the HLR of the home system about the location of this mobile so that the home area of this subscriber can route the call correctly when necessary.

The authentication center verifies the identity of the user at call inception times for security purposes, and contains such parameters as authentication keys and other required parameters. The EIR contains the *International Mobile station Equipment Identity* (IMEI) numbers of all mobile stations that are registered. Each mobile is assigned a unique IMEI number that can be used to determine whether the equipment is genuine or not. The OMC is a centralized network management system that provides the capability of remote system administration and maintenance of all equipment and databases, and may also perform such functions as billing and so on.

Figure 5-3 shows the functional block diagram of a GSM system. A brief description of the system is presented here. Each functional block will be further explained in the next section.

Figure 5-3(a) represents the transmitter of a mobile station. The speech signal from a mobile terminal is encoded into 13-bit uniform *pulse code modulation* (PCM) samples and applied to the input of a RPE-LTP coder [1]–[5]. It is in fact a linear predictive coder, or a vocoder as it is called, that attempts to predict the incoming speech by modeling the speech-generating system by a finite-order, time-varying digital filter.[2] The RPE-LTP encoder actually consists of two

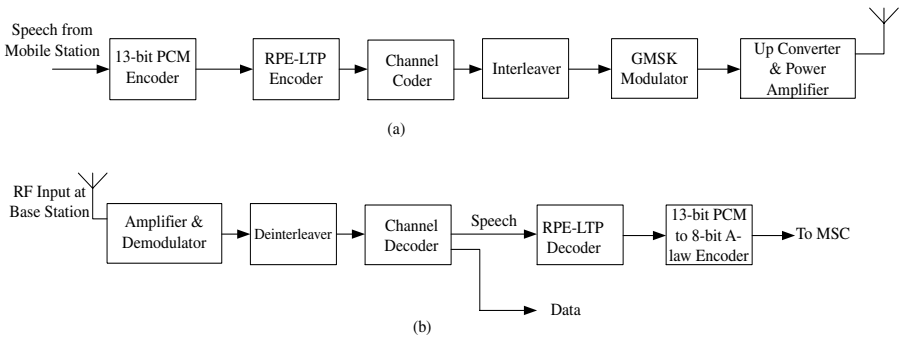


Figure 5-3 The functional block diagram of a GSM system

[2]In other words, the vocoder tries to mimic the vocal system of a human being. See References [1] to [5].

linear predictive filters, and generates a residual excitation (that is, a difference signal between the incoming speech and predicted speech), which is multiplexed with the coefficients of the two filters and passed to the channel encoder. Because there are 260 bits at the output of the encoder for every 20 ms of the speech input, the bit rate of a full-rate coder is 13 kb/s.

The purpose of the channel encoder is to provide some protection against impairments that are likely to be introduced by the channel. This is done by encoding the incoming frames into error-correcting codes. GSM uses a combination of block codes and convolutions codes. An (n, k) block code encodes a k-bit message block into an n-bit code by adding $n - k$ parity bits. A convolutional code of rate $1/2$ with a constraint length of 5 is used for speech. For 9.6 kb/s data, convolutional codes are the same, but a few bits are deleted from each frame (resulting in punctured codes) so as not to exceed the data rate of the physical channel. If the data rate is 2.4 kb/s, the convolutional code is rate $1/6$, but has the same constraint length. Not all bits of the speech encoder output are equally critical to the subjective quality of speech at the receiver. Bits that are most essential to intelligibility of speech are protected against channel errors by encoding them into both block codes and convolutional codes. If there are any errors in these bits that the receiver cannot correct, they are discarded. Bits that are less critical are encoded into convolutional codes only. The remaining bits are transmitted without any channel coding. A 20 ms frame at the output of the channel encoder contains 456 bits.

The output of the channel encoder is applied to an interleaver, which simply rearranges the order of the incoming bits.[3] As mentioned previously, speech or data is transmitted in frames, as shown in Figure 5-4. Each frame consists of eight slots, each of which is assigned to an individual user. Later on, we will see how exactly the data structure of a slot is constructed. For the time being, though, it is sufficient to say that each slot contains 114 bits of the interleaver output and 42.5 bits of overhead.

[3]Correlated signal fading, which is a characteristic of a mobile radio channel, results in burst errors. The purpose of the interleaver is to spread out in time these burst errors so that the receiver can detect and correct them with a higher probability. Notice that the data rate at the output of the interleaver is the same as the input.

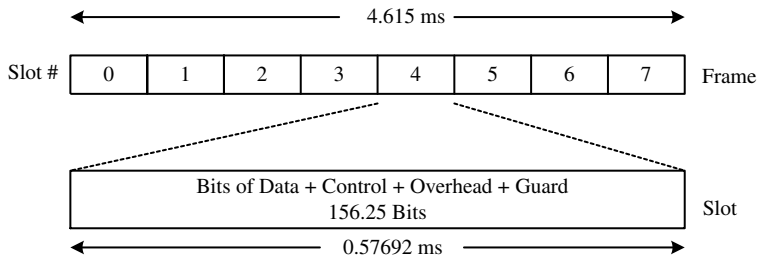


Figure 5-4
Frames and slots in GSM

The output of the interleaver feeds into the modulator. Prior to modulation, however, the rectangular pulses of the encoder output are filtered with a Gaussian pulse-shaping filter. These filters are usually specified in terms of their 3 dB bandwidth B and the bit period T of the input data. In GSM, this parameter $BT = 0.3$.

Because the data rate is 270.83333 kb/s, $T = 3.692$ μs. Thus, the 3 dB bandwidth of this filter $B = 81.26$ kHz. The filter is designed with this bandwidth because when its output modulates an RF carrier, most of the spectral energy (say, about 95 percent) at the output of the modulator is contained in an RF bandwidth of 200 kHz (which is the channel spacing in GSM).

The baseband modulation scheme used in GSM is *minimum shift keying* (MSK), which is a special case of the continuous phase *frequency shift keying* (FSK) or the more general type called the *continuous phase modulation*. A property of this modulation technique is that the phase trajectory of the modulator output is always continuous and that the maximum amount of the phase change during any bit period is ± 90 degrees. Because the modulator output has a constant amplitude and does not have any abrupt phase changes, the signal can be efficiently amplified using a Class C power amplifier.

Transmitter functions of the base station are similar to those of a mobile station except for the way in which the speech signal received from the MSC is treated. Since speech is encoded in the PSTN using the 8-bit A-law companding scheme, the speech signal from the MSC is first converted into 13-bit uniform PCM and then encoded with the RPE-LTP encoder in exactly the same way as described in the previous paragraph.

Figure 5-3(b) shows the receiver function of a base station. The received signal is detected using phase-coherent demodulation techniques. The demodulated signal at the baseband is deinterleaved and passed through a channel decoder. Convolutional codes are decoded using soft decision decoding. The output of the channel decoder is passed through an RPE-LTP decoder to obtain 13-bit uniform PCM samples. These PCM samples are transcoded into 8-bit A-law PCM and sent out to the MSC.

Speech Encoder

The speech encoder used in GSM belongs to the vocoder class of encoders, which, unlike a waveform quantizer, model the vocal tract as a time-varying digital filter such that when it is excited with an appropriate input, the output is a speech signal [3], [4]. In essence, it is a linear predictive coder [5]. Its coefficients representing the pitch period and the gain factor are determined by performing a time-domain analysis of an incoming speech frame at the transmitter end. These coefficients together with a residual excitation, which is a difference signal between the actual speech input and the output of the predictor, are transmitted to the remote end. At the receiver, the coefficients are used to construct linear predictors, which are then excited with the received residual signal to synthesize the desired speech.

Figure 5-5 shows an overly simplified functional block diagram of an RPE-LTP speech encoder [1]. It consists of two linear predictive filters, one a *short-term predictor* (STP) and the other a *long-term predictor* (LTP). Once every 20 ms, the encoder reads 160 13-bit, uniform PCM samples, preprocesses them to remove any DC offset that might be present, analyzes them, and derives the coefficients of an STP. The 160 input samples are passed through the filter thus constructed, and its output applied to the input of an LTP filter. The predicted output of this filter is compared to the 160 samples of the short-term residual signal. The difference signal, which is marked *long-term residual excitation* in the figure, is low-pass filtered and applied to a decimator. Here the signal is down-sampled, and an optimum sequence is selected. The resulting output is sent to the channel encoder and also used to update the LTP coefficients.

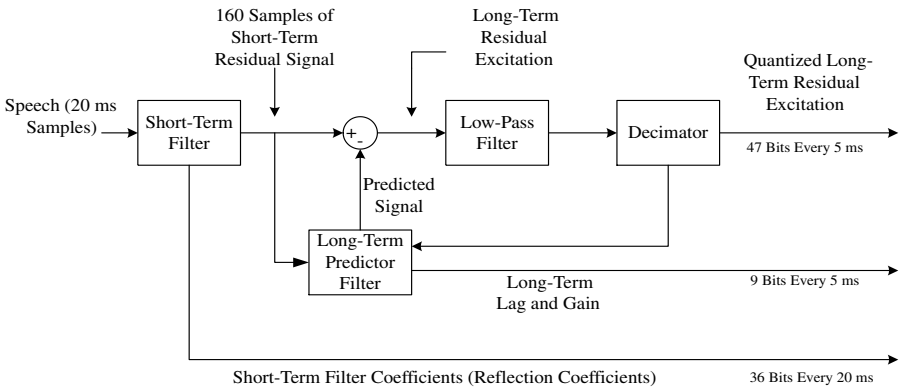


Figure 5-5
A simplified block diagram of the RPE-LTP speech encoder

The output of the encoder consists of three components: (i) the STP coefficients, also known as the reflection coefficients, (ii) the quantized long-term residual excitation, and (iii) the LTP coefficients. Because it is these parameters and not the actual speech samples that are eventually transmitted to the remote end, the bit rate of these kinds of coders is generally much less than those of any waveform coders, such as PCM, *Adaptive Differential PCM* (ADPCM), and so on, which try to replicate the speech signal directly at the remote end.

Actually, the 20 ms short-term residual signal is divided into four 5 ms blocks, which are then analyzed one at a time to compute the LTP coefficients and generate the long-term residual excitation. Similarly, the LTP coefficients are also updated once every 5 ms. Thus, once every 5 ms, there are 47 bits of the long-term residual output and 9 bits of LTP coefficients. To reduce quantization errors, each reflection coefficient x is mapped to a value $\log\left(\frac{1+x}{1-x}\right)$ before transmission. A 20 ms speech signal results in 36 bits of STP coefficients. Thus altogether, there are $47\times4 + 9\times4 + 36 = 260$ bits in a 20 ms frame, resulting in a bit rate of 13 kb/s.

Channel Encoder

Besides speech, various types of signaling and control information as well as circuit-switched data at different rates (such as 9.6, 2.4 kb/s,

and so on) are transmitted over the air interface. Examples are paging messages, synchronization information, information that a mobile station can use to correct its frequency, and so on. In each case, it may be necessary to employ a different channel coding scheme. By way of examples, we will describe the channel-encoding procedures only for full-rate speech and 9.6 kb/s data.

Channel Encoding for Full-Rate Speech Every 20 ms, the speech encoder generates 260 bits. These bits are classified in order of their importance as Class 1a, Class 1b, and Class 2 types. Because Class 1a bits are the most important, they are first encoded in a block code to add 3 bits of parity so that errors not corrected by the convolutional decoder can be detected. Any block that fails the parity check is discarded by the receiver. Class 1b bits are also channel-encoded, but do not have any parity bits. However, four tail bits are added so that the last bit of the input data is correctly encoded by the convolutional coder. Class 2 type bits are transmitted over the channel without any encoding. This is shown in Figure 5-6.

Channel Encoding for 9.6 kb/s Data The encoding procedure for circuit-switched data on a traffic channel is shown in Figure 5-7. Here, every 20 ms, 240 bits of user data are multiplexed with 4 tail bits and applied to the input of a convolutional encoder of rate $1/2$ and constraint length 5. Thirty-two bits are then removed from the output of the encoder. This is called the *punctured coding*.

In either case, the output of the channel encoder is 456 bits, which are interleaved and transmitted over 8 TDMA frames using a fixed slot in all frames.

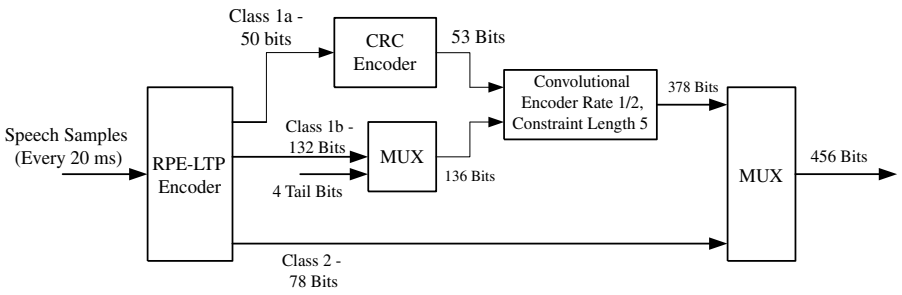


Figure 5-6
Channel coding of full-rate speech

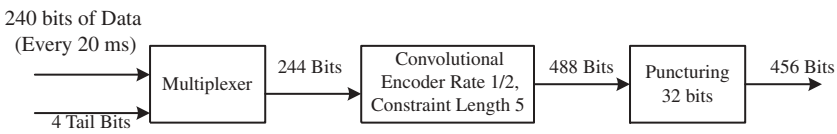


Figure 5-7
Channel coding of 9.6 kb/s data

Interleaving

The purpose of the interleaver is to scramble, or spread out in time, clustered errors resulting from correlated signal fading. We shall describe the interleaving procedure by way of an example. Consider first the speech traffic. The output of the speech coder appears in 20 ms blocks, each consisting of 260 bits. Each block, when encoded by the channel encoder, results in 456 bits at its output.

Figure 5-8 depicts how contiguous blocks can be interleaved. The 456-bit output from the encoder corresponding to a 20 ms interval is written into a 57×8-bit memory by row from left to right and from top to bottom, and read out column by column, as shown in Figure 5-8(b). The result is 8 subblocks, which are numbered 1 through 8 in Figure 5-8(c), each containing 57 bits. Thus, any two adjacent bits of the interleaver input are now spaced apart by 57 bit periods.

To achieve further interleaving, the eight subblocks of any block, say, m, are transmitted over eight frames, the first four frames being shared between blocks, say, m-1 and m, and the last four frames being shared by blocks m and m+1 as depicted in Figure 5-8(c). Assuming that slot k is assigned to a given user, subblock 5 of block m and subblock 1 of block m+1 are inserted into slot k of frame n. Similarly, slot k of frame n+1 contains subblock 6 of block m and subblock 2 of block m+1, and so on for frames n+2 and n+3. In the same way, although it is not shown in the figure, subblocks 1 through 4 of block m and subblocks 5 through 8 of block m-1 of the interleaver output are transmitted over slot k of frames n-4, n-3, n-2, and n-1. Because each block of the channel encoder output is transmitted in eight bursts using eight consecutive TDMA frames, the interleaver is said to have depth eight. A similar interleaving process is used for user data or signaling and control information.

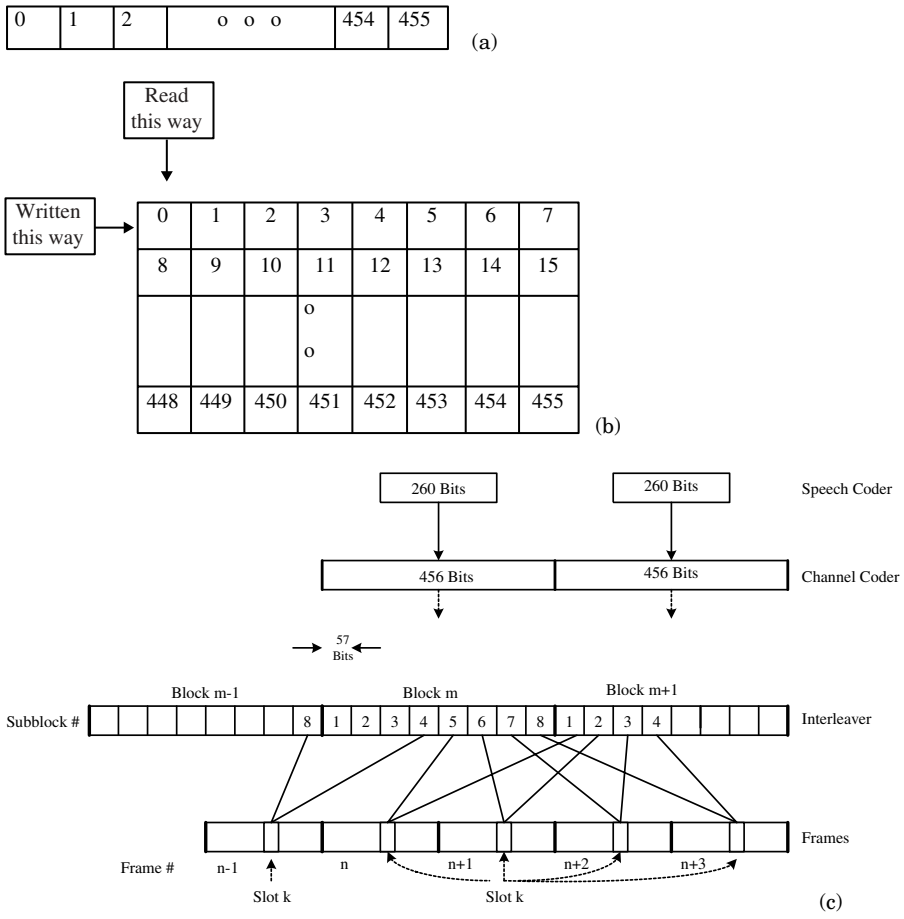


Figure 5-8
Interleaving scheme for full-rate speech coding

Modulation Technique—GMSK

The modulation technique used in GSM is MSK, which is a special form of *continuous-phase frequency shift keying* (CPFSK) [6]–[8], [1], [2]. To understand CPFSK, consider the simple case of binary FSK. There are only two symbols, -1 and +1, and the symbol rate is the same as the bit rate. To transmit +1, the frequency of the carrier is set to $f_c + \frac{1}{2}\Delta f$, where f_c is the center frequency and Δf is the peak frequency deviation. Similarly, the symbol -1 is transmitted

by setting the carrier frequency to $f_c - \frac{1}{2}\Delta f$. Notice that in this case, depending upon the incoming data, the carrier frequency may change abruptly once every symbol period. Similarly, the phase of the carrier also changes in discrete steps. For example, it changes by $\pi\Delta f$ if the symbol is $+1$ and by $-\pi\Delta f$ if the symbol is -1. A consequence of this discontinuity is that the power spectral density of the modulated signal contains large side lobes outside the main lobe. Thus, to preserve the integrity of the signal at the receiver, very often a larger bandwidth is needed than is usually available in the system.

To prevent the occurrence of these side lobes and thus minimize the bandwidth requirement, the carrier frequency may be changed continuously by a carefully chosen modulating signal. For example, the FM signal may be written as

$$y(t) = A\cos\left(2\pi f_c t + 2\pi\Delta f\int_0^t m(t')dt'\right) \qquad (5\text{-}1)$$

where $m(t)$ is the modulating signal. Since the baseband signal in GMSK is passed through a Gaussian filter before applying it to the modulator, $m(t)$ will not have any discontinuity. The important thing to observe in equation 5-1, though, is that even if $m(t)$ is discontinuous, its integral is continuous, and thus the phase of the modulated signal is always continuous. That is why it is called *continuous-phase FSK*.

In the above expression, the modulation index is defined as $\beta_{FSK} = \dfrac{2\Delta f}{R_b}$, where R_b is the bit rate. For orthogonal detection of FSK signals, the minimum frequency deviation is $\Delta f_{\min} = 0.25R_b$ corresponding to $\beta_{FSK} = 0.5$. The two carrier frequencies are, therefore, $f_c + 0.25R_b$ and $f_c - 0.25R_b$. In this case, the bandwidth required is minimized, and the CPFSK modulation is termed MSK.

There are a number of ways to construct an MSK signal. For example, the modulated signal during any symbol period may be given by

$$y(t) = m_I\cos\left(\frac{\pi t}{\tau}\right)\cos\omega_c t + m_Q\sin\left(\frac{\pi t}{\tau}\right)\sin\omega_c t \qquad (5\text{-}2)$$

where T is the symbol period and is equal to twice the bit period T_b. m_I and m_q are either $+1$ or -1 corresponding to the odd and even bits of the incoming data stream, respectively. It can be shown that the phase changes linearly with time, the maximum over any bit period being $\pm\pi/2$.

MSK offers the following advantages. The side lobes of the power spectral density of an MSK signal are considerably reduced, and most of the energy appears in the main lobe. Consequently, the bandwidth required is much less. Because there are no abrupt changes in the phase trajectory of the modulated signal, its envelope remains virtually constant. As such, the signal can be amplified with a class C power amplifier, which is generally much more efficient than a linear amplifier.

Rectangular pulses, when transmitted over a channel with a limited bandwidth, cause *intersymbol interference* (ISI), and may also lead to significant adjacent channel interference. To minimize this interference, the baseband signal is passed through a pulse-shaping filter before modulating the RF carrier. The filter that is used in conjunction with MSK is a Gaussian filter [2] whose transfer function is given by

$$H(\omega) = Ae^{-k^2\omega^2} \tag{5-3}$$

where A and k are constants, and ω is the frequency in radians. Notice that equation 5-3 is similar to the density function of a Gaussian random variable. The parameter k determines the normalized, 3-dB bandwidth ω_0 through the following expression:

$$\omega_0 = \frac{1}{1.6986k} \tag{5-4}$$

The frequency response of the Gaussian filter has a sharp cutoff. Its impulse response spreads out in time as bandwidth is decreased (resulting from larger values of k) and has virtually no overshoot, properties that make the output of the modulator suitable for amplification with nonlinear class C power amplifiers.

Figure 5-9 shows the modulator in a functional block diagram form.

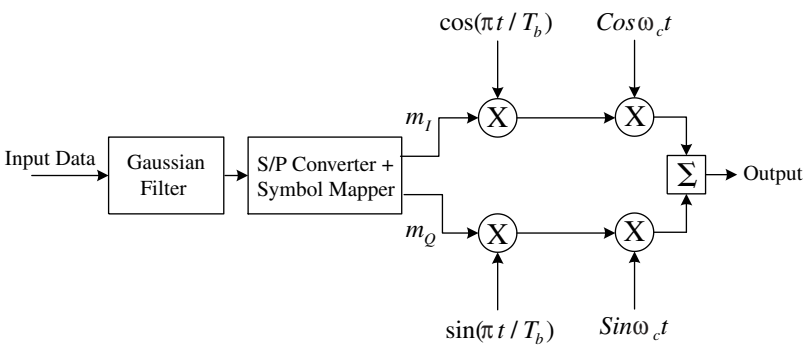


Figure 5-9
A functional block diagram of a GMSK modulator using a digital Gaussian filter

Logical Channels

As we have indicated before, many different types of information are required to be transmitted in GSM. To do this, the concept of logical channels is used. Broadly, there are two types of logical channels, traffic and control.

- *Traffic Channels* (TCH) These channels may be full-rate or half-rate. The full-rate traffic channels are used for encoded speech at 22.8 kb/s or user data at rates of 12 kb/s, 6 kb/s, and 3.6 kb/s. The half-rate channels are intended for encoded speech at a lower bit rate, such as 11.4 kb/s, and user data at 4.8 and 2.4 kb/s.
- *Signaling Channels* There are three categories of these channels:

  A. *Broadcast Control Channel* (BCCH) As the name implies, it is a point-to-multipoint channel in the downlink direction, and carries system and cell-related information identifying the network and cells, information used in cell selection and handoff, the current control channel configuration parameters, and so on. There are actually two channels here:
     - *Synchronization Channel* (SCH) that broadcasts information so that a mobile station can synchronize to the base station

- *Frequency Correction Channel* (FCCH) that enables mobile stations to adjust their frequencies when necessary.

B. *Common Control Channels* (CCH) There are three channels in this category:
   - *Paging Channel* (PCH) This downlink, point-to-multipoint channel is used to broadcast paging messages to mobile stations.
   - *Random Access Channel* (RACH) This is an uplink, point-to-point channel over which mobile stations send a call origination request or response to a page using a multiple-access, slotted Aloha scheme. Because more than one user may access the channel simultaneously, there is some non-zero probability of collision, in which case users must back off and wait a random period before seizing the channel again.
   - *Access Grant Channel* (AGCH) The purpose of this downlink, point-to-point channel is to indicate to a requesting mobile station which traffic channel or which stand-alone dedicated control channel it is being assigned.

C. *Dedicated Control Channels* (DCCH) There are two types of these channels:
   - *Stand-alone Dedicated Control Channel* (SDCCH) This is a point-to-point channel and operates on both uplinks and downlinks at 1/8 of the data rate of a traffic channel.[4] A mobile station may use it for such things as registration, authentication, or location updates before it is assigned a traffic channel. After the assignment of the traffic channel, SDCCH must be relinquished.
   - *Associated Control Channel* (ACCH) This type of channel is associated with a traffic channel. There are two channels of this type—the slow and the fast. The *slow associated control channel* (SACCH) is used by a mobile station to

[4]This rate is achieved by transmitting the information only once every eight TDMA frames.

send, for example, signal strength measurements that are required in a handoff decision.

The purpose of the *fast associated control channel* (FACCH) is to send signaling and control information after a connection has already been established with a mobile station. This is done by momentarily "stealing" the traffic channel, as it were, by interrupting its voice transmission for about 20 ms and sending signaling messages during that period.

GSM Frame and Slot Structure

In GSM, information is sent out in frames. Each frame is 4.615 ms long and is divided into eight slots. Each slot is 0.577 ms wide and contains 156.25 bits. Thus, the data rate on each physical channel is 270.8 kb/s. The slot and frame structures are shown in Figure 5-10. Multiframes are constructed with multiple frames. One such multiframe containing 26 frames is shown in this figure. It carries traffic channels, the slow associated control channel, and the fast associated control channel. A second multiframe, which is not shown here,

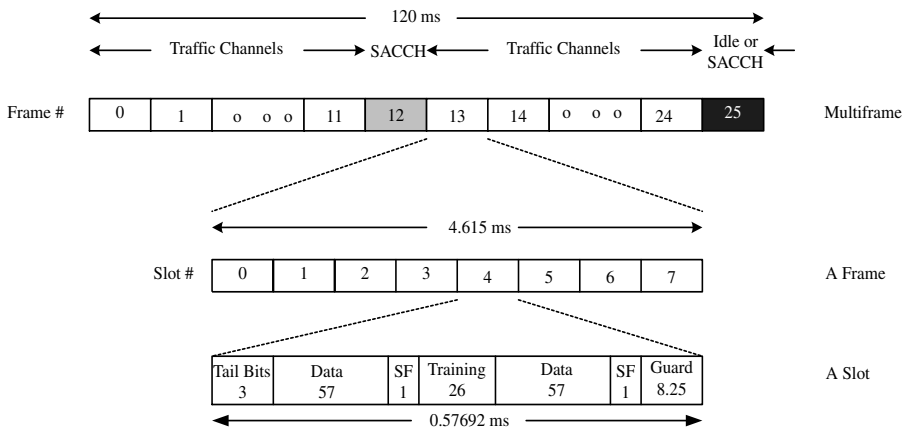


Figure 5-10 The TDMA slot and frame hierarchy in GSM

consists of 51 frames and includes, in addition to the traffic channels, a broadcast control channel and a dedicated control channel.

Each user is generally assigned one slot per frame, during which it transmits its data in the form of a burst. Five kinds of bursts have been defined in GSM:

- *Normal burst* This type of burst is used for all traffic and control channels except the synchronization channel, the frequency correction channel, random access channel, and dummy channel. The normal burst is shown in Figure 5-11(a). The slot begins with three tail bits, which are all zeros, and then follows with a 57-bit segment of user data or encoded speech. The *Stealing Flag* (SF) bit indicates whether this particular slot is being transmitted by interrupting the normal user data flow on a traffic channel. The next field is a 26-bit training sequence that, along with the tail bits, is used by an equalizer to correct ISI at the receiving end. The tail bits—three at the beginning and three more towards the end—are needed so that the equalizer knows when to start and when to stop. The training sequence is followed by 1 more SF bit and another 57 bits of user data. Adjacent slots are separated by 8.25 bits of guard period to minimize interference.
- *Synchronization channel burst* The format of this burst is shown in Figure 5-11(b). It transmits, among other things, a 64-bit binary sequence that a mobile station can use for timing recovery and synchronization. An exact copy of this sequence is stored in the mobile memory. Thus, by comparing the received sequence with this copy, a mobile station can determine the timing and correctly synchronize with the base station. The 39-bit data segments preceding and following the training sequence

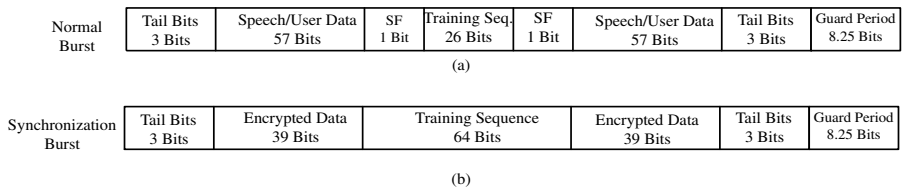


Figure 5-11 The burst formats in GSM

include the base station identification code and the TDMA frame number. This data is encrypted to protect against eavesdropping.

- *Frequency correction channel burst*
- *Random access channel burst*
- *Dummy burst*

A multiframe may consist of 26 or 51 TDMA frames, whereas a superframe contains 1,326 frames and is 6.12 seconds long. 2,048 superframes make one hyperframe.

Data Services in GSM

The data services in GSM include the following:

- *Circuit-switched data* As mentioned before, each slot of a GSM frame can be used to transmit circuit-switched data at rates up to 9.6 kb/s. Higher rates are possible if more than one slot is allocated to an application. For example, when all 8 slots of a frame are bundled together, a data rate of 76.8 kb/s is achieved. A GSM network that provides circuit-switched data services is shown in Figure 5-12. The interworking function, labeled IWF in

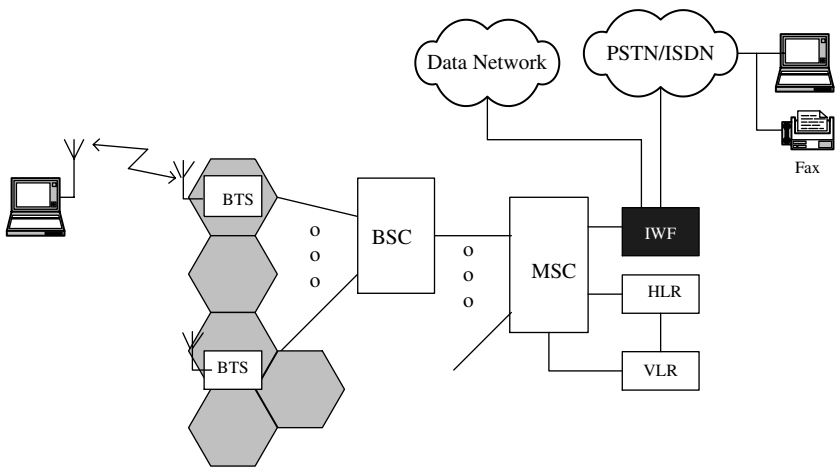


Figure 5-12 A GSM network that provides circuit-switched data services

the figure, interfaces GSM to a private or public network such as a PSTN, ISDN, circuit-switched public data network, packet-switched public data network, and so on. Because the data communication protocols of a mobile station may be different from those of devices it is in communication with across a network, entity IWF may perform protocol conversion, rate adaptation, and so on.

- *Short messaging service* A mobile station in GSM may transmit or receive short alphanumeric messages during both idle and active call states.
- *GPRS* This is a new service available with GSM Phase 2+ that enables multiple users to transmit packet data over a single slot. In this section, we will present a brief description of GPRS.

General Capabilities and Features of GPRS

In circuit-switched data services, when a user wants to transmit or receive any data, first a physical channel is set up using the normal GSM call control procedures. Because data usually comes in bursts separated by variable periods of inactivity, the channel may remain idle for a considerable length of time, depending upon the type of data services being used. One could, of course, release the channel during inactive periods of data and reestablish the connection when user data is ready. However, this approach is not very efficient or practical, because delays associated with call control procedures for setting up a physical channel are relatively long. A packet switching system, where multiple users may transmit their data over the same physical channel using the so-called *virtual circuits*, overcomes this problem by taking advantage of the statistical nature of the traffic arrival process. The virtual circuits may be either permanent or switched. But even when they are switched, call control delays for setting up or tearing down a virtual circuit are usually very small.

As we mentioned before, GPRS is a new feature of GSM that provides the capability of packet mode transmission of user data and signaling information using the existing GSM network and radio resources. Each physical channel is shared by multiple users. The

channel access mechanism has been optimized for intermittent, short bursts as well as large volumes of data, allowing data to be transmitted within about 0.5 to 1.0 seconds of a reservation request. It supports both IP and X.25 protocols and real-time as well as non-real-time data. Both point-to-point and point-to-multipoint communications are possible. There is no restriction on the transfer of SMS messages over GPRS channels.

In packet switching, it is necessary to use a set of data communication protocols so that the transmission is efficient and error-free. Protocols that are of interest here are usually the lower-layer protocols such as the *logical link control* (LLC) and *medium access control* (MAC).

Users are allowed to request a desired *quality of service* (QoS) from the network. However, only a limited number of QoS profiles are supported. Different modes of operation are possible. For example, in one mode, a mobile station can receive both GSM and GPRS services simultaneously (that is, a voice call and packet mode data transfer at the same time). In another mode, it can only receive the GPRS service. In the third mode, the mobile station monitors control channels of both GSM and GPRS, but can receive services from only one of them at a time (that is, either a voice or packet mode data). Four channel-coding schemes, designated CS-1, CS-2, CS-3, and CS-4, with coding rates of $1/2$, $2/3$, $3/4$, and 1, respectively, are supported. The throughput depends on the coding scheme used: with CS-1, the maximum throughput is about 9 kb/s, whereas with CS-4, it is 21.4 kb/s. Because a user may be assigned all eight slots of a frame, the per-user throughput may be in excess of 160 kb/s.

GPRS Network Architecture

Figure 5-13 is the architecture of a general GPRS network. The interface points between different elements of the network have also been indicated. To see the difference between a GSM and a GPRS network, compare this figure with Figure 5-2. Notice that in GPRS, there are only two new entities:

- *Serving GPRS Support Node* (SGSN) As the name implies, the SGSN provides GPRS services to a mobile station in the serving

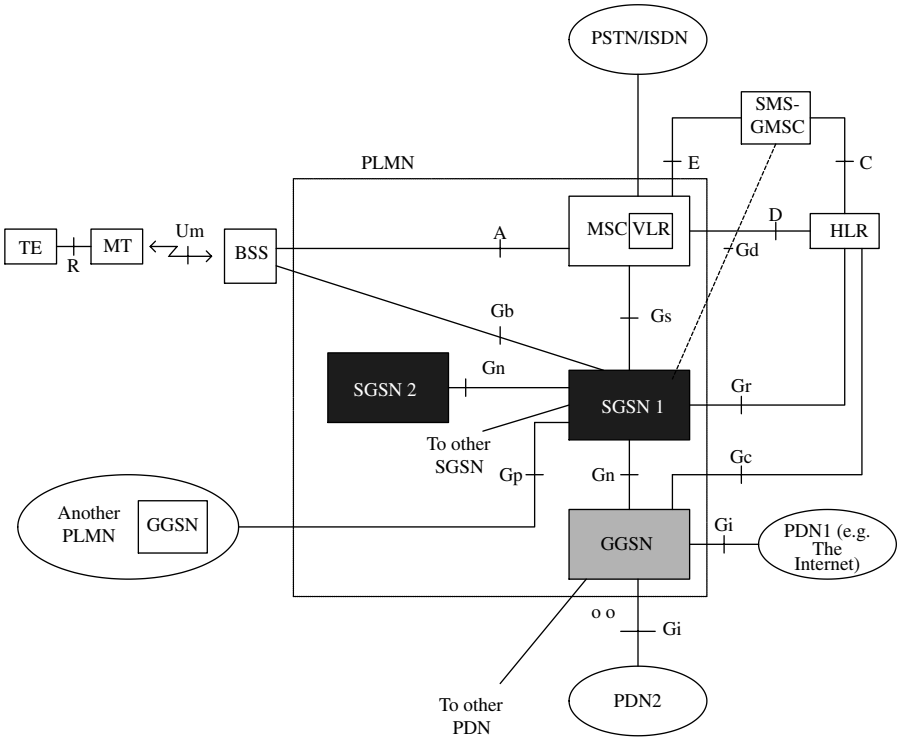


Figure 5-13
The architecture of a GPRS network

area of its associated MSC. A PLMN may have more than one SGSN, in which case, the SGSNs are connected together over an IP-based Gn interface. Two different PLMNs, on the other hand, are connected over a Gp interface. A serving GSN connects to a gateway GSN via a Gn interface and to its BSS over a Gb interface that uses the Frame Relay protocol at the link layer.

An SGSN node locates mobile stations subscribing to GPRS services and adds this information to the HLR. Another function of the SGSN is to control user access to the network by performing authentication using the same encryption keys and algorithm as in GSM. Optionally, it can also perform signaling and control for non-GPRS services. For example, it can support short messaging service over a GPRS radio channel and efficiently process paging messages and mobile location information required in GSM circuit-switched calls.

- *Gateway GPRS Support Node* (GGSN) GGSN provides an interface between a GPRS network and any external network such as a *packet-switched public data network* (PSPDN). Thus, as an example, whenever a PSPDN has a packet to send to a PLMN, it comes first to the GGSN. The gateway GSN contains the routing information of all mobile stations attached to it and forwards an incoming packet appropriately en route to its destination. It may request information from an HLR or provide information to the HLR when necessary. Both SGSN and GGSN have IP routing functionality, and as such may be connected together by an IP router.

In the current version of cellular systems (that is, 2G+), GPRS is supported by adding packet-handling capabilities to the base station controller. This is done by means of an interface called *packet control unit* (PCU) as shown in Figure 5-14. In a fully evolved 3G system, the interface to a GPRS network would be integrated into the UTRA BSS.

GPRS Protocol Stacks

The GPRS protocol stacks required in a mobile station, BSS, SGSN, and GGSN, are shown in Figure 5-15. Although the networking

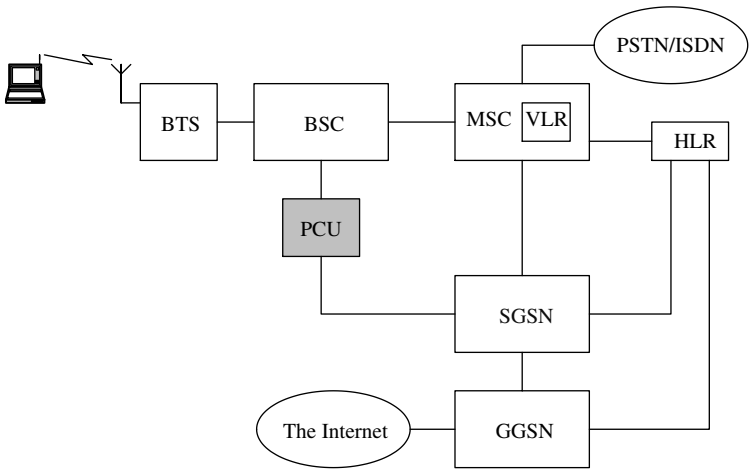


Figure 5-14
Support of GPRS in 2G+ networks

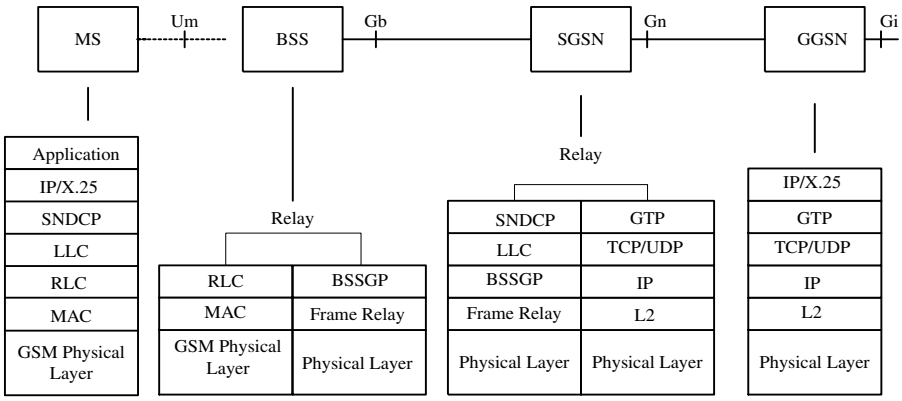


Figure 5-15 GPRS protocol stacks at a few reference points

protocol is shown in the figure to be either IP or X.25, GPRS is fully capable of supporting applications based on any standard data protocol.

GPRS protocols at various layers are thoroughly described in References [14]–[22]. Here, we provide only a short description of the protocol at each layer:

- *Subnetwork Dependent Convergence Protocol* (SNDCP) SNDCP, which, in the protocol hierarchy, lies between the network layer (that is, IP/X.25) and the LLC layer, takes the network layer PDUs (corresponding to different protocols) and converts them into a format that is suitable for transmission over the underlying radio interface network. For example, if the protocol at the layer above it is IP, the SNDCP will take the IP packet, compress its header, and pass it to the LLC layer. Similarly, when it receives a packet from the LLC layer, it may decompress the header and pass it to the IP layer. User packets may have variable lengths and are segmented, if necessary. Both acknowledged and unacknowledged data transfer is possible. Other functions performed at this layer include
  - Data transfer using negotiated QoS profiles
  - Security and encryption of user data and control to provide protection against eavesdropping
- *Logical Link Control* (LLC) The data link layer at the mobile station (the Um reference point) consists of two sublayers: the

upper sublayer known as LLC and the lower sublayer consisting of a *radio link control* (RLC) and a MAC sublayer. The LLC sublayer is based on the *link access procedures of the ISDN D channel* (LAPD) and supports procedures for the following:

- Unacknowledged data transfer. The Frame Relay protocol is a subset of LAPD procedures using the unacknowledged information transfer mode.
- Acknowledged data transfer.
- Flow control.
- Error recovery using sequence numbering in the acknowledged transfer mode.
- Ciphering of logical link PDUs in both acknowledged and unacknowledged transfer modes.

- *RLC* The RLC protocol provides a reliable transmission of data blocks over the air interface using a selective *automatic repeat request* (ARQ)-type procedure, where data blocks received in error are retransmitted by the source.
- *MAC* The MAC sublayer controls access of the physical medium by mobile stations using a slotted Aloha scheme by resolving contention among multiple users or among multiple applications of an individual user and then granting the requested access in a manner that ensures efficient utilization of bandwidth.
- *GPRS Tunneling Protocol* (GTP) In GPRS, address and control information are added to protocol data units so that they can be routed within a PLMN or between two PLMNs. The protocol that defines this process is known as the GTP.[5] Simultaneous

[5]The term *tunneling* is used to mean encapsulating an original packet with a new header. Its use is quite common in packet-switched networks. Suppose that an IPv6 packet has to be sent over a network that is using the older IPv4 protocol. In this case, we could take the original IPv6 packet, add the IPv4 header to it, and send the resulting packet over the network. That would be called tunneling. Another example is IP over ATM where an IP packet, when it first enters an ATM device, is encapsulated with an 8-octet header before it is sent out over ATM.

operation of two modes is possible—unacknowledged mode for UDP/IP and acknowledged mode for TCP/IP.

- *Relay function* It provides a procedure for forwarding a packet received at a node to the next node *en route* to its destination. In the BSS, LLC PDUs are relayed between Um and Gb interfaces. In the SGSN, packet data protocol (that is, IP and X.25) PDUs are relayed between interfaces Gb and Gn.
- *Base Station System GPRS protocol* (BSSGP) The function of this protocol is to provide multiple, connectionless, layer 2 links and to transfer data, QoS-specifying parameters, and routing information between a base station and an SGSN.
- *Frame Relay* This is the link layer protocol on the Gb interface. Data is transmitted over one or more *permanent virtual circuits* (PVCs). Frames received in error are discarded. The data link connection identifier is two octets long. The maximum frame size is 1,600 octets.

The physical layer on the Um interface includes the typical, GSM radio link functions such as framing, channel encoding, interleaving, modulation, wave-shaping, synchronization, timing recovery, and so on.

For a description of TCP and IP protocols, see Reference [10].

Figure 5-15 also indicates the need for protocol conversion at different points in the network. For example, consider the serving GSN. After receiving a packet from the base station system, it must terminate the five lower layers—physical, frame relay, BSSGP, LLC, and SNDCP—and retrieve the network *protocol data units* (PDUs). These IP/X.25 PDUs must then be encapsulated in GTP, TCP/UDP, IP, and L2 in that order and sent out over its physical layer to GGSN.

Packet Structures

The packet structure at each layer of the Um interface is shown in Figure 5-16. PDUs received from the IP or X.25 layer for transmission over the air interface are segmented at the SNDCP layer into smaller packets and passed to the LLC layer where a header and

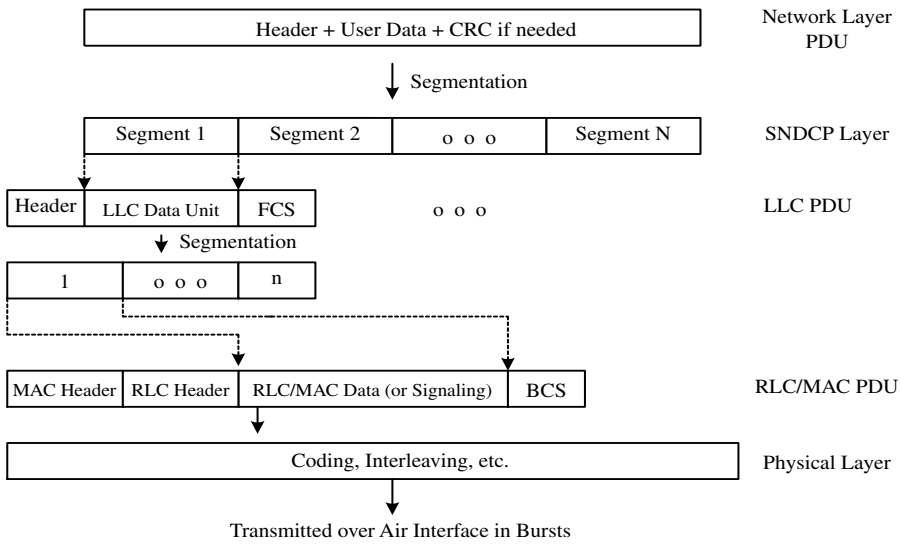


Figure 5-16
Packet structure at different protocol layers at the Um interface

frame check sequence are added to each segment. The maximum size of the LLC data unit is 1,600 octets. Each LLC PDU is further segmented, if necessary, into smaller blocks before passing it to the RLC/MAC layer. To each of these blocks are added an RLC header, a MAC header, and a block check sequence (BCS). The resulting frame, after the usual physical-layer processing, is sent out in normal bursts, each consisting of 156.25 bits, of which 114 bits are from an RLC/MAC PDU.

Logical Channels

Broadly speaking, there are three types of logical channels for transmitting packets in GPRS. They are packet broadcast control channel, *packet common control channel* (PCCH), and traffic channels. Some operate only on uplinks, some on downlinks, and the rest on both uplinks and downlinks (that is, they're bidirectional).

Uplink Channels *Packet Random Access Channel* (PRACH) This is a common control channel and is used by a mobile station to start a packet transfer process or respond to a paging message.

Downlink Channels *Packet Broadcast Control Channel* (PBCCH) It broadcasts system-specific parameters to all mobile stations in a GPRS serving cell.

The following are common control channels:

- *Packet Paging Channel* (PPCH) The GPRS network uses this channel to transmit paging messages before sending user packets.
- *Packet Access Grant Channel* (PAGCH) When a mobile station wants to initiate a data transfer, it transmits a Packet Channel Request message on a PRACH or on a RACH in the absence of a PRACH. In reply, the base station sends a Packet Immediate Assignment message on a PAGCH, reserving one or more packet data transfer channels for that mobile station. Similarly, the network may send on this channel a resource assignment message to a mobile station.
- *Packet Notification Channel* (PNCH) This channel is used to notify a group of mobile stations prior to sending packets to those stations in a point-to-multipoint fashion.

Bidirectional Channels A *Packet Data Transfer Channel* (PDTCH) is allocated to a mobile station for transferring their data packets. A given user may request, and be granted, more than one PDTCH.

A *Packet Associated Control Channel* (PACCH) carries signaling information, such as an *acknowledgment* (ACK), in response to a data block transfer, a resource assignment message in response to a resource request, or power control information. Only one PACCH is assigned to each mobile station, and is associated with all packet data transfer channels that may be allocated to that station.

Logical channels are multiplexed at the MAC layer onto physical channels on a block-by-block basis. Physical channels used for GPRS packet data transmission are known as *packet data channels* (PDCH).

Packet Transmission Protocol

Multiple users may transmit packets on a PDCH on a time-shared basis. Each PDCH consists of one time slot of a TDMA frame. How-

ever, a mobile station may be assigned up to eight PDCHs for packet data transmission.

A cell may permanently set aside a fraction of its available physical channels exclusively for packet data transmission and the rest for the usual voice traffic. Alternatively, it may use a dynamic allocation scheme whereby one or more channels out of its available pool of channels are allocated to packet data transmission on a demand basis, and are deallocated and returned to the pool when there is no longer any need for them. The number of packet data channels active at any time depends on the number of simultaneous users and the volume of traffic generated by each user. However, there must be at least one PDCH to enable transfer of control and signaling information (as well as user data if necessary). It is not necessary that the same PDCH be used to send packets to/from a given mobile station.

Multiple users transmit on a PDCH using a slotted Aloha, multiple-access reservation scheme. In the event of transmission errors, an ARQ protocol is used that provides error recovery by selective retransmissions of RLC blocks. To this end, GPRS employs the concept of a *temporary block flow* (TBF), which is actually a physical connection between a mobile station and the network, allowing the transfer of RLC/MAC blocks.[6] Each RLC data block or RLC/MAC control block includes in its header a *temporary flow identifier* (TFI) that indicates the TBF to which the block belongs.[7] Furthermore, all downlink RLC/MAC blocks contain in their header an *uplink state flag* (USF) that indicates which mobile station (or application) can use the next uplink RLC block on the same time slot. In this way, different mobile stations may be multiplexed on the same PDCH when necessary.

A mobile station transfers packets to an SGSN following the state diagram of Figure 5-17. The corresponding state machine representation of an SGSN is similar.

In the IDLE state, a mobile station may select or reselect a cell, but its location or routing information is not available to the SGSN.

---

[6]It is temporary because it exists only as long as there is an RLC/MAC block to send and is removed when it is no longer needed.

[7]On any PDCH, the same TFI may be used in the uplink and downlink directions. Similarly, different PDCHs may use the same TFI.

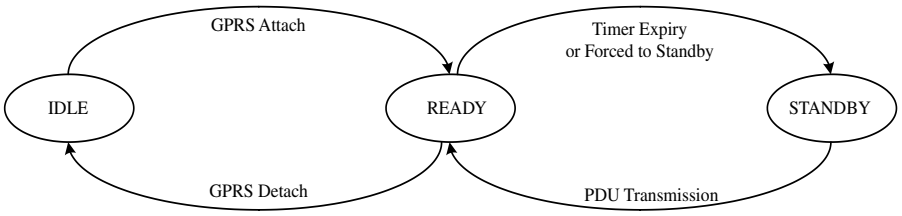


Figure 5-17
A state machine model of the packet data transfer function of a mobile station

In other words, it is not attached to the mobility management function, and therefore cannot receive or originate a call.

When the mobile station establishes a logical link to an SGSN, it enters the READY state. The mobile is now attached to the mobility management function and can initiate a mobile-originated call on a PRACH (that is, a packet random access channel) or monitor the packet-paging channel to see if there is any packet transfer request from the network. If there is no PRACH available yet, the GPRS-attached mobile station may use the GSM common control channel. After being allocated appropriate resources from the network, the mobile station can begin to transmit and receive packets.

The mobile station remains in the READY state as long as there is any packet to send. Even when there is no packet pending in its buffer, it may continue in the READY state for a certain length of time that is marked by starting an associated timer. As the timer is running, the network has the capability to preempt the timer and force the mobile station into the STANDBY state. When the timer expires, the mobile station changes to the STANDBY state. While in the READY state, the mobile station may power down by performing a GPRS-detach procedure. It then enters the IDLE state, whereupon the SGSN deletes the location and routing information of the mobile.

In the STANDBY state, the mobile station is still GPRS-attached and sends the SGSN its location and routing information periodically and each time it moves into a new *routing area* (RA). While in this state, it can transmit a PDU and then transition to the READY state.

The packet transfer procedure when initiated by a mobile station is shown in Figure 5-18. The mobile station sends a packet channel request over a packet random access channel (or in its absence, a

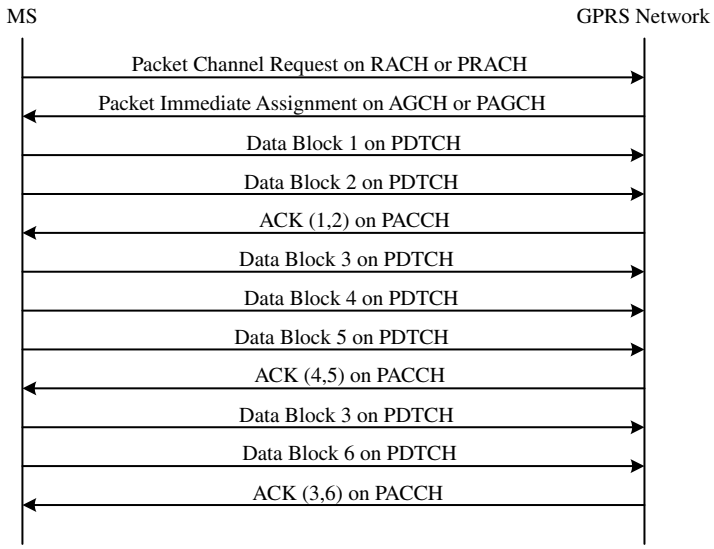


Figure 5-18
Packets in a mobile-originated transfer in a GPRS system

random access channel). The network replies by reserving the necessary resources required by the MS and sending a packet immediate assignment message. In this case, the access method completes in a single phase. In a two-phase access procedure, when the network sends a packet immediate assignment, it reserves only the resources required by a mobile station to transmit a packet resource request. Consequently, the mobile station sends this request message indicating resources it needs, whereupon the network makes the necessary reservation and replies with a packet resource assignment message.

After receiving the packet immediate assignment, the MS can begin to send data packets. The network may withhold acknowledgment until after receiving a few packets. When it receives a block in error (say, block 3 in this example), it sends an ACK (4,5), excluding block 3 from this acknowledgment, as shown in Figure 5-18. In this case, the mobile station performs a selective retransmission of block 3 only (and not blocks 3, 4, and 5), transmitting it along with block 6. Alternatively, the network could send a NACK in the event of an error.

The packet transfer procedure initiated by the network (that is, an SGSN) is shown in Figure 5-19. The mobile station monitors the

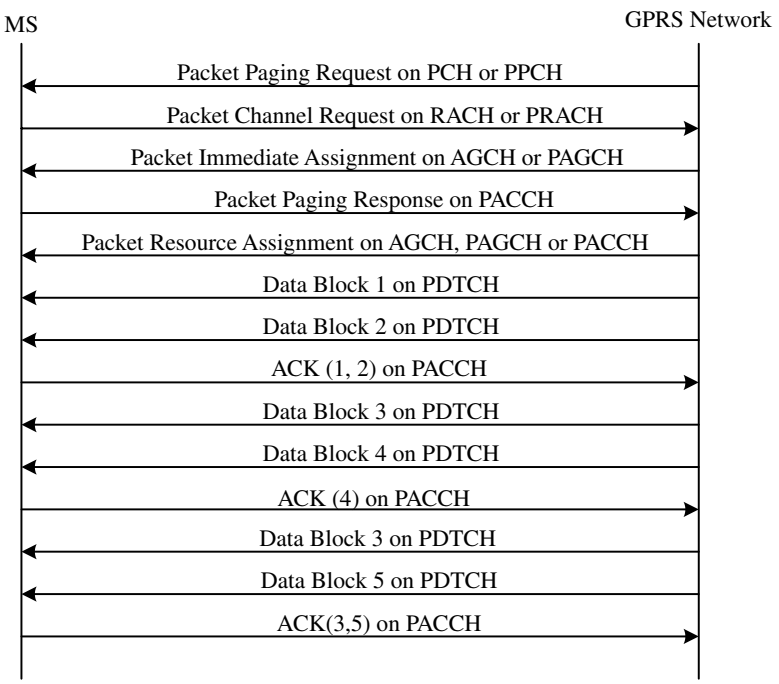


Figure 5-19 *Packets in a network-initiated transfer in a GPRS system*

packet paging-channel (or in its absence, the paging channel). When it receives a packet-paging request, it sends a packet channel request. The network answers by sending a packet immediate assignment. This is followed by a packet-paging response from the MS and a packet resource assignment. At this point, the network may begin to transmit data blocks to the MS.

Summary

In this chapter, we have presented a brief description of the GSM system. Its features and capabilities have been summarized, and some technical detail has been provided about the speech encoder, channel encoder, interleaver, modulator, TDMA slot, frame formats, and logical channels. One of the important aspects of GSM is its data service capability such as the short messaging service and circuit-switched

data. In the short messaging service, users can transmit messages of about 160 alphanumeric characters in both point-to-point and point-to-multipoint fashion. The circuit-switched data rate per slot may be 2.4, 4.8, or 9.6 kb/s. By bundling multiple channels, a user can be provided much higher data rates, say, up to about 76.8 kb/s. The GPRS is a relatively new feature of GSM Version 2.5+ that provides packet mode data services at rates of 8 to 20 kb/s per slot. In this chapter, we have described the general capabilities and features of GPRS, its network architecture, protocols at various layers, logical channels, packet structures, and the packet transmission protocol.

References

[1] A. Mehrotra, *GSM System Engineering*. Norwood, MA: Artecth House, 1997.

[2] T.S. Rappaport, *Wireless Communications*. New Jersey: Prentice Hall, 1996, pp. 501–529.

[3] N.S. Jayant, "High-Quality Coding of Telephone Speech and Wideband Audio," *IEEE Comm. Mag.,* Vol. 28, No. 1, pp. 10–19, January 1990.

[4] P. Vary, et al., "Speech Codec for the European Mobile Radio System," *Proc. ICASSP '88*, pp. 227–230, April 1988.

[5] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, Vol. 63, pp. 561–580, April 1975.

[6] R.W. Lucky, et al., *Principles of Data Communications*. New York: McGraw Hill, 1968, pp. 200–202.

[7] J.G. Proakis, *Digital Communications*. New York: McGraw Hill, 1968, pp. 172–186.

[8] C. Sundberg, "Continuous Phase Modulation," *IEEE Comm. Mag*., pp. 25–38, April 1986.

[9] J. Cai and D.J. Goodman, "General Packet Radio Service," *IEEE Comm. Mag*., pp. 122–131, October 1997.

[10] M. Naugle, *Network Protocol Handbook*. New York: McGraw-Hill, 1994.

ETSI Standards

[11] GSM 03.60: GPRS Service Description, Stage 2.

[12] GSM 03.64: Overall Description of the GPRS Radio Interface, Stage 2.

[13] GSM 04.60: GPRS, Mobile Station—Base Station System (BSS) Interface, Radio Link Control/Medium Access Control (RLC/MAC) Protocol.

[14] GSM 04.64: GPRS, Logical Link Control.

[15] GSM 04.65: GPRS, Subnetwork Dependent Convergence Protocol (SNDCP).

[16] GSM 07.60: Mobile Station (MS) Supporting GPRS.

[17] GSM 08.08: GPRS, Mobile Switching Center—Base Station Subsystem (MSC-BSC) Interface: Layer 3 Specification.

[18] GSM 08.14: Base Station Subsystem—Serving GPRS Support Node (BSS-SGSN) Interface; Gb Interface Layer 1.

[19] GSM 08.16: Base Station Subsystem—Serving GPRS Support Node (BSS-SGSN) Interface; Network Service.

[20] GSM 08.18: Base Station Subsystem—Serving GPRS Support Node (BSS-SGSN); Base Station Subsystem GPRS Protocol (BSSGP).

[21] GSM 09.60: GPRS Tunneling Protocol (GTP) Across the Gn and Gp Interface.

[22] GSM 09.61: General Requirements on Interworking Between the Public Land Mobile Network (PLMN) Supporting GPRS and Packet Data Network (PDN).

[23] GSM 2.01, Version 4.2.0, January 1993.

[24] ETSI/GSM Section 4.0.2, "European Digital Cellular Telecommunication System (Phase 2); Speech Processing Functions: General Description," April 1993.

CHAPTER 6

Universal Mobile Telecommunications System (UMTS)

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

As we indicated in Chapter 1, "Introduction," the *European Telecommunications Standards Institute* (ETSI)/*Special Mobile Group* (SMG) developed two standards for *International Mobile Telecommunication in the year 2000* (IMT-2000). One of them is the *Universal Mobile Telecommunications System* (UMTS) *Wideband Code Division Multiple Access* (W-CDMA), which is based upon a *direct-sequence CDMA* (DS-CDMA) technology and operates in the *frequency division duplex* (FDD) mode. The other is the UMTS TDD system, which is based on *time-division CDMA* (TD-CDMA) principles. The purpose of this chapter is to present an overview of the W-CDMA UMTS system as specified in the ETSI standards documents [1]–[40].

The chapter is organized as follows. We begin with a synopsis of the UMTS system features and follow it up with the *third-generation* (3G) wireless network architecture. The UMTS uses a layered protocol architecture at different interface points, each layer performing a set of specific functions. We present an overview of the radio interface protocol stack. The next few sections describe each of the constituent protocols of this stack, namely the physical layer, the medium access control, radio link control, the packet data convergence protocol, the broadcast multicast protocol, and the radio resource control protocol. Topics, such as the synchronization procedure, power controls, and handovers, are also described. The material of this chapter has been drawn from a series of standards documents. In many instances, our descriptions have been necessarily brief and comprehensive. However, we have included relevant references at the end of the chapter so that the interested reader may consult them for greater detail.

System Features

The UMTS operates in two modes—FDD and *Time Division Duplex* (TDD). In both modes of operation, the information is transmitted usually in 10 ms frames. In FDD, two distinct frequency bands, separated by a guard band, are used—one for the uplink and the other for the downlink transmission. In TDD, on the other hand, the same frequency band is used for transmissions in both directions. More

specifically, in this mode, each frame consists of a number of synchronized time slots, some of which are dedicated to uplink and the rest to downlink transmissions. The difference between the two modes is illustrated in Figure 6-1.

The UMTS has been allocated a bandwidth of 120 MHz in the FDD mode and 35 MHz in the TDD mode in the 2000 MHz spectrum range. When operating as paired bands as shown in Figure 6-1(a), the transmitter and receiver frequencies in all *user equipment* (UE) must be spaced apart by 190 MHz. As we mentioned earlier in the book, CDMA uses *Direct Sequence Spread Spectrum* (DSSS). Because one of the goals of 3G systems is to provide multimedia and high-speed data services at rates up to 2 Mb/s, the nominal channel bandwidth is 5 MHz. A service provider may, however, adjust the channel bandwidth if necessary to optimize the spectrum utilization. The center frequency must be an integer multiple of 200 kHz.[1] The chip rate for spectrum spreading is 3.84 Mc/s.

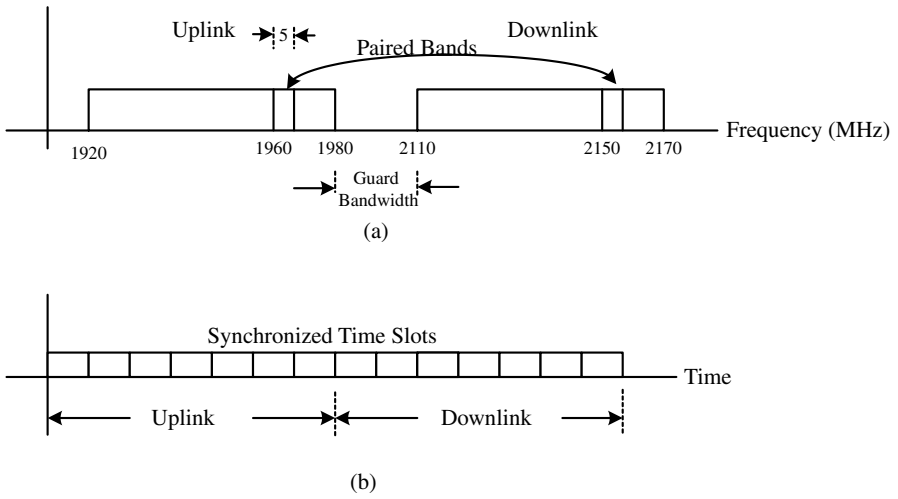


Figure 6-1
The two modes of UMTS: (a) The FDD mode and (b) The TDD mode

[1]This is called channel raster.

W-CDMA is an asynchronous system where base stations do not have to maintain a system-wide reference time scale. However, each cell or each sector of a cell must now use a different scrambling code. Because there is no global timing reference, the time offsets between signals received from multiple users by a base station in such a system may be quite significant. Since the cross-correlation between scrambling codes assigned to different users is no longer zero, the received signal from any user depends not only on the signal from that user but also on the signals received from all other users over a number of consecutive symbol periods. Thus, multiuser detection would be useful in such a system.[2] In contrast, cdmaOne is synchronous because all base stations in the system use a reference time that is based on the *Global Positioning System* (GPS) time derived from the Universal Coordinated time. More specifically, the I and Q channels at any base station in cdmaOne are spread by two maximal-length pilot pseudonoise sequences with an offset that is unique for that base station. This simplifies and accelerates cell searching at a mobile station.

The following specifications apply to the radio transmission and reception in the UMTS FDD mode. The separation between the uplink and downlink frequency bands must be in the range of 134.8 to 245.2 MHz. The maximum transmitter power of the user equipment is in the range of 21 to 33 dBm (that is, 125 mW to 2 W). The receiver sensitivity, which is nominally defined as the minimum receiver input power at the antenna port such that the *bit error ratio* (BER) is 0.001 or less, is −117 dBm for the UE and −121 dBm for a base transceiver station.[3] With *transmit power control* (TPC) commands, the UE adjusts its transmitter power output by 8 to 12 dB in steps of 1 dB, by 16 to 24 dB in steps of 2dB and by 16 to 26 dB in steps of 3 dB. A base station, on the other hand, adjusts its transmit power by 8 to 12 dB in steps of 1 dB and by 4 to 6 dB in steps of 0.5 dB. These features are summarized in Table 6-1.

---

[2]Multiuser detection principles are briefly described in Chapter 3, "Principles of Wideband CDMA (W-CDMA)."

[3]The receiver sensitivity at a base station may be less because its performance can be improved using multipath diversity, adaptive antenna arrays, or multiuser detection techniques.

Table 6-1

W-CDMA system features

Spectrum allocation	FDD mode: 1850–1910 MHz for uplink, 2110–2170 for downlink. TDD mode: 1900–1920 MHz and 2010–2025 MHz. Each of these bands for the TDD mode is used for both uplink and downlink transmissions.
Channel spacing	5 MHz
Center frequency	Integral multiples of 200 kHz
Separation between uplink and downlink frequency bands	134.8–245.2 MHz
Chip rate	3.84 Mc/s
Modes	FDD and TDD
Transmitter power output of UE	21, 24, 27, or 33 dBm
Receiver sensitivity	−121 dBm for base stations and −117 dBm for UE at a bit error rate of 10^{-3}
Power control steps	1, 2, or 3 dB for UE and 0.5 or 1 dB for base stations
Maximum possible change in the transmit power level on TPC commands	26 dB for UE and 12 dB for base stations
Data rates	144 kb/s in rural outdoor, 384 kb/s in urban/suburban outdoor, 2 Mb/s in indoor or low-range outdoor

Wireless Network Architecture

In many instances, standards documents describe protocols and interfaces at some well-defined reference points. A general network architecture with these reference points is shown in Figure 6-2. The network may be partitioned into two broad entities—the *Universal Terrestrial Radio Access Network* (UTRAN) and the core network. The UTRAN is responsible for establishing connections between UE and the rest of the network. A *Radio Network Controller* (RNC) is

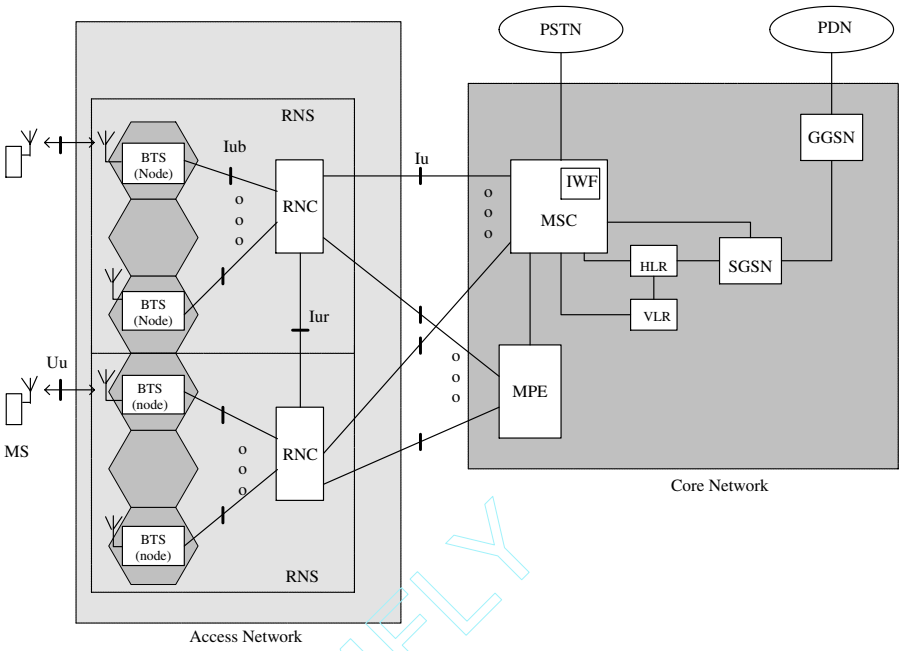


Figure 6-2
A general UMTS network architecture with various interface reference points

connected to one or more *Base Transceiver Stations* (BTS) or nodes, each of which serves a cell. The function of an RNC is to control radio resources. For example, it would assign frequencies, spreading and scrambling codes, the power levels of the various channels, and so on. The interface point between an RNC and a node is Iub. The user equipment accesses the UTRAN via the base station located in the serving cell. A *Radio Network Subsystem* (RNS), consisting of an RNC with its associated nodes, connects to the *core network* (CN) at a reference point Iu. Similarly, notice the interface point between two RNCs. An RNS is either the whole UTRAN, or only a part thereof that provides connections to a UE and includes only one RNC. For each UE connected to the network, only one RNS (called the *serving RNS*) controls the connections. However, other RNSs may assist the serving RNS as the mobile moves from one cell to another. Such an RNS is called a *drift RNS*.

With the exception of the *Multimedia Processing Equipment* (MPE), the 3G-core network is very similar to the core network of a GSM with *General Packet Radio Service* (GPRS) capabilities. A detailed description of a GSM and GPRS network may be found in

Chapter 5, "The GSM System and General Packet Radio Service (GPRS)." The MPE performs such functions as code conversion between audio coding, video coding, and control and signaling standards that might be used in a completely general network to provide interoperability amongst different multimedia terminals used in the system.

A UTRAN may consist of a number of RNSs. In the previous figure, there are two.

Radio Interface Protocol Stack—An Overview

The UMTS uses a layered protocol architecture at different interface points, each layer performing a set of specific functions. These architectures are usually described in terms of the control plane and user plane protocols. The control plane protocols are concerned with the signaling and control required to establish a connection between a UE and the network, or request specific services or resources from the network. The user plane protocols, on the other hand, specify how the user data is to be transferred across an interface after a connection has been established between the UE and the network.

Figure 6-3 shows the protocol architecture of the UTRAN. It is also the lower-layer protocols on UE. On the UTRAN side, the physical layer, which is responsible for carrying the information bits, is provided by the BTS while the other layers reside in an RNC. The *Radio Resource Control* (RRC) is a layer 3 protocol in the control plane that interfaces with the *radio link control* (RLC) sublayer of layer 2 and terminates in the UTRAN. It deals with two types of control and signaling messages—those that are generated at the higher layers and others that are generated in the RRC itself. The first type is *nonaccess stratum* (NAS) messages, which originate at a UE and terminate in the core network or vice versa. The RRC layer simply passes these NAS messages to the higher layers or to the RLC layer below en route to the physical layer. Examples of these messages include system information provided by the core network that needs to be broadcast to all UEs, signaling functions such as call control

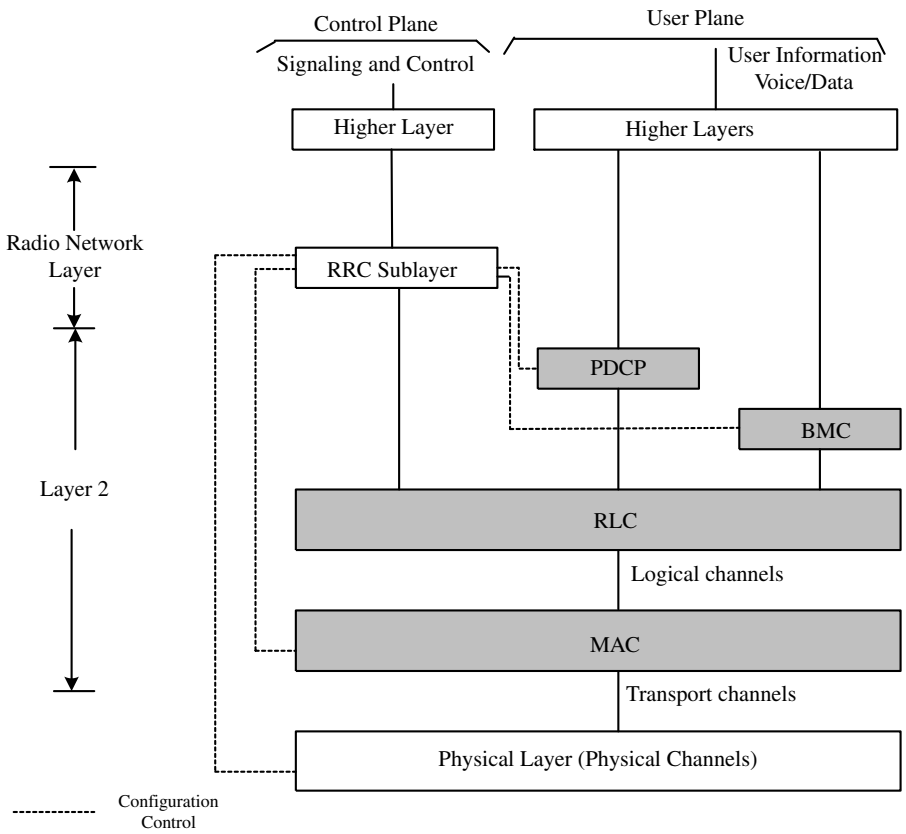


Figure 6-3
The layered protocols at the UTRAN

and mobility management, and so on. Signaling messages of the second type that originate at the RRC layer are associated with such functions as assigning radio resources (such as a common packet channel or codes), requesting a UE to perform measurements, reporting the results, and so on. Another function of the RRC layer is to configure all lower layers—all layer 2 sublayers as well as the physical layer, that is, the *Packet Data Convergence Protocol* (PDCP), *broadcast/multicast control* (BMC), RLC, *media access control* (MAC), and the physical layer. It does this by setting up direct communication links to these layers in addition to the interfaces shown in the figure.

User information is transferred between the higher layers and the physical layer via layer 2, which can be partitioned into a number of

sublayers—the PDCP, the BMC, the RLC, which is similar to the well-known *logical link* (LLC) protocol, and the MAC sublayer. Separate logical channels are used to transfer different types of information between RLC and MAC layers. Similarly, the MAC sublayer and the physical layer exchange information by means of a number of transport channels, each of which may have some defining characteristic. For example, some transport channels carry only uplink data, while others are configured only on the downlinks. Or, one transport channel may carry only system and cell-specific information that needs to be broadcast over an entire cell, while another is used to broadcast only the paging information, and so on. As we shall see, it is at the physical layer that these channels are multiplexed, or a coded, composite transport channel is demultiplexed.

The user data (such as the packet mode data) originates at the application layer and is encoded according to some higher-layer protocols such as the *Transmission Control Protocol* (TCP) at the transport layer and *Internet Protocol* (IP) at the network layer. The encoded data is eventually passed via layer 2 to the physical layer where the data stream is further processed before it is sent out over the radio link. For example, it is encoded into a forward error-correcting code, interleaved, spread out with an orthogonal channelization code, and then modulates a carrier. Similarly, the signal received over the radio interface is demodulated, despread, deinterleaved, and decoded for error detection and correction. The physical layer delivers the resulting data to the MAC layer where it is further processed and then forwarded to the upper layers. Other functions performed at the physical layer include multiplexing various transport channels, demultiplexing coded composite transport channels, frequency and time synchronization, power control, and rate matching.[4]

---

[4]As we shall see later in greater detail, power control is the process of controlling the transmitter power of a mobile station so that a base transceiver station receives an equal signal level from all mobile stations in the cell, thereby preventing a single mobile (which may happen to be closer to the base station) from swamping out the signals from other mobiles.

Rate matching means adjusting the data rate on a transport channel by repeating or deleting some bits so that the resulting data rate is equal to the data rate of the physical channel to which the transport channel is being mapped.

Any given layer provides services to its immediate upper layer via a logical point, called a *service access point* (SAP), using a set of primitives that includes REQUEST, INDICATION, ACK, and CONFIRMATION in acknowledged data transfers but only REQUEST and INDICATION in the unacknowledged mode. In Figure 6-4, when the MAC layer wants to send a packet to its peer entity at the other end of the air interface, it sends a REQUEST to the layer below. The physical layer at the receiving end, on receiving the packet, sends an INDICATION to its MAC layer, which then sends an ACK to its physical layer. On receipt of this ACK, the physical layer at the sending end sends a CONFIRMATION to the previous MAC layer.

Physical Layer

The purpose of the physical layer is to condition the digital data from higher layers so that it can be transmitted over a mobile radio channel reliably. In the transmit direction, it performs such functions as channel coding, interleaving, scrambling, spreading, and modulation. In the receive direction, these functions are reversed so that the transmitted data is recovered at the receiver. The MAC layer delivers user data and signaling over a number of transport channels.

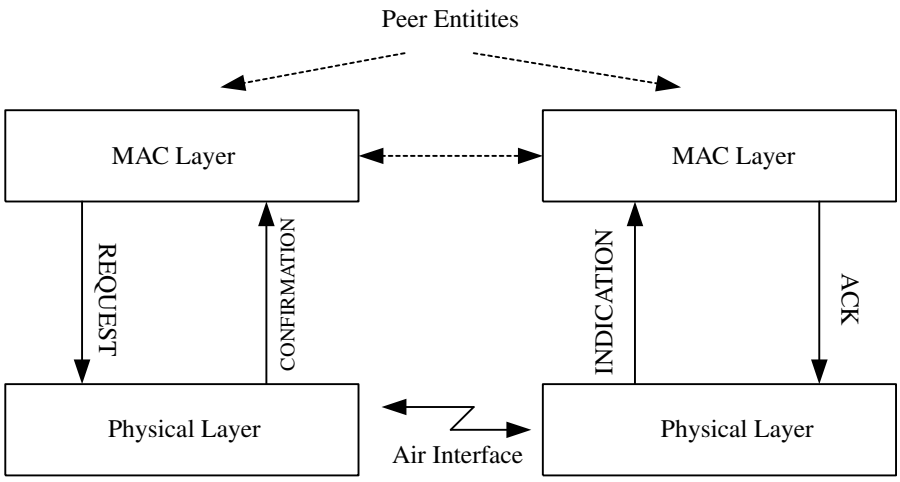


Figure 6-4
Use of primitives to transfer data between peer entities in a system with layered protocols

The physical layer maps each of these channels into a physical channel and transmits the information over the radio interface. A physical channel is characterized by its associated carrier frequency, scrambling, and channelization codes, the radio frame length, and the relative phase angle when meaningful. A radio frame is 10 ms long and consists of 15 time slots. Because the chip rate is 3.84 Mc/s, there are 38,400 chips in a frame and 2,560 chips in a slot. Except where otherwise indicated, the description of this section applies only to the FDD mode of UMTS.

Overview of Physical Layer Functions

The physical layer of the UMTS can be best explained in terms of the functions performed by a transmitter. Figure 6-5 shows a functional block diagram of a transmitter in the UE. In UMTS, the data arrives on a transport channel in blocks. We begin by encoding each of these blocks of a transport channel into a *cyclic redundancy check* (CRC) code and then serially concatenating all encoded blocks of that channel.[5] Because the length of the resulting output may exceed the maximum size, it may be necessary to segment it into a number of smaller code blocks and then pass them through a channel encoder, where the input stream is encoded into either a convolutional code of rate 1/3 or 1/2 or turbo code of rate 1/3. Sometimes it may not be necessary to use any channel coding at all, in which case there is no restriction on the size of the code block, and therefore the segmentation step may be omitted.

The output of the channel coder is applied to the first of two interleavers, where incoming bits are interleaved following certain rules such that the relative position of each bit in the output stream is different from its position in the input.

---

[5]The generator polynomials of these codes may be one of the following:

$g_1(x) = x^{24} + x^{23} + x^6 + x^5 + x + 1$

$g_2(x) = x^{16} + x^{12} + x^5 + 1$

$g_3(x) = x^{12} + x^{11} + x^3 + x^2 + x + 1$

$g_4(x) = x^8 + x^7x^4 + x^3 + x + 1$

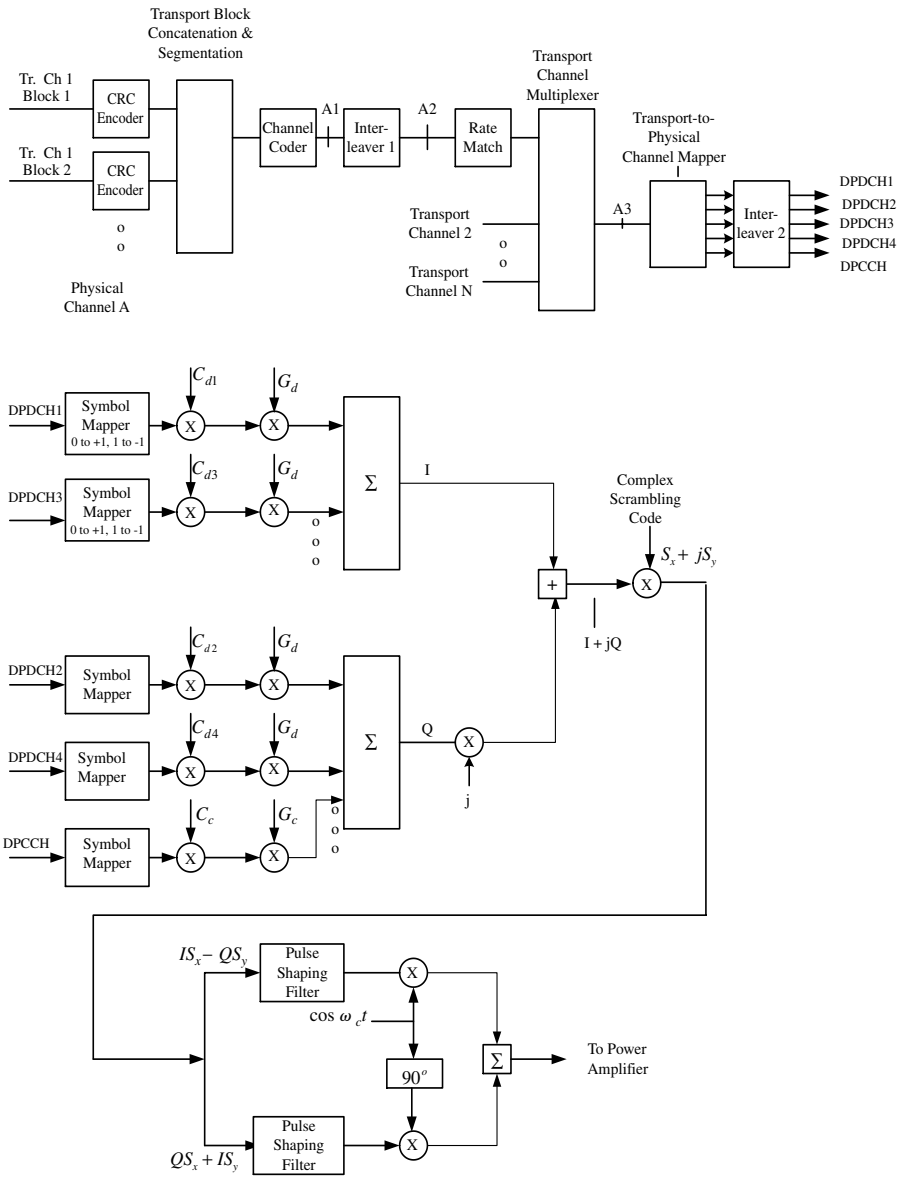


Figure 6-5
A functional block diagram of a transmitter in the UE

The number of bits in a transport block is not constant, but generally varies from one transmission time interval to another. In the downlink direction, these variable-length blocks can be handled by interrupting transmission when there are fewer bits to transmit. In

an uplink, however, the resulting bit rate after transport channels are multiplexed should be exactly matched to the available physical channel rate. This is achieved by passing the first interleaver output through the functional block labeled *Rate Match* where, depending upon the channel rate, bits are either repeated or deleted without loss of information (that is, by deleting or puncturing some coding bits).

The remaining physical-layer functions will be described in greater detail later in this chapter. For the purpose of this overview, however, we shall present a brief description. All transport channels are processed in exactly the same way as described previously, then multiplexed together, and interleaved using a second interleaver. It is worth mentioning here that as far as the broadcast channel is concerned, there may be only one transport block during each transmission time interval.

Because the transmission time interval of any transport channel is usually longer than 10 ms, which is the length of a radio frame, a number of consecutive radio frames may be required to transmit a given transport block set. Thus, it is necessary that the bit sequence at A1 contain an integral number of data segments of the same size. In the uplink direction, this is achieved, in a process called *radio frame equalization*, by appropriately padding the output of the channel coder. Although not shown in the figure, it is this padded bit sequence that gets applied to the input of the first interleaver. Consequently, the bit stream at A2 can now be divided into a number of equal segments that are suitable for transmission over consecutive radio frames.

When more than one physical channel is involved, which is generally the case, the bit sequence at A3 is split into the required physical channels before applying it to the input of the second interleaver. This process is known as *physical channel segmentation*.

Different transport channels are mapped to different physical channels and separated into two distinct sets, say, the I and Q sets. The data associated with channel set I is transmitted on the *in-phase* (I) component of the signal. Similarly, the Q set data is transmitted on the *quadrature* (Q) component of the signal. In this figure, there are only four *dedicated physical data channels* (DPDCHs) and one *dedicated physical control channel* (DPCCH). To balance the

signal, the odd-numbered data channels are placed in the I set and the even-numbered data channels together with the DPCCH in the Q set. The incoming data on each physical channel in either set is converted into real-valued symbols—a binary one mapped to −1 and a binary zero to +1, thus in essence performing the BPSK modulation. However, because there are two channel sets, the net result is QPSK modulation.

Outputs of the symbol mappers are spread to the chip rate by real-valued channelization codes. These codes are *orthogonal variable spreading factor* (OVSF) codes or Walsh sequences, and are constructed with the elements of a row of an orthogonal matrix, whose entries are either +1 or −1, so that when channels are spread with different OVSF codes, they become mutually orthogonal. To uniquely identify all physical channels at the receiver, each is spread with a separate OVSF code. Different data rates are supported on a physical channel by simply changing the spreading factor of the associated code. The symbols, after spreading, are multiplied with a gain factor that depends upon the relative power level at which the channel is to be transmitted. The resulting outputs corresponding to all different physical channels of the two sets are summed together, added in quadrature, and scrambled by a user-specific complex scrambling code. A complex code is selected for this purpose because it increases the average-to-peak power ratio. The real and imaginary parts of the scrambler output are passed through wave-shaping filters and modulate a CDMA carrier in the usual way.

The general structure of the BTS transmit functions (that is, on the downlinks) is almost the same as Figure 6-5 and therefore will not be discussed separately. Some of the differences between the two are as follows. To provide balanced spreading, the data stream on each downlink physical channel is split into two substreams, I and Q, corresponding, respectively, to the I and Q components of the signal. The two substreams are spread by the same channelization code before being scrambled by a cell-specific complex scrambling code. The radio frame equalization is not necessary here because higher layers at the UTRAN can schedule the transmission at appropriate instants or use proper spreading factors to ensure that the bit sequence after rate matching is an integral multiple of a radio frame.

Transport Channels

As we previously mentioned, the MAC layer periodically transfers higher-layer data to the physical layer in one or more transport blocks over a transport channel. A block size is given by the number of bits in the block. The size of a transport block may be different on different transport channels. The number of blocks exchanged each time between the MAC layer and the physical layer is called a *transport block set*. The interarrival time between two successive transport block sets is called the *Transmission Time Interval* (TTI). The format used for each transport block depends on the specific application. For speech, each transport block set contains only one transport block. For packet mode data, a set may consist of a variable number of transport blocks. For circuit-switched data, the number of blocks per set is usually fixed during a call, but may be greater than one. The TTI may also vary. For example, it is either 10 ms or 20 ms for speech, depending upon the coding rate of speech and 10/20/40/80 ms for circuit or packet mode data. Figure 6-6 shows transport blocks of different sizes and their interarrival times.

A transport format consists of two parts—a dynamic part and a semistatic part. The dynamic part indicates the transport block size in bits and the number of blocks in a set. The semistatic part specifies TTI, the type of channel coding (such as whether any channel coding is to be used and, if so, the coding type) and the rate matching parameter. When a transport block is delivered to the physical layer, it includes a label called a *Transport Format Indicator* (TFI) that specifies its format. The physical layer uses these TFIs associated with all parallel transport channels to construct a *Transport Format Combination Indicator* (TFCI) and sends it as part of a

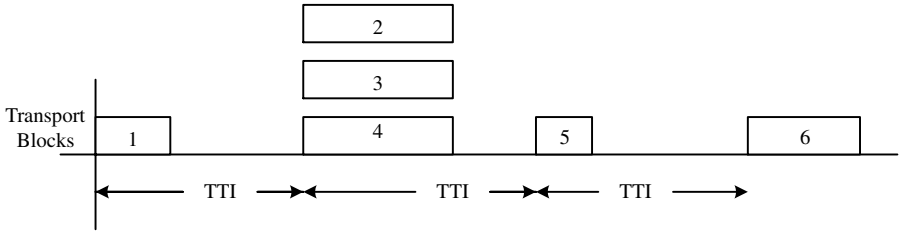


Figure 6-6
Transport blocks and the TTI between successive blocks

control channel over the air interface so that the physical layer at the receiving end knows how to demultiplex the transport channels.

The transport channels are summarized in Table 6-2. Broadly speaking, there are two types of transport channels: *dedicated transport channels* (DCH) that are unique to each user and *common transport channels* that are shared by all users in a cell or a UTRAN. The dedicated channels employ inherent addressing of UE. Common channels use explicit addressing of UE when such addressing is required.

DCH DCHs are defined on both uplinks and downlinks, are available in both FDD and TDD modes, and carry higher-layer data and control information such as voice, video, handoff orders, and signal measurements destined to any desired user.

Common Transport Channels

- *Broadcast Channel* (BCH) This is a downlink transport channel that carries UTRAN system and cell-specific parameters

Table 6-2

Summary of transport channels

Transport Channel	Direction	Function
DCH	Bidirectional	Carries user data and control.
BCH	Downlink	Broadcasts system and cell-specific parameters.
PCH	Downlink	Transmits paging messages.
FACH	Downlink	Carries control information and short data packets.
RACH	Uplink	Transmits signaling messages and short data packets. Transmission is subject to collision.
CPCH	Uplink	Carries user data packets. May experience collision.
DSCH	Downlink	Carries user data and control.

TEAM-Y

such as random access codes, access slots in a cell, or diversity types used. The associated physical channel broadcasts these information types over an entire cell. Each UE must decode this channel before it can register with a cell. The channel usually operates at a low data rate and high transmit power level.

- *Paging Channel* (PCH) This downlink transport channel is designed to carry paging messages for mobile-terminated calls. The physical layer transmits the information over one or more cells.
- *Forward Access Channel* (FACH) This is a downlink transport channel and carries control information such as an acquisition indication from a base station after a UE has randomly selected an access slot on an available random access channel and transmitted a preamble on that slot. It may also be used to carry a limited amount of packet data. The information is transmitted over an entire cell or part of a cell if the network design permits it. There may be more than one FACH in a cell, each operating at a different data rate.
- *Random Access Channel* (RACH) This uplink transport channel is available only in the FDD mode and carries signaling messages (such as a mobile-originated call request) and some limited amount of user data. The associated physical layer sends the information content of this channel by randomly selecting an available uplink access slot. Because this channel is accessed by all users in a cell, it is generally designed to operate at a low bit rate.
- *Common Packet Channel* (CPCH) The purpose of this uplink transport channel is to carry packet data from a user. It operates in much the same way as the random access channel. For example, a UE can begin to transmit during any predefined interval with respect to a received BCH frame. A downlink dedicated channel provides power control and control commands for this channel. It is only available in the FDD mode.
- *Downlink Shared Channel* (DSCH) This downlink transport channel is associated with one or more downlink dedicated

channels and carries user data and control information. It is thus shared by several UEs. The physical layer may transmit the information over an entire cell or part of a cell.

- *Uplink Shared Channel* (USCH) These channels, which are only used in the TDD mode, are shared among multiple users and carry both control and user data.
- *Fast Uplink Signaling Channel* (FAUSCH) These channels, which are used in conjunction with an FACH, carry signaling information when allocating dedicated channels.

Physical Channels

At the physical layer, each transport channel is mapped to a physical channel. Data is sent over the air interface in consecutive frames. Each frame is 10 ms long and consists of 15 slots, each of duration 0.66667 ms. Because the chip rate is 3.84 Mc/s, the length of each slot is 2560 chips. The frame structure is shown in Figure 6-7.

The number of bits that can be sent over a slot depends on the spreading factor. For example, if the spreading factor is 4, the number of bits per slot is 2560/4 = 640. If a given user is assigned only one slot per frame, the user data rate is 640/(10 ms) = 64 kb/s. If the spreading factor is 256, the user data rate on this channel is only 1 kb/s. If the user data rate is 4 kb/s, a spreading factor of 64 is to be used. Thus, a mobile station can be assigned virtually any data rate by using different spreading factors and allocating the required number of slots and channels.

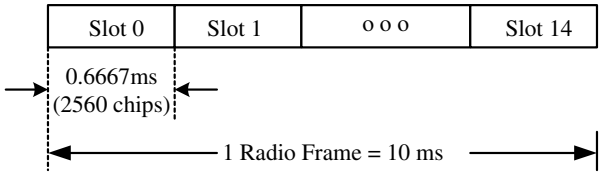


Figure 6-7
An uplink frame structure for a DPDCH or DPCCH

Physical channels are different in different directions—the uplinks and downlinks, as listed in the following section.

Uplink Physical Channels *Uplink Dedicated Physical Channels* (Uplink DPCHs) fall into two categories—dedicated physical data channels and dedicated physical control channels:

- *Uplink DPDCH* It carries the user information originating at the higher layers. At any given time, a radio link may not have any uplink DPDCH, or may have one or more of these channels, each using a different bit rate and hence a different spreading factor. For example, a DPDCH may have a data rate of 15, 30, 60, 120, 240, 480, or 960 kb/s, using respectively, a spreading factor of 256, 128, 64, 32, 16, 8, or 4. The data rate, together with the associated spreading factor to be used at a mobile station, is configured by the higher layers.
- *Uplink DPCCH* This uplink physical control channel transmits control information at a rate of only 10 bits per slot with a spreading factor of 256. Thus, the data rate on this channel is 15 kb/s. Its per-slot data structure is presented in Figure 6-8. It consists of four fields:
  - Pilot bits
  - The *transport format combination indicator* (TFCI)
  - The *transmit power control* command (TPC)
  - A feedback information field

  There are 3 to 8 pilot bits that provide the receiver with a reference carrier for coherent demodulation. The TFCI field,

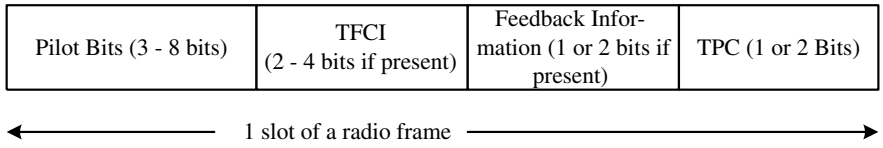


Figure 6-8 The per-slot data structure of DPCCH

which was previously explained, indicates the format of a transport block set that is being transmitted over a DPDCH at the same time as this DPCCH. This field may sometimes be absent. If present, it may be 2, 3, or 4 bits long. The feedback information bits provide information from UE to the UTRAN on such things as the closed loop mode transmit diversity and site selection diversity transmission [9]. This field also may at times be absent. When present, it may have one or two bits. The TPC field is always present and may be one or two bits long—either all 0s or all 1s. Unlike a DPDCH, there is only one uplink DPCCH on each radio link. Clearly, a number of slot formats are possible depending upon the values of these four fields. Which of these formats is being used is indicated by the higher layers.

- *Physical Random Access Channel* (PRACH) Mobile stations use this channel to send signaling information such as a call origination request to the UTRAN and, if necessary, a small amount of user data, such as short messages, alphanumeric texts, and so on. The access mechanism is based on the slotted Aloha scheme [44]. There are 15 access slots in every two 10 ms frames. Thus, the duration of each slot is 5,120 chips. Users wait until the beginning of a slot and examine if it is idle. If it is, they start transmitting. Clearly, in this case, there is some nonzero probability of collision; however, it is less than if the users had started transmitting at any arbitrary point in a slot.

  The information structure transmitted on PRACH is depicted in Figure 6-9. The transmission begins with one or more preambles at slot boundaries and ends with a message part. Each preamble is 4,096 chips long and consists of a 16-chip signature, repeated 256 times. The message part on this channel may be transmitted in a single 10 ms frame or in two 10 ms consecutive frames.

  For a 10 ms radio frame, the length of each slot is 2,560 chips. Each message part itself consists of two parts—a data part and a control part. The spreading factors permitted on the data part are 256, 128, 64, and 32. Thus, the number of bits transmitted per slot in the data part is 10, 20, 40, or 80 bits depending upon the spreading factor.

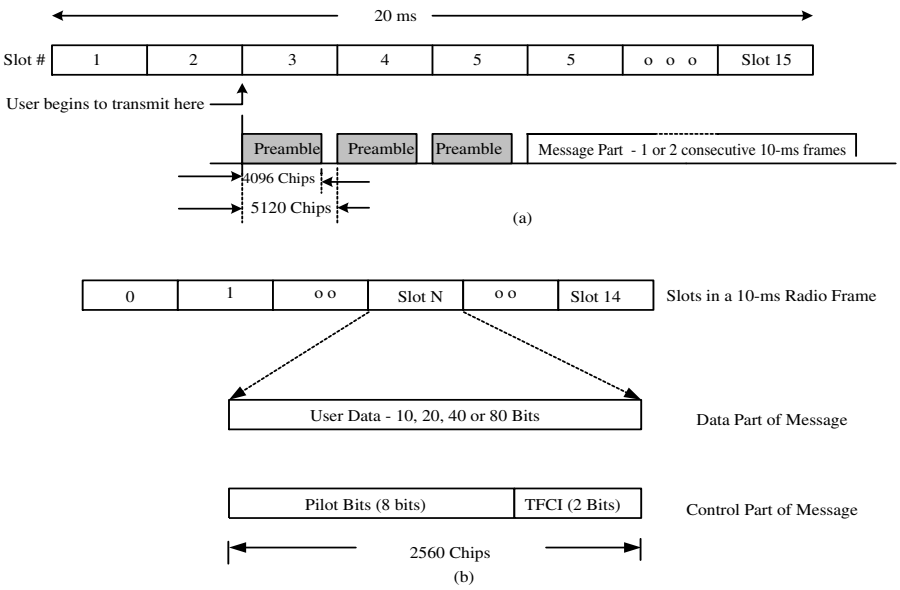


Figure 6-9
The operation of a PRACH

The control part of the message, on the other hand, uses a spreading factor of 256.[6] Thus, each slot in the control part carries only 10 bits, of which 8 are pilot bits that enable coherent detection at the receiver, and 2 are TFCI bits. The data part and the control part are transmitted simultaneously.[7]

- *Physical Common Packet Channel* (PCPCH) This is a multiple-access channel that carries the information on the CPCH transport channel. The access scheme is based upon *digital sense multiple access with collision detection* (DSMA-CD), whereby users can start transmission at the beginning of any of a number of time slots of a radio frame. The transmission structure is

[6]The spreading factor, which is actually the spreading gain, determines the capacity of a CDMA system, that is, the number of simultaneous users for a given *signal-to-interference ratio* (SIR). The higher the spreading factor, the larger the capacity.

[7]The data and control parts of the message are transmitted simultaneously using two different PN codes.

shown in Figure 6-10. All preambles as well as the message part must start at slot boundaries. The access and collision detection preambles, which are each 4,096 chips long, are similar to the preambles used on a PRACH that was just described. There may be one or more access preambles but only one collision detection preamble. The power control preamble may at times be omitted, but if present, as indicated by the higher layers, it is 8 slots long. The message part may consist of one or more 10 ms frames. As in PRACH, the message has two parts—the higher-layer user data and the physical-layer control information. The data part uses the same spreading factors as a DPDCH—4, 8, 16, 32, 64, 128, and 256. As in DPCCH, the control part operates at 15 kb/s with a spreading factor of 256. Both parts of the message are transmitted simultaneously using different codes.

Downlink Physical Channels The following are the physical channels on a downlink:

- *Downlink Dedicated Physical Channel* (Downlink DPCH) This channel time-division multiplexes the user data from the higher layers and the control information that is generated at the physical layer. The data portion forms a downlink DPDCH and the control portion a downlink DPCCH. The data format of a downlink DPCH slot is shown in Figure 6-11.

  The DPCCH consists of three fields: TPC, TFCI, and pilot bits. The TPC field is always present and is 2, 4, 8, or 16 bits long. The TFCI field is not necessary for an application that has a fixed data rate for the duration of a call. If multiple applications are

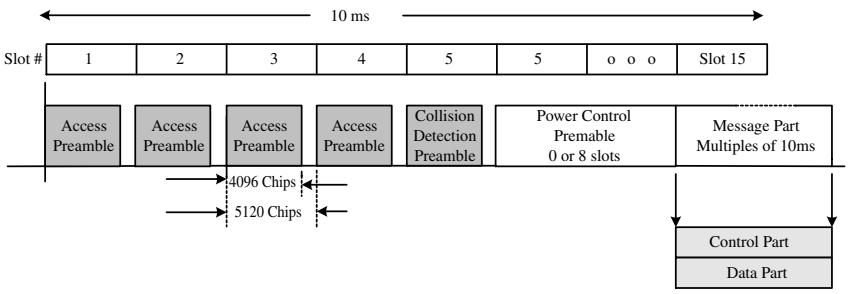


Figure 6-10 The transmission structure on the PCPCH

active simultaneously, the TFCI field may have 2, 4, 8, or 16 bits. The pilot bits are always included, and there may be 2, 4, 8, 16, or 32 bits of this field. The spreading factor used depends on the desired data rate and varies from 4 to 512.

- *Common Pilot Channel* (CPICH) This channel transmits an unmodulated carrier that can be used by the receiver to estimate the channel parameters. As in other cases, each radio frame is 10 ms long and consists of 15 time slots, each transmitting a predefined sequence of 10 symbols (with 2 bits per symbol) with a spreading factor of 256. Thus, the data rate is fixed at 30 kb/s. If transmit diversity is used on any downlink channel, both antennas are required to transmit the common pilot channel.[8] However, the symbol patterns on the two antennas are slightly different.

  There are two types of pilot channels—the primary and secondary. A primary pilot channel employs a fixed channelization code and a primary scrambling code. Each cell is assigned only one primary common pilot channel, which is transmitted over the entire cell. The secondary channel, on the other hand, may use any channelization code of length 256 and a primary or secondary scrambling code and may be used over an entire cell or part of a cell.

- *Primary Common Control Physical Channel* (PCCPCH) This channel, which maps the broadcast channel, is transmitted over

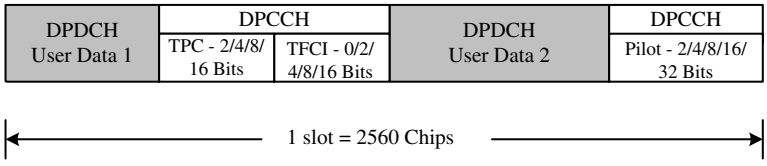


Figure 6-11 The data structure transmitted over a slot in a downlink dedicated physical channel

[8]Transmit diversity on a downlink channel can be implemented by arranging the incoming data into two streams and spreading them with different orthogonal codes. In a multicarrier system, such as cdma2000, transmit diversity is obtained by transmitting the streams over different antennas.

an entire cell. Its per-slot data structure is shown in Figure 6-12. Each slot is 2,560 chips long. The spreading factor used on this channel is 128, and a total of only 20 bits is transmitted per slot. However, the transmitter is turned off for the first 256 chips so that the primary and secondary synchronization channels can be transmitted during that period. Eighteen bits of data are then transmitted during the remaining 2,304 chips. Because there are 15 slots in a 10 ms frame, the effective rate on this channel is 27 kb/s. The broadcast channel, which is mapped by this physical channel, uses a fixed, predetermined transport format combination.

- *Secondary Common Control Physical Channel* (SCCPCH) This physical channel transmits the information contents of two transport channels—the FACH and the PCH. Unlike the primary common control physical channel, the secondary common control physical channel may be transmitted in a narrow lobe and may use any transport format combination as indicated by the TFCI field. The two transport channels may be mapped either to the same SCCPCH or to two different SCCPCHs.
- *Synchronization Channel* (SCH) This channel is used by mobile stations for cell search. There are two synchronization channels —the primary and the secondary. The primary synchronization channel transmits a modulated code, called the *primary synchronization code*, with a length of 256 chips during the first 256-chip period of each slot of a 10-ms, 15-slot radio frame (refer to Figure 6-12). The PCCPCH is transmitted during the remaining period of each slot. Every cell in a UTRAN uses the same primary synchronization code.

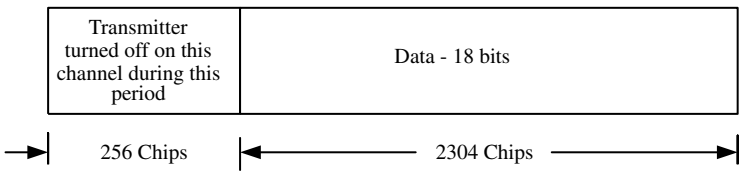


Figure 6-12 The per-slot data structure for the PCCPCH

The secondary SCH is constructed by repeating a sequence of modulated codes of 256 chips and is transmitted in parallel with the primary SCH, that is, on a different physical channel at the same time. There are 64 scrambling code groups for the secondary SCH.

- *Acquisition Indicator Channel* (AICH) This downlink channel indicates whether a UE has been able to acquire a PRACH. It operates at a fixed rate with a spreading factor of 128, using a 20 ms frame containing 15 slots, each with a length of 5,120 chips. Each access indicator is 32 bits long and is transmitted during the first 4,096 chips of each slot. Transmission is turned off during the last 1,024 chips so that another channel, such as the common packet *channel status indicator channel* (CSICH), can be transmitted during this period. See Figure 6-13.
- *Paging Indicator Channel* (PICH) This channel is associated with the secondary common control physical channel, uses a spreading factor of 256, and carries 288 bits of paging indication over each 10 ms radio frame. Transmission is turned off during the rest of the frame.[9]

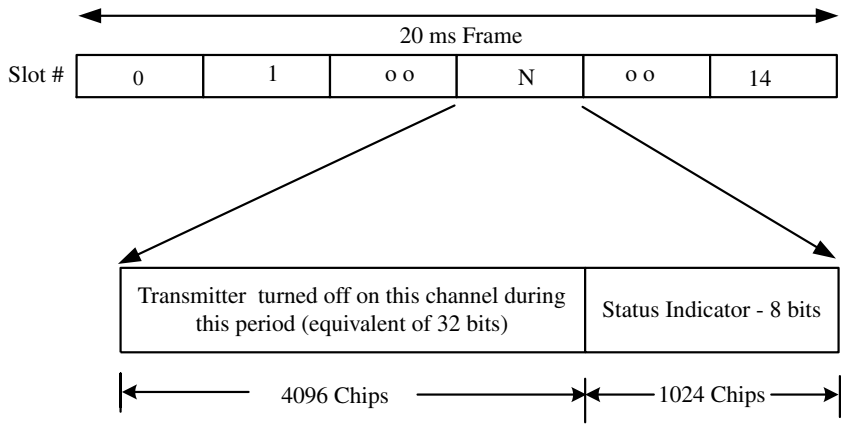


Figure 6-13 The data structure of the CSICH

[9]A 10 ms radio frame with a spreading factor of 256 can carry 300 bits of data.

- *Common Packet Channel (CPCH) Status Indicator Channel* (CSICH) As the name implies, this channel carries the CPCH status information. More specifically, the UTRAN uses it to notify the user which slots are available, indicating the data rates supported on those channels. It operates at a fixed rate with a spreading factor of 128. Its data structure is shown in Figure 6-13. This channel is deactivated during the first 4,096 chips so that another channel, such as the *acquisition indicator channel* (AICH), the CPCH *Access Preamble Acquisition Indicator Channel* (AP-AICH), or the *collision detection/channel assignment indicator channel* (CD/CA-ICH) can be activated during the same period.
- *Physical Downlink Shared Channel* (PDSCH) This channel, which maps a DSCH transport channel, is always associated with one or more downlink DPCH (that is, downlink dedicated physical channels). It consists of 10 ms frames, each containing 15 slots. The spreading factors used range from 2 to 128.

Packet Mode Data

It is clear from the previous description that packet mode data from the user plane may be transmitted over a number of channels. If the packets are short and infrequent, they may be transmitted over a RACH, CPCH, or FACH rather than a dedicated channel where the associated overhead may be unacceptably high. The RACH and CPCH are multiple-access channels and use the slotted Aloha scheme. If packets are long and relatively more frequent, a dedicated channel is established. In this case, after transmitting all packets that have arrived at the input, the channel may be either released immediately or held only for a short period thereafter. If there are any new packets during this period, they are transmitted; otherwise, the channel is released at the end of that period.

Mapping of Transport Channels to Physical Channels

As we indicated in the last section, the physical layer, on receiving the data over a transport channel, transmits it over a radio frame using a particular physical channel. In other words, transport channels are mapped to specific physical channels. This mapping is summarized in Figure 6-14.

Physical Layer Procedures

The standards documents specify procedures for synchronization, power control, accessing common channels, transmit diversity, and the creation of idle periods in the downlink. In this section, we will present a brief description of some of these procedures.

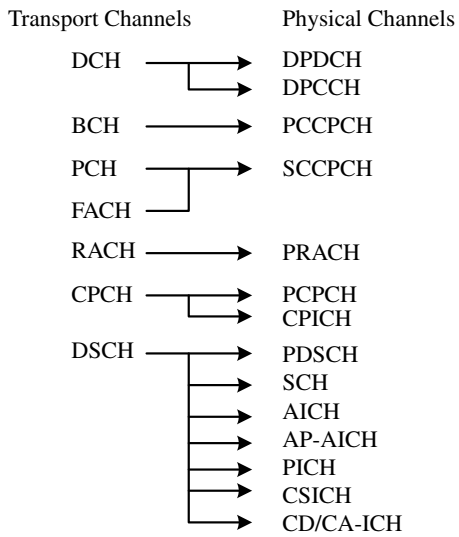


Figure 6-14 Mapping of transport channels into physical channels

Synchronization Procedures Synchronization procedures include the cell search mechanism and synchronization on the dedicated channels—the common physical channels as well as the dedicated physical control and data channels.

Cell Search Procedure By cell search, we mean searching for a cell, identifying the downlink scrambling code, achieving the frame synchronization, and finding the exact primary scrambling code used in the desired cell. The procedure is outlined in the following steps:

1. Because the primary synchronization code is the same for all cells in a system and is transmitted in every slot of the primary synchronization subchannel, the slot boundaries can be determined by passing the received signal through a filter that is matched to the primary synchronization code and observing the peaks at its output.
2. Notice that it is not possible to identify the frame boundary in step 1.[10] To do this, the received signal is correlated with each of the 64 secondary codes, and the output of the correlator is compared during each slot. The code for which the output is maximum is the desired secondary synchronization code. Similarly, the sequence of 15 consecutive slots over which the correlator output is maximum provides the frame synchronization.
3. The last step is concerned with the determination of the primary scrambling code. Because the common pilot channel is scrambled with the primary scrambling code, the latter can be determined by correlating the received signal over this channel with all codes within the code group determined in step 2. After having found the scrambling code, it is now possible to detect the primary common control physical channel that maps the broadcast channel.

[10]At this point, only slot boundaries have been found, but we do not know yet which slot belongs to which frame.

Synchronization on the Physical Channels Once frame synchronization has been achieved during the course of the cell search procedure, the radio frame timing of all common physical channels is known. Thereafter, layer 1 periodically monitors the radio frames and reports the synchronization status to the higher layers.

The status is reported to the higher layers using the following rules:

1. During the first 160 ms following the establishment of a downlink dedicated channel, the signal quality of the DPCCH is measured over the last 40 ms. If this measured signal is better than a specific threshold Q_{in}, the channel is reported to be in sync. At the end of this 160 ms window, go to step 2.
2. Measure the signal quality of the DPCCH over a 160-ms period. Also check transport blocks with attached CRCs. If the signal is less than a threshold Q_{out}, *or* if the last 20 transport blocks as well as all transport blocks received in the previous 160 ms have incorrect CRCs, declare the channel as out of sync. If, on the other hand, the quality exceeds Q_{in}, *and* at least one transport block received in the current frame has a correct CRC, the channel is taken to be in sync. Similarly, if the signal exceeds Q_{in} *but* no transport blocks or no transport blocks with a CRC are received, the status is taken to be in sync.

Setting Up a Radio Link When setting up a radio link, there are two cases to consider depending on whether or not there already exists a radio link for the UE:

- To establish a radio link when there are none initially, the UTRAN starts transmitting on a downlink DPCCH. If there is any user data to send, it may also start transmitting that data on a downlink DPCCH.[11]

  The UE monitors the downlink DPCCH and first establishes frame synchronization using a PCCPCH. Thereupon, it can begin

[11]Recall that the downlink DPCCH and the downlink DPDCH are time-division multiplexed.

to transmit on the uplink DPCCH either immediately or, if necessary, after a delay of a specified activation time following the successful establishment of the downlink channel. Transmission on the uplink DPDCH can start only after the end of the power control preamble.

The base transceiver station monitors the uplink DPCCH and establishes chip and frame synchronization on that channel. Once the higher layers in the UTRAN have determined that the link is in sync, the radio link is considered established.

- To set up a radio link when there are other radio links already established, the UTRAN begins to transmit on a new downlink DPCCH and, if necessary, on a new downlink DPDCH with appropriate frame timing.

  The UE monitors the new downlink DPCCH, establishes frame synchronization on this channel, begins to transmit on an uplink DPCCH, and, if necessary, on an uplink DPDCH as well.

  The base transceiver station monitors the uplink DPCCH and establishes chip and frame synchronization on that channel. Once the higher layers in the UTRAN have determined that the link is in sync, the new radio link is considered established.

  It is possible that the receive timing of a downlink DPCH may drift significantly over time so that the time difference between downlink and uplink frames exceeds the permissible value. When this is the case, the physical layer reports the event to higher layers so that the network can be requested to adjust its timing.

Power Control As we mentioned, power control is an important feature of a CDMA system. Its objective is to ensure a satisfactory signal-to-interference ratio at the receiver for all links in the system. In UMTS, different power control procedures are used for uplink and downlink physical channels. Because our goal is to acquaint the reader with the general concept of the power control in UMTS, we will briefly describe only some of these procedures [12]. First, however, definitions of a few terms are in order.

Open Loop Power Control This is a process by which the UE sets its transmitter power output to any specific level. The open loop power control tolerance is ±9 dB under normal conditions and ±12 dB under extreme conditions.

Inner Loop Power Control in the Downlink This procedure enables a base station to adjust its transmit power in response to TPC commands from the UE. Power is adjusted using a step size of 0.5 or 1 dB. The objective here is to maintain a satisfactory signal-to-interference ratio at a UE using as little base station transmitter signal power as possible.

Inner Loop Power Control in the Uplink This procedure is used by the UE to adjust its transmit power in response to a TPC command from a base station. With each TPC command, the UE transmit power is adjusted in steps of 1, 2, or 3 dB in the slot immediately following the decoding of TPC commands.

A TPC command may be either 0 or 1. If it is 0, it means that the transmitter power has to be decreased. If it is 1, the transmitter power is to be increased.

Uplink Inner Loop Power Control Procedure on Dedicated Physical Channels The dedicated physical channels use the uplink inner loop power control. Briefly, the procedure is as follows. The UE starts transmitting on the uplink DPCCH at a power level that is initially set by the higher layers. Serving cells measure the received SIR and compare it with a target threshold. If the measured SIR exceeds the threshold, the UTRAN sends a TPC command 0, indicating that the mobile station should decrease its power level using a step size of 1 or 2 dB as specified by the higher layers. If the measured SIR is less than the threshold, TPC command 1 is transmitted, requiring the mobile to increase its power level. If both data and control channels are active at the same time, the power level of both uplink channels is changed simultaneously by the same amount. For a DPCCH, this change should be affected at the beginning of the uplink DPCCH pilot field immediately following the TPC command on the downlink

channel. This is shown in Figure 6-15. Notice the timing offset between the downlink DPCH and the uplink DPCCH. It is also worth mentioning that the TPC command on the uplink starts 512 chips after the end of the pilot field on the downlink channel.

When a mobile station is being served by a single cell and is not in a soft handoff state, it receives only one TPC command in each slot. Because there are 15 slots in a radio frame and each frame is 10 ms long, it may receive 1,500 TPC commands per second.

However, if the mobile is in a soft handoff state, more than one TPC command may be received in each slot of a radio frame from cells in an active set that participate in the handoff process. The physical layer parses these commands, and if it finds all TPC commands to be 1, it increases the transmitter power by the selected step size. Similarly, if all commands are 0, the power is decreased by the same amount. Otherwise, if the commands are all random and uncorrelated, they are interpreted based on a probabilistic model [12]. The same procedure is used to adjust the power level during the uplink DPCCH power control preamble.[12]

The procedure that we have just described adjusts the power level in accordance with the TPC commands received during each slot,

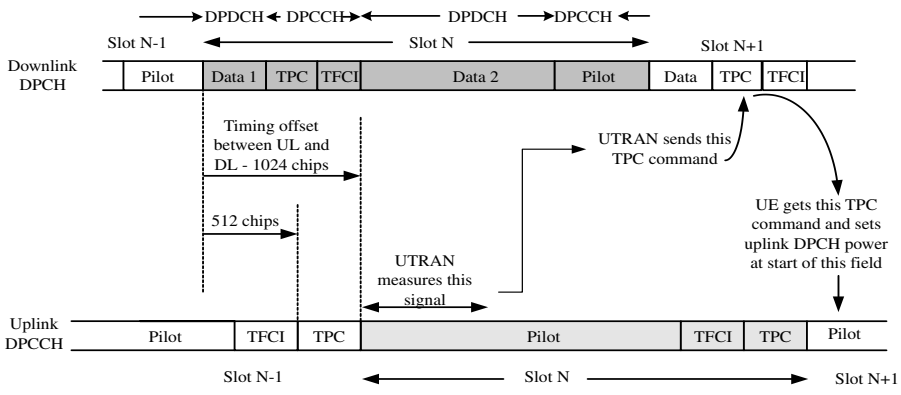


Figure 6-15
The sequence of events and their timing during the uplink power control

[12]The transmission on a DPDCH starts only after the end of this preamble.

using a step size of 1 or 2 dB. This is referred to in the standards document as *Algorithm 1*. Using a slight variation of this algorithm, we can emulate a smaller step size and thus effect a finer adjustment. This is called *Algorithm 2*, which is briefly described here. Assume that the mobile is being served by a single cell and is not going through any handoff process. For each set of five slots aligned to the frame boundaries, no action is taken on those commands that were received in the first four slots. During the fifth slot, the receiver determines if the TPC commands in all of these five slots are the same. If they are, the power level is increased or decreased by the previous step size, depending on whether they are all 1 or 0. Otherwise, the commands are ignored. Because the power level is now being changed by the same amount every five slots, the net result is the equivalent of a smaller step size.

The procedure to emulate a smaller step size when the mobile is undergoing a handoff process is similar.

Downlink Inner Loop Power Control on DPCCH and DPDCH The operation of the downlink inner loop power control is quite similar. Assuming that the mobile is being served by a single cell and is not going through a handoff process, the UE measures the SIR on the downlink physical channels and compares it with a desired target value. If the measured SIR is less, the UE sets the TPC command to 1 in the next available TPC field of the uplink DPCCH. The UTRAN responds by increasing the power of the downlink DPCH at the beginning of the next pilot field on that channel following the TPC command on the uplink.

If the measured SIR is more than the desired value, a TPC command 0 is sent in the next available TPC field of the uplink DPCCH, thus requesting a reduced power level. In response, the UTRAN decreases the power level of the downlink DPCH at the beginning of the next pilot field on that channel following the TPC command on the uplink. The downlink power control timing is shown in Figure 6-16.

Depending on the downlink power control mode, the UE may send either a unique TPC command in each slot or the same command over three slots while making sure that a new command appears at the beginning of a radio frame. On receiving a TPC command, the

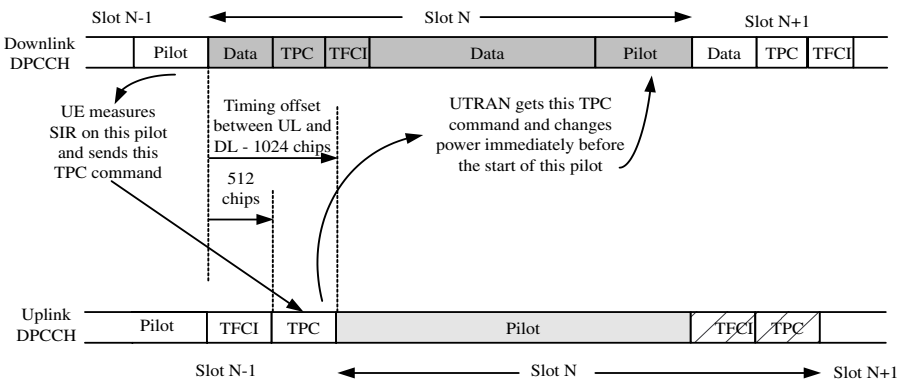


Figure 6-16
The sequence of events and their timing during the downlink power control

UTRAN estimates the necessary change in the transmit power as required by the command, but modifies it to some extent before actually making the adjustment of the transmitter power. The purpose of this modification is to balance the radio link powers so as to maintain a common reference level in the UTRAN. Reference [12] describes how the inner loop power control is usually estimated, and also gives an example of a power-balancing procedure.

Uplink Inner Loop Power Control on PCPCH The uplink inner loop power control procedure for the message part of the PCPCH is very similar to the inner loop power control for the dedicated physical channels.[13] A PCPCH message has two parts—the data and the control—which are usually associated with different power levels that depend upon their gain factors. The uplink PCPCH inner loop power control adjusts the powers of the two parts simultaneously and by the same amount. Thus, assuming that their gain factors remain unchanged, the power difference or the power offset, as it is called, between the data and control parts remains the same after the transmit power has been adjusted in accordance with TPC commands.

[13]Notice that here we are not talking about the power control during the CPCH access procedure.

The UTRAN measures the SIR on the received PCPCH and compares it with the desired SIR objective. If the measured SIR is less, the network sends a TPC command 1, requesting that the power be increased. If it is more, the network sends a TPC command 0, indicating that the power should be decreased. The UE may process the TPC commands and adjust the uplink transmit power in steps of 1 or 2 dB using either Algorithm 1 or Algorithm 2 that was described earlier.

Random Access Procedure Random access procedures are used to transmit data (that is, the signaling information and/or user data) on the two uplink physical channels: the PRACH and the PCPCH. The procedures are initiated when the physical layer receives a service request from the MAC layer. These procedures, which are described in great detail in the standards documents, are similar for the two physical channels. We will illustrate them by providing a brief description of the access mechanism on the PRACH only.

Random Physical Access Channel There are 12 RACH subchannels, each containing five access slots. The UE may select any one of these slots on the available RACH subchannels within an access service class and commence transmission. The procedure uses a number of system parameters including, among others, preamble scrambling code, available RACH subchannels for each access service class, the maximum number of preamble retransmissions, and the initial preamble power. The UE receives these parameters from the radio resource control layer of the UTRA. A brief description of the procedure is presented in Figure 6-17.

Spreading and Modulation

In UMTS, the signal is spread in two steps. First, all physical channels with the exception of the downlink synchronization channels are spread by unique channelization codes so that they can be separated at the receiver. The spreading factor is defined as the number of chip periods into which each incoming symbol is spread. The

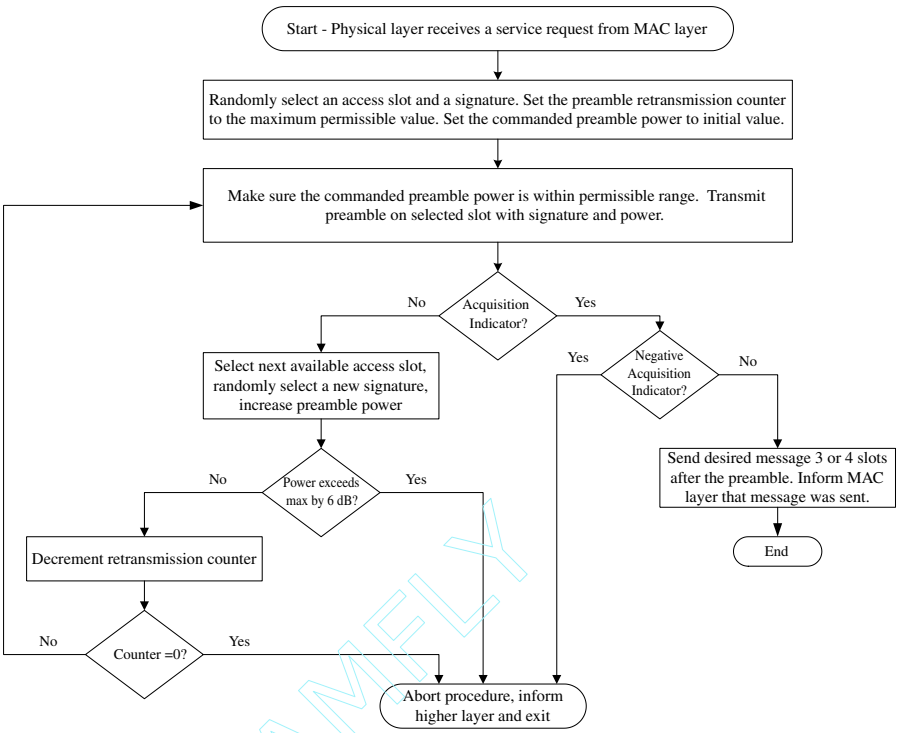


Figure 6-17 *Procedure to access the random access physical channel*

TEAMFLY

channelization codes are mutually orthogonal and may spread each physical channel by a variable spreading factor. As such, the codes are known as *Orthogonal Variable Spreading Factors* (OVSF). In the second step, the physical channels thus spread are summed together and scrambled by unique, complex-valued scrambling codes so that the sources of the physical channels (such as different mobile stations in a cell or various sectors of a cell) can be unambiguously identified at the receiver.

The general principles of spreading and modulation were presented in Chapter 3. For UMTS, the uplink and downlink channels are treated in a slightly different way.

Uplink Channels The spreading and modulation technique for uplink channels is shown in Figure 6-18. The incoming binary data on each physical channel is converted into symbols, a binary 0 being

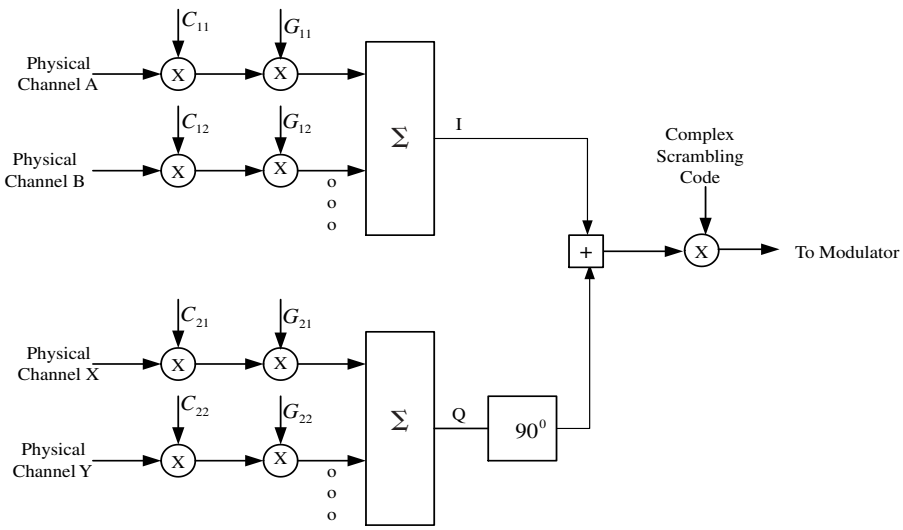


Figure 6-18
Spreading multiple physical channels from a mobile station

represented by +1 and binary 1 as −1.[14] Now assume that a number of these channels have to be transmitted using a single CDMA carrier. As an example, a mobile station may have a number of uplink DPDCHs as well as a DPCCH. In this case, the channels are divided into two sets—say, channels A, B, and so on in one set, and channels X, Y, and so on, along with the DPCCH, in another. The physical channels are split this way so that one set of channels modulates an in-phase (that is, I) carrier and the other set a quadrature (that is, Q) carrier.[15]

Continuing with Figure 6-18, each of the physical channels is spread by a unique OVSF code. The spreading factor is 256 for a control channel and varies from 4 to 256 for a data channel. Because different channels are usually transmitted at different relative power levels, the spread symbols are multiplied with appropriate gain factors and summed together. The gain factors vary from 0 to 1 in steps of 1/15. Because the resulting real-valued symbol sequences, indicated

[14]In other words, we are going to use binary phase shift keying.

[15]Each set modulates the carrier using BPSK. The net result, of course, is QPSK.

as I and Q in the figure, are to be scrambled by a complex scrambling code, the I and Q sequences are transformed into a complex sequence by first advancing the phase of Q by 90 degrees and then adding it to I.[16] The output of the scrambler is separated into real and imaginary parts, and applied to the modulator, as shown in Figure 6-5.

Because the symbols from the second channel set are shifted by 90 degrees before adding them to the symbols of the first set, the effective modulation is QPSK with the constellation of Figure 6-19.

Downlink Channels The spreading of downlink channels is slightly different, as shown in Figure 6-20. Incoming symbols on all downlink channels except AICH may be +1, -1 or 0. Symbol 0 corresponds to the situation when the transmission is to be discontinued. The incoming data on each physical channel, with the exception of a synchronization channel, is converted into parallel form and separated into two streams, one with the odd bits and the other with the even bits. Each of these streams is spread by a channel-specific, orthogonal spreading code shown as C_1 in this figure. The spreading factor is 256 for common downlink physical channels and varies from 4 to 512 for a downlink DPCH.

The I and Q channels are added in quadrature, scrambled by the cell-specific downlink scrambling code S_m, and multiplied by the gain factor G_1. The synchronization channel, on the other hand, is simply

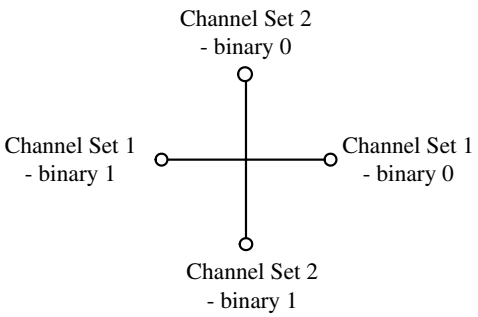


Figure 6-19 *The signal constellation in QPSK modulation used in UMTS*

[16]In other words, I and Q are added in quadrature.

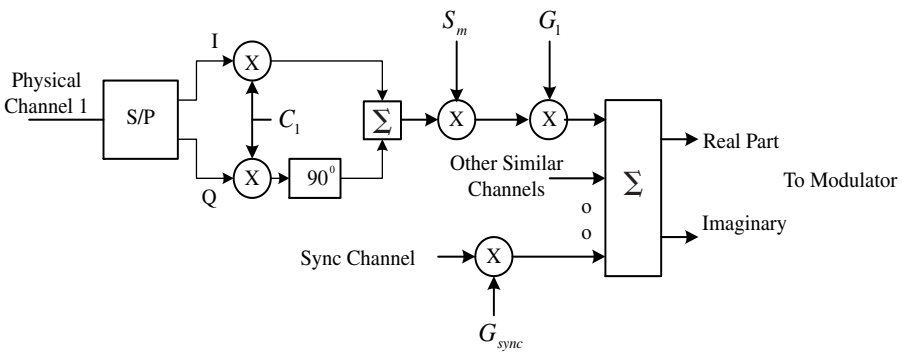


Figure 6-20
Spreading of downlink channels

multiplied by gain G_{sync}. The complex-valued outputs from all channels after the gain multiplication are added, separated into real and imaginary parts, and passed to the modulator.

Channelization Codes The channelization codes are mutually orthogonal and are obtained from an Nth order, orthogonal Walsh, or Hadamard matrix:

$$H_N = [h_{ij}], i,j = 0,1,2,\ldots,N-1$$

where each h_{ij} is either $+1$ or -1. This matrix is called orthogonal because the inner product of any two different rows is 0:

$$\sum_{k=0}^{N-1} h_{ik}h_{jk} = 0 \text{ for } i \neq j$$

Thus, the entries of each row of such a matrix can be used as a channelization code. Before modulating with the channelization code, however, the incoming binary data is first subjected to a level transformation whereby a 0 is converted to $+1$ and a binary 1 to -1.

Orthogonal Hadamard matrices can be generated recursively as discussed in Chapter 3 when their dimension N is given by $N = 2^n$ with n as an integer. They can also be represented in the form of a tree. For example, in Figure 6-21, the entries of matrices H_1, H_2, and H_4 are shown alongside the branches of a tree. Notice that matrix H_4 corresponds to codes associated with branches emanating from nodes C1 and C2, rows 1 and 3 representing codes at C1, and rows 2 and 4 at C2.

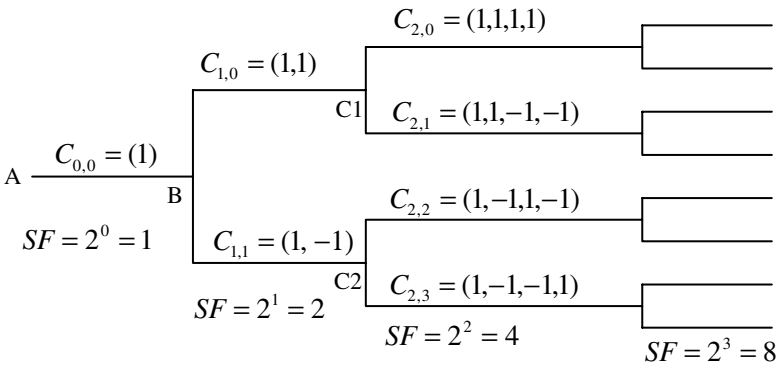


Figure 6-21 Orthogonal channelization codes arranged in the form of a tree

Scrambling Codes A scrambler maps an incoming data sequence into a different sequence such that if the input is periodic, the output is also periodic with a period that is usually many times the input period. Scramblers are built using a series of shift registers where certain outputs are added module 2 and then fed back to the input of the register array.

The theory of PN and scrambling codes was presented in Chapter 3. For the purpose of this chapter, it is sufficient to point out that the feedback path in a shift register array may be represented by a polynomial, say, *f(x)*.[17] It can then be shown that the period of the output sequence from the scrambler is the smallest integer p such that *f(x)* divides $x^p + 1$ using, of course, modulo 2 addition when performing the polynomial division. If there are m registers in the array, *f(x)* is of degree m, and the maximum possible period of the scrambler output is $2^m - 1$. In this case, we say that the shift register sequence has a maximum length. However, to achieve this length, it is necessary that *f(x)* be irreducible; that is, it should be divisible only by itself and by 1.[18]

[17] *f(x)* is also known as the characteristic polynomial.

[18] In the literature, this polynomial is sometimes referred to as a primitive polynomial over a Galois field $GF(2^m)$. To understand it, suppose that we want to construct the Galois field $GF(2^m)$ that has 2^m elements where m is an integer. If we use arithmetic

Let us illustrate these ideas by a simple example. Consider the scrambler of Figure 6-22. Because it has five shift registers, $m = 5$. The maximum possible period of this scrambler is $2^m - 1 = 31$. The feedback tap polynomial is

$$f(x) = 1 + x^2 + x^5$$

It can be shown that *f(x)* is a primitive polynomial, and the output of the scrambler has indeed the maximum length of 31.[19]

UMTS uses complex uplink scrambling codes. There are two types of these codes: long codes and short codes. The long codes are derived from two long sequences in the following manner. First, two shift register sequences are generated with primitive polynomials $f_1(x) = x^{25} + x^3 + 1$ and $f_2(x) = x^{25} + x^3 + x^2 + x + 1$ over Galois field $GF(2^m)$. The two sequences are added using modulo 2, yielding the first of the two long sequences. The second one is obtained by shifting the first by 16,777,232 chips. Long, complex scrambling codes are then formed using these two sequences as a basis. Short scrambling codes

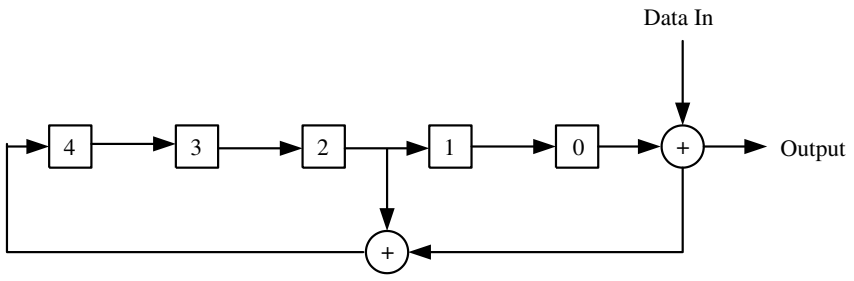


Figure 6-22
A scrambler with a 5-stage shift register

addition modulo 2, then 0 and 1 are two of the elements of that field. Now assume that $f(x)$ is a polynomial of degree m, and β is a root of the polynomial so that $f(\beta) = 0$. If we have selected $f(x)$ properly, we can construct all the other elements of the field by taking different powers of β such as $\beta, \beta^2, \beta^3, \ldots \beta_2^{m-2}$ and simplifying the arithmetic using the relation $f(\beta) = 0$. In this case, we say that β is a primitive element. If $f(x)$ is irreducible, we say that it is a primitive polynomial.

[19]Reference [45] shows that if $2^m - 1$ is a prime number, every irreducible polynomial of degree m generates a maximal length sequence.

are generated from three sequences, each using an array of 8 registers and a feedback polynomial of degree 8. For details, see Chapter 3 and Reference [11].

Uplink scrambling codes may be either long or short. The long codes have a length of 38,400 chips (that is, 10 ms), whereas short codes are only 256 chips long. The use of short codes on an uplink channel requires advanced multiuser detection techniques at the base station.

Downlink scrambling codes are also complex valued and, like the long uplink scrambling codes, are generated using two constituent sequences, which are derived from two shift register arrays with primitive polynomials: $f_1(x) = x^{18} + x^7 + 1$ and $f_2(x) = x^{18} + x^{10} + x^7 + x^5 + 1$ over $GF(2^m)$. There are a total of $2^{18} - 1$ of these codes. However, only 8,192 are used on downlinks. They are divided into 512 groups, each containing one primary scrambling code and 15 secondary scrambling codes. Each code is of length 38,400 chips.

Physical Layer Measurements

From time to time, the user equipment and UTRAN are required to perform signal measurements and, if necessary, report the results to higher layers. These measurements are required for a number of reasons. For example, the UTRAN may use them to determine if it is necessary to handover a mobile to another base station using the same carrier, to another base station using a different carrier, to another system (for example, a GSM network), or to another service provider, if necessary. Measurements are performed periodically, or on demand, or may be triggered by some events (for example, the current CCPCH is no longer the best one). They are evaluated and filtered at different layers before they are reported to the higher layers. These measurements are done during idle slots inserted in a radio frame for this purpose. The mechanism by which these inactive slots are built in a radio frame so that the UE can perform these measurements is called the *compressed mode*.

Measurements may be divided into a few types:

- Measurements on downlink physical channels. These measurements may involve a single W-CDMA frequency,

different W-CDMA frequencies (such as when a UE is near the boundary of two W-CDMA systems), or a W-CDMA frequency and the operating frequency of another system such as GSM.

- Traffic volume measurements on uplink channels
- UE transmit power and received signal level
- Measurements of *quality of service* (QoS) parameters such as block error rates and delay variations

Some of the parameters that are measured by the UE are listed in the following list. For a more detailed description of these parameters, see [14].

- Time difference between the *system frame number* (SFN) of the target neighboring cell and *connection frame number* (CFN) in the UE. It is given in terms of chips for the FDD mode and frame numbers for the TDD mode.
- Difference between the timing of a given UTRA and the timing of a GSM cell.
- E_c/N_0 for the *common pilot channel* (CPICH), where E_c is the received energy per chip and N_0 the noise power density.
- SIR for CPICH.
- Received signal code power on the PCCPCH after despreading.
- Interference with the received signal on the common pilot channel after despreading.
- Received *signal strength indicator* (RSSI), which is the wideband signal strength measured in the desired bandwidth.
- Transport channel block error rate.
- The UE transmitter power at the antenna connector.
- Time difference between an uplink frame and the first significant path of a downlink frame on a dedicated physical channel. This measurement is valid only for the FDD mode.
- Time difference in the system frame numbers between a specific cell and a target cell on CPICH and PCCPCH.

The following is a partial list of the parameters measured by a UTRAN:

- Received power over an uplink channel. For a diversity receiver, this is a linear average on the diversity branches.
- The transmitted power on a downlink carrier with respect to the maximum power possible on that carrier.
- Transport channel *block error rate* (BLER) and BER.
- The physical channel bit error rate.
- Receive timing deviation. Valid only for TDD mode, this is the propagation delay of an uplink signal.
- Received signal code power on a dedicated physical channel or a PRACH after despreading. Valid only for TDD mode.
- Round-trip delay between the start of a downlink frame and the beginning of a corresponding uplink frame.
- Frequency offsets between two nodes.
- SIR. As defined in the standards documents, it is actually the signal-to-interference ratio multiplied by the spreading factor. Measurement is done on the dedicated physical control channel in the FDD mode and dedicated physical channel in the TDD mode.
- Propagation delay in accessing the PRACH or a PCPCH.

MAC Layer Protocol

Overview

The MAC layer, as the name implies, determines how different types of information coming from the higher layers over different logical channels should be transmitted over a physical channel on a radio frame (that is, the medium), and controls the timing of those transmissions [7], [15]. It provides the following services to the upper layers: data transfer, reallocation of radio resources and redefinition of MAC parameters, and measurement of the traffic volume and signal quality, and reporting the results to the RRC layer.

The MAC layer interacts with the RLC sublayer over a number of logical channels. Data flows on each logical channel are associated with a certain priority based on the attributes of the radio bearer service and the RLC buffer status. For example, if a particular UE is running two applications simultaneously, say, voice and a file transfer in the background, the RRC sublayer may assign a different priority to each of the two applications. Similarly, multiple UEs may be assigned relative priorities as well. To meet these priority requirements, the MAC layer may use some scheduling algorithms and map, say, high-priority data to a high-bit-rate transport format and low-priority data to a low-bit-rate transport format. Thus, the responsibility of the MAC layer is to map each logical channel onto a transport channel, selecting, on the basis of the associated priorities, an appropriate transport format within a *Transport Format Combination* (TFC) set that is assigned by the RRC layer. When multiple users access a RACH, the MAC layer informs the physical layer of the RACH resources assigned to each user (such as access slots, channelization codes, back-off parameters, and so on). Other functions of the MAC sublayer include the following:

- Multiplexing higher layer *protocol data units* (PDUs) onto transport blocks and delivering them to the physical layer on a common transport channel or a dedicated channel
- Demultiplexing the transport blocks received from the physical layer into higher-layer PDUs and presenting them to the higher layer
- Measuring traffic volumes on a logical channel and reporting the information to the RRC layer so that it can control the admission of new users and provide required QoS
- Ciphering of data for transparent mode operation of the RRC layer

The MAC layer supports only unacknowledged data transfer without any segmentation or reassembly. It interfaces the RLC layer over logical channels and the physical layer over transport channels. To avoid repetition, it is sufficient to say that the various logical channels are similar in concept to the transport channels, which

were discussed earlier. The MAC layer maps them into transport channels as shown in Figure 6-23.

MAC Procedures

The MAC layer follows a set of procedures that enables different UEs to access a RACH or a CPCH. We will describe only one of them, namely the one that controls transmissions on a RACH.

Transmission Control on a RACH For the purpose of accessing this channel, the access slots and preamble signatures (or the time slots and channelization codes in a TDD system), which are referred to in the standards documents as RACH resources, are assigned a set of relative priorities. Based on these priorities, the resources can be divided into a number of Access Service Classes, each using a certain partition of the RACH slots with a given transmission probability (persistence value).

The medium access control procedure for a RACH is briefly the following:

1. The MAC layer receives from the RRC layer RACH transmission control parameters such as available access service classes, backoff time intervals after which a UE can resume

TEAMFLY

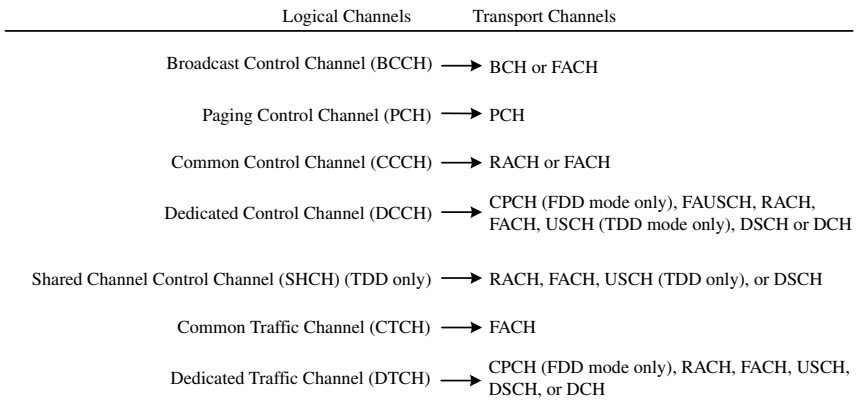


Figure 6-23 Mapping of logical channels into transport channels

transmission following a negative acknowledgment on the acquisition indication channel, and so on.

2. When the MAC layer receives from the higher layer a service data unit that has to be transmitted on the RACH, it selects an available access service class and initializes a preamble transmission counter to 0.
3. The preamble counter is incremented by 1. If its value is less than or equal to the maximum permissible value, go to step 4. Otherwise, terminate the access procedure and report error conditions to the higher layers.
4. The MAC layer selects a random number in the range of 0 to 1. If this number is greater than the transmission probability associated with the selected access service class, the MAC layer shall wait for the next transmission timing interval and repeat step 4. Otherwise, send an access request primitive to the physical layer and go to step 5.
5. If there is no acknowledgment from the physical layer, wait for the next transmission timing interval and go to step 3.

If the MAC layer receives a NACK (that is, a negative acknowledgment) on the *Acquisition Indication Channel* (AICH), wait until the next transmission timing interval. Then select a random backoff interval, set a backoff timer to this value, start the timer, wait until this backoff timer has expired, and go to step 3.

If, on the other hand, the MAC layer has received an ACK (that is, a positive acknowledgment), send the higher-layer data to the physical layer via a data request primitive.

Traffic Volume Measurement Before configuring a new radio bearer, the Radio Resource Control layer may require information on the traffic volume carried by each transport channel. The RRC indicates to the MAC layer, via a Measure-REQ primitive, which traffic parameters—buffer occupancy, its average value, and its variance on each radio bearer—should be reported, whether the report should be sent periodically, and, if so, at what interval, or whether the report should be generated only when certain criteria are met (for example,

the amount of data to be transmitted on a transport channel is outside a range set by the RRC). Similarly, the RLC may also notify the MAC layer about the amount of data queued in its buffer. If the mode is periodic, the MAC layer gathers the requested information during each transmission timing interval and sends it to RRC at prescribed intervals. Otherwise, it compares the amount of traffic on a transport channel with the thresholds specified by the RRC, and reports only the result of this comparison.

MAC Layer Data Formats

On receiving an upper-layer PDU, the MAC layer adds a header and passes the resulting PDU to the physical layer. The MAC PDUs, arriving at the physical layer during any transmission timing interval, are transmitted in the order in which they arrive. Similarly, the MAC PDUs corresponding to a given logical channel are multiplexed in the same order in which they originate at the higher layers.

The general format of the MAC PDU is presented in Figure 6-24. There are four fields in the header. The first field called the *Target Channel Type Field* (TCTF) indicates the logical channel and may be 1, 2, 3, 4, 5, or 8 bits long depending upon the associated transport channel and the mode (that is, FDD or TDD). For example, on a RACH in the FDD mode, this field is 00 for a common control channel and 01 for a dedicated control or traffic channel.

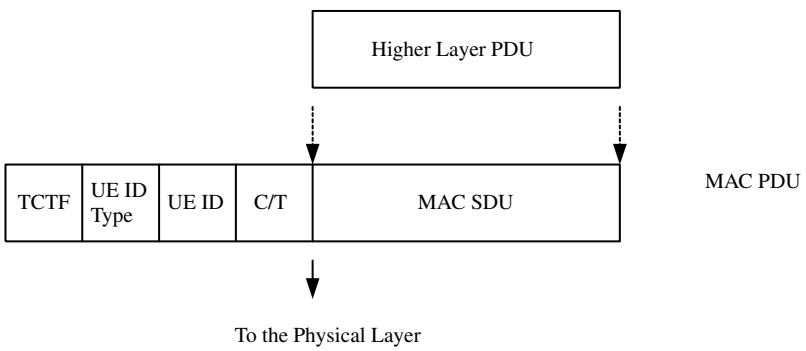


Figure 6-24 The MAC PDU format

On the other hand, on a RACH in the TDD mode, that field is 00 for a common control channel and 0100 for a dedicated control or traffic channel.

Because multiple users may be transmitting on a common transport channel, it is necessary to separate them using UE Identification numbers. There are two types of UE IDs—the UTRAN *Radio Network Temporary ID* (U-RNTI) and the *Cellular Radio Network Temporary ID* (C-RNTI). They are indicated by the 2-bit UE ID Type field. The UE IDs are 32 bits long for U-RNTI and 16 bits for C-RNTI.

The C/T field is four bits long and identifies the logical channel in a MAC PDU when multiple logical channels are multiplexed on the same transport channel.

Depending upon the logical and transport channels, some or all of the four fields of the MAC PDU header may be missing. For example, when a dedicated traffic or control channel is mapped to a dedicated channel, the entire header is omitted. On the other hand, when they are mapped to a RACH or a FACH, all four fields of the header are included.

Radio Link Control Protocol

RLC Functions

As shown in Figure 6-3, the RLC layer interfaces the RRC, PDCP, and BMC layers on one side and the MAC layer on the other. Its main functions are

- Data transfer
- Segmentation and reassembly
- Error detection and correction
- Flow control
- Ciphering for the purpose of providing security

If the PDUs coming from the higher layers to the RLC sublayer are too long, they are first segmented into smaller units of equal size and then transferred to the MAC layer. If the higher-layer PDU or a segment thereof is smaller than the given size, it may be concatenated with the first segment of the next incoming higher-layer PDU. Alternatively, it may be extended by adding padding bits. A header is added to each segment that includes, among other things, the sequence number and the length of the PDU, and the resulting packet is passed to the MAC layer.

There are three modes in which the data transfer can take place between peer RLC layers (that is, the transmit and receive RLC layers of a communication link)—acknowledged, unacknowledged, and transparent. In the acknowledged transfer mode, the RLC provides guaranteed, error-free delivery. On receiving a packet from the MAC layer, the RLC layer checks to see if its sequence number is correct. If it is not, the receiver discards the packet and attempts to recover from that error in one of the following ways. It may use an *automatic repeat request* (ARQ) whereby the sender is requested to repeat a transmission. Or, the receiving end may ask the transmitter to go back *N* frames and retransmit all packets starting from that frame. Lastly, it may request the transmitter to selectively retransmit the desired packets. In the acknowledged transfer mode, the receiving RLC entity reassembles the received segments and delivers a unique copy of each PDU to the higher layers in the correct sequence. Out-of-sequence packets may also be forwarded to the higher layers. In this case, the higher layers should have the capability of arranging them in the correct sequence. If the RLC entity determines that error recovery is not possible, it must notify its higher layers and its peer to that effect.

In the unacknowledged transfer mode, the data delivery is not guaranteed to the sending end. After receiving a packet, the RLC layer examines the sequence number. If it is erroneous, the packet is discarded. Otherwise, it is delivered to the higher layers. In neither case, however, does the receiving RLC send an acknowledgment to its peer. Consequently, in this mode, it is the responsibility of the

higher layers to recover the missing packets using an appropriate error recovery mechanism.

The RLC layer, on receiving a PDU from the higher layers, segments it and usually adds a header to each segment before passing it to the MAC layer. Similarly, the MAC layer, after receiving the RLC PDU, adds a MAC header, and the resulting packet, which is actually a transport block, is then passed to the physical layer. This is explained in Figure 6-25.

Another function performed by the RLC sublayer is the flow control. When the receiving RLC sublayer cannot service the incoming packets fast enough, it may want to control the rate at which the transmitter sends packets over the channel. When the data transfer takes place using the acknowledged mode, the flow control can be implemented simply by withholding acknowledgment at proper instants.

Ciphering, which prevents unauthorized reception of data, is implemented only in the nontransparent mode.

In the transparent mode, the RLC layer segments a PDU, but does not add any header to a segment (see Figure 6-26).

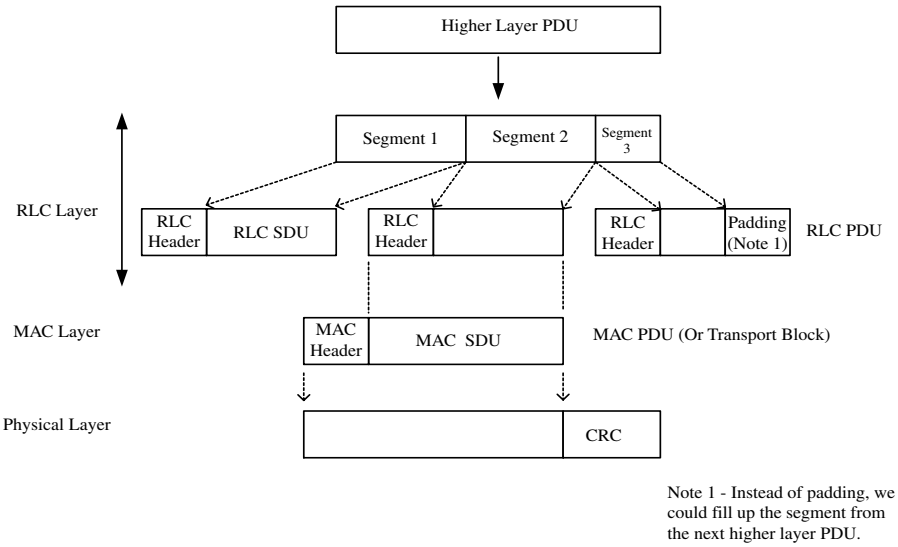


Figure 6-25
Packet construction at different layers for acknowledged and unacknowledged data transfer

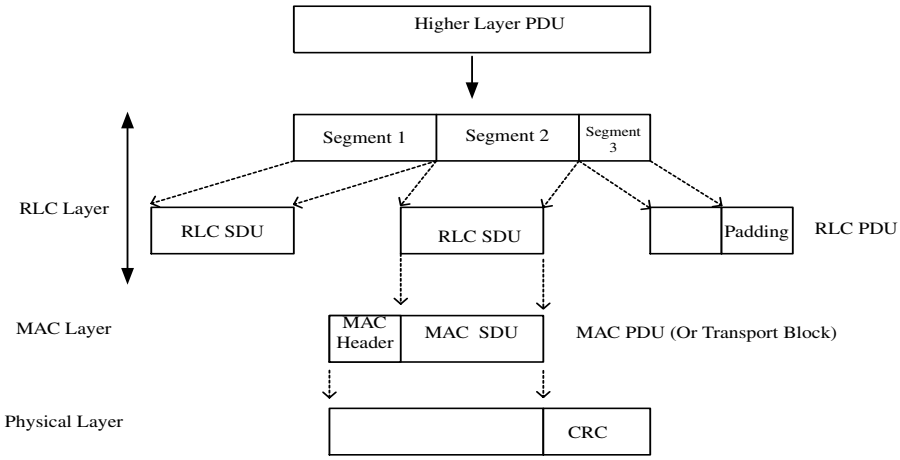


Figure 6-26
Packet construction with transparent RLC

RLC Protocol Description

In this section, we will provide a brief description of the protocol. Specifically, we shall discuss the RLC PDU types and their formats.

Broadly speaking, there are two types of PDUs—the data PDUs, which are used to transfer user data and control PDUs, which carry control information such as status. There are three data PDUs corresponding to the three data transfer modes—the transparent PDU, the unacknowledged PDU, and the acknowledged PDU. The transparent PDU is used to transfer any length of user data without adding any RLC header. Here the upper-layer PDU may be segmented, but neither a header nor a sequence number is added to any segment. An unacknowledged PDU transfers user data sequentially using a sequence number in the header, but does not provide for any acknowledgment to the sender. An acknowledged PDU transfers user data sequentially using a sequence number in the header and also provides for an acknowledgment to the sender by means of a control PDU, such as STATUS, which may be sent as a completely separate PDU or as part of the data PDU in a piggyback fashion.

There are three control PDUs—STATUS, RESET, and RESET ACK—which are used in the acknowledged mode only. A receiving RLC sublayer uses the STATUS PDU to inform the transmitter which PDUs it has received correctly and which PDUs are missing,

and may also indicate the window size that it is using.[20] It can also be used by a transmitting entity to request its peer to adjust its window size. The RESET PDU is used to command the other end to reset its protocol state variables and timers so that the two ends may be synchronized with each other. RESET ACK is simply an acknowledgment of the RESET command.

Because the transparent PDU does not contain any header, the RLC PDU has the same format as the higher-layer PDU. The PDU format for the unacknowledged mode is shown in Figure 6-27. The sequence number is seven bits long. The E bit allows a field to be extended—if it is 0, it means that the next octet is a data field. If it is 1, it means that the next octet includes the length indicator field and the E bit. The length indicator field gives the size of the data field in octets. To illustrate the use of the E bit, we notice that it should be set to 1 in the first octet. If E is 1 in the second octet and 0 in the third, we have a 14-bit length indicator field, and in that case, the user data begins from the fourth octet. Similarly, if the E bit is 0 in the second octet, the length indicator field is 7 bits wide. The purpose of the padding field is to ensure that the PDU has the required length.

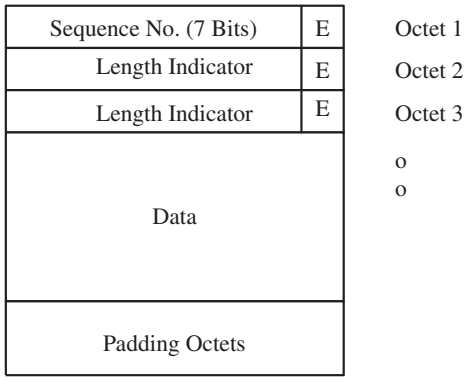


Figure 6-27
The RLC PDU format for the unacknowledged mode

[20]The receiver may acknowledge each packet separately as it comes or may wait until it receives a number of them, say *W*, and then acknowledge all of them using a single acknowledgment message. In that case, *W* is called the window size.

The PDU format for the acknowledged mode is shown in Figure 6-28. The D/C field indicates if it is a data PDU or a control PDU in the acknowledged mode—a 0 indicates a control PDU and a 1 a data PDU. The sequence number (for the acknowledged mode) is 12 bits long. The P bit is set to 1 when the transmitter wants to request a status report from the receiver. The HE field consists of 2 bits and is used in the same way as the E bit—if it is 00, it means that the next octet is a data field. If it is 01, it means that the next octet includes the length indicator field and the E bit. The other two values are reserved for future use.

The STATUS PDU provides status indication in the acknowledged mode. The status information of both the receiver and transmitter may be included in the same PDU. As we said before, it is used only in the acknowledged transfer mode and may be sent either as a completely separate PDU or as part of the data PDU in a piggyback fashion. Its format, when sent separately, is depicted in Figure 6-29.

The one-bit D/C field distinguishes the control PDU from the data PDU—it is 0 for the former and 1 for the latter. The 3-bit PDU type field is 000 for a STATUS PDU, 001 for a RESET command, and 010 for RESET ACK. Its other possible values are reserved.

Superfields (SUF) are actually variable-length information elements of the STATUS PDU and perform many functions. For exam-

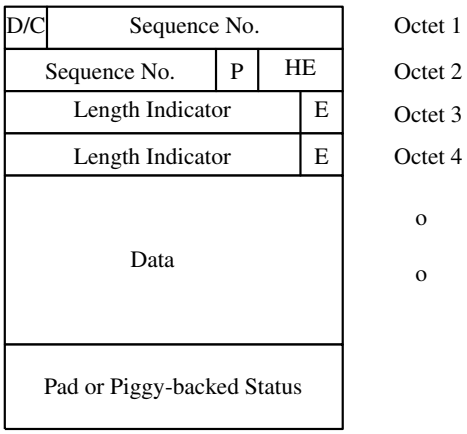


Figure 6-28 The RLC PDU format for the acknowledged mode

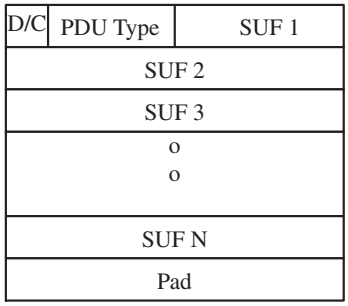


Figure 6-29
The RLC STATUS PDU format

ple, they may indicate if there is more data to follow, inform the peer RLC which PDUs have been received correctly so that the sending end can repeat a transmission if necessary, and so on. Each superfield may have as many as three subfields:

- Type
- Length
- Value

See Figure 6-30 for an illustration. The type is a four-bit identifier and must always be included. The other subfields are optional. For example, the particular SUF, which indicates that there is no more data to follow, has only the type field (which is 0000). This SUF is placed at the end of the list of superfields, immediately ahead of the padding bits, if any.

Another superfield is acknowledgement—it has two subfields: a 4-bit type (0010) and a 12-bit length field called the *last sequence number* (LSN). The receiver uses this superfield to indicate to the transmitter that it has received correctly all PDUs with sequence numbers less than the value of the LSN field except for those which are explicitly indicated as erroneous by the superfield named List (type = 0011). The superfields are shown in Table 6-3.

The piggybacked STATUS PDU has the same format as shown in Figure 6-29 except that the D/C bit is now a reserved bit because it is not needed any more when the status information is included in the data PDU.

Figure 6-30
Superfield format

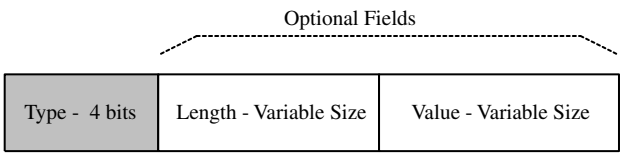


Table 6-3

The superfields of the STATUS PDU

Superfield Description	Type	Length	Value
No more data	0000	None	None
Acknowledgment	0001	12 bits	None
Window	0010	12 bits	None
List	0011	4 bits	A number of 16-bit fields—the first 12 bits indicate the sequence number received incorrectly and the next 4 bits the corresponding list number
Bitmap—the position of a bit indicates a sequence number with respect to a reference. A 0 in a given position indicates that the corresponding sequence number has been received incorrectly. A 1 indicates otherwise.	0100	4 bits	Variable number of octets
Relative list—gives an alternative way to indicate which sequence numbers are received incorrectly.	0101	12 bits	Variable number of octets
Move receiving window—the receiver is requested to change its window size and indicate the discarded SDUs.	0110	4 bits	Variable number of octets
Move receiving window	0111	4 bits	Variable number of octets

AMFLY

Packet Data Convergence Protocol (PDCP)

Overview

Currently, the most common protocols for data services are IPv4 and IPv6 at the network layer and TCP and *User Datagram Protocol* (UDP) at the transport layer. Because new services and applications are continually being introduced, it is conceivable that in the future, new network layer protocols would be developed supporting these services. As such, it is desirable that the UTRAN be able to interwork with these emerging protocols without requiring any changes to its RLC or MAC layers. The packet data convergence protocol helps achieve this goal by adapting different network layer protocols to the RLC layer so that the user data can be transferred across the UTRAN transparently. Another function of PDCP is to improve the transmission efficiency for delay-sensitive information such as voice or video. It does this by reducing the size of the header fields of upper-layer packets using different header compression and decompression algorithms that have been standardized by IETF. In the initial version of this protocol, a separate PDCP entity is used for each radio bearer. In the later version, however, it will be possible to multiplex two or more radio bearers onto the same RLC PDU using a single PDCP entity.

PDCP receives the user data from the higher layers, performs the header compression if requested, and sends the resulting PDCP PDUs to the RLC layer. Notice that these PDUs are nonaccess stratum because they are exchanged between the UE and the core network, and do not terminate in the UTRAN. The data transfer between peer RLCs may take place in acknowledged, unacknowledged, and transparent modes. In the acknowledged mode, the receiving RLC, after receiving the data, passes it to its PDCP layer and returns an acknowledgment to its peer. In the other two modes, there is no acknowledgment involved.

Header Compression

The transmission efficiency depends, to some extent, on the size of the packet header compared to the length of the entire packet. Consider, for example, IPv6 where the header is 48 octets long. If the user data in the packet is also 48 octets, the transmission efficiency is only 50 percent. If 13 kb/s coded speech is being transmitted, it requires about 29.5 ms of speech samples to construct a packet.[21] Thus, large headers not only require higher channel bandwidths and decrease the transmission efficiency but also increase the delay.

Compression algorithms are based on the fact that many of the header fields remain the same during the life of a call and do no change from one packet to another. For example, most of the fields of an IP header remain constant. Similarly, with TCP and UDP, the source and destination address and port numbers do not change. Thus, it is sufficient to transmit a packet with the entire header only at the beginning of a sequence of packets and thereafter send only those fields that have changed in the meantime. The receiving end saves the full header and uses it to decompress the received header. If, however, the entire header changes at some point, it will be necessary to transmit the full header to the receiving end.

Broadcast/Multicast (BMC) Protocol

This layer 2 sublayer is responsible for transferring messages from the network that are to be broadcast or multicast to all mobile stations in a cell. The user data that comes from the higher layers are temporarily saved in the BMC, if necessary, until they are scheduled for transmission. The transfer takes place using the unacknowledged mode service of the RLC on a common traffic channel to the

[21]At 13 kb/s, the time required to collect 48 octets of speech samples is $(48 \times 8)/13{,}000$ seconds or 29.5 ms.

MAC layer.[22] User data that is not intended to be broadcast over a cell is passed transparently by the BMC. As for other functions performed, this entity on the network side periodically estimates the volume of the cell broadcast traffic and forwards the information to the RRC layer using an indication primitive.

There are two types of messages in BMC—the cell broadcast message from the network and the schedule message that gives the location of broadcast messages in the next schedule period as well as the location of the schedule message for the following period. The BMC layer, on the UE side, determines which messages to forward to its upper layer and sends an indication to the RRC layer.

The BMC message format is very similar to that of an ISDN call control message and consists of a number of information elements. As an example, the information elements of a cell broadcast message are shown in Figure 6-31.

Radio Resource Control Protocol

RRC Functions

Radio resources comprise W-CDMA frequencies, different channel types, channelization codes, spreading factors, scrambling codes, the capability to control transmitter power, and so on. The RRC layer manages these radio resources so as to establish, maintain, and release connections between UTRAN and the UE. Reference [19] describes in detail functions performed by the RRC, procedures,

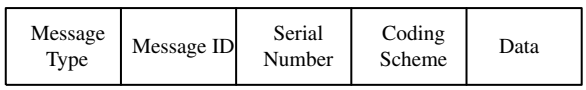


Figure 6-31 The cell broadcast message format

[22]This is a logical channel.

messages, and their constituent information elements, specifies timers and counters, and discusses specific functions such as traffic volume measurements and how they are to be reported. In this section we shall provide a brief overview of the RRC layer.

Among the functions performed at this layer are the following:

- Broadcasting nonaccess stratum user data such as cell broadcast messages to UE that are transferred transparently through the UTRAN.
- Broadcast of system information. The system information includes, among other things, NAS system information, timers and counters used by the UE in the idle mode, *user registration area* (URA) identity and information for periodic cell and URA updates, parameters for cell selection and reselection, configuration parameters for the common physical channels in the cell or for the common and shared physical channels in the connected mode, uplink interference and dynamic persistence levels for PRACHs, and CPCH parameters to be used in the cell.
- Paging and notification. Paging messages are usually sent by higher layers to establish a connection. However, paging may also be initiated by the UTRAN to force UEs to read the latest system information or transition, if necessary, to a known state.
- Initial cell selection and reselection. The UE is required to initiate cell selection on a designated UTRA carrier and camp on it.
- Management of RRC connections between UE and UTRAN and associated radio resources.
- Management of radio bearers. A radio bearer is an RLC layer service for transferring user data between UE and the serving radio network controller. The UTRAN may establish a new radio bearer by sending, to a UE, a RADIO BEARER SETUP message from its RRC layer indicating, among other things, the desired uplink and downlink transport channels, the frequency, and the maximum allowed uplink transmit power.

- Support of mobility functions.
- Routing higher-layer PDUs.
- Management of links between *opportunity-driven multiple access* (ODMA) relay nodes.
- Providing an interworking capability between UTRAN and a gateway node.
- Contention resolution in the TDD mode.
- Arbitration of radio resources on the uplink dedicated control channel.
- UE measurements and their reporting.
- Admission control or, its broader equivalent, QoS control.

Management of RRC Connections

Because signaling and control information flows through the RRC, it is necessary to establish an RRC connection between the UE and UTRAN. Connections are established, removed, or reestablished using layer 3 call control messages. The message constructs and procedures are very similar to those of other network layer protocols such as ISDN Q.931. For example, both protocols specify a set of messages, each consisting of a number of information elements and associated with a set of rules that govern how the particular message is to be initiated and handled.

Here is an example to illustrate the RRC connection procedure. Assume that a UE wants to send some user data to a destination endpoint. As we previously mentioned, the user data packets do not terminate in the UTRAN, but go all the way to the core network. To do this, it is necessary, first of all, to establish a connection between the peer RRCs of the UE and UTRAN. To that end, the UE requests the UTRAN to initiate an RRC connection. The messages exchanged between them are shown in Figure 6-32.

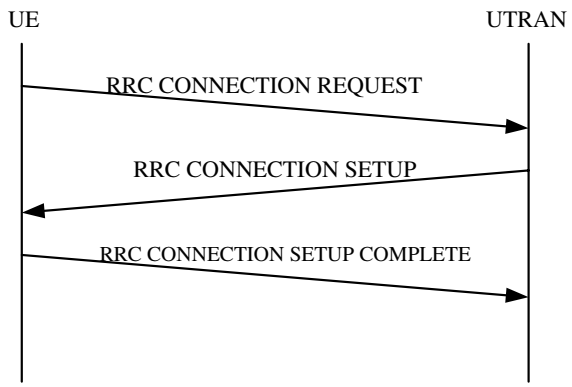


Figure 6-32
RRC connection procedure

Handover

Handover is the process by which the communication with a mobile station is transferred from one radio channel to another. As in narrowband CDMA systems, there are different types of handovers—hard handover and soft handover. The hard handover takes place when the base stations participating in the handover process operate on different CDMA carriers, thus requiring that all old radio links be released before new ones are established. Consequently, this type of handover causes the received signal to be interrupted even though it may be for a short time. In a soft handover, on the other hand, a mobile station can receive signals from two or more base stations or two or more sectors of one or more base stations at the same time. As such, the received signal is not interrupted. A soft handover is possible only when the participating base stations use the same CDMA carrier frequency, which is usually the case. The IMT-2000 supports intracell, intercell, and multibearer handovers. In fact, seamless handover without any perceptible degradation in the received signal quality is a desired goal of 3G systems.

The handover in W-CDMA is similar in concept to the handoff procedure in cdmaOne based on the standard IS-95. For example, like cdmaOne, it is also triggered by a measurement of the pilot strength.

But there are some significant differences. Recall that in cdmaOne, if the signal strength of a pilot exceeds a given threshold, that pilot is taken to be a candidate for handover and is added to the candidate set. In other words, we do not compare the pilots and then select one that is relatively stronger. The threshold may be set to different values by different base stations but does not change dynamically. In W-CDMA, on the other hand, a pilot is selected on the basis of its relative signal strength compared to other pilots.

Handover Types As in cdmaOne, there are three types of handovers in W-CDMA:

- Soft handover where all cells in a serving area use the same W-CDMA frequency.
- Hard handover where the participating cells operate on different W-CDMA frequency bands. Here, not only is it necessary to change the frequency, but also the mode may have to be changed, say, from FDD to TDD.
- Intersystem handover, for example, with GSM. The network initiates this handover by issuing a handover command.

For the purpose of handover, the UE maintains a list of the cells that it is currently using or may likely use at some point during a call. This list includes the following:

- *Active set* It consists of all cells that are simultaneously involved in a communication during a soft handover. The UE demodulates the received signals from these cells and coherently combines them to provide diversity. The net gain in performance depends, among other things, upon the relative path loss from the participating base stations to the UE and may be as much as 2 dB or so. An active set contains two or more cells for an FDD system but only one in a TDD mode.
- *Monitored set* These are cells that are not in the active set but are monitored by the UE because they are part of the neighbor list.

- *Detected set* These are cells that are neither in the active nor in the monitored set but are detected by the UE anyway.

In what follows, we shall only describe the soft handover [43].

Soft Handover The soft handover concept is illustrated in Figure 6-33. As we just said, the UE maintains an active set for handover purposes. The permissible number of cells in an active set is a system parameter. Assume that cell 1, being the strongest for a given UE, is the only cell in the active set. If, at a certain instant t_1, the pilot associated with cell 2 is sufficiently strong that the difference Δ_1 between the signal strengths of pilot 1 and pilot 2 is less than a threshold, we can say that pilot 2 is usable, and can therefore include it in the active set. So, from this point on, the UE is in communication with two cells and may, as a result, use diversity combining. This threshold is $L - H_1$, where L is the reporting range, and H_1 is the *addition hysteresis*. If, at some later time, say, t_2, pilot 1 has degraded enough that the difference Δ_2 between pilot 2 and pilot 1 is greater than another threshold, pilot 1 is no longer usable and can, there-

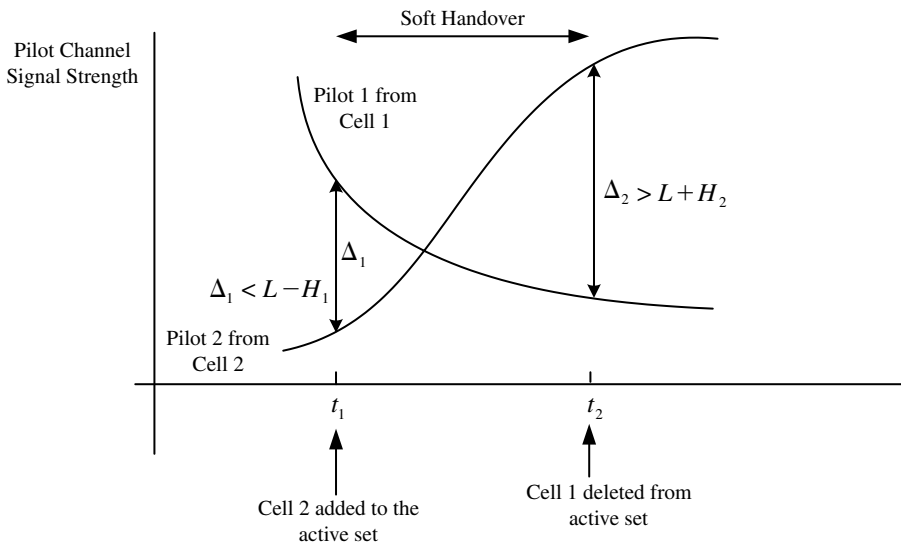


Figure 6-33 Soft handover in UTRAN

fore, be removed from the active list. Thus, from now on, the UE is in communication with only one cell, namely cell 2. This second threshold is $L + H_2$, where H_2 is the *removal hysteresis*.

As the mobile moves away from its present cell into the coverage area of another, the signal level from the present cell will fall with respect to the signal from the new cell as shown in Figure 6-34. At instant t_0, the signal strength of the best candidate exceeds pilot 1. Consequently, at that point, we can replace pilot 1 with the new one. This would mean that if pilot 1 was the only member of the active set, the UE will now be communicating with the new cell exclusively, instead of cell 1. If, on the other hand, there were two or more pilots in the active set, the weakest pilot is compared with the new and subsequently replaced if the criterion indicated in the figure is fulfilled. As a result, the UE is now involved in communication with this new cell as well as all old cells except cell 1 (or a particular sector of a cell).

The UE updates the active set on command from the UTRAN and sends an acknowledgement back. The messages exchanged when the active set is to be updated are shown in Figure 6-35.

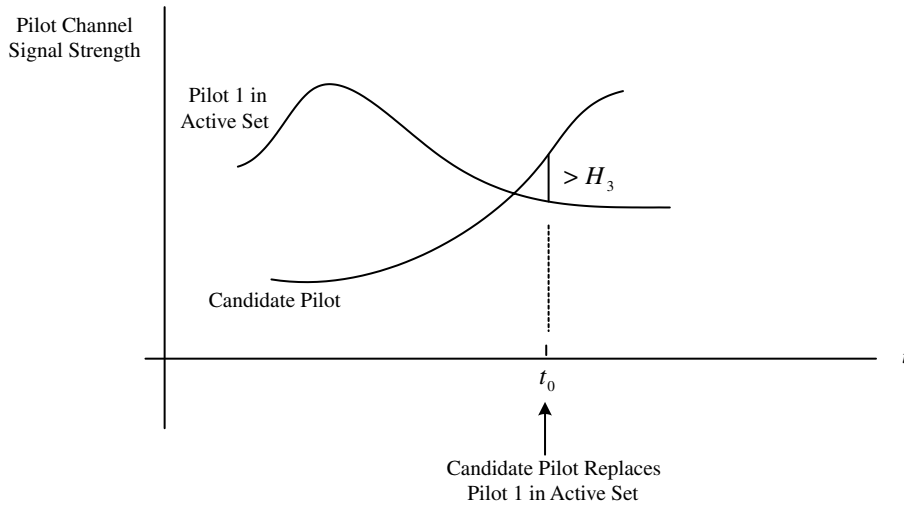


Figure 6-34
Intercell handover. In this case, all cells have the same W-CDMA carrier.

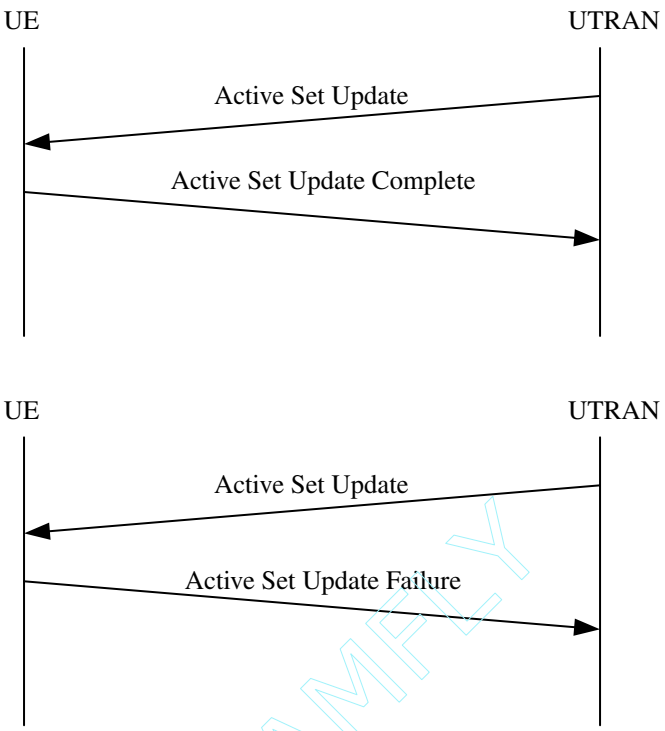


Figure 6-35 *Procedure to update the active set*

TEAMFLY

Summary

In this chapter, we have presented a brief description of the UMTS system, its features, and the UTRAN network architecture. The radio interface protocol stack of the UTRAN, which is also the same as the lower-layer protocols of UE, has been presented in some detail. More specifically, we have described the physical layer, the medium access control layer, radio link control, the packet data convergence layer, the broadcast multicast protocol, and the radio resource control protocol. Procedures such as those used in synchronization, power controls, and handovers are also described.

The key features of UMTS W-CDMA may be summarized as follows:

- *Wider bandwidth* This is a direct-spread CDMA system with a nominal bandwidth of 5 MHz. The chip rate is 3.84 Mc/s. A radio

frame is usually 10 ms long and consists of 15 slots, each of duration 2,560 chips.

- *Asynchronous operation* There is no need for cell sites to be synchronized to each other using a global timing reference. Each cell site may operate in a fully asynchronous manner. However, this requires a different scrambling code for each cell or each sector of a cell.
- *Channel coding* Incoming data, depending upon applications, may not be encoded at all or may be encoded into either a convolutional code of rate 1/3 or 1/2, or turbo code of rate 1/3.
- *Spreading codes* Physical channels are separated at the receiver by spreading them with channel-specific OVSF codes. A spreading factor of 256 is used for control channels. For user data channels, spreading factors vary from 4 to 256 on uplinks and 4 to 512 on downlinks.
- *Scrambling codes* Uplink scrambling codes are complex valued and may be either long or short. The long codes have a length of 38,400 chips (that is, 10 ms), whereas short codes are only 256 chips long. The short codes are particularly useful for multiuser detection at base stations. Downlink scrambling codes are also complex-valued. There are a total of $2^{18} - 1$ of these codes. However, only 8,192 are used on downlinks. They are divided into 512 groups, each containing one primary scrambling code and 15 secondary scrambling codes. Each code is of length 38,400 chips.
- *Complex spreading* W-CDMA uses complex spreading that reduces the amplitude variations of the baseband filter output, thus making the signal more suitable for nonlinear power amplifiers. It also provides better efficiency by reducing the difference between the peak power and the average power.
- *Variable bandwidth* Any user equipment may be assigned a variable bandwidth by simply changing the spreading factors and allocating one or more slots and one or more dedicated channels to the UE. Similarly, the system supports multiple applications simultaneously for the same UE.

- *Packet mode data services* W-CDMA UMTS supports a highly flexible packet mode data service. The multiple-access procedure is based upon the slotted Aloha scheme. Channels that may be used for this purpose include the RACH, CPCH, dedicated channels, and FACH.
- *Coherent demodulation, multiuser detection, and adaptive antenna arrays* The system has been designed to facilitate coherent demodulation using pilot bits and supports such advanced technologies as beam forming with adaptive antennas and multiuser detection techniques.
- *Transmit diversity* In contrast to GSM, the performance of W-CDMA can be improved to some extent by implementing transmit diversity on a downlink channel.

References

General Systems Descriptions

[1] 3G TS 22.105, Service Aspects; Services and Service Capabilities.

[2] 3GPP TS 23.107, QoS Concept and Architecture.

[3] 3GPP TS 25.401, UTRAN Overall Description.

[4] 3GPP TS 25.101, UE Radio Transmission and Reception.

[5] 3GPP TS 25.104, UTRA (BS) FDD, Radio Transmission and Reception.

[6] 3GPP TS 25.105, UTRA (BS) TDD, Radio Transmission and Reception.

Overview of the UE-UTRAN Protocols

[7] 3GPP TS 25.301, Radio Interface Protocol Architecture.

Physical Layer

[8] 3GPP TS 25.201, Physical Layer—General Description.

[9] 3GPP TS 25.211, Physical Channels and Mapping of Transport Channels onto Physical Channels (FDD).

[10] 3GPP TS 25.212, Multiplexing and Channel Coding.

[11] 3GPP TS 25.213, Spreading and Modulation (FDD).

[12] 3GPP TS 25.214, Physical Layer Procedures.

[13] 3GPP TS 25.215, Physical Layer—Measurements.

[14] 3GPP TS 25.302, Services Provided by the Physical Layer.

Layer 2 and Layer 3 Protocols

[15] 3GPP TS 25.321, MAC Protocol Specification.

[16] 3GPP TS 25.322, RLC Protocol Specification.

[17] 3GPP TS 25.323, Packet Data Convergence Protocol (PDCP) Specification.

[18] 3GPP TS 25.324, Broadcast/Multicast Control (BMC) Protocol Specification.

[19] 3G TS 25.331, RRC Protocol Specification.

[20] 3G TS 25.303, Interlayer Procedures in Connected Mode. Also, 3GTS 25.304, UE Procedures in Idle Mode and Procedures for Cell Reselection in Connected Mode.

Protocols at Different Interface Points

[21] 3GPP TS 25.410, UTRAN Iu Interface: General Aspects and Principles.

[22] 3GPP TS 25.411, UTRAN Iu Interface: Layer 1.

[23] 3GPP TS 25.412, UTRAN Iu Interface: Signaling Transport.

[24] 3GPP TS 25.413, UTRAN Iu Interface: RANAP Signaling.

[25] 3GPP TS 25.414, UTRAN Iu Interface: Data Transport and Transport Signaling.

[26] 3GPP TS 25.415, UTRAN Iu Interface: CN-RAN User Plane Protocol.

[27] 3GPP TS 25.420, UTRAN Iur Interface: General Aspects and Principles.

[28] 3GPP TS 25.421, UTRAN Iur Interface: Layer 1.

[29] 3GPP TS 25.422, UTRAN Iur Interface: Signaling Transport.

[30] 3GPP TS 25.423, UTRAN Iur Interface: RNSAP Signaling.

[31] 3GPP TS 25.424, UTRAN Iur Interface: Data Transport and Transport Signaling for CCH Data Streams.

[32] 3GPP TS 25.425, UTRAN Iur Interface: User Plane Protocols for CCH Data Streams.

[33] 3GPP TS 25.426, UTRAN Iur and Iub Interface Data Transport and Transport Signaling for DCH Data Streams.

[34] 3GPP TS 25.427, UTRAN Iur and Iub Interface User Plane Protocols for DCH Data Streams.

[35] 3GPP TS 25.430, UTRAN Iub Interface: General Aspects and Principles.

[36] 3GPP TS 25.431, UTRAN Iub Interface: Layer 1.

[37] 3GPP TS 25.432, UTRAN Iub Interface: Signaling Transport.

[38] 3GPP TS 25.433, UTRAN Iub Interface: NBAP Signaling.

[39] 3GPP TS 25.434, UTRAN Iub Interface: Data Transport and Transport Signaling for CCH Data Streams.

[40] 3GPP TS 25.435, UTRAN Iub Interface: User Plane Protocols for CCH Data Streams.

Miscellaneous Specifications of Interest

[41] 3G TR 23.922, Architecture of an All IP Network.

[42] 3G TR 25.990, Vocabulary.

[43] 3G TR 25.922, Ver. 0.5.0, Radio Resource Management Strategies.

Other References

[44] N. Abramson, "The Throughput of Packet Broadcasting Channels," IEEE Trans. Comm., Vol. COM-25, No. 1, January 1977, pp. 117-128.

[45] S.W. Golomb, *Shift Register Sequences*. Revised Edition, Aegean Park Press, Laguna Hills, CA, 1982.

Web Sites

http://www.itu.int/publications/

http://www.itu.int/imt/2-rad-devt/index.html

http://www.itu.int/brsg/ties/imt-2000/index.html

This page intentionally left blank.

CHAPTER 7

Evolution of Mobile Communication Networks

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

As the access part of a mobile communication network is evolving towards *third generation* (3G) to support new air interfaces, so is the architecture of the core network. In order to get the maximum return from their investment, service providers want a network that would be adequate for current customer needs, but at the same time be able to provide new, emerging services by simply adding some new capabilities in the form of a hardware and/or software upgrade to their existing equipment. Because the second generation wireless systems are required to support only limited data, such as short messaging services and slow-speed circuit-switched or packet mode data, the current network is principally circuit-switched, but includes an entity called the *interworking function* to provide the data capabilities. Now, however, the demand for higher data rates is growing at a rapid rate. Because packet-switched networks are inherently more efficient for data services, networks are evolving that combine the more common, ubiquitous circuit-switched fabric with elements of a packet-switched network. One such example is the *general packet radio service* (GPRS) that can support packet mode data at rates up to about 160 kb/s [9], [10]. In view of the requirements of the 3G systems for both constant and variable bit rate services, the need for such a network appears to be even more compelling than ever before. In fact, because of these 3G requirements and emerging applications (such as conversational voice and video, interactive data, high volume data transfer, and so on) with a guaranteed quality of service, the network is gradually evolving to an all-IP architecture [12].

In this chapter, we will discuss this evolution of wireless networks. But first we will review the 3G system requirements so that we can understand the driving forces behind the network evolution.

Review of 3G Requirements [1]–[4]

3G wireless systems are required to provide traditional voice, enhanced voice, multimedia services, and high-speed circuit and packet mode data to mobile users as well as special services such as paging and address dispatch or fleet operation. A mobile station may run multiple applications at any time; however, the network is required to support, for each mobile station, a total bit rate of

- 144 kb/s or more in vehicular operations
- 384 kb/s or more for pedestrians
- About 2.048 Mb/s for indoor or low-range outdoor applications

Some user applications may require bandwidth on demand and a guaranteed *quality of service* (QoS) from networks. Thus, the core network should be capable of reserving resources based on user requests and making sure that all users get the requested quality. 3G standards call for efficient utilization of the spectrum and, in some cases, phased introduction of these services. Open interfaces should be used wherever possible. The service quality to be provided to the mobile users is intended to be comparable to that available from fixed networks and should be maintained even when more than one service provider is operating in a serving area. All these services should be provided to each user with an acceptable degree of privacy and security that would be at least as good as or better than what is currently available over a *Public Switched Telephone Network* (PSTN). Finally, the 3G networks should be synergistic with the architecture of future networks.

The user traffic in 3G may be

- Constant bit rate traffic such as speech, high-quality audio, video telephony, full-motion video, and so on, which are sensitive to delays and more importantly, delay variations.
- Real-time variable bit-rate traffic such as variable bit-rate encoded audio, interactive MPEG video, and so on. This type of traffic requires variable bandwidths and is also sensitive to delays and delay variations.
- Nonreal-time variable bit-rate traffic such as interactive and large file transfers that can tolerate delays or delay variations.

From these requirements it appears that a hybrid architecture such as the one GSM with GPRS enhancements is a possibility for 3G systems. However, because more and more of the emerging applications require bandwidth on demand, an all-packet fabric is an attractive alternative, particularly if it can be designed to support delay-sensitive real-time applications.

Network Evolution

First-Generation Network

We begin with the network reference model of the *Telecommunications Industry Association/Electronics Industry Association* (TIA/EIA) standard IS-41 [5], which is shown in Figure 7-1. This also represents the network for the first generation systems that support only voice and no data. This reference model is similar to the GSM architecture.

The *mobile switching center* (MSC) performs mobile switching functions and interfaces the cellular network to a PSTN, *Integrated Services Digital Network* (ISDN), or another MSC. *Home Location Register* (HLR) contains a centralized database of all subscribers to the home system. This database includes such information as the *electronic serial number* (ESN), *directory number* (DN), the service profile subscribed by this user (such as roaming restriction, if any, supplementary services that this mobile has subscribed to, and so on), and its current location. Similarly, *Visitor Location Register* (VLR) contains a database of all visitors to this particular system. Whenever a mobile station moves into a foreign service area, its

TEAMFLY

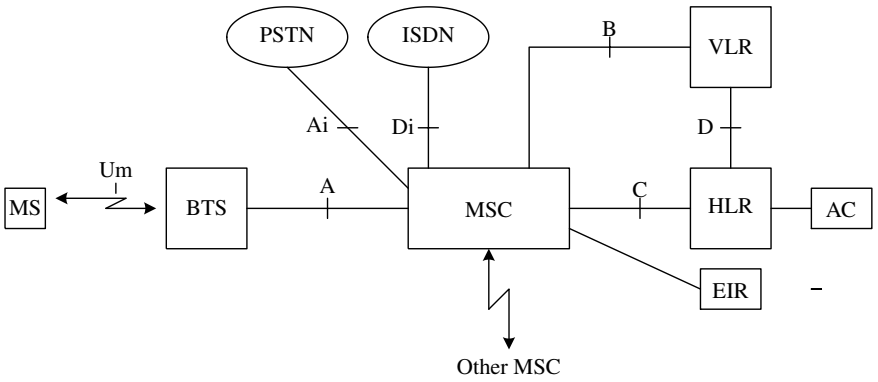


Figure 7-1 The reference model of a mobile communication network

MSC saves all the pertinent information of that mobile station in its VLR. The home MSC is also notified so that incoming calls to this mobile can be forwarded to the foreign MSC. The information in the VLR is really the same as that of the HLR. However, when the mobile moves out of this foreign serving area, its MSC removes the database of this visitor from its VLR. The *equipment identity register* (EIR) contains the equipment identification number. The *authentication center* (AC) manages user data-encryption-related functions such as ciphering keys, and so on.

The intersystem operations between entities at reference points B, C, and D are specified in the EIA Standard IS-41, which, more specifically, define procedures for handoff as a mobile moves from the service area of one MSC to another and automatic roaming.

IS-634-A [6], [7] defines the interface at reference point A between an MSC and a base station. It specifies the interface requirements for all types of user traffic and signaling information exchanged over this reference point. The *Asynchronous Transfer Mode* (ATM) protocol is used to transport the following information:

- The coded user traffic (such as user data or low bit-rate speech) and the signaling information between an MSC and a *base station* (BS). Separate logical channels carry the user traffic and the signaling information. These interface functions are designated as the A3 interface.
- The signaling information between a source BS that initially serves a call and any other BS that supports this call (that is, the target BS). This interface function is designated as the A7 interface.

Figure 7-2 shows the protocol stack for these interfaces. Notice that at the ATM adaptation layer, AAL5 is used for signaling and packet mode data, and AAL2 for the user traffic. AAL Type 2 is intended for variable-bit-rate, circuit-switched applications where the source timing information may have to be transmitted to the receiving end. AAL Type 5, on the other hand, is used in connectionless, variable-bit-rate services where the receiving end-point does not require this timing information.

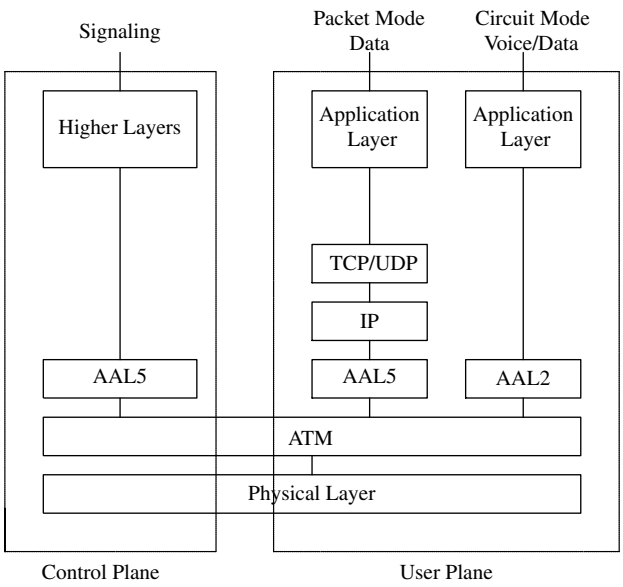


Figure 7-2
The protocol stack for A3 and A7 interfaces

Second-Generation Networks

An important feature of the *second-generation* (2G) systems is their data service capability. For example, IS-95 supports circuit-switched data and digital fax, IP, mobile IP, and *cellular digital packet data* (CDPD). GSM provides the short messaging service and circuit-switched data at rates up to 9.6 kb/s per slot. Figure 7-3 shows a network architecture that supports these data services as well as voice. Notice that it is very similar to that of Figure 7-1 except for its interface to a *public data network* (PDN). This interface to the PDN is via an interworking function labeled IWF, which actually performs some protocol conversion that might be necessary because of the differences in the protocols used on the mobile stations and the PDN.

To see what kind of protocol conversion is performed by IWF, consider Figure 7-4, where we show the protocol stacks between a mobile station and a base station and between IWF and PDN for packet data transmission in an IS-95 system.

The *radio link protocol* (RLP) accepts a packet from the link layer (that is, the layer above it), segments it into smaller sizes that fit the

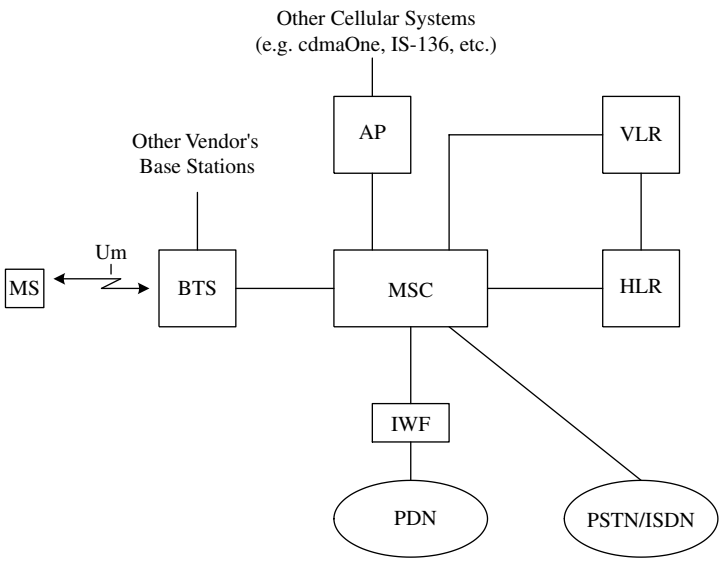


Figure 7-3
2G wireless network with packet data services

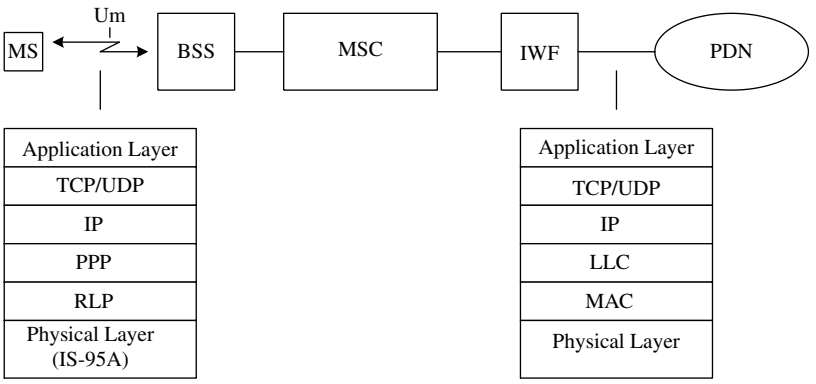


Figure 7-4
Protocol stacks at the reference point Um between a mobile station and a base station, and between IWF and PDN

20-ms frames of a traffic channel, and then passes them to the physical layer where they are transmitted over the radio interface. The *point-to-point protocol* (PPP) is a byte-synchronous, data link layer protocol, which takes a datagram packet from the IP layer, adds a frame check sequence, encloses it between two flags, and passes it to the layer below. The well-known IP layer protocol interconnects two packet switching nodes and routes an incoming packet to a next node *en route* to its destination. The *Transmission Control Protocol*

(TCP) at the transport layer guarantees reliable data transfer by providing error recovery on an end-to-end basis. The MAC layer protocol is IEEE-802.3 or IEEE-802.5, which along with the logical link control (IEEE-802.2) forms the link layer protocol on the fixed side.

The 2G GSM network was shown in Figure 5-12 [8]. There is really not much difference between that network and the one shown in Figure 7-3 except for the fact that the MSC of the GSM network may additionally include an echo canceler and an audio transcoder.

2G+ Networks

Notice that in Figure 7-3, except for the IS-634-A interface between a BS and an MSC, the core network is circuit-switched. Equipment from many different manufacturers is now available in the market that can support packet mode data in a core network. One possible architecture around which many new networks are being built is shown in Figure 7-5.

The salient features of this architecture are the following: First, it consists of a backbone network that is based on IP/ATM. The use of ATM appears to be almost natural because the interface on the radio

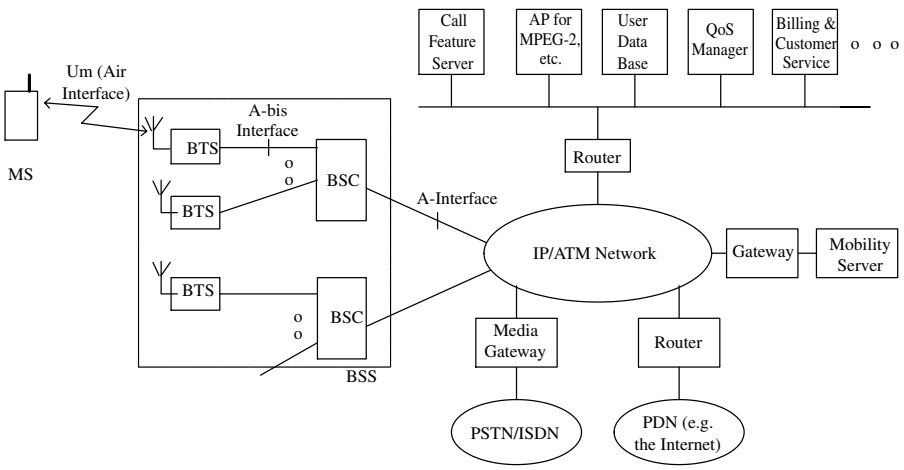


Figure 7-5
The evolution of the core network. The core network is packet switched. Recall that the A interface is based on IS-634-A.

network controllers, which as we said before is IS-634-A, already uses ATM at the link layer. Furthermore, ATM has high-bandwidth capability, and provides low delays and bandwidth-on-demand with guaranteed QoS.

Second, it interfaces to legacy networks in a rather straightforward way. For example, the media gateway performs the necessary protocol conversion between the backhaul ATM network and the circuit-switched PSTN or ISDN. The IP routers are used to route packets to or from IP-based packet data networks. The mobility server, which is based on IP, supports mobility management, connection control, and signaling gateway functions to help provide seamless roaming capability across different networks with centralized directory management and, if needed, end-to-end security. As such, the functional entities of the mobility server would include, among other things, call control, HLR and VLR databases, and radio resources management.

Third, it allows for distributed processing, thus offloading the core network, and provides a platform where new services, features, and applications (such as a new call feature, an MPEG-2, or MPEG-4 application) can be developed, tested, and installed in the network when necessary. Finally, the architecture is compatible with an all-IP network that appears to be the trend of the future.

It's worth mentioning here that GPRS, which has already been introduced in the 2G+ version of GSM, supports packet mode data at rates up to 160 kb/s [9], [10]. The GPRS network, which was discussed in Figures 5-13 and 5-14, is redrawn in Figure 7-6 for the convenience of the reader. The core network consists of a number of *serving GPRS support nodes* (SGSNs), a *gateway GPRS support node* (GGSN), and a *packet control unit* (PCU). The SGSN, which is actually a router, connects to a BSC via a PCU, which implements the link layer protocol. There may be more than one serving GSN in any *public land mobile network* (PLMN) as shown. Two separate PLMNs are connected through a GGSN.

The GGSN is also a router and is the first entry point of the core network from any external packet data network (such as the Internet). The *short messaging service gateway MSC* (SMS-GMSC) provides the necessary protocol conversion for handling SMS through the GPRS network (instead of the traditional GSM network).

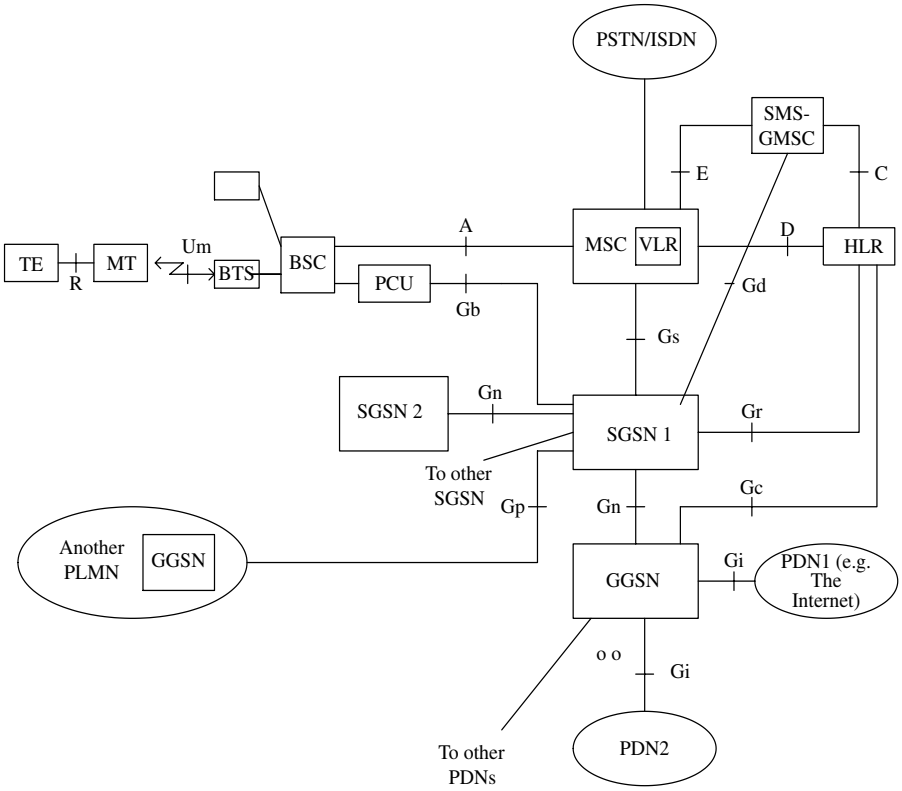


Figure 7-6
The GPRS network architecture

3G Network

The 3G UMTS network architecture was shown in Figure 6-2 [11]. As maybe seen there, the core network combines the circuit-switched MSC of the 2G network with the packet-switched elements of the GPRS network of Figure 7-6, thus providing both voice and packet mode data services with an integrated fabric. Notice how the GPRS network is evolving into the 3G, where the packet control unit of Figure 7-6 no longer exists as a separate entity, its function now being integrated into a radio network controller.

All-IP Network

A core network with an all-IP architecture appears very attractive for a number of reasons. First, because of the tremendous growth in the use of the Internet over the last few years, new applications are constantly emerging, supporting terminals that are IP clients. With an all-IP architecture, it will be easier to bring the benefits of these applications to 3G wireless customers. Similarly, there are many higher-layer protocols that have been already developed or are in the process of development that are IP based, which would be useful in wireless networks. An example of one such protocol is *Resource Reservation Protocol* (RSVP), which is used to make a reservation request in connection with the QoS in a packet-switched network. Thus, if the network is all-IP, we can take advantage of these protocols when implementing the QoS.

Second, some protocols designed for a multimedia terminal use IP at the network layer. For example, multimedia terminals in a 3G system may be based on *International Telecommunications Union* (ITU) standards H.324 and H.320. In either case, the control and indication signals, when transported across a 3G network, use *User Datagram Protocol* (UDP) at the transport layer and IP at the network layer. Similarly, the H.323 protocol for video conferencing over traditional *local area networks* (LANs) also uses IP at the network layer. Besides, as the offered load in the system increases beyond its rated capacity, a packet-switched network is inherently capable of serving all users with only a slight, almost unnoticeable degradation in the service quality.

In view of these considerations, the 3G Partnership Projects are defining an all-IP packet-switch architecture for the core network [12]. This architecture, as we shall see shortly, is actually an evolution of the GPRS network. Some of the fundamental requirements that the IP network must meet are the following:

- Mobiles must be able to roam seamlessly from an IP network to a 2G or 2G+, GSM/GPRS network and to earlier versions of a 3G network, and vice versa.

- Handovers must be supported (a) between any two IP networks, (b) between any two radio access networks within the same IP network, (c) between any two radio network controllers within the same *UMTS Terrestrial Radio Access Network* (UTRAN) in any given IP network, and (d) between an IP network and any 2G or 2G+ network.

Figure 7-7 shows one of the architectures that the *European Telecommunications Standards Institute* (ETSI) is considering as a reference model to provide 3G services to mobile subscribers. For simplicity, the interface to legacy signaling networks and reference points on interfaces have been omitted.

Notice, first of all, that this network is very similar to the GPRS network in Figure 7-6. The serving GSN and the GGSN perform the same functions as in the GPRS network. However, now, the PCU is not a separate unit any more, but is integrated into the radio network controller. The GGSN forwards packets to the legacy PSTN via the *media gateway* (MGW). The purpose of the MGW is to provide some protocol conversion between the packet-switched core network and the circuit-switched PSTN. For example, user information associated with real-time applications, such as voice, video, real-time

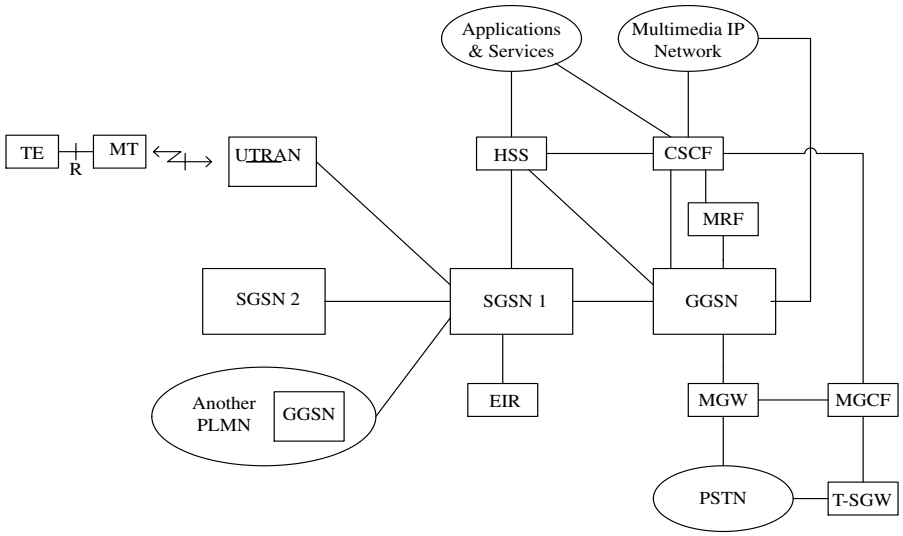


Figure 7-7
A simplified view of the all-IP wireless networ

data, and so on, is transported across a packet-switched network using *real-time protocol* (RTP), TCP/UDP, and IP. The PSTN, on the other hand, transports the user information as circuit-switched data without any overhead or encapsulation that is implicit in a packet mode transmission. Similarly, some transcoding function may be required in the MGW.

The *home subscriber server* (HSS), which is similar to the HLR of 2G and 2G+ networks, is a database of all subscribers to the home system. It contains the user identity, call features and services subscribed by the user, authorization information, and so on. However, unlike the HLR, it must now communicate with other elements of the network via IP.

The *call state control function* (CSCF) works in conjunction with HSS to perform call control procedures. For example, it establishes, maintains, and tears down a call, analyzes and translates the address when there is an incoming call, performs call screening and call forwarding, checks call restrictions (if any) on outgoing calls, and so on.

The *multimedia resource function* (MRF), which is similar in concept to a multi point control unit in traditional video conferencing using H.323, performs multi party and multi media conferencing.

The *media gateway control function* (MGCF) works in conjunction with the CSCF and *transport signaling gateway* (T-SWG), and performs protocol conversion on the signaling information for calls that originate or terminate in a PSTN. This is necessary because the call controls in the core network are IP based, whereas those in the PSTN may use different protocols such as R1, R2, Q.931, and so on.

Summary

In this chapter, we have discussed the evolution of the core network of a mobile communication system. Because the second generation systems are required to have only limited data capabilities, such as short messaging services and slow-speed circuit-switched, or packet mode data, current networks are principally circuit switched. Data services are provided in these networks by means of an entity called

the interworking function that interfaces the core network to an external PSTN or PDN. With the increasing demand for higher data rates, networks are emerging that still include a circuit-switched mobile switching center, but use elements of a packet-switched network more extensively than before. One such example is the GPRS that can support packet mode data at rates up to about 160 kb/s. Because in 3G, a subscriber may run multiple applications simultaneously involving conversational voice and video, interactive data, high volume data transfer, and so on, with a guaranteed QoS, it appears that an all-packet fabric is a possibility for 3G networks, particularly if it can be designed to support delay-sensitive real-time applications. In fact, ETSI is defining a standard adopting an all-IP architecture to deliver 3G mobile telephone services to subscribers.

References

[1] IMT-2000: Recommendations ITU-R M.687-2, 1997.

[2] IMT-2000: Recommendations ITU-R M.816-1, "Framework for Services Supported on International Mobile Telecommunications-2000 (IMT-2000)," 1997.

[3] IMT-2000: Recommendations ITU-R M.1034-1, "Requirements for the Radio Interface(s) for International Mobile Telecommunications-2000 (IMT-2000)," 1997.

[4] 3G TS 22.105 Release 1999, Services and Service Capabilities.

[5] EIA/TIA/IS-41.1-B, Cellular Radio—Telecommunications Intersystem Operations: Functional Overview, 1991.

[6] EIA/TIA/IS-634-A, MSC-BS Interface (A-Interface) for Public 800 MHz, 1998.

[7] M.R. Karim, *ATM Technology and Service Delivery.* New Jersey: Prentice Hall, 1999, Chapters 1 and 2.

[8] A. Mehrotra, *GSM System Engineering.* Norwood, MA: Artecth House, 1997.

[9] GSM 03.60: GPRS Service Description, Stage 2.

TEAMFLY

[10] GSM 03.64: Overall Description of the GPRS Radio Interface, Stage 2.

[11] 3GPP TS 25.401: UTRAN Overall Description, 2000.

[12] 3GPP TR 23.922: Architecture for an All IP Network, 1999.

This page intentionally left blank.

CHAPTER 8

Call Controls and Mobility Management

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

The purpose of this chapter is to provide the reader with a general, high-level concept of *call controls* (CC) and *mobility management* (MM) in wireless networks. Call controls are concerned with signaling procedures for establishing or releasing a call and sending miscellaneous messages (such as a status enquiry, status, information, congestion control, notifications, and so on) when a call is active. The term MM refers to location updates and location reporting as a mobile moves around from the coverage area of one *radio network controller* (RNC) to another or from one system (that is, a core network) to another and includes the registration of *mobile stations* (MS) and their authentication at appropriate instants. Authentication is the process by which the network validates the identity of a mobile subscriber. Whenever an MS performs a location update when relocating from the serving area of one *mobile switching center* (MSC) to another MSC or when it requests service, it may be required to go through the authentication process.[1] Some networks may also provide, on an optional basis, encryption of user data exchanged over the air interface, using a ciphering key that is derived during authentication.

To understand call control procedures, knowledge of signaling protocols is helpful. As such, we begin with a brief description of protocol stacks at reference points between an MS and a *Base Transceiver Station* (BTS), between a BTS and a MSC, and, for some systems, between an MSC/*Visitor Location Register* (VLR) and a *Home Location Register* (HLR). *The Radio Access Network Application Part*

[1]Authentication is performed in the following way. When the VLR determines that it is necessary to authenticate a subscriber's identity, it requests the *Authentication Center* (AUC) (via the HLR) to send the required authentication parameters. Thereupon, the AUC generates (1) a random number and, using the authentication key that has been assigned to the MS at the subscription time, derives (2) a data sequence, say, S1, and (3) a ciphering key. On receiving these three parameters from the AUC, the VLR forwards them to the MSC. The MSC, in turn, sends the random number to the MS. In response, the MS fetches the authentication key that is stored in its memory and uses it and the received random number to derive a data sequence S2, and sends it to the MSC. The MSC compares S1 and S2, and if the two match, the authentication process is successful. In that case, a network may provide encryption of user data over the air interface as an optional service. The user data in either direction may be encrypted using the ciphering key that was generated by the AUC in the authentication process.

(RANAP) provides signaling and control procedures between a *UMTS Terrestrial Radio Access Network* (UTRAN) and the core network. Functions performed by this protocol, messages exchanged when assigning radio resources (such as channels) during a handover, and CC messages exchanged when a call is initiated are briefly described. The chapter concludes with an example CC scenario.

Protocol Stacks in Access and Core Networks

Because the core network and protocols used in *Universal Mobile Telecommunications System* (UMTS) are similar to those of *Global System for Mobile Communications* (GSM), we will begin the chapter with a brief description of the GSM protocols.

GSM

The signaling protocol stacks in the GSM core and access network are shown in Figure 8-1. The interface between a *base station controller* (BSC) and an MSC (that is, *the interface A*) is based on the *Signaling System 7* (SS7) protocol [10], [11]. At the physical layer, it uses *Message Transfer Part* (MTP) Level 1, and generally operates at

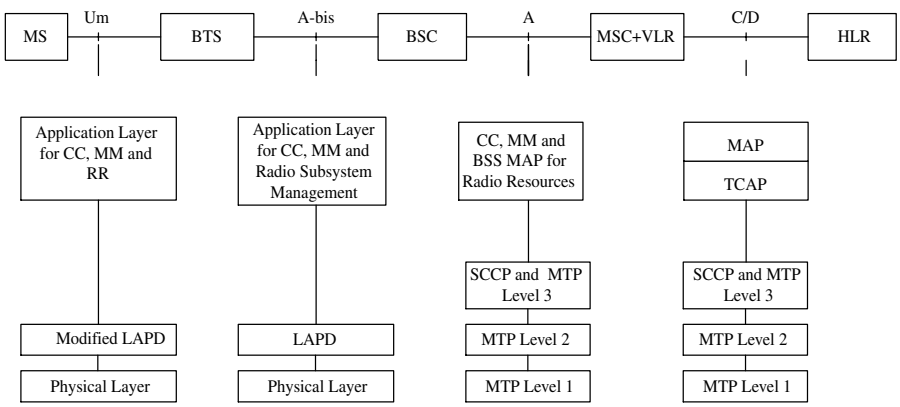


Figure 8-1 The signaling protocol stacks in GSM

64 kb/s on a DS0 time slot of an E1 interface. MTP Level 2 is a link layer protocol that specifies the frame format, functions, and procedures for transferring user part information over a signaling data link. MTP Level 3 provides for the distribution and routing of connectionless signaling messages to different nodes and is also responsible for managing signaling links.

The *Signaling Connection Control Part* (SCCP) provides additional functions to MTP Level 3, supporting both connectionless and connection-oriented services to transfer signaling information between nodes (such as MSC, RNCs, exchanges, and specialized centers).[2] It controls logical signaling connections in an SS7 network, allowing signaling messages to be routed to a signaling point based upon, for example, the dialed digits.

The transport, session, and presentation layers are not used in the access or core network. The application layer performs call controls and mobility management, and manages *radio resources* (RR) such as radio channels, spreading codes, scrambling codes, and so on. CC and MM messages originate at an MSC and terminate on an MS, and vice versa. Thus, they are carried transparently over the access network.

The RR protocol manages physical channels (that is, radio connections) and specifies procedures for broadcasting system parameters, setting up radio connections between the network and an MS, reporting signal-level measurements and performing a handoff when a call is in the active state, and releasing all dedicated channels at the end of a call. The system parameters include the network identifiers, control channels to be used for signaling purposes within the system, options supported by the system, cell selection parameters, and so on. The protocol responsible for providing RR messages between an MSC and a BSC is the *Base Station System Mobile Application Part* (BSSMAP).

The protocol stack on the *A-bis interface* consists of the physical layer, the data link layer, and the application layer. It does not have any network, transport, session, or presentation layer protocol. The

[2]Specialized centers may be databases that translate special numbers or contain information to invoke an *Advanced Intelligent Network* (AIN) function.

signaling links on this interface operate at 64 kb/s at the physical layer, and use the *Integrated Services Digital Network* (ISDN) *Link Access Procedure on the D channel* (LAPD) protocol at the link layer [9]. The application layer protocols for CC and MM are the same as those of the A interface. RR messages between a BTS and BSC are provided by the *Radio Subsystem Management* (RSM) protocol.

The protocol stack on the *Um interface*, like the A-bis interface, consists of the physical layer, the data link layer, and the application layer only. The physical layer is the GSM air interface. The link layer protocol is similar to the ISDN LAPD except that it now uses the framing for the air interface. The application layer protocols for CC and MM are the same as those of the A interface. RR messages between a BTS and MS are provided by the *Radio Interface Layer* (RIL) 3, while those between the MSC and an MS are supported by the *Data Transfer Application Part* (DTAP) protocol.

The lower three layers of the protocol stack at the interface between MSC/VLR and HLR (that is, the C and D interfaces) are the same as for the A interface. Recall that an HLR contains a database of home subscribers such as their identification numbers, their service profile, and their present location. A VLR contains a database of all subscribers who are currently visiting this system, and includes the identification number and current location of the visiting subscribers and parameters used in the authentication process. The *Mobile Application Part* (MAP) is an application layer protocol [9], and interfaces the *Transaction Capabilities Application Part* (TCAP) [12]. The term *transaction capabilities* refers to functions provided by protocols that lie above the network layer and extend to the application layer. These functions are not application specific, but are rather common to applications. The TCAP protocol performs transaction-based services between node pairs. To execute a transaction, the participating nodes exchange data using a query and response mechanism. MAP messages are not CC messages, but are nevertheless required for the successful completion of a call. For example, when an MSC detects that a roaming mobile has registered in its serving area, its VLR sends a MAP message (that is, a Registration Notification message) to the HLR of the visiting mobile, and must receive all relevant information from the HLR database before this MSC can provide service to the roaming mobile.

The signaling protocol for the interface between a base station and an MSC (that is, the A interface) in IS-136, IS-95, and cdma2000 also uses MTP and SCCP in the lower three layers. Messages on this interface are specified by IS-634 [15], [13]. The intersystem operations (that is, operations required by handoff, automatic roaming, and so on) are specified in IS-41. For a description of these protocols, see Reference [14].

UMTS

The UTRAN protocol model [1], [2] was discussed in Chapter 6, "Universal Mobile Telecommunications System (UMTS)" and is redrawn in Figure 8-2 for ease of reference.

The RNC controls radio resources of the access network through one or more BTSs (also called *node B*). An RNC usually connects to multiple BTSs on one side and to an MSC and a *serving GPRS support node* (SGSN) (not shown in this figure) on the other side. A UTRAN may consist of one or more RNCs. An RNC, with its associated BTSs, forms a *radio network subsystem* (RNS).

Protocols at Uu and Iub Reference Points The UTRAN protocol stack is conceptually similar to the GSM protocol stack. The *Radio Resource Control* (RRC) protocol is a layer 3 protocol in the signaling plane that provides management of radio resources. Above it is the application layer protocol for CC and MM functions. The *Packet Data Convergence Protocol* (PDCP), *Broadcast/Multicast Control* (BMC), *Radio Link Control* (RLC), and *Media Access Control* (MAC) are link layer protocols, and were discussed in some detail in connection with UMTS. Recall that the RLC protocol supports three modes of data transfer: transparent, unacknowledged, and acknowledged. The physical layer of the Iub interface may operate at different data rates, such as E1, *Synchronous Optical Network* (SONET), and so on, over coaxial cables or fiber optics, among other things.

The protocol stack at the Uu interface is similar to the one shown in Figure 8-2 with some differences. For example, its physical layer is the W-CDMA air interface. As discussed in Chapter 6, the physical layer functions, which are provided by a BTS, include channel

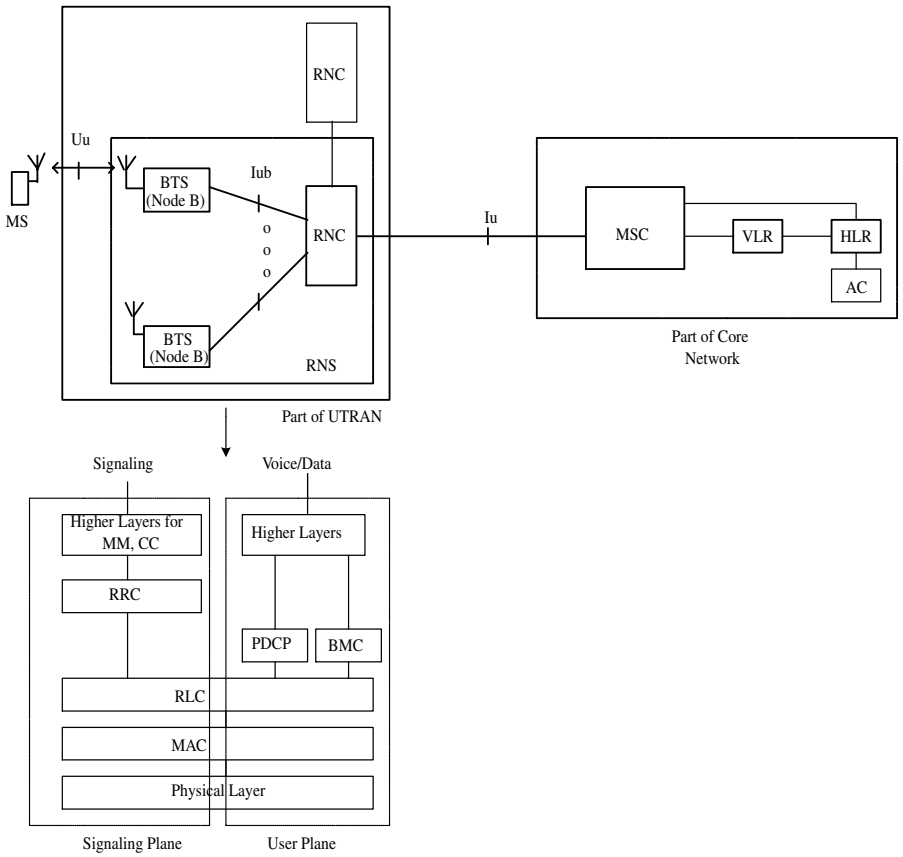


Figure 8-2
The UTRAN protocol stack

coding, interleaving, rate matching, spreading, scrambling, and modulation at the transmitter and inverse functions at the receiver. The RRC functions at a BTS may also be somewhat different from those of an RNC. As an example, a BTS is required to provide an inner loop power control. As in the case of GSM, the CC and mobility functions are not terminated by a BTS or RNC.

Protocol Stack at Iu Reference Point The Iu interface is located between the UTRAN and the core network as shown in Figure 8-2. For a general description of this interface, see References [3], [5]. *Asynchronous Transfer Mode* (ATM) is used over the transport network to transfer signaling and user information between an MSC

and RNCs. Because the UMTS supports *circuit-switched* (CS), *packet-switched* (PS), and *broadcast* (BC) services, the core network may be thought of as being composed of three logical domains: CS, PS, and BC. Clearly, to provide these services, separate signaling and user data connections are required over the Iu interface to each of these domains. In what follows we will only describe the protocol architectures that govern the signaling and user data connections to the CS and PS domains.

In the protocol description, use is made of the control plane and user plane. The control plane is concerned with the establishment, maintenance, and release of links for signaling and control purposes. The user plane, on the other hand, is associated with the exchange of user data such as voice or data. In the same way, the transport network makes use of a control plane for signaling information and a user plane for transferring user data across the transport network. In this description, we will only be concerned with the transport network user plane protocols. For a description of the transport network control plane protocols, see Reference [7].

The *signaling protocol* stack at the Iu interface for *circuit-switched services* is shown in Figure 8-3(a). Signaling messages are specified by RANAP [6], which is a user function of SCCP. SCCP provides signaling connections between an RNC and the CS domain of the core network. Both connectionless and connection-oriented procedures are provided, and a separate connection is required between an RNC and the core network for each active *user equipment* (UE).

Message Transfer Part Level 3 of Broadband (MTP3-B) SS7 is responsible for routing and distributing a message and the management of signaling links so as to ensure proper load sharing.

The next three layers—the *Service Specific Coordination Function* (SCCF), the *Service Specific Connection-Oriented Protocol* (SSCOP) and the *ATM Adaptation Layer 5* (AAL-5)—constitute the *Signaling ATM Adaptation Layer-Network Node Interface* (SAAL-NNI). Because the application layer protocols are service-specific, these three layers adapt the requirements of the upper layers to those of the ATM layer. Thus, for example, the SSCF layer processes the packets from the layer above it so that they become suitable for transmission over the layer below. The next layer, the SSCOP, provides for connection management, such as setting up and discon-

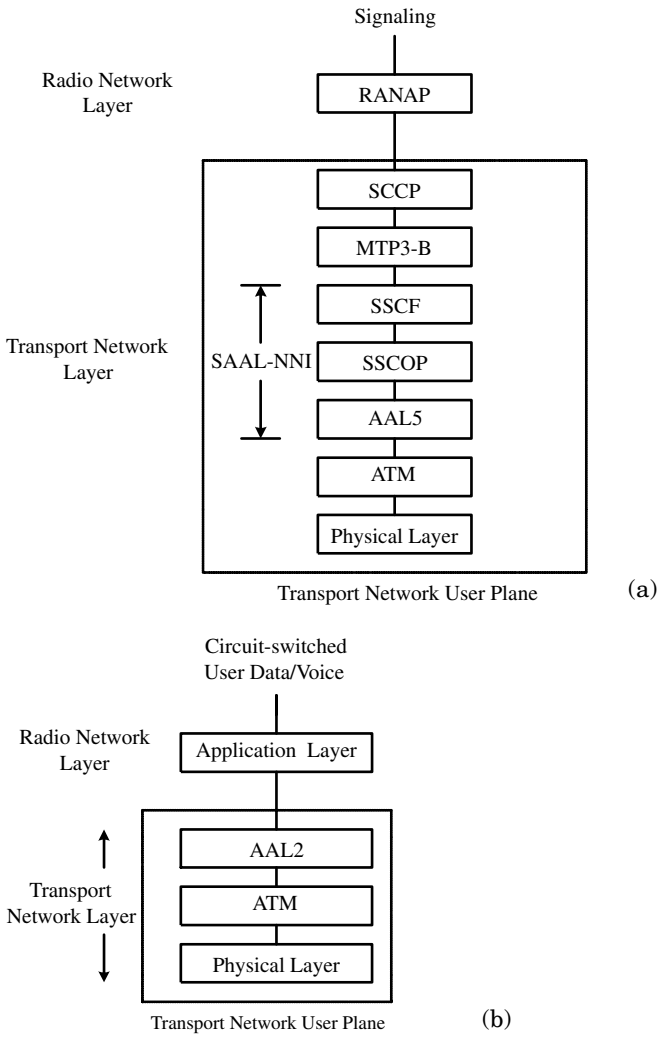


Figure 8-3
The protocol stack for the Iu interface in UMTS supporting circuit-switched services. (a) Signaling plane. (b) User plane.

necting a call, and allows for a reliable exchange of signaling messages. AAL-5 may be used for connectionless or connection-oriented services that do not require any timing information to be sent from the source to the destination. Two operational procedures are available: assured and nonassured. In the first, each *AAL service data unit* (AAL-SDU) is delivered to its destination error-free using

retransmission and flow control if necessary. In the second procedure, there is no guarantee that a given AAL-SDU will be delivered correctly.

The *user plane protocol stack for a CS-domain* is shown in Figure 8-3(b). The *ATM Adaptation Layer Type 2* (AAL2) performs circuit emulation of connection-oriented services. It allows the timing information to be transferred from the source to the destination so that the user data can be delivered at precise instants as required by the application (such as voice), and provides a means for tracking lost and misinserted cells. If the incoming data blocks are large, the *segmentation and reassembly* (SAR) sublayer of AAL-2 segments each block into a number of smaller blocks so that they can fit into ATM cells and sends them out in a sequence [13]. The physical layer for user data may operate at different rates such as 1.5, 2.0, 51, 155, 622 Mb/s, and so on over fiber optics or coaxial cables. The signaling protocols to establish or tear down AAL-2 connections over the transport network are provided by the *access link control applications part* (ALCAP), and may be found in Reference [7].

The protocol stack on the Iu interface point for packet-switched services is shown Figure 8-4. The signaling protocol model, which is shown in Figure 8-4(a) is similar to that of Figure 8-3(a). M3UA is the SS7 Message Transfer Part 3—User Adaptation Layer. The *Stream Control Transmission Protocol* (SCTP) permits different signaling protocols to be transported over IP networks.

The IP protocol in the control and user planes is actually IP over ATM, where an IP packet is encapsulated with an 8-octet header before it is sent out over ATM. The header is constructed so as to uniquely identify each protocol type (such as IP, ATM Address Resolution Protocol, and so on). Consequently, a receiving ATM device can forward each incoming packet to the proper entity (such as an IP router) based on this header alone. Furthermore, it is now possible to use a single ATM *virtual circuit* (VC) for different applications. For a detailed description of IP over ATM, see [13].

The *GPRS Tunneling Protocol* (GTP-U) is used to transfer IP datagrams between two GTP-U endpoints in the user plane. The GTP-U tunnels are set up and torn down using RANAP. The concept of a tunneling protocol was introduced in connection with GPRS in

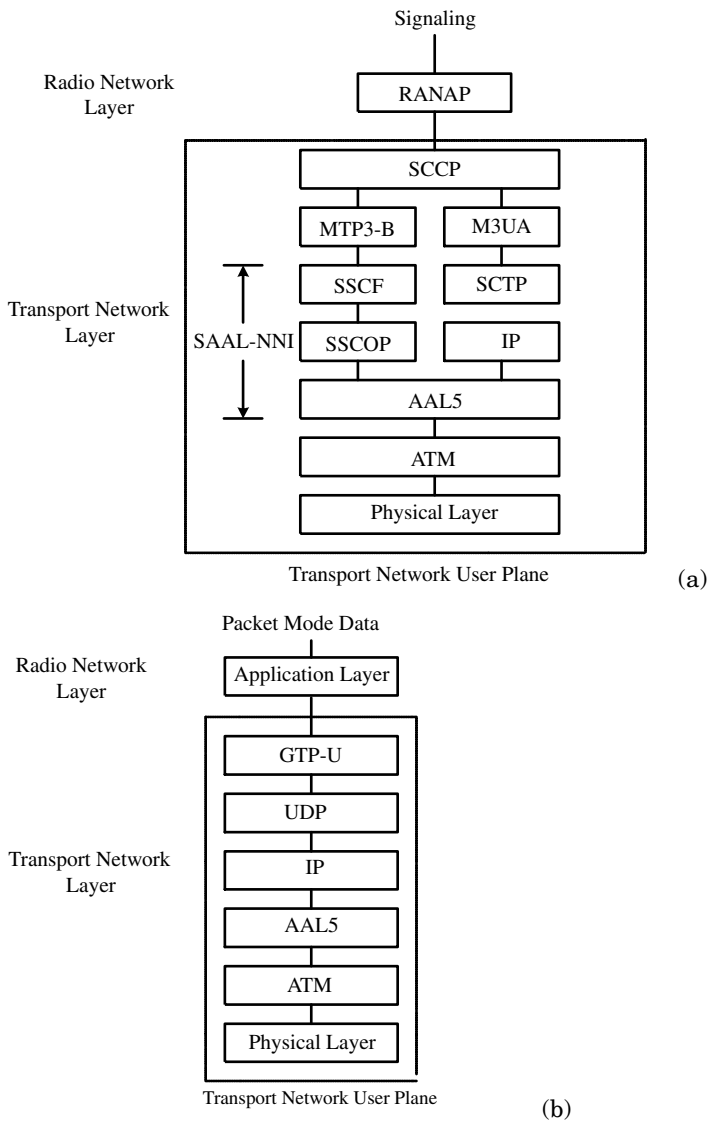


Figure 8-4
The protocol stack for the Iu interface in UMTS supporting packet-switched services. (a) Signaling plane. (b) User plane.

Chapter 5, "The GSM System and General Packet Radio Service (GPRS)." No signaling protocol is required in the transport network control plane over the Iu interface to the PS domain.

RANAP [6] RANAP provides signaling between a UTRAN and a *core network* (CN), and supports both connection-oriented and connectionless transfers of user data. For connection-oriented services, signaling links are established or removed dynamically. RANAP is notified if any of these connections fail at any time. Some of the functions performed by RANAP are listed here:

- *Management of radio access bearers* (RAB) RANAP provides for the establishment, maintenance, and release of RABs. The term RAB is used to indicate the service provided by the UTRAN when user data is to be transferred between a UE and a CN. For example, it may include physical channels over the air interface, logical channels for packet mode data, etc. Although the overall responsibility for managing RABs and signaling connections resides with the CN, an RNC may request their release at any time.
- *Relocation of a serving RNC* As an MS moves from the domain of one RNC to another, or from one serving area to another, the MS is to be handed over to a new RNC (or another system), and so radio resources must be reassigned. To maintain the continuity of service, it may be necessary to establish some new signaling connections or tear down some old ones. Similarly, RABs may have to be added or deleted, depending upon the new location of the UE.
- *Resetting of Iu interface* RANAP may reset an Iu interface under certain conditions.
- *Load balancing of an Iu interface* RANAP provides procedures for controlling the load on an Iu interface so that it remains within its prescribed limits.
- *Paging* RANAP provides for paging a UE.
- *Transferring non-access stratum* (NAS) *signaling messages between a CN and a UE* NAS messages are those messages that are exchanged between a CN and a UE transparently to the UTRAN. In other words, these messages do not terminate on the UTRAN. A signaling message may be sent to set up a new connection. On the other hand, a signaling message could also be sent over an already existing connection.

- *Location reporting from an RNC to a CN* RANAP allows an RNC to report the location information of a mobile station to a CN. Similarly, it includes procedures to activate or deactivate location reporting.
- *Security functions* RANAP supports transmission of encryption keys that provide protection against eavesdropping. Similarly, there are procedures in the protocol for enabling or disabling the security mode.
- *Reporting error conditions* The protocol includes procedures for reporting general error conditions. As an example, when the system fails to transmit a data segment associated with an RAB, the likely cause for failure may be reported to the management layer.

RANAP defines a set of elementary procedures that can be used to realize any of the previous functions. When a source end sends a message requesting the receiving end to perform a specific function, the receiver executes the command and reports the result to the source. To illustrate the general ideas, suppose that the CN requests an RNC to assign an RAB. On receiving the message, the RNC attempts to configure the requested RAB if it is available and sends an RAB assignment response message that includes the following information:

- RABs that have been successfully connected
- RABs that have been successfully released
- RABs that have been queued
- RABs that the RNC has failed to configure
- RABs that the RNC has failed to release

The exchange of messages is shown in Figure 8-5. The figure assumes that the requested operation was successful. If, however, the RNC fails to configure the requested RAB, it includes in the response message as precise a cause of failure as possible. For example, the cause may be "Invalid RAB Parameter Value," "Requested Maximum Bit Rate Not Available," "Requested Transfer Delay Not Achievable," and so on.

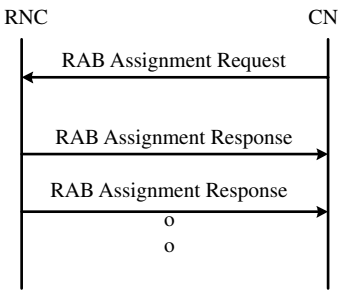


Figure 8-5
A typical RANAP procedure. Here, the CN requests the RNC to assign an RAB for a particular application. The RNC successfully assigns the RAB.

A relocation or handoff is handled by RANAP in the following way. When the source RNC (that is, the presently serving RNS) determines that a handoff is necessary, it sends a RELOCATION REQUIRED message to the CN. If it is an intrasystem handoff, in other words, if the mobile is merely being transferred from one RNC to another within the same core network, the source RNC includes in the message its own ID as well as the ID of the target RNC. If it is an intersystem handoff, that is, if the relocation involves another CN, the message must indicate the identifier of the current service area as well as the identifier of the cell in the target system. The message also provides a cause value for the handoff. For example, this cause value may be "Time Critical Relocation," "Resource Optimization Relocation," or "Desirable Relocation for Radio Reasons," and so on. The message may also contain other information such as the number of Iu signaling connections to the UE and so on.

On receiving the RELOCATION REQUIRED message, the CN sends a RELOCATION REQUEST message to the target RNC (or target CN), requesting it to schedule necessary resources required by the source RNC. If the target RNC (or target CN) is able to support the requested service, it sends a RELOCATION REQUEST ACKNOWLEDGE message to the CN, indicating that necessary resources in the target RNC have been prepared.

On receiving the acknowledgment from the target RNC, the CN sends a RELOCATION COMMAND to the source RNC. If the target RNC does not support all the RABs that are required by the UE, the CN may include in the RELOCATION COMMAND a list of all the

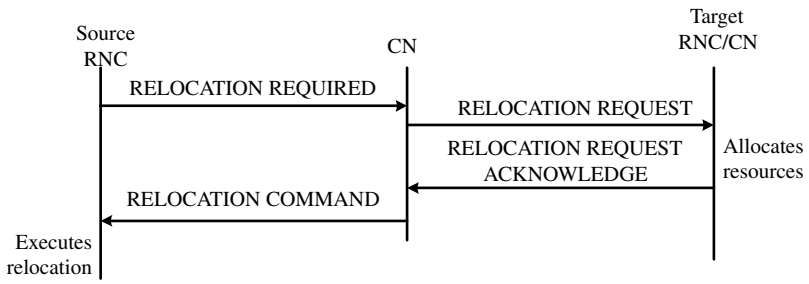


Figure 8-6 Relocation procedure

RABs that are not going to be supported by the target RNC. At this time, the source RNC may actually handoff the UE to the target RNC.

If the CN or the target RNC is unable to honor the relocation request from the source RNC, the CN sends a RELOCATION PREPARATION FAILURE message that includes the cause for the failure. The failure may be due to the target RNC or the target CN not being able to support the relocation, or the source CN not receiving any response from the target RNC/CN within a specific timeout period. These message flows are depicted in Figure 8-6.

Call Controls

Call controls in different systems are conceptually similar. Suppose that an MS is visiting a new serving area. Before the subscriber can initiate or receive a call, it must register with the new system. In this case, the following sequence of events takes place:

1. To request or receive the service from an MSC, an MS must register with the system by sending its identification number. When an MS has moved into a new area, it may autonomously register with that system. Alternatively, if the mobile originates a call after moving into the new area before the registration process begins, the VLR may ask the MS to register. In either case, after the registration process is complete, the VLR of the new serving area has the identity of the mobile subscriber.

The MS may use the following procedure to determine that it has moved into a new location area. The MS fetches from its memory the identifier of the location area in which it was last registered, and compares it with the corresponding information that is being broadcast by the system along with other system parameters. If the two do not match, the MS knows that it is visiting a foreign serving area, and initiates an autonomous registration.

2. This VLR notifies the HLR of the visiting subscriber about the registration that just took place, whereupon the HLR updates the present location of the subscriber so that if there is an incoming call to this mobile, the HLR knows how to route the call.
3. The HLR also sends a registration cancellation message to the VLR of the area that this subscriber had last visited. This latter VLR may then request its associated MSC to delete all information about this subscriber from its memory.
4. The VLR of the new serving area may send a qualification request message to the HLR to determine if the visiting subscriber is authorized to receive the service. In response, the HLR checks its database, and sends the required information along with other optional, relevant parameters.
5. Notice that the VLR of the new serving area does not have the database of the MS yet, and so requests the HLR to send the profile of the MS. When it receives the requested information, it saves that information in its database. The visited area is now ready to serve the MS.

The message flow that takes place during the registration phase is shown in Figure 8-7. These messages are specified by the MAP protocol [14]. There are other MAP messages that perform different functions. For example, an MSC may send a message to another MSC, requesting it to take measurements for a required handoff. Or, it may send a location request message to an HLR, seeking information about the current location of an MS, and so on. Message formats are specified by TCAP. A TCAP message is initiated by a TCAP invoke, which is followed by a TCAP result. When an entity, such as

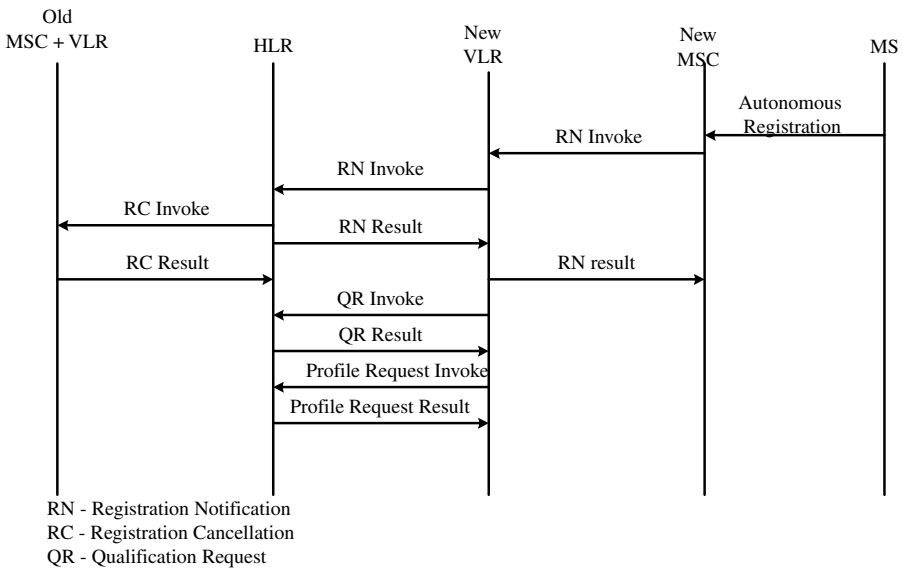


Figure 8-7 The registration process invoked when an MS visits a new serving area

an HLR or VLR, receives this message, it executes the task specified or implied by the message and sends the result to the source.

Many different call scenarios are possible. For example, there may be a land-originated call to an idle MS in a home system or in a visited system. In a second scenario, a call comes to an MS that has an unconditional call-forwarding feature activated. Or, a call is delivered to an MS that is inactive while visiting a service area, and so on. In what follows, we will only illustrate the sequence of events that take place when a land-originated call is directed to an idle mobile station in a visited system. The message flow is shown in Figure 8-8.

1. A land-originated call destined to an MS comes to the MSC of the home location, that is, the originating MSC. Because the HLR contains the information on the current location of the MS, this MSC sends a location request to the HLR, asking for the present location of the (roaming) mobile.
2. On receiving the location request, the HLR retrieves the location information from its database and sends a routing request to the VLR of the visited system, asking how the call may be best routed to the called MS.

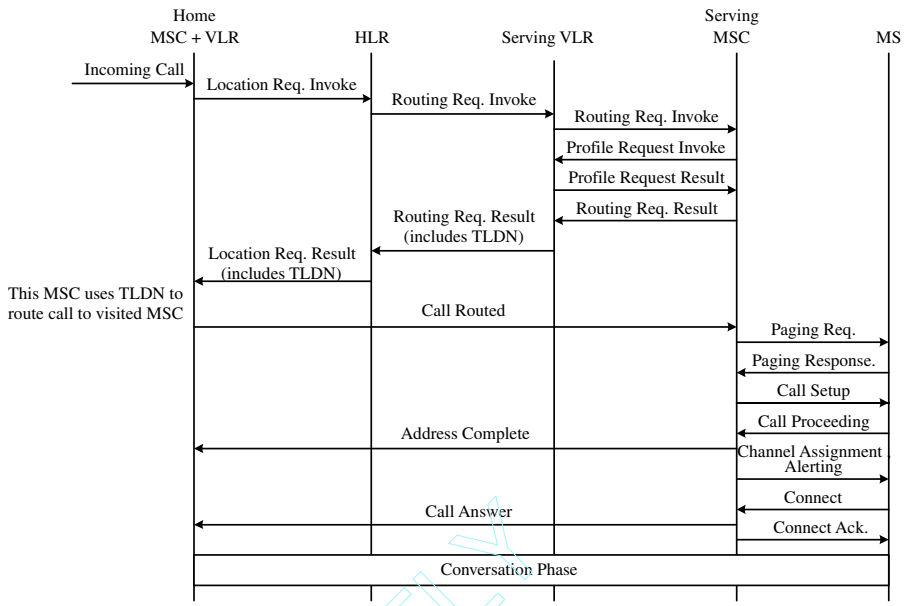


Figure 8-8
A land-originated call to a roaming mobile

TEAMFLY

3. When the VLR of the visited area receives the routing request, it forwards the routing request to the MSC of the visited area, that is, the serving MSC.[2]
4. On receiving the routing request, the MSC may request the VLR asking for the subscriber profile. Recall that the VLR has acquired the profile information of the visiting mobile from its HLR during the registration process.
5. The serving MSC assigns a *temporary local directory number* (TLDN) or equivalently a *temporary mobile subscriber identity* (TMSI), constructs the location response, and sends it to the originating MSC. This temporary identity is used by the originating MSC to route the incoming call to the serving MSC.
6. The serving MSC sends a paging message to the desired base station that is providing the coverage to the MS.

[2]The routing information is usually held in the MSC.

7. The base station pages the mobile.
8. The MS sends a page response message to the base station.
9. The serving MSC presents a call setup message to the mobile, which confirms it by sending a call proceeding message, whereupon the serving MSC sends an SS7 address complete to the originating MSC.
10. The serving MSC requests the base station to assign a traffic channel.
11. The MS sends an acknowledgment to the base station, which relays it to the serving MSC, indicating that the channel assignment is completed (not shown).
12. The mobile is alerted.
13. The mobile station goes off-hook and sends a connect message to the base station indicating that a connection has been established. This message is relayed to the serving MSC, which then replies back with an SS7 answer message. The conversation may now begin.

Summary

A general, high-level concept of call controls and mobility management in wireless networks has been presented in this chapter. Call controls are concerned with signaling procedures required to establish or tear down a call. MM refers to location updates and location reporting, registration of MSs, and authentication. The protocol stacks at a few reference points in the access and core networks have been briefly described. RANAP provides signaling between a UTRAN and the core network. Functions performed by this protocol and messages required to assign radio channels when a call is initiated or when an MS visits another system are described. The chapter concludes with some simple call control scenarios.

References

[1] 3G TS 25.301, "Radio Interface Protocol Architecture," 1999.

[2] ETSI TS 125 401 (3GPP TS 25.401 Version 3.5.0), UTRAN Overall Description, 2000.

[3] ETSI TS 125 410 (3GPP TS 25.410 Version 3.3.0), UTRAN Iu Interface, General Aspects and Principles, 2000.

[4] ETSI TS 125 411 (3GPP TS 25.411 Version 3.3.0), UTRAN Iu Interface Layer 1, 2000.

[5] ETSI TS 125 412 (3GPP TS 25.412 Version 3.6.0), UTRAN Iu Interface Signaling Transport, 2000.

[6] ETSI TS 125 413 (3GPP TS 25.413 Version 3.4.0), UTRAN Iu Interface RANAP Signaling, 2000.

[7] ETSI TS 125 414 (3GPP TS 25.414 Version 3.6.0), UTRAN Iu Interface Data Transport and Transport Signaling, 2000.

[8] ETSI TS 125 415 (3GPP TS 25.415 Version 3.5.0), UTRAN Iu Interface User Plane Protocols, 2000.

[9] M. Pautet and M. Mouly, "GSM Protocol Architecture: Radio Sub-System Signaling," *IEEE Veh. Technol. Conf., 1991.*

[10] ANSI T1.111-1988: Signaling System No. 7 (SS7)—Message Transfer Part (MTP).

[11] ANSI T1.112-1988: Signaling System No. 7 (SS7)—Signaling Connection Control Part (SCCP).

[12] ANSI T1.114-1988: Signaling System No. 7 (SS7)—Transaction Capabilities Application Part (TCAP).

[13] M.R. Karim, *ATM Technology and Services Deliver.* New Jersey: Prentice Hall, 2000.

[14] EIA/TIA IS-41.1–41.5, Cellular Radio-Telecommunications Intersystem Operations, December 1991.

[15] TIA IS-634, MSC-BS Interface for 800 MHz, 1995.

CHAPTER 9

Quality of Service (QoS) in 3G Systems

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

Introduction

The Internet was originally designed for nonreal-time data services such as interactive burst or interactive bulk transfer. In these applications, there are no requirements on the maximum amount of delays that a packet may encounter during its transit to the destination. Similarly, bandwidths required by an end user are never specified. As such, the network accepts all incoming packets without using any admission control mechanism, forwards them using a simple, first-come-first-served algorithm, and delivers them on a best-effort basis.[1] Thus, issues concerning the *quality of service* (QoS) delivered to an end user are rather straightforward. The QoS in present-day mobile IP is also minimal because, once again, data is delivered using the best-effort scheme. With the emergence of real-time multimedia services as envisaged by *third-generation* (3G) wireless systems, new QoS requirements are imposed on the networks. For example, with interactive video conferencing or streaming video and audio, the network must be able to deliver these services to the destination on a timely basis. Because flow control or retransmission is not possible for these applications, the bit error rate or packet loss ratio must be kept below a certain level; otherwise, the QoS may suffer. For instance, if the bit error rate is too high, the video in an MPEG application may never synchronize at a receiver.[2]

The *Internet Engineering Task Force* (IETF) is developing standards that provide QoS in an IP network. To this end, it has defined two models. One of them, called *integrated services* (IntServ), allows a receiving terminal (or host) to reserve network resources along a route to the sender [3]. The traffic coming into a node *from each user* is classified on the basis of its characteristics, and resources are reserved explicitly on an end-to-end basis for each flow (that is, a sequence of packets between any sending and receiving application).

---

[1]In other words, the network delivers as many packets with as little delays as possible within the constraint of its resources.

[2]Similarly, the IP and *User Datagram Protocol* (UDP) alone are not adequate for these real-time applications. The *Real-time Transport Protocol* (RTP) [6], which works above the UDP layer and allows for the identification of payload types, sequence-numbering, time-stamping, and so on, has been designed specifically for these services.

This reservation is made using the *Resource Reservation Protocol* (RSVP) defined by the IETF for real-time services over virtual circuits [1], [2]. The other model is known as *differentiated services* (DiffServ). It divides incoming traffic (*from any user*) into a few classes [14]. For example, one class of traffic may require the network to assign the associated packets the highest priority and therefore forward them first. Another traffic class may be such that associated packets can wait a while before they are forwarded to the next hop, but in the event of congestion, they must be dropped last and so on. In DiffServ, unlike IntServ, the receiving end point does not make any explicit reservation of network resources.

QoS concepts and issues as they relate to ATM, frame relay, and IP-based networks have been extensively studied by many authors [6], [7]. Concepts and QoS architectures germane to 3G cellular networks have been published in [9], [10]. In providing QoS, we are only concerned with the UTRAN (that is, the radio link) and the core network (that is, the routers), and not the entire networks from one end to the other. In other words, referring to Figure 9-1, the QoS concepts and procedures developed by 3G only apply to the UMTS radio bearer service. The objective of this chapter is to provide the reader with a basic understanding of the subject.

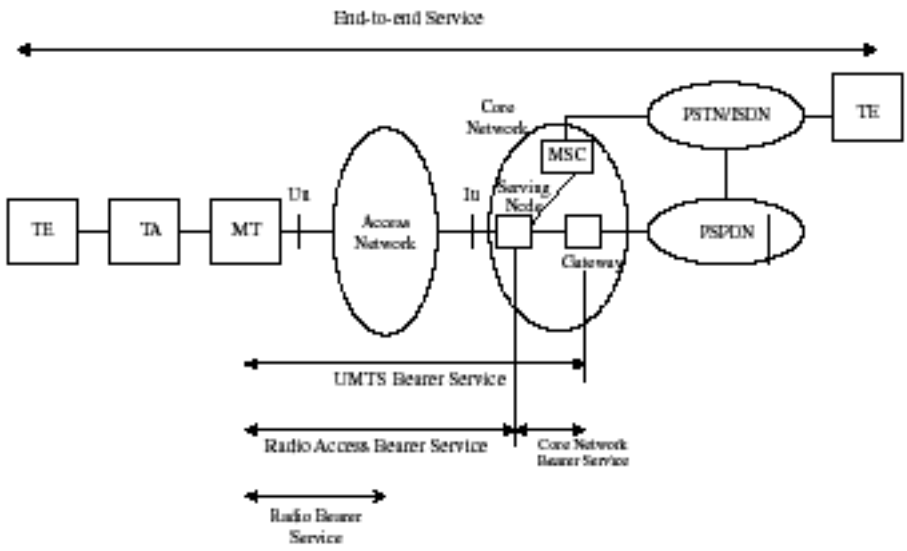


Figure 9-1
Provision of QoS in 3G networks

The organization of this chapter is as follows. It begins with an overview of how QoS is usually implemented following the IntServ model. To request and provide the desired QoS, it is necessary to classify the traffic emitted by a source and characterize the service requested by a user. These topics are discussed in sections 3 and 4, respectively. A brief description of RSVP is provided in section 5. Admission control and servicing strategies that ensure the requested QoS are presented in sections 6 and 7. Section 8 gives a short description of the DiffServ model and indicates how IntServ and DiffServ regions can be connected together in an attempt to combine the features of the two models. The chapter concludes with a discussion of how a 3G network can provide QoS to a mobile station as it is handed over from one cell to another in the same serving area, or from one serving area to another.

Overview of the Concepts

In the IntServ model, there are actually four aspects to the QoS issues. This is shown in Figure 9-2. First, the network must be designed so as to provide a means for the user application to request the QoS desired. Second, once the network receives the service request, it should be able to analyze the request, determine the network resources (such as the bandwidth, buffers, spreading codes, error-detecting codes, algorithms, and so on) required to provide the requested quality, and admit the user only if it is authorized to receive the service or request the QoS, and, at the same time, the network has enough resources. Third, the network must now set aside the required resources for the incoming user and mark packets that should receive the requested quality. Finally, the network must be able to police and enforce the service contract of each user by monitoring traffic rates, inform any user that violates its traffic contract, and take appropriate action to prevent congestion in the network so that the packet loss ratio is within the advertised limits of the system. If incoming packets meet their service contract, the network must forward them towards their destination, still guaranteeing the requested quality. Otherwise, it may send the packets on a best-effort basis.

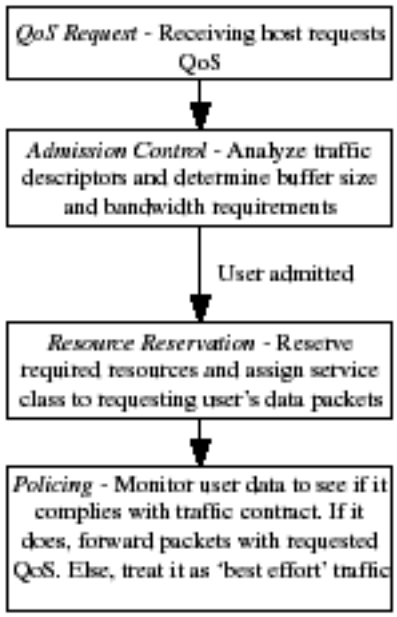


Figure 9-2
Functions involved in providing QoS

Classification of Traffic

One of the distinctive features of a 3G system is its capability to provide different services such as video conferencing, real-time process control and telemetry, streaming audio and video, high-speed data transfers, and so on. More specifically, it is required to support data rates of at least 384 kb/s in urban or suburban areas, 144 kb/s in rural areas, and up to 2.048 Mb/s in indoor or limited-range, outdoor environments. Because, in a general case, a mobile station may run several applications simultaneously, it is necessary to characterize the traffic in some meaningful way so that each application can request the desired QoS from the network in a straightforward manner.

One way to classify the user traffic is based upon how the network should assign its resources, say, bandwidth, to transport that traffic across the network. For example, some traffic, such as video telephony or *voice over IP* (VoIP), requires the bandwidth to be allocated at regular intervals, whereas no such regular allocation is necessary for nonreal-time, delay-insensitive data. On the basis of this requirement, then, the following types of traffic are possible:

- *Constant bit rate* (CBR) *traffic* Sensitive to delays, this type of traffic generates fixed-size packets on a periodic basis. Examples are speech, high-quality audio, video telephony, full-motion video, and so on. Here, each individual application knows the amount of bandwidth it requires for the duration of the call and may use it to request the QoS.
- *Real-time variable bit rate* (VBR) *traffic* This traffic generates variable-size packets on a periodic basis. Examples include variable bit-rate encoded audio, interactive video encoded into an MPEG standard, and so on. In this case, it is not possible for the application to know the exact bandwidth it needs. However, it may have the knowledge about the sustainable traffic rate, maximum traffic rate, and maximum traffic burst. The sustainable traffic rate may be defined as the maximum amount of traffic per unit time that the network has agreed to carry over time. The phrase *over time* is important because although the network can carry a much larger traffic load for a relatively short interval, the user traffic should be within this value most of the time. That being the case, the sustainable traffic rate can equal but never exceed the access rate, that is, the transmission rate of the physical medium. Similarly, if multiple logical channels are defined on a physical channel, the sum of the sustainable traffic rates of all these channels cannot exceed the access rate.

  The maximum traffic rate is not meaningful unless we also define the maximum burst size. As an example, suppose that the access rate on a physical channel is 64 kb/s (that is, the maximum bit rate), and the sustainable traffic rate is 32 kb/s. The user could also specify to the network that it would emit traffic at a rate of 64 kb/s for, say, 100 ms every second. At other times, it agrees to keep its traffic rate below 32 kb/s so that over time the sustainable traffic rate would be around 32 kb/s as subscribed by the user.[3] Because there may be many logical channels defined on a single physical channel, the sum of the peak traffic rates may well exceed the access rate. In this case,

[3]To be precise, the source would emit 25.6 kb over the next 0.9 s.

the network may congest during these peak traffic times and drop excess packets unless it has sufficient buffers to hold the incoming burst. Because the network has limited buffers, it is necessary for it to know the burst size so that it can allocate a buffer of appropriate size.

With the specification of these parameters, the network may allocate the maximum amount of bandwidth on a regular basis to ensure that packets of all sizes would be able to get through. Or, better yet, the network may measure the input traffic and, using past allocation, predict the required number of slots so that the frame error rate is within the limits specified by the user.

- *Nonreal-time variable bit rate traffic* This type of traffic can tolerate delays or delay variations. An example is an interactive and large file transfer service. The source may indicate to the network its minimum sustained traffic rate and the maximum tolerable delay between successive transfers. If the source does not specify any of these parameters, the network may, in the event of congestion, reduce its bandwidth allocation.

The traffic may also be classified according to the extent to which it can tolerate end-to-end delays and delay variations. Based on this criterion, Reference [9] lists the following types of traffic:

- *Real-time conversational traffic* This real-time traffic is bi-directional, involving human users at the two ends of a communication link, and is characterized by low end-to-end delays. Since the perceptual quality of the received signal greatly depends on how well the silent or inactive periods between adjacent information entities are preserved, delay variations should also be kept very small (about 1 ms or less). Applications that fall in this category are conversational voice, video phone, and video conferencing. Also belonging to this category, but with somewhat less stringent requirements on the transfer delay, are interactive games and two-way process control and telemetry information.
  - *Interactive traffic* This class of traffic, which involves man and machine, is based on a request/response from end-points.

It is nonreal-time and may be unidirectional or bidirectional. Examples are web browsing, e-mail, data transfer to or from a server, transaction services (that is, e-commerce), and so on. Because applications generating this traffic are nonreal-time, delays may be longer but usually have an upper limit. However, payload contents of *protocol data units* (PDUs) must be transferred without any modification and with low bit error rates. Delay variations must be kept very small (1 ms or less).

- *Streaming traffic* This traffic is associated with real-time applications. However, unlike the conversational type, it is unidirectional (between man and machine) and has a somewhat more continuous flow with fewer and shorter inactive/silent periods between information entities. Because the traffic is unidirectional, end-to-end transfer delays may be large as long as their variations are small (1 ms or less). Examples are audio streaming, one-way video, still images, large-volume data transfers, and telemetering information for monitoring purposes at an operations and maintenance center.
- *Background traffic* As the name implies, the data transfer for this kind of traffic takes place in the background only when the computer has some real-time left after finishing high-priority tasks. Because there is no requirement on when it should arrive at the destination, permissible delays or delay variations are not specified. Bit error rates, however, must be very low and PDU contents must be preserved. Applications giving rise to this traffic have usually low priorities. For example, the *short messaging service* (SMS) in GSM, or the delivery of e-mails from one server to another, falls in this category.

UMTS Service Attributes

In 3G, the term *service attributes* means services provided by the network. As such, the traffic type may be considered a service attribute. For example, when requesting a QoS, a mobile station may specify the traffic class to be conversational voice. The network may then use it as a basis for scheduling necessary resources to meet the

quality of that class of traffic. Some of the service attributes in 3G are listed below [10]:

- *Guaranteed bit rate* It is given by the number of bits guaranteed to be delivered by the UMTS per unit time over the duration of a call. It must be at least equal to the sustainable bit rate, which is equal to the number of bits/second averaged over the life of a call.
- *Burst size* A burst is said to have occurred when two or more packets appear in quick succession such that the associated minimum interarrival times are less than the packet arrival time averaged over the duration of a call. The number of bits in the burst is called the *burst size*. The term *maximum service data unit (SDU) size* used in the UMTS literature may be considered as equivalent to the burst size. Specification of the SDU format indicating possible exact sizes of SDUs enables the UTRAN to operate in the transparent *radio link control* (RLC) protocol mode.
- *Maximum bit rate* It is defined as the number of bits in a burst averaged over the burst duration. The network may use this parameter to select the type of coding and coding rate on a downlink channel on the radio interface.

  For the definition of these parameters, see Figure 9-3. Assuming that we have $t_1 < t_2 < t_3 < t_4$, the four packets appearing in time period T_b form a burst because their interarrival time t_1 is the shortest. If T, the duration of the call, is large compared to t_1, ..., t_4 and T_b, and b_1, b_2, ..., b_7 are the packet size in bits, then the sustainable bit rate for this example is $(b_1 + ... + b_7)/T$ bits/s. The burst size is $b_1 + b_2 + b_3 + b_4$, and the peak traffic rate is $(b_1 + b_2 + b_3 + b_4)/T_b$.
- *Delays and delay variations* A packet may encounter delays at a number of points as it travels from the source to the destination. They include:
  - Propagation delays.
  - Delays due to buffering at an input or output port.
  - Serialization delay.

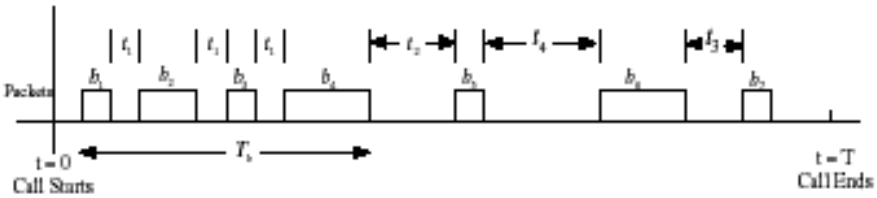


Figure 9-3
Definition of sustainable traffic rate, burst size, and peak traffic rate

- Switching delays. This component increases with the number of nodes in a network.

Delays in packet-switched networks are not constant but vary because incoming packets arrive at input ports randomly and are processed according to some priority queuing schemes. Furthermore, each packet may be treated differently for forwarding purposes. For example, packets with the same destination address may be forwarded along different routes at different instants during the life of a call depending upon the priority of each packet and the incoming traffic volume at a node.

Real-time conversational traffic (such as voice, video telephony, and so on) is sensitive to delays and delay variations (that is, jitter). For instance, round-trip delays of 600 ms or more, even with echo cancellers, may cause confusion on the part of listeners and may even cause both parties to talk simultaneously. But delays in the range of 100 ms or so may not have any noticeable effect on the perceptual quality of speech. On the other hand, the human ear is extremely sensitive to delay variations. For example, if packets of length, say, 100 ms or so, are subjected to variable delays of 10 to 170 ms, speech becomes unintelligible.

As for the interactive video, it is also sensitive to delays and delay variations. For example, if delay variations exceed a few tens of milliseconds, the video may not synchronize at all at the receiving end. Because video telephony or, for that matter, one-way video also includes audio, and because the video and audio must always be synchronized, the allowable jitter for these applications is generally quite low.

3G system requirements specify that the maximum transfer delay must lie in the range of 20 to 300 ms with little or no jitter

for all real-time applications in urban or suburban outdoor environments at velocities up to 120 km/h, rural outdoor at speeds up to 500 km/h, and indoor or low-range outdoor at speeds up to about 10 km/h. For nonreal-time applications, delays may be 1200 ms or more. As for delay variations, they are subject to an upper limit of 1 ms. Once delays are specified, transport formats on various transport channels can be defined.

- *SDU error ratio* It is the ratio of SDUs either lost or received in error to the total number of SDUs. The UTRAN can use this parameter to select protocols, error-detecting codes, algorithms, and so on.
- *Residual bit error ratio* It is defined as the undetected bit error ratio in the delivered SDUs and may be used to select protocols and error-detecting codes.
- *Bit error rate* Actually, bit error rates are not considered to be service attributes in UMTS. However, an upper limit on this parameter is sometimes specified as a desired goal. For example, in a 3G system, the maximum bit error rate is specified to be $10^{-7} - 10^{-3}$ for real-time applications in urban, suburban, rural, indoor, and outdoor environments and $10^{-8} - 10^{-5}$ for nonreal-time applications.

  It is worthwhile to mention here that some applications, such as file transfers, e-mails, and so on, cannot tolerate any errors. Voice and video traffic, on the other hand, are much more tolerant of channel errors. For example, with some coding schemes (such as waveform quantizers), an acceptable speech quality is obtained even when the error rate is as high as 10^{-2}. As for video, bit error rates on the order of 10^{-8} or less have negligible or no effect. Error rates in the range of $10^{-8} - 10^{-3}$, if not corrected, will have a significant effect. If they are any higher, the system may not even operate at all.

  As indicated earlier, the recommended jitter for audio and video regardless of the traffic type in 3G is 1 ms or less. There is no specification of this parameter for data. Table 9-1 gives the one-way transfer delays for different media. The permissible frame error rates for audio, video, and data of various traffic types are presented in Table 9-2.

Table 9-1 One-way transfer delays of different media and examples of relevant applications

Medium	Examples of Applications	One-Way Transfer Delay
Conversational audio	Conversational voice	150–400 ms
Interactive audio	Voice messaging	1–2 s
Streaming audio	High-quality audio	<10 s
Conversational video	Video telephony	150–400 ms
Streaming video	One-way	<10 s
Conversational data	Two-way telemetry, process control, interactive games, and so on	250 ms
Interactive data	Web browsing, e-commerce, e-mail (from an end user to a local server), and so on	<4 s
Streaming data	High-volume data transfer, file transfer, still image, telemetry information, and so on	<10 s
Background	SMS, e-mails (from server to server), and so on	Not applicable

Table 9-2 Recommended frame errors for different media in 3G systems

Medium	Frame Error Rate
Conversational audio	<3%
Interactive audio	<3%
Streaming audio	<1%
Video (conversational or streaming)	<1%
Data	0%

Table 9-3

Suggested attributes for the four traffic types

Service Attributes	Conversational Type	Interactive Type	Streaming Type	Background
Maximum bit rate	x	x	x	x
Burst size	x	x	x	x
Guaranteed bit rate	x		x	
Delays	x		x	

In many instances, particularly in initial versions of 3G, a subset of the service attributes that we discussed earlier may be adequate. Table 9-3 shows these attributes for the four traffic types. Based on a meaningful combination of these attributes, the network may create a set of profiles. For example, the network may have two profiles for the conversational type of traffic. Profile 1 provides a maximum bit rate of 144 kb/s, a maximum SDU size of 456 bits, a guaranteed bit rate of 64 kb/s, and a transfer delay of 100 ms. Profile 2 supports a maximum bit rate of 384 kb/s, while its remaining attributes are the same as for profile 1.

Requesting QoS—RSVP Protocol

As we saw before, multimedia services in 3G systems are IP-based. For IP networks, IETF has defined a resource reservation protocol, called *RSVP*, that enables a receiving terminal (or host) to request the desired QoS. This protocol operates at the transport layer above IPv4 [4] or IPv6 [5]. However, it does not transport user data, but rather works like the *Internet Control Message Protocol* (ICMP).[4]

[4]ICMP handles error conditions that may arise when delivering an IP packet. For a description of ICMP, TCP, and UDP, see M. Naugle, *Network Protocol Handbook,* New York: McGraw-Hill, 1994.

Because future wireless networks are also likely to be based on IP, we will present a brief overview of RSVP here.

Before describing this protocol, it is necessary to define a few terms used in connection with RSVP. The term *flow* refers to a set of packets from the source to a destination during a session. Packets could come to a destination from many different sources; similarly, the same source could send packets to multiple destinations. In other words, many-to-many multicasting is possible in RSVP. A *template* specifies the format of data packets. Mechanisms that implement a QoS for a particular data flow are called *traffic control*. In RSVP, the following terms are called *objects*:

- *TSPEC* This object describes the traffic that is likely to be generated by an application at the source end, and indicates such things as the peak rate at which the sender expects to emit, its maximum packet size, the minimum packet size that can be monitored (by a router), and so on. The sender's TSPEC travels downstream (that is, towards the receiver) unchanged from one hop to the next until it reaches the receiver.
- *RSPEC* This is used by a receiving host to specify the requested level of service.
- *ADSPEC* This indicates the QoS requirements of the sender, expressed in terms of the bandwidth estimate, minimum path latency, and so on. Generated by the sending host, it travels downstream, and may be changed, if necessary, at an intermediate node.

  A reservation request includes two parameters—*flowspec* and *filterspec*, which together are known as a *flow descriptor*.
- *FLOWSPEC* This describes the QoS requested by the receiving host in terms of the token bucket rate, token bucket size, peak data rate (that the receiving end-point can support), minimum packet size, maximum packet size, and so on.[5] Upon receiving

[5]In this characterization of the traffic, the token rate, r, is the sustainable data rate over time. The bucket size, b, indicates a data burst that may exceed the sustainable rate for a short period compared to the duration of a call. Thus, for any period t during the life of a call, the amount of data is less than $r\,t + b$. These things are explained later in this chapter.

the sender's ADSPEC, the receiver may use it as a guide to choose these parameters in its resource reservation request. It flows upstream towards the sending host.

- *FILTERSPEC* This object specifies the group of packets that are to be specially treated by the network so that they get the requested QoS. The group of packets may be selected on the basis of some upper-layer protocol specification or on the basis of some header values. For example, we could say that it would be all those packets that are associated with the *Address Resolution Protocol* (ARP).[6]

RSVP defines a set of seven messages: PATH, RESV, PATH ERROR, RESERVATION ERROR, PATH TEAR, RESERVATION TEAR, and RESERVATION CONFIRMATION. The first two of these messages—PATH and RESV—are used in setting up a QoS reservation request.

The working of the protocol is described in the form of flow charts in Figure 9-4. Before the receiving host initiates a resource reservation request, the sending host must send a PATH message downstream towards the receiver that should include, among other things, the sender's template, TSPEC, and ADSPEC. As each router receives the PATH message, it may modify some parameters that are specific to it and then forward the message to the next hop. The PATH message follows exactly the same path as the data packets.

On receiving the PATH message, the receiving host constructs an RESV message, taking into consideration the QoS requirements of the sender as indicated in the PATH message as well as its own capabilities, and sends it upstream to the nearest router. The router checks to see if the requesting terminal has the administrative permission to make the reservation. If the answer is positive, the router analyzes PATH and RESV messages and determines if it has enough resources to meet the desired QoS. If it does, it assigns a service class to the user's packets corresponding to the requested QoS, reserves necessary resources by setting some parameters in the scheduler (such as the bandwidth on the outgoing interface, the right buffer size, priorities with which different queues are serviced, and so on),

[6]See the reference in footnote 4.

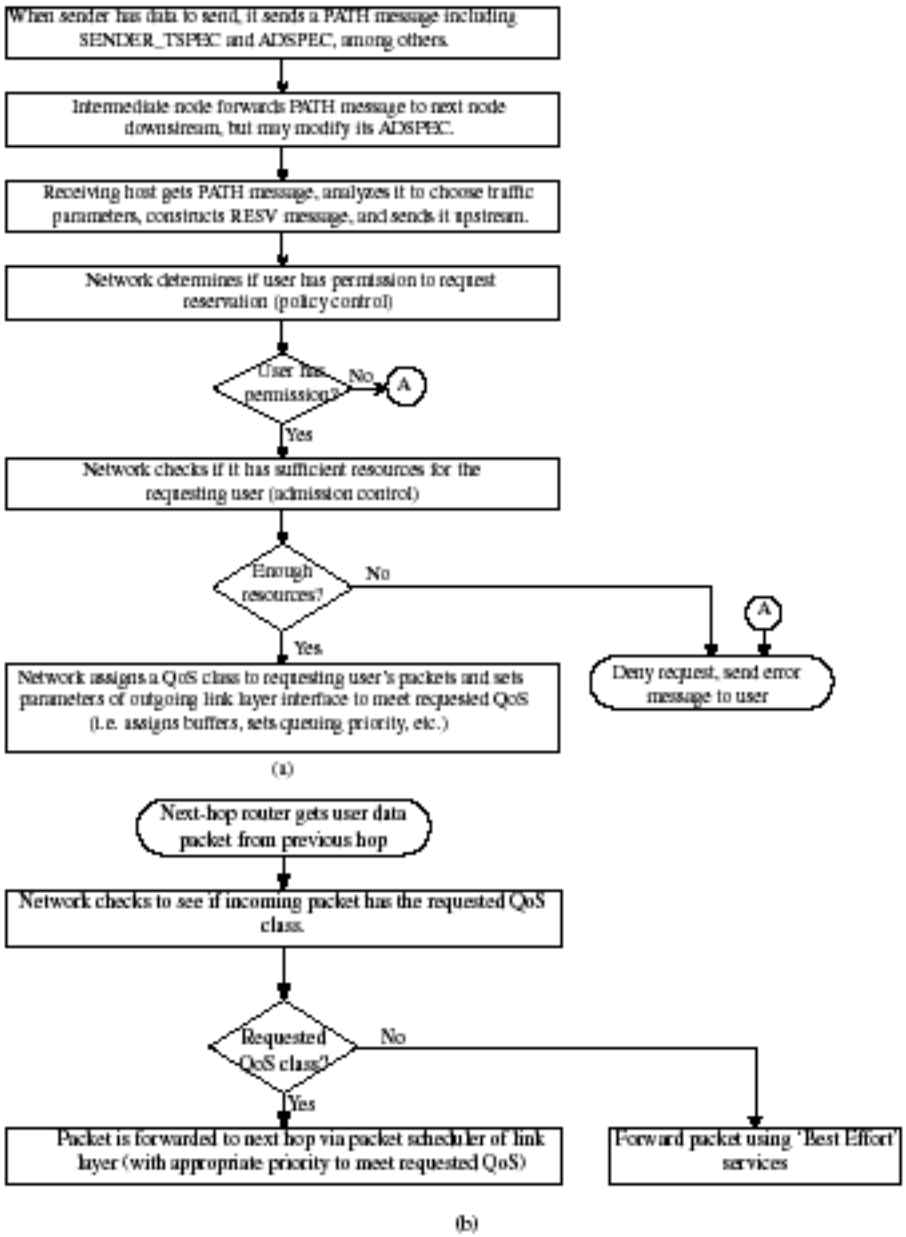


Figure 9-4
The RSVP protocol mechanism

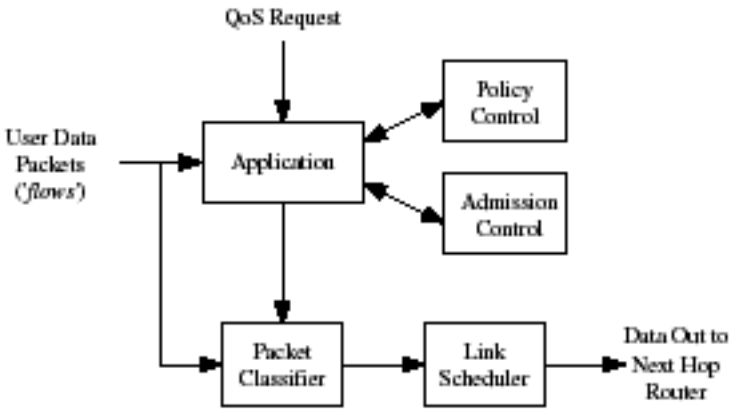


Figure 9-5
Functional entities in a router or host that process a reservation request and subsequently provide the desired QoS to incoming data packets

and then forwards the reservation request backwards to the previous-hop router. This process continues until the reservation request reaches the sending host.

Figure 9-5 shows functional entities in a router that processes a resource reservation request and subsequently provides the desired QoS to incoming data packets. The term *packet classification* is used to indicate how a particular packet is going to be treated when transmitting it over an outgoing link so that the requested QoS can be ensured. For example, it might specify what fraction of the total bandwidth of that link can be assigned to that packet. Similarly, if different packets are held in different queues while the outgoing interface is busy, packet classification would indicate how the node should schedule the transmission of each packet from different queues when the interface becomes subsequently idle.[7]

As user data packets come into a router from the source (or another upstream router), the service class of each packet is examined. If the associated service class matches the service class assigned in the request phase, the packet is treated for forwarding purposes using the scheduler parameters that were previously set during the reservation request process. Otherwise, the packet is forwarded to the next hop downstream as the best-effort or available bit rate traffic.

[7]The 3-bit precedence levels of the *Type of Service* field in IPv4 or the 20-bit *flow label* in conjunction with the 8-bit *Traffic Class* field in IPv6 can be used as packet classifiers, which in essence is a mapping of the QoS to an appropriate number.

Figure 9-6 *PATH and RESERVATION (RESV) messages in RSVP*

The flow of PATH and RESV messages is shown in Figure 9-6. As each hop receives the RESV message, it may reject the request or accept it by reserving resources according to the *flowspec* content of the RESV message. Finally, when the sender receives it, it sets the necessary traffic control parameters for the next-hop router. QoS is controlled at the point where the data enters the link layer at the source end.

Formats of RSVP messages are shown in Figure 9-7. Each message consists of a common header with eight octets and a variable-length description of each object to be included in the message. As shown in the figure, the data structure of each object begins with a four-octet header and a variable length description of the object.

A PATH message may contain the following objects: INTEGRITY, SESSION, RSVP_HOP, TIME_VALUES, POLICY_DATA, SENDER_TEMPLATE, SENDER_TSPEC, or ADSPEC. The POLICY_DATA may include user class and credentials, account numbers, limits, and so on, that indicate whether this user has the authority or permission to make the reservation request.

Similarly, an RESV message may contain the following objects: INTEGRITY, SESSION, RSVP_HOP, TIME_VALUES, RESV_CON-

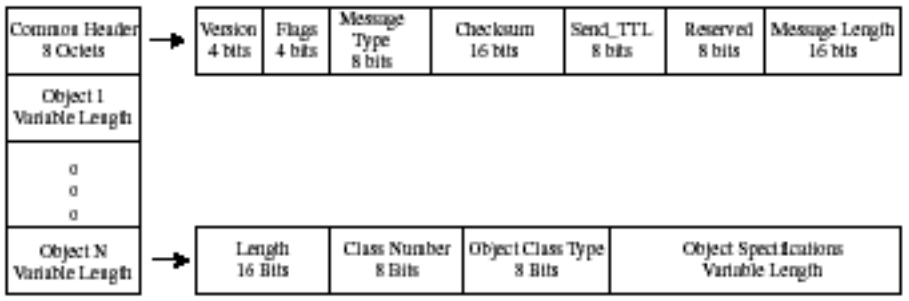


Figure 9-7 *Message formats of RSVP messages. Examples of objects are ADSPEC, FLOWSPEC, and so on. Each object specification is of variable length.*

FIRM, SCOPE, POLICY_DATA, STYLE, and flow descriptor list. The flow descriptor list includes the FLOWSPEC and FILTER_SPEC objects.

Admission Control

Admission Control Strategies

The idea behind admission control strategies is as follows. When a user originates a call to the network requesting a desired QoS, the network must do two things before accepting the call request. First, it must make sure that it has sufficient bandwidth to allocate to that user. Second, it must determine if, after admitting the user, it can continue to provide the same QoS for all existing connections. One QoS objective may be a desired packet loss probability. Thus, before the network admits the new user, it should determine whether it can meet that packet loss probability goal for all connections, old as well as new.

The network can proceed to do it in two ways. The simplest way is to set aside a portion of its available bandwidth that is equal to (or perhaps slightly greater than) the requested peak data rate of the incoming call and reserve that bandwidth for the duration of the call. This is possible only if the sum of the bandwidth requirements of the existing and new connections does not exceed the total available bandwidth. Or, considering a little more complex situation, assume that each user provides the network with information about its traffic characteristics during call establishment, such as its peak and sustainable data rates. The network can now analyze the traffic using a fixed reference model that is known as *a priori*, determine the mean data rate, data rate variance, and packet loss probability, and then, based on some criteria, decide whether to admit the requesting user. Once admitted, it is allocated a fixed bandwidth. This is called the *nonstatistical or fixed allocation*. This control scheme is suitable for constant bit rate services.

In the previous allocation, there are some inaccuracies. For one thing, a source may not be able to describe its traffic accurately. For another, the real-time traffic from a source is generally quite complex in nature (such as the variable-bit-rate encoded audio or video), and very often cannot be represented by a fixed reference model. To correct these inaccuracies, the network may set the packet loss probability objective slightly lower than the desired value. In this case, the QoS can be maintained with a higher probability but only at the expense of lowering the throughput.

An alternative approach would be for the network to actually analyze the packet arrival processes from all existing connections, estimate the desired QoS parameter, assuming that the requesting source is admitted into the network, and then determine if it can meet the packet loss probability objective for all users. If it cannot, it rejects the incoming call. Otherwise, it connects the user to the network. This is called a *statistical allocation scheme*. This is summarized in Figure 9-8. An advantage of this scheme is that the network

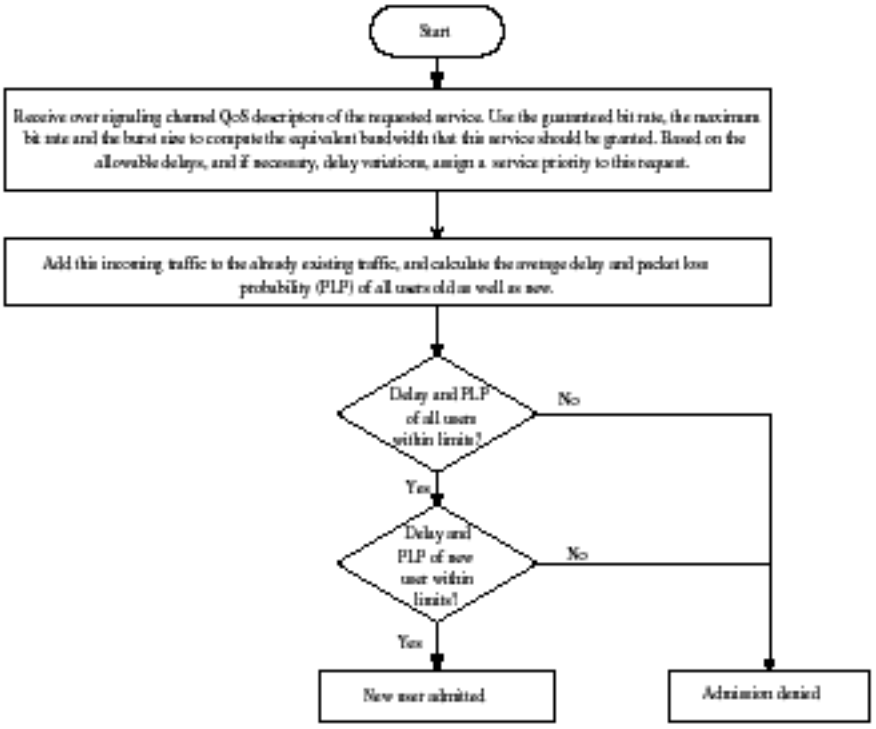


Figure 9-8
An admission control policy

can now serve more users than would be possible with the fixed allocation scheme. However, because the network must now do the traffic analysis in real-time, these schemes are generally very difficult to implement. The problem becomes even more complex if the network consists of many switches over geographically dispersed areas.

Resource Allocation

Because the packet arrival process is purely random, it is often not easy to compute the required bandwidth for an arbitrary sustainable data rate, peak data rate, and the burst size except for very simple cases. As an example of a simple traffic pattern, consider Figure 9-9 where packets arrive in bursts periodically with a period T. Each burst contains n packets of b bits each with an interarrival time t between them such that $t << T$. This represents the worst-case situation because all incoming packets are treated as bursts. Assuming the duration of each packet it is much less than t,

$$peak\ traffic\ rate = b/t,$$

$$Sustainable\ traffic\ rate\ R = nb/T$$

Then the bandwidth B required to accommodate this peak traffic rate is $B \leq nb$.[8]

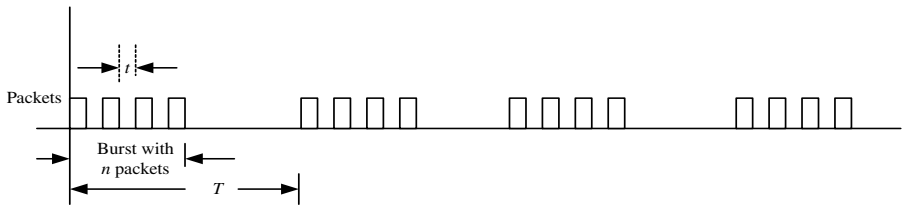


Figure 9-9
An overly simplified traffic pattern used in calculating bandwidth in admission control

[8]In fact, it is shown in [8] that $B = [n - R(n - 1)t]b$.

It has been shown in [8] that if the bandwidth is allocated on the basis of this worst-case traffic pattern as permitted by the leaky bucket algorithm for ATM networks, QoS cannot be guaranteed.[9] Notice that the situation depicted here is overly simplified. In reality, the traffic pattern would hardly, if ever, be so well defined. A number of similar, although more complex, traffic patterns have been considered in [8].

Policing

If the network uses admission control and if no user violates its traffic contract, there should not be any congestion in the network. However, users do sometimes exceed their negotiated rates and cause congestion in the network. Consequently, it is necessary to monitor the sustainable and peak traffic rates and the burst size on each *virtual circuit* (VC). If these rates for any user exceed the negotiated rates, the network can react by dropping its packets. When the user application detects the increase in the packet loss, it may adjust its traffic rate, and thus alleviate congestion in the network. So, the overall policy to prevent congestion and thereby maintain the QoS for all users would be to employ an appropriate admission control scheme, monitor each user's maximum data rate, the burst size, and the sustainable data rate, and drop packets if these parameters exceed their negotiated values.

A simple way to police these parameters is to use the so-called *leaky bucket algorithm*. The idea is illustrated in Figure 9-10. Newly arrived packets enter a bucket (that is, a buffer) of finite depth, say, B. The leak rate of the bucket is R, which is the rate at which the input queue is being serviced by the network. If the bucket is full, it overflows, and consequently, the incoming packet is dropped. Clearly, if packets come in at a rate faster than R, the bucket will overflow after a while, causing packets to be rejected.

The leaky bucket algorithm can be implemented using a single up-down counter—the counter is initially set to zero and is incre-

[9]See the next section.

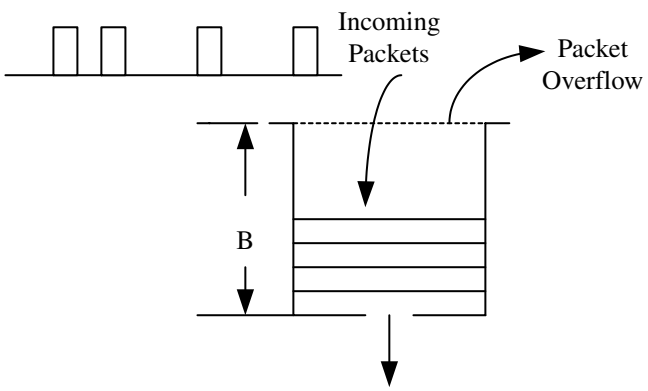


Figure 9-10
Leaky bucket algorithm to control congestion by monitoring the sustainable traffic rate

mented by one each time a packet comes in and decremented by a clock of frequency R, making sure that the counter is never decremented below zero. When the counter reaches the maximum count, we stop incrementing the counter and mark the associated packet for discard.

To police the maximum bit rate, it is only necessary to set the bucket depth B to 1, as shown in Figure 9-11.

To police the average bit rate or the burst size, the token bucket concept is used. Refer to Figure 9-12. Tokens are being generated at a constant rate, say, R, and placed in a bucket that can hold, say, only N tokens. When a packet comes in, it must take a token to be admitted into the network. If there are no tokens left in the bucket, the newly arrived packet is dropped. Clearly, if the rate of the incoming traffic exceeds R, at some point, tokens will be exhausted, and the incoming packets will be denied access into the network. Similarly, suppose that the input source has been inactive for a while so that N tokens have accumulated in the bucket. At this point, the source may transmit N packets in a burst, all of which will be accepted into the network. Thus, the token bucket scheme can police both the average bit rate and the burst size (that is, the number of packets in a burst, which is N in this example).

In most cases, it is necessary to monitor both the maximum and sustainable rates. This can be done easily by means of two buckets in tandem—the simple leaky bucket of Figure 9-10 for the maximum

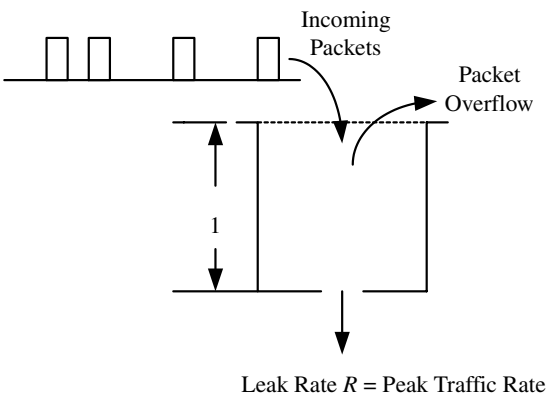


Figure 9-11
Leaky bucket algorithm to police the peak traffic rate

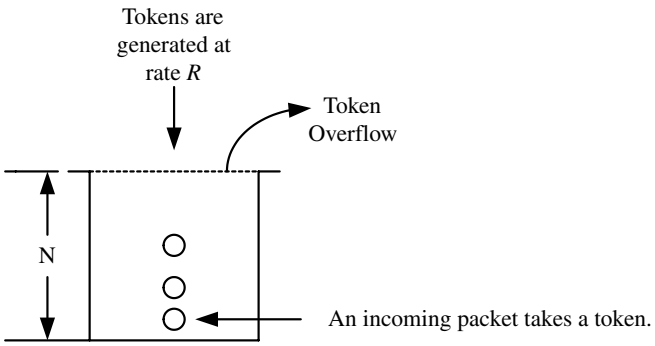


Figure 9-12
The token bucket algorithm

rate and the token bucket for the sustainable rate. Although leaky bucket algorithms are simple to implement, they can sometimes lead to inaccurate results.

Providing Requested QoS

To ensure the QoS, the network must do two things. First, it must allocate enough bandwidth to each user based upon its traffic descriptors, such as the peak traffic rate, the sustainable traffic rate, and the burst size. If users are admitted into the network following

strict admission control strategies, this requirement would be satisfied automatically. Second, many real-time applications—audio, video, process control information, and so on—specify the maximum delay they can tolerate. Because a physical channel may contain a number of logical channels and because packets on each of these logical channels arrive at a node randomly, it is likely that the output link would be busy transmitting data from a particular logical channel when packets are coming in over other logical channels. In this case, it is necessary to buffer the packets temporarily in a queue until the link is idle. Or, if at a given point in time the network is congested, we could drop the packets, or better yet, save then in a buffer until that condition clears. In any case, because packets from many different users are buffered in the queue, the delay encountered by each would depend on how the packets are being serviced.

A number of servicing strategies are available. For example, when the output link is busy, packets of all classes could be buffered in a single *first-in, first-out* (FIFO) order. When the link becomes idle, they are read out of the FIFO and transmitted over the link on a first-come, first-served basis. There are two problems in this approach. First, if packets for a particular application are small (such as Telnet) and appear behind large packets of other applications (such as FTP), they would be subjected to rather long delays. Second, it would be impossible to guarantee the QoS this way because if the maximum delay is one of the QoS parameters, and if packets with low delay requirements happen to be near the end of the FIFO, they would be the last ones to be transmitted over the link.

In an effort to solve the previous problems, a separate buffer is assigned to each application (that is, packet class). These buffers are then serviced following certain rules. For example, the scheduler could read the buffers in a round robin fashion and transmit the entire contents of each before moving on to the next. This, again, cannot guarantee the requested QoS. An alternative way is to transmit only a certain number of packets—rather than all—from each buffer according to their service classifiers, move to the next, and continue the cycle until all buffers have been fully serviced. This is called the *weighted fair queuing scheme.* In this scheme, because packets are classified according to the requested QoS, the number of packets transmitted at

a time from any buffer (that is, the bandwidth allocated to the associated application) depends on the QoS. This scheme is shown in Figure 9-13. In essence, the total available bandwidth of the link is being divided into a number of small time slots or units. Each user requesting service from the network is assigned a certain number of these basic units according to its associated QoS metrics.

For constant bit rate services that cannot tolerate any variable delays, the associated packets must be sent over the link at regular intervals using the required fraction of its bandwidth. Thus, in general, when both constant bit rate and variable bit rate services are to be supported, the bandwidth assignment in the weighted fair queuing scheme appears as in Figure 9-14.

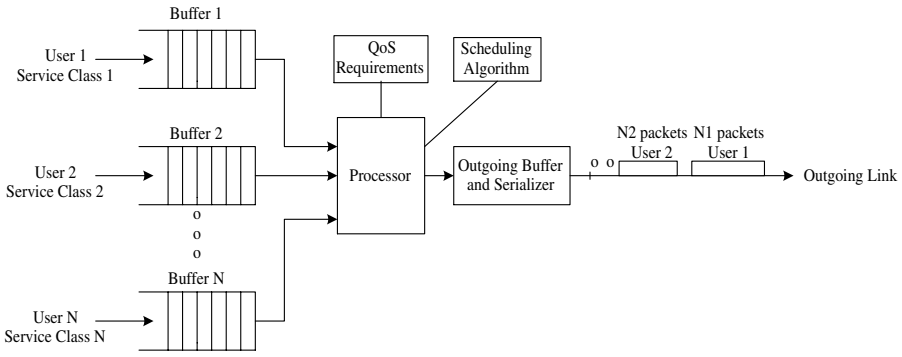


Figure 9-13 *Weighted fair queuing scheme*

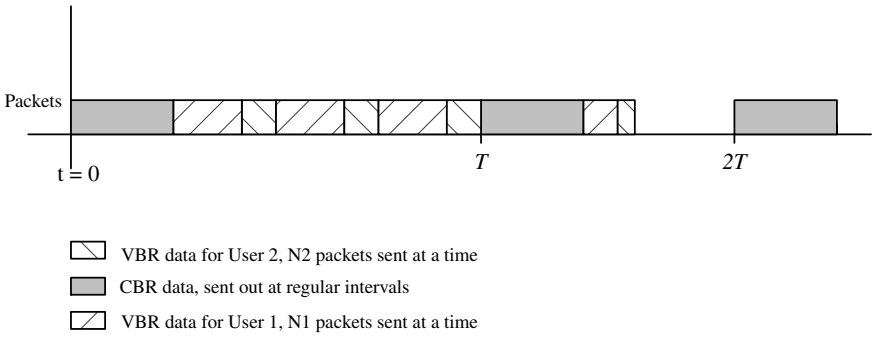


Figure 9-14 *CBR and VBR packet data services handled at a node*

Differentiated Services (DiffServ)

Because in IntServ, an explicit reservation of resources is made for each flow, the integrated services model is not suitable for a large system. The DiffServ model overcomes this problem by avoiding classification on a per-user or per-flow basis. Instead, it differentiates the incoming packets into a small number of classes and forwarding each packet on the basis of its class.

In DiffServ, packets are differentiated using a one-octet *Differentiated Service* (DS) field, which is actually the same as the *Type of Service* (TOS) field of IPv4 header [4] or the Traffic Class field of IPv6 header [5] except for the fact that the meaning of each bit is now redefined. The TOS field of IPv4 and the significance of its subfields are shown in Figure 9-15.

The DS field in DiffServ is shown in Figure 9-16. Instead of defining different service types available in the network, the six-bit *DS codepoint* (DSCP) indicates how a collection of packets (referred to in the RFC as a *behavior aggregate*) should be treated for forwarding purposes on a per-hop basis. The forwarding treatment for a group of packets in the DiffServ model is termed the *Per-Hop Behavior* (PHB). There are many possible DSCPs and corresponding PHBs. For example, DSCP 000 000 may indicate a collection of packets that are delivered by the network on a best-effort basis. Similarly, DSCP 010 000 may be defined such that packets marked by this codepoint are forwarded on a timely basis with a higher probability than packets designated with, say, DSCP 001 000. Yet another DSCP may indicate that associated packets are to use only, say, 50 percent of the bandwidth of a shared link, and so on. Notice that the so-called *default PHB* mapped to by DSCP 000 000 is the same as the forwarding strategy used in present-day IP routers, and thus provides for interoperability between existing routers and a DiffServ network. For a detailed description of how codepoints may be mapped to PHBs, see Reference [15].

Procedures to provide QoS in DiffServ are conceptually similar to those of IntServ that have been discussed earlier. When packets

3 Bits Precedence	D 1	T 1	R 1	Unused 2

(a)

Precedence Bits	
000	Routine
001	Priority
010	Immediate
011	Flash
100	Flash Override
101	CRITICAL/ECP
110	Internetwork Control
111	Network Control

(b)

Delay (D), Throughput(T) and Reliability (R) Bits	
000	Normal Delay, Throughput and Reliability
100	Low Delay
010	High Throughput
001	High Reliability

(c)

Figure 9-15
The TOS field in the IPv4 header. (a) The field format. (b) Significance of precedence bits. (c) Meaning of D, T, and R bits.

originating from a sender enter a DiffServ network, they are classified on a PHB basis and conditioned (that is, metered, shaped, and marked) at an edge device. As packets travel downstream, the aggregate traffic is policed at each intermediate node in the network. New users are admitted only if the network capacity is not exceeded.

TEAMFLY

Figure 9-16 The DS field in DiffServ

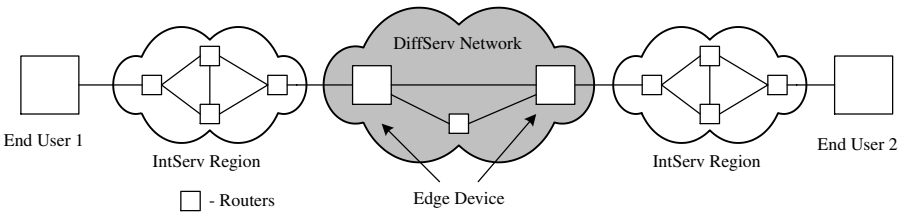


Figure 9-17 Providing an end-to-end QoS over any network

Evidently, an advantage of the DiffServ model is that it lends itself easily to a large network such as the Internet. On the other hand, because it does not support per-user flows, it cannot provide QoS on an end-to-end basis (that is, for end users). Because an IntServ network does not scale well, but supports per-user flows, an end-to-end QoS can be delivered over any network by providing IntServ in the access part and DiffServ in the core, as shown in Figure 9-17. Here, flows between two IntServ regions may be considered as being transported across the DiffServ network over virtual connections between two IntServ-capable routers [11], [16].

RSVP for Mobile Systems

Problems arise when standard RSVP is used in a mobile communications network. Reference [11] has studied these problems in detail, and also summarizes the various RSVP-based protocols that have been suggested as a possibility for wireless mobile networks. To understand these problems, consider Figure 9-18, where the cellular network is connected to the *packet-switched public data network* (PSPDN) via three routers—R1, R2, and R3. R2 provides

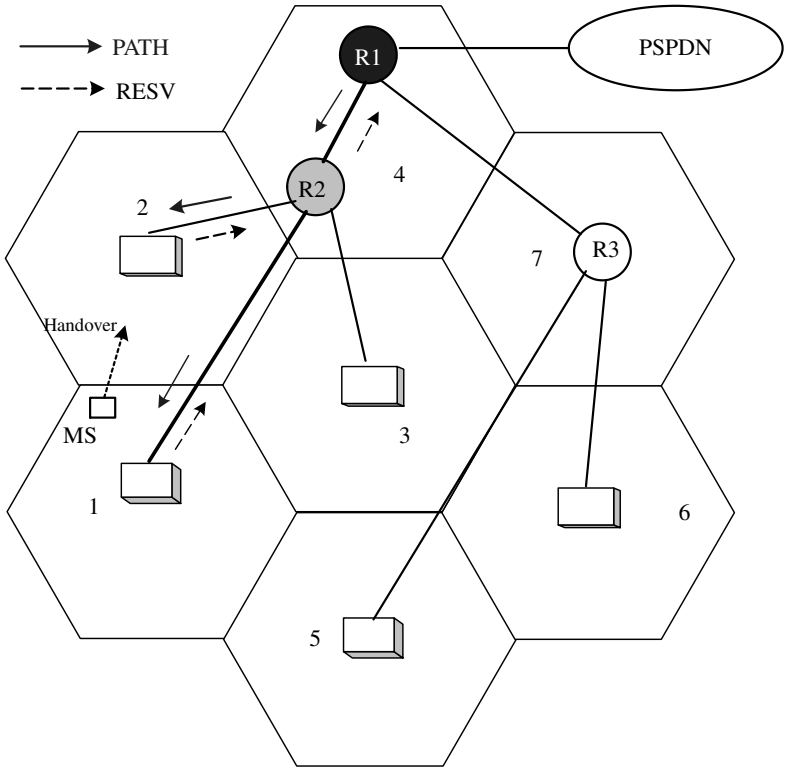


Figure 9-18
RSVP in a mobile environment

services to cells 1–4, and R3 to cells 5–7.[10] Initially, when the *mobile station* (MS) is in cell 1, it reserves resources along the path shown by the thick lines from base station 1 to R2 to R1. If the MS now moves into cell 2, it is attached to base station 2 and makes a reservation along the new route from base station 2 to R2. In RSVP, because a receiving host cannot initiate a reservation request (using the RESV message) until it has received a PATH message and because a data source sends the PATH message periodically, the MS must wait before it can make the new reservation. Meanwhile, R2 continues to send packets along the old route to cell 1,

[10]Notice that router R1 is functionally equivalent to a *Gateway GPRS Support Node* (GGSN) of a GPRS or all-IP wireless network. Similarly, routers R2 and R3 correspond to *Serving GPRS Support Nodes* (SGSN).

and so these packets are lost by the MS. This problem may be somewhat mitigated by modifying RSVP so that as soon as the MS is handed over to the new cell, R2 issues a PATH message. However, since the resource reservation process is rather slow, the associated delays may still be too long to avoid any packet loss. The problem becomes particularly severe if the MS moves into a cell where the traffic is already quite high. In this case, the MS must renegotiate with the network for the allocation of required resources, leading to even longer delays and further loss of packets. And depending upon the requested QoS, the network may not be able to guarantee the desired quality, and consequently, may even deny the reservation request.

This problem can be solved if the network has some prior knowledge of how a mobile station is going to move around in a serving area. Reference [12] suggests an extension of RSVP, called *Mobile RSVP* (MRSVP), based upon this idea. In this protocol, the network maintains in its database a list of all nodes that are likely to be visited by an MS. The MS reserves resources at all of these nodes in advance, even though it will be using resources at only one of them at any time. In this way, delays caused by the negotiation and reservation of resources can be eliminated. The flow of RSVP messages in this protocol is shown in Figure 9-19. Because the MS may visit neighboring cells 1, 2, and 3, resources are reserved along three different routes. If the MS is located in cell 2, only the route, shown by the heavy lines, from the MS to base station 2 to R2 through R1 to the PSPDN is active, while the other two routes are passive. Although this protocol is conceptually simple, there are several disadvantages:

- Although only one of the routes is active at any time, the system must reserve resources at many other nodes. Much of these resources may never be used by this MS, but could have been used by other mobile stations for a more efficient utilization of the bandwidth.
- An incoming call is blocked if requested resources are not available at all nodes. Therefore, the call-blocking probability increases with the number of cells where the reservation is to be made.

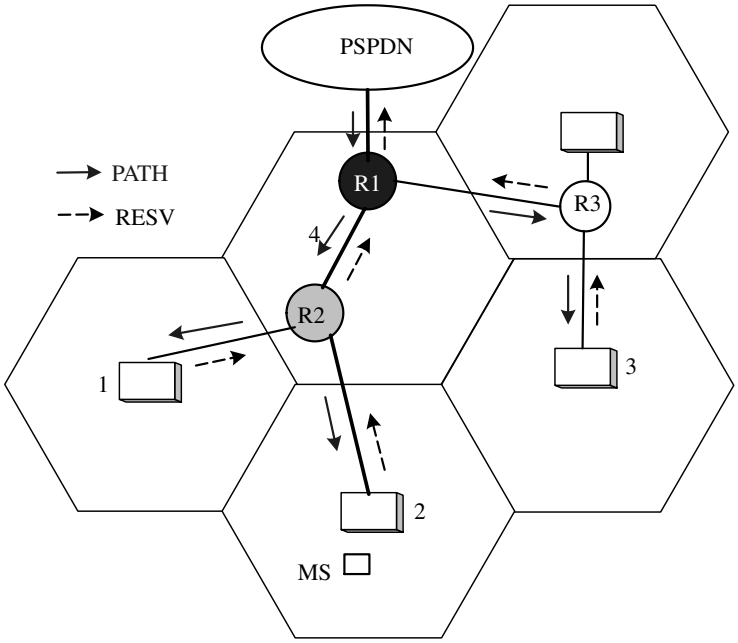


Figure 9-19
Reservation in MRSVP proposal

- In many instances, an MS may not know in advance exactly how it is going to roam. As such, this protocol is not very practical.
- If the number of cells that a mobile station may likely visit is large and if there are many mobiles in a serving area, the database that must be maintained by the network also becomes very large.

Reference [13] has suggested another protocol called *Mobile IP with Location Register* (MIP-LR) for mobile wireless networks. According to this protocol, when a mobile station moves into a foreign serving area, its new location is saved in the HLR. Subsequently, when a source node has a packet to send to this MS, it receives the location address of the mobile from the HLR, and sends the packet directly to that address. However, it is still necessary to

reserve resources along the new route, and packets may be lost as resources are being reserved.

Summary

In this chapter, we have discussed QoS issues and concepts and described how it can be provided in 3G UMTS networks. Providing the QoS usually consists of four steps: requesting resources from the network in accordance with a desired quality, admission control of the newly arrived user, resource reservation by the network, and policing the incoming packets to ensure that users are not violating their contract. In order to request the QoS, the user must know how to characterize its traffic. With this end in view, we have classified the traffic that is likely to originate in UMTS and described the traffic attributes that can be used to create a reasonable set of simple QoS profiles. The RSVP protocol, admission control procedures, and policing schemes have been presented in some detail. Although many of the ideas of the QoS that are applicable to fixed networks extend to mobile networks, standard RSVP is not quite suitable. Problems that arise when RSVP is used in a mobile network are discussed. A number of protocols based on the modification and extension of RSVP have been suggested. A brief description of some of these protocols is presented.

References

[1] R. Braden, et al., "Resource Reservation Protocol (RSVP) — Version 1 Functional Specification," RFC 2205, September 1997.

[2] J. Wroclawski, "The use of RSVP with IETF Integrated Services," RFC 2210, September 1997.

[3] R. Braden, et al., "Integrated Services in the Internet Architecture: An Overview," RFC 1633, June 1994.

[4] *Internet Protocol,* RFC 791, September 1981.

[5] S. Deering, et al., "Internet Protocol, Version 6 (IPv6) Specification," RFC 2460, December, 1998.

[6] M.R. Karim, *ATM Technology and Services Delivery.* New Jersey: Prentice Hall, 2000, pp. 87–98.

[7] D.C. Lee, *Enhanced IP Services for Cisco Networks.* Indiana: Cisco Press, 1999, pp. 115–177.

[8] N. Yamanaka, Y. Sato, and K. Sato, "Performance Limitation of the Leaky Bucket Algorithm for ATM Networks," *IEEE Trans. Commun.*, Vol. 43, No. 8, August 1995, pp. 2298–2300.

[9] 3G TS 22.105 Release 1999, Services and Service Capabilities.

[10] 3GPP TS 23.107: QoS Concept and Architecture, Release 1999.

[11] B. Moon and H. Aghvami, "RSVP Extensions for Real-Time Services in Wireless Mobile Networks," *IEEE Commun. Mag.*, Vol. 39, No.12, December 2001, pp. 52–59.

[12] A.K. Talukdar, et al., "MRSVP: A Resource Reservation Protocol for an Integrated Services Network with Mobile Hosts," Dept. Comp. Sci. *Tech. Rep. TR-337*, Rutgers University.

[13] R. Jain, et al., "Mobile IP with Location Registers (MIP-LR)," *Internet Draft*, draft-jain-miplr-01.txt, July 2001.

[14] S. Blake, et al., "An Architecture for Differentiated Services," RFC 2475, December 1998.

[15] K. Nicholas, et al., "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," RFC 2474, December 1998.

[16] Y. Bernet et al., "A Framework for Integrated Services Operation over Diffserv Networks," RFC 2998, November 2000.

See the following web site for RFCs:

http://www.cis.ohio-state.edu/

CHAPTER 10

Network Planning and Design

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

The objective of network planning and design is to provide wireless telephony services in a serving area in the most cost-effective manner. In the case of an existing system, the objective is to expand and augment its facilities so as to add new features and capabilities or increase its capacity in case the system has reached its coverage limit. The design usually involves determining the number of base stations and their locations that would provide the necessary coverage in a serving area, meet the desired grade of service, and satisfy the required traffic growth so that the total startup cost is minimized and the rate of return maximized. Clearly, if the network is well planned, it would be able to meet the traffic growth within the limits of its design until a point is reached where it becomes necessary to add new cells, assign new frequencies to an existing cell, or supplement the old system with a new one in a manner that is consistent with the quality of service objectives. Because base stations are to be connected to a mobile switching office, the design also requires the capacity and type of the connecting links to be specified. Sometimes elements of the core network may not be able to support new services, such as high-speed multimedia applications or gateway access to a packet data network. In these cases, the network planner must consider upgrading the switching systems.

The design process is roughly the following [5], [6]. Service providers who want to own and operate the network generate a set of system requirements concerning the type of the desired system (such as analog, *Global System for Mobile Communications* [GSM], *Code Division Multiple Access* [CDMA], and so on), the expected traffic, and the desired service quality. In general, received signal-to-interference ratios and bit error rates are used as the quality of service indicators. Based on the above requirements, an appropriate propagation model is used to calculate a link budget that indicates the maximum allowable path loss for a given transmitter power so that the received signal-to-interference ratio at any point in the designated serving area is sufficient to ensure the desired quality. Maps and the terrain of the serving area are inspected, and assuming approximate locations of base stations, the signal distribution over that area is calculated.

One of the design goals is to provide coverage on the entire serving area with a minimum number of base stations consistent with the requirement of the projected traffic growth. Currently, software tools

are available that take into consideration design requirements as well as terrain features of the serving area, and given any number of base stations and their locations, they can plot the signal distribution over that area. If the coverage or the number of base stations is not optimum, the user can alter their number and locations and run as many iterations as necessary until the design objective is satisfied. The design might make assumptions about the maximum available transmitter power, commercially available transmitters types, specific antenna (such as omnidirectional or sectored), their heights, and so on. It may be reviewed by engineers in consultation with the customer, modified if necessary, and when it is complete, all system components may be installed. Because the system, when actually installed, may not be exactly the same as the design (for example, the actual base station locations or their antenna heights may be slightly different), it may be useful to verify the design by running some field tests.

System requirements fall usually into the following categories:

- *The coverage area* This generally involves areas to be served, e.g. counties comprised by the area. It may also be necessary to include information on such things as the terrain and clutter, e.g., the average height and density of buildings, streets, hills, forests, large water bodies if any, highways, population distribution, and so on.
- *System-related requirements* They should specify the following:
  - The technology type, indicating, for example, whether it should be CDMA, GSM, *Wideband CDMA* (W-CDMA), *Time Division Multiple Access* (TDMA), cellular, and so on
  - The allocated bandwidth, e.g., the number of available channels and the number of channel sets (that is, the reuse factor)
  - The type of antennas to be used (for the link budget calculation)
  - Maximum cell size and so on
  - The cost objective
- *The traffic* This should include the following:
  - The number of mobile stations to be served

- The amount of traffic, e.g., the offered load per mobile and the holding time
- The geographical distribution of the traffic if it is not uniform over the whole serving area
- Specification of the traffic types (such as constant bit rate, variable bit rate, delay-intolerant data, elastic data, and so on) and relevant traffic descriptors (e.g., the maximum tolerable delay during the busy-hour period)
- The probability of calls being blocked or the grade of service
- Ratio of the total daily traffic to the busy-hour traffic

For satisfactory service, the system should be designed so that the mobile stations receive a sufficiently strong signal inside buildings or vehicles where the penetration loss may be significant, outside buildings where there is no such penetration loss, and on highways. The system may then be designed so as to optimize one of the following parameters or any combination thereof:

- The signal distribution as received by mobiles or base stations
- The S/I ratio at base stations
- The S/I ratio at mobiles or any combination of these parameters

However, the usual practice is to design the system such that both forward and reverse links have a balanced signal distribution.

Network Design

Spectrum Requirements

Once the system requirements are known, the first step is to ensure that the service provider has licensed sufficient spectrum for the expected amount of traffic and the call-blocking probability. The system must be designed so that it can carry the peak traffic, that is, the traffic during a busy-hour period. The traffic is determined by the call arrival rate and the holding time of each call. The unit of traffic is the *Erlang*, which is defined as the traffic that a circuit can carry

if it is utilized 100 percent of the time during a busy hour. The holding time varies from applications to applications, but for telephone-type conversations during a busy hour, it lies in the range of 60 to 80 seconds. The probability that a call is blocked depends on the number of traffic channels (circuits) available and the total amount of traffic coming into the network (the offered load), and is given by the well-known Erlang B formulation. Call-blocking probabilities for various values of the offered load and circuits are available as tables and graphs, where it is assumed that calls arrive at the system randomly with a Poisson distribution and that blocked calls are cleared; that is, when a call is blocked, it is not reinitiated [8], [9]. The traffic capacity in Erlangs as a function of the number of circuits for a few values of the call-blocking probability is given in Figure 10-7 of Appendix A to this chapter. The determination of the bandwidth is best explained by an example. Consider the following.

Example. Suppose that it is necessary to design a cellular system for 50,000 subscribers. On the average, each subscriber makes about two calls during a busy hour, and the average holding time of a call is two minutes. Let us assume that the serving area is to have about 14 3-sector cells and that the traffic is uniformly distributed over the entire serving area.[1] If the call-blocking probability is to be 1 percent, how much bandwidth is required to provide the service?

Solution for Case 1. First, let us consider an analog system. The total traffic during the busy hour = No. of Subscribers × No. Calls/hour × Holding Time in hours = 50,000 × 2 × (2/60) = 3,333.3 Erlangs. So the traffic per sector of a cell = 3,333.3/(No. of Cells × No. of Sectors) = 3,333.3/(14 × 3) = 79.36 Erlangs.

The number of channels or circuits per sector required to support this traffic for a call-blocking probability of 1 percent (from Figure 10-7) = 95. In other words, 95 × 3 or 285 channels are needed per cell.

[1]A few comments are in order here. First of all, the traffic is rarely uniform over a serving area. It generally depends on the population distribution and is much higher in urban and suburban areas, gradually decreasing toward outlying areas. Secondly, the number of cells is not known at the outset. In fact, given the requirements, our goal is to determine this number.

If each channel has a bandwidth of 30 kHz, the total bandwidth per cell amounts to 8.55 MHz. Because seven-cell clusters are being used, the total spectrum required = 8.55 × 7 = 59.85 MHz, which is well beyond the spectrum allocated by the FCC to U.S. cellular systems.

In the previous example, the bandwidth required is not going to be significantly decreased by simply increasing the call-blocking probability. To be able to meet the spectrum requirements, many more cells must be used so that each cell is now smaller than before. As an example, if the number of cells is increased to 70 and if the blocking probability is again 1 percent, the reader can easily verify that the bandwidth required is about 15.75 MHz.

Case 2. Let us now consider a CDMA system. As in case 1, the number of channels per sector for a 1 percent call-blocking probability = 95. Recall that the number of simultaneous users per sector of a CDMA cell is given by

$$N = 1 + \frac{G_p}{\frac{E_b}{N_0}(1 + \beta)v} \tag{10-1}$$

where

N = The number of simultaneous users per sector of a cell.

G_p = Process gain. It is given by $G_p = B/R_b$, where B is the bandwidth and R_b is the information bit rate.

E_b = Energy per bit.

N_0 = Spectral density of noise plus interference. Thus, E_b/N_0 = required signal-to-interference ratio.

β = Interference due to mobiles in other cells transmitting on the same channel. β is 0.85 for 3-sector cells and 0.6 for omnidirectional ones.

ν = Voice activity factor. The value of ν is generally taken to be 0.4.

Notice first of all that equation (1) gives an upper limit on the number of users because it assumes that the power control of the mobiles in this cell is perfect and that the interference to mobiles in any sector due to mobiles in the other two sectors is zero, assumptions that are usually not satisfied in practice. Second, the mobiles that are in cells other than the desired cell are not under power control by the base station of this cell. So, the interference they cause varies randomly. However, because there are many of these mobiles,

it is possible to consider an average value of that interference, which in fact is being represented here by β.

A satisfactory value of E_b/N_0 for speech is about 7 dB or $E_b/N_0 = 10^{0.7} = 5.1$. Because 3-sector cells are being used, $\beta = 0.85$. So, $G_p = (95 - 1) \times (5.1) \times (1.85) \times (0.4) = 348.5$. If the bit rate $R_b = 14.4$ kb/s, $B = 5$ MHz, which is well within the realm of the allowable spectrum. Thus, it is possible to achieve a significantly higher capacity with the CDMA technology than with a corresponding analog system (or for that matter, a TDMA system).

Link Budget Calculation

The link budget calculation is fundamental to the design of cellular systems. We shall illustrate the procedure involved by way of examples.

Analog System

Example 1. Let us determine the transmitter power output P_{BTS} of a base transceiver station that will provide a 30 dB *signal-to-noise ratio* (SNR) at the baseband in an urban coverage area. The system is assumed to be analog FM. The following parameters are assumed:

Carrier frequency	900 MHz
Base station antenna height	30 m
User Equipment (UE) antenna height	1.5 m
Distance, d, between UE and base station	2.0 km
Base station antenna gain, G_{BTS}	9 dB
UE antenna gain, G_{UE}	3 dB
UE receiver noise figure, NFR[2]	5 dB
Radio frequency (RF) bandwidth	30 kHz
Bandwidth of speech	~3 kHz

[2]The receiver noise figure indicates the noise that the receiver adds, usually in its first amplifier stage. This is in addition to the noise that comes with the signal at the input to the receiver.

According to Hata-Okumura model that was discussed in Chapter 2, "Propagation Characteristics of a Mobile Radio Channel," the path loss P_L in a typical urban area at 900 MHz with respect to a reference point at a distance of 1 km from the transmitter antenna is given by the following relation [2], [3]:

$$P_L = 123.33 + 33.77 \log r \text{ dB}, r \geq 1 \text{ (km)}$$

So, in this example,

$$P_L = 123.33 + 33.77 \log (2.0) = 133.5$$

Referring to Figure 10-1, the input to the mobile receiver is

$$p_r = P_{BTS} + G_{BTS} - P_L + G_{UE} = P_{BTS} + 9.0 - 133.5 + 3.0 = P_{BTS} - 121.5$$

The receiver noise floor is given by 10 log (KTB), where K is the Boltzman constant, T is the absolute temperature and is taken to be 290 degrees Kelvin, and B is the RF bandwidth. Here, kT = $1.38 \times 10^{-20} \times 290 = 4 \times 10^{-18}$ mW/Hz. So, the receiver noise floor is 10 log (KT) = −174 dBm/HZ. Because the noise figure of the receiver is 5 dB, the receiver noise density is actually −174 + 5.0 = −169 dBm/Hz.

So,

$$P_{BTS} = p_r + 121.5 - 169.0 + 10 \log (\text{B}) = p_r - 47.5 + 10 \log (\text{B}) \text{ dBm}$$

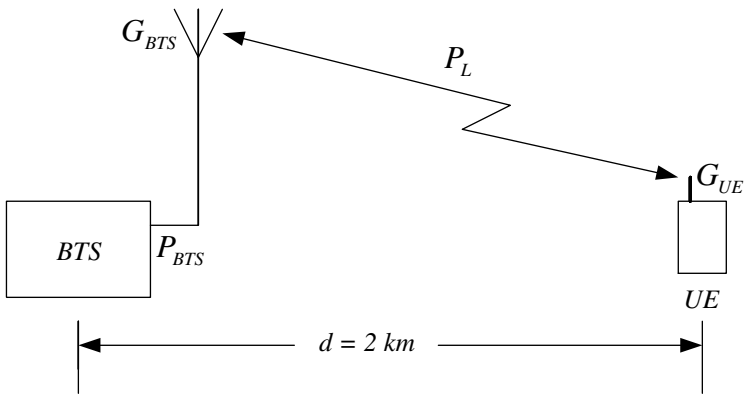


Figure 10-1
Link budget calculation of the example

The baseband SNR of an analog FM system depends on the average value of the *carrier-to-noise ratio* (CNR) at the input to the receiver and the RF channel bandwidth. Reference [1] describes how this average CNR for any desired SNR at the baseband can be calculated by averaging the fading signal at the receiver input over the fading distribution. It is shown that to achieve a 30 dB SNR at the baseband with an RF bandwidth of 30 kHz, the required CNR at the receiver is 33.0 dB. In other words, $p_r = 33.0$ dB. This assumes that the mobile velocity is about 100 km/h. So,

$$P_{BTS} = 33.0 - 47.5 + 10 \log (30{,}000) = 30.27 \text{ dBm} = 1.06 \text{ W}.^{3}$$

In this calculation, random FM has been ignored. Also, there is no diversity in the system.

CDMA System It was shown in Chapter 2 that the local mean signal level varies randomly with a log-normal distribution with a standard deviation of about 8 to 12 dB. As the mobile moves into a region where the signal level received from another base station is higher, it will be handed off to the new base station because the signals from the two base stations are generally uncorrelated in the analog system. In a CDMA system, when a mobile station is in a soft handoff state, there will be a 2 to 3 dB gain due to diversity combining. However, in some cases, signals from the two base stations may not be completely uncorrelated,[4] and as a result, soft handoff may not take place. Thus, for a CDMA system, a log-normal fade margin of 8 to 12 dB as well as a soft handoff gain of 2 to 3 dB should be included in the link budget calculation.

It is also necessary to include a receiver interference margin to help avoid an overly optimistic estimate of the allowable path loss. This margin depends upon cell loading that indicates the percentage of the maximum number of users that the cell has been designed for,

[3]It has been shown in Reference [1] that as the RF bandwidth is increased, the transmitter power required for a 30 dB SNR at the baseband first decreases, reaches a minimum, and then begins to increase.

[4]This is true of IS-95, which, unlike the UMTS, is a synchronous system.

who are actually using the system. Clearly, the greater the loading, the larger the interference margin should be. One way to estimate this margin is to use the following formula:

$$\text{Receiver Interference Margin} = 10 \log \frac{1}{1 - l_f}$$

where l_f is the loading factor. As an example, suppose the capacity per sector of a 3-sector cell = 30 (users). If, on the average, only 15 users are in the system, $l_f = 0.5$ and so the receiver interference margin = 3 dB. So, in this case, a fade margin of 3 dB should be used in the link budget calculation.

In UMTS, the closed loop power control on an uplink channel takes place at a rate of 1,500 times a second. In IS-95, this rate is 800 per second. The fast power control can be used effectively to overcome fading, but only for slow-moving vehicles. At higher speeds, say, 120 km/h or higher, the number of fades per second is significantly higher, and the average fade duration is lower than for vehicle speeds of, say, 10 km/h. Consequently, the fast power control cannot compensate for fading at higher speeds. To compensate for inaccuracies in power control algorithms, a fast fading margin of 2.0 to 5.0 dB should be included in the link budget calculation for low vehicle speeds.

Reverse Channel (Uplink) To illustrate these ideas, we will consider a narrowband CDMA system and calculate the link budget for the reverse channel.

Example 2. Let us assume the following parameters:

The mobile transmitter power is 250 mW = 24 dBm. The transmitter antenna gain, and the cable and connector losses at the mobile station will be ignored in this example.

Body loss = 2 dB
In-vehicle penetration loss = 8 dB
Base station receiver antenna gain = 15 dB
Receiver cable loss = 1 dB
Receiver noise figure = 5 dB
Receiver interference margin = 3 dB
Information rate R_b = 14.4 kb/s for a 13 kb/s vocoder
E_b/N_0 = 7 dB
Soft handoff gain = 2 dB

Log-normal fade margin = 8 dB

The previous values of E_b/N_0 and fade margin have been chosen with a view to providing 90-percent coverage[5] near cell boundaries and 96-percent coverage over an entire cell.

We are required to calculate the maximum allowable path loss.

The effective radiated power of the mobile transmitter = 24 dBm (that is, 250 mW). Suppose that the path loss from the mobile station to the base station is P_L. Then the input to the base station receiver, p_{in} = 24 − Body Loss − Penetration Loss − Path Loss + BTS Receiver Antenna Gain − Receiver Cable Loss = 24 − 2 − 8 − P_L + 15 − 1 = 28 − P_L dBm.

Because the receiver noise figure is 5 dB and the receiver is required to provide an interference margin of 3 dB, the noise and interference density of the receiver = 5 + 3 − 174 = −166 dBm/Hz. The input to the base station receiver must provide a 7 dB E_b/N_0 and 8 dB log-normal fade margin and must also support a data rate 14.4 kb/s. So, the required input signal = −166 + 7 + 8 + 10 log (14,400) = −109.4. Because the soft handoff gain is 2 dB, the required input = −109.4 − 2 = −111.4. Therefore,

$$p_{in} \geq -111.4 \text{ dBm, or } 28 - P_L \geq -111.4; \text{ that is,}$$

$$P_L \leq 28 + 111.4 = 139.4 \text{ dB.}$$

Thus, the maximum path loss is 139.4 dB.

If a propagation model is known, the previous path loss can be used to determine the maximum cell size. For example, if the base station antenna height is 50 m, the mobile station antenna height is 1.5 m and the carrier frequency 900 MHz, then following the Hata-Okumura model for a large city, the propagation loss is given by

$$L = 123.33 + 33.77 \log r$$

So, for a maximum path loss of 139.4 dB, r = 2.99 km. In other words, the maximum cell radius is 2.99 km. This value can then be

[5]Here the term coverage means the fraction of the total cell area where the QoS is satisfactory.

used to determine the number of cells required to provide the desired coverage in a serving area.

Clearly, the signal strength will be different at different points within a cell depending on the terrain and clutter. However, because necessary margins have been included in the design, the signal strength everywhere in the cell will be within the prescribed limit. Notice that if the requirement on the signal-to-interference ratio is relaxed so that the QoS is slightly lowered for all users, the radius of a cell will increase.

Forward Channel (Downlink)

W-CDMA W-CDMA supports a number of bearer services at various data rates. For example, one of them is the delay-tolerant interactive data service (such as web browsing) or a file transfer at 384 kb/s or more in urban or suburban environments for pedestrians. Consider, therefore, a W-CDMA application involving a nonreal-time data transfer at 256 kb/s in an urban area at low vehicle speeds, say, 3 to 10 km/h. Because at these speeds the level crossing rate and consequently the bit error rate are low, a relatively smaller value of E_b/N_0 can be used for this service than for speech or real-time mulitmedia applications. Even though the vehicle speed is low, a fast fading margin of about 3 dB is included in the link budget.

Example 3. Assuming the following parameters, let us calculate the maximum allowable path loss.
Chip rate = 3.84 Mc/s
Mobile transmitter power = 24 dBm
In-vehicle penetration loss = 8 db
Base station receiver antenna gain = 16 dB
Receiver cable loss = 2 dB
Receiver noise figure = 5 dB
Receiver interference margin = 3 dB
Information Rate R_b = 256 kb/s
E_b/N_0 = 6 dB
Soft handoff gain = 2 dB
Fast fading margin = 3 dB
Log-normal fade margin = 8 dB

The previous values of E_b/N_0 and fade margin provide 90 percent coverage near cell boundaries and 95 percent coverage over the entire serving area.

Suppose that the path loss from the mobile station to the base station is P_L. Then the input to the base station receiver p_{in} = 24 − Penetration Loss − Path Loss + BTS Receiver Antenna Gain − Receiver Cable Loss = 24 − 8 − P_L + 16 − 2 = 30 − P_L dBm.

Noise and interference density of the receiver = Noise Figure + Interference Margin + Noise Floor = 5 + 3 − 174 = −166 dBm/Hz. So, the required input to the receiver = −166 + E_b/N_0 + Log-normal Fade Margin + 10log (R_b) = −166 + 6 + 8 + 10 log (256,000) = −97.9.

So, the required input = −97.9 − Soft Handover Gain = −97.9 − 2 = −99.9. Hence,

$$p_{in} \geq -99.9 \text{ dBm, or } 30 - P_L \geq -99.9; \text{ that is,}$$

$$P_L \leq 30 + 99.9 = 129.9 \text{ dB}$$

Thus, the maximum allowable path loss is 129.9 dB.

Frequency Planning

Analog and TDMA Systems

When a new analog or TDMA cellular system is being built, an obvious requirement is to achieve the desired system capacity while maintaining a satisfactory signal-to-co-channel-interference ratio over the entire serving area with a high probability. The capacity can be increased by decreasing the cell size. As shown in Chapter 1, the signal-to-co-channel-interference ratio depends on the co-channel reuse ratio, D/R:

$$\frac{D}{R} = \sqrt{3N}$$

where R is the radius of a cell, D is the distance between two cells using the same channel set, and N is the number of cells in a cluster [4]. The assumption here is that all cells are equal in size and

hexagonal in shape. A satisfactory signal-to-co-channel-interference ratio is obtained with $N = 7$, in which case, $D = 4.6R$. This means that the available spectrum block is to be divided into seven sets, each of which can then be reused in other cells outside a cluster. This was shown in Chapter 1 and is depicted again in Figure 10-2 for the convenience of the reader. The numbers in the figure represent the channel sets assigned to the individual cells. For example, the shaded cells all use channel set 2.

Another important consideration in the assignment of channels is the adjacent channel interference. Interference from adjacent channels may sometimes be serious in a mobile radio environment. Consider, for example, Figure 10-3 where two mobile stations, M1 and M2, are transmitting on two adjacent channels [4]. If the distance r2 from mobile M2 to the base station is, say, 10 times or more greater than the distance r1 of mobile 1, the signal received at the base station from mobile M1 may be over 40 dB higher than the signal from M2. Furthermore, because of fading, it is possible that the signal from M2 is in a deep fade while the other signal is above its local mean. Thus, even though the *Intermediate Frequency* (IF) filter in the base station (or the mobile station) receiver is designed to provide significant attenuation outside its bandwidth, there may be enough energy from the adjacent channel in the band of interest that the signal-to-adjacent-channel-interference ratio would be unacceptably low. Consequently, it becomes necessary to assign the channels in such a way that there are no adjacent channels in the same set. This can be done very easily in the following way. Suppose that there are only 21 voice channels in the available spectrum block. The channels may then be divided into seven sets, each with three channels, using channels 1, 8, and 15 in set 1; channels 2, 9, and 16 in set 2; channels 3, 10, and 17 in set 3; and so on.

Notice, however, that in the previous assignment, even though no cell contains an adjacent channel, any two adjacent cells will always have some adjacent channels between them. For example, cell 1 has channels 1, 8, and 15, and cell 2 has channels 2, 9, and 16. This situation would not arise if larger clusters were used, that is, if the value of N were, say, 12, 13, 27, and so on. However, if N = 7, and if base stations are located at the center of a cell and use omnidirectional antennas, adjacent channels will cause interference in any cell. If, on

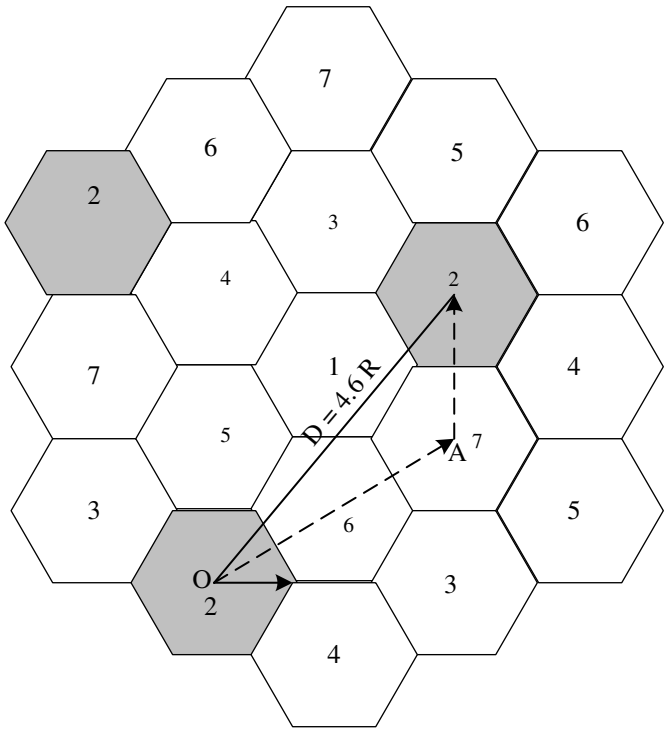


Figure 10-2
Reuse of channels in a cellular system. Each number represents a channel set. The co-channel reuse distance is D = 4.6 R.

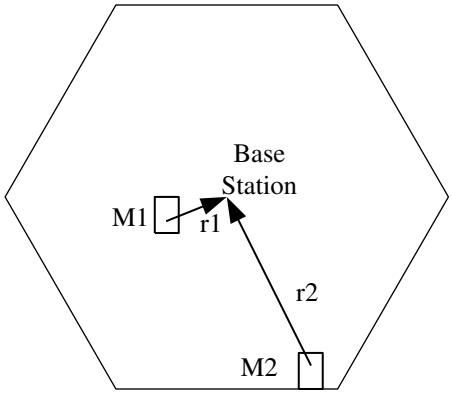


Figure 10-3
Adjacent channel interference

the other hand, base stations are placed at alternate corners of each cell, and 120 degree-antenna sectors are used at each site, it is then possible to allocate the channels to the three sectors in a way that

will minimize the adjacent channel interference [4]. This allocation is shown in Figure 10-4 where it is assumed that there are only 21 channels in the available spectrum block. Notice that adjacent channels belonging to two different channel sets are assigned to sectors facing away from each other. For example, channel set 1 (consisting of channels 1, 8, and 15) and channel set 2 (consisting of channels 2, 9, and 16) are still assigned to two cells as before, but now channels 1 and 2 are assigned to sectors that face away from each other toward different cells so that the interference to a mobile receiver in cell A on channel 1, caused by, say, a base station transmitting on channel 2 in cell B, is attenuated by the front-to-back ratio of the antenna. Thus, by simply allocating the channels to the directional antennas in a specific way, the interference due to adjacent channels can be reduced to a satisfactory level.

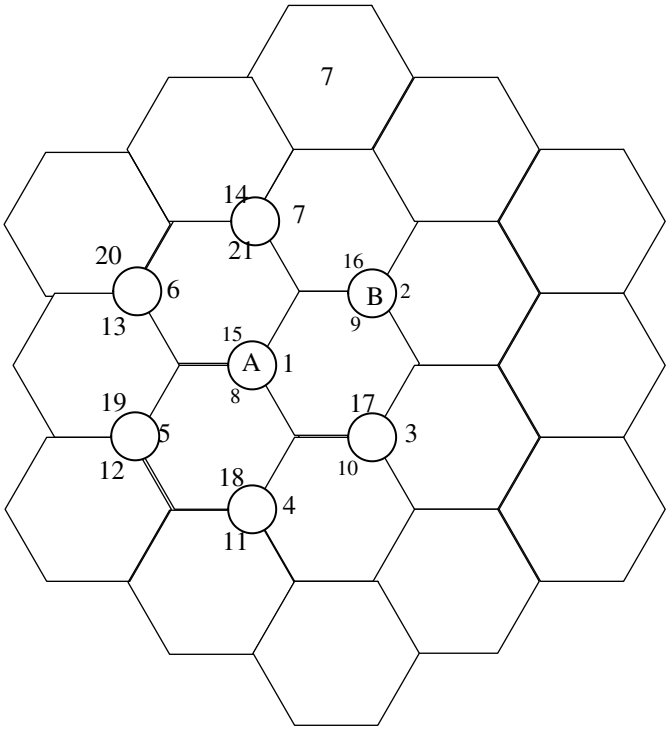


Figure 10-4
Channel allocation that minimizes adjacent channel interference using corner cells and three-sector directional antennas

CDMA System

Because N = 1 in a CDMA system, its frequency planning is quite straightforward. There are, however, a few things that must be taken into consideration. For example, if a CDMA system is going to be installed in an area that already has another system, say, GSM, and if the spectrum blocks allocated to the two systems overlap, clearly not all channels licensed by this service provider can be used in the CDMA base stations near the boundaries. Furthermore, if the signals in either system are sufficiently strong, in order to reduce interference an appropriate guard band may also have to be provided between the two systems. As a result, it is possible that some CDMA base stations, particularly those at the boundary, may not have all the channels that other base stations away from the existing system usually have [6]. If that is the case, these base stations will not be able to provide soft and softer handoff.

The other consideration is concerned with the time offsets for the I-channel and Q-channel pilot *pseudonoise* (PN) sequences that are used in synchronous CDMA systems (such as IS-95 and cdma2000) to spread the forward and reverse channel transmissions. Recall that these PN sequences, which are maximal length shift register sequences with a period of 2^{15} chips, are first offset in time (or phase) by an amount that is unique for each base station and then used to spread out the in-phase and quadrature components of the modulator output at a chip rate of 1.2288 Mc/s (or some integer multiples thereof). Because the offsets have to be different for different sectors in a cellular system, their allocation must be managed and coordinated in much the same way as the channels for cellular and TDMA systems [5].

Cellular System Growth

If the number of subscribers increases every year, at some point the traffic will reach or even exceed the capacity of the system. If that happens, the percentage of blocked calls will increase, thus leading to a degradation in the service quality. This problem can be overcome in a number of ways. For example, if there are any unused channels in

the available spectrum block, they may be added to the core of the serving area that experiences the highest growth in traffic. If the traffic is unevenly distributed over the serving area, some channels can be taken away from the low-usage cells and assigned to the core. For analog and TDMA systems, another possibility is to split cells and install base stations at the new sites. Finally, a new system, possibly using a different technology, may be added as an overlay on the existing system to handle the growing traffic. In fact, the new system may be planned and designed in such a way that it carries more and more of the traffic until eventually it serves all of the incoming traffic. In this case, the old system is no longer needed and may be discontinued. Obviously, there are other possibilities, but our discussion in this section is restricted to cell splitting and the overlay design only.

Cell Splitting

By cell splitting, we mean reducing the radius of coverage to half the older value so that the coverage area is reduced by a factor of 4. This is achieved by installing a new base station midway between two existing cells [4]. See Figure 10-5. Because the number of cells is increased four times, obviously the capacity also increases by the same amount. Notice that it is easier to perform cell splitting when base stations are located at alternate corners of the cells and use 120-degree directional antennas, because in that case, the hexagonal geometry of the new cells is identical with that of the larger cells. To see how channel sets should be assigned to the new cell sites, observe that a new cell site is added halfway between two old co-channel cells. For example, in Figure 10-5, cell site A has been placed at a midpoint between co-channel cells 2. Consequently, cell site A must be assigned channel set 2. This rule is followed for assigning channels to other new cells as well.

Overlay Design

When the traffic exceeds the system capacity, we could, instead of splitting cells, overlay a number of new, smaller cells on the existing system built around larger cells. When designing such a system, it is

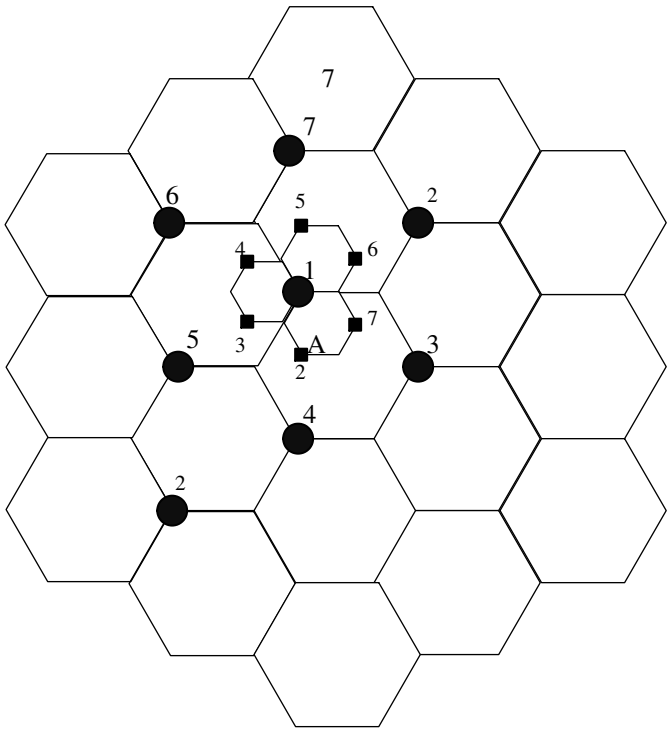


Figure 10-5
Cell splitting to increase the system capacity. The shaded circles represent the old cell sites with the larger radius of coverage. The shaded squares are the new cells with one-half the larger radius, each appearing midway between two adjacent cell sites.

necessary to understand how the co-channel interference is going to be affected in the new arrangement. The issues involved are explained in Figure 10-6 [4]. Here, cell sites A and B of the underlying system with larger cells use the same channel set. Now suppose that new cell site C, being located halfway between A and B, has also been assigned the same channel set. If mobile M1 is being served by C, the co-channel interference caused to M1 by cell sites A and B would be in the acceptable range because each of them is at a distance of 4.6r, where r is the radius of smaller cells. But if mobile M2 is being served by A, the co-channel interference that M2 experiences due to C may be quite severe because C is at a much shorter distance than 4.6R where R is the radius of larger cells. Consequently, the channels that are being used in the new cells cannot be used in the old cells. So, as the traffic grows, the overlay system is built by taking more and more of the channels from the existing system and allocating them to the new.

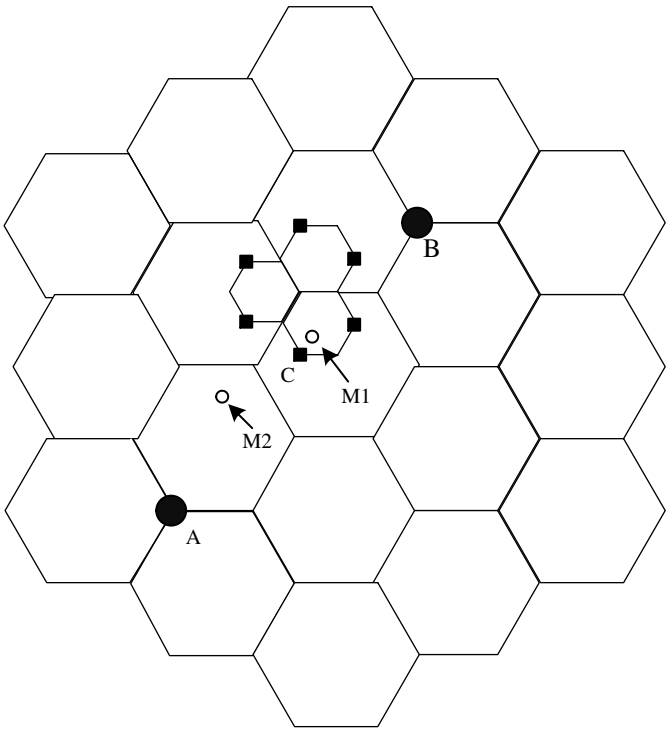


Figure 10-6
Overlay of a few small cells over an existing system of larger cells to meet increased traffic

Reference [7] gives an elaborate example of how a CDMA system can be designed as an overlay on an analog system so that the portion of the traffic carried by the new system is gradually increased. The idea is rather simple. Initially, the entire traffic is carried by the old system. When the traffic begins to grow, a new CDMA system is built with larger cells and is overlaid on the existing system. At the beginning, only a small portion of the spectrum block, say, the equivalent of a few channels, is taken from the analog system and used in the CDMA network. As the traffic increases, more and more channels from the old system are assigned to the CDMA system, thus enabling it to carry a larger portion of the incoming traffic. The number of channels used in the new system at any time depends on the traffic to be carried and the call-blocking probability.

Summary

In this chapter, fundamental principles of the planning and design of a cellular network have been described. The planning begins with a set of requirements such as the total traffic, the average signal-to-interference ratio for a desired coverage criterion, the data rate, the propagation characteristics of the serving area, the desired technology, the available spectrum, and so on. These requirements are specified by the service provider. The goal of the design is to provide the required coverage with a minimum number of cell sites. The design involves calculating the link budget to estimate the allowable path loss for a given transmitter power and SNR at the receiver. These ideas are explained using a number of numerical examples. Other aspects of the network design include frequency planning, coordination of the time offsets of the I and Q pilot PN sequences of a synchronous TDMA system, and cellular system growth. Frequency planning of a CDMA system is quite simple because the same channels can be used at every base station. If there are two or more adjacent systems in a serving area, the frequencies of the cell sites at their boundaries must be carefully planned so as to avoid co-channel and adjacent channel interference. We have also discussed how the system capacity can be increased by splitting cells or overlaying a new system on an existing system without requiring any new spectrum allocation.

Appendix A — Traffic Capacity of a Network

Suppose that a network consists of n circuits and that calls are arriving randomly so that the offered load is L Erlangs. Then, assuming that blocked calls are not redialed into the network, the call-blocking probability is given by the following Erlang B formula:

$$p_b = \frac{L^n/n!}{\sum_{i=0}^{n} L^i/i!}$$

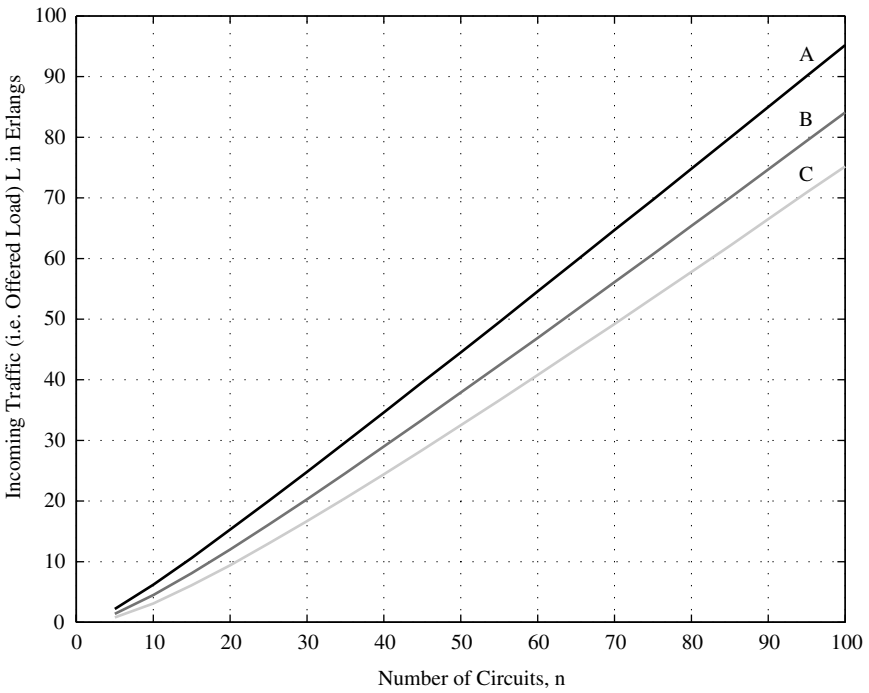


Figure 10-7
Offered load in Erlangs as a function of the number of circuits for a few values of the call-blocking probability p_b. (A – p_b = 5%, B – p_b = 1%, C – p_b = 0.1%)

The offered load (L) is plotted in Figure 10-7 as a function of the number of circuits for a few values of the call-blocking probability p_b.

References

[1] W.C. Jakes, Jr., "A Comparison of Specific Space Diversity Techniques for Reduction of Fast Fading UHF Mobile Radio Systems." *IEEE Trans. Veh. Tech,* Vol. VT-20, No. 4, November 1971, pp. 81–92.

[2] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, "Field Strength and Its Variability in VHF and UHF Land-Mobile Radio Service," *Rev. Elec. Communication Lab*., Vol. 16, pp. 825–873, 1968.

[3] M. Hata, "Empirical Formula for Propagation Loss in Land Mobile Radio Services," *IEEE Trans. Veh. Tech*., Vol. 29, May 1980.

[4] V.H. MacDonald, "The Cellular Concept," *Bell. Syst. Tech. J.,* Vol. 58, No.1, January 1979, pp. 15–41.

[5] Lucent Technologies, "Personal Communication Services, CDMA RF Engineering," *Internal Publications,* 1998.

[6] Lucent Technologies, "AUTOPLEX Cellular Telecommunications System, System 1000, CDMA RF Engineering Guidelines," *Internal Publications,* 1998.

[7] V.K. Garg, *IS-95 and cdma2000.* New Jersey: Prentice Hall, 2000, pp. 255–268.

[8] D. Bear, *Principles of Telecommunication Traffic Engineering.* London: Peter Peregrinus Ltd., 1988, pp. 236–237.

[9] R.D. Rosner, *Packet Switching.* Belmont, California: Lifetime Learning Publications, 1982, p. 59.

This page intentionally left blank.

TEAMFLY

CHAPTER 11

Beyond 3G

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

Third-generation (3G) wireless systems have been made possible by the concerted efforts of research and development organizations, equipment manufacturers, service providers, and above all, standards bodies that took the initiative in coordinating the works of these organizations, harmonizing the specifications coming out of the laboratories, and producing a set of international standards. Many companies are now in the process of manufacturing 3G wireless equipment so that service providers can begin to roll out 3G services to their customers in the near future. NTT DoCoMo of Japan has been actively developing 3G technologies. Lucent Technologies of the United States has developed base station equipment (such as 3G1X and 3G3X) that will provide 3G services for cdma2000, UMTS, and UWC-136, and is currently conducting field trials of their equipment with DoCoMo, Vodafone, Sprint, Bell Atlantic, AT&T, and SBC. NTT DoCoMo had targeted year 2001 for nationwide deployment of a commercial 3G system.

Service providers, equipment manufacturers, and research laboratories have already begun looking beyond 3G. NTT DoCoMo is working on their vision of the *fourth-generation* (4G) system, and has started defining and developing 4G services and technologies [1]. Many research and development projects in Europe are studying broadband communication issues (such as the air interface standards, cell coverage characteristics, provision of the *quality of service* [QoS] in all IP networks, and so on) that would be relevant to 4G [2]–[5].

Driving Force Behind 4G

The demand for mobile telephone services has been phenomenal. Since the introduction of cellular telephony in 1981, the annual growth in the number of mobile subscribers has been about 40 percent, whereas telephony services over fixed networks has grown at a modest rate of about 5 to 7 percent over the same period. Most of the traffic in present day mobile telephony consists of voice. However, the demand for mobile data has gone up steadily, spurred to a large

extent by the availability of the Internet-based applications. This is apparent from the fact that the data service capability of earlier *Global Systems for Mobile Communications* (GSMs) was limited to short messaging service and circuit-switched data at rates of up to 9.6 kb/s. As the demand for data services began to grow, the *European Telecommunications Standards Institute* (ETSI) developed a standard for *General Packet Radio Service* (GPRS), which is now being deployed in many countries of the world to provide packet mode data services at 12–20 kb/s per slot. The forecasted growth in the voice and data traffic in mobile communications is shown in Figure 11-1.[1] Although the numbers are approximate, the important thing to notice here is that the mobile data traffic is expected to be double the voice traffic by around 2010 and increase by a factor of about 24 in the year 2015. If the traffic does indeed continue to grow at this rate, 2G and 3G systems may run out of capacity by that time, and so it may be necessary to consider allocating new radio spectrum to wireless communications for next generation systems.

The source of most of the traffic in the above scenario is expected to be multimedia services. Even though there would be some applications where the traffic is interactive, the data transport will be asymmetrical—the traffic downlink will be much more than the uplink traffic. In 3G, multimedia services for mobile and outdoor applications will most likely operate at 384 kb/s. On fixed networks, video conferencing routinely uses 384 kb/s over fractional T1 lines or a primary ISDN line using bundled B channels. Many customers, currently are also using multimedia services over these networks at 1.536 or 2 Mb/s. If, as is usually the case, mobile customers expect to be provided the same services as are routinely available over fixed networks, mobile networks will be required to support multimedia services at 2 Mb/s. Similarly, there are applications for indoor environments, such as full-motion video, home entertainment systems, and so on that require data rates at 10–20 Mb/s. Since the maximum data rate for fixed and indoor applications in 3G only goes up to

[1]However, it is important to mention here that the past growth is no guarantee for future growth and that extrapolation is very speculative.

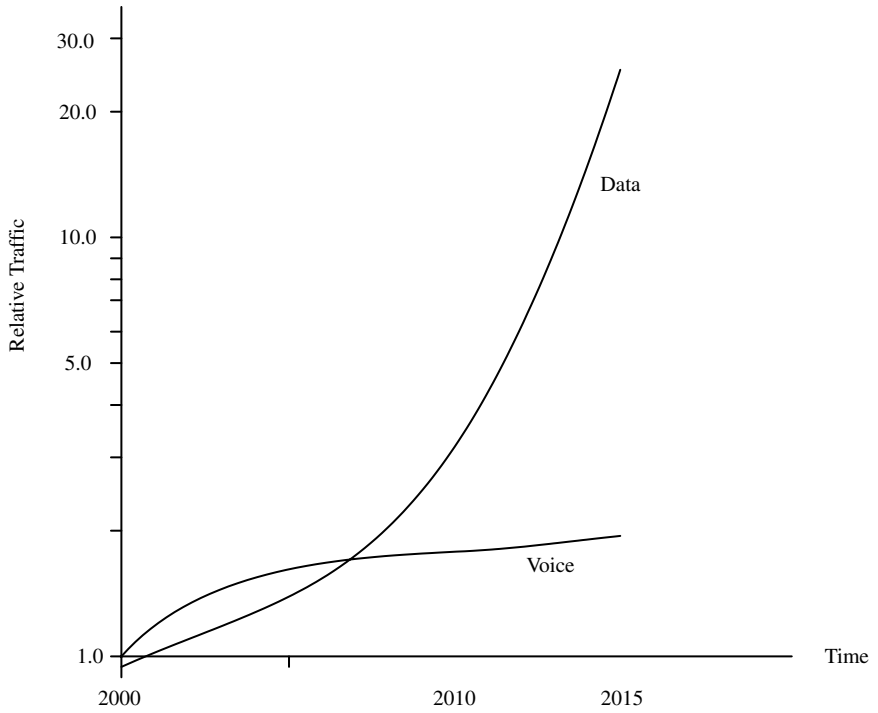


Figure 11-1
Growth of voice and data traffic in mobile telephony

2 Mb/s, that would be another reason to consider 4G systems. Assuming that the 4G system is indeed a possibility, we have highlighted the features and applications of various systems in Figure 11-2.

Applications and Features of 4G

Possible applications of 4G include multimedia services for mobile environments (such as typical vehicular, high-speed vehicular, aeronautical, satellite, and so on) at rates up to 2 Mb/s, *compact disc* (CD)-quality radio broadcasting, video surveillance of one's home when travelling, full-motion video and home entertainment at rates up to 20 Mb/s for indoor applications, position locating systems, and so on. The goal here is to provide multimedia services to anyone, anywhere, anytime. The usual data rates for high-quality video are

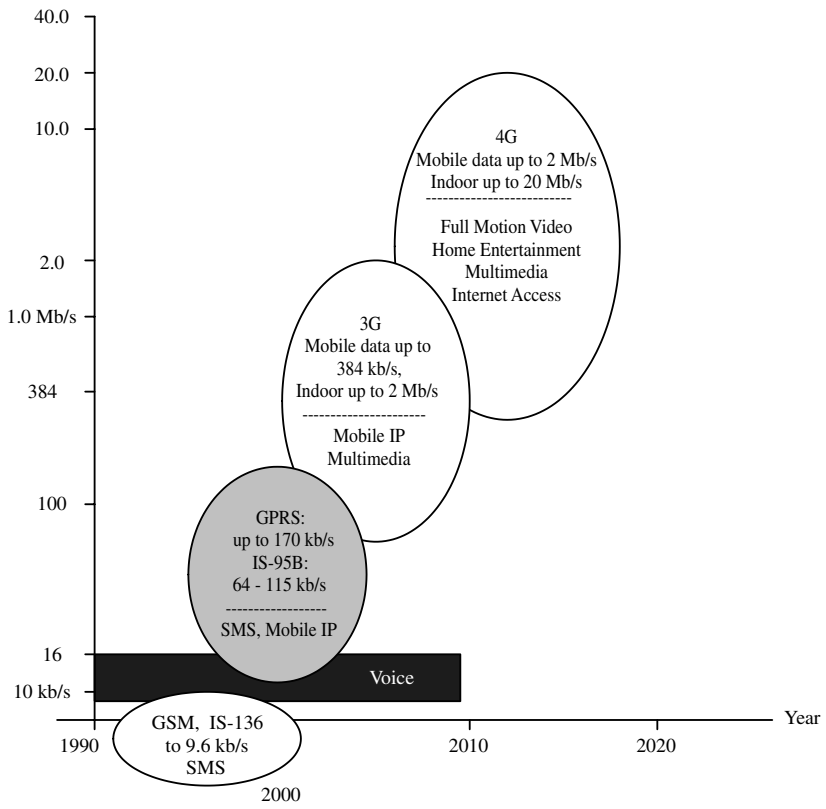


Figure 11-2
Various systems with their services and applications

shown in Table 11-1, where, for comparison, ISDN and PSTN video phones have also been included.

Some of the 4G features are

- Unrestricted, seamless roaming and global mobility not only for voice, but also for data services over regional and global networks that would, in all likelihood, have an all-IP architecture.
- Interoperability between 3G and 4G and between 2G and 4G.

As for the time frame of 4G services, analog systems were introduced in the United States and Europe in 1981. Within 10 years, around 1991, digital systems were deployed. The 3G systems are targeted for introduction in 2001 and 2002. Thus, it is reasonable to

Table 11-1

Data rates for different types of video

	Video Resolution (pels[2] × line × frames/s)	Uncompressed Bit Rate	Compressed Bit Rate
Film (USA and Japan)	480 × 480 × 24	133 Mb/s	3–6 Mb/s
NTSC video	480 × 480 × 29.97	168 Mb/s	4–8 Mb/s
Phase Alternating Line (PAL) video	576 × 576 × 25	199 Mb/s	4–9 Mb/s
High-Definition Television (HDTV) video	1920 × 1080 × 30	1493 Mb/s	18–30 Mb/s
ISDN video phone	352 × 288 × 29.97	73 Mb/s	64–1920 kb/s
PSTN video phone	176 × 144 × 29.97	18 Mb/s	10–30 kb/s

[2]A digital image is represented by a two-dimensional array of samples. Each sample is called a picture element (PEL) or pixel. A color image consists of red, green and blue. There are 24 bits or more per sample.

assume that a probable time frame for 4G would be around 2010–2012. Even though the present systems have not reached a saturation point, it is necessary to consider allocation of new spectrum for 4G. Since it takes about three to four years to develop standards, the standards work might begin in the next two to four years.

Technologies

Technologies that are likely to play a key role in the development and eventual success of 4G and, to a lesser extent, 3G are

- *Software radio [6], [7]* In software radio, most of the processing is done with digital signal processors. For example, baseband data processing (such as pulse shaping, error coding, and so on), modulation and up-conversion at the transmitter, channel separation, demodulation, detection, and baseband data processing at the receiver are all performed in the digital domain. In addition, digital signal processing is used to characterize the channel and adjust the power level as needed,

analyze the received signal to determine the quality (such as the *bit error rate* [BER], *forward error correction* [FER], etc.), reduce or cancel interference from specific sources, implement multipath diversity, and so on.

- *Adaptive antenna arrays [8], [9]* Adaptive antennas may be used for beam forming in a specific direction so as to provide extended coverage in certain areas or for selected mobiles and steering nulls in the direction of co-channel cells, and so on.
- *Development of suitable multimode configurable terminals, special keyboards, and video displays [10]* We might consider wearable PCs with voice-activated, hands-free operation, or specially designed keyboards that can be strapped to a wrist, and small LCD displays with magnifying optics attached to, or reflected onto, the user's glasses.
- *QoS* For efficient utilization of bandwidth, the network must implement a flexible resource management scheme to provide mobile stations with an end-to-end QoS across all-IP networks. Users must be able to roam seamlessly across multiple serving areas or across networks of different service providers.
- *Reduced power levels* Since the data rates in a 4G system are much higher than for 2G or 3G, the power consumption of the terminals will increase. Because an increased battery life is a desirable user feature, terminals should be designed to operate at reduced power levels.

Other Considerations

The discussion so far has been strictly from a technical point of view. Clearly, technologies already exist or are in the process of development that are necessary for successful deployment of 3G and 4G services. There is, however, an economic perspective as well. For example, nearly one hundred billion dollars were spent for 3G licenses in Western Europe. In addition, a considerable amount of capital is needed for the 3G infrastructure. Thus, to ensure a reasonable rate of return from this capital investment, it is necessary to generate sufficient customer demand for 3G services and continue to use the infrastructure for 4G and possibly beyond. For this to be pos-

sible, it would be necessary to develop applications that are meaningful and attractive to customers, and at the same time commercially viable from the standpoint of service providers.

References

[1] "The Path to 4G Mobile," Interview with N. Nakajima, NTT DoCoMo, *IEEE Commun. Mag.,* Vol. 39, No. 3, March 2001, pp. 38–41.

[2] M. Dinis and J. Fernandes, "Provision of Sufficient Transmission Capacity for Broadband Mobile Multimedia: A Step Toward 4G," *IEEE Commun. Mag.,* Vol. 39, No. 8, August 2001, pp. 46–54.

[3] T. Tjelta, et al., "Future Broadband Radio Access Systems for Integrated Services with Flexible Resource Management," *IEEE Commun. Mag.,* Vol. 39, No. 8, August 2001, pp. 56–63.

[4] T. Robels, et al., "QoS Support for an All-IP System Beyond 3G," *IEEE Commun. Mag.,* Vol. 39, No. 8, August 2001, pp. 64–72.

[5] L. Becchetti, et al., "Enhancing IP Service Provision over Heterogeneous Wireless Networks: A Path Toward 4G," *IEEE Commun. Mag.,* Vol. 39, No. 8, August 2001, pp. 74–81.

[6] J. Mitola, "The Software Radio Architecture," *IEEE Commun. Mag.,* Vol. 33, No. 5, May 1995, pp. 26–38.

[7] R.J. Lackey and D. W. Upmal, "Speakeasy: The Military Software Radio," *IEEE Commun. Mag.,* Vol. 33, No. 5, May 1995, pp. 56–61.

[8] J. Kennedy, et al., "Direction Finding and 'Smart Antennas' Using Software Radio Architectures," *IEEE Commun. Mag*., May 1995, Vol. 33, No.5, pp. 62–68.

[9] G. Tsoulos, et al., "Wireless Personal Communications for the 21st Century: European Technological Advances in Adaptive Antennas," *IEEE Commun. Mag.,* Sept. 1997, Vol. 35, No. 9, pp. 102–109.

[10] S. Ditlea, "The PC Goes Ready-to-Wear," *IEEE Spectrum*, October 2000, pp. 34–39.

APPENDIX

List of Abbreviations and Acronyms

AAL	ATM Adaptation Layer
AAL2	ATM Adaptation Layer 2
AAL5	ATM Adaptation Layer 5
AC	Authentication Center
ACCH	Associated Control Channel
ACELP	Algebraic Code Excited Linear Prediction
ACK	Acknowledgment
ADPCM	Adaptive Differential Pulse Code Modulation
AGCH	Access Grant Channel
AI	Acquisition Indicator
AICH	Acquisition Indicator Channel
AIN	Advanced Intelligent Network
ALCAP	Access Link Control Application Part
AMPS	Advanced Mobile Phone Service
AMR	Adaptive Multirate
ANSI	American National Standards Institute
AP	Access Preamble; Applications Processor
AP-AICH	Access Preamble Acquisition Indicator Channel
ARIB	The Association of Radio Industries and Business
ARP	Address Resolution Protocol
ARQ	Automatic Repeat Request
AS	Access Stratum
ASC	Access Service Class
ATM	Asynchronous Transfer Mode
AUC	Authentication Center
AWGN	Additive White Gaussian Noise

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

BC	Broadcast
BCCH	Broadcast Control Channel
BCH	Broadcast Channel; Bose-Chaudhuri-Hocquenghem
BCS	Block Check Sequence
BER	Bit Error Rate
BFSK	Binary Frequency Shift Keying
B-ISDN	Broadband ISDN
BLER	Block Error Rate
BMC	Broadcast/Multicast Control
BPSK	Binary Phase Shift Keying
BS	Base Station
BSC	Base Station Vontroller
BSS	Base Station Subsystem
BSSGP	Base Station System GPRS Protocol
BSSMAP	Base Station System Mobile Application Part
BTS	Base Transceiver Station
CBR	Constant Bit Rate
CC	Call Controls
CCCH	Common Control Channel
CCH	Control Channel
CCITT	International Telegraph and Telephone Consultative Committee
CCPCH	Common Control Physical Channel
CD/CA-ICH	Collision Detection/Channel Assignment Indicator Channel
CDMA	Code Division Multiple Access
CELP	Code Excited Linear Prediction
CFN	Connection Frame Number
CN	Core Network
CNR	Carrier-to-Noise Ratio

TEAMFLY

CPCH	Common Packet Channel
CPCS	Common Part Convergence Sublayer
CPFSK	Continuous Phase Frequency Shift Keying
CPICH	Common Pilot Channel
CRC	Cyclic Redundancy Check
CRNC	Controlling Radio Network Controller
C-RNTI	Cellular Radio Network Temporary ID
CS	Circuit Switched
CSCF	Call State Control Function
CSICH	CPCH Status Indicator Channel
CTCH	Common Traffic Channel
CTD	Cell Transfer Delay
DCA	Dynamic Channel Allocation
DCCH	Dedicated Control Channel
DiffServ	Differentiated Services
DL	Downlink (Forward Link)
DN	Directory number
DNS	Domain Name System
DPCCH	Dedicated Physical Control Channel
DPCH	Dedicated Physical Channel
DPDCH	Dedicated Physical Data Channel
DQPSK	Differential Quadrature Phase Shift Keying
DRNS	Drift Radio Network Subsystem
DS	Differentiated Service, Direct Sequence
DS-CDMA	Direct-Sequence Code Division Multiple Access
DSCH	Downlink Shared Channel
DSCP	Differentiated Service Code Point
DSMA-CD	Digital Sense Multiple Access with Collision Detection
DSP	Digital Signal Processing
DSSS	Direct-Sequence Spread Spectrum

DTAP	Direct Transfer Application Part
DTCH	Dedicated Traffic Channel
DTX	Discontinuous Transmission
EDGE	Enhanced Data Rates for GSM Evolution
EIA	Electronic Industry Association
EIR	Equipment Identity Register
EIRP	Effective (or Equivalent) Isotropic Radiated Power
ETSI	European Telecommunications Standards Institute
FACCH	Fast Associated Control Channel
FACH	Forward Access Channel
FAUSCH	Fast Uplink Signaling Channel
FCCH	Frequency Correction Channel
FCS	Frame Check Sequence
FDD	Frequency Division Duplex
FEC	Forward Error Correction
FER	Frame Error Rate
FIFO	First In First Out
FN	Frame Number
FSK	Frequency Shift Keying
FTP	File Transfer Protocol
GF	Galois Field
GGSN	Gateway GPRS support node
GMSK	Gaussian Minimum Shift Keying
GP	Guard Period
GPRS	General Packet Radio System
GSM	Global System for Mobile Communication
GSN	GPRS Support Node

GTP	GPRS Tunneling Protocol
GTP-U	GPRS Tunneling Protocol in the User Plane
HCS	Header Check Sequence, Hierarchical Cell Structure
HLR	Home Location Register
HO	Handover
HSS	Home Subscriber Server
ICMP	Internet Control Message Protocol
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IMEI	International Mobile Station Equipment Identity
IMT-2000	International Mobile Telecommunication 2000
IntServ	Integrated Services
IP	Internet Protocol
IPv4	IP Version 4
IPv6	IP Version 6
IS	Integrated Services, Interim Standards
IS-54	Interim Standard 54 for U.S. Digital Cellular
IS-95	Interim Standard 95 for U.S. Code Division Multiple Access
IS-136	Interim Standard 136 for U.S. Digital Cellular with Digital Control Channels
ISDN	Integrated Services Digital Network
ISO	International Standards Organization
ISP	Internet Service Provider
ITU	International Telecommunication Union
ITU-R	ITU — Radio Communications Sector
ITU-T	ITU — Telecommunications Standardization Sector
IWF	Interworking Function

kb/s	Kilobits per second
ksps	Kilo symbols per second
LAC	Link Access Control
LAN	Local Area Network
LAPD	Link Access Procedures on the D Channel
LLC	Logical Link Control
LP	Linear Prediction
LPC	Linear Predictive Coding
LSA	Local Service Area
LSF	Line Spectral Frequency
LSN	Last Sequence Number
LSP	Line Spectral Pair
LTP	Long Term Predictor
M3UA	Message Transfer Part 3 — User Adaptation layer
MAC	Medium Access Control
MAP	Mobile Application Part
Mb/s	Megabits per second
Mc/s	Million chips per second
MER	Message Error Rate
MGCF	Media Gateway Control Function
MGW	Media Gateway
MIP-LR	Mobile IP with Location Register
MM	Mobility Management
MPE	Multimedia Processing Equipment
MPEG	Moving Picture Experts Group
MRF	Multimedia resource function
MRSVP	Mobile RSVP
MS	Mobile Station
MSC	Mobile switching center
MSK	Minimum Shift Keying

MT	Mobile Terminal
MTP	Message Transfer Part
MTP3-B	Message Transfer Part Level 3 of Broadband Signaling System 7 for Q.2140
MUD	Multiuser Detection
MUI	Mobile User Identifier
NACK	Negative ACK
NAS	Nonaccess Stratum
NMT	Nordic Mobile Telephone
NNI	Network-Node Interface
NS	Network Service
NSAPI	Network (layer) Service Access Point Identifier
NSS	Network Subsystem
NTSC	National Television Standards Committee
OA&M	Operations, Administration, and Maintenance
ODMA	Opportunity-Driven Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OMC	Operations and Management Center
OQPSK	Offset QPSK
OVSF	Orthogonal Variable Spreading Factor
PACCH	Packet Associated Control Channel
PAD	Packet Assembler and Disassembler
PAGCH	Packet Access Grant Channel
PBCH	Packet Broadcast Channel
PC	Power Control
PCCH	Paging Control Channel
PCCPCH	Primary Common Control Physical Channel
PCH	Paging Channel
PCM	Pulse Code Modulation

PCPCH	Physical Common Packet Channel
PCS	Personal Communications System
PCU	Packet Control Unit
PDCH	Packet Data Channel
PDCP	Packet Data Convergence Protocol
PDN	Public Data Network
PDP	Packet Data Protocol (IP, X.25, etc.)
PDSCH	Physical Downlink Shared Channel
PDTCH	Packet Data Transfer Channel
PDU	Protocol Data Unit
PHB	Per-Hop Behavior
PHY	Physical (Layer)
PI	Paging Indicator
PICH	Paging Indicator Channel
PLDCF	Physical Layer-Dependent Convergence Function
PLICF	Physical Layer-Independent Convergence Function
PLMN	Public Land Mobile Network
PLP	Packet Loss Probability
PMD	Physical Medium-Dependent
PN	Pseudonoise
PNCH	Packet Notification Channel
PPCH	Packet Paging Channel
PPM	Parts per million
PRACH	Packet Random Access Channel; Physical Random Access Channel
PS	Packet Switched
PSCH	Physical Shared Channel
PSPDN	Packet Switched Public Data Network
PSTN	Public Switched Telephone Network
PTM	Point-to-Multipoint

PTP	Point-to-Point
PVC	Permanent Virtual Circuit
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RA	Routing Area
RAB	Radio Access Bearer
RACH	Random Access Channel
RAI	Routing Area Identity
RANAP	Radio Access Network Application Part
RESV	Reservation (Message)
RF	Radio Frequency
RFC	Request for Comments
RL	Radio Link
RLC	Radio Link Control
RLCP	Radio Link Control Protocol
RLP	Radio Link Protocol
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RNTI	Radio Network Temporary Identity
RPE-LTP	Regular Pulse Excitation with Long-Term Predictor
RR	Radio Resources, Resource Reservation
RRC	Radio Resource Control
RRM	Radio Resource Management
RSM	Radio Subsystem Management
RSSI	Received Signal Strength Indicator
RSVP	Resource Reservation Protocol
RTP	Real-time Transport Protocol

SAAL-NNI	Signaling ATM Adaptation Layer — Network Node Interface
SACCH	Slow Associated Control Channel
SAP	Service Access Point
SAPI	Service Access Point Identifier
SAR	Segmentation and Reassembly
SCCH	Synchronization Control Channel
SCCP	Signaling Connection Control Part
SCCPCH	Secondary Common Control Physical Channel
SCH	Synchronization Channel
SCP	Service Control Point
SCTP	Stream Control Transmission Protocol
SDCCH	Stand-alone Dedicated Control Channel
SDU	Service Data Unit
SF	Spreading Factor; Stealing Flag
SFN	System Frame Number
SGSN	Serving GPRS Support Node
SIM	Subscriber Identity Module
SIR	Signal-to-Interference Ratio
SM	Short Message
SMG	Special Mobile Group
SMS	Short Messaging Service
SM-SC	Short Messaging Service Center
SMS-IWMSC	SMS Interworking MSC
SMS-GMSC	SMS Gateway MSC
SNDC	Subnetwork-Dependent Convergence
SNDCP	Subnetwork-Dependent Convergence Protocol
SRBP	Signaling Radio Burst Protocol
SRLP	Signaling Radio Link Protocol
SRNC	Serving Radio Network Controller
SRNS	Serving Radio Network Subsystem
SS7	Signaling System 7

SSCF-NNI	Service-Specific Coordination Function — Network Node Interface
SSCOP	Service-Specific Connection-Oriented Protocol
SSCS	Service-Specific Convergence Sublayer
STP	Short-Term Predictor; Signalling Transfer Point
SUF	Superfield
TACS	Total Access Communication System
TBF	Temporary Block Flow
TC	Transmission Convergence; Transactions Capabilities
TCAP	Transaction Capabilities Application Part
TCH	Traffic Channel
TCP	Transmission Control Protocol
TCTF	Target Channel Type Field
TD-CDMA	Time Division—Code Division Multiple Access
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TE	Terminal Equipment
TF	Transport Format
TFC	Transport Format Combination
TFCI	Transport Format Combination Indicator
TFCS	Transport Format Combination Set
TFI	Transport Format Indicator; Temporary Flow Identifier
TIA	Telecommunications Industry Association
TLDN	Temporary Local Directory Number
TMSI	Temporary Mobile Subscriber Identity
TOS	Type of service
TPC	Transmit Power Control
TrCH	Transport Channel
T-SWG	Transport Signaling Gateway

TTA	Telecommunications Technology Association
TTI	Transmission Time Interval
TLV	Type, Length, and Value
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink (Reverse Link)
UMTS	Universal Mobile Telecommunications System
UNI	User-Network Interface
U-RNTI	UTRAN Radio Network Temporary ID
USCH	Uplink Shared Channel
USF	Uplink State Flag
USIM	UMTS Subscriber Identity Module
UTRA	Universal Terrestrial Radio Access
UTRAN	Universal Terrestrial Radio Access Network
UWC	Universal Wireless Communications
UWCC	Universal Wireless Communications Consortium
VBR	Variable Bit Rate
VC	Virtual Circuit (or Channel)
VCC	Virtual Channel Connection
VCI	Virtual Channel Identifier
VLR	Visitor Location Register
VQ	Vector Quantization
VSELP	Vector Sum Excited Linear Prediction
W-CDMA	Wide-band Code Division Multiple Access
WFQ	Weighted Fair Queuing

TEAMFLY

INDEX

Symbols

A

B

Copyright 2002 M.R. Karim and Lucent Technologies. Click Here for Terms of Use.

C

D

E

F

G

H

I

J–L

M

N

O

P

Q

R

S

T

U

V

W

TEAM-DOC