mobile communications series

Toni Janevski

Traffic analysis and design of wireless IP networks

Traffic Analysis and Design of Wireless IP Networks

For a listing of recent titles in the *Artech House Mobile Communications Series*, turn to the back of this book.

Traffic Analysis and Design of Wireless IP Networks

Toni Janevski

Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data
Janevski, Toni.
Traffic analysis and design of wireless IP networks / Toni Janevski.
p. cm. — (Artech House mobile communications series)
Includes bibliographical references and index.
ISBN 1-58053-331-0 (alk. paper)
1. Wireless communication systems. 2. Telecommunication—Traffic. 3. Mobile communication systems. I. Title II. Series.

TK5103.2.J38 2003
621.382'15—dc21 2003041890

British Library Cataloguing in Publication Data
Janevski, Toni
Traffic analysis and design of wireless IP networks. — (Artech House mobile communications series)
1. Mobile communication systems—Design and construction 2. Wireless Internet
3. Telecommunication—Traffic I. Title
621.3'8456

ISBN 1-58053-331-0

Cover design by Igor Valdman

685 Canton Street
Norwood, MA 02062

International Standard Book Number: 1-58053-331-0
Library of Congress Catalog Card Number: 2003041890

10 9 8 7 6 5 4 3 2 1

To my wonderful sons, Dario and Antonio, and
to the woman of my life, Jasmina

Contents

Preface

Wireless networks have penetrated almost a billion subscribers worldwide with first and second generation mobile networks. The main service was voice, and more recently modem-based low-rate data services. Because of the voice-oriented traffic and circuit-switching technology, these networks are dimensioned and designed using the traditional traffic theory in telecommunications. Their design is based on high-cost centralized switching and signaling equipment and base stations as wireless access points. Another technology dominated the world in the wired local telecommunication networks: IP technology. The transparency of the *Internet Protocol* (IP) to different traffic types and low-cost switching equipment made it very attractive to operators and customers.

The *third generation* (3G) of mobile networks introduces wide spectrum and high data rates as well as variety of circuit-switched and packet-based services. It provides IP connectivity besides the circuit switching. Future generation mobile systems are expected to include heterogeneous access technologies, such as wireless LAN and 3G, as well as end-to-end IP connectivity (i.e., an all-IP network). The diversity of traffic services and access technologies creates new possibilities for both operators and users. On the other hand, it raises new traffic and design issues.

This book provides traffic analysis, dimensioning, *quality of service* (QoS), and design aspects for wireless IP networks with multiple traffic classes.

In Chapter 2 we provide a description of existing mobile systems, installed or standardized, from *second generation* (2G) towards the 2G+ and 3G mobile systems.

Internet protocols are the main subject in Chapter 3. We consider IP protocol version 4 and version 6, as well as the *Transport Control Protocol* (TCP), which is the most commonly used protocol on the transport layer in accordance

to OSI. We also describe mechanisms and protocols for introducing mobility and QoS support to the Internet.

Chapter 4 models telecommunications networks and provides the basis of the teletraffic theory (i.e., traffic theory for telecommunications).

Characterization and classification of IP traffic is the main issue in Chapter 5. Based on the statistical analysis of traffic traces from real measurements, IP traffic is classified into two main classes, A and B, and several subclasses.

Chapter 6 proposes architectures for wireless IP networks. It also provides traffic and mobility models that can be applied for traffic analysis.

An analytical framework for traffic analysis in mobile networks is given in Chapter 7. We considered single-class and multiclass mobile networks. Analyses are provided for different access technologies, such as *frequency/time division multiple access* (FDMA/TDMA) and *code division multiple access* (CDMA).

A hybrid admission control algorithm for wireless IP networks is proposed and discussed in Chapter 8. The proposed algorithm considers both call-level and packet-level.

Because of the burstiness of some traffic types (e.g., video traffic) and the random mobility of users, as well as a lack of analytical analysis in a closed form, we perform simulation analysis. Simulation analyses of wireless IP networks under different mobility and traffic parameters in the network are shown in Chapter 9.

Micromobility and location management in wireless IP networks are addressed in Chapter 10. We propose a handover scheme that locates handover management at the base stations by using handover agents.

Chapter 11 discusses scheduling and service differentiation in wireless IP networks. Existing solutions for wireless LANs and 3G networks are considered. Also, we give a design proposal for scheduling in multiclass wireless IP networks based on the traffic classification made in Chapter 5.

The main conclusions from the book are given in Chapter 12.

The material provided in this book is mainly targeted to telecommunications students, members of corporate mobile communications research and development departments, network designers, capacity planners, and anyone who finds the contents of this book helpful.

1

Introduction

1.1 Evolution Process

Cellular mobile networks made unforeseen development in the telecommunications field during the last decade of the twentieth century and the beginning of the twenty-first. Mobile communications are less pragmatic, and continue to demand higher bandwidths and different multimedia services for the end users. In addition, the *Internet Protocol* (IP) is technology that started to penetrate the world in the 1990s, as a result of the development of the *World Wide Web* (WWW) and the popularization of electronic mail (e-mail) communication on the Internet. The Web browser was the first widespread application to provide different multimedia services, such as browsing text and images, and streaming audio and video. Technological development in the 1990s and 2000s made computers smaller and smaller, thus allowing users to carry them while moving. The integration of wireless cellular networks and the Internet becomes a foreseen scenario, one that is being realized from the 3G standardization process and initiatives for future generations mobile networks (e.g., 4G and beyond), as well as from the introduction of mobility to the Internet, which was initially created for hosts attached to interconnected wired local computer networks.

Considering the development of telecommunications technology, one may distinguish among three key events (i.e., revolutions):

1. The introduction of automatic telephone exchange (at the end of the nineteenth century);
2. The digitalization of telecommunications systems from the 1970s to the 1990s;

3. The integration of circuit-switched connection-oriented telecommunications and packet-based connectionless Internet in the 1990s and 2000s.

The above path, in the last two steps, was also followed by mobile systems. Hence, *first generation* (1G) mobile cellular systems appeared in the 1980s. It provided only classical analog voice service. The *second generation* (2G) in the 1990s introduced digitalization of the communication link end-to-end as well as additional *Integrated Services Digital Network* (ISDN)-based services and modem-based data services. Data communication in 2G is provided with data rates of maximum 9,600 bps or 14,400 bps, which depends upon coding redundancy. The third generation mobile systems appeared in the 2000s (i.e., the first commercial systems started in 2002 in Japan and South Korea). The global initiation for standardization of 3G was placed within the *International Telecommunication Union's* (ITU) International Mobile Telephony–2000 (IMT-2000), which was created to coordinate different initiatives for 3G mobile systems from various developed countries: for example, *Universal Mobile Telecommunication System* (UMTS) in Europe and cdma2000 in the Americas. The 3G is created to support Internet connectivity and packet-switched services besides the traditional circuit-switched ones, with data rates ranging from 144 Kbps for fast moving mobiles to 2 Mbps for slow moving mobile users.

Future mobile networks are expected to provide end-to-end IP connectivity (i.e., they are expected to be wireless IP networks).

1.2 Why Wireless IP Networks?

The answer is not straightforward, and with each attempt one can include something either for or against them. The circuit-switched wired and wireless networks (e.g., 2G cellular networks) provide QoS support with appropriate signaling and control information. They are very well defined, robust, and hence very expensive systems. They are created mainly for deterministic voice service, although they can be also used for modem-based data communication. In addition, technological development in the 1990s made computers available for the mass market in developed countries, and the Internet gained momentum in the past 10 years by offering different multimedia content able to be accessed through *personal computers* (PCs).

In the telecommunications sector, the basic philosophy is always towards the balance between the costs and the quality (i.e., network operators and service providers tend to provide higher quality of service for lower costs so that end users can buy such services). Hence, it is not only a matter of whether the technology can support some services, but at what costs.

A telecommunications system is composed of two main parts: switching part and transmission part. Switching systems may be exchanges in circuit-switched telecommunications or routers in packet-based networks such as the Internet. Transmission systems are wired or wireless links that interconnect the switching systems. Also, there are links that connect users, fixed and mobile, to the switching systems, which forms the access network.

Then, there are two main costs for the network operators:

1. Equipment and installation costs;
2. Operation and maintenance costs.

For different media types and applications the above costs are lower when all content is carried over a single network than through different specialized networks because of the statistical multiplexing that reduces transmission and switching costs. Accordingly, in the early 1990s European countries began to develop *Asynchronous Transfer Mode* (ATM) as a technology that would provide a single network for different traffic types. The idea was to take the concept of "a single socket in the wall" for telecommunication services, similar to an electrical-power distribution network where different appliances can be plug into a same socket. Although well-defined, ATM had high network costs, so it mainly lost the battle with a simpler and cheaper solution. That solution is the Internet Protocol, which is transparent to different multimedia types. Furthermore, IP provides simple interconnection and maintenance of IP networks (i.e., local area networks) as well as low-cost switching systems (i.e., IP routers). Also, together with its main overlaying protocols, TCP and *User Datagram Protocol* (UDP), it provides support for different traffic types. Gaining global popularity via the WWW and e-mail, IP emerged as the clear winner over its opponents such as the ATM concept. The Internet provided a new type of economy in telecommunications via support of new multimedia services, as we discuss in Chapter 3.

Regarding voice service, mobile networks have largely reached market saturation in developed countries (e.g., European Union), so the introduction of IP services to existing mobile networks was considered a driving force, and it started with 2G+. The trend continued in 3G systems, which offer higher bandwidths than 2G but lower than wireless LANs. Wireless resources are limited over a given geographical area. Hence, the future generation of mobile networks is considered as an integration of the existing cellular networks and wireless LANs with added personalized mobile networks (e.g., WPAN) and broadband radio access networks. Only end-to-end IP networks with wireless access can accomplish such a task, and that is the answer to the question of why wireless IP networks should be considered.

Definition of a Wireless IP Network. A wireless IP network is an all-IP network with wireless access. All data, signaling, and control information are carried using IP packets. (*Note:* This definition is related to this book, and other authors may use the same term in a different manner.)

1.3 Traffic Issues

The Internet was created to be simple and transparent to different traffic types. But, considering the QoS, Internet basically supports one traffic type for all, which is called best-effort traffic. The creators of IP, however, have left options for introducing multiple traffic classes via the *Type of Service* (ToS) field in IPv4 header format, and lately via the *Differentiated Services* (DS) field in IPv6 headers. Integration of IP (i.e., Internet) and telecommunication networks for voice service highlights the QoS support in the Internet like never before. One traffic type for all does not well suit all applications. Also, some users may be willing to pay more for guaranteed QoS. The QoS support is especially important in wireless IP networks where resources are scarce and should not be wasted.

Dimensioning precedes initial network deployment. After the start of a network, the operator should perform traffic analysis and optimization of the network to maintain given QoS constraints. The design of a circuit-switched network with single traffic class (i.e., voice) is carried in telecommunications by using a traditional approach based on the Erlang-B formula. Traffic distribution and its parameters in wireless networks depend upon user mobility, cell size, bit rate of the wireless link (i.e., cell capacity), network load, scheduling at the base stations (i.e., wireless access points), handover, and location management. A multiclass environment requires network planners and designers to consider different traffic parameters for different classes. Hence, packet-based multiclass wireless networks raise new demands on the traffic analysis and network dimensioning.

In a wireless IP network there would simultaneously exist different traffic types, such as voice, audio, video, multimedia, and data. Applications can be classified into real-time (e.g., voice service) and nonreal-time (e.g., e-mail and Web browsing). Different traffic types have different characteristics. For example, voice service has low correlation and it is predictable. This is not the case with the bursty traffic, such as Web or video traffic. Therefore, one should use statistical analysis to obtain traffic characteristics. Furthermore, different traffic types have different QoS demands. Statistical characteristics and QoS requirements of different traffic types should be the main parameters for classification of the aggregate IP traffic.

The QoS requirements may be analyzed on different time scales and different levels (i.e., call-level and packet-level). However, best-effort traffic should

coexist with higher-class traffic, which has QoS demands. To provide certain quality within the given constraints on the quality measures, wireless IP networks need an appropriate admission control algorithm that will admit/reject calls depending upon the traffic conditions in the cell and its neighboring cells. So far, most of the admission control algorithms in multiclass networks are based only on a call-level or on a packet-level. But in heterogeneous IP networks one may find as the most appropriate solution to use hybrid admission control algorithms that consider call-level parameters (e.g., call blocking probabilities) and packet-level parameters (e.g., packet loss, delay). Also, different traffic types have different traffic parameters (e.g., bandwidth requirements, call rate, and so forth), which requires an analytical framework for dimensioning and optimization of multiclass wireless networks. In some cases where an analytical approach is not tractable, one should proceed with simulation analysis of traffic scenarios.

1.4 Design Issues

Wireless networks have their own characteristics. The two most important differences between the wired and wireless networks are mobility of the users and location-dependent bit errors on the wireless link. These specifics create significantly different conditions for QoS support.

Considering the QoS support for the Internet, there are several concepts proposed, analyzed, and implemented. First, chronologically, is the concept of Integrated Services, which is based on the end-to-end reservation of resources. To provide unified QoS support for different protocols, such as IP and ATM, which were developed independently, the *Multiprotocol Label Switching* (MPLS) concept was introduced. Finally, there is a Differentiated Services concept, which specifies by definition per-hop-behaviors instead of end-to-end services. This mechanism differentiates the aggregate traffic per class, and hence is scalable. All of these mechanisms are created for wired IP networks. But, integration of mobile networks and the Internet is a foreseen process. Therefore, QoS mechanisms are mapped from wired to wireless access networks.

Mobile Internet is already present via existing wireless LANs and 3G mobile networks. However, wireless LAN is based purely on the Internet principle in wired local networks, supporting best-effort class only. On the other hand, 3G mobile systems are a combination of circuit-switching and packet-switching technology. Simplified, 3G gets all the features of 2G systems and adds IP accessibility, as well as larger bandwidth than 2G cellular networks, but smaller than wireless LANs. In the future, mobile systems are expected to include heterogeneous access networks.

Future generation mobile networks are going to be all-IP networks; thus, all signaling, control, and data information should be carried using IP packets.

In such a situation an important issue at the network design level is micromobility management. Mobile IP protocol is defined as a standard for macromobility management (i.e., global mobility), but it is not efficient for local mobility. Several different solutions are proposed for micromobility management in IP-based wireless networks, such as Cellular IP, HAWAII, and others. There are several important design issues within the micromobility concept, including handover scheme, routing algorithm, and location control. Handover is a process of transiting an ongoing connection from one service area (i.e., cell) to another, and hence, it influences the flow and the ongoing traffic in the network. Therefore, one of the main goals of the design of wireless networks is a fast and transparent handover mechanism. It is closely related to the routing in the wireless access network and to the location control, both functions that should be adapted to the IP environment.

The second important characteristic of wireless networks is bit error ratio in the wireless channels (a definition of the wireless channel is given below). In circuit-switched cellular networks, mobile hosts measure the *bit error ratio* (BER) and signal strengths and send periodic reports to the base stations. Using the BER and signal strengths in the wireless channel, a centralized controller of the wireless access points decides whether to initiate a handover or not. Errors in the wireless channels influence the QoS of the affected flow(s). In wireless IP networks we have flows with variable data rate and different QoS requirements. Hence, service differentiation with appropriate scheduling of IP packets onto the wireless link is a challenging problem.

By default, wired routers on the Internet today use the *first-come first-serve* (FCFS) scheduling discipline. But this mechanism does not offer QoS support. Therefore, we should implement a more advanced scheduling discipline to provide service and flow differentiation. While scheduling in wired IP networks has reached its maturity, it is not the case with the wireless networks. Due to error-prone wireless channels, one should propose different or adapted scheduling mechanisms for wireless networks. There are also different proposals for design of scheduling mechanisms in wireless IP networks, such as *Idealized Wireless Fair Queuing* (IWFQ), *Channel-condition Independent Fair Queuing* (CIF-Q), and *Wireless Fair Service* (WFS). The design issue to consider is the provision of efficient service differentiation in a multiclass wireless IP network.

Definition of a Wireless Channel. A wireless channel is the amount of bandwidth that is allocated to a mobile user at a given time. The bandwidth allocation may be provided as frequency band(s), time slot(s), access code(s), or their combination(s). It does not mean that cell capacity is divided into circuit-switched channels. (*Note:* This definition is related to this book, and other authors may use the same term in a different manner.)

Overall, traffic analysis and design of wireless IP networks is not so straightforward. There are different possibilities and different solutions that can be applied. However, each solution might enhance certain parameters and worsen others, so there is no best single solution. In this book we provide existing solutions to the problems, as well as propose some methods, algorithms, and concepts that are helpful for traffic analysis and design of wireless IP networks.

2

Third Generation Wireless Mobile Communications and Beyond

2.1 Introduction

At the beginning of the twenty-first century, we are facing very fast development and deployment of two communication technologies: mobile networks and Internet.

Wireless communications had remarkable development in the last decade of the twentieth century. Figure 2.1 shows the exponential increase in the number of mobile subscribers in recent years. This growth was made possible due to the high-tech development of communication tools, which are no longer only voice-oriented as they were in the past. There are many nonvoice services that network providers offer to users. So, the paradigm of communication anywhere, anytime has become realistic. Today, telephony is still the primary service type in mobile networks, although low bit rate data services are also being supported. The lower prices, however, of laptop computers, palm devices, pagers, communicators, and personal organizers are increasing the requirements for multimedia services in mobile systems.

At the same time, Internet technology has been developing as fast as wireless networks. From the beginning of the Internet (formerly known as ARPANET), the number of users and host computers attached to the Internet doubled each year. Figure 2.2 shows the exponential growth of the Internet (for more details and precise numbers, a reader may refer to [1]). The spreading of the Internet throughout the world is hastened by the invention of the World Wide Web in 1993, which supports user-friendly browsing and retrieval of different types of information [2]. Total Internet traffic, however, increases faster

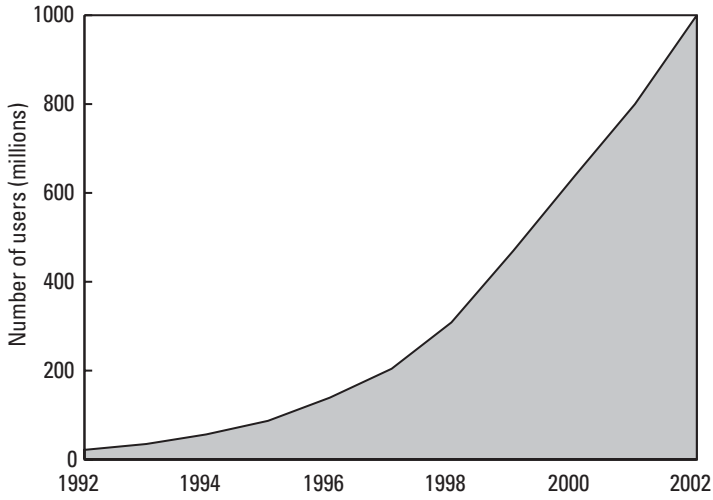


Figure 2.1 Growth of mobile users (a sketch).

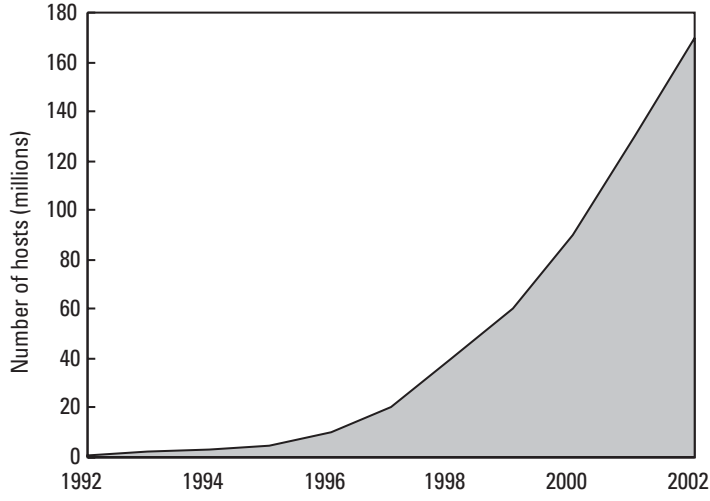


Figure 2.2 Growth of hosts on the Internet (a sketch).

than the number of hosts. The Internet is global and supports a variety of multimedia services. Hence, the Internet is becoming an integrated part of society and culture: in the science world, as well as in the community, entertainment, newspapers, administration, governments, interactive TV, and many more. No one can determine the boundaries of the Internet, if there are any. Of course, the development of the Internet became possible due to development of low-cost

personal computers and data networks for interconnection of individual PCs for exchanging data and sharing resources. As a technology, the Internet is based on the IP, which is robust enough and transparent enough to support transmission of different type of information (audio, video, data, and multimedia) by using IP packets. We refer to IP and TCPs in more detail in Chapter 3.

The fast development of these two technologies, wireless mobile communication and the Internet, goes towards their integration. Mobile network operators now seek new services to offer to users besides the voice service, because the mobile telephony market is almost saturated in the developed world (almost everyone has a mobile phone). On the other hand, Internet users are seeking connection to the Internet when they are on the move. In most of the cases, people are users of mobile networks and the Internet at the same time. Naturally, the users and the providers have interest in integrating these two technologies. So, although mobile networks and the Internet started separately—the first generations of mobile cellular systems (first and second generation) were created mainly for telephony service, while Internet was created for global exchange of data and communication between wired (fixed) hosts offering the same service level to all users—the development of these technologies leads toward their integration in mobile Internet or wireless IP networks. We notice this trend in the standardization processes of the third generation of mobile networks and beyond [3–7], as well as in mobility proposals for Internet technology [8–10].

2.2 Evolution of Wireless Communication

The early origins of wireless communication date back to 1861, when J.M.C. Maxwell at King's College in London proposed a mathematic theory of electromagnetic waves. Later, this theory was practically demonstrated by H. Hertz in 1887 at the University of Karlsruhe. Several years later, Guglielmo Marconi (at age 21) built and demonstrated the first real wireless communication device in summer of 1895 at the University of Bologna. It was the first radiotelegraph. Officially, it marks the start of the era of wireless communications.

The civilian use of wireless technology began with the 2-MHz land mobile radiotelephone system developed in 1921 by the Detroit Police Department for police car dispatch. Soon, the advantages of mobile communication were realized, but its wider use was limited due to a lack of channels in the low frequency band that was used at that time. Hence, higher frequencies were used. Armstrong made key progress in 1933 with the invention of *frequency modulation* (FM), which made possible high-quality two-way communication. Extending such technology (FM) to a large number of users required excessive bandwidth. The solution was found in dividing the service area into several

smaller service areas called cells, and using same subsets of radio channels in different cells. This cellular concept started in Bell Laboratories in 1947, by D. H. Ring.

With the invention of the cellular networks, the next problem that was faced was that of handover (or handoff, which we treat as a synonym for handover) between the cells. It should be transparent to the users. Seamless handover was successfully implemented by AT&T in 1970 in their analog cellular system, *Advanced Mobile Phone Service* (AMPS), which was placed in the 800-MHz band. The first commercial AMPS service did not begin until 1983.

In Europe, the first mobile systems started in the Scandinavian countries, in order to cope with their sparsely distributed population. The first such mobile system in Scandinavia started in 1978, but the real boom happened with analog *Nordic Mobile Telephony* (NMT) mobile system, which started in 1981. There are two versions of NMT: NMT 450 operating on the 450-MHz band, and NMT 900 operating on the 900-MHz band [11]. These systems marked the start of the first-generation mobile systems. Parallel to the NMT, the United Kingdom developed their *Total Access Communication System* (TACS), while Germany developed C-system, which was more advanced than NMT or AMPS systems due to its digital signalization and advanced power control in mobiles and base stations. Japan implemented a modified version of the British TACS mobile system, called *Japanese TACS* (JTACS). All the systems mentioned belong to the first generation. The basic characteristics for this generation include analog transmission of information and incompatibility of the systems in different countries.

Thus, each developed country in Europe developed its own system and standards for it, but different systems were incompatible with each other. This, of course, was less than ideal, since it limited the movement of the users and segmented the market for mobile equipment. European countries collectively realized this problem and decided to create a pan-European public land mobile system. Therefore, in 1982 the *Conference of European Posts and Telegraphs* (CEPT) formed a study group called *Groupe Speciale Mobile* (GSM) for that purpose. Later, in 1989 the standardization of GSM was transferred to the *European Telecommunication Standards Institute* (ETSI). Phase-I standards for GSM were published in 1990.

2.3 Second Generation Mobile Networks

The second generation of mobile systems (2G) was under way at the beginning of the 1990s. The first trials with GSM began in 1991, which changed its name for market reasons to Global System for Mobile communications. Soon, GSM overtook the wireless market, having around 700 million subscribers and more than 400 GSM operators by April 2002 [12]. These figures include

GSM 900, GSM 1800, and GSM 1900 mobile systems (we refer to GSM later in this chapter).

Applying the ISDN concept in the design of the GSM, it became a fully digital system. The main characteristic of GSM, besides the digital subscriber line, is roaming. With the introduction of roaming, GSM allows subscribers of one GSM network to use services in other GSM networks worldwide. These characteristics of GSM made it the world leader in 2G mobile systems considering the number of subscribers and network operators.

GSM technology is a combination of *frequency division multiple access* (FDMA) and *time division multiple access* (TDMA). GSM 900 systems were the first digital ones. They use the 900-MHz band. For each direction, uplink and downlink, 25 MHz of frequency spectrum was allocated. FDMA is used to divide the available 25 MHz of bandwidth into 124 carrier frequencies of 200 kHz each. Each frequency is then divided into eight time slots by using the TDMA technique (Figure 2.3). Two-way communication is made possible by assigning the same time slots on carriers 45 MHz apart from each other. Each pair of carriers is called the *absolute radio frequency channel number* (ARFCN). For example, ARFCN=1 uses 890.2 MHz in uplink and 935.2 MHz in downlink, while ARFCN=124 uses 915 MHz in uplink and 960 MHz in downlink direction. Uplink frequency spectrum is in the 890- to 915-MHz band (890.0 MHz is used as a guard channel), while downlink frequency spectrum is in the 935- to 960-MHz band (935.0 is also a guard channel). Each cell has one or more frequency carriers. A couple of time slots on one of the carriers in each cell are dedicated to signaling, while all others are used to carry traffic. In one logical channel several logical channels may be multiplexed. For example, usually 10 different logical signaling channels are multiplexed on two time slots in each cell.

Today, the GSM system operates in the 900-MHz and 1,800-MHz bands throughout the world, due to capacity demands, with the exception of the Americas where they operate in the 1,900-MHz band, due to frequency spectrum regulations.

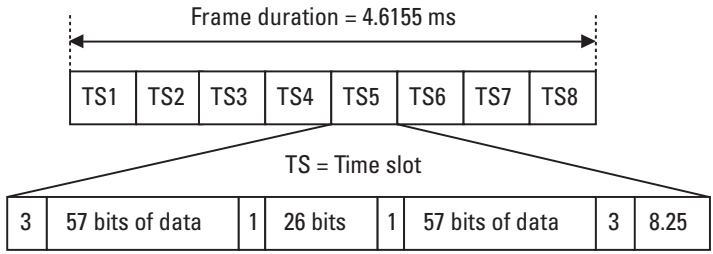


Figure 2.3 TDMA frame structure in GSM system.

Parallel to GSM, Japan developed similar TDMA-based technology called *Personal Digital Communications* (PDC). In North America *Digital AMPS* (D-AMPS) was launched, a successor to the analog AMPS. The upgrade of AMPS to D-AMPS is made by introducing three time slots per frequency carrier in AMPS, which are separated by 30 kHz. D-AMPS is known as the IS-54 standard, and it is also based on TDMA. Later, this system transited to IS-136, where IS-54 was improved by adding better performances and new services, such as *Short Message Service* (SMS). These two TDMA systems, D-AMPS and PDC, have been deployed worldwide and share the rest of the market, which is several times smaller than the GSM market share in 2G.

Very quickly, the capacity needs of 2G cellular mobile systems increased, and new approaches to cellular technology were needed. In 1993, the United States approved a new standard IS-95, proposed by Qualcomm, named *code division multiple access* (CDMA). IS-95 uses 1.25-MHz bandwidth that can be simultaneously used by many subscribers. The CDMA technique spreads the narrowband signal into wideband signal and assigns a unique code to each telephone or data call (we refer to CDMA technology in more detail in Section 2.5). It allows the use of the same frequency bands in the adjacent cells, simplifying the planning of the cellular network.

The roots of CDMA are in military communications, several decades ago. It was very popular due to its robustness to signal jamming. Because the signal occupies larger bandwidth, CDMA is known as spread spectrum technique. Each signal is spread over the whole dedicated bandwidth. In the receiver, the signal is extracted from the wideband signal by correlating it with the user code, which is unique for each traffic stream. In the downlink, the base station uses orthogonal spreading codes to communicate with multiple users using the same bandwidth. The mobile receives the signal by correlating the wideband signal with the known user code. In CDMA, multipath propagation of the signal (due to reflection of buildings, trees, and so forth) is found to be useful, due to diversity gain (i.e., accumulating signal power from different paths gives better signal-to-noise ratio). So-called RAKE receivers collect the energy from different paths. In the uplink direction each mobile spreads the signal using the user code. The base station extracts signals from individual connections by correlating the wideband signal with the user codes. To be able to perform multiple receptions simultaneously, it is important to have power control in the radio network, so each signal from mobiles arrives at the base station at the same power level.

The primary service in all 2G mobile systems is telephony. However, several data services are also supported, such as low data rate modem connections (up to 9,600 bps), fax, SMS, as well as supplementary services such as calling line identification presentation, call forwarding, conference call, call barring, and closed user group.

In the following section we give an architectural view of the GSM mobile systems.

2.3.1 GSM—State of the Art

The GSM system consists of the following subsystems: *base station subsystem* (BSS), *network and switching subsystem* (NSS), and *operation and support subsystem* (OSS), as shown in Figure 2.4.

BSS consists of *base transceiver stations* (BTSs) and the *base station controller* (BSC) [13]. The role of the BSS is to provide transmission paths between the mobiles and the NSS. The BTS is the radio access point, which has one or more transceivers. Each transceiver operates on one ARFCN at a given moment. The BSC monitors and controls several base stations (the number of BTSs under the control of a single BSC depends on the manufacturer, and it can be up to several hundreds of stations). The main functions of the BSC are cell management, control of a BTS, and exchange functions. The hardware of the BSC can be located on the same site with the *mobile switching center* (MSC), or at its own stand-alone site (e.g., in a case of several BSCs connected to a single MSC).

NSS includes switching and location management functions. It consists of the MSC, databases for location management [*home location register* (HLR) and *visitor location register* (VLR)], the *gateway MSC* (GMSC), as well as the *authentication center* (AuC) and *equipment identity register* (EIR) [13]. GMSC provides

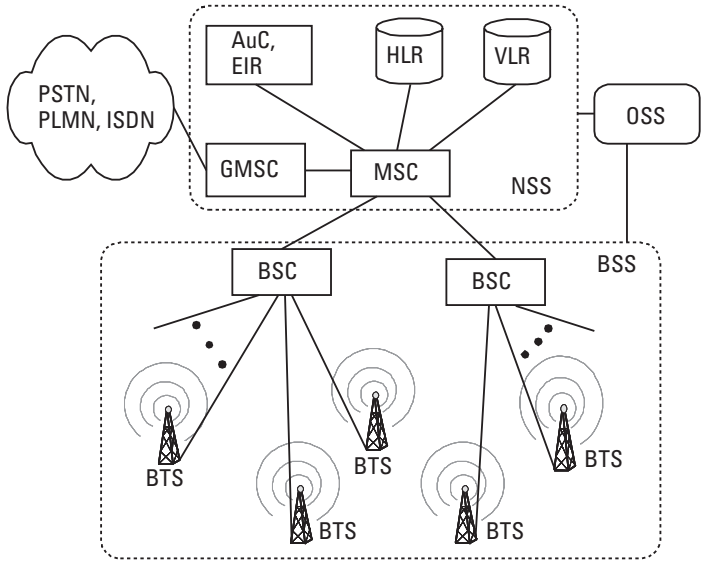


Figure 2.4 GSM network architecture.

interface between the mobile network and *Public Switched Telephone Network* (PSTN). MSC is a complete exchange with switching and signaling capabilities. It is capable of routing the calls from the BTS and BSC to mobile users in the same network (again via BSC and BTS) or to users in the PSTN (via GMSC) or to answering machines integrated within the MSC. Physically, MSC and GMSC may be integrated into one network element.

The HLR stores the identity and user data of all subscribers belonging to the mobile operator, no matter if they are currently located in the network or abroad (i.e., roaming). This data is permanent, such as the unique implicit number *International Mobile Subscriber Identity* (IMSI), explicit user's phone number the so-called *mobile station ISDN* (MS-ISDN) number, which is different than IMSI, the authentication key (necessary to protect the network from fraud), the subscriber's permitted supplementary services, and some temporal data. The VLR contains the permanent data as found in the HLR of the user's origin network, of all subscribers currently residing in its MSC serving area. Temporary data slightly differs from that of HLR. Thus, VLR contains data of its own subscribers of the network that are in its service area, as well as that of roamers from other GSM networks. Also, VLR tracks the users considering their residing location area. On the user's side, permanent and temporal user data is stored on *subscriber identification module* (SIM) cards, which are placed in the mobile phone.

The AuC is related to HLR and contains sets of parameters needed for authentication procedures for the mobile stations. EIR is an optional database that is supposed to contain the unique *International Mobile Equipment Identity* (IMEI), which is a number of the mobile phone equipment. EIR is specified to prevent usage of stolen mobile stations or to bar malfunctioning equipment (e.g., from certain manufacturer).

GSM is a system created mainly for telephony service, but it also supports low data rate modem connections up to 9,600 bps. For support of higher data rates in the radio access network (which are demanded by some multimedia services, such as Internet applications), GSM, on its way towards the third-generation mobile systems, is extended to the *General Packet Radio Service* (GPRS).

2.4 Evolution from 2G to 3G

The explosion of Internet usage has had a tremendous impact on the demand for advanced wireless communication services. However, the effectively rate of 2G mobile systems is too slow for many Internet services. As a result, in a race for higher speeds, GSM and other TDMA-based technologies from 2G developed so-called 2G+ mobile systems. In this group we classify the following

systems: *High Speed Circuit Switched Data* (HSCSD) and GPRS. One may also classify in the 2G+ group *Enhanced Data Rates for Digital Evolution* (EDGE), but it is somewhere referred to as 3G technology.

2.4.1 HSCSD

HSCSD is a software upgrade to the GSM networks. No extra hardware is required. In the GSM network, single time slots are allocated to each user for voice or data (via modem) connection. Standard data transfer rate in GSM is 9,600 bps, although by reducing the redundancy in the channel coding it may go up to 14,400 bps. HSCSD gives a single user simultaneous access to multiple channels (time slots), up to four of eight in a single TDMA frame. However, it is more expensive for end users to pay for multiple simultaneously occupied time slots.

Assuming a standard transmission rate of 14.4 Kbps and using four time slots with HSCSD allows a theoretical data rate of 57.6 Kbps. This enables Internet access at the same speed of many dial-up modem (56K) services across the fixed access network with 64-Kbps digital transmission lines. Although HSCSD is easy to be implemented in 2G networks, the drawback is the lack of statistical multiplexing (i.e., four time slots are occupied all the time during the connection). A potential problem in HSCSD is handover, which is complicated unless the same time slots are available end-to-end throughout the duration of the call.

While HSCSD is still circuit-switched technology, GPRS is complementary for communication with other packet-based networks such as the Internet.

2.4.2 GPRS—Tracing the Way to Mobile Internet

The fast growth of the Internet increased the user demands for wireless data services. The data rates in 2G were too slow to support Internet-like services, and also circuit-switched technology is too expensive to be used for bursty traffic (i.e., at the air interface, a complete traffic channel is allocated for a single user for the entire call duration). Hence, packet-switched services were needed to introduce statistical multiplexing (i.e., sharing of a single channels by multiple users). For that purpose GPRS is defined as an upgrade to the GSM system. Parallel with GPRS, *Cellular Digital Packet Data* (CDPD) is a similar upgrade for AMPS, IS-95, and IS-136 mobile systems. GPRS is the first step towards integration of the Internet and mobile cellular networks.

GPRS differs from HSCSD because it applies a packet radio principle to transfer user data packets. It is packet-based technology designed to work in parallel with 2G GSM, PDC, and TDMA systems. GPRS uses a multiple of one to eight time slots in a TDMA frame on 200-kHz carriers.

GPRS is created as a hardware and software upgrade to the existing GSM system. In order to integrate GPRS into existing GSM architecture, two new network nodes should be added: *serving GPRS support node* (SGSN) and *gateway GPRS support node* (GGSN), as shown in Figure 2.5. SGSN is responsible for the delivery of packets from/to mobile stations within its service area. Its main tasks are mobility management (including location management, attach/detach), packet routing, logical link management, authentication, and charging functions. GGSN acts as an interface between the GPRS packet network and external packet-based networks (i.e., Internet). It converts *protocol data packet* (PDP) addresses from the external packet-based networks to the GSM address of the specified user and vice versa. For each session in GPRS, so-called PDP context is created, which describes the session. It contains the PDP type (e.g., IPv4), the PDP address assigned to the mobile station for that session only, the requested QoS profile, and the address of the GGSN that is the access node to that packet network.

There may exist several SGSNs or GGSNs. All GPRS support nodes are connected via an IP-based GPRS backbone network. In the case of GPRS, HLR stores the user profile, the current SGSN address, and the PDP address(es) (e.g., IP address for communication with Internet) for each user. MSC/VLR is extended with additional functions that allow coordination between GSM circuit-switched services (e.g., telephony) and GPRS packet-switched services.

Due to the variety of packet-switched services, such as real-time multimedia, WWW, file download, and e-mail, each with different QoS requirements,

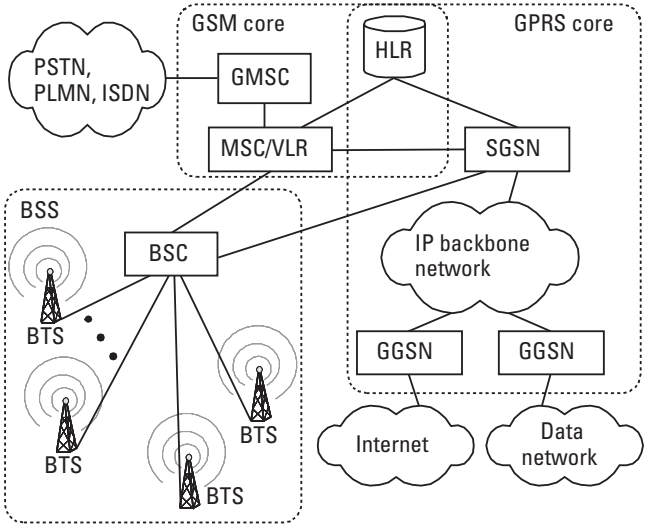


Figure 2.5 GPRS network architecture.

GPRS allows defining QoS profiles using the parameters service precedence, reliability, delay, and throughput [14]. The first parameter is priority of the service. There exist three types of priority: high, normal, and low. Reliability describes transmission characteristics of the GPRS network, such as loss probability, duplication, misinsertion, and corruption of packets. The delay defines average delay and maximum delay in 95% of all transfers. The throughput refers to maximum bit rate and mean bit rate.

For location management GPRS has three possible states: idle, ready, and standby. In idle state, the network does not know the location of the mobile station and no PDP context is associated with the station. When the mobile station sends or receives packets, it is in ready state. In this state the network knows which cell the user is in. After being silent for a period of time, MS reaches standby state. For location management in standby state, a GSM location area is divided in several so-called *routing areas* (RAs). To locate the mobile station in standby state, the network performs paging in the current routing area. In ready state there is no need for paging, while in idle state the network is paging all BTSs in the current location area of the mobile station.

While GPRS utilizes the same radio access network as GSM does, the third-generation mobile networks have defined different radio interfaces to provide higher bit rate services to users.

2.4.3 EDGE

EDGE was created to provide higher data rates for packet-based services with higher bandwidth demands using the existing 2G mobile networks. It is supposed to provide an update to GSM systems as well as to the ANSI-136 TDMA system.

EDGE technology was created to enhance throughput per time slot for both HSCSD and GPRS. It uses a new modulation scheme 8-PSK (phase shift keying) in addition to the *Gaussian minimum shift keying* (GMSK) modulation scheme in GSM/GPRS networks, and it enables data rates up to 384 Kbps. Hence, the EDGE upgrade to a GPRS network is also known as *Enhanced GPRS* (EGPRS), while enhancement of HSCSD is called ECSD. In ECSD, the data rate per time slot will not increase from 64 Kbps due to air interface limitations, but the data rate per time slot will triple when using all time slots for single connection in EGPRS, and the peak throughput will exceed 384 Kbps.

EDGE technology is also used over the D-AMPS systems (i.e., ANSI-136 TDMA-based networks), where it provides data rates over 473 Kbps per 30-kHz carriers. This is referred to as EGPRS-136HS. In this way EDGE offers the possibility of convergence of GSM and ANSI-136 systems.

EDGE technology is an option for 3G services. Additionally, EDGE can coexist with UMTS to provide high-speed services for wide area coverage, while UMTS in such scenarios may be used for the urban hot spots.

2.5 Third Generation Mobile Networks

The 3G systems should provide convergence of the existing standards in 2G, such as CDMA, GSM, and TDMA. The main reasons for the standardization of 3G are higher data rates in the air interface via implementation of a wideband technology, and introduction of new packet-based services to the end-users (i.e., Internet connectivity). Because GPRS (or CDPD) and EDGE already introduced packet-switched services, 3G is created to provide higher data rates and the possibility for creation of various services over the same network architecture (i.e., separating the service creation from the network operation). The network should be transparent and open to new services created by the service and content providers.

2.5.1 Standardization

The process of standardization of 3G mobile networks has several forms and bodies included with it. First, there are regional standardization bodies, such as ETSI in Europe and ANSI in North America. Furthermore, there are global standardization efforts, such as ITU standards for 3G called International Mobile Telephony 2000 (IMT-2000) as well as the 3G Partnership Project (3GPP) and 3GPP2, which include standardization bodies, industry, and academia members.

2.5.1.1 ITU's International Mobile Telephony—IMT-2000

ITU made efforts for harmonization and convergence in 3G mobile networks through the envelope of 3G mobile systems. Through a consensus ITU decided how much convergence was needed in 3G. In the mid-1990s, ITU created a framework for 3G mobile systems called IMT-2000. The concept of IMT-2000 includes the following aspects:

- Global, seamless access to mobile systems;
- Compatibility with major 2G systems;
- Convergence between the mobile and fixed network;
- High data rates for wireless communication;
- Circuit-switched and packet-switched data transfers;
- Introduction of multimedia applications.

By itself, IMT-2000 covers both third generation mobile terrestrial and mobile satellite systems.

The radio interface was the most interesting element for global standardization, because that is needed to provide universal access of mobile terminals to

different 3G networks. For the terrestrial radio access network, the choice was made on *wideband CDMA* (WCDMA) [6]. Within the framework of IMT-2000, ITU defines five different terrestrial radio interfaces. They are listed in Table 2.1 together with associated standards. ETSI's WCDMA and *time division CDMA* (TD-CDMA) are foreseen as the main users of the first two modes, respectively, while cdma2000 is foreseen as main user of the third one. The last two are (1) Universal Wireless Communications-136 (UWC-136) developed by the Telecommunication Industry Association (TIA) TR 45.3 subcommittee, which is based on TDMA single-carrier, and (2) Digital Enhanced Cordless Communication (DECT) developed by ETSI, which is based on FDMA/TDMA technology.

Thus, the main idea behind IMT-2000 is global roaming. Although only 3% of the calls involve intercountry roaming [6], the percentage of revenue is higher as these are expensive calls. In addition, global roaming sells mobile terminals around the globe. However, there are different interests for both industry and operators, so it is hard to expect that all 3G cellular systems will be compatible. But the number of subscribers within the global roaming cloud is expected to increase (within 2G systems, GSM is the world leader, considering the number of subscribers and network operators).

2.5.1.2 3G Partnership Project for UMTS

The ETSI began development of 3G mobile systems in the mid-1990s. The standard was named the *Universal Mobile Telecommunication System* (UMTS), and it is standardized as European terrestrial 3G system.

ETSI completed different studies on the choice of UMTS radio interface in 1996 and 1997. In June 1998 ETSI decided to select wideband CDMA (WCDMA) as the standard for the *UMTS Terrestrial Radio Access* (UTRA) air interface for *frequency division duplex* (FDD) operation, and TD-CDMA for *time division duplex* (TDD) operation. So, UTRA-FDD and UTRA-TDD were

Table 2.1
Standards for IMT-2000 Interface Operation Adopted by ITU

Mode	Standard
Direct spread (DS) CDMA	UTRA-FDD
Time division duplex (TDD)	UTRA-TDD
Multicarrier (MC) CDMA	cdma2000
Single carrier (SC)	UWC-136
Frequency time (FT)	DECT

created, and at the same time UTRA was submitted to the ITU as the ETSI proposal for IMT-2000.

In parallel, similar activities started in different regions of the world for standardization of technology like WCDMA. To ensure compatibility of the equipment as well as global standardization for 3G, the standardization organizations involved in the creation of the *3G Partnership Project* (3GPP). The partners in 3GPP are ETSI (Europe), ARIB/TTC (Japan), CWTS (China), T1 (United States), and TTA (South Korea) [15]. The original scope of 3GPP was to introduce technical specifications for 3G mobile networks based on the evolved GSM core networks and radio access technologies for both FDD and TDD modes. Additionally, 3GPP was amended to include GSM technical specifications as well as GPRS and EDGE, which evolve from GSM as transition to 3G. For more details on 3GPP, the reader may consult [15].

2.5.1.3 3GPP2 for cdma2000

For comprising American and Asian interests on 3G systems, their standardization bodies ANSI/TIA/EIA-41 started an initiative for the creation of 3GPP2, running parallel with 3GPP. It was born from ITU's initiative for IMT-2000. Although 3GPP started by an ETSI initiative in Europe, there was effort to consolidate collaboration efforts of all ITU members. In the end, 3GPP2 was created as a solution for American interests and that of some Asian countries. It includes as partners ARIB/TTC from Japan, TIA from North America, TTA from South Korea, and CWTS from China. The 3GPP2 efforts are based on standardization of cdma2000 for the air interface and an IP-based core network with Internet connectivity.

2.5.2 UMTS

The ETSI candidate for 3G is UMTS. This standardization body has defined the strategy for the third generation mobile systems [5] as follows:

- Core network of UMTS should be compatible with IP;
- Should be compliant with IPv4 as well as IPv6;
- Data rates up to 2 Mbps;
- Global roaming—between UMTS and GSM, and between UMTS and other systems from the IMT-2000 family;
- Support for mobility of users, terminals, and services.

Thus, the main ideas in UMTS are new services (e.g., multimedia services), content provision, and global roaming.

2.5.2.1 QoS Concept in UMTS

UMTS is planned to include variety of services, each with different QoS characteristics. Hence, four QoS classes are defined for UMTS [16] as follows:

- Conversational class;
- Streaming class;
- Interactive class;
- Background class.

When defining UMTS QoS classes, which are referred to as traffic classes, one should take into account the characteristics of the air interface (i.e., bandwidth limitations and error characteristics).

The main distinguishing factor between the QoS classes is the requirement for real-time service. In that sense, the parameter that defines real-time traffic is delay. Conversational class is defined for very delay-sensitive traffic, while the most delay-insensitive traffic is background traffic class. The first two classes, conversational and streaming, are specified to carry real-time traffic. The others, interactive and background classes, are mainly defined for nonreal-time applications.

A typical example of services in conversational class are circuit-switched telephony (e.g., GSM-like), but IP telephony and videoconferencing belong to this traffic class as well. Also, some other real-time communication that includes live end users may be added to the conversational class. Streaming class is created for one-way real-time transport, when a user is looking at (or hearing) a real-time video (or audio) stream. By the term "stream" we denote one-way communication flow to a live human destination. This class is also delay sensitive, but without strict delay requirements. Low delay variations may be neutralized by the receiving end. For real-time services, retransmission of lost or corrupted traffic packets is not desirable due to delay sensitivity. This is not the case with control packets for this type of application, which usually use some transport control mechanism (e.g., TCP). Interactive class is defined for applications where the end user (either a machine or a human) is requesting data from a remote end (e.g., a server). Examples of such services are Web browsing (WWW), database retrieval, and server access. Round-trip delay is one of the key attributes for the interactive class. Interactive applications require low delay, but are less sensitive to delay than conversational class. On the other hand, they have requirements for low bit error rate, and hence some transport control mechanism should be applied (e.g., for retransmissions of the lost packets). Finally, background class is created for sending and receiving data by a computer (no direct human interaction or presence is needed on either end of the

communication). Examples of background applications are e-mail, SMS, download of databases, and reception of measurement records. Table 2.2 shows the QoS attributes defined for each traffic class.

To describe the QoS level for a given service, one needs definitions of QoS parameters (or attributes, as noted in [16]). Attributes defined for UMTS are as follows: traffic class (conversational, streaming, interactive, or background), maximum bit rate (Kbps), guaranteed average bit rate (Kbps), delivery order (yes, or no), maximum *service data unit* (SDU) size, residual bit error ratio, *SDU error ratio* (SER), transfer delay (ms), traffic handling priority, and some other less important attributes (the reader may go to the ETSI Web site http://www.etsi.org for more details on its recommendations).

Let us briefly go through QoS attributes in UMTS. Maximum bit rate is the maximum number of bits transmitted over a time interval. The traffic is conformant with this parameter as long as it follows a token bucket algorithm, where token rate is equal to maximum bit rate and bucket size is equal to maximum SDU size parameter. The traffic is conformant with the guaranteed bit rate as long as it follows a token bucket algorithm where token rate equals guaranteed bit rate and bucket size equals maximum SDU size. The general token bucket algorithm is shown in Figure 2.6. Tokens represent the allowed data volume (e.g., in bytes for IP, or in packets for ATM). They are generated periodically according to the traffic contract and are stored in a token bucket (ATM terminology), or we may say a *token bucket counter* (TBC) is increased by a fixed value in each small time unit (IETF terminology). If the token bucket is full, arriving tokens are discarded (TBC is equal to the bucket size). If TBC is bigger than the incoming packet length, then the packet arrival is judged complaint

Table 2.2
UMTS QoS Attributes Defined for Each Traffic Class

Traffic Class	Conversational	Streaming	Interactive	Background
Maximum bit rate	X	X	X	X
Guaranteed bit rate	X	X		
Delivery order	X	X	X	X
Maximum SDU size	X	X	X	X
Residual bit error ratio	X	X	X	X
SDU error ratio	X	X	X	X
Transfer delay	X	X		
Traffic handling priority			X	

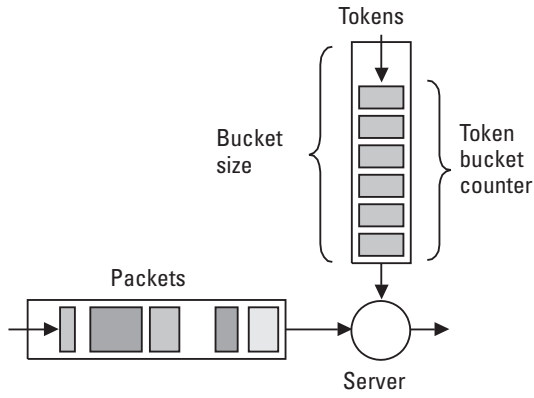


Figure 2.6 Token bucket traffic shaper.

(i.e., the traffic is conformant). Otherwise, the packet is marked as noncompliant (i.e., the traffic is not conformant).

The delivery order specifies whether out-of-sequence packets are acceptable or not to the destination. Maximum SDU size is defined for admission control and policing mechanisms (e.g., for policing the admitted bit rate). The residual bit error ratio indicates the undetected bit error ratio, or if no detection of errors is requested, it indicates the bit error ratio for the delivered SDUs. The SER indicates the fraction of SDUs lost or detected as erroneous. It is used in error detection schemes. Transfer delay indicates a maximum delay for the ninety-fifth percentile of the distribution of delay for all delivered SDUs within UMTS network. Traffic handling priority is defined to provide the possibility for differentiation of the traffic within interactive traffic class (it is used for scheduling purposes in the UMTS network nodes).

2.5.2.2 UMTS Architecture

UMTS architecture is described in [17, 18]. According to [17], UMTS's basic architectural split is between the user equipment (mobile terminals) and the infrastructure. There are two trivial domains: the *user equipment* (UE) domain and the infrastructure domain. UE is used users to access UMTS services. It includes the identity module and mobile equipment, which may include several functional software groups and hardware devices. The mobile equipment performs radio communication with the network and contains applications for the services.

The infrastructure domain is further split in two domains: the *network access* (NA) domain and the *core network* (CN) domain. The CN domain should have capability to use any NA technique (at least, all global access techniques). The NA domain consists of physical entities (nodes), which manage the radio

resources. The CN domain consists of physical entities, which provide support for the features and telecommunication services (e.g., mobility management, call management, and so forth).

The core network consists of the *circuit-switched* (CS) domain and *packet-switched* (PS) domain, as defined by [17]. These two domains in CN are overlapping in some common elements. CS mode is the GSM mode of operation, while PS is the mode supported by the GPRS. The entities specific to CS domain are MSC and GMSC. Of course, there are other entities used by the CS domain, but they are shared with the PS domain. Specific entities for the PS domain only are the GGSN and SGSN, which are introduced for the first time in GPRS (i.e., 2G+). To distinguish between 2G and 3G entities we usually write 3G-SGSN for SGSN in UMTS, while 2G-SGSN or just SGSN for GPRS, and so on for other domain-specific entities, either for CS or PS domain. The entities common to both domains in the core network, CS and PS, are *home subscriber server* (HSS), AuC, VLR, EIR, and SMS-support nodes [17]. HSS is master database for a given user, which contains user identification (numbering, addressing information), user security information (authentication, authorization), user location information, and user profile information (to which services the user has access). In the previous releases of UMTS, instead of HSS, the HLR was used. From now on, HLR for the CS and HLR for PS domain are considered as subsets of HSS, where HSS additionally provides IP multimedia functionality in the core network. Other common entities have similar functions as previously described in the GSM and GPRS sections in this chapter. The UMTS Network architecture is shown in Figure 2.7.

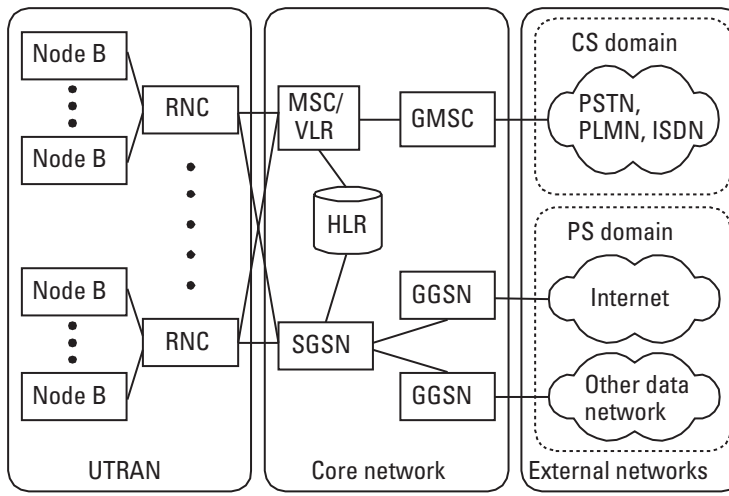


Figure 2.7 UMTS network architecture.

Considering the access network, two different types are specified for UMTS: the BSS and the *radio network system* (RNS). The BSS is the GSM radio access network solution (also used for GPRS and EDGE). BSS consists of the BSC and BTSs, where each BTS serves one cell. Usually several BTSs are grouped in a base station and placed on a single site. For UTRAN we need network elements responsible for radio resource management, handover management, and power control. This network system, which corresponds to the GSM BSS, is the RNS, but it significantly differs from the GSM access operation. RNS consists of the *radio network controller* (RNC), which controls the radio access nodes, called Node B. A Node B is a network component that serves one cell. We have different types of Node B, such as macro, micro, and picocells, where we face different requirements in traffic, coverage, and services. There are two types of Node B for UMTS: Node B FDD and Node B TDD. The latter is targeted to hot spots in coverage, while FDD is planned for wider coverage area (micro, macro).

For lowering the costs of 3G system implementation, it is planned to colocate BTS/Node B, and BSC/RNC sites. UMTS/GSM colocation ensures greater efficiency by sharing space and infrastructure. An overall comparison of 2G, 2G+, and 3G mobile networks architectures, services, and terminal's capabilities is given in Table 2.3.

Table 2.3
Comparison of 2G and 3G Mobile Networks

Network	Second Generation (2G)	Second Generation + (2G+)	Third Generation (3G)
Core network	MSC/VLR, GMSC, HLR, AuC, EIR	MSC/VLR, GMSC, SGSN, GGSN, HLR, AuC, EIR	3G-MSC/VLR, 3G-GMSC, 3G-SGSN, 3G-GGSN, HLR, AuC, EIR
Radio access network	BSC, BTS, MS	BSC, BTS, MS	RNC, access node, mobile station
Services	Voice, SMS, ISDN supplementary services	Voice, SMS, e-mail, WAP services	Voice, Internet, multimedia services, videotelephony
Data rates	Up to 9,600 bps (or up to 14,400 bps)	Up to 57.6 Kbps for HSCSD; Up to 115 Kbps for GPRS; Up to 384 Kbps for EDGE	Up to 2 Mbps
Mobile terminals	Voice-only terminals	User-friendly terminals, enhanced service capabilities	Voice, data, and video terminals, multiple modes

2.5.2.3 UMTS Frequency Bands

In 1992, the World Administrative Radio Conference (WARC-92) identified the 1,800 to 2,200-MHz frequency band for IMT-2000 [19]. 3GPP has specified frequency bands for UMTS for both radio access modes, FDD and TDD.

In Europe 12 carrier pairs are available in FDD mode (5 MHz for uplink and 5 MHz for downlink). So, in FDD, duplex connection is realized by using different frequency carriers for uplink and downlink direction. In TDD mode, uplink and downlink are implemented in the same frequency band (same carrier). It is achieved by defining time frames and time slots. In TDD mode, the network allocates radio resources on a time-slot basis in both uplink and downlink where time slots are grouped into frames. A certain number of time slots within a time frame is allocated to uplink, and the remaining time slots to downlink. So, the transmission occurs quasi-simultaneously. Seven 5-MHz carriers are available in the TDD mode, as shown in Figure 2.8.

2.5.3 WCDMA

WCDMA is a UTRA-FDD mode of operation. It uses direct sequence CDMA. The term *wideband* is used to differentiate WCDMA from 2G CDMA based on technology pioneered by Qualcomm, called cdmaOne (or IS-95 CDMA). WCDMA uses approximately three times wider bandwidth than cdmaOne (i.e., it uses bandwidth of approximately 5 MHz per carrier). The same carriers may be reused in neighboring cells. Radio access network separates each user flow (voice, data, and so forth) by multiplying the user information with pseudo-random bits called chips. The chip rate specified for WCDMA is 3.84 Mcps (millions of chips per second).

2.5.3.1 CDMA Operation

In CDMA operation the narrowband signal of the user is spread across the whole bandwidth of the carrier, which is much wider. For this reason, CDMA technology is sometimes referred to as *spread spectrum*. An example of the spreading of the user signal is shown in Figure 2.9.

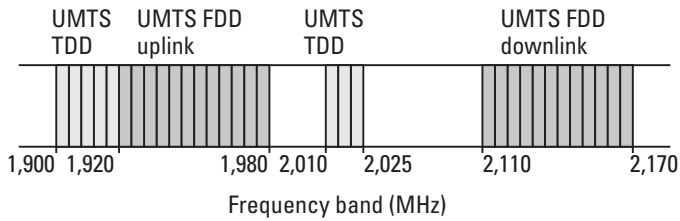


Figure 2.8 Frequency bands for UMTS.

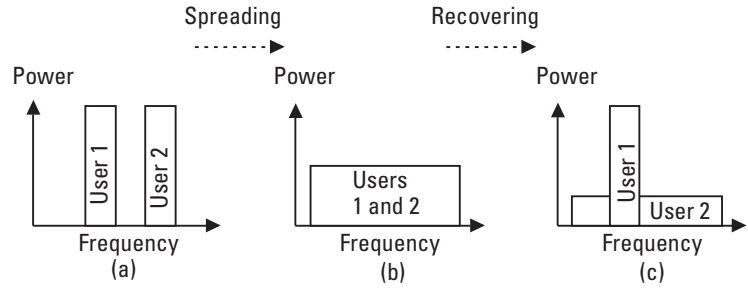


Figure 2.9 Spreading the signal by CDMA: (a) unspread signal; (b) spread signal; and (c) recovered signal.

At the receiver end despreading of the wideband signal is needed. The despreading process converts the wideband-spread signal back to the original narrowband signal by multiplying the spread signal with the same pseudo-code. This way the original narrowband signal is reconstructed, while other spread signals (from other users) are considered as noise, called interference. More user traffic means more interference, which results in lower quality.

2.5.3.2 Why CDMA?

What is the advantage of CDMA over FDMA/TDMA techniques? In FDMA and TDMA the common channel space is partitioned in orthogonal single-user subchannels, which are not overlapping. In FDMA each user uses a certain frequency band per call, which is not shared with other users during the call. In TDMA the time is divided into time slots, and each user is given a time slot. The problem arises when we have bursty traffic accessing the network (e.g., Web traffic). In such cases we may need to transmit a larger volume of information data in shorter time periods, and then silent period to follow, and so on. For example, voice contains talk-spurts and silent periods. Also, a Web connection contains active periods of browsing, and silent period for looking at (or hearing) the information content. TDMA allows flexible rates in multiples of basic single channels and subchannels (submultiples) for low bit rate transmitting. However, it requires additional signal processing to cope with synchronization. WCDMA supports rates up to 2 Mbps, utilizing variable spreading factor and multicode links. User data is transmitted using 10-ms frames during which the user data remains constant. By variable spreading factor, we address actual carrier bandwidth, which may be between 4.4 and 5 MHz, by using grid of 200 kHz (e.g., it may be 4.4 MHz or 4.6 MHz). With the use of multicode links, we address the assignment of additional codes when users demand more bandwidth on the link (more codes gives more bandwidth). Spreading codes are designed to allow the symbols from multiple users to occupy the same spectrum at the same

time. WCDMA uses asynchronous transmission at the base station. For comparison, IS-95 CDMA uses synchronous transmission where synchronization is made possible by using *Global Positioning System* (GPS). Due to FDD mode, WCDMA uses separate frequency bands to provide duplex connections.

2.5.3.3 Characteristics of WCDMA

Here we address some features that are specific to WCDMA, or to CDMA in global terms. In CDMA, the individual connections between mobiles and base stations are separated by the codes, while transmission takes place simultaneously on the same frequency band. It is always possible to establish an additional connection with the use of a new code. Hence, CDMA has soft capacity. However, the more data being transmitted by the radio interface (in the cell or in adjacent cells), the more noise disturbs the connection, thus reducing the quality of the call. Because the available bandwidth per carrier is up to 5 MHz, data transmission rates from 8 Kbps to 2 Mbps can be realized. Also, due to application of codes in CDMA, the same frequency bands can be used in the neighboring cells, resulting in frequency reuse factor of 1. This makes frequency planning easier than in GSM. Furthermore, multipath propagation in WCDMA is considered as an advantage (it was the opposite case in GSM).

Soft Handover

User equipment and base stations use special RAKE receivers that allow each UE to simultaneously communicate with multiple base stations. In WCDMA we define two types of "soft" handovers: soft and softer handover. The former refers to handover between the same carriers in cells belonging to neighboring base stations, while the latter refers to soft handover between cells belonging to the same base station. However, some hard handovers are still required in CDMA networks. For example, for handover between FDD and TDD modes in UMTS, only hard handover is possible.

Multipath Reception

The RAKE receivers also allow the UE to decode multiple signals that have traveled over different physical paths from the base station. For example, one signal may travel directly from the base station to the UE, and another may reflect off a large building or woods and then travel to the UE. This phenomenon, called *multipath propagation*, also provides a diversity gain. The same effect occurs on the uplink from the UE to the base station. WCDMA and cdma2000 have three times bigger bandwidth than 2G CDMA (IS-95 standard), and hence have a higher diversity gain.

Power Control

Transmissions by the UE must be carefully controlled so that all transmissions are received with roughly the same power at the base station. If power control is

not used, a "near-far" problem occurs. In this case mobiles close to the base station overpower signals from mobiles farther away. The base station uses a fast power control system to direct the mobile to power up or power down as its received signal level varies due to changes in the propagation environment. Similar, on the downlink, transmissions from the base stations are power-controlled to minimize the overall interference throughout the system and to ensure a good received signal by the UE. For example, in WCDMA fast power control is applied with 1,500 Hz (for comparison, in GSM, power control has an update frequency of only 2 Hz—that is, transmitting power level is changed two times during one second).

Frequency Reuse of 1

Due to the application of codes in CDMA, the same frequency bands (carriers) can be used in neighboring cells, so no frequency planning is required. But, since every site causes interference to every other site, careful attention must be paid to radio propagation for each site.

Soft Capacity

Capacity and coverage are intertwined in CDMA, depending on the number of users in the system and the amount of interference allowed before access is blocked for new users. By setting the allowed interference threshold lower, coverage will improve at the expense of capacity. By setting the threshold higher, capacity will increase at the expense of coverage. Because of the fundamental link between coverage and capacity, cells with light traffic loads inherently share some of their latent capacity with more highly loaded surrounding cells.

2.5.4 TD-CDMA

TD-CDMA is a solution for UTRA-TDD mode. It operates in time division duplexing using the same frequency carrier for uplink and downlink (see Figure 2.10). Uplink and downlink time slots are grouped into sequences. So, the communication is quasi-duplex because at a given time slot, the mobile terminal only transmits data in the uplink, or receives data in the downlink. Here, spreading codes separate user signals within one or more time slots. In TD-CDMA we define a physical channel by a frequency carrier, a time slot, and a code. For comparison, in FDD we use a carrier and a code to define a physical channel. Each time slot can be assigned either to uplink or downlink, depending on the demand. Users may occupy several time slots in a frame to obtain variable transmission rates. Furthermore, we may achieve variable rates by varying the spreading of a single code allocated to the given connection, or by adding more codes (multicode) to the connection with fixed spreading.

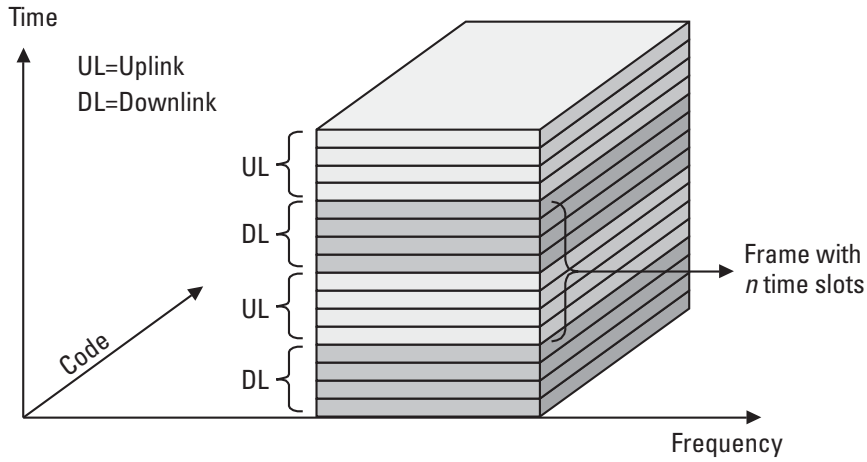


Figure 2.10 UTRA-TDD mode (TD-CDMA).

2.5.5 cdma2000

The other of the ITU's main candidates for 3G, besides UMTS, is cdma2000. Considering the market, cdma2000 is oriented to the Americas and in part to Asia. This standard is compatible with 2G CDMA IS-95 mobile systems, but it is not compatible with 2G GSM. While WCDMA is asynchronous, cdma2000 is based on a synchronous architecture similar to IS-95 systems. The cdma2000 can be deployed in several phases [20]. The first phase, cdma2000 1x, supports up to 384 Kbps packet data (theoretically) and doubles voice capacity of IS-95. It operates in the 1.25-MHz channel. In the second release of 1x, two alternatives are currently proposed: 1xEV-DO (1x Evolution–Data Only) and 1xEV-DV (1x Evolution–Data and Voice). The 1xEV-DO provides pure data over the network (i.e., a carrier will be reserved for data only). This way it is possible to achieve data rates above 2 Mbps. A disadvantage of such approach is inefficient frequency space utilization. For example, if the network is loaded with "heavy" voice traffic and low data traffic, the free resources from the data-only carrier cannot be allocated to voice traffic. In 1xEV-DV, cdma2000 provides more flexibility by mixing data and voice traffic on the same carrier. The second phase of cdma2000 is called 3x, which introduces higher data rates at the expense of cell coverage. The chip rate of cdma2000 3x is 3.6864 Mcps, which is slightly lower than in WCDMA. But the chip rate is three times the chip rate in an IS-95 system that provides compatibility between the systems and easy migration from IS-95 to cdma2000. Also, there is a potential CDMA Nx ($N > 3$) ought to support much higher data rates than 2 Mbps.

In the next section we refer to the Wireless IP standard defined for cdma2000, as well as its QoS concept.

2.5.5.1 Wireless IP Standard for cdma2000

Within cdma2000 is defined a standard called "Wireless IP standard" [21], which defines requirements for the support of wireless packet data networks in cdma2000. Its reference model is given in Figure 2.11. This standard defines two methods for accessing packet data networks:

- Simple IP;
- Mobile IP.

The main tendency in Wireless IP standard is usage of already existing IETF protocols for mobility support. In Simple IP the user is assigned a dynamic IP address from the *Public Data Service Network* (PDSN). The user retains its IP address as long as it is in the service area (radio network) of the assigning PDSN. Simple IP does not provide IP mobility beyond the PDSN. Mobile IP is defined in RFC 2002 [9]. We refer to it in more detail in the next chapter. The mobile station is assigned either a nonzero static IP address or a dynamically assigned IP address belonging to its home IP network, called the

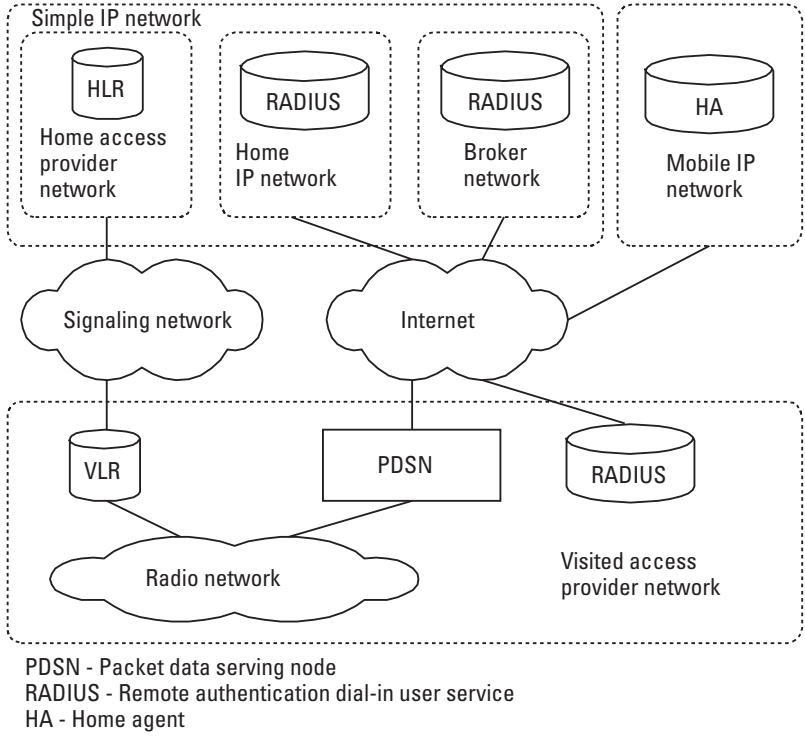


Figure 2.11 Reference model for access in the Wireless IP network.

home address (HA). In this case the user is able to have a persistent IP address, even at handover between separate PDSNs.

Both Simple IP and Mobile IP utilize *Remote Authentication Dial-In User Service* (RADIUS) servers, which use the RADIUS Protocol for carrying authentication, authorization, and configuration information between a network access server that desires to authenticate its links and a shared authentication server [22]. RADIUS is built over the UDP/IP protocol stack. There are three types of RADIUS servers: home, visited, and broker RADIUS. Home RADIUS resides in a Home IP network, and visited RADIUS resides in Visited IP network. Broker RADIUS is an intermediate optional server (or servers) that has security relations with the visited RADIUS and the home RADIUS, and is used to transfer messages between the Visited IP network and Home IP network.

Considering the mobility management (i.e., support for continuality of the connection during the radio interface change), and by using Figure 2.12, the standard defines two types of handovers (or handoffs):

1. *Packet call function* (PCF)-to-PCF handover (refer to Figure 2.12), where a *Point-to-Point Protocol* (PPP) session continues if the handover occurs between access points in the same PDSN; if the handover is to different PDSNs, a new PPP session needs to be established.
2. PDSN-to-PDSN handover, which requires Mobile IP. This type of handover needs detection of the FA and establishment of new PPP session after successful authentication in the target PDSN.

In 3GPP2 proposals for Wireless IP, QoS is limited to support for differentiated services (we refer to them in next chapter). Mobile stations are allowed

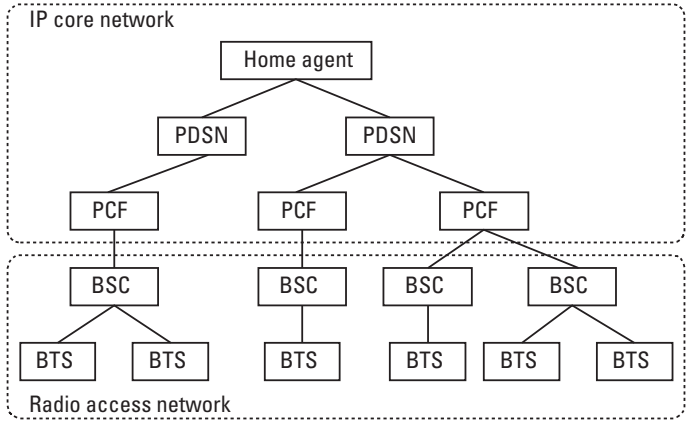


Figure 2.12 Packet data mobility concept in cdma2000.

to mark IP packets with a *differentiated services* (DS) class, which they transmit to network. However, PDSN may accept this mark or it may remark the packet based on user profile's DS class options at the home server.

2.5.5.2 QoS in cdma2000 Systems

QoS is defined from the perspective of the users and/or the system operator [23]. From the user perspective, cdma2000 defines different QoS service levels, where each level has an associated profile for the QoS requirements of that service. All applications are classified into QoS groups. Here, QoS refers to the ability of a network operator to support the end user requirement with regards to four QoS parameters:

1. Bandwidth;
2. Latency (delay);
3. Jitter (delay variation);
4. Traffic loss.

Network QoS consists of two parts:

1. The first part consists of a mechanism for negotiation of the QoS service level between applications on the user's end and application servers, for some or for all data traffic carried between them through the network. The aim of such a mechanism is to ensure constraint on the user QoS parameters (i.e., bandwidth, latency, jitter, and loss), with a specified probability, which is not negotiated.
2. The second part consists of set of protocols between network elements, which are used to negotiate QoS characteristics on each link (e.g., a hop) or path through a single network carrier.

Similar to the UMTS QoS concept, cdma2000 supports four traffic classes: conversational class, streaming class, interactive class, and background class. Attributes of traffic for defined classes are identical to those of UMTS.

We may conclude that UMTS and cdma2000 are converging towards a unified QoS concept (i.e., both standards have the same traffic classification considering the QoS).

2.6 Third Generation Mobile Applications and Services

So far, we conclude that there are two main concepts for 3G mobile systems: UMTS standardized by 3GPP, and cdma2000 standardized by 3GPP2. In both concepts the mainstream is towards:

1. Packet-based radio-access (with IP technology) for the purpose of statistical multiplexing and integration of heterogeneous services over the wireless access network and core network of the mobile operators;
2. Introduction of new services and content, which increases the revenue from the users.

According to the 3GPP proposal [24], a network offers three types of services:

1. Bearer services;
2. Teleservices;
3. Supplementary services.

Bearer services involve only low layer functions, in reference to the OSI model. In 3G they should support both domains, CS and PS. Bearer services for UMTS may be connection-oriented or connectionless services. They are required to provide guaranteed (constant) bit rate and real-time dynamically varying bit rate. Real-time and nonreal-time services should be supported. This implies that bearer services should have the ability to provide guaranteed real-time service with guarantees on bit rate, delay, and delay variations. Because nonreal-time services will be on the same link as real-time services (at least in UMTS), bearer services should provide the ability for QoS differentiation between different users. Also, multimedia applications should be supported. By definition [24], multimedia refers to several user flows to/from the user having different traffic types, such as real time and nonreal time.

Applications that are on higher layers according to the OSI model should specify their traffic requirements to the network by requesting appropriate bearer service. Considering the entities included in communication, bearer services may be:

- *Point-to-point*, which can be unidirectional or bidirectional. In the latter case we may further divide the services into symmetric or asymmetric;
- *Point-to-multipoint*, which may be multicast (in this case receiving ends are specified) or broadcast (receiving end points are not explicitly specified by the sender, but any receiver should have ability to admit or reject such service).

We previously introduced four QoS classes, which refer to bearer services. According to requested bearer service, each application may be classified into one of the QoS (traffic) classes. In UMTS, an application requests bearer service

by specifying any of the following: traffic type, traffic characteristics (e.g., point-to-point), maximum transfer delay, delay variation, bit error ratio, and data rates. Each mobile terminal may have several bearer services simultaneously (each of which may be connection-oriented or connectionless). In such a case, we may refer to the cumulative bit rate of the mobile terminal. In [24], maximum cumulative bit rates for UMTS are specified, which are:

- Up to 144 Kbps for rural environment or for satellite radio access;
- Up to 384 Kbps for urban/suburban outdoor environments;
- At least 2 Mbps for indoor radio environment.

However, radio and network interfaces have granularity limitation considering the supported bit rates. To allow for the support of flexible bandwidth on-demand services, bearer services should be provided with the finest possible granularity that can be efficiently supported. The aim is to increase bandwidth usage efficiency through the means of statistical multiplexing.

By definition [24], teleservices provide full capability of communication by using terminal equipment, network equipment, and some dedicated centers to that service. The teleservice and its meaning is also defined by ITU-T Recommendation F.700 [25]. Teleservices may be grouped into single media services (e.g., speech) and multimedia services. There are six specified categories of multimedia services:

1. Multimedia conference services;
2. Multimedia conversational services;
3. Multimedia distribution services;
4. Multimedia retrieval services;
5. Multimedia messaging services;
6. Multimedia collection services.

Teleservices require association of terminal and network capabilities. Here, upper layer capabilities (in reference to the OSI model) are necessary. However, lower layer capabilities are always needed (i.e., there is mapping between each teleservice and some of the bearer services in the system). Teleservices supported in 2G mobile systems are:

- Speech (telephony);
- Emergency call;
- SMS.

Speech is a service that should be supported by any mobile system. Emergency call has a speech component, but it has significantly reduced authentication requirements compared to telephony (usually, it is always allowed for a user to make an emergency call). SMS may be point-to-point (SMS-PP) or cell broadcast (SMS-CB). Additionally, UTRAN shall provide interworking with external data networks, of which the Internet is seen as the most important one. Therefore, Internet access is considered by 3GPP as one of the main teleservices. The most important benefits from the definition of Internet access are: optimized transmission of IP traffic over the radio interface considering the scarce radio resources, and implementation of QoS mechanisms. QoS mechanisms should be compliant with those defined for the Internet. A typical QoS mechanism, foreseen as a main candidate for 3G and beyond, is differentiated services (which is described in Chapter 3).

The third generic type of services is supplementary services. A supplementary service supplements a basic telecommunication service (i.e., teleservice). Such services are offered in 2G mobile systems, such as GSM, and they continue to be offered in 2G+ and 3G systems. Examples of supplementary services include the following: calling/connected line identification presentation/restriction, call forwarding (on subscriber busy, no reply, not reachable, or unconditional), call waiting, call hold, multiparty service, closed user group, advice of charge, and barring outgoing/incoming/roaming calls.

2.6.1 New Killer Applications

The answer cannot be easily given to the question: What will be or what are the killer applications for 3G and beyond over the next several years? Additionally, one may ask whether such applications would be spread worldwide, or be locally based.

Because users choose applications that are of value to them considering some merit (e.g., a need, a personal image, popularity), it should be noted that applications are not always driven by the technical merit (e.g., bandwidth). So, it is not simply a question of mobilizing the content found on Internet, but it is a question of creating applications that capture the needs of the users in terms of mobility. Some analysis leads to the conclusion that, in the future, revenue from voice services will slowly decline due to regulation, saturation of the user market, new services, as well as competition among the operators over a same geographic area. On the other hand, one may expect broadband packet-based data services to be offered to the users.

Key services in future mobile systems are in the areas of multimedia messaging, m-commerce, and location-based services. Of course, voice will continue to be the dominant service in the next several years. Predictions show that by

2010, revenue from voice will decline below one-half of the total revenue from mobile networks.

The deployment of 3G mobile networks should provide a set of broadband services at different data rates using CDMA-based techniques. However, the radio interface's bandwidth can be further extended by wireless LAN technologies, such as IEEE 802.11, *High Performance Radio LAN* (HIPERLAN) 1 and 2, as well as Bluetooth and satellite networks. The bit rate provided to the user in next generation mobile networks depends on several factors. One of them is the velocity of the user, which is inversely proportional to the bit rate that can be allocated to the user by the network.

There are standardization efforts for the creation of protocols that will support bearer transparency at the application layer. Examples of such protocols are the *Wireless Application Protocol* (WAP) and *Freedom of Multimedia Access* (FOMA). The latter is created for support of services offered by NTT DoCoMo in Japan. The WAP standard [26] is introduced for 2G+ mobile networks; 3G networks should also support it. However, WAP needs the so-called WAP-gateway network node (for conversation between HTTP and WAP), and therefore it is an optional protocol for the mobile network. The trend is to integrate WAP with current Internet standards that are proposed by IETF. The target is to create optimized HTTP/TCP protocol stack that will allow presentation of Web content to any browser on any device, from PCs and laptops to mobile terminal and PDAs.

One of the starters in 3G is NTT DoCoMo from Japan [27]. They have created the very successful i-mode mobile Internet service, which provides easy e-mail and Internet access using the existing HTTP protocol (created for wired Internet). FOMA is an advanced version of i-mode. It provides multimedia messaging such as e-mail with attached sound file and still images, downloading JPEG pictures, downloading and playing network games (based on Java technology), and support for video images via i-mode. FOMA can be used as a packet-switched service (data rates up to 384 Kbps in downlink), but it can also be used in circuit-switched mode (64 Kbps in each direction, uplink and downlink). So, we may classify killer applications in 3G in several groups:

- *Personal services*, such as gaming (playing games by using a mobile terminal and a mobile network), multimedia messaging services (messages with attached audio, video files, and/or still images), location-based services (e.g., city guide, local information services, GPS mapping information), mobile instant messaging (e.g., "chat"), on-line shopping (viewing and purchasing merchandise via mobile), and many others.
- *Business services*, such as videoconferencing (video and audio conference over distance), mobile commerce (receiving/sending customer and sales data), and on-line stock trading via the mobile.

- *Home services*, such as mobile office (intranet access from mobile terminal, database access), home surveillance (i.e., distance control of the home security system), distance control of home appliances (e.g., one may choose to start the heating system while driving home on the highway).
- *Entertainment services*, which include information and audio or video clips from sport events, theaters, concerts, as well as previews of cinema/TV movies. Here, we must mention audio and video streaming as an option. However, streaming services also refer to viewing/listening to shorter clips rather than watching movies or TV via the mobile, which are not considered to be killer applications due to high bandwidth requirements (especially for video streaming) in time-continuous flows during longer time periods (such service will be too costly to be attractive to the users).

However, these new services require mobile terminals with certain capabilities. The mobile terminal has a significant effect on the usage of services (easy to use, quality of reproduction, support for multiple services and operation modes). Dual-mode and multimode terminals are needed for the transition between 2G and 3G, and such will be needed for the transition between 3G and next generation mobile networks (e.g., 4G). Terminals should support simultaneous voice and data capability (e.g., voice conversation and downloading a file at the same time).

An important issue is the charging (billing) of these killer applications. While in voice service, billing is based on the duration of the call, GPRS allows charging based on transmitted information volume (for data services). However, sometimes it is not convenient to charge per megabyte downloaded. The user will rarely download a song if it is cheaper to purchase a CD in a store. On the other hand, tracking real-time events while on move (e.g., downloading a video clip immediately after a scored goal in a football game) or previewing a clip from cinema movie is a different situation, where the user will pay for recent news, interesting information, or entertainment. However, many operators are moving from a traditional time-based billing approach to a model of billing that sets a fee for a certain service over a longer period (e.g., a week or a month), while giving users freedom to use that service as much as they want or within certain limits (in number of events, in time usage, or in data volume). Of course, different tariff models and different associated QoS classes should be used to optimize the utilization of the network resources.

Finally, another issue that is important is security. For example, mobile terminals with Web browsing capabilities may experience a new wave of viruses and hackers. There is no 100% protection, but certain security mechanisms

should be applied. Traditional approaches include authentication (using unique user number), authorization (checking for admitted services to the user), and encryption of transmitted data (protecting data on the link by applying some type of encryption of data bits). In contrast to 2G networks, in this case (with many new services), security and trust is not addressed to network operators/providers only, but also to service providers, which shall provide services by using the network resources. However, a network operator may also be a service provider.

2.6.2 Real-Time Services

First, let us define a real-time service. Real-time refers to services that use one-way or two-way transmission of information where:

- The time relation between information entities shall be preserved (i.e., different media streams should be synchronized);
- Conversational services have very stringent requirements on delay.

Both requirements should be satisfied for real-time conversational services, while only the first one is required for real-time streaming services.

2.6.2.1 Real-Time Conversational Services

The most well known and the oldest of real-time services is telephony speech. In GSM it is basic service (it is based on circuit-switching). In packet-switched networks, such as 3G mobile networks and beyond, new applications require this scheme, such as IP telephony, videophone, and videoconferencing tools.

IP telephony has similar QoS requirements as CS telephony (e.g., GSM). The human ear is highly intolerant to delay variation (it should be less than 1 ms). Also, it has stringent tolerance on delay (delay 30 ms: the user does not notice any delay; delay 100 ms: the user does not notice it when echo cancellers are applied in the network). ITU-T has specified in Recommendation G.114 that the preferred delay for one-way voice communication is in the range 0 to 150 ms, while up to 400 ms is also acceptable. Above 400 ms, it is not. On the other hand, experimental studies suggested that *frame error ratio* (FER) less than 3% is acceptable for telephony.

Videoconferencing services provide two-way transmission of real-time video, audio, and other media. Latency (delay) is the key parameter for these services. In the case of a multimedia conference, an important issue is synchronization between different media streams (e.g., video and audio). Also, it is not feasible to recover retransmission errors or losses by retransmission mechanism because of added delays. However, the use of error-recovery and error-

concealment techniques are necessary to ensure graceful performance degradation over the wireless link, which has time-variable error ratio. According to [28], videoconferencing services are symmetric and include encoding, transmission, interpreting, and decoding components. Such services are allowed in PS mode and CS mode. Videoconference service should coexist with other data service options on the same wireless link. However, QoS requirements may be applied only to the portion of the communication link within the wireless network. The videoconferencing service shall be compatible with ITU-T standards for video conferencing, such as H.324M/H.323 and its adaptation to circuit-switched multimedia telephony service [29]. Data rates for videoconference may range from 32 Kbps up to 384 Kbps in 3G standardization groups, both 3GPP (with UMTS) [24] and 3GPP2 (with cdma2000) [28]. Videophone shall apply the same delay requirements as for conversational voice. However, audio and video must be synchronized within certain limits to provide "lip-synch" (synchronization of speaker's lips with the speech at the receiving end). So, lip-synch 100 ms is specified for UMTS, while it shall be below 20 ms (inter media skew) for cdma2000.

Other real-time services are interactive games (i.e., on-line gaming), two-way telemetry, and telnet [24]. Game requirements is dependent on the type of game (e.g., a slow game like chess versus a fast game such as racing), but in all cases delay should be kept below 250 ms. Two-way telemetry and telnet have similar delay requirements as on-line gaming.

2.6.2.2 Real-Time Streaming Services

In this section we classify audio streaming, video streaming, and multimedia streaming services. These services shall be allowed for the first time in the wireless environment in 3G mobile networks. Streaming refers mainly to unidirectional stream with high continuous utilization of bandwidth.

Audio streaming is intended to provide better quality than conversational telephony service (it includes music, surround noise). Because there is no conversational element involved, audio streaming is allowed to have higher delay, but this is limited because of the limited buffer capacity at the receiving mobile terminal.

Video streaming (one-way video) has similar performance requirements as audio streaming. It includes audio and video streams. Video streaming enables users to view videos within certain playout delay, instead of downloading the whole video file before it can be viewed. Thus, this service allows the user to start viewing video soon after downloading begins. Considering the delay, the maximum playout delay recommended by [28] is 30 seconds, while the maximum delay for this service specified in [24] is 10 seconds. In both standards it is specified that FER $\leq$ 1%. For circuit-switched systems it corresponds to a BER $\leq$ 0.1%.

Multimedia streaming includes streaming of audio, video, and multimedia data. It is defined as separate service by 3GPP2 [30]. This service is generally used in multimedia and message retrieval, video-on-demand, pay-TV, interactive news retrieval, and other multimedia broadcasting. It requires similar performances as video streaming. On the other side, multimedia streaming includes content creation and transmission. The receiving side decodes the data by sending packets from each media stream to corresponding audio/video codec and by performing synchronization between streams and content reproduction.

Also, we may provide multimedia services by using the *push service*, which delivers information initiated from a network (which may be external to the mobile network) to the mobile terminal. This service will cause a PDP context (refer to Section 2.4.2) to be activated if needed. Consumers should be able to choose to accept/reject push services through pop-up menus that should make navigation of services more efficient. Push services should allow customers to continuously refine their customer profiles.

All streaming services usually include control protocols for setting up connection between parties and negotiating various options for the particular service.

2.6.3 Nonreal-Time Services

In nonreal-time services we classify all 3G interactive services and background services.

Typical interactive services are voice messaging, Web browsing, high-priority transactions (e-commerce), and e-mail retrieval (server access). For interactive services round-trip delay is one of the most important parameters because the receiving-end peer expects the requested information within a reasonable time period. Experimental studies have shown that for these services acceptable delays are in the range 2 to 4 seconds. However, lower delays are always desirable (e.g., delay of 0.5 second is desirable for Web browsing).

Another group of nonreal-time services are background services. Typical examples are delivery of e-mail (server to server), SMS, fax, and low-priority transactions. These services have no specified requirements on delay or delay variation. However, the information should be delivered to the user within a reasonable time, which ranges from tens of seconds to several hours. Reasonable time for SMS and fax is up to 30 seconds, while for e-mail is up to several hours (also, it is dependent on the end user). The only strict requirement is that information should be delivered to the user error-free. For that purpose, a mechanism for retransmissions of lost or corrupted packet is usually applied (e.g., TCP). Table 2.4 gives an overview of 3G services considering the QoS class and performance requirements.

Table 2.4
Performance Comparisons of Main 3G Services

Service	Traffic Class	Real Time	Data Rate (Kbps)	Loss (FER)	Delay	Jitter
Voice	Conversational	Yes	4–25	<3%	<150 ms	<1 ms
Videophone	Conversational	Yes	32–384	<1%	<150 ms	N/A
Gaming	Conversational	Yes	<32	0	<250 ms	N/A
Two-way telemetry	Conversational	Yes	<32	0	<250 ms	N/A
Voice messaging	Interactive	No	4–13	<3%	<2 seconds	<1 ms
WWW	Interactive	No		0	<4 seconds	N/A
E-commerce	Interactive	No		0	<4 seconds	N/A
E-mail access	Interactive	No		0	<4 seconds	N/A
Audio streaming	Streaming	Yes	32–128	<1%	<10 seconds	N/A
Video streaming	Streaming	Yes	32–384	<1%	<10 seconds	N/A
SMS	Background	No		0	<30 seconds	N/A
Fax	Background	No		0	<30 seconds	N/A
E-mail (between servers)	Background	No		0	Variable	N/A

2.7 Future Wireless Communication Networks Beyond 3G

There is continuous evolution of mobile networks. The 2G introduced digital wireless access and many supplementary services based on ISDN technology. The 2G+ networks, such as GPRS, used a combination of CS and PS modes in the core network. The 3G networks have IP-based *radio access networks* (RANs), and the so-called 3G-core network, which has CS and PS capabilities. The next generation mobile networks are going towards all-IP networks, including an all-IP core network and IP RAN. The first move towards all-IP networks is made in 3G by the definition of the *IP Multimedia* (IM) core network [31]. The evolution of mobile networks towards an all-IP network is shown in Figure 2.13.

The next generation wireless networks are supposed to continue the evolution of the mobile world beyond 3G. There is a lot of ongoing work considering the next generation (e.g., 4G). In this section we will refer to it as *FutureG* (future generation mobile networks). If we follow the evolution path of mobile networks until 3G, we may extract two main drivers:

- Higher data rates;
- New multimedia services.

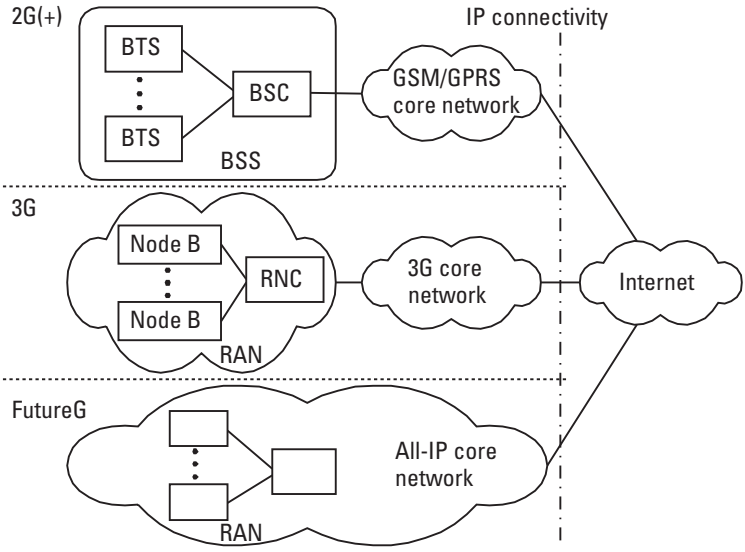


Figure 2.13 Evolution towards an all-IP mobile network.

Using this approach, FutureG shall provide more bandwidth (beyond 2 Mbps) and find new frequency bands for a new standard [32]. The path to FutureG is not so straightforward, however. One of the main concerns in FutureG is that it will be deployed in a situation when there will be variety of wireless and wired communication networks (e.g., 3G, GSM, wireless LAN, Bluetooth, or xDSL for wired access). According to the current expectations of the vendors and researchers in the wireless community, FutureG will be more focused on services and user needs by using wireless infrastructure in a more transparent way. Using the standards, a few large global companies, vendors, and organizations control the wireless market. It is expected that interface standards and frequency bands will become of secondary concern, although they are needed and should be specified.

If mobile networks have evolved from a 10-Kbps data rate in 2G (e.g., 9.6 Kbps in GSM) to 2 Mbps in 3G (which is almost a 200 times higher data rate than in GSM), then we may expect data rates in FutureG radio interface of 100 Mbps or more [32]. An example of an application that may need such a high data rate is telepresence, which is a foreseen application that will be used to create virtual meetings between individuals, and provide full simulation of all senses required to provide an illusion of actually being somewhere else. In that sense, mobile terminals will become extremely computational capable.

A mobile terminal in FutureG will not only present information to the user, but it should also be capable of gathering information about the user or the environment (e.g., by using a camera or some sensor placed on the front of the

mobile, generating content by using a specific application in the mobile terminal). So, the user would be a service provider in some cases. For example, a user may leave a message about a car crash on the highway with which it will inform other drivers to avoid that road. Furthermore, in FutureG the user should not rely on a single type of access network, or on a single operator. Even if there is no presence of an operator's network, the user should be able to use ad hoc applications to communicate with other users in the area of coverage of the mobile terminal.

The discussion on FutureG leads to the following possible key drivers (or demands) for the evolution of mobile networks beyond 3G:

- Open mobile service architecture;
- Adaptive personal mobility;
- Device mobility;
- Direct access to Internet.

We briefly describe each of the FutureG drivers given above in the following section.

Open Mobile Service Architecture

FutureG should have an open service architecture that will allow deployment of new services by service providers or by users anytime and at anyplace [33]. Also, different wireless access networks in FutureG should support service mobility—that is, allowing the user to use same services independently of the type of wireless access network (e.g., a user should be able to make a voice call in a wireless LAN in the same manner as in a GSM network).

Adaptive Personal Mobility

The user should have the possibility to move to other wireless networks and at the same time remain reachable using the same address (e.g., IPv6 address). For support of adaptive personal mobility, FutureG will incorporate personal agents (virtual software objects) that will reside in the Internet and should provide access and media control for the user considering the network condition (e.g., signal strength from the network, current network capacity, QoS demands) and user equipment (e.g., installed software in mobile terminal and its capabilities).

Device Mobility

Mobile terminals in FutureG are expected to become more complex because they should support different radio interfaces. Furthermore, they should have multiple sensors for gathering information from the user and environment, and they should have capabilities for presentation as well as for creation of

multimedia contents. Thus, FutureG will incorporate multinetwork and multifunction terminals. Of course, cheap single-service terminals will remain on the market, targeting different user profiles.

Direct Access to the Internet

FutureG should also incorporate direct access to Internet. This will be necessary because broadband Internet access is being provided in a rapidly increasing way, and not only by telecommunication operators and vendors (in many cases it is provided by transportation companies and power companies). Wireless LAN and/or wireless *broadband access network* (BAN) are considered part of the FutureG. Furthermore, there are existing solutions for direct access to the Internet (without any subscription, authorization, or authentication to the network operator), such as e-cash. In this case operators provide IP access, while billing of the used services is transferred to a third party (it should be a trusted party). This way, network operators as well as service providers may be paid directly or indirectly. In such a situation, users can use any service from any third party without any intervention by the network operator, which in such cases only provides access to the network.

Finally, we may define the FutureG mobile networks (e.g., 4G) considering their foreseen characteristics as described above.

Definition of FutureG Mobile Networks. FutureG (beyond 3G mobile network) will be an end-to-end wireless IP network with high bandwidth gathered by exploiting heterogeneous wireless access, and with the capability for direct IP connectivity to different multimedia services provided by a third party. The concept of FutureG is information anywhere, anytime, and in any form.

2.7.1 All-IP Mobile Network

The main drivers for an all-IP network are resource efficiency (by means of statistical multiplexing), operational costs, and the transparency of IP technology to different types of services. Therefore, telephony will also migrate from CS to PS (i.e., IP telephony). However, the inertia from 2G and the first 3G mobile systems will be an intrinsic element in this area during the first decade of the twenty-first century.

Internet services are made closer to the mobile terminal by the introduction of the GPRS system and WAP. The market will change with the emergence of new killer applications, which is happening now. Mobile Internet is the business model for the future of cellular technology. IP mobility will give the operators unlimited possibilities for service differentiation. Such an approach allows unlimited possibilities for the introduction of new services and the creation of new content, as shown in Figure 2.14.

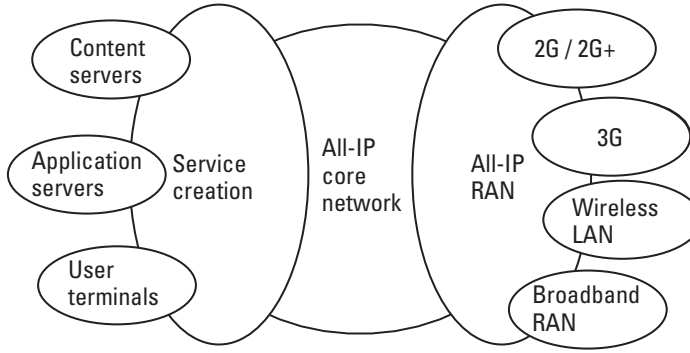


Figure 2.14 All-IP mobile network concept.

Thus, the mobile service model mirrors the fixed Internet service model, but also adds new location-based services. Such a model will significantly increase the total traffic volume. Therefore, the IP solution should be cost-effective. Also, high capacity in hot spots (e.g., business building) should be provided via WLAN or broadband RAN through IP radio access network and the same mobile terminal.

However, the network operator role will change, as shown in Figure 2.15. There will be portal gateways to the Internet that should provide advertising and content revenue. Furthermore, content providers should provide value added information. For trusted transactions (e.g., e-commerce) there will be needed trusted providers. Finally, future mobile networks shall incorporate application providers, such as wireless Web access and intranet access.

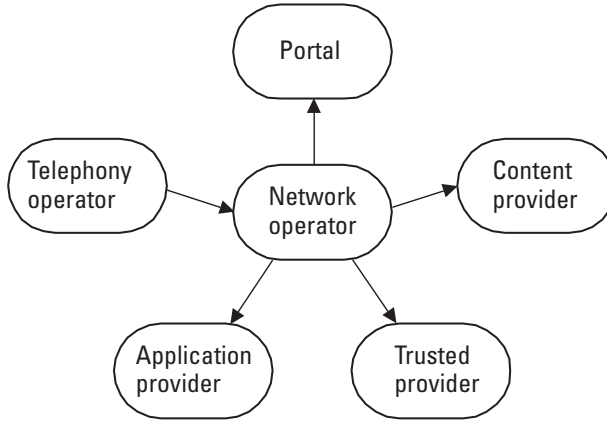


Figure 2.15 Scenario for operators in mobile networks beyond 3G.

2.8 Discussion

Wireless networks are very attractive due to possibility of communication from any place at any time. The cellular concept, as well as digital radio access and roaming introduced in 2G, made possible a global expansion of mobile networks and services. 2G networks are based on circuit switching and primarily voice service.

Development and popularization of Internet technology in the 1990s introduced many new multimedia services and contents. Internet users need mobility of the services, or in other words, mobile users want new multimedia services similar to those found on the Internet. Such a situation led to the creation of 2G+ and additional 3G mobile networks, which introduce more bandwidth and more flexibility by using wideband CDMA-like access techniques as well as IP RAN. Due to the inertia in telecommunications, circuit-switched operation continues to coexist with the packet-switched mode in 3G networks. Also, 3G networks provide many new multimedia services, such as real-time services (IP telephony, video-telephony, multimedia streaming services) and nonreal-time services (Web browsing, fax, multimedia messaging services, and location-based services).

Considering the research activities on the future wireless cellular communications, mobile networks beyond 3G should provide adaptive personal communication with direct IP connectivity, allowing migration of users, devices, and services between heterogeneous wireless networks ranging from GSM to wireless LAN and broadband radio access networks.

References

[1] Internet Host and Traffic Growth, http://www.cs.columbia.edu/~hgs/internet/.

[2] Eldering, C. A., and J. A. Eisenach, "Is There a Moore's Law for Bandwidth?" *IEEE Communications Magazine*, October 1999.

[3] ETSI TR 101 458, *Universal Mobile Telecommunications Services (UMTS): Future Direction of Standards Work on UMTS/IMT-2000*, v1.0.0, October 1999.

[4] Trillium Digital Systems, "Third Generation (3G) Wireless White Paper," http://www.trillium.com, March 2000.

[5] ETSI Guide, *Universal Mobile Telecommunications Systems (UMTS): Strategies*, ETSI EG 201721 v1.1.2, February 2000.

[6] *Universal Mobile Telecommunications Systems (UMTS): Future Direction of Standards Work on UMTS/IMT-2000*, ETSI Technical Report, ETSI EG 201721 v1.1.2, May 2000.

[7] ITU-T Recommendation, *Network Functional Model for IMT-2000*, ITU-T Q.1711, March 1999.

[8] Valko, A. G., "Cellular IP: A New Approach to Internet Host Mobility," *ACM Computer Communication Review*, January 1999.

[9] Perkins, C., (ed.), *IP Mobility Support*, RFC 2002, proposed standard, IETF Mobile IP working group, October 1996.

[10] Oliphant, M. W., "The Mobile Phone Meets the Internet," *IEEE Spectrum*, August 1999.

[11] Mehrotra, A., *Cellular Radio Analog and Digital Systems*, Norwood, MA: Artech House, 1994.

[12] http://www.gsmworld.com.

[13] Redl, S. M., M. K. Weber, and M. W. Oliphant, *An Introduction to GSM*, Norwood, MA: Artech House, 1995.

[14] Bettstetter, C., H. -J. Vogel, and J. Eberspacher, "GSM Phase 2+ General Packet Radio Service GPRS: Architecture, Protocols, and Air Interface," *IEEE Communications Surveys*, Vol. 2, No. 3, Third Quarter 1999.

[15] http://www.3gpp.org.

[16] 3GPP TS 23.107, *Technical Specification Group Services and System Aspects; QoS Concept and Architecture (Release 5)*, V5.3.0, January 2002.

[17] 3GPP TS 23.002, *Technical Specification Group Services and System Aspects; Network Architecture (Release 5)*, V5.5.0, January 2002.

[18] 3GPP TS 23.101, *Technical Specification Group Services and System Aspects; General UMTS Architecture (Release 4)*, April 2001.

[19] Chaudhury, P., W. Mohr, and S. Onoe, "The 3GPP Proposal for IMT-2000," *IEEE Communications Magazine*, Vol. 37, No. 12, December 1999, pp. 72–81.

[20] Liew, J., et al., *3G Wireless in the US: cdmaOne to cdma2000*, Harvard University, Massachusetts Institute of Technology, Tufts University, May 8, 2000.

[21] 3GPP2 P.S0001-A, *Wireless IP Network Standard*, Version 3.0.0, July 2001.

[22] Rigney, C., et al., *Remote Authentication Dial In User Service (RADIUS)*, RFC 2138, April 1997.

[23] 3GPP2 S.R0035, *Quality of Service*, Version 1.0, October 2001.

[24] 3GPP TS 22105-500, *Services Aspects: Services and Service Capabilities (Release 5)*, V5.0.0, October 2001.

[25] ITU-T Recommendation F.700.

[26] http://www.wapforum.org.

[27] http://foma.nttdocomo.co.jp.

[28] 3GPP2 S.R0022, *Video Conferencing Services – Stage 1*, Version 1.0, July 2000.

[29] 3GPP TS 26.111, *Codec for Circuit Switched Multimedia Telephony Service; Modifications to H.324*, March 2001.

[30] 3GPP2 S.R0021, *Multimedia Streaming Service – Stage 1*, Version 2.0, April 2002.

[31] 3GPP TS 22.228, *Service Requirements the IP Multimedia Core Subsystem (Stage 1) (Release 5)*, December 2001.

[32] Bria, A., et al., "4th Generation Wireless Infrastructures: Scenarios and Research Challenges," *IEEE Personal Communications*, Vol. 8, No. 6, December 2001.

[33] Kanter, T., "An Open Service Architecture for Adaptive Personal Mobile Communication," *IEEE Personal Communications*, Vol. 8, No. 6, December 2001.

3

Wireless Mobile Internet

3.1 Introduction

The Internet has experienced exponential growth in the number of hosts as well as in the number of sites during the past 10 years, as shown in Figure 2.2. One of the main reasons for such growth is the invention of World Wide Web, based on the HTTP/TCP/IP protocol stack, which provides easy creation of different content in different areas and easy access to them by simply clicking on links in HTML pages. It is based on plug-and-play and easy-to-use approaches that make the Internet the driving force of humanity at the beginning of the twenty-first century. The fixed Internet is perhaps entering a period of adolescence after a spectacular growth period, reaching almost 180 million hosts by 2002. Mobile Internet, however, is gaining momentum since the introduction of Internet connectivity is one of the main features in 3G mobile networks and beyond. In that sense, the number of Internet users will incorporate most of the mobile users. Currently, the total number of mobile users on the planet is around 1 billion (refer to Figure 2.1) and it is expected to increase further. So, the total number of Internet users (we mean users attached to the Internet by wired or wireless links) would likely exceed 1 billion within the next few years.

The common feature of the Internet is the Internet Protocol, which is deployed in all Internet protocol stacks—it is the heart of the network. IP is the layer-3 protocol in reference to the OSI model. Today's Internet basically provides only one type of service, so-called best-effort service. It does not provide guarantees on QoS parameters, such as throughput, loss, or delay.

Another characteristic of the Internet is its heterogeneity. There is heterogeneity in end nodes, either servers or clients, which can be personal computers, communicators, mobile terminals, or powerful servers. Heterogeneity also exists

in the bandwidth of links deployed in the Internet, which range from several kilobits per second up to gigabits per second. Then, there are heterogeneous protocols created and implemented over the IP, which range from connectionless protocols [e.g., *Unicast Delivery Protocol* (UDP); it is also referred to as the *User Datagram Protocol*] to connection-oriented protocols [e.g., *Transport Control Protocol* (TCP)] and multicast protocols. Also, IP runs over different underlying protocols, such as Ethernet and ATM. Finally, in the Internet there exists heterogeneity in application types, which range from nonreal-time applications (e.g., e-mail, file downloading) to real-time applications with constraints on the QoS parameters (e.g., voice over IP, audio/video streaming).

The most used transport protocol over the IP is TCP (refer to Chapter 5). Therefore, we usually refer to the Internet Protocol stack as TCP/IP. Both protocols are described in the following sections.

3.2 IP

The native Internet is built over IP version 4 (IPv4) [1]. It is designed for interconnected systems of packet-switched computer networks. The IP provides transmitting blocks of data called datagrams, from a source to a destination. Usually, we use the term "datagram" for connectionless delivery of IP packets, while for connection-oriented delivery we use the term "segment" (e.g., TCP segment).

The starvation for IP addresses due to limited address space of IPv4 led to creation of IP version 6 (IPv6) [2], which offers several improvements over IPv4. The most important of which is much larger address space.

3.2.1 IPv4

IPv4 (we refer to IPv4 as IP) provides transmission of datagrams as well as fragmentation and reassembly of long datagrams, if necessary (depending on the type of network). IP provides only functions necessary to transmit a packet (a package of bits) from its source to a destination host.

IP has two basic functions: addressing and fragmentation. Internet addresses of the source and destination are carried in IP headers. Each IP address has 4 bytes (i.e., 32 bits), which are necessary to allow routing of IP packets to the destination. The path selection for transmission is called routing. The nodes that provide routing of IP packets are called routers. Each packet is routed independently of any other packet from that connection or some other (there are no logical or virtual circuits).

The data and header from higher layers become payload for the IP packets. To this payload, IP adds a header, which is shown in Figure 3.1. IP uses four key mechanisms in providing its service: type of service, time to live, options, and header checksum.

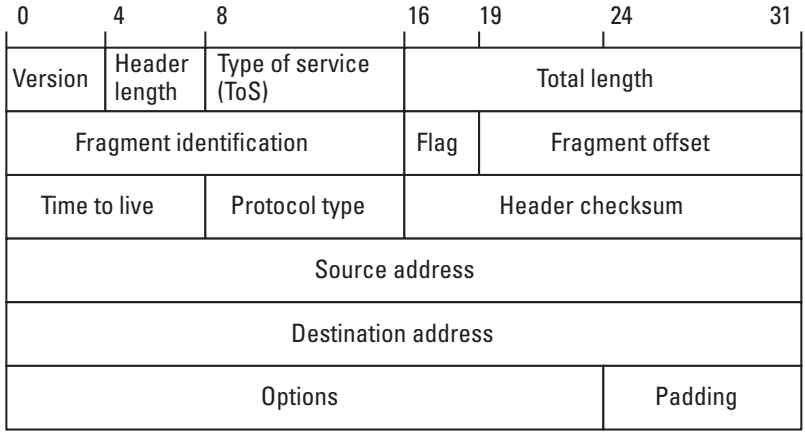


Figure 3.1 IPv4 header format.

The IP does not provide reliable transmission of the datagrams. There are no acknowledgments either end-to-end or hop-by-hop. Also, there is no error control of data, although there is a header checksum. Furthermore, IP does not provide flow control, sequencing, or any traffic control mechanism. Such control is left to the higher layer protocols (e.g., reliability, retransmissions of lost packets) or lower layer protocols (e.g., error control). But through the type of service field in the IP header, it provides the possibility for networks to apply traffic management (e.g., QoS support). It is a service indication for the network nodes when they are selecting actual transmission parameters for a particular network. *Time to live* (TTL) indicates the upper bound of the time an IP packet is allowed to reside in the Internet. It is set by the sender and reduced at each network node; thus, when it reaches zero, it is destroyed. Because TTL has 8 bits, its maximum value is 255, while usually each network node reduces its value by one. This mechanism eliminates the possibility of endless traveling of IP packets through the Internet. The header checksum is used for error check of the header. The header is more sensitive to errors because it may result in delivery of the IP packet to a wrong destination, or it may result in reset of the TTL value. So, each router in the Internet necessarily performs two functions first: error checking of the IP header, and reducing the TTL value. If header checksum fails, the datagram is discarded at once by the entity that detects the error. The identification field contains an identifying value assigned by the sender to help in assembling the fragments of a datagram.

IP exists in each node attached to the Internet, either a host or a gateway that interconnects networks. But, it does not provide any reliability or flow control. For reliable delivery of packets, we need such functions to be implemented in a higher layer protocol such as TCP.

3.2.2 IP Version 6

IPv4 has several drawbacks, the most important of which is its small address space, especially considering the exponential growth of the users. IPv6 is a new version of the IP [2]. It was created as a successor of IPv4, providing several main changes in the following categories:

- *Expanded addressing capabilities:* IPv6 extends the address size from 32 bits in IPv4 to 128 bits. Also, it supports more levels of addressing hierarchy. The address space is expected to be enough in a longer time period. Also, besides existing (but improved) multicast address, IPv6 adds a new type of address called "anycast" address, which is used to send a packet to any one in a group of nodes. However, in the transition period over the next several years both protocols will coexist. Therefore, IPv6 addressing is compatible with IPv4 (i.e., when IPv4 packets travel through IPv6 network the last 32 bits of IPv6 address corresponds to IPv4 address).
- *Flow labeling capabilities:* It allows possibility of labeling packets that belong to a same flow, for which the sender requires specific handling, such as real-time communication.
- *Authentication and privacy capabilities:* IPv6 introduces extensions to support authentication, data integrity and confidentiality (if required).

Some IPv4 fields are dropped in the IPv6 header, which is shown in Figure 3.2. For example, the header checksum is dropped due to its redundancy (i.e., lower layer protocols always use an error control mechanism). Also, the identification field is reduced because IPv6 does not provide segmentation/reassembly possibility for the intermediate network nodes (it is allowed only at the end points of the communication: receiver and sender).

Also, IPv6 allows less stringent limits on the length of options as well as greater flexibility for the introduction of new options in the future. Considering the QoS support, IPv6 header has traffic class field, which corresponds to ToS field in IPv4 header format.

If we compare header formats of both Internet Protocols so far, IPv4 and IPv6, we may conclude that IPv6 has higher header redundancy due to increased address length (to cope with higher header redundancy, some fields from IPv4 are dropped in IPv6 headers). But overall, IPv6 provides a solution to lack of IP addresses as well as new authentication and flow control capabilities, which allow support of real-time services (per flow).

Similar to IPv4, IPv6 also needs an upper layer protocol to provide transmission reliability (e.g., TCP).

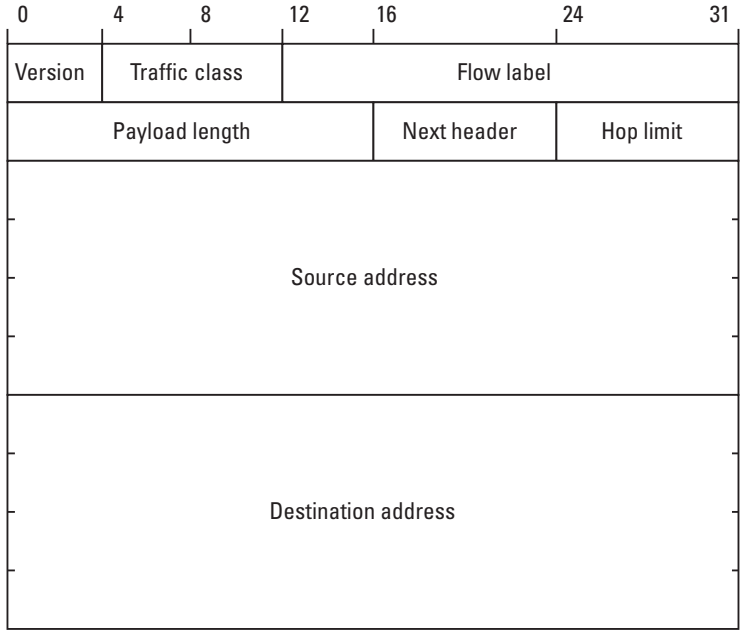


Figure 3.2 IPv6 header format.

3.3 Transport Control of IP Packets

For reliable data transport in today's Internet, the de facto standard is the TCP [3]. Most of the popular Internet services today are based on the TCP/IP protocol stack. Examples of such services are WWW, *File Transfer Protocol* (FTP), and Telnet. Measurements given in Chapter 4 show that TCP traffic covers 95% of all bytes, 85% to 95% of all IP packets, and 75% to 85% of all flows. From the rest, the main share belongs to UDP, which has unreliable data transfer, and it is usually used in real-time communication where retransmissions of lost packets is not desirable due to unacceptable delays. Such a situation in Internet traffic is mainly due to the WWW, which is so far the most used service on Internet accounting for 65% to 80% of all bytes.

The original specification that introduced TCP is RFC 793 [3]. However, variants of TCP have been developed and implemented in the past 20 years, such as Tahoe and Reno [4].

TCP is based on acknowledgments of the successfully received packets. So, it belongs to the *automatic repeat request* (ARQ) family of transport protocols, where acknowledgments are sent after one or several successfully received TCP packets (i.e., data blocks). We refer to a TCP continuing sequence of bytes as a segment. Each TCP segment is identified with a 32-bit long start and end sequence number. Sequence numbers are byte based (i.e., they represent the

number of bytes successfully transmitted from the start of the TCP communication). TCP at the receiving side acknowledges the last successfully received byte (in order). So, TCP provides a fully reliable, byte-stream in-order delivery of IP packets. Each TCP connection is connected to higher layer protocols by using so-called socket interface, which is the software-defined communication point between transport and session layers in reference to the OSI model. In the opposite direction, towards the IP-layer, TCP segments are placed as a payload into IP packets. The typical size of a TCP segment is 536 or 512 bytes.

3.3.1 TCP Mechanisms

TCP incorporates mechanisms for reaction in a case of packet losses. Losses occur in the Internet mainly because of congestion at network nodes (routers, switches), due to the principle best-effort service and the bursty nature of IP traffic. When TCP discovers that data has been lost in the network, it recovers from it by retransmitting the missing segments. TCP will discover loss by receiving duplicate acknowledgments at the sender's side, or if the sender does not receive any acknowledgement in the time period longer than a predefined timeout.

Because TCP was originally created for wired packet networks, it always assumes that losses are occurring due to congestion only. It reacts to the congestion by decreasing its data rate based on a congestion avoidance mechanism. TCP uses window-based congestion avoidance [5] with two windows: congestion window (*cwnd*) and receiver advertised congestion window (*rcvwnd*). The receiver advertises congestion window size at the sender. The sender determines the congestion window from receiver advertised window and congestion information (e.g., missing acknowledgement for sent packets). For explanation of congestion avoidance, it is appropriate to introduce the TCP slow start mechanism first [6].

The slow start mechanism operates by observing that the rate at which new packets should be sent is the rate at which acknowledgments are returned by the receiver side. Here, the congestion window is initialized to one segment (i.e., the segment size announced by the other end, or by default, is typically 536 or 512 bytes). Each time an acknowledgment is received, the congestion window is increased by one segment. After a certain threshold—called the slow start threshold (*ssthresh*)—is reached, the connection moves into the congestion avoidance phase. In this phase, the congestion window effectively increases by one segment for each successfully transmitted window. The sender can send up to the minimum of the congestion window and the advertised window. The congestion window is flow-control imposed by the sender, while the advertised window is flow-control imposed by the receiver. The sender sets the congestion window size based on the perceived network congestion, while the

advertised window is related to the amount of buffer space in the receiver reserved for that connection. When the sender discovers a packet loss, it halves its congestion window. If timeout occurs, the congestion window is set to one segment and the connection goes through the slow start once again. TCP congestion avoidance and loss recovery mechanisms are shown in Figure 3.3.

By increasing the congestion window size, the TCP connection increases the data rate. At some point the capacity of the shared link (or hop) will be reached, resulting in discarding packets by the congested router. When TCP at the sender discovers packet loss, it retransmits the lost packets. There are two types of loss recovery: timer-driven retransmissions and data-driven retransmissions. Timer-driven retransmissions happen when the sender does not receive a positive cumulative ACK for a segment within a certain timeout interval (cumulative ACK acknowledges several consecutive segments—that is, it is not necessary for each single segment to be separately acknowledged). To determine the timeout, we usually use an estimation of the round-trip time, which we may get by using the *exponential weighted moving average* (EWMA) formula:

$$srtt = \alpha * rtt + (1 - \alpha) * srtt \tag{3.1}$$

where *srtt* is smoothed round-trip time, *rtt* is the round-trip time, and α is the EWMA constant with a value typically set to 0.125. A timeout occurs if the

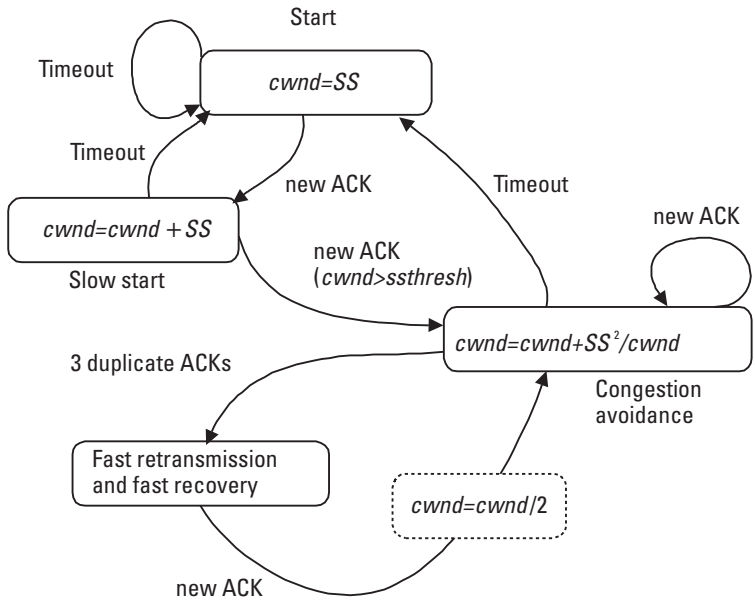


Figure 3.3 TCP congestion avoidance and loss recovery mechanisms.

sender does not receive an ACK for that segment within time period *rto* (round trip timeout):

$$rto = srtt + 4 * rttvar \tag{3.2}$$

where *rttvar* is round-trip time variance, which we may calculate using (3.1), but now with $\alpha = 0.25$. Timer-driven retransmissions are typically with a 500-ms granularity. However, timeouts should be the last mode of recovery when all other methods fail.

Data-driven recovery uses a technique called fast retransmission. It is usually followed by the fast recovery mechanism. Both are now outlined.

Fast Retransmission

This mechanism of TCP is proposed to deal with duplicate ACKs due to reordering of segments. Since TCP does not know whether a duplicate ACK is caused by a lost segment or a reordering of the segments, the fast retransmission mechanism forces TCP to wait for a small number of duplicate ACKs to be received at the sender. It is assumed that if there is just reordering of segments, with a high probability that there will be one or two duplicate ACKs before the reordered segment is processed, this will then generate a new ACK. On the other hand, if three or more duplicate ACKs are received in a row, it is almost certain that a segment has been lost. Then, TCP performs retransmission of the missing segment before the retransmission timer expires. Thus, this mechanism is called fast retransmission.

Fast Recovery

Fast recovery allows the invocation of congestion avoidance instead of slow start after retransmission of the missing segment by fast retransmission algorithm. Since the arrival of duplicate ACKs signals to the sender that data is flowing between the two ends of the communication path, there is no reason to reduce flow suddenly by going into slow start. The fast retransmit and fast recovery mechanisms are usually implemented together as follows:

1. If sender receives three duplicate ACKs in a row, then *ssthres* = min{(1/2)**cwnd*;2}. The sender retransmits the missing segment. Then, the sender takes into consideration that three duplicate ACKs means that three packets have reached the receiving end and have left the network, thus *ssthresh* = *ssthresh* + 3**segment_size*.
2. Another duplicate ACK increments the congestion window size *cwnd* by the segment size, because a new ACK means that an additional segment has left the network. If new *cwnd* allows, sender transmits a packet.

3. When the next positive ACK arrives (that acknowledges the new data), then *cwnd* = *ssthresh* (the value from the first step). This ACK should acknowledge all the intermediate segments sent between the lost packet and the receipt of the first duplicate ACK. So, here TCP is in congestion avoidance.

Fast retransmissions are efficient for single packet losses, but they are not sufficient for recovery from multiple losses in a single window [4]. This usually results in coarse-grained timeout before the packet is retransmitted. There are several variants of TCP depending upon the included mechanisms. We outline the most commonly used TCP implementations in the following section.

3.3.2 TCP Implementations

There are different implementations of TCP. The most used versions are Tahoe and Reno.

TCP Tahoe includes slow start, congestion avoidance, and fast retransmission mechanisms. In Tahoe, slow start follows fast retransmission. If we additionally include the fast recovery mechanism to TCP Tahoe, we obtain the TCP Reno version. The mechanisms described in the previous sections are all implemented in Reno.

TCP Tahoe functions well at single loss within the congestion window. But it follows the congestion by invoking slow start. TCP Reno improves the performances of the TCP stream at a single loss per window, but problems occurs when multiple packets are dropped from a window of data. Such behavior at multiple dropped packets from a window is overcome by some changes implemented in latter versions of TCP, such as: TCP NewReno and TCP *selective acknowledgments* (SACK).

TCP NewReno makes simple changes to the Reno version to avoid waiting for the retransmit timer when multiple packets are lost from a window. It uses partial ACKs to retransmit missing packets (i.e., each duplicate ACK indicates that the following segment is lost and it is retransmitted until TCP receives a positive ACK). At all times TCP remains in fast retransmission and fast recovery phases. This way, TCP NewReno allows TCP to recover X multiple packet losses from a window of data within X round-trip time intervals.

TCP may experience poor performance when multiple packets are lost from one window of data. For such situations one proposed solution is TCP SACK [7]. There are several ways of implementing SACK. But in all of them the common characteristic is an additional SACK packet sent by the receiver at each duplicate ACK, together with the duplicate ACK. By using SACK, the sender keeps track on the missing segments more precisely, even if it is more aggressive. In the case of cumulative ACKs only, a TCP sender can only learn about a single

lost packet per round-trip time. One way of implementing SACK is described in [7]. In this scheme, the receiver reports up to three of the last received, out-of-order, maximal contiguous blocks of data, in addition to the cumulative ACK. That way, the sender can accurately know which segments have reached the receiver side. So, TCP SACK allows recovery from multiple lost packets in a window of data within one round-trip time, which is not the case with Tahoe and Reno versions of TCP. In a mobile environment, packet losses may occur due to wireless link errors, which are location-dependent and time-varying. These errors are usually bursty in nature, thus producing multiple packet losses within one window.

In that sense, one may find SACK appropriate for wireless links. Additionally, TCP-like congestion control is considered as one alternative in *Reliable Multicast Transport* (RMT) protocols [8]. There are also many other modifications of TCP that attract more or less attention of the researchers and industry.

3.3.3 Stream Control Transmission Protocol

Stream Control Transmission Protocol (SCTP) is the most recent IP transport protocol that is standardized by IETF [9]. It exists on an equivalent level as the UDP and TCP protocols, which provide transport layer to most Internet applications. SCTP is designed to transport signaling messages from the PSTN over IP networks, but it also can be used in broader applications.

SCTP is a result of the study conducted within IETF that started in 1998, targeted to create an Internet equivalent to ITU-T *Signaling System 7* (SS7) transport services. The original protocol framework was initially named *Common Signaling Transport Protocol* (CSTP), the requirements of which are listed in [10].

Unlike TCP, SCTP provides a number of functions that are essential for telephony signaling transport, and at the same time it can potentially benefit other applications needing transport with additional performance and reliability.

SCTP also has similarities with TCP. For example, SCTP provides reliable transport service and a session-oriented mechanism (i.e., communication between the end points is established prior to data being transmitted). Also, it provides TCP-friendly congestion and flow control. SCTP uses the SACK version of TCP protocol (one SACK per every received packet at the receiver). Flow and congestion control mechanisms follow TCP algorithms: slow start, congestion avoidance, fast recovery, and fast retransmit. Thus, SCTP is rate adaptive as TCP, although for some application it may be likely that adequate resources will be allocated to SCTP traffic to ensure prompt delivery of time-sensitive data. One should know that TCP is byte oriented while SCTP is message oriented. Message-based orientation of the protocol is advantageous over

TCP, which is connection oriented, ensuring a more reliable and flexible transmission of small amounts of data, like signaling information.

Another important feature of SCTP, which provides reliability, is multihoming. This is the ability of a single SCTP endpoint (each SCTP session is between exactly two endpoints) to support multiple addresses. This approach increases survivability of the SCTP session in the presence of network failures. Due to the importance of signaling information, multihoming is used for redundancy, and not for load sharing of signaling traffic (e.g., one IP address is used as primary address for normal transmission, while additional IP addresses are used at the retransmissions to improve the probability of reaching the remote end).

Unlike TCP, which assumes a single stream of data, SCTP allows data to be partitioned into multiple streams (the name SCTP is derived from this streaming feature), so that messages lost in any one stream will affect the delivery within that stream only, and not the other streams. In this approach multiple streams belong to a single SCTP session. For example, multistreaming can be used for delivery of multimedia documents, such as a Web page, over a single session. Another example of multistreaming is telephony signaling over IP network, where one should maintain sequencing of messages that affect the same call or channel.

Due to its characteristics, SCTP is considered as an alternative to provide signaling over IP core network in UMTS in preference to TCP, and in parallel to SS7 used in the circuit-switched core network.

3.4 QoS Provisioning in the Internet

Although the Internet was created as a network with one-type service for all, the rapid development of the Internet into its present commercial infrastructure raised demands for QoS support. This is due to the variety of Internet applications and the increased number of users, which have different demands for content, type of information, and quality of service. Many times has it been debated whether QoS provisioning is needed for the Internet. One opinion is that fiber technology, such as *wavelength division multiplexing* (WDM) shall provide cheap bandwidth as much as it is needed. On the other hand, the experience of the development of applications in recent years shows that no matter how much bandwidth is provided, new applications will be invented to consume it. In a mobile environment, however, we have limited resources due to limited frequency spectrum available for wireless communications over a given geographical area.

The IETF has proposed several mechanisms for QoS provisioning in Internet. The most attention is given to *Multiprotocol Label Switching* (MPLS),

Integrated Services with Reservation Protocol (RSVP), and Differentiated Services [11–13]. All of them are defined for the wired Internet. However, the number of mobile users grows even faster than the number of Internet users. As we already discussed in Chapter 2, the convergence of mobile networks and the Internet is a foreseen process. Such convergence raises new demands on wireless access to Internet considering the QoS provisioning. In the following sections we go through QoS mechanisms proposed for the Internet, and then we consider such mechanisms in a cellular wireless network.

3.4.1 MPLS

MPLS is a scheme that utilizes a fixed-length label for packet handling. Each packet that enters an MPLS-enabled network domain obtains an added MPLS header, which is encapsulated between the link layer header and the network layer header. The MPLS capable router is called the *label switching router* (LSR). Such a router analyzes the label only in forwarding the packets. Thus, MPLS is packet-forwarding scheme. The network protocol can be IP or another (e.g., ATM). Therefore, this scheme is called Multiprotocol Label Switching.

For each packet, the router that adds the label is called ingress router, while the router that extracts the label is called egress router. The header of a MPLS packet contains a 20-bit label, where 3 bits are defined for the *class of service* (CoS) field, 1 bit is for indication of the label stack, and 8 bits are used to specify TTL for the packet within the MPLS domain only.

MPLS uses protocols to distribute labels within the domain, to set up so-called *label switched paths* (LSPs), which are paths between the ingress LSRs and egress LSRs. They are similar to the virtual circuits in ATM networks. For LSP setup, MPLS uses RSVP protocol (we refer to this later in this chapter) or a specialized protocol for label distribution called *Label Distribution Protocol* (LDP) [12]. Each MPLS-enabled router LSR has a routing table for the labels, which is managed by the LDP. When an LSR receives a labeled packet, it will use the label as the index to look up the forwarding table. The packet is processed according to the table entry. The LSR is allowed to change the label of the packet, if necessary. So, each packet gets a MPLS label at the entrance of a MPLS domain (Figure 3.4), which is used by the internal routers for routing and traffic control. Before a packet leaves the MPLS domain, the egress router removes its MPLS label.

MPLS may also provide efficient tunneling of the packets between two network nodes (ingress and egress routers), where the path is completely determined by the label assigned by the ingress router [14]. This requires a protocol that will refresh the routing tables of internal routers (e.g., RSVP). Since the label applied at the ingress router of the LSP defines a traffic that flows along the label-switched path, these paths can be treated as tunnels, and we refer to them

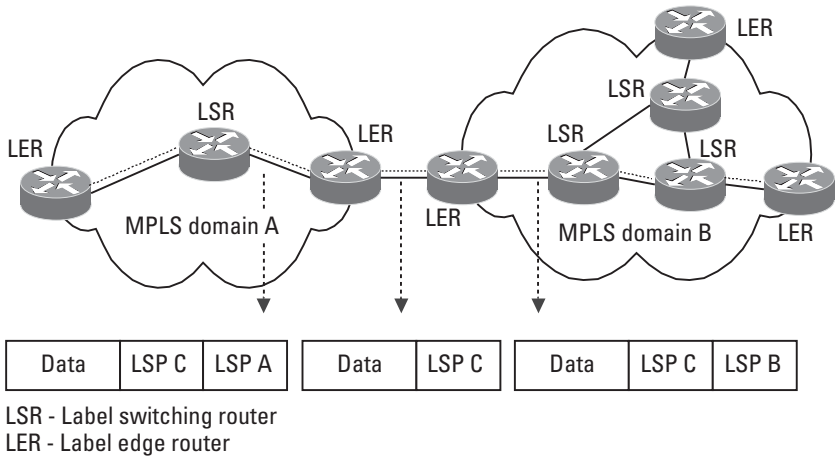


Figure 3.4 MPLS architecture.

as LSP tunnels. Each LSP is established with a set of traffic parameters (i.e., constraints), such as bandwidth. To provide certain QoS we need to perform *constraint-based routed label switched paths* (CR-LSPs) [15]. After CR-LDP is set up, its bandwidth may be dynamically changed upon new requirements for the traffic on that path.

Overall MPLS provides means for traffic engineering in the Internet (i.e., performance optimization of the network). Two main advantages of MPLS are:

- Faster forwarding;
- Efficient tunneling of packets.

Also, we may apply MPLS in wireless IP-based networks. In this case, the basic requirements put on MPLS from the underlying wireless IP access technology are:

- Mapping of all incoming IP packets into the MPLS domain at the edge routers, and removal of the labels for outgoing IP packets;
- Establishment of LSP through the network routing protocols. There are two possibilities for routing within MPLS domain: hop-by-hop routing or explicit routing (using predefined path);
- LSRs need to support label swapping for forwarding IP packets and IP merging for multicast. Also, LSRs need to process each packet, such as decrementing TTL, next hop determination, and so forth;

- LSR needs to support label distribution through LDP. All labels are stored in a base called *label information base* (LIB).

In a cellular network one type of label edge router may be a base station. Another possible type of edge router is a gateway-node of the wireless network to the wired Internet. In this situation it is suitable to perform classification of the traffic in the wireless network and its differentiation to/from mobile users, which should be performed at the wireless access nodes (e.g., base stations). Therefore, implementation of MPLS in a wireless network will not have an impact on the radio access network, which is a primary interest. It may, however, be applied in the wireless core network.

3.4.2 Integrated Services

Integrated Services architecture called Int-Serv is defined by IETF in RFC 1633 [16]. The main idea behind this proposal is support of real-time services in the Internet.

Integrated Services introduces a fundamentally new concept for the Internet. This protocol assumes that resources are reserved for every flow requiring QoS at every router hop in the path between the sender and the receiver. To be able to support per-flow traffic management, the network needs to establish an end-to-end path by using signaling, which is provided by RSVP. This is in contrast to the traditional approach in the Internet, where intermediate routers do not store routing information for each flow. Integrated Services provides two additional QoS classes (besides the best-effort traffic class):

1. *Guaranteed service* [17] for applications requiring bounded end-to-end queuing delay of packets and bandwidth guarantees. The delay has two parts: fixed and queuing delay. Fixed delay is a property of the chosen path by the setup scheme. Hence, only the queuing delay is determined by the guaranteed service. In this concept a flow is described using a token bucket; and given this description of the flow, a service element (e.g., a router) computes various parameters describing how the service element will handle the flow's data. However, a setup mechanism (e.g., RSVP) must be used for guaranteed reservations. To achieve bounded delay requires that every service element (i.e., node) in the path supports guaranteed service, although one may benefit also with its partial deployment.
2. *Controlled load service* [18] (or controlled link sharing) for applications requiring reliable and enhanced best-effort service. This service uses admission control to assure that this service is received even

when the network element is overloaded. In other words, the controlled load does not accept or provide specific target values for delay and loss, but it provides a commitment by the network element to provide service equivalent to that provided by uncontrolled (best-effort) traffic under lightly loaded conditions. For example, a possible implementation of this service is to provide a queuing mechanism with two priority levels: a high priority for controlled load traffic, and a lower priority for best-effort traffic.

To be able to provide such QoS classes, network nodes must maintain a per-flow *soft state* (i.e., flow-specific state). A soft state is a temporary state governed by the periodic expiration of resource reservations. Soft states are refreshed by periodical RSVP messages called PATH messages (Figure 3.5). Usually, a PATH message is sent every 30 seconds to maintain the reservations [19]. It is routed through the Internet as an ordinary IP packet. PATH messages contain the traffic characteristics of the source. After reception of the PATH message, the receiver sends a so-called RESV message back to the sender. When this packet passes through the intermediate routers on the path between the sender and the receiver, it performs reservation of resources. Each router may accept or reject such reservation request (if some router rejects the reservation request, it sends a notification packet to the source). If all intermediate routers accept the reservation request, then each of them allocates resources for the flow (i.e., link bandwidth and buffer space at the router).

Integrated Services are implemented by four components in the intermediate routers: the signaling protocol (e.g., RSVP), the admission control mechanism, the classifier, and the packet scheduler. We now describe all four components considering the wireless access networks.

Reservation Protocol

This protocol makes reservations in the routers along the path of the packets from the sender to the receiver. There are two types of reservation protocol:

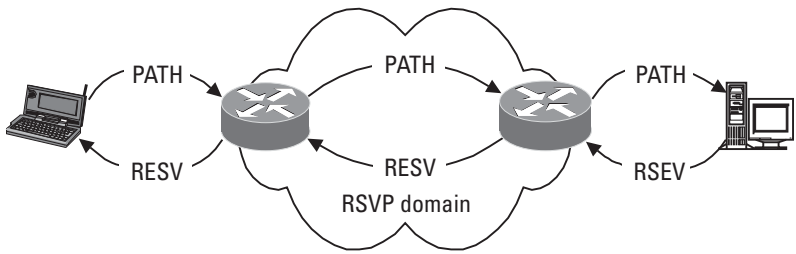


Figure 3.5 Resource reservations in Integrated Services scheme.

- *Hard state:* This type is connection-oriented, and all packets go through the same intermediate nodes. In this case, the connection is made and removed completely.
- *Soft state:* This is a connectionless state, where the reservation for a specific flow is saved in a cache at intermediate routers, and it is updated periodically as discussed above. The most used reservation protocol for Integrated Services is RSVP, which uses the soft-state method.

Integrated Services allow unicast and multicast reservations. So, the wireless access technology must be able to do such reservations, as well as to change a reservation (style and reserved resources) during a session.

Admission Control Mechanism

The admission control mechanism decides whether a request for resources can be granted. This mechanism is invoked at each node to make a local accept/reject decision. It also has a role in accounting and administration. When we consider wireless access technology, we must support mobility. In relation to admission control, the wireless network must be able to find out if a negotiated QoS can be guaranteed when handovers are likely to happen. However, the negotiating access point (e.g., base station) together with the core network nodes must make this decision.

Classifier

When a router receives a packet, the classifier performs a classification and puts the packet in a specific queue based on the classification result. All packets from the same class get the same treatment from the packet scheduler. A class in this model may correspond to a variety of flows, attributed by a QoS or to a particular organization. Furthermore, a class might hold a single flow (i.e., separate class for each flow) like in routers near the periphery (e.g., access network). Backbone routers may choose to map many flows into a few aggregate classes.

Packet Scheduler

This schedules the packets to meet their QoS requirements. The packet scheduler manages the forwarding of different streams using a set of queues and timers. It is implemented at the point where the packets are queued.

Policing and traffic shaping functions differ from the admission control. Because wireless resources are very scarce, it is recommended that the policing function (e.g., the token bucket algorithm, as given in Figure 2.6) be implemented in the wireless access point (i.e., node). However, it is not always possible to implement a policing function at the wireless access node. A similar

discussion holds for traffic shaping. Packet policing does not change interpacket distance, it just marks the packets as conformant (packets that comply to the SLA) and nonconformant (packets that do not comply to the SLA).

Integrated Services has several disadvantages, as given here:

- The amount of information increases proportionally with the number of flows. This places a huge storage and processing overhead in the routers. So, scalability is the main problem. It can be dealt by limiting the number of classes, at least in the backbone networks.
- It places high demands on routers. All of them must implement the RSVP, the admission control module, the classifier, and the packet scheduler.
- Guaranteed service requires ubiquitous deployment (in all routers in the path between the sender and the receiver). In the case of the controlled-load service we may utilize an incremental deployment (i.e., only at bottleneck routers and tunneling the RSVP messages in the rest of the domain).
- Time-varying and location-dependent bandwidth (e.g., due to interference and bit errors) of the wireless link is also a problem for the Integrated Services model. For example, a user that is experiencing a temporary higher error ratio may suffer a forced termination of the RSVP connection.

3.4.3 Differentiated Services

The Differentiated Services architecture [20] is proposed as a response to the scalability problems in the Integrated Services concept. DS architecture reduces the state of information stored in the network compared to the IS architecture, by providing QoS to limited number of classes.

DiffServ is based on class identification by using the DS header field, which is intended to supersede the existing definitions of the IPv4 ToS octet and IPv6 traffic class octet [21]. In the DS field, 6 bits out of 8 bits are used as a *DS code point* (DSCP) to specify the QoS requirements, while 2 remaining bits are currently unused (Figure 3.6). DSCP is used to differentiate aggregate flows from different traffic classes. It is incompatible with IPv4 ToS, where the first 3 bits are used to specify the precedence, and the next 4 bits are used to specify the requirements on delay, throughput, reliability, and cost. The presumption is that DS domains protect themselves by deploying demarking boundary nodes.

The basic principle of DS is packet-forwarding treatment, which is defined by *per-hop behavior* (PHB) [21]. Basic service in DS, when nothing else is specified, is the best-effort service (all DSCP bits are zeros). By marking the

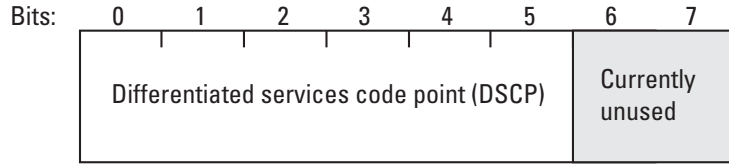


Figure 3.6 Differentiated Services field in IP headers.

DS field differently and handling packets based on their DS fields (e.g., by traffic conditioners), we may create several differentiated service classes. Therefore, one may refer to DS as a relative priority scheme.

In order for a customer to receive DS from his or her *Internet service provider* (ISP), the customer must have a *service level agreement* (SLA) with the ISP. SLA can be static or dynamic. Static SLA is made on daily, weekly, or monthly bases. Dynamic SLA requires a signaling protocol, such as RSVP, for requesting services on demand. The network under control of one ISP is usually called a domain. With the aim to provide DS, edge routers of the DS domain should classify, police, and shape the traffic entering the network domain. When a certain packet enters one domain from another, its DS field may be re-marked according to the SLA between the two domains. A classifier selects the packet based on the DSCP value in the packet header. Using the QoS mechanisms, such as classification, policing, shaping, and scheduling, different service classes can be provided. Such examples include: *premium service* for applications requiring low delay and low jitter; *assured service* for applications requiring better service than best-effort service; *olympic service*, which is further divided into three service types (gold, silver, and bronze) with decreasing quality.

DS conceptually differs from IS. The number of classes is limited within DS due to the limited size of the DS (or ToS) field in IP headers. Furthermore, DS does not have the scalability problem as IS does. The amount of information stored at a network node is proportional to the number of classes rather than to the number of flows. Another advantage of DS is in that classification, policing, shaping, and admission control should be performed only at the boundary routers of an ISP's domain. This way, intermediate routers can easily perform fast forwarding of packets, while boundary routers do not need to forward packets very fast because user access links are many times slower than the core network links. Because wireless resources are also limited and scarce, DS mechanisms seems to be convenient for such environment, while for the core network we can add bandwidth as required (we are not bandwidth limited in the wired part of the network).

So far, IETF has proposed two PHB proposals as standards: *expedited forwarding* (EF) [22] and *assured forwarding* (AF) [23]. Any wireless access network, part of a DS domain, should support at least one of these PHBs.

3.4.3.1 AF Service

The assured forwarding service is created for customers that demand reliable communication even in the presence of network congestion. We may use AF for flexible applications that can tolerate some QoS degradations (e.g., packet loss). This service provides delivery of IP packets in four different AF classes (class 1 to 4). Each DS node allocates a certain amount of resources (i.e., buffer space and bandwidth) for each AF class.

Classification and policing are performed at the ingress routers of the ISP network. All packets that do not exceed the negotiated QoS profile are considered as in-profile, while the excess packets are considered as out-of-profile. All packets, in-profile and out-of-profile, are buffered in the same queue to avoid out-of-order delivery. In a case of network congestion, out-of-profile packets are discarded first.

An AF mechanism must detect and respond to long-term congestion in terms of minimizing it for each traffic class. Short bursts may be handled by buffering the packets. But long-term congestion should be dealt with by dropping packets. However, we want the dropping of packets to be independent of short-term traffic characteristics. This way, all flows with equal data rates, but with different burstiness, should experience equal probability of dropping packets in longer time periods. One way to perform such queue management is random packet dropping.

A typical scheme that uses random dropping is *random early detection* (RED) [24]. This uses two congestion thresholds. When congestion is below the first threshold, none of the packets is dropped. But, if the congestion (expressed in the length of the queue) increases beyond the threshold, then the router drops packets randomly with probability p, which increases linearly with the congestion, going from the first to the second threshold. When congestion reaches the second threshold (e.g., queue size), all arriving packets are dropped ($p = 100\%$). This queue management will trigger all TCP flow control mechanisms at different end hosts and at different times. This way, the RED scheme prevents queues from overflowing, thus avoiding tail-drop behavior (in that case, a router drops all subsequent packets when a queue overflows). The drop-tail scheme is typical for the *first-in first-out* (FIFO) scheduling mechanism. It is inconvenient for Internet traffic because it triggers TCP flows to decrease and then to increase their rate simultaneously.

Each of the four AF classes has the possibility of three different priorities for packet dropping: low, medium, and high drop precedence [23]. Each node in a DS domain should have separate queues for each AF traffic class. Network nodes with DS capability perform class differentiation by matching the DSCP field to a particular packet handling mechanism. Packets received with an unrecognized code point are forwarded as if they were marked for the default behavior (e.g., best-effort service).

3.4.3.2 EF Service

Expedited forwarding service (or premium service) [22] is targeted to applications that have stringent requirements on packet delay and jitter, as well as assured bandwidth, such as Internet telephony, videoconferencing, and *virtual private networks* (VPNs). The delay and delay variation (jitter) occur due to queuing packets at the network nodes. Increase in the traffic queue occurs when the departure rate is close or slower than the arrival rate in the same node. EF sets up the nodes in such a way that the aggregate traffic has a minimum departure rate that is independent of the intensity of the other traffic at the node. It uses PHB as AF does. However, EF PHB does not provide quantified guarantees on jitter or delay, but these parameters are assumed to be sufficiently low (to support the applications).

The EF service is implemented as follows. At the ingress nodes traffic policing and shaping is applied. So, all nodes within the EF-capable domain assume that traffic is conditioned (i.e., there is minimum departure rate at each intermediate node in the DS domain). To provide small delay and jitter, EF traffic should always see an almost empty queue (i.e., the average length of EF queues should be kept small). The percentage of the traffic in the network is kept low enough to provide constraints on delay and jitter by applying SLA. There are two types of SLA for the EF service: static and dynamic SLA. Static SLA is usually provided via a subscription. Dynamic SLA allows customers to request EF service on demand without a subscription to it. In this case, admission control should be applied at the network nodes. For control of the conformance of the flows to their SLAs, network nodes do traffic policing and shaping. All nonconformant packets (at traffic policing) should be already discarded at the ingress nodes.

We may provide EF service by using priority over services such as AF. To avoid low QoS for the less demanding services, we usually use a small part of the link bandwidth for EF traffic (e.g., 10%). However, unevenly distributed traffic within the DS domain may cause bottlenecks in some parts of the network. Therefore, although EF traffic is limited, ISP cannot guarantee that there will be no starvation for AF and best-effort services during some time periods [11]. This situation may be solved with an appropriate packet scheduling mechanism for EF and AF service classes, such as *weighted fair queuing* (WFQ).

3.4.3.3 Differentiated Services in Wireless Access Networks

The ISP controls service allocation in a DS domain. There are two types of service allocation:

- Each host decides which service to use.
- There is a resource controller called the *bandwidth broker* (BB), as shown in Figure 3.7. The bandwidth broker may be a host, router, or active software process in some of the edge routers.

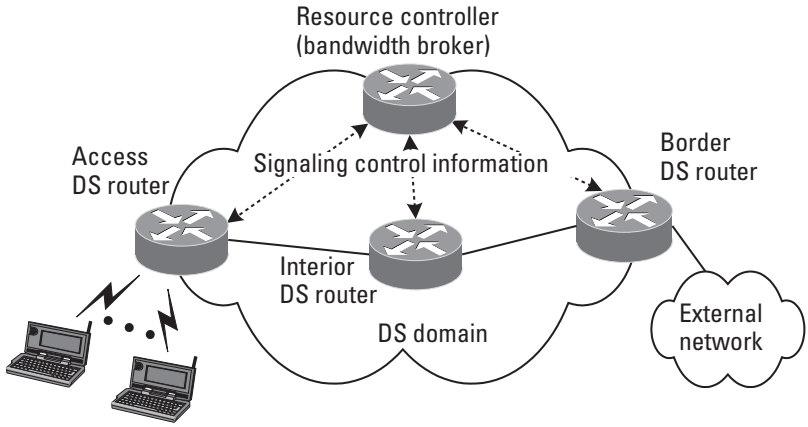


Figure 3.7 Differentiated Services architecture.

Also, in a case of wireless access to the network, there is a need for SLA between the wireless access network and interconnection network [13]. IP packets should be marked according to the SLA. Usually, the source (e.g., mobile host) marks the packets, but at least the ingress node (e.g., a base station) needs to re-mark or mark the packets. Ingress nodes also perform classification, policing, and shaping of the incoming traffic. Wireless access network needs admission control in a case of dynamic SLA to allow support of different QoS demands. Admission control is a task for the bandwidth brokers. Furthermore, wireless access networks with DS capability require PHB (i.e., packet forwarding treatment). PHB is suitable for wireless networks because of its characteristics (i.e., it does not provide quantitative guarantees on the QoS, but it provides higher QoS for one class than for a lower level class). This approach is suitable for wireless networks, where the wireless interface with time-variable BER does not allow quantitative guarantees on the QoS. Usually, cellular networks use only one wireless hop (in a case of communication between a mobile terminal and a fixed node) or two hops (in a case of communication between two mobile terminals). Thus, we have consecutive wireless and wired hops within one communication link end-to-end. In such a case, DS is one of the most suitable QoS mechanisms. So, for wireless access network we may prefer DS to other QoS mechanisms, such as MPLS and Integrated Services.

3.5 Introduction of Mobility to the Internet

Although development of both technologies, cellular mobile networks and the Internet, began separately without an idea for their interconnection, today we

are facing a need for their integration. This can be seen from the IETF's proposals for introducing mobility to the Internet, as well as from the requirements of the cellular mobile systems for packet-based communication and different multimedia services, on the way from 2G towards 3G and beyond.

3.5.1 Mobile IP Protocol

The main problem in the process of introducing mobility to the Internet is IP addressing. The IP address is a unique address for each network access point (e.g., in a router, a terminal, and so forth). Furthermore, the IP address is used for routing packets in the intermediate routers between the source and the destination. So, the main problem for mobility in the Internet is how to handle the mobile terminal's IP address and routing information when the mobile host makes handover between two wireless access points (e.g., base stations) or when it roams between two network domains (i.e., between two network operators). A solution to this problem is provided through the Mobile IP protocol [25]. This protocol provides mobility support and at the same time is transparent to the transport and higher protocol layers. Therefore, implementation of Mobile IP does not require changes in the existing nodes and hosts on the Internet. In the following we define the Mobile IP.

In Mobile IP all required functionalities for handling mobility information are embedded in three major subsystems: a *home agent* (HA), a *foreign agent* (FA), and a *mobile node* (MN). The original Mobile IP is defined for IPv4, and therefore it is also referred to as Mobile IPv4.

The Mobile IP protocol allows the MN to retain its IP address regardless of the point of attachment to the network. IP addresses are primarily used to identify the end system. Popular transport protocols, such as TCP, keep track of their session by using end IP addresses of the two endpoints (with appropriate port numbers). Also, routers use IP addresses to route the traffic from the source to the destination. The route does not have to be the same in both directions (for bidirectional communication). Routing in the Internet is based on a packet's destination address and some congestion information in the network nodes. A mobile terminal needs a stable IP address to be identifiable to other Internet hosts and nodes. Therefore, Mobile IP provides two IP addresses for the MN: a home address and a *care-of address* (CoA). The home address is a static IP address that is used to identify higher layer connections (e.g., TCP). The care-of address is used for routing purposes. While the mobile is roaming among different networks, the care-of address changes. In this way, the care-of address represents the IP address of the mobile terminal attachment to the network. In Mobile IPv4 management of CoA is performed by the FA in the visiting network for the mobile terminal. However, the CoA is registered by the HA.

Internet hosts, which communicate with an MN, do not need to know a terminal's location. The MN, using its home address, is able to receive data on

its home network through the HA. When the MN roams in a new network (or domain), it needs to obtain new CoA via the FA in that network. The new CoA will be registered in the HA. Thus, a packet addressed to the MN first reaches the HA, which then tunnels the packets to the FA by using the CoA as the destination address of the packets. At the end of the tunnel, FA decapsulates the packets, such that packets will appear to have the mobile's home address as the destination IP address. After decapsulation, the packets are sent to the MN. Because packets arrive at the MN with their home address as a destination address, the Mobile IP is transparent to higher layer protocols.

Packets sent by the MN are routed by using standard IP routing mechanisms. In this case, MN uses its unique home address as a source address in the IP header (CoA is a temporary address that is used for tunneling from HA to FA when the mobile is roaming in a foreign network). The routing of packets according to the Mobile IP protocol forms a triangle routing among the HA, FA, and the *correspondent node* (CN), as shown in Figure 3.8.

Open Issues in Mobile IP

Mobile IP supports global mobility (i.e., when mobile terminals are roaming among different networks). However, there are open issues in Mobile IPv4.

One of them is related to macromobility management. That is the triangle routing and inefficient direct routing (considering the number of hops). Also, handover procedure is inefficient since HA should be notified during each inter-domain handover. Furthermore, Mobile IPv4 has inefficient binding deregistration (i.e., when an MN moves to a new FA, the previous FA does not release resources immediately, but it waits until a binding registration lifetime expires).

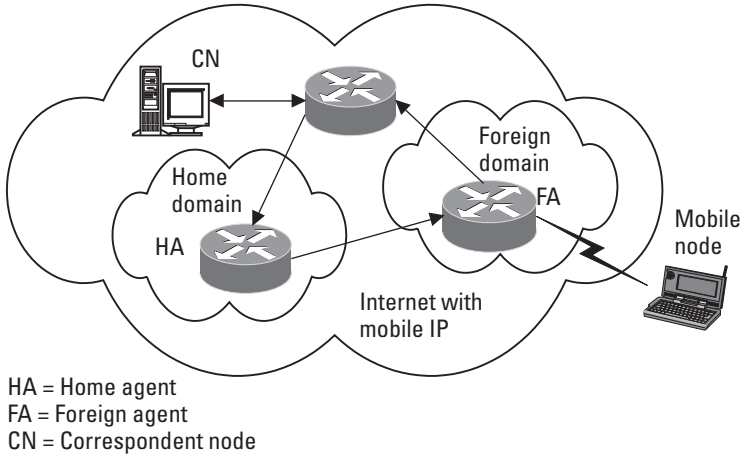


Figure 3.8 Mobile IP protocol.

Mobile IP does not provide solutions for micro-mobility management procedures. Intradomain handovers should be kept as local as possible. Also, after the intradomain handover, the IP data stored in the previous base station should be transferred to the new one. The router crossings should be avoided as much as possible.

Furthermore, Mobile IP does not provide capabilities for QoS provisioning. On the other hand, we expect IP-based mobile networks to provide QoS guarantees for some real-time services (e.g., IP telephony). Due to the heterogeneity of the traffic and QoS demands, Mobile IP should incorporate mechanisms for QoS support (e.g., RSVP).

Another important issue for Mobile IP is security. Standard security measures include authentication (determines the originator of the IP packet), authorization (determines who may access the network and the resources), and encryption of the data. Mobile IPv4 does not provide reliable authentication. Additional security features include ingress filtering (ingress nodes of an ISP filter the packets based on the source address), and location privacy (a sender should be able to control which receivers, if any, may know the sender's location of physical attachment to the network). Firewall protected private Internet networks may cause problems to Mobile IP connections by rejecting IP packets. This may be avoided by ingress filtering (i.e., disallowing datagram entry from any leaf domain).

The introduction of Mobile IPv6 solves some of these open issues. The basic idea of Mobile IP remains the same in IPv6: The MN is reachable by sending packets to its home network, and the HA sends the packets to the mobile's current care-of address by using encapsulation. IPv6 comes with its address configuration protocols: neighbor discovery and stateless address autoconfiguration [26]. By using these configuration protocols, the MN has a greatly enhanced capability to obtain a CoA, thus reducing the need for FAs, which have been eliminated from Mobile IPv6. Also, destination options defined in IPv6 headers simplify binding updates overhead, because now binding updates may be included in any normal data packet. Considering the security, IPv6 offers enhanced authentication. In IPv6, the MN is the only node that can send binding updates to its correspondence nodes, and usually it sends the updates after moving to a new point of attachment to the network. Even after the introduction of Mobile IPv6, however, the micromobility issue will still remain open. It should be dealt with by applying additional local mechanisms, as discussed in the following section.

3.5.2 Micromobility

The Mobile IP protocol solves the macromobility issue (interdomain mobility). In a case of frequent handovers, however, the Mobile IP mechanism introduces

significant network overhead in terms of increased delay, packet loss, and signaling. For example, many real-time services (e.g., IP telephony) would experience noticeable degradation of the quality of service with frequent handovers. Therefore, a number of IP micro-mobility protocols [27] have been proposed that complement the base Mobile IP protocol. Micromobility is directly connected to handovers between cells that belong to a same domain or subnetwork. Also, QoS support in Mobile IP networks is closely related to successful handover management.

One solution for the micromobility problem is given in the recently proposed Cellular IP protocol, which provides mobility and handover support for frequently moving hosts [28, 29]. However, there are several other protocols for micromobility support in wireless IP networks (we refer to them later in this chapter). We choose Cellular IP as the most appropriate example because it considers almost all location and mobility management issues. Other protocols with similar functionalities might be created in the future.

3.5.2.1 Cellular IP

Cellular IP is defined as an extension to the Mobile IP protocol. It is intended for application on a local level (i.e., in the cellular access network). Cellular IP can interwork with Mobile IP to support wide-area mobility—that is, mobility between Cellular IP networks. A typical Cellular IP network architecture is shown in Figure 3.9.

Cellular IP optimizes the cellular network for fast handovers. This protocol provides integrated mobility control and location management functions at the wireless access points.

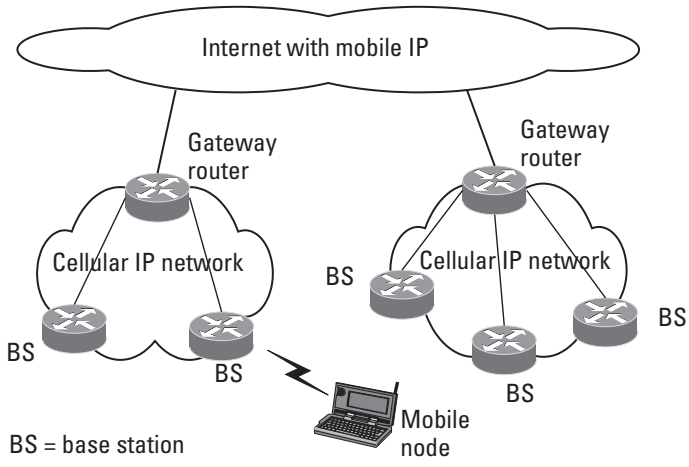


Figure 3.9 Cellular IP network architecture.

Cellular IP Network Architecture

Cellular IP networks are connected to the Internet via gateway routers. Mobile terminals are identified to the network by using the IP address of the base station (access router) as a CoA. Because Cellular IP assumes that Mobile IP manages macromobility, the home agent tunnels the IP packets to the gateway router of the Cellular IP network. Within the network domain, packets are routed upon the home address of the mobile terminal. In the reverse direction, packets from mobile terminal are routed to the gateway router hop-by-hop. After reaching the gateway router, packets are routed through the Internet by Mobile IP.

Routing

In a Cellular IP network, the gateway router periodically sends a beacon packet to the base stations in the wireless access network [30]. Base stations record the interface through which they last received this beacon and use it to route packets toward the gateway. Furthermore, base stations forward the beacon to mobile terminals. Each base station maintains a routing-cache. Packets that are transmitted by mobile nodes are routed to the gateway using standard hop-by-hop routing. Each node in the Cellular IP network that lies in the path of these packets should use them to create and update routing-cache mappings. This way, routing-cache chain mappings are created, which can then be used to route the packets addressed to the mobile node along the reverse path. As long as the mobile node is regularly sending data packets, nodes along the path between the mobile node's actual location and the gateway maintain valid routing entries. Information in the routing-cache, which includes the IP address of the mobile and the interface from which the packets arrive, disappears after a certain time, called route-timeout. Every consecutive packet refreshes the routing information stored at the network nodes. Also, a mobile terminal may prevent a timeout from occurring by sending route-update packets at regular intervals, called route-update time. These are empty data packets. They do not leave the Cellular IP networks (i.e., they are discarded at the gateways).

Location Management

Cellular IP uses two caches at each node in the access network. One is the routing-cache (already discussed above). The other one is a paging-cache, which is optionally implemented at the base stations. While routing-cache is primarily used to keep routing information for the ongoing connections, the paging-cache is primarily used for idle users. Cellular IP defines an idle mobile host as one that has not received data packets for a system-specific time, called active-state timeout. Mobile nodes that are not regularly transmitting or receiving data (i.e., idle nodes) periodically transmit paging-update packets to maintain the paging caches, which may be used to route IP packets (when routing-cache mapping for

that node is expired). Paging-update packets are empty packets addressed to the gateway and are distinguished from a route-update packet by their IP type parameter. These updates are sent to the base station that offers the best signal quality. Similar to data and route-update packets, paging-update packets are routed on a hop-by-hop basis to the gateway. So, maintaining the paging-caches is accomplished similarly to the routing-caches, except for two differences. First, any packet sent by the mobile updates paging-cache mappings, while paging-update packets do not update routing-cache mappings. Second, paging-caches have a longer timeout than routing-caches. Therefore, idle mobile hosts have mappings in paging-caches but not in routing-caches. In addition, active mobile hosts will have mappings in both types of cache. All update-packets are discarded by the gateway, to isolate Cellular IP–specific operations from the Internet. After the paging-timeout, paging mappings are cleared from the cache (e.g., when mobile terminal is turned off).

Mappings always exist in the paging-cache when the mobile node is attached to the network. If routing-cache mappings do not exist, incoming packets may be routed by the paging-cache. However, paging-caches are not necessarily maintained in all nodes.

Handovers

In Cellular IP networks the mobile node initiates a handover [31]. Mobile hosts listen to beacons transmitted by base stations and initiate handover based on signal strength measurements. To perform a handover, a mobile node has to tune its radio to the new base station and transmit a route-update packet. These update packets create routing-cache mappings and thus configure the downlink route from the gateway to the new base station. During the handover the mobile node redirects its data packets from the old to the new base station. At the handover, for a time equal to the routing-cache timeout, packets addressed to the mobile node will be delivered to both the old and new base stations. If the wireless access technology allows listening to two different logical channels simultaneously, then the handover is soft. If the mobile node can listen to only one base station at a time, then the handover is hard (in this case performances of the handover will be more dependent on the radio interface). The routing-cache mappings will be automatically cleared at the moment timeout elapses.

Two parameters define the handover performances: handover delay (i.e., latency) and packet loss. Handover delay is decomposed into rendezvous and protocol time [30]. Rendezvous time refers to the time needed for a mobile node to attach to a new base station after it leaves the old base station. This time is closely related to wireless link characteristics (i.e., the rate of beacons transmitted by the base stations). Protocol time refers to the time spent to restore the connection once the mobile host has received a beacon from the new base station. Usually, rendezvous time is small and we may approximate handover delay

with protocol time. The second parameter is packet loss during the handover. Let us explain how losses occur. Packets are routed through the old base station until the arrival of the first packets through the new route. For hard handover, during this time some packets may be lost. These losses are proportional to the *handover loop time* [30], which is defined as the transmission time from the crossover node to the old location of the mobile node plus the transmission time from the new location to the crossover node, which is the gateway in the worst case. The traffic flow at handover in a Cellular IP network is shown in Figure 3.10.

Although IP packets may be lost at handover, Cellular IP has lower handover delay than Mobile IP. This is due to the local management of the handover (i.e., only local network nodes should be notified at the intradomain handover). There is no need for communication with the home agent that may be located far away from the mobile node's current network.

To reduce packet losses during the handover in a Cellular IP network, a possible solution is semi-soft handover [31]. In this case, the routing-cache mappings are created before the actual handover takes place. So, before the handover to a new base station, the mobile node sends a semi-soft packet to the new base station and immediately returns to listen to the old base station. The idea with semi-soft packets is to establish the new route between the gateway and new base station before the handover execution. During this time the mobile node is still connected to the old base station. After a time period called semi-soft delay (e.g., 100 ms), the mobile node performs a regular handover. The semi-soft approach, however, does not ensure a smooth handover. In reality, the transmit time from

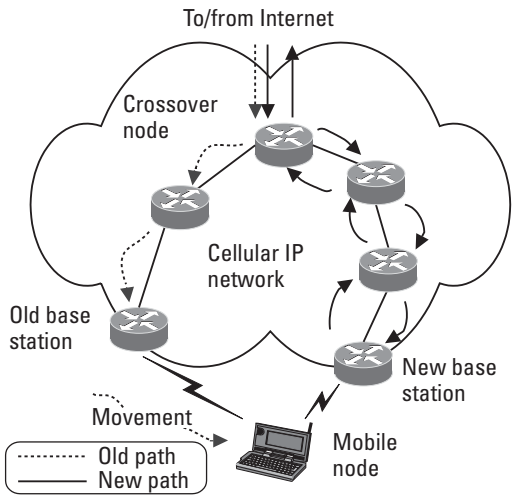


Figure 3.10 Handover in a Cellular IP network.

the crossover node to the old and new base station may differ. If the new base station is "behind" the old one, the mobile node will receive duplicate packets (they usually do not disrupt traffic flows, such as TCP). In the opposite case, when the new base station is "ahead," then some packets may be lost. Also, semi-soft handover adds additional delay.

Open Issues in Cellular IP

Cellular IP is a protocol and concept that integrates location management functions and fast handovers, which are usually found in today's mobile systems, with typical Internet routing and addressing mechanisms. Cellular IP solves micromobility, while Mobile IP handles the macromobility. However, there are several open issues.

First, the handover mechanism assures local management of intradomain handovers (i.e., micromobility), but it is not persistent to packet losses. Losses disrupt typical Internet traffic, such as TCP flows. Semi-soft handover reduces the losses, but still it does not guarantee zero loss.

Second, Cellular IP does not provide mechanisms for QoS support, which is very important for some applications (e.g., real-time services). The protocol is basically proposed for the best-effort service, which is the dominant type of traffic in the Internet today. To be able to support multiple traffic classes with different QoS demands, we should integrate Cellular IP with some of the QoS mechanisms.

3.5.2.2 Handover Mechanisms for Cellular Wireless Packet Networks

Besides Cellular IP, there are several other proposed solutions to micromobility as an extension to the Mobile IP. We refer to some of them, such as the multicast-based Mobile IPv4 algorithm [32] and IP micromobility support using *Handover-Aware Wireless Access Internet Infrastructure* (HAWAII) [33]. There are other micromobility proposals, such as vertical handoffs in wireless overlay networks [34], hierarchical foreign agents [35], as well as recent Internet drafts: fast handovers for Mobile IPv6 [36] and low latency handovers in Mobile IPv4 [37]. A handover mechanism for wireless IP networks is proposed in Chapter 10.

Multicast-Based Intra-Handover Algorithms

The multicast-based intra-handover algorithm has been implemented and tested in the Daedalus project at Berkeley [32]. The algorithm is created for Mobile IPv4, but after some minor modifications it can also be used for Mobile IPv6. This algorithm is active in the time period when the home agent forwards the packet to the mobile node's CoA.

In this algorithm, the MN is also assigned a second address, which is a temporary multicast address. When the home agent receives a packet addressed

to the mobile node, it forwards the packet to the associated multicast group, which consists of the serving base station and some neighboring base stations, which are determined according to the signal strength of the recent received beacons by the MN as well as communication quality. The serving base station is called the primary one. At a given instant of time there is only one primary base station. Other base stations from the multicast group (which are identified as potential targets for a handover by the MN) do not forward the packets to the wireless access network, but they buffer the last few packets transmitted from the HA. After a handover, the MN is supposed to send control messages to all base stations within the multicast group as a request to begin or end forwarding or buffering of packets. In the reverse direction, packets sent by the correspondent node are directly routed via the new base station, without having them forwarded from the previous base station. This algorithm has minimal packet losses and has negligible delays. Therefore, it is seamless to the data flow. On the other hand, it requires extra buffer space at the base stations and additional signaling traffic.

Another protocol, the Hierarchical Mobile IP protocol [38], uses multicast of the IP packets in the downstream using the *gateway foreign agent* (GFA), which is a network entity that manages certain regions with several FAs. Hierarchical Mobile IP offers the possibility for the host to perform local registrations in the visited wireless network. In this case, the GFA will have the binding of the mobile host's CoA and the host's home address. So, considering the handover, the only difference between this scheme and Daedalus multicast is the entity that performs multicasting in the downstream, which is GFA in the Hierarchical Mobile IP protocol.

Another multicast-based mobility approach is the Intra-Domain Mobility Management Protocol [39], which provides fast handovers by using a hierarchical structure with a mobility agent on top of it with several subnetwork foreign agents interconnected to it. The top-level mobility agent in the hierarchy is the gateway to the Internet.

HAWAII

In this approach, host-based forwarding entries are installed in specific routers to support intra-domain mobility [33, 40]. The HAWAII-capable network is divided into hierarchies called domains. Each domain has a root router. Also, each mobile node has a home domain. When an MN is in the home domain, it retains the IP address. The packets that are addressed to the MN can reach the domain root router based on the subnetwork address of the domain. From the root router, packets are routed to the MN by using specially established paths. When an MN roams in a foreign network, Mobile IP mechanisms are used to handle macromobility. In the foreign network an MN gets a COA from the foreign domain. Within the foreign domain the MN retains its COA while it is

moving. Due to this fact, notifications to the HA are significantly reduced. HAWAII supports different path setup schemes [i.e., forwarding scheme: MN can transmit/receive to/from one base station at a time (e.g., TDMA wireless technology); and non forwarding scheme: MN is able to transmit/receive data to/from several base stations at a time (e.g., CDMA wireless technology)].

3.6 QoS Specifics of Wireless Networks

Wireless networks differ from wired networks in terms of access technology and in the characteristics of the transmission medium. In this section we point to some important characteristics of the wireless medium that have influence on the communication quality.

3.6.1 Cellular Topology

One of the main problems for wireless networks is limited frequency spectrum. Therefore, the number of simultaneous connections over a particular geographical area is bounded by the capacity of the specific wireless access system. On the other hand, the capacity of wired (fixed) networks is not an issue, because if we need capacity we may invest into additional infrastructure (e.g., by adding more twisted pairs or fiber).

In order to allow a greater number of users for a specific wireless technology, we need to use a cellular principle, as shown in Figure 3.11. Thus, a wireless network consists of wireless access points called base stations, where each base station covers particular geographical area. Due to fading (the power of the radio waves decreases with the distance), we may reuse the same frequencies by using appropriate frequency planning. For better frequency reuse, we group the available frequency carriers or bands into groups. The number of cells within a group defines the reuse factor. For example, in the TDMA-based GSM system we have different frequency reuse patterns, such as 3/9, 4/12, and 7/21. The notation x/y has the following meaning: all available frequency carriers (or bands) are divided into groups of y frequencies each, which are distributed in x different cells. Then, the pattern is repeated through the network. Some 3G systems, such as WCDMA, do not require frequency planning (i.e., they have reuse factor equal to 1 which is discussed in Chapter 2).

In a dense area (with a large number of mobile users) we must use smaller cells due to the frequency reuse and capacity requirements.

3.6.2 Mobility

User mobility and cellular topology are the reasons that handovers are necessary. Also, a mobile node frequently changes its location within a single cell, thus

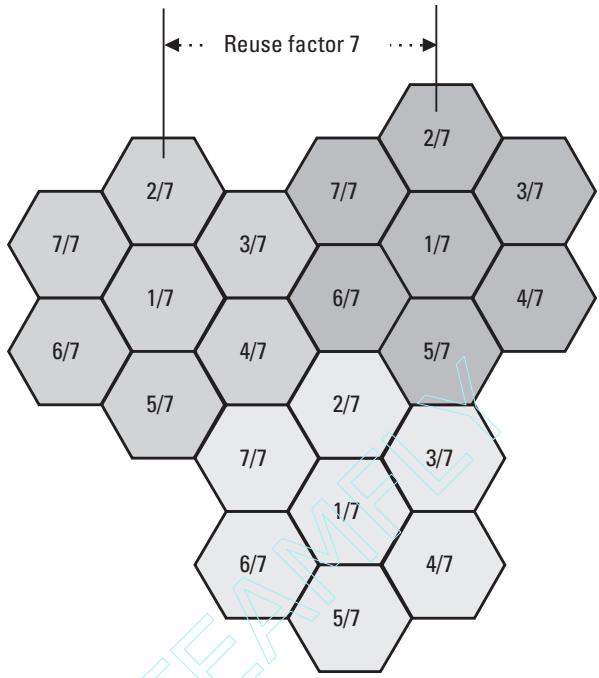


Figure 3.11 Cellular concept of a mobile network.

resulting in time-varying bit error ratio and interference, which directly define the QoS for that connection.

Handover schemes have so-called handover latency. This is a time period during which the mobile node is unable to send or receive IP packets. In certain scenarios, the handover latency resulting from Mobile IP handover procedures may be greater than what is acceptable for real-time services [36].

Also, handovers may cause packet losses. Such losses may disrupt both real-time and nonreal-time services, and hence are undesirable.

User mobility introduces one additional problem: location control. It is necessary to track the users within the network. However, storing the exact location of each user in the mobile network (e.g., the current cell) reduces the mobile node's battery recharge time (due to increased demands for location-updating), inefficiently utilizes the scarce wireless resources (due to signaling messages), and increases the overall cost of the system. Therefore, the existing cellular systems define two main user states: (1) busy users, which have an ongoing connection (e.g., allocated time slot), and (2) idle users, which are attached to the network (i.e., mobile terminal is turned "on") but are not active at a particular time. Thus, the network keeps track of the mobile's current cell while it is in the busy state. When a mobile is idle, the network stores the information of

the mobile's current location area, which usually includes several tens of cells. Some mobile systems introduce intermediate states between the two main mobile's states. An example is the standby state, which is defined in GPRS and UMTS systems (refer to Chapter 2). When the network receives a call or a packet addressed to a mobile node that is in idle state, it performs paging through all the base stations in the mobile's current location area. If the mobile node is in some intermediate state, then network does paging in a particular area that is defined for that state (e.g., in GPRS, if the mobile is in standby state, the network performs paging through all base stations in the mobile's current routing area). When the mobile node replies to the paging message, the network determines its current serving cell and establishes a communication link.

3.6.3 BER in the Wireless Link

Bit errors in the wireless interface may occur as a result of several different causes. According to [41], bit errors in wireless links are caused by interference, noise, fading, and shadowing.

Fading is one of the main characteristics of a signal's propagation over wireless links. From the aspect of noise and shadowing, fading is not desirable. But, considering the interference and frequency reuse concept, the fading is useful. It bounds the coverage of a single wireless network access point (e.g., a base station) over a limited geographical area. Thus, the fading allows the cellular concept in the wireless networks. One general formulation of the fading is given with the following relation [41]:

$$L = \frac{P_R}{P_T} = c\frac{1}{f^2 d^k} \tag{3.3}$$

where P_R is received power at the receiver (a base station or a mobile node), P_T is transmitted power at the transmitter, f is frequency, d is distance between the transmitter and receiver, and c is a constant. Factor k depends upon the characteristics of the wireless medium. For a free-space propagation of radio waves, a typical value is $k = 2$. But empirical studies have shown that the value of the factor k in a cellular mobile environment is typically between 3 and 5, due to the characteristics of the wireless link, such as shadowing.

Shadowing is a consequence of obstacles on the path of the radio waves (e.g., there is no line of sight between the mobile node and base station). Furthermore, due to the reflection of the signal from surrounding objects (e.g., buildings, houses, and so forth) different parts of the same signal may reach the receiver via different paths. This effect is called multipath. It is not desirable in systems like GSM, but it is helpful in systems such as WCDMA.

Interference is a consequence of the reuse of the same or adjacent frequency bands in the same or neighboring cells. The design of radio networks in 2G mobile systems, based on FDMA/TDMA technology, tends to minimize cochannel and adjacent channel interference. The 3G technologies, such as WCDMA, are robust to the interference due to the spreading of the narrowband signal over wide frequency spectrum.

These characteristics of the wireless medium, as discussed above, cause higher bit error ratio in wireless links than in wired ones. Also, BER is dependent upon the location of the mobile node. Furthermore, bit errors occur in bursts due to the inertia of the mobile's movement and actual state of the wireless link at a specific location in the cell. We need to design QoS mechanisms to deal with the location-dependent and time-variable bit errors in the wireless links.

3.7 Discussion

The growth of the Internet is similar to that of cellular mobile networks. The idea for Mobile Internet is already widely accepted by the Internet service providers and cellular operators. In order to design a cost-effective wireless IP network, however, we need to create many small network domains that should be interconnected as well as connected to the commercial cellular networks (e.g., GSM).

IETF has defined the Mobile IP protocol, which is the de facto standard for macro-mobility in the Internet. Also, there is a significant research effort towards QoS support and efficient micromobility management in mobile IP networks.

So far, we have several proposals on mechanisms for QoS provisioning in the Internet, such as: MPLS, Integrated Services, and Differentiated Services. These QoS mechanisms were initially created for wired IP networks. After minor modifications, however, we may also apply them in wireless access networks. The Differentiated Services scheme is foreseen as the most suitable QoS mechanism for the wireless access networks because of its limited processing and space requirements at the network nodes.

The introduction of mobility to the Internet requires the creation of mechanisms to deal with handovers, location management (i.e., tracking the users within the network), and location-dependent bit errors (they are much higher in wireless than in wired networks). Over the past several years a number of IP micromobility protocols have been proposed, such as Cellular IP, HAWAII, and multicast-based intra-handover management. Most of them are created for Mobile IPv4, but they may be applied in Mobile IPv6 networks. Generally, IPv6 offers some improvements over IPv4 considering the mobility management, QoS, and security issues.

We may conclude that second and third generation mobile networks offer seamless mobility and QoS support, but are built on complex and costly connection-oriented networking. On the other hand, IP offers robustness, scalability, and flexibility as well as transparency to different services. Their convergence is the way towards future wireless mobile networks.

References

[1] Postel, J., (ed.), *Internet Protocol*, RFC 791, September 1981.

[2] Deering, S., and R. Hinden, *Internet Protocol, Version 6 (IPv6) Specification*, RFC 2460, December 1998.

[3] Postel, J., (ed.), *Transmission Control Protocol*, RFC 793, September 1981.

[4] Fall, K., and S. Floyd, "Simulation-Based Comparisons of Tahoe, Reno, and SACK TCP," *Computer Communication Review*, Vol. 26 No. 3, July 1996, pp. 5–21.

[5] Jacobson, V., "Congestion Avoidance and Control," *ACM SIGCOMM'88*, August 1988.

[6] Stevens, W., *TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms*, RFC 2001, January 1997.

[7] Mathis, M., S. Floyd, and A. Romanow, *TCP Selective Acknowledgement Option*, RFC 2018, October 1996.

[8] Mankin, A., et al., *IETF Criteria for Evaluating Reliable Multicast Transport and Application Protocols*, RFC 2357, June 1998.

[9] Stewart, R., et al., *Stream Control Transmission Protocol*, RFC 2960, October 2000.

[10] Ong, L., et al., *Framework Architecture for Signaling Transport*, RFC 2719, October 1999.

[11] Xiao, X., and L. M. Ni, "Internet QoS: A Big Picture," *IEEE Network Magazine*, March/April 1999.

[12] Xiao, X, et al., "Traffic Engineering with MPLS in the Internet," http://web.cps.msu.edu/~xiaoxipe/researchLink.html.

[13] Zee, M., and G. Heijenk, *Quality of Service over Specific Link Layers*, Open Report, Ericsson, 09-07-1999.

[14] Awduche, D., et al., *RSVP-TE: Extensions to RSVP for LSP Tunnels*, RFC 3209, December 2001.

[15] Ash, J., et al., *LSP Modification Using CR-LDP*, RFC 3214, January 2002.

[16] Brade, B., D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: an Overview*, RFC 1633, June 1994.

[17] Shenker, S., C. Partridge, and R. Guerin, *Specification of Guaranteed Quality of Service*, RFC 2212, September 1997.

[18] Wroclawski, J., *Specification of Controlled-Load Network Element Service*, RFC 2211, September 1997.

[19] Zee, M., and G. Heijenk, *Quality of Service Routing – State of the Art Report*, Open Report, Ericsson, 08-07-1999.

[20] Blake, S., et al., *An Architecture for Differentiated Services*, RFC 2475, December 1998.

[21] Nichols, K., et al., *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*, RFC 2474, December 1998.

[22] Jacobson, V., and K. Poduri, *An Expedited Forwarding PHB*, RFC 2598, June 1999.

[23] Heinanen, J., et al., *Assured Forwarding PHB Group*, RFC 2597, June 1999.

[24] Floyd, S., and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, August 1993.

[25] Perkins, C., (ed.), *IP Mobility Support*, RFC 2002, October 1996.

[26] Perkins, C., "Mobile IP," *IEEE Communications Magazine*, 50th Anniversary Issue, May 2002.

[27] Campbell, A. T., et al., "Comparison of IP Micro-Mobility Protocols," *IEEE Wireless Communications*, February 2002.

[28] Valko, A. G., "Cellular IP: A New Approach to Internet Host Mobility," *ACM Computer Communication Review*, January 1999.

[29] Kim, S., et al., "A Cellular IP Demonstrator," *Sixth IEEE International Workshop on Mobile Multimedia Communications (MOMUC'99)*, San Diego, CA, November 1999.

[30] Valko, A. G., et al., "On the Analysis of Cellular IP Access Networks," *IFIP Sixth International Workshop on Protocols for High Speed Networks (PfHSN'99)*, Salem, MA, August 1999.

[31] Campbell, A. T., et al., "Design, Implementation and Evaluation of Cellular IP," *IEEE Personal Communications*, Special Issue on IP-based Mobile Telecommunications Networks, June/July 2000.

[32] Seshan, S., H. Balakrishnan, and R. H. Katz, "Handoffs in Cellular Wireless Networks: The Deadalus Implementation and Experience," *Kluwer International Journal on Wireless Communications Systems*, 1996.

[33] Karagiannis, G., and G. Heijenk, *Mobile IP – State of the Art Report*, Open Report, Ericsson, 13-07-1999.

[34] Stemm, M., and R. H. Katz, "Vertical Handoffs in Wireless Overlay Networks," *ACM Mobile Networking (MONET)*, Special Issue on Mobile Networking in the Internet, 1998.

[35] Caceres, R., and V. Padmanabhan, "Fast and Scalable Handoffs for Wireless Networks," *ACM Mobicom*, 1996.

[36] Dometry, G., (ed.), *Fast Handovers for Mobile IPv6*, Internet Draft, March 2002.

[37] El Malki, K., (ed.), *Low Latency Handoffs in Mobile IPv4*, Internet Draft, June 2002.

[38] Gustafsson, E., A. Jonsson, and C. Perkins, *Mobile IP Regional Registration*, Internet Draft, draft-ietf-mobileip-reg-tunnel-03, work in progress, July 2000.

[39] Misra, A., et al., "IDMP-Based Fast Handoffs and Paging in IP-Based 4G Mobile Networks," *IEEE Communications Magazine*, Vol. 40, No. 3, March 2002, pp. 138–145.

[40] Ramjee, R., et al., "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks," *IEEE ICNP*, 1999.

[41] Rappaport, T. S., *Wireless Communications: Principles and Practice*, Englewood Cliffs, NJ: Prentice Hall, 1996.

4

Teletraffic Theory

4.1 Introduction

For the purpose of design and analysis of telecommunications networks, we need to have a well-defined traffic theory. The theory strongly depends on the type of traffic in the network. So, we need first to classify telecommunications networks and then to establish appropriate theory for various traffic types. Usually, telecommunications networks are divided into two main groups: circuit-switched and packet networks. Furthermore, one can classify telecommunications networks based on traffic homogeneity into homogeneous and heterogeneous networks. The homogeneous type is used to describe classical telecommunications service based on voice transmission and switching. The heterogeneous type includes integrated traffic streams from different sources (voice, audio, video, data) into a single network. Using these classifications, globally we can find four types of telecommunications networks:

- Circuit-switched networks with homogeneous traffic;
- Circuit-switched networks with heterogeneous traffic;
- Packet networks with homogeneous traffic;
- Packet networks with heterogeneous traffic.

Based on the type of access network, we categorize telecommunications networks into wired (fixed) access networks and wireless (mobile) access networks. We can combine all four types of networks, defined above, with both access types, wired and wireless.

In this chapter, we assume that the reader is familiar with basic probability and statistics theory. However, in the next section we consider some important random processes for traffic theory. Going through Markov chains and birth-death processes, we cover the traffic theory for voice as well as for networks with integrated traffic. We further focus our interest in wireless networks with heterogeneous traffic.

4.2 Some Important Random Processes

For the analysis of many physical events we need to use variables, which have random values during the observation of some time interval. A random (probabilistic) event or process is a general mathematical form for a description of such behavior. The theory concerned with random events is called probability theory. Thus, a random event lacks prediction of its result. On the other hand, many random events show some behavior. One simple example of a random event is throwing of a coin. There are two possible results of this event, a head or a tail. We can define a random variable ξ_n, which will describe the behavior of the random event. We can set $\xi_n = 0$ when nth throw is heads and $\xi_n = 1$ when nth throw is tails. In this example we expect to have half of the results heads and half tails. When the number of events is larger, then the appearance of each result will be closer to one-half. This means that random events have regularity that manifests itself in a collection of random events. If we can make very precise statement about large collections of random events, then we have statistical regularity. In this book, we are interested in random events with statistical regularity.

By definition, a random process is a group of random variables $\xi_t(S)$ where t denotes the family of events and S denotes the sample space. Each event from family t is a set of sample points as shown in Figure 4.1. We can also use $\xi(t, S) \equiv \xi_t(S)$ as a notation for a random process. A family of events is a collection of mutually exclusive exhaustive outcomes. To get a random variable we observe the values of all events from a family in a moment of time. All possible outcomes form the sample space S. In probability theory [1] we map all possible outcomes into real numbers called probabilities P of the outcomes. At a given moment t_0, $\xi(t_0, x)$ is a random variable. This mapping must satisfy some properties:

- All outcomes are mutually exclusive.
- Probability P of one or more outcomes of a single random process is always in the interval $0 \leq P \leq 1$.
- $P\{S\} = 1$, meaning the sum of probabilities of all possible outcomes from a random process is equal to one.

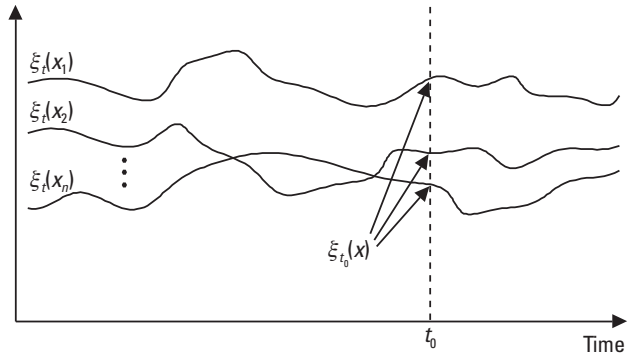


Figure 4.1 Definition of a random process.

We can categorize random variables into two groups [1]: discrete and continuous. If the range of fluctuations of the random variable is a discrete set of values, the variable is a discrete random variable. When the range of variation of a random variable is continuous, the variable is called continuous random variable. For example, in traffic analysis in telecommunications we consider time as a random event, such as waiting time, busy time, and interarrival time. Time is continuous by its nature, so this kind of random variable should be continuous. On the other hand, if we observe a number of busy channels or number of events in given time interval (e.g., call arrivals, call servicing) as random variables, then we have a discrete set of values that we can assign to that variable.

We can describe a random variable by its statistical properties [1]: distribution function, mean value (first moment), and variance (second central moment). For completeness we consider time as a random variable ξ and we define its statistical properties. Here we introduce a probability distribution function $F_\xi(t)$:

$$F_\xi(t) = P\{\xi \le t\} = \int_{0-}^{t} dF_\xi \text{ for } 0 \le t < \infty$$
$$F_\xi(t) = 0 \text{ for } t < 0 \tag{4.1}$$

If the distribution function is differentiable, then probability density function $p_\xi(t)$ exists and it is defined by:

$$dF_\xi(t) = p_\xi(t)dt = P\{t < \xi \le t + dt\}, t \ge 0 \tag{4.2}$$

which equals the probability of ξ taking a possible value x as a function of its complete set of possible values at a given parameter t. Every distribution is characterized by its moments. The mean value (expectation) is the first moment:

$$m = m_1 = E\{\xi\} = \overline{\xi} = \int_0^{\infty} t p_{\xi}(t)dt \tag{4.3}$$

Generally, the ith central moment is defined as

$$E\left\{(\xi - m)^i\right\} = \int_0^{+\infty} (t - m)^i p_{\xi}(t)dt \tag{4.4}$$

The variance is identical to the second central moment:

$$\sigma^2 = E\left\{(\xi - m)^2\right\} = \overline{\xi^2} - \overline{\xi}^2 = m_2 - m_1^2 \tag{4.5}$$

Analytically, we can carry out different calculations for different time distributions. We always assume that mean value exists.

Teletraffic theory covers specific types of random processes. Examples of random processes in telecommunications include the following: average connection duration, average number of users in the system, busy time, service time, and call arrival process. All variables considered in this book are nonnegative stochastic variables.

We can classify random processes based upon their characteristics. Basically, the classification of a random process depends upon three quantities [2]:

1. Time (index) parameter: It is possible to have discrete time parameter process or continuous time parameter process.
2. The state space: As in (1), it is possible to have discrete or continuous state space.
3. Statistical dependence among the random variables $\xi_t(x_i)$, $i = 1, 2, 3, \ldots$, for different values of the index parameter t.

The main distinguishing feature of a random process is the statistical dependence among random variables used for that process. We classify random processes upon their last feature. There are several basic types of random processes [2] that we provide in the rest of this section.

Stationary process is invariant to shifts in time for all values of its argument; that is, given any τ, it must hold that

$$\xi_{t+\tau}(\mathbf{X}) = \xi_t(\mathbf{X}) \tag{4.6}$$

where $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and $t = (t_1, t_2, \ldots, t_n)$ are row matrices, while x_i is a member of the family $\mathbf{X}$ of the random process.

Independent process is the most trivial random process, which can be described by set of independent random variables:

$$f_X(t) = f_{x_1, x_2, \ldots, x_n}(t_1, t_2, \ldots, t_n) = f_{x_1}(t_1) f_{x_2}(t_2) \ldots f_{x_n}(t_n) \qquad (4.7)$$

The *Markov process* (A. A. Markov, 1907) consists of a set of random variables $\{x_n\}$, which forms so-called Markov chain where the probability of the next value x_{n+1} (each value of the random variable defines a particular state in the chain) depends only upon the current value (state) x_n and not upon any previous state (or the history of the process):

$$\begin{aligned} &P\left[\xi_n = x_n \middle| \xi_{n-1} = x_{n-1}, \xi_{n-2} = x_{n-2}, \ldots, \xi_1 = x_1\right] \\ &= P\left[\xi_n = x_n \middle| \xi_{n-1} = x_{n-1}\right] \end{aligned} \qquad (4.8)$$

The *Birth-death process* is a special case of Markov chains (Markov process) that is very important to queuing theory [2]. Here, a state transition in the Markov chain takes place between adjacent states only, or in other words, the probability of a transition between not-neighboring states is negligible, either for discrete or continuous process.

The *Semi-Markov process* is a generalization of the Markov process (Figure 4.2). At every unit interval of time the process is required to make a transition from the current state to some other state (there is possibility of going back to the same state in a loop). Upon its definition, we allow an arbitrary distribution of time that a semi-Markov process may stay in a state, either in discrete or continuous domain.

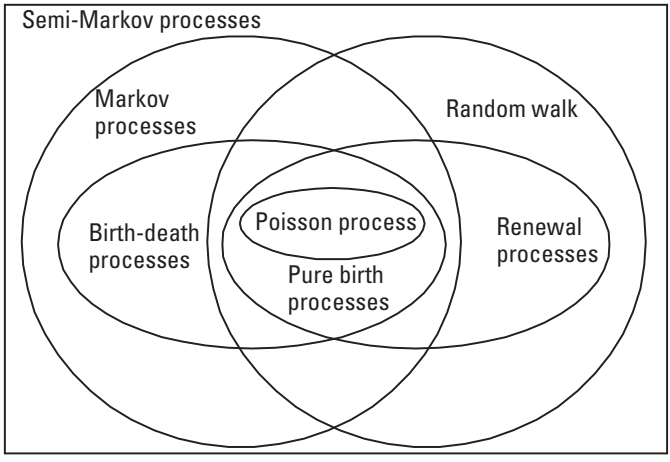


Figure 4.2 Relationship among random processes.

Random walk may be defined as a particle moving among the states in some state space. For this random process it holds that the next position the process occupies is equal to a sum of the previous position of the process and arbitrarily distributed random value, which distribution does not change with the state of the process (except maybe at boundary states).

Renewal process is closely related to the random walk process. Its variables get values from counting transitions that take place as a function of time. The distribution of time between adjacent transitions is an arbitrary common distribution. A typical example of this kind of process is G/G/m queue (for queue notation refer to Section 4.6.4), which has an arbitrary (or G General) arrival process and general distribution of servicing time (the second G in the notation of the queue).

4.3 Discrete Markov Chains

A random process is called Markov when the actual system's state depends only of the previous state of the system and does not depend of the history of the process. One can express this characteristic of the Markov process using the relation (4.8). If we are able to determine the state of the Markov chain in one moment in time, then we can predict its behavior in the future or in the past.

We will restrict our attention to situations where the random variable ξ_n may take only a finite set of possible values. These discrete-time, finite, Markov processes sometimes are called finite Markov chains. Let us restrict our attention to first-order Markov chains where the outcome at any time t_{n+1} depends only upon the outcome at the immediately preceding time t_n. For each pair of random variables (ξ_n, ξ_{n+1}) we can define the transition probability $p_{ij}^{n,n+1}$ as follows:

$$p_{ij}^{n,n+1} = P\left\{\xi_{n+1} = x_j \middle| \xi_n = x_i\right\} \tag{4.9}$$

where x_i and x_j are realizations of the random variables ξ_n and ξ_{n+1}. In traffic theory, we are interested in cases where transition probabilities are stationary (independent of the time at which the transition takes place). For the stationary case we can rewrite last equation as

$$p_{ij} = P\left\{x_j \middle| x_i\right\} \tag{4.10}$$

Sometimes it is convenient to average these probabilities, given by (4.10), into an $N \times N$ transition matrix:

$$\mathrm{T} = \left[p_{ij} \right] = \begin{bmatrix} P_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ . & . & \cdots & . \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix} \tag{4.11}$$

where N is the number of possible realizations of ξ_n. It is trivial to say that considering probability theory [3] each element of the matrix satisfies the conditions

$$p_{ij} \geq 0, \text{for } i, j = 1,2,\ldots,N \tag{4.12}$$

$$\sum_{j=1}^{N} p_{ij} = 1 \text{ for } i = 1,2,\ldots,N \tag{4.13}$$

Let us denote with $p_i^{(n)}$ the unconditional probability of state i occurring at time t_n; that is,

$$P\{\xi_n = x_i\} = p_i^{(n)} \tag{4.14}$$

Also, we can define the state distribution vector $\Pi^{(n)}$ as

$$\Pi^{(n)} = \left[p_1^{(n)}, p_2^{(n)}, \ldots, p_N^{(n)} \right] \tag{4.15}$$

The initial state distribution, at the start of the random process or at least at the start of the observation of the random events, is given by

$$\Pi^{(0)} = \Pi = \left[p_1^{(0)}, p_2^{(0)}, \ldots, p_N^{(0)} \right] \tag{4.16}$$

Let us now define r-step transition probabilities:

$$p_{ij}{}^{(r)} = P\left\{ \xi_{r+m} = x_j \middle| \xi_r = x_i \right\} \tag{4.17}$$

In that case, one-step transition probability is defined by (4.10), or by

$$p_{ij}{}^{(1)} = p_{ij} \tag{4.18}$$

The absolute state probability at time $t = t_1$ is

$$p_j{}^{(1)} = \sum_{i=1}^{N} p_i{}^{(0)} p_{ij} \tag{4.19}$$

or, by using matrix notation:

$$\prod^{(1)} = \prod \cdot \mathrm{T} \tag{4.20}$$

Furthermore, for two-state transition probabilities we obtain

$$p_j^{(2)} = \sum_{k=1}^{N} \sum_{i=1}^{N} p_i^{(1)} p_{ik} p_{kj} \tag{4.21}$$

After rearranging last equation, we get

$$\prod^{(2)} = \prod^{(1)} \mathbf{T} = \prod \mathbf{T}^{(2)} \tag{4.22}$$

If we continue in the same manner, we can generalize by

$$\prod^{(r)} = \prod \mathbf{T}^{r} \tag{4.23}$$

If vector $\prod^{(r)}$ approaches limiting distribution **P**, independent of the initial state, then it has reached equilibrium:

$$\lim_{r \to \infty} \prod^{(r)} = \mathbf{P} \tag{4.24}$$

where **P** is a fixed probability vector $\mathbf{P} = [p_1, p_2, \ldots, p_N]$, and p_i is the stationary probability distribution of state *i*.

A Markov chain that has the property (4.24) is called regular [3]. We may write the following:

$$\lim_{r \to \infty} \prod^{(r)} = \lim_{r \to \infty} \prod^{(r+1)} = \mathbf{P} \tag{4.25}$$

Using (4.25) and (4.23), we get

$$\mathbf{P} = \mathbf{P} \cdot \mathbf{T} \tag{4.26}$$

We now have an eigenvalue problem, the solution of which does not depend upon the initial state vector.

If we allow transitions between the states of the Markov chain to take place at any point in time, then we are talking about a continuous-time Markov chain. When the transitions take place only at fixed points in time, then the Markov chain is a discrete-time chain.

In discrete-time Markov chains, the time that the system spends in the same state is geometrically distributed [2]. We can easily prove this statement. Let us assume that the system has entered a state i. Then, the probability that the system will remain in the same state is p_{ii}. The probability that the system will leave its state at the next step is $(1 - p_{ii})$. Due to the memoryless property of the Markov chains, we may write the following:

$$P\{\text{system remains in state } i \text{ after } m \text{ consecutive steps}\} = (1 - p_{ii})p_{ii}^{m} \tag{4.27}$$

For continuous-time Markov chain, we have exponential distribution of the time in single state (discrete-state continuous-time Markov process, see Figure 4.3), and we may write the following:

$$F(t) = P\{\xi \le t\} = 1 - e^{-\lambda t} \tag{4.28}$$

where λ is a parameter of the exponential distribution. The density function of the exponential distribution (Figure 4.4) is given by

$$f(t) = \lambda e^{-\lambda t} \tag{4.29}$$

The probability that the interarrival time between two consecutive arrivals will be up to t after it was t_0 may be calculated by

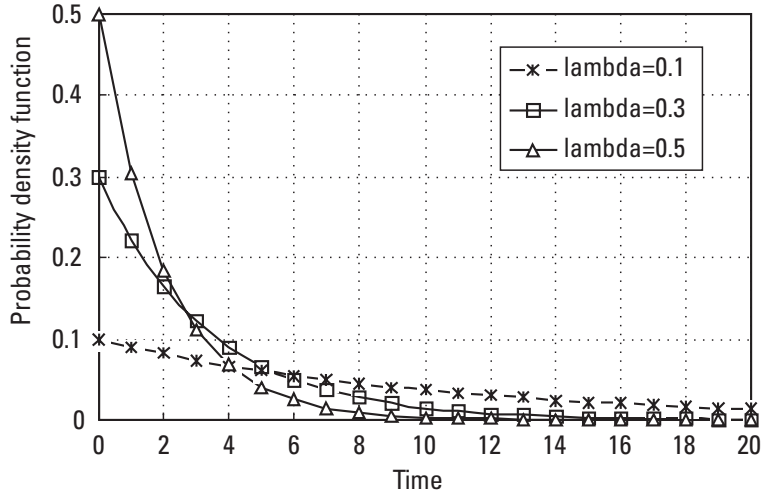


Figure 4.3 Probability density functions of discrete-state continuous-time Markov chain.

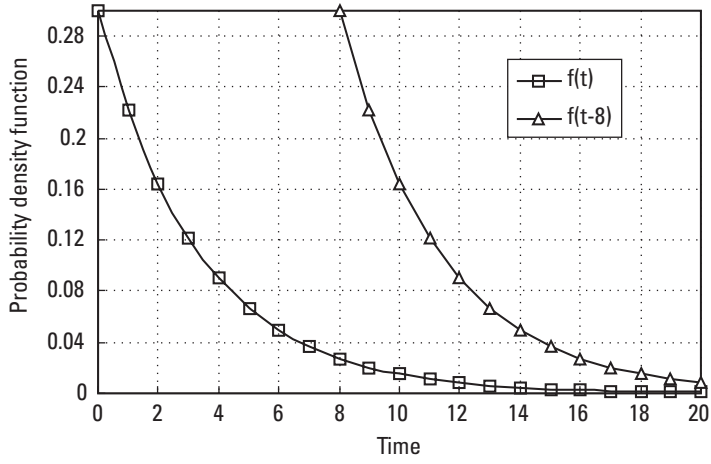


Figure 4.4 Probability density function of the exponential distribution.

$$P\{\xi \le t + t_0 | \xi > t_0\} = \frac{P[t_0 < \xi \le t + t_0]}{P\{\xi > t_0\}}$$

$$= \frac{P\{\xi \le t + t_0\} - P\{\xi \le t_0\}}{1 - P\{\xi \le t_0\}} = \frac{\left(1 - e^{-\lambda(t+t_0)}\right) - \left(1 - e^{-\lambda t_0}\right)}{e^{-\lambda t_0}} = 1 - e^{-\lambda t} \tag{4.30}$$

If there is only one event from $t = 0$ to time $t = t_0$, then the probability for a new event to occur in next time period t (from t_0 to $t + t_0$) does not depend upon t_0.

We will further apply Markov processes in telecommunications because most of the random events can be considered in a Markov chain fashion.

4.4 The Birth-Death Process

The birth-death process is a special case of the Markov processes. Here, the transitions are permitted between adjacent states only. We are mainly interested in continuous-time processes, so we consider birth-death processes in that fashion. The probability that more then one event will occur in an infinitesimal time interval is zero:

$$\lim_{\Delta t \to 0} \frac{P\{> 1\}}{\Delta t} = 0 \tag{4.31}$$

This feature is called *ordinarity*. We usually write

$$\frac{P\{>1\}}{\Delta t} = o(\Delta t) \tag{4.32}$$

We consider only continuous-time birth-death process, which are of primary interest to us in this book. Birth-death processes are often used for the analysis of mass systems in telecommunications and computer networks (which means a large number of users in the system). They are often appropriate for modeling changes in the size of population. In a telecommunications network the population is the number of users in the system. Therefore, we can refer to some state E_k according to the number of users. So, without losing generality, we may denote with E_k the state of the system when the population is of size k. From the state k, birth-death process may transit only in state $k+1$ and state $k-1$, or remain in the state k during time interval Δt. We introduce the notion of birth rate λ_k as well as death rate μ_k in a state k. Due to the memoryless property of the birth-death process, these birth and death rates are independent of time, but depend upon the current state E_k only. Possible transitions in a birth-death process are shown in Figure 4.5.

Because the birth-death process permits transitions among neighboring states only, we can describe it by a state transition diagram shown in Figure 4.6. We will refer to this type of diagram as a one-dimensional Markov chain.

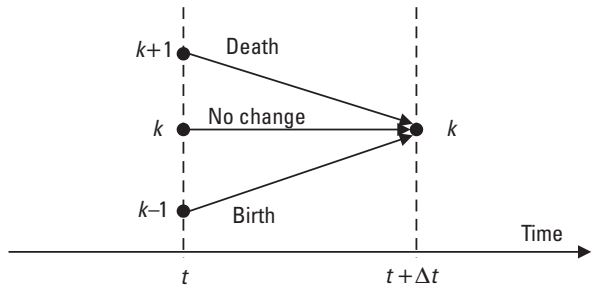


Figure 4.5 State transitions in a birth-death process.

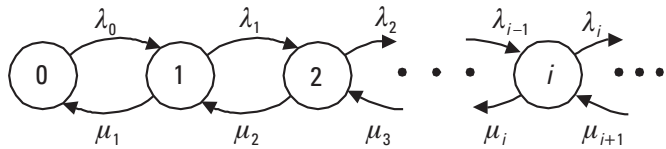


Figure 4.6 One-dimensional Markov chain for a birth-death process and infinite population in the system.

The state E_k can be reached in time interval Δt from states E_{k-1}, E_k, and E_{k+1}. Considering that birth and death are independent and using Figure 4.5, we may write:

1. The probability of exactly one birth in $(t, t + \Delta t)$ when the process is in state E_{k-1} is $\lambda_{k-1}\Delta t + o(\Delta t)$.
2. The probability of exactly one death in $(t, t + \Delta t)$ when the process is in state E_{k+1} is $\mu_{k+1}\Delta t + o(\Delta t)$.
3. The probability of exactly zero births in $(t, t + \Delta t)$ when the process is in state E_k is $1 - \lambda_k\Delta t + o(\Delta t)$.
4. The probability of exactly zero deaths in $(t, t + \Delta t)$ when the process is in state E_k is $1 - \mu_k\Delta t + o(\Delta t)$.

Let us denote with p_{ij} the probability for a transition from state i to state j. Using the Kolgomorov-Chapman approach, we analyze possible transitions of our particle. In this case, in a time interval $(t, t + \Delta t)$ we can enter state E_k only by three mutually exclusive possibilities:

1. $P\{\text{no state change occurred in state } k\} = [1 - \lambda_k\Delta t + o(\Delta t)]\,[1 - \mu_k\Delta t + o(\Delta t)]$;
2. $P\{\text{the system was in state } k - 1 \text{ and we had one birth}\} = \lambda_{k-1}\Delta t + o(\Delta t)$;
3. $P\{\text{the system was in state } k + 1 \text{ and we had one death}\} = \mu_{k+1}\Delta t + o(\Delta t)$.

If we use $P_k(t)$ to denote the probability that the system was in state k at time t, and we use $p_{k,j}(\Delta t)$ to denote the probability for a transition from state k to state j during time t, then we may write

$$P_k(t + \Delta t) = P_k(t)\,p_{k,k}(\Delta t) + P_{k-1}(t)\,p_{k-1,k}(\Delta t) + P_{k+1}(t)\,p_{k+1,k}(\Delta t) \tag{4.33}$$

If $p_{i,j}(\Delta t)$ are expressed using birth and death rates, we obtain

$$P_k(t + \Delta t) = P_k(t) - (\lambda_k + \mu_k)\Delta t P_k(t) + \lambda_{k-1}\Delta t P_{k-1}(t) + \mu_{k+1}\Delta t P_{k+1}(t) + o(\Delta t) \tag{4.34}$$

From the last equation, with some algebra, we may write

$$\frac{P_k(t+\Delta t)-P_k(t)}{\Delta t} = -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) + o(\Delta t) \tag{4.35}$$

If we allow $\Delta t \to 0$, then we have

$$\begin{aligned} \frac{dP_k(t)}{dt} &= -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t),\ k \geq 1 \\ \frac{dP_0(t)}{dt} &= -\lambda_0 P_0(t) + \mu_1 P_1(t),\ k = 0 \end{aligned} \tag{4.36}$$

We have lower boundary at the state 0 (no population). Also, it is possible to have an upper boundary if we have specified the maximum number of users in the system. Because birth-death process is a special case of the Markov processes, we can apply the same matrix notation introduced by (4.11) to (4.13), and we obtain the following transition matrix:

$$\mathrm{T} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & . & . & 0 \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & 0 & . & . & 0 \\ 0 & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix} \tag{4.37}$$

Let us now assume, for simplicity, that the system starts at state E_0 at time $t = 0$:

$$P_k(0) = \begin{cases} 1, k = 0 \\ 0, k > 0 \end{cases} \tag{4.38}$$

From (4.36) and (4.38) we get a differential equation, the solution to which is

$$P_0(t) = e^{-\lambda t} \tag{4.39}$$

Then, it is easy to continue for values $k \geq 1$:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t},\ k \geq 0, t \geq 0 \tag{4.40}$$

The last relation is called Poisson distribution. It characterizes the Poisson process, which, in fact, is a pure birth process with

$$\lambda_k = \begin{cases} \lambda, k \geq 0 \\ 0, k < 0 \end{cases} \text{ and } \mu_k = 0 \text{ for every } k \tag{4.41}$$

The Poisson process is significant in traffic theory in telecommunications, especially for circuit-switched networks, as we shall see later in this chapter. But the Poisson process has an even wider significance. It was shown by Palm that in many cases a large sum of independent stationary renewal processes tends to a Poisson process.

4.4.1 Stationary System

In practice we are interested in a stationary regime of processes because it is convenient for the analysis due to unique distribution of the state probabilities, independent of the initial condition. For a stationary system it holds that

$$P_k = \lim_{t \to \infty} P_k(t) \tag{4.42}$$

Now, we may write

$$\lim_{t \to \infty} \frac{dP_k(t)}{dt} = 0 \tag{4.43}$$

We define a system that satisfies the last equation as a system in statistical equilibrium. Furthermore, by using (4.36) we obtain

$$\lambda_{k-1} P_{k-1} + \mu_{k+1} P_{k+1} - (\lambda_k + \mu_k) P_k = 0, \; k \geq 1 \tag{4.44}$$

$$\lambda_1 P_1 - \lambda_0 P_0 = 0, \; k = 0 \tag{4.45}$$

Needless to say, because we cannot have a negative number of users in the system, $p_i = 0$ for $i < 0$, and $\lambda_i = 0$, $i < 0$; $\mu_j = 0$, $j < 1$. Because the birth-death process in statistical equilibrium is a case of Markov processes, we can apply the general equation for a Markov chain in equilibrium, which can be derived directly from the state-diagram given in Figure 4.7.

To obtain dependences among state probabilities, we draw arbitrary boundaries as shown in Figure 4.7. The total outgoing rate from a closed boundary should be equal to the total incoming rate into the boundary in a state of equilibrium. Using boundary 1 from Figure 4.7, we obtain

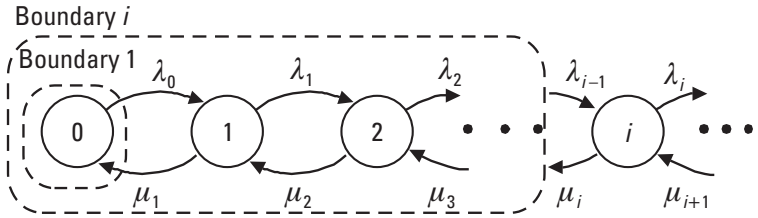


Figure 4.7 State-diagram of a birth-death process in equilibrium.

$$\mu_1 P_1 = \lambda_0 P_0 \tag{4.46}$$

and so on; from ith boundary we get

$$\lambda_{i-1} P_{i-1} = \mu_i P_i \tag{4.47}$$

Also, we require the conservation relation to hold for state probabilities:

$$\sum_{i=0}^{\infty} P_i = 1 \tag{4.48}$$

From (4.46) and (4.47), after some simple algebra, we obtain

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \tag{4.49}$$

Then, if we apply (4.48) and (4.49), we obtain

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \tag{4.50}$$

Because for all probabilities must hold $0 \leq P_k \leq 1$, there is a restriction on the values of the rates λ_k and μ_{k+1}, $k = 0, 1, \ldots$. To address the existence of the state probabilities P_k, we define two sums as follows:

$$S_1 = \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \tag{4.51}$$

$$S_2 = \sum_{k=1}^{\infty} \frac{P_0}{P_k} = \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\mu_{i+1}}{\lambda_i} \tag{4.52}$$

All states will be ergodic only and only if $S_1 < \infty$ and $S_2 = \infty$. Because we need an ergodic process to have equilibrium, it is of most interest to our analysis. The condition is fulfilled if there exists some k_0 such that for all $k > k_0$ it holds that

$$\frac{\lambda_k}{\mu_{\kappa+1}} < 1 \tag{4.53}$$

This condition is usually true in telecommunications systems that we design.

4.4.2 Birth-Death Queuing Systems in Equilibrium

Let us consider the importance of the statistical equilibrium of a birth-death process. We can define two types of equilibrium: a global balance and a local balance.

Global balance may be defined by using (4.44) and by applying it in infinitesimal time interval Δt:

$$\lambda_k \Delta t P_k + \mu_k \Delta t P_k = \lambda_{k-1} \Delta t P_{k-1} + \mu_{k+1} \Delta t P_{k+1} \quad k \geq 1 \tag{4.54}$$

By analyzing the last equation, we may observe that the left side of the relation gives the probability of a transition to neighboring states with respect to k—that is, to $k + 1$ (a new birth), and to $k - 1$ (a death). The right side of (4.54) gives the probabilities of transition from adjacent states to the state k. We can say that total outgoing traffic intensity from a particular state k is equal to the total incoming traffic intensity to that state. This is referred to as a global balance.

Local balance is defined by multiplying (4.47) by Δt, which leads to

$$\lambda_{i-1} \Delta t P_{i-1} = \mu_i \Delta t P_i \text{ for } i = 1,2,3,\ldots \tag{4.55}$$

From the last equation it is obvious that the possibility of a transition from state $k - 1$ to state k (the left side) is equal to the transition probability in the reverse direction (the right side of the equation). This is called a *local balance*, and (4.55) is referred to as a *local balance relation*.

4.5 Teletraffic Theory for Loss Systems with Full Accessibility

We covered the basics of queuing theory in previous sections of this chapter. Now, let us go through traditional way of design and analysis in telecommunications represented by the famous Erlang's loss formula (elsewhere it is referred to as Erlang-B formula or Erlang's first formula).

In the early decades of telephony and switching systems, Erlang made an extensive analysis of the traffic data such as telephone calls initiated by users connected to a switching system (telephone exchange), blocking of the calls and their duration. He found that call arrivals suit well into a Poisson process. Also, call duration was shown to be easily modeled by using the exponential distribution for the call duration times. According to the above statement, we may say that Erlang's loss formula is based on the following model:

- Arrival process is Poisson and service times are exponentially distributed.
- We consider a circuit-switched system with servers (channels, trunks, or time slots) working in parallel.
- An arrival is accepted for service if any channel is idle. The system allocates one channel per call. We say the group (of channels) has full accessibility when every incoming user competes with other users for all idle channels (not allocated resources).

Using the queuing theory and Kendall notation for queuing systems [2], we can describe Erlang's conclusions by using M/M/*n*/*n* queuing system. In this case *n* servers are *n* channels that may serve up to *n* users at the same time. Usually, the number of potential users is many times higher than the number of available channels (resources) due to the economic aspects of telecommunications networks design. This statement holds for both analog and digital circuit-switched networks, because in both cases we have one type of traffic only (voice telephony) and the system allocates equal resources for each call. Of course, because telephony is bidirectional, we have occupancy of two channels per call (one for each direction), but in traditional traffic theory it is enough to consider only one direction in calculations due to symmetrical resource allocation in both directions. This picture will change, however, if we introduce packet-based communication and heterogeneous services, as we shall later see.

The state diagram for M/M/*n*/*n* is given in Figure 4.8. For this system we have

$$\begin{aligned} \lambda_k &= \lambda,\ k = 0,1,2,\ldots,n-1 \\ \mu_k &= k\mu,\ k = 1,2,\ldots,n \end{aligned} \tag{4.56}$$

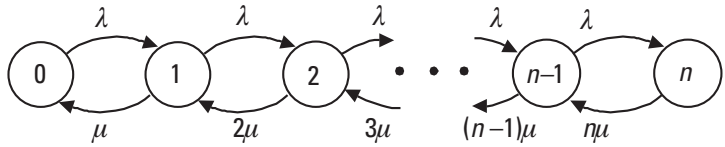


Figure 4.8 State-diagram for Erlang's loss formula.

Using (4.49) and (4.50), which we proved for the general case of the birth-death processes, and by replacing λ_k and μ_k according (4.56), we obtain

$$P_k = \frac{\frac{\left(\frac{\lambda}{\mu}\right)^k}{k!}}{\sum_{i=0}^{n}\frac{\left(\frac{\lambda}{\mu}\right)^i}{i!}} = \frac{\frac{A^k}{k!}}{\sum_{i=0}^{n}\frac{A^i}{i!}} \tag{4.57}$$

where $A = \lambda/\mu$ is intensity of the offered traffic. It is expressed in units Erlangs in honor of Erlang. Relation (4.57) is called *Erlang distribution* or truncated *Poisson distribution.*

Definition of the carried traffic: We define the carried traffic per single channel as a sum of busy times for each channel during time interval T divided by the time interval. For a pool of resources, it is given by

$$A = \frac{\int_0^T \overline{x(t)}dt}{T} \tag{4.58}$$

where $\bar{x}$ is the average number of busy channels:

$$\overline{x(t)} = \sum_{x=0}^{n} xP(x,t) \tag{4.59}$$

If there are many users in the system, then we can apply the statistical equilibrium where the number of simultaneously occupied channels does not depend upon the moment in time—that is, $P(x, t) = P(x)$. Then, the offered traffic may be calculated by using the following equation:

$$Y = \sum_{x=1}^{n} xP(x) \tag{4.60}$$

Also, we may define offered traffic as average intensity of calls C_A (calls/second) multiplied by average call duration time t_μ:

$$A = C_A t_\mu \tag{4.61}$$

It is obvious that $C_A = \lambda$, while $t_\mu = 1/\mu$ due to exponential distribution of the call duration, so we get

$$A = \frac{\lambda}{\mu} \tag{4.62}$$

The probability that n channels are busy at a random point in time can be obtained from (4.57) for $k = n$:

$$E_n(A) = \frac{\dfrac{A^n}{n!}}{\displaystyle\sum_{i=0}^{n} \frac{A^i}{i!}} \tag{4.63}$$

The last relation is called Erlang's loss formula, written for the first time in 1917 by Erlang. It is also denoted as $E_{1,n}(A) = E_n(A)$ where index 1 indicates that it is Erlang's first formula (Erlang-B formula).

For the sake of completeness, we briefly refer to two other situations regarding the number of users N and the number of trunks n—that is, $N \leq n$ and $N > n$ (not $N >> n$). In the Erlang's case, the number of users $N >> n$ and hence number of idle users is $N - n >> n$. However, for these two cases, we do not have independence of the offered traffic from the number of busy trunks. Therefore, for both of them the call arrival rate depends upon the number of active users [i.e., $\lambda_k = (N - k)\lambda$, $\mu_k = k\mu$], for k busy trunks. In these cases, we use an additional parameter $\beta = \lambda/\mu$, which is offered traffic per idle source. If we consider that the source changes between idle and busy states, the offered traffic per source is

$$\alpha = \frac{T_{busy}}{T_{idle} + T_{busy}} = \frac{1/\mu}{1/\lambda + 1/\mu} = \frac{\beta}{1 + \beta} \tag{4.64}$$

Then, the offered traffic to the system is $A = N\alpha$. Using the state transition diagram given in Figure 4.8, we can derive the distribution functions. Hence, for the case $N \leq n$ we get the Binomial distribution:

$$P_k = \binom{N}{k} \alpha^k (1 - \alpha)^{N-k} \tag{4.65}$$

For the case $N > n$, we get Engset distribution (it is also called truncated binomial distribution):

$$P_k = \binom{N}{k} \frac{\beta^k}{\displaystyle\sum_{i=0}^{n} \binom{N}{i} \beta^i} \tag{4.66}$$

The Erlang-B formula has been used throughout the last century, and it continuous to be used today for design and analysis of circuit-switched telephone networks with wired (fixed) access. With some approximations, it may be used in mobile cellular networks. We will refer to wireless networks later in this chapter.

When we design a system we need to specify some *grade of service* (GoS) or QoS requirements (we define these later in this chapter). For that purpose, we define performance parameters, which indicate the QoS (or GoS) level of the designed system. The main QoS parameter in circuit-switched networks is blocking. We define three types of blocking: call congestion, time congestion, and traffic congestion. In the following we consider the Erlang distribution.

Call congestion is the probability that a random call is lost due to blocking; that is, all channels are busy at call arrival:

$$B_C = \frac{\lambda P(n)}{\sum_{i=0}^{n} \lambda P(i)} = E_n(A) \tag{4.67}$$

Time congestion is equal to the percent of time during which all available resources are busy:

$$B_t = \frac{t_{congestion}}{t_{congestion} + t_{free}} = \frac{t_{congestion}}{T} = E_n(A) \tag{4.68}$$

The carried traffic may be calculated by using

$$Y = \sum_{i=0}^{n} iP_i = \sum_{i=0}^{n} i \frac{\left(\frac{\lambda}{\mu}\right)^i}{i!} P_0 = A\left[1 - E_{1,n}(A)\right] \tag{4.69}$$

We define the lost traffic as

$$Y_R = A - Y = AE_n(A) = AB \tag{4.70}$$

Lost traffic is intuitive because there is no traffic when we have lost calls. Then, we may define traffic congestion as

$$B_T = \frac{A - Y}{A} = E_n(A) \tag{4.71}$$

When number of users is many times greater then the number of channels (assumption holds for Erlang's loss formula), we have

$$B_c = B_t = B_T = E_{1,n} \tag{4.72}$$

This is a characteristic of all systems with the Poisson arrival process and a large population. In Figure 4.9 we show blocking as a function of the offered traffic.

We have three related parameters in Erlang's first formula: number of channels n, offered traffic A, and blocking $E_{1,n}(A)$. If we know two of them, we can calculate the third parameter. For example, if we design a switching system, we need to predict the offered traffic to the system and specify the desired GoS. Then, we can obtain the number of needed channels by applying Erlang's first formula. Although we derived the formula starting from the exponential distribution of connection holding times, it can be easily shown that it is independent from the holding time distributions. The basic assumption in Erlang's loss formula is Poisson arrival process, which is fulfilled only when we have a large number of sources, according to Palm's theorem. This assumption is valid in telephone systems and that is why the formula has been widely used since the second decade of twentieth century.

4.6 Teletraffic Theory for Loss Systems with Multiple Traffic Types

Today, in telecommunications networks we usually have more than one traffic type (the voice telephony). Different traffic types have different characteristics considering call arrivals, intensities, and call duration times. Networks with

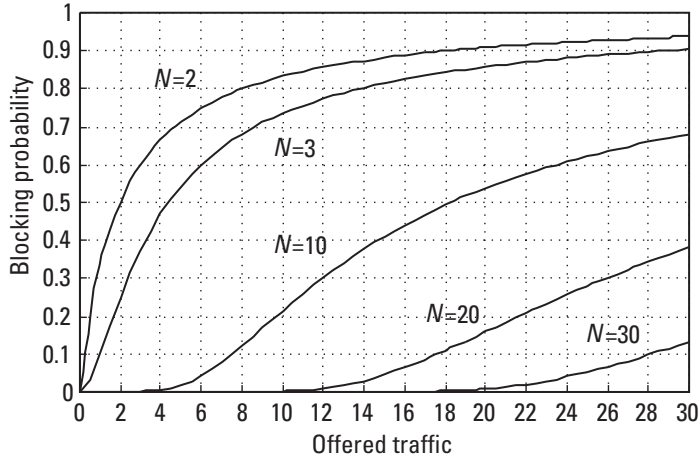


Figure 4.9 Blocking as a function of offered traffic by using Erlang's first formula.

integrated heterogeneous traffic sources are called integrated networks (it is a notation for a circuit-switched networks). Each service has different traffic characteristics. In networks with asynchronous transport of the information, different traffic services have different intensities. Typical examples of asynchronous transmission are packet-based networks.

4.6.1 Loss Systems with Integrated Traffic

In packet-based networks as well as in integrated networks, we often have arrivals with different rates and different resource demands from different traffic types. If we assume that a channel (or we may say a bandwidth allocation unit) is the smallest unit that can be allocated by the system, then we can have allocation of two, three, or more bandwidths units (e.g., channels) for some connections (e.g., video streaming).

We analyze the teletraffic system M[ξ]/M/n/k. This is full accessibility group of resources with possibility of losses. The distribution of the connections interarrivals times and connection holding times are assumed to be exponential. The ξ is a random variable, which shows the occupied resources as a function of time. Let b_i denote the probability that i bandwidth units are allocated to a call:

$$b_i = P\{\xi = i\} \tag{4.73}$$

Then, the average number of occupied lines is

$$b = \sum_{i=1}^{\infty} i b_i \tag{4.74}$$

A blocking (loss) occurs when a new connection arrives and there are not enough resources to be allocated for that connection. For example, if a connection demands four channels, and there are three or fewer channels available, then it will be rejected or terminated.

We will refer again to resources as channels. So, from every state $i = 0, 1, \ldots, n-1$ we have arrival rate , but varying number of requested channels per call (i.e., 1 channels/call, 2 channels/call, and so on). Using the condition for a global balance, from Figure 4.10 we may write

$$\begin{aligned} &\lambda \sum_{i=0}^{j-1} P_i b_{j-1} + (j+1)\mu P_{j+1} = (\lambda + j\mu) P_j \quad j = 0,1,2,\ldots,n-1 \\ &\lambda \sum_{i=0}^{n-1} P_i \sum_{k=n-i}^{\infty} b_k = n\mu P_n \end{aligned} \tag{4.75}$$

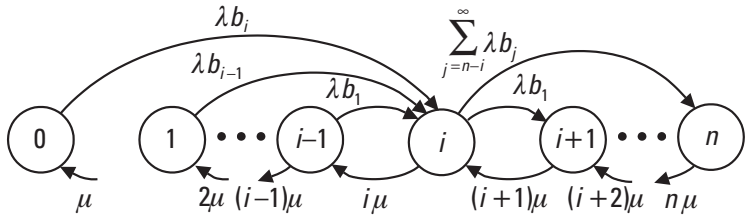


Figure 4.10 State diagram for the group analysis of integrated traffic.

Of course, the conservation of state probabilities must hold:

$$\sum_{i=0}^{\infty} P_i = 1 \tag{4.76}$$

By summing the equations (4.75) from 0 to $j-1$ we obtain a recurrent relation for the state probabilities:

$$P_j = \frac{\lambda}{j\mu} \sum_{i=0}^{j-1} \left(P_i \sum_{k=j-1}^{\infty} b_k \right) j = 1,2,3,\ldots,n \tag{4.77}$$

We can calculate the offered traffic in this system by using

$$A = b\frac{\lambda}{\mu} \tag{4.78}$$

Then, the carried traffic is

$$A_0 = \sum_{i=0}^{n} iP_i \tag{4.79}$$

Finally, we may calculate call losses by using

$$B_c = 1 - \frac{A_0}{A} \tag{4.80}$$

Probability distribution for ξ may vary. For an example, ξ may be geometrically distributed [4]. Then, one can calculate the probability that i-channels are simultaneously occupied by using

$$b_i = p(1-p)^{i-1} \; i = 1,2,3,\ldots \tag{4.81}$$

where p is the probability that a call requests one channel. When using geometrical distribution, the average number of busy channels allocated per call is $b = 1/p$. After some algebra, for the state probabilities we obtain

$$P_j = \frac{pA}{j}\sum_{i=0}^{j-1} P_i(1-p)^{j-i-1} \quad j = 1,2,3,\ldots,n \tag{4.82}$$

We also may use other distribution for the description of the number of bandwidth units allocated per connection.

In the next sections we extend the analysis to systems with multiple classes and different call arrival rates and call durations per class. But, for such analysis we need to introduce combinations of exponential distributions.

4.6.2 Phase-Type Distributions

Exponential distribution is the basis of time interval distributions (i.e., it is used for modeling time durations of different events in telecommunications—for example, call duration and intercall arrival time). Time distribution within teletraffic theory is most important. Hence, we may find it suitable to combine exponential distributions in series or in parallel. Also, we may combine exponential distributions both in series and in parallel and thus obtain a class of general distributions called phase-type distributions.

4.6.2.1 Steep and Flat Distributions

Sometimes, if we cannot describe a random process (e.g., call duration) with single exponential distribution (i.e., one parameter), then we can use two or more exponential distributions. There are two basic types of such combinations of exponential distributions (introduced by Palm):

1. *Steep distribution*, which is a set of stochastic independent exponential distributions in series (Figure 4.11);
2. *Flat distribution*, which is a set of stochastic independent exponential distributions in parallel (Figure 4.12).

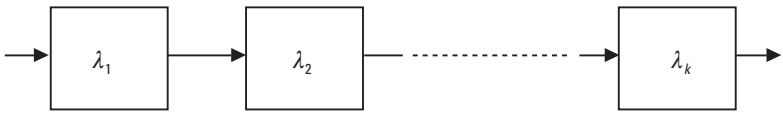


Figure 4.11 Steep distribution.

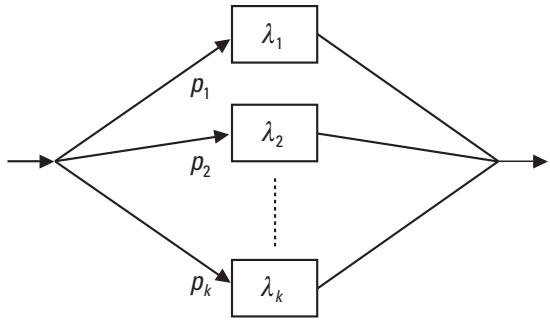


Figure 4.12 Flat distribution.

We refer to steep distribution as generalized Erlang distribution. It is obtained by convolving k exponential distributions. If all k exponential distributions are identical, we get an Erlang-k distribution (we refer only to this case) with the following probability distribution function:

$$f(t) = \frac{(\lambda t)^{k-1}}{(k-1)!}\lambda e^{-\lambda t},\ k = 1,2,\ldots \tag{4.83}$$

The distribution function equals

$$F(t) = 1 - \sum_{j=0}^{k-1}\frac{(\lambda t)^j}{j!}e^{-\lambda t} \tag{4.84}$$

Flat distribution, in general, is the weighted sum of exponential distributions (i.e., parallel combination), the distribution function of which is given by

$$F(t) = \int_0^{+\infty}\left(1 - e^{-\lambda t}\right)dW(\lambda) \tag{4.85}$$

where $W(\lambda)$ is weight function, which can be discrete or continuous. If $W(\lambda)$ is a discrete function, the obtained distribution is called hyper-exponential. If we have parallel combination k exponential distributions with parameters, $\lambda_1, \lambda_2, \ldots, \lambda_k$, and discrete positive weights $W(\lambda)$ (i.e., $p_1, p_2, \ldots, p_k$) where

$$\sum_{j=1}^{k} p_j = 1 \tag{4.86}$$

then we can write the hyper-exponential distribution function as

$$F(t) = 1 - \sum_{j=1}^{k} p_j e^{-\lambda_j t} \tag{4.87}$$

For example, hyper-exponential distribution can be used for modeling the call-holding time of a system with multiple traffic types, where each traffic type has exponentially distributed call duration but different mean value.

4.6.2.2 Coxian Distributions

Coxian distribution is a general distribution obtained by combining steep and flat distributions. Each exponential distribution that is used in the Coxian distribution (or in other specific cases, such as steep or flat distributions) is called an exponential phase. Therefore, we refer to Coxian distribution as a general class of phase-type distributions. Elsewhere it is called Erlang distribution with branches.

The Coxian distribution function can be written as a weighted sum of exponential distributions. But, different from flat distributions where the sum of all weights equals one, as given by (4.85), and all weights are positive, in Coxian distribution we have the following conditions for weights:

$$0 \le \sum_{j=1}^{k} p_j \le 1 \tag{4.88}$$

Any distribution function can be approximated arbitrarily close by a function that consists of series of steps, as Coxian distribution does. We can explain Coxian distribution by using k-stage server, as shown in Figure 4.11. Let service time at stage j be described by a random variable X_j, $j = 1, 2, ..., k$. Then, service time, described by the random variable X, can have values $X = X_1$, or $X = X_1 + X_2$, or .... $X = X_1 + X_2 + ... + X_k$. So, a random variable has a Coxian distribution of order k if it has to go through up to at most k exponential phases. The mean length of phase n is λ_n, $n = 1, 2, ..., k$. It starts in phase 1. At each stage n, p_n is probability that a job leaves the server immediately after completing stage n, and $q_n = 1 - p_n$ is the probability that a job requires more service after stage n, as shown in Figure 4.13.

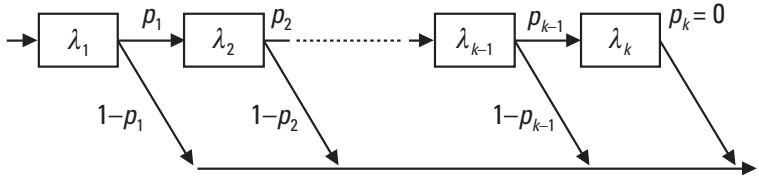


Figure 4.13 Coxian distribution.

Coxian distribution gets a lot of attention due to its wide applicability to distributions of practical interest and also because of the possibility of using the theory of Markov processes (phase-method), which do not require advanced mathematics. For example, Erlang-k distribution and hyper-exponential distribution are special types of Coxian distribution. We may also show exponential distribution through its Coxian distribution equivalent.

4.6.3 Multidimensional Erlang Formula

We generalize the teletraffic theory to systems with multiple traffic types. In such systems we have different classes of services (i.e., services with various traffic and QoS demands). Each class corresponds to a traffic stream. We will expand the Erlang loss formula to a general case of a network with multiple traffic classes.

For that purpose, we consider a group of n bandwidth units (e.g., channels, time slots), which is offered to two independent traffic streams with arrival/service rates (λ_1, μ_1) and (λ_2, μ_2). Then, we have the offered traffic $A_1 = \lambda_1/\mu_1$ for the stream 1, and $A_2 = \lambda_2/\mu_2$ for the stream 2.

Let us denote with (i, j) the state of the system when there are i connections from stream 1 and j connections from stream 2. We limit allocation to one channels or slot per connection. Then, the following restrictions hold: $0 \le i \le n$, $0 \le j \le n$, and $0 \le i + j \le n$. In this case we have a two-dimensional transition-state diagram, given in Figure 4.14, that corresponds to a reversible Markov process.

By assuming equilibrium in the system, we may apply global balance equations. If we denote with $P(i, j)$ the probability that the system is in state (i, j), then we may write

$$P(i, j) = P(i)P(j) = k_0 \frac{A_1^j}{i!} \frac{A_2^j}{j!} \tag{4.89}$$

The constant k_0 in the last equation is needed to conserve the state probabilities. Because birth processes are Poisson, the total arrival process will be Poisson also. The same holds for the exponential service time distribution. If we consider circuit-switched networks, then models with Poisson arrivals are insensitive to service time distribution, and we obtain

$$P(k = i + j) = k_0 \frac{(A_1 + A_2)^k}{k!} = k_0 \frac{A^k}{k!} \tag{4.90}$$

From probability conservation law, one may obtain the constant k_0:

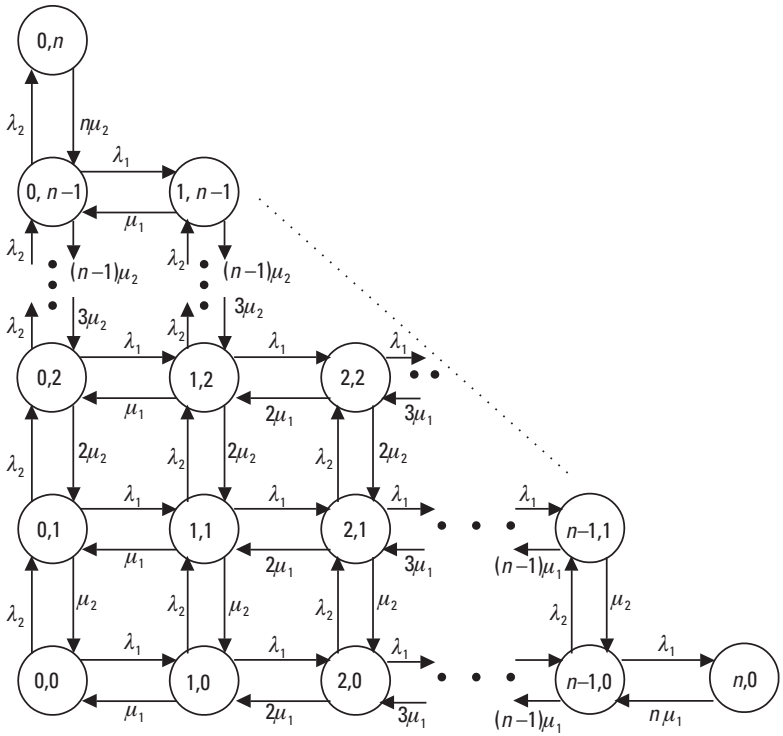


Figure 4.14 Two-dimensional transition-state diagram for a loss system with two traffic types and *n* bandwidth units.

$$k_0 = \frac{1}{\sum_{i=0}^{n} \frac{(A_1 + A_2)^i}{i!}} \tag{4.91}$$

The offered traffic is $A = A_1 + A_2$. The time congestion, call congestion, and traffic congestion are all identical for both traffic streams due to the Poisson arrivals, and all are equal to $P(n) = P(i+j=n)$. We may generalize this conclusion to s streams with rates (λ_1, μ_1), (λ_2, μ_2), ..., (λ_s, μ_s). The total offered traffic into the system is

$$A = \sum_{i=1}^{s} \frac{\lambda_i}{\mu_i} = \sum_{i=1}^{s} A_i \tag{4.92}$$

We may also interpret the system behavior by using the hyper-exponential service time distribution. In that case, we may again have an overall Poisson arrival process with total arrival rate:

$$\lambda_T = \lambda_1 + \lambda_2 + ... + \lambda_s \tag{4.93}$$

The holding times are hyper-exponentially distributed with density function:

$$f_\mu(t) = \frac{\lambda_1}{\lambda_T}\mu_1 e^{-\mu_1 t} + \frac{\lambda_2}{\lambda_T}\mu_2 e^{-\mu_2 t} + ... + \frac{\lambda_s}{\lambda_T}\mu_s e^{-\mu_s t} \tag{4.94}$$

The mean service time for the case of hyper-exponential distribution is

$$\overline{t_\mu} = \frac{\lambda_1}{\lambda_T}\frac{1}{\mu_1} + \frac{\lambda_2}{\lambda_T}\frac{1}{\mu_s} + ... + \frac{\lambda_s}{\lambda_T}\frac{1}{\mu_2} = \frac{\sum_{i=1}^{s} A_i}{\lambda_T} = \frac{A_T}{\lambda_T} \tag{4.95}$$

The average service rate may be calculated as

$$\mu_T = \frac{1}{\overline{t_\mu}} = \frac{\lambda_T}{A_T} \tag{4.96}$$

So, for generalized model with *s* traffic streams into the systems, we obtain

$$P(j_1, j_2, ..., j_s) = k_0 \frac{A_1^{j_1}}{j_1!}\frac{A_1^{j_2}}{j_2!}...\frac{A_s^{j_s}}{j_s!}$$
$$0 \le j_i \le n, 1 \le i \le s,\ j_1 + j_2 + ... + j_s \le n \tag{4.97}$$

We may write that the probability that *j* bandwidth units are busy in a loss system with *n* bandwidth units is

$$P(j) = \frac{\frac{\left(\sum_{i=1}^{s} A_i\right)^j}{j!}}{\sum_{k=0}^{n}\frac{\left(\sum_{i=1}^{s} A_i\right)^K}{k!}} \tag{4.98}$$

The last relation is, in fact, multidimensional Erlang's loss formula. The bandwidth requested by a call may depend on the type of service. Thus, voice service may require one bandwidth unit per call, and video stream may require two or more. Let us denote with c_i the requested bandwidth units per call of service type *i*. Then we may write

$$c_i j_i \le n_i \le n,\ i = 1,2,\ldots,s$$
$$\sum_{i=1}^{s} c_i j_i \le n \tag{4.99}$$

where j_i is number of calls from service type i.

4.6.4 Priority Queuing

So far, we have considered only classical queuing systems where all traffic processes are birth and death processes. In systems with different traffic types, however, we may need to apply priorities between classes. We consider M/G/1 queue with r traffic types (i.e., customer types). We may choose M/M/1 queue as well, but it is only a special case of more general M/G/1 queue (refer to Kendall's notation given in this section). The type i customer arrives according to Poisson stream with rate λ_i, $i = 1, 2, ..., r$. The mean service time is denoted as S_i. The offered traffic from stream i is $A_i = \lambda_i S_i$. The type 1 customer has highest priority, the type 2 customer the second highest priority, and so on. Basically, there are two types of priorities:

1. *Nonpreemptive priority:* current service of lower priority is not interrupted by arrivals of higher priority customers, but they have to wait until the service of lower priority customers has been completed;
2. *Preemptive-resume priority:* an ongoing service is interrupted by arrival of higher priority customer. Later the service time of the lower priority customer resumes at the point where it was interrupted.

Kendall's Notation

Kendall introduced a notation to characterize different groups of queuing models: $A/B/N$, where A is arrival process, B is service time distribution, and N is number of servers. For a more complete description of a queuing model, we use notation $A/B/N/K/S/X$, where K is the total capacity of the system (default value $= \propto$), S is population size of customers (default value $= \propto$), and X is queuing disciplines (default = FCFS). If the last three letters in the complete notation are omitted, we assume their default values.

4.6.4.1 Nonpreemptive Priority

Let us denote with W_i and L_i the mean waiting time in the queue of type i customers and number of type i customers waiting in the queue, $i = 1, 2, ...r$, respectively. All customers are divided into r priority classes. Service time and residual time for the i type customer is denoted by S_i and R_i, respectively. Then, waiting time for the highest priority customers (i.e., type 1) is given by

$$W_1 = L_1^q S_1 + \sum_{j=1}^{r} A_j R_j \tag{4.100}$$

A very important relation between the mean number of customers in the system L, the mean sojourn time T, and the average number of customers entering the system λ, is given by Little's law that states $L = \lambda T$ [5]. Applying Little's law to a queue (without the server) gives the following relation between the queue length L_i^q and the waiting time W_i:

$$L_i^q = \lambda W_i \tag{4.101}$$

Combining the last two equations, we get the mean waiting time for the highest priority class:

$$W_1 = \frac{\sum_{j=1}^{r} A_j R_j}{1 - A_1} \tag{4.102}$$

Determination of the minimum waiting time for lower priority customers is more complicated. It is suitable to divide this waiting time in portions as shown in Figure 4.15. The first portion s_1 is the amount of work associated with the customer in service and all other customers of the same or higher priority that are present in the queue. The second portion s_2 is the amount of higher priority work arriving during s_1. The third portion is the amount of higher priority work arriving during s_2, and so on. Then, the mean waiting time is

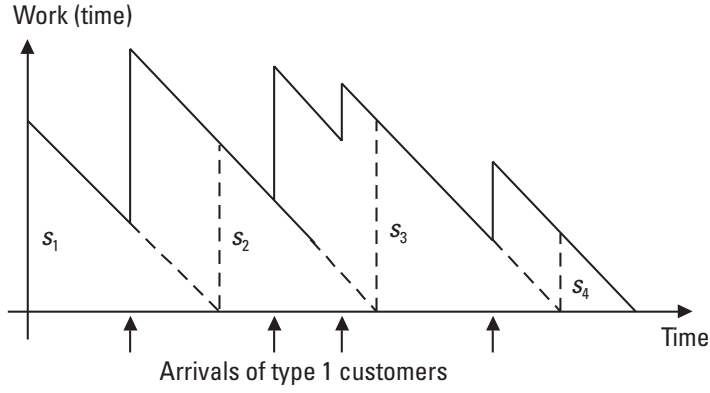


Figure 4.15 Waiting time of type 2 customer.

$$W_i = E\left(s_1 + s_2 + s_3 + \ldots\right) = \sum_{k=1}^{\infty} s_k \tag{4.103}$$

Because each class j has offered traffic A_j, it is easy to show that for i type customers the following relation for the consecutive time portion holds:

$$E\left(s_{k+1}\right) = \left(A_1 + A_2 + \ldots + A_{i-1}\right) E\left(s_k\right) \tag{4.104}$$

Using (4.104) for each traffic type i we get a geometrical progression of the time portions:

$$E\left(s_{k+1}\right) = \left(A_1 + A_2 + \ldots + A_{i-1}\right)^k E\left(s_1\right),\ k = 0,1,2,\ldots \tag{4.105}$$

From (4.103) and (4.105) we obtain the mean waiting time of i traffic type:

$$W_i = \frac{E\left(s_1\right)}{1 - \left(A_1 + A_2 + \ldots + A_{i-1}\right)} = \frac{\sum_{j=1}^{i} L_j^q S_j + \sum_{j=1}^{r} A_j R_j}{1 - \left(A_1 + A_2 + \ldots + A_{i-1}\right)} \tag{4.106}$$

Using (4.102) and (4.106), it is straightforward to show that

$$W_i = \frac{\sum_{j=1}^{r} A_j R_j}{\left[1 - \left(A_1 + A_2 + \ldots + A_i\right)\right]\left[1 - \left(A_1 + A_2 + \ldots + A_{i-1}\right)\right]} \tag{4.107}$$
$$i = 1,2,\ldots,r$$

The mean sojourn time for i traffic type is

$$T_i = W_i + S_i \tag{4.108}$$

Considering the residual times, Kleinrock has proved that when the service time of a customer is independent of the queuing discipline (i.e., it is work conserving), average waiting time for all traffic types weighted by the traffic intensity A_i of the observing class is also independent of the queue discipline. It is expressed through the following relation, which is referred to as Kleinrock's conservation law:

$$\sum_{i=1}^{r} A_i W_i = \frac{AV}{1 - A} \tag{4.109}$$

where $A = A_1 + A_2 + ... + A_r$ and V is the mean residual service time for the customer when higher priority customer arrives, that is,

$$V = \sum_{i=1}^{r} A_i R_i \tag{4.110}$$

The mean waiting time given by (4.107) can also be derived using the last two relations.

Prioritization is very likely to be applied in wireless packet networks (e.g., in scheduling algorithm) to assure higher QoS for a particular service type(s).

4.6.4.2 Preemptive Priority

Derivation of the service delay for preemptive priority follows directly from the previous one for the non-preemptive priority. This priority rule is more typical for computer systems, and therefore we just briefly refer to it.

The waiting time W_i of i traffic type is calculated using the same relation (4.107) as for non-preemptive priority, because the total amount of work in the system does not depend upon the order in which the customers are served. But, there is a difference in the sojourn time, which can be derived in the similar manner as relation (4.106). It is given by

$$T_i = W_i + \frac{S_i}{1 - \left(A_1 + A_2 + ... + A_{i-1}\right)}, \; i = 1,2,...,r \tag{4.111}$$

4.6.5 Error Control Impact on Traffic

To cope with errors in the transmission links different and often complex coding techniques are used. While in wireline links errors are independent, in wireless links errors tend to group in bursts due to fast fading radio channels. However, error control causes a small traffic increment. For error control two basic coding methods are used:

1. Error correcting [*forward error correction* (FEC)];
2. Error detection and retransmission [*automatic repeat request* (ARQ)].

By sending redundant information over the wireless channel, errors can be corrected at the receiving end without retransmissions of data (e.g., FEC). Two traditionally used methods for error correction are: blocking codes and convolutional codes. Lately, turbo codes are also applied in the wireless channel (e.g., for

high data rates in UMTS). Due to errors in bursts, we use interleaving and concatenation of codes. Interleaving sends information out of order (or mixed with information from other users) to spread the effect of a burst error. Concatenated codes protect the information using several codes in parallel. Better protection of information data, however, requires higher redundancy in coding, thus leading to a slight increase in traffic demands. Usually, FEC is applied for real-time flows that cannot afford additional delay due to retransmissions. Also, redundancy is higher for traffic types that are more sensitive to losses. For example, WCDMA for UMTS uses 1/2-rate and 1/3-rate convolutional codes for low data rates and 1/3-rate turbo code for high data rates. The advantage of FEC is fixed delay and utilization of resources, while implementation complexity is the drawback.

For flows that demand error-free transmission, but are not too sensitive to delays (e.g., nonreal-time flows), we usually use error detection and retransmission techniques (e.g., ARQ). There are three basic ARQ schemes: stop-and-wait, go-back-N, and selective ARQ. The main difference between the three ARQ techniques lies in the decision when and what to retransmit. In stop-and-wait, the transmitter after sending a frame waits for an ACK from the receiver and then continuous with the next frame. Inefficiency of stop-and-wait ARQ can be overcome with go-back-N, which allows continuing transmission without waiting for ACK for each frame. Up to N consecutive frames can be transmitted without receiving ACKs (i.e., a sliding window technique). In a case of transmission errors, the go-back-N starts retransmissions of all frames from the frame in error (i.e., goes back up to N frames). In channels with a high error rate, go-back-N is inefficient because it also eventually retransmits error-free frames that are sent after the frame in error. Selective ARQ overcomes this problem by allowing the receiver to accept frames out of order, and by allowing retransmissions of individual frames.

ARQ is simpler than FEC, but is targeted to environments with very low error probabilities. Usually, ARQ schemes are applied at the link level for data transfers. Also, combinations of FEC and ARQ can be used for better performance.

Both redundancy of FEC and retransmissions of ARQ increase the traffic in the wireless channel due to error control. Generally, we may write the probability that a packet is being retransmitted as

$$\begin{aligned} P_r &= P[\text{Retransmission}] \\ &= P[\text{Error}]P[\text{Error is detected and not corrected}] \end{aligned} \tag{4.112}$$

Considering the error probability P[Error], we may use the Markov error model for modeling error-state and error-free state of the channel (refer to Section 6.5).

Considering the error control, if the user data rate is R_{user} (in bps), then the actual traffic demand for network resources will be $R > R_{user}$; that is, we can write

$$R_{user} = \left(1 - \frac{n_{red}}{n_{frame}}\right) R_{link} \eta_r = R \eta_r \tag{4.113}$$

where n_{red} is the number of overhead and redundancy bits in the frame due to error control coding (for both FEC and ARQ), n_{frame} is the total number of bits including data bits and redundancy bits in a frame, $0 < \eta_r \leq 1$ is transmission efficiency due to retransmissions, R_{link} is the actual bit rate on the link, and R is user data rate without retransmissions. The choice of the coding scheme is a matter of a network design, and it depends upon the characteristics of the traffic type (e.g., voice or Web) and sensitivity to errors and delay (e.g., refer to Table 2.4).

We can derive the transmission efficiency η_r for all three ARQ schemes. The probability that a frame is successfully transmitted after $i - 1$ retransmissions can be obtained as

$$P_i = (1 - P_r) P_r^{i-1} \tag{4.114}$$

Then, the average total time to transmit a frame is given by

$$t_{total} = \begin{cases} t_0 + t_{out} N_r, & \text{Stop-and-wait} \\ t_{frame}(1 + NN_r), & \text{Go-back-N} \\ t_{frame}(1 + N_r), & \text{Selective ARQ} \end{cases} \tag{4.115}$$

where $t_0 = 2t_{pp} + (n_{frame} + n_{ack}) / R_{link}$ is the basic delay (t_{pp} is one-way total propagation and processing delay), t_{out} is timeout for stop-and-wait (for simplicity we may assume $t_{out} = t_0$), t_{frame} is transmission time for a frame, N is window size for go-back-N, and N_r is average number of retransmissions per frame. One may calculate N_r using the following:

$$N_r = \sum_{i=1}^{\infty} (i-1) P_i = \sum_{i=1}^{\infty} (i-1)(1 - P_r) P_r^{i-1} = \frac{P_r}{(1 - P_r)} \tag{4.116}$$

Using the last two relations and being aware that $n_{frame} / t_{frame} = R_{link}$, it is straightforward to derive the following ARQ efficiencies:

$$\eta_r = \frac{n_{frame}/t_{total}}{R_{link}} = \begin{cases} (1-P_r)\dfrac{n_{frame}}{n_{frame}+n_{ack}+2t_{pp}R_{link}}, & \text{Stop-and-wait} \\ (1-P_r)\dfrac{1}{1+(N-1)P_r}, & \text{Go-back-N} \\ (1-P_r), & \text{Selective ARQ} \end{cases} \tag{4.117}$$

where n_{frame} is length of a frame, and n_{ack} is length of an ACK.

From the above analysis it follows that if we want to allocate to the user effective bit rate R_{user}, then we need to allocate link rate $R_{user} + \Delta R$, where ΔR is the traffic increment due to retransmissions; that is,

$$\Delta R = \frac{1-\eta_r}{\eta_r} R_{user} \tag{4.118}$$

4.7 Teletraffic Modeling of Wireless Networks

We cannot apply directly the teletraffic theory developed for wireline networks to cellular wireless networks [6]. In mobile networks different channels may be occupied and released several times during one call. This phenomenon is due to the cellular topology of mobile and wireless networks, where each cell is a service zone with limited capacity. Here, we define a handover or a handoff as a process of carrying an ongoing connection from one cell to one of its adjacent cells. This means allocation of resources in the new target cell by some algorithm and release of the resources in the earlier cell. If an idle channel is available in the target cell, the handover call is resumed nearly transparently to the user. Otherwise the handover call is dropped. So, a channel in cellular wireless networks may be occupied by an arrival of a new call or a handover call. Also, a channel may be released either by ending of the call or by handing the call over to one of the neighboring cells.

For simplicity, in this chapter we restrict the analysis to wireless cellular networks with one traffic type. In Chapter 7 we extend the theory to multiple traffic types. In cellular networks traffic parameters are related with mobility parameters, such as velocity of users and characterization of their movement within a cell. In wireline networks we considered performances parameters: call arrival process, call departure process, and blocking. For analysis of cellular

wireless networks we need to introduce additional wireless-specific parameters due to handover phenomenon, such as average channel holding time, new call and handover call intensities, new call blocking probability, and handover call blocking probability. The last two parameters define the QoS level in the network.

First, consider the following simple scenario: users in a certain cell 1 initiate fresh calls as a Poisson process at rate $\lambda_1 = \lambda_n$ and receives no handovers. So, the distribution function for arrival of new calls is $p_n(t) = \lambda_n e^{-\lambda_n t}$. However, the calls are allowed to perform handover to the neighboring cells. It is assumed that the number of users in a cell is $N >> c$, where c is the number of channels in the cell, so the calls from different users may be considered as independent. We start by using a simple resource allocation scheme (one channel per call). Also, we shall assume that each mobile user in the originating cell may complete the call in the cell or may handover to one of the neighboring cells after certain time periods that are exponentially distributed with mean values $1/\mu_c$ and $1/\mu_h$, respectively. Thus, an ongoing call (new or handover) completes service at rate μ_c, and a mobile engaged in a call departs the cell at rate μ_h. Finally, we obtain the total call termination rate in the cell by

$$\mu_T = \mu_c + \mu_h \tag{4.119}$$

Using Little's result [2], for the case of only fresh calls in the observed cell, traffic intensity is $A_1 = \lambda_1/\mu_T$. Using the $M/M/c/c$ queuing system, with channel policy blocked calls cleared (when all channels are busy, an incoming call is lost), we can calculate blocking probability for new calls by using the Erlang-B formula:

$$P_{B_n} = \frac{\frac{A_1^c}{c!}}{\sum_{i=0}^{c} \frac{A_1^i}{i!}} \tag{4.120}$$

Carried traffic in the cell may be calculated by

$$Y_1 = \frac{\lambda_1}{\mu_T}\left(1 - P_{B_n}\right) \tag{4.121}$$

However, we must consider incoming handover calls in a cellular environment, so the simple Erlang-B formula cannot be directly applied. Because of the memoryless property of exponential distribution, one can assume that the handover process is also a Poisson process with mean intensity λ_h, so that the total call arrival intensity will be

$$\lambda_T = \lambda_n + \lambda_h \tag{4.122}$$

Effective offered traffic to a cell for a generalized case is

$$A_e = \frac{\lambda_T}{\mu_T} = \frac{\lambda_n + \lambda_h}{\mu_c + \mu_h} \tag{4.123}$$

The handovers are not independent processes. They depend upon the new call arrivals in the cells of the mobile network. If we denote with P_B the overall blocking probability in a cell (including new calls blocking and handover blocking), then the carried traffic in the cell is

$$Y = \frac{\lambda_T}{\mu_T}\left(1 - P_B\right) \tag{4.124}$$

Assuming equilibrium in a cell, handover intensity from the observed cell to its neighboring cell is equal to incoming handover intensity to that cell from other cells [7]. We may write

$$\lambda_h = \lambda_n \left(1 - P_B\right) \frac{\mu_h}{\mu_c + \mu_h} \cdot P_B \tag{4.125}$$

Using the last two equations, after some simple algebra, we obtain the effective offered traffic to a cell:

$$A_e = \frac{\lambda_n}{\mu_c + \mu_h P_B} \tag{4.126}$$

If the handover rate from a cell is similar to or lower than call completion rate, and blocking probability is small, then we may approximately calculate the effective offered traffic to the cell by using $A_e \cong \lambda_n/\mu_c$. If there is no reservation of channels for handovers, new call blocking probability P_B will be equal to handover call blocking probability and can be calculated by using Erlang-B formula with effective offered traffic:

$$P_B = P_{Bn} = P_{Bh} = \frac{\dfrac{A_e^c}{c!}}{\displaystyle\sum_{i=0}^{c} \frac{A_e^i}{i!}} \tag{4.127}$$

where c is the number of channels in the cell. One may solve the last equation for P_B by considering (4.126) and using iterations.

4.8 Principles of Dimensioning

The basic principle in telecommunications networks is balance of *quality of service* (QoS) requirements against economic costs. There are several measures to characterize the services. The most important is QoS, which includes different aspects of a connection such as delay, loss, and reliability. The subset of these measures, which refers to the traffic performances, is called *grade of service* (GoS). We have already used these terms throughout the text. Usually, GoS is considered as blocking probability in the network. To have a well-functioning system, we should keep blocking at lower values (e.g., GoS ≤ 1%). Let us denote GoS by using the Erlang-B formula for loss systems. To observe the constraints on the GoS, it is convenient to apply the Erlang-B formula to systems with different numbers of channels and different traffic loads, as shown in Table 4.1.

Analyzing the data from Table 4.1 we may notice two different trends:

1. Large groups (e.g., $n = 100$ channels) have higher utilization α of the resources (channels) compared to smaller groups (e.g., $n = 10$ channels).
2. The blocking probability due to overload increases faster for large groups compared to small ones.

The economic cost of the system is related to the amount of resources and their utilization. Low utilization means bad economy. Also, losing traffic

Table 4.1
Offered Traffic for a Given GoS ≤ 2% and Average Channel Utilization at Normal Traffic Load and at Overload of 30%

N	1	10	50	100
A [Erlang]	0.02	5.08	40.24	87.95
$E_{1,M}(A) \leq 0.02$				
$\alpha = Y/N$	0.0196	0.4978	0.7887	0.8619
$A' = 1.3*A$ [Erlang]	0.026	6.60	52.31	114.34
$E_{1,M}(A')$	0.0253	0.0634	0.1306	0.1626
$\alpha' = Y'/N$	0.0253	0.6182	0.9096	0.9575

reduces revenue. We want to optimize the system to maximize the revenue. To obtain such a relation we need to know the additional traffic that system can carry when we add new resources.

For a group of channels or slots we define an improvement function at offered traffic A and n channels in the system as follows:

$$F_n(A) = Y_{n+1} - Y_n \tag{4.128}$$

$$\begin{aligned} F_n(A) &= A\left[1 - E_{1,n+1}(A)\right] - A\left[1 - E_n(A)\right] \\ &= A\left[E_n(A) - E_{n+1}(A)\right] \end{aligned} \tag{4.129}$$

It is easy to note that the following relation holds:

$$0 \le F_n(A) \le 1 \tag{4.130}$$

The improvement function for different A and n is given in Figure 4.16. We may define the improvement function as $\Delta A/\Delta n$, or approximately by $dA/dn \approx \Delta A/\Delta n$. Let us consider the cost as a function of time. We define income per carried traffic in time by unit price g that corresponds to income from 1 Erlang per time unit (we may arbitrarily chose the time unit). Then, let the cost function of the equipment be $g_e(n)$, a function of the available resources (we consider channels without loosing the generality of the approach). The total cost of the system for a given number of channels n is a sum of the cost of resources and the costs due to lost traffic:

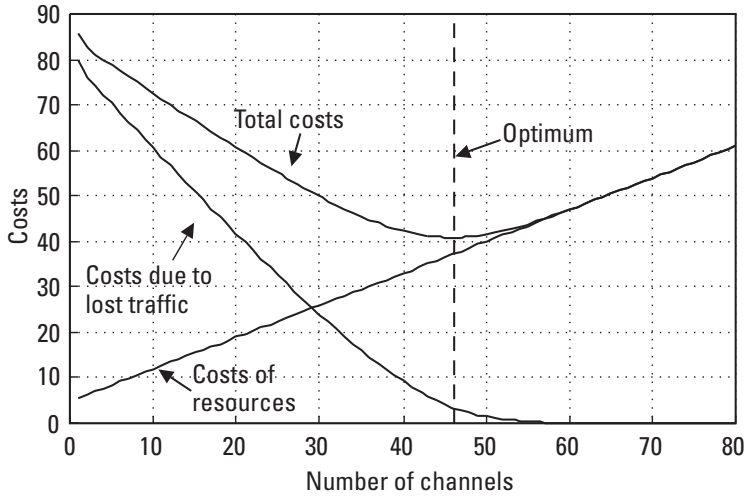


Figure 4.16 Total cost of the system as a sum of resources costs and lost traffic costs.

$$G(n) = g_e(n) + gY_R = g_e + gAE_{1,n}(A) \tag{4.131}$$

For example, let us assume linear cost function for the $g_e(n)$:

$$G_n = G(n) = c_0 + cn + gAE_{1,n}(A) \tag{4.132}$$

If we want to optimize the system we need to minimize the last function, as shown in Figure 4.16. The optimum value of the total cost function $G(n)$ must hold:

$$\begin{aligned} G_n &\le G_{n-i}, i = 1,2,..,n-1 \\ G_n &\le G_{n+i}, i = 1,2,..,N-n \end{aligned} \tag{4.133}$$

where N is number of channels in the system. By combining the last two equations, we obtain

$$\begin{aligned} E_{1,n-i}(A) - E_{1,n}(A) &\ge \frac{c}{gA} i, \; i = 1,2,\ldots,n-1 \\ E_{1,n}(A) - E_{1,n+i}(A) &\le \frac{c}{gA} i, \; i = 1,2,\ldots,N-n \end{aligned} \tag{4.134}$$

Because $E_{1,0}(A) = 1$, there must be at least one value for n that satisfies the relation

$$gA\left[1 - E_{1,n}(A)\right] \ge c \tag{4.135}$$

Also, for some n the following relation should hold:

$$gA\left[1 - E_{1,n}(A)\right] \ge g_e(n) \tag{4.136}$$

If the last equation does not hold, then the system is not profitable and it should not be deployed in such manner. By summarizing, one may conclude that there are two basic reasons for a nonprofitable system:

1. The cost per traffic unit is too low and it should be increased well above the limit given by the last equation.
2. The cost of the system is too high at the time being, so it should be purchased or deployed later.

Besides GoS in the access network, one should consider end-to-end QoS, which depends on the access networks at either side and on the core network. In

circuit-switched networks we usually use more stringent GoS (e.g., GoS≤0.1%) for the interconnection backbone network (e.g., trunks between two telephone exchanges) than for the access networks. The efficiency of the transmission is much higher on the backbone networks due to accumulation of the traffic on a group of trunks, and hence we may provide better quality at lower costs. But, providing end-to-end QoS over IP-based networks (i.e., Internet) is a tough venture, because it introduces problems starting from statistical multiplexing, different applications, different architectures, and different protocols. It becomes even tougher when we have wireless access networks and different access technologies. In Chapter 3, however, we discussed current end-to-end QoS solutions for fixed Internet, such as Differentiated Services, Integrated Services, and MPLS. The QoS approach in IP networks can be divided into: fixed (backbone) network QoS and core network QoS. In this manner, the wireless IP core network is compatible with QoS solutions for the fixed IP network. The gateway of wireless IP networks should provide mappings between the Internet and the core network, as exchanges in circuit-switched networks provide mappings between the channels in the access network and the channels in the backbone interconnection network.

4.9 Discussion

Modeling of telecommunications networks is essential for their design and analysis. In this chapter we covered the queuing theory basics, and we developed the teletraffic theory. It is mainly based on assumption of Poisson arrival processes and exponential distribution of service time. These assumptions usually hold in classical circuit-switched networks. We introduced the famous Erlang-B formula for loss systems with full accessibility. Then, we extended the scenario from one traffic type to multiple traffic types, common for integrated services networks and packet networks. This resulted in the multidimensional Erlang loss formula. Basically, we are interested in mobile/wireless cellular networks. So, we further extended the approach to the mobile environment, where some changes are required due to phenomenon of handover process. Teletraffic theory for wireless networks given in this chapter is targeted to circuit-switched mobile networks such as 2G. However, wireless LANs and 3G mobile networks experience heterogeneous and packet-based traffic. Although modeling of real-time services (e.g., telephony) in these wireless networks can be performed using the traditional teletraffic theory, some services, such as Web browsing and video streaming require different approaches (we refer to them in the following chapters).

Finally, we introduced the fundamental principles in the design of telecommunication networks, which holds either for wireline or wireless networks:

the design of telecommunications networks is based on the balance between the offered quality of service and costs of the system.

References

[1] Lee, Y. W., *Statistical Theory of Communication*, New York: John Wiley & Sons, 1960.

[2] Kleinrock, L., *Queuing Systems Volume I: Theory*, New York: John Wiley & Sons, 1975.

[3] Thomas, J. B., *Statistical Communication Theory*, New York: John Wiley & Sons, 1969.

[4] Mirchev, S. T., *Teletraffic Engineering*, Technical University, Sofia, 1999.

[5] Adan, I., and J. Resing, *Queueing Theory*, Eindhoven University of Technology, February 14, 2001.

[6] Haring, G., et al., "Loss Formulas and Their Application to Optimization for Cellular Networks," *IEEE Transactions on Vehicular Technology*, Vol. 50, No. 3, May 2001, pp. 664–673.

[7] Lam, D., D. C. Cox, and J. Widom, "Teletraffic Modeling for Personal Communications Services," *IEEE Communication Magazine*, Vol. 35. No. 2, February 1997.

5

Characterization and Classification of IP Traffic

5.1 Introduction

IP traffic is fundamentally different compared to classical voice telephony traffic. Therefore, in order to design a mobile IP network, we need to characterize and then classify IP traffic.

There are existing measurements of IP traffic from wired networks. With the introduction of a mobile Internet, however, the services offered on the Internet will not change drastically overnight. Of course, many Internet sites that will target mobile users should offer services adapted to the mobile terminal's display sizes. Low prices of universal mobile communication devices, communicators, mobile phones, and laptop computers promise usage of similar Internet services as those offered to the wire Internet.

The first step towards analysis and design of wireless IP networks is to clearly specify the description of the network traffic and the performances requirements of the applications. A complete and definite specification is problematic for two reasons. First, most of the applications adapt to the facilities provided by the network. Second, it is debatable whether it is appropriate to design a future network using information on the traffic from current applications, due to the possibility of new and radically different applications in the future than those considered today. The approach taken in this chapter, however, assumes that applications posses certain generic inherent properties that result from human behavior and interaction, which are therefore independent of the network infrastructure and are not likely to change in the future.

Furthermore, on the Internet exist heterogeneous services with different traffic and QoS demands. Some services demand real-time communication; others do not have such requirements. Also, there is heterogeneity in the bandwidth needed by traffic flows generated from different applications. Therefore, we need to determine the nature of IP traffic and to consider its characteristics. Traditional techniques in telecommunications for QoS support are based on voice traffic, where resource allocation is deterministic (allocation and switching of channels). Compared to the traditional networks, IP traffic has two main differences:

- Resource allocation is dynamic, and resources are allocated on a per-packet basis.
- There is no explicit support for the allocation of a specific quantity of network resources.

On the other hand, this simplicity of IP networks provides transparent transmission of different traffic types: voice, video, audio, data, and multimedia, over the same network. There is no need to build parallel network infrastructure for each traffic type. These characteristics of IP networks, together with the development of different services offered to the end users, have introduced IP as the main new concept for the telecommunication networks [1, 2].

5.2 Characterization of IP Traffic

Internet traffic, based on IP, includes many multimedia services that have different characteristics considering their traffic parameters: bit rate, burstiness, connection duration time, as well as their demands upon performance metrics (i.e., packet delay, loss, and throughput).

We are interested in network aspects in our analysis of IP traffic. Therefore, we need to capture the behavior of individual streams as well as the behavior of aggregate Internet traffic. Here, we analyze the aggregate Internet traffic by using traces from real traffic measurements. Furthermore, we filtrate different components from the aggregate traffic depending on the applications that generate that traffic.

5.2.1 Aggregate Internet Traffic

The dominant type of service for today's Internet is best-effort service, meaning that equal bandwidth is shared among all traffic flows. All transport control is moved to the end nodes of the communication path. On the other hand, the Internet is initially based on client-server interaction. There are network nodes that provide particular services for Internet users, called servers, as well as

applications running on users' terminals (personal computers and mobile terminals) that demand services. Information streams in digital systems consist of a series of bits: zeros and ones. Network nodes and terminals segment the information stream into packets. Then, we add headers and tails to the packets' information data to include addressing and control information, on the way from application down to the physical medium. In the opposite way we extract headers and tails to provide the information to the target application. Various applications use various transport protocols depending upon their traffic demands (e.g., TCP and UDP). These protocols use sockets to communicate with the application layer. Between transport protocols and link layer protocols on the Internet (e.g., SONET and ATM) we have the IP protocol. Therefore, we reference the aggregate traffic on the Internet as IP traffic or Internet traffic.

In [3] the authors reported measurements from trunks in a commercial Internet backbone over two ranges: 24-hour and 7-day. They captured aggregate Internet traffic as well as traffic per protocol. It shows that Web traffic dominates as the single largest Internet application, with TCP accounting for the most of the traffic: 95% or more of the bytes, 85% to 95% of the packets, and 75% to 85% of the flows. Most of the TCP traffic is actually Web traffic, which dominates as the single largest Internet application, with client-server accounting for more than half of the bytes (65–80%), packets (55–75%), and flows (65–75%) seen on the measured links. Before the invention of the Web, most of the TCP traffic was due to file transfer (FTP), electronic mail, and some interactive applications [4]. After the introduction of the WWW, which is based on *Hypertext Transport Protocol* (HTTP) on the application layer and TCP on the transport layer, Web traffic became dominant in the aggregate Internet traffic composition [5]. So far, all analyses of Internet traffic show that TCP traffic is the dominant one. However, one should expect such results based on the principles of today's Internet, which was created to provide basically one service type (best effort) for all services and does not have proven mechanisms for QoS support. In such a scenario of only best-effort service, one may expect users to prefer applications that are based on reliable protocols at the end-peers of the communication path.

Figure 5.1 shows traffic measurements on a link between an ISP and the worldwide Internet. These measurements show traffic separation upon transport protocol used by the application. The same conclusion for the dominant role of TCP traffic on Internet may be found in other analyses [3, 4].

5.2.2 Internet Traffic Components

We usually classify Internet traffic upon the transport protocol (TCP and UDP) or application (Web, telnet, FTP, or e-mail) used. Furthermore, each of these traffic segments consists of many multiplexed streams from different

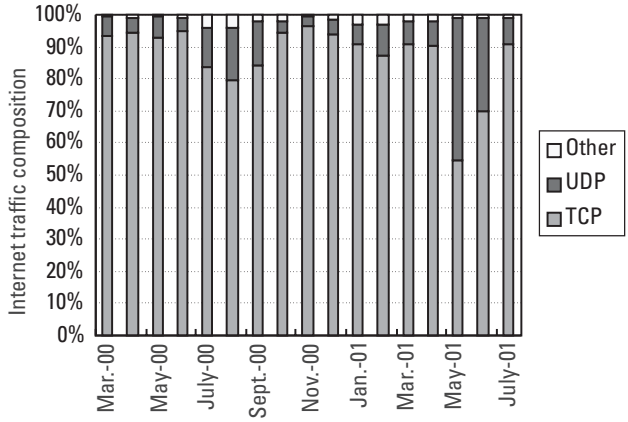


Figure 5.1 Internet traffic analysis on a protocol basis.

connections. One user may initiate one or more streams simultaneously (e.g., parallel connections for one session due to acceleration goals, or more than one session initiated from single browser).

We have shown that TCP is the dominant protocol on the Internet today. Figure 5.2 shows the distribution of aggregate TCP traffic upon application type. According to the given data, WWW accounts for 55% to 90% of the TCP traffic. A smaller segment of TCP traffic is generated from FTP, *Simple Mail Transfer Protocol* (SMTP), and other protocols on the application layer.

Although TCP traffic is dominant on the Internet today, there is also a large segment of UDP traffic. Today, UDP traffic is mainly used for interserver

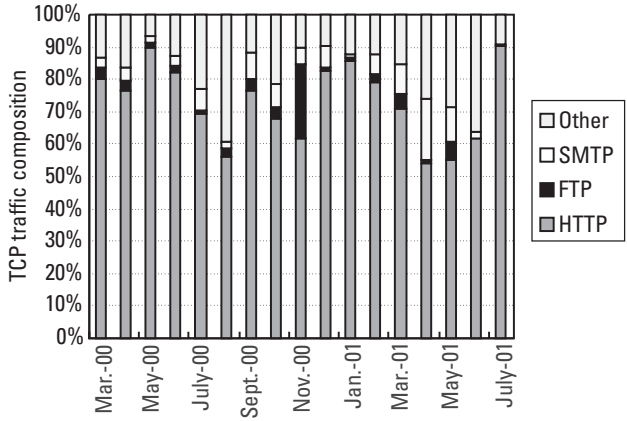


Figure 5.2 TCP traffic analysis.

communication and for *Domain Server Name* (DSN) traffic. UDP is convenient for real-time services and it may be used in combination with the *Real-Time Protocol* (RTP). However, here we need QoS support on the Internet, especially on the access network.

5.3 QoS Classification of IP Traffic

The analysis of IP traffic shows the heterogeneity of the network considering different types of services and applications. The result is a wide range of services with various characteristics and different demands to the network. To provide network design, especially when we have wireless access to the Internet, we need to classify the traffic that exists today as well as the traffic expected to occur on the network in the future. We are going to make classification of IP traffic upon QoS demands from different services.

In Table 5.1 we show services that exist on the Internet as well as services that we expect to exist when QoS support is given. We characterize services based upon:

1. Service type (audio, video, data, and multimedia);
2. Distribution of information.

Table 5.1
Classification of Internet Applications by Information Type, Real-Time Requirements, and Demands for QoS Support

Application	Audio	Video	Data	Real Time	QoS
WWW	—	—	X	2	3
IP telephony	X	—	—	1	1
Multimedia conference	X	X	X	1	1
Audio streaming	X	—	—	2	2
Video streaming	X	X	—	2	2
File download	—	—	X	3	3
Electronic mail	—	—	X	3	3
Multimedia mail	X	X	X	3	2
E-commerce	—	—	X	1	1
Services on demand	X	X	X	2	2

Requirements: 1–high; 2–medium; 3–low.

Table 5.1 does not list all possible services—it is not even possible to do so. However, we consider services with different QoS demands and different types, what seems to be enough to perform classification of the traffic. Today's most common applications on the Internet do not have requirements for real-time service, neither strict QoS support. Examples include WWW and e-mail. These applications use best-effort service, which is the basic service of the current Internet.

Most of the applications given in Table 5.1 are multimedia applications, containing audio, video, and data/images. From the user perspective, one may classify applications in three main groups:

- Interactive applications (e.g., IP telephony);
- Distributive services (e.g., audio or video streaming and Web TV);
- Services on demand (e.g., e-mail, video or audio on demand, and data transfers).

We classify service's requirements based on packet loss, packet delay and delay variation (jitter), and throughput. We approach the problem first through discussions, and then by statistical analysis of traces from real traffic measurements.

Let us look at the interactive applications first. They have very stringent requirements on packet delay and delay jitter. When people are interactive in real time, introduced delay or jitter more than few hundred milliseconds causes a significant impact on the perceived quality of communication. One example is voice telephony over an IP network. According to [6], a delay of 0 to 150 ms is acceptable for telephony; between 150 ms and 400 ms can also be acceptable, but more than 400 ms is not. The total acceptable delay must be divided into a delay budget for each node on the path between the sender and receiver. Speaking in that fashion, other audio and video interactive communications also have very stringent delay and delay variation requirements. Furthermore, losses are not desirable, although limited losses can easily go unnoticed by using error-concealment techniques. The main interactive real-time service, which is telephony, requires low bandwidth due to statistical characteristics of the human voice: it is placed in 3.1 kHz bandwidth (it is narrowband service), there are silent periods between talk spurts (one may apply ON-OFF model for voice sources), and it is predictable. Due to above listed characteristics of telephony—such as sensitivity to packet delay and delay jitter, sensitivity to packet loss (although lower compared to delay), and low bandwidth requirements (compared to other services, such as multimedia)—it is necessary for packets from these applications to enter almost empty buffers. This is possible if packets from IP telephony and similar services are not mixed with other traffic

in the buffers (e.g., TCP traffic). If we put all packets in same buffers, that would break the queuing theory irreparably and in real networks add unmanageable delays and possible losses to the time-sensitive audio data. This is the situation we have today. A simple solution is to place "higher priority" data, such as IP telephony packets, into separate buffers and serve this queue before other data. This would be a priority scheme. It should be mentioned here that many other schemes exist to isolate and protect time, or even loss of sensitive data from interactive real-time applications, but the priority scheme is the simplest one.

On the other hand, services such as e-mail do not have stringent requirements on packet delay and jitter. Reliable transport of information may be made by retransmission of lost packets. Therefore, e-mail should be sent over the link when some resources would be free. Other applications, such as WWW, do not demand low delay and jitter, but they are not tolerant to these parameters as e-mail is. This is because WWW applications are client-server interactive services. From its WWW client, the user sends a request to some server on the network, and then waits for the response. If losses or delay on the network are higher, it will deteriorate the service by causing discontinuity of data transmission and unacceptable delays in the communication (what is the acceptable delay is also more or less a relative issue). Therefore, we may say that WWW services demand higher QoS than the classical best-effort service found on the Internet today. However, best-effort suits well e-mail and scheduled file transfers.

Distribution services, such as audio and video distribution, are rather tolerant to delay and delay variation. Acceptable delays are in the range of several seconds, which depends on the playback buffers in receivers. These delays are higher than the delay thresholds for interactive communication. Loss toleration depends upon type of service. For example, video distribution requires lower losses than videoconferencing. Packet losses reduce video perceived quality because the information is already compressed when it is sent over the transmission link. Video coders use spatial and temporal coding to remove redundancy information within video frames. For example, the widespread standards for video compression and coding are *Moving Pictures Experts Group* (MPEG) and H.261/263 (from ITU-T). Video applications, based on these standards, are widespread on the Internet today. Most video services on the Internet are on-demand. A typical example is the MPEG-4 standard, which supports flexible video transmission: it adapts to the available bandwidth on the link and provides transport of data in the error-prone environment [7]. It is important to note that video applications have the highest bandwidth requirements per connection, as well as the bursty nature of the traffic [8]. Therefore, we do not give the same priority to video services as we give to interactive services, which are less sensitive to QoS and require less bandwidth.

Considering the above discussion on QoS requirements of today's and future Internet applications, as well as the traffic characterization of the Internet (with two main traffic types according to the transport protocol: TCP and UDP traffic), we propose in Table 5.2 a global classification of Internet applications into two main traffic classes:

- *Class-A:* traffic with QoS support, serviced with higher priority;
- *Class-B:* traffic without QoS support, serviced with lower priority.

Within class-A, we further divide the traffic into three subclasses:

- *Subclass-A1:* traffic with highest priority of all;
- *Subclass-A2:* traffic with variable bit rate and support for real-time communication (VBRrt);
- *Subclass-A3:* best-effort traffic with minimal QoS guarantees, but higher than best-effort traffic, which is defined as class-B.

The mapping of Internet applications from Table 5.1 to the proposed traffic classes is given in Table 5.2. Subclass-A1 is the most demanding one, which includes IP telephony, bank transactions, or high-quality multimedia conferencing. It is handled by giving it reserved peak bandwidth, and it is differentiated from other traffic by using priorities. Subclass-A2 traffic has higher QoS constraints on packet loss and delay, but it is more tolerable than subclass-A1. This traffic commonly has time-variable bandwidth demands. Subclass-A3 is proposed for applications with constraint on packet delay, such as Web surfing and immediate file transfers. Class-B does not request any explicit QoS guarantees

Table 5.2
Classification of Internet Applications

Class	Subclass	Flow type	Application
A	A1	Highest priority	IP telephony, videoconference, e-commerce
	A2	VBR real-time	Video/audio streaming, service on demand
	A3	BE-min	WWW, immediate file download, multimedia mail
B		BE (best-effort)	E-mail, scheduled file download

from the network. It is equivalent to the best-effort service model, the basic service model of the current Internet.

We classify the traffic into a limited number of classes, the number of which does not depend on the load of the network or the number of established connections at the moment. Therefore, we do not have scalability problems in the network by adding more IP traffic and increasing the number of flows, because network nodes should store information on QoS parameters per traffic class only, not per flow. To remind the reader, the Integrated Services model for QoS support on the Internet has problems with scalability due to resource reservations for each flow. More likely for existing carriers would be to allocate a part of their bandwidth for this service and through mechanisms such as Differentiated Services provide QoS support. It should be followed by adequate charging model (i.e., higher prices for services with higher QoS requirements). Class differentiation in the wired access network may be done by using the DiffServ model and exploiting the DS field in IP headers. An alternative way is to use differentiation of the traffic by defining other fields in the packet's headers. Because we have a limited number of classes (Table 5.2), and for compatibility with IP standards (IPv4 and IPv6), we should use the ToS field in IPv4 and the DS field in IPv6 for traffic differentiation (refer to Section 3.4.3).

5.4 Statistical Characteristics

For the purpose of our analysis we use traces from traffic measurements. According to the previous discussions, we use traces of aggregate TCP traffic because TCP is dominant on the Internet today. From TCP traces we extract the WWW traffic, which is dominant application on Internet. Also, we extract traces of single WWW connections from the aggregate Web traffic (client-server communication). Besides Internet traces taken from measurements, we also use traces generated from VBR video traffic. We analyze video traces due to the specific character of video information considering the bandwidth requirements (higher than most of other services) and variable bit rate.

For modeling purposes and for analysis, we use the set of traces given in [9]. But first let us define the terms that we are using here. A TCP or WWW trace file is a sequence of rows of data, where each row contains data for a single IP packet, such as: time when packet arrives at the network node that collects the data, IP address (usually they are masked due to users confidentiality), TCP port numbers at both end nodes on the communication path, and length of the information field of the packets (e.g., ACK packets have length zero). These traces have been used many times by the science community [5, 10–12], and therefore can be trusted. However, one should be aware that there are many other trace collections from different networks. We also

use VBR video traces due to the specifics of this traffic. These traces are produced from MPEG coded movies. Video traces are sequences containing the size of each video frame, where frames are generated from video coders every 40 ms when using frame rate 25 Hz (PAL standard, common for Europe), or every 33 ms when using frame rate 30 Hz (NTSC standard, common for North America).

So far we have considered the traffic without specifying the type of access network: wired or wireless. We use this approach because one may expect that at the maturity of wireless IP networks there should be offered all services found in the wired Internet, and ISDN-based services that are offered by commercial telecommunication networks today, either wired or wireless. In this chapter we focus on the analysis of current Internet services because they are less predictable than current commercial telecommunication services, such as circuit-switched telephony, SMS, and other teleservices or supplementary services supported by ISDN standards.

5.4.1 Nature of IP Traffic

One may define the nature of traffic using its statistical parameters and time dynamics. To capture the nature of Internet traffic, we analyze traces from TCP traffic and VBR video sources.

When compared to the voice, data and multimedia traffic are characterized with much less predictability and higher burstiness [13]. For example, many voice streams may be multiplexed over a single link by assigning fixed bandwidth to each stream. We usually reference fixed bandwidth allocations as communication channels. When a new voice call is initiated, the network allocates channels in both directions, to and from the user. Furthermore, we will refer to resources allocated for a single connection as a single channel, but implicitly we should have in our minds that there are always two channels, one for each direction. In circuit-switched networks, other users cannot use a channel that is allocated to a call until that call is terminated or handed over to a neighboring cell (in wireless cellular networks).

Each network node on the communication path should store information on all established connections through that node. Also, all information within a single connection follows the same path through the network from the sender to the receiver. On the other hand, data traffic is characterized with a wide range of time durations of calls, ranging from very short (a couple of seconds) to very long (hours) with high burstiness, various bandwidth requirements (from very low to very high), and with different QoS demands due to heterogeneity of applications and information types. In such an environment, the current Internet provides mechanisms where each packet may be routed through the Internet independently from other packets belonging to the same connection.

Figure 5.3 shows time-dependent activity in TCP traces, which we use in this chapter. Traces are obtained from [9]. We reference TCP traces as *tcptrace1* and *tcptrace2*. Each trace is 60 minutes long. To capture the nature of Internet traffic, we show time dependence of traffic at different time scales (e.g., bytes per 10 ms and bytes per 1 second). To obtain Figure 5.4, we used TCP trace *tcptrace1*. One may notice that traffic in Figure 5.4 is bursty, independent of the time scale. For instance, in Figure 5.4(b) the time-scale is 1,000 times longer (10 seconds), but we notice the same traffic behavior. This multiscale burstiness

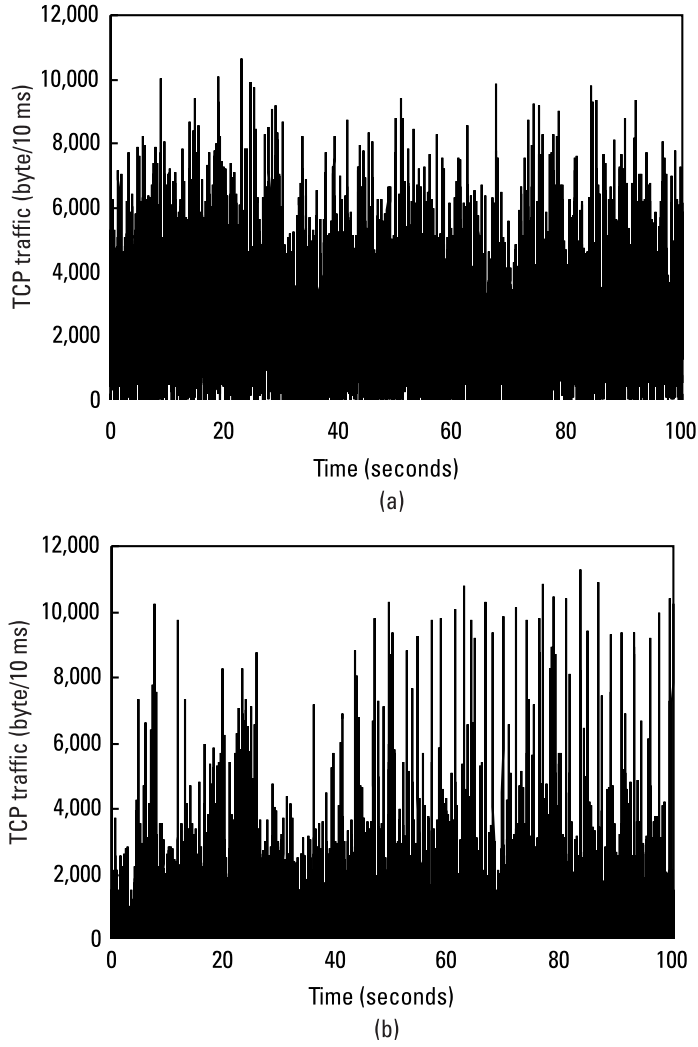


Figure 5.3 TCP traces at time scale 10 ms: (a) *tcptrace1*; and (b) *tcptrace2*.

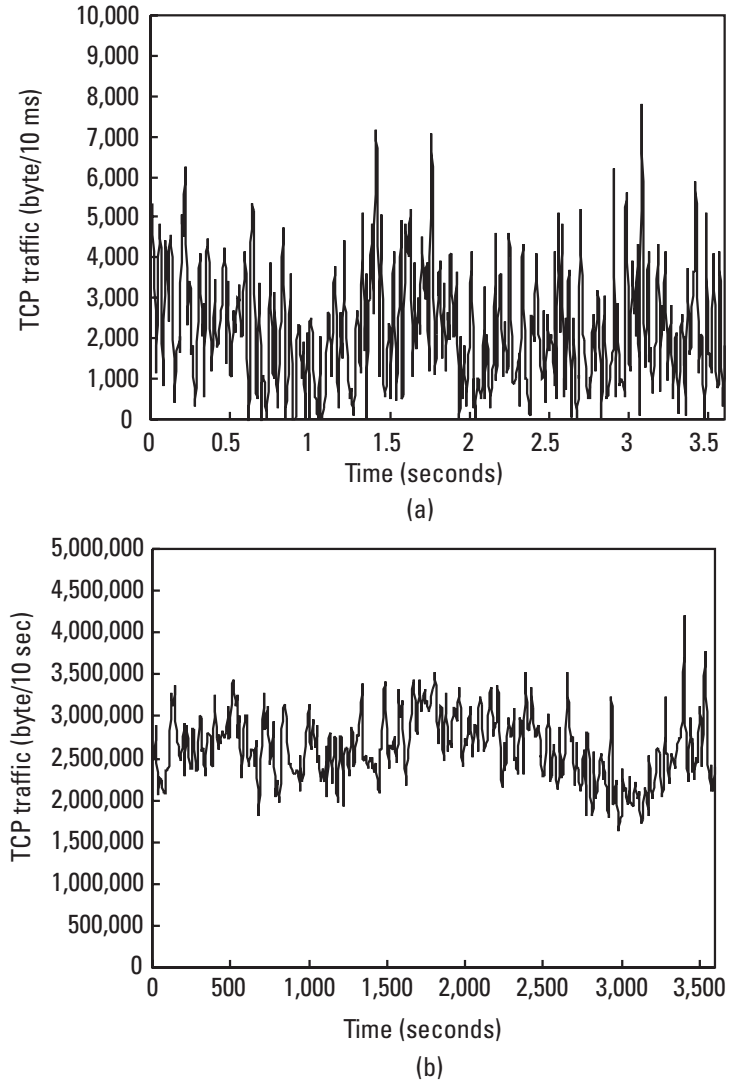


Figure 5.4 TCP traffic at different time scales, *tcptrace1*: (a) 10-ms aggregation periods; and (b) 10-second aggregation periods.

does not fit the traditional Poisson process, which is successfully used for modeling and design of traditional voice-based telecommunication networks. Voice traffic is predictable, while TCP traffic is not. The Poisson process fails to capture the burstiness in the traffic [10, 13].

TCP traffic looks the same (similar) over time scales ranging from milliseconds to hours. Such processes are known as self-similar processes [5, 14] or

fractals (the word is used for the first time by Benoit Mandelbrot to denote a mathematical object whose appearance remains unchanged regardless of the distance from which it is viewed).

Previously we showed that WWW is currently the dominant traffic type on the Internet. Therefore, we extract WWW traffic from the aggregate TCP traces, using the information about the ports at destination to which a packet is addressed (e.g., TCP uses port 80 for WWW applications). In Figure 5.5 we show the extracted WWW traffic from *tcptrace1*. Aggregate WWW traffic is a multiplex of many WWW connections, which are characterized by active periods: surfing the data from the network and downloading Web pages and images; and passive periods when user absorbs the information from the content by looking, hearing, or reading the contents.

From the aggregate WWW traffic, we may extract traces of individual WWW connections by using analysis of IP address in packet headers. Time dynamics of individual WWW connections are shown in Figure 5.6 (it is usually server-client communication). One may notice different traffic intensity of the WWW connections. The second characteristic of WWW traffic is that packet length is usually a multiple of 500 bytes. This may be explained by the typical size of TCP segments of 512 or 536 bytes, because Web traffic is utilizing TCP on the transport layer.

For analyses of real-time video transmission, we use two VBR video traces, obtained from MPEG coded movies. For the analyses, we use hour-long traces, obtained from the movies *Armageddon* and *The Truman Show*. We reference these traces by *video1* and *video2*, respectively. Time diagrams for both

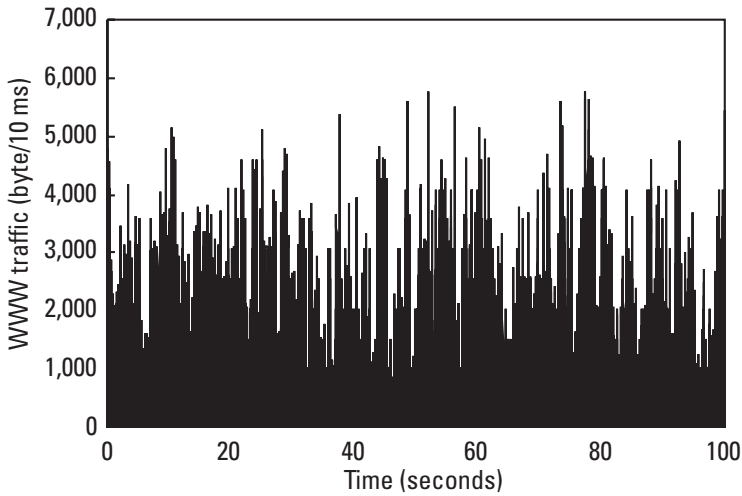


Figure 5.5 WWW trace extracted from *tcptrace1*.

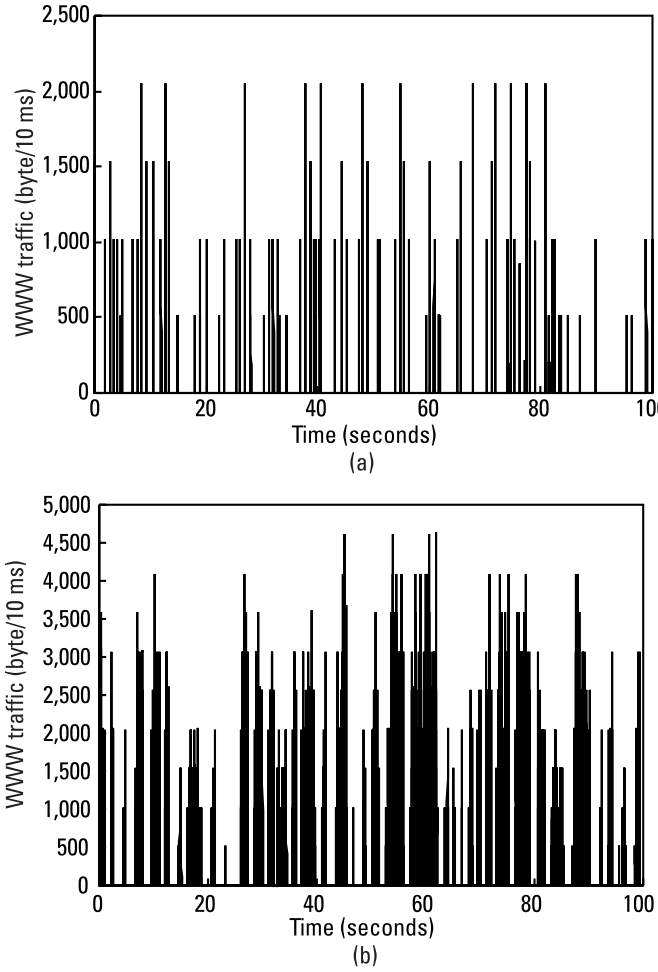


Figure 5.6 Single WWW connections, extracted from aggregate WWW traffic by random choice: (a) WWW flow with lower intensity; and (b) WWW flow with higher intensity.

sequences are given in Figure 5.7. One may notice a bursty nature of the video traffic, similar to that observed at TCP and WWW traces, for aggregate traffic as well as individual connections. The burstiness in video stream is result of the content changing, from one frame to the next one. For example, MPEG coding includes different types of video frames such as intraframe coding, based on entropy coding, and frames that additionally include motion compensation to previous or next frames. The different ways of coding result in different traffic characteristics for different frame types. So, video traffic may be also considered as bursty and therefore we may use self-similar processes to describe it [15].

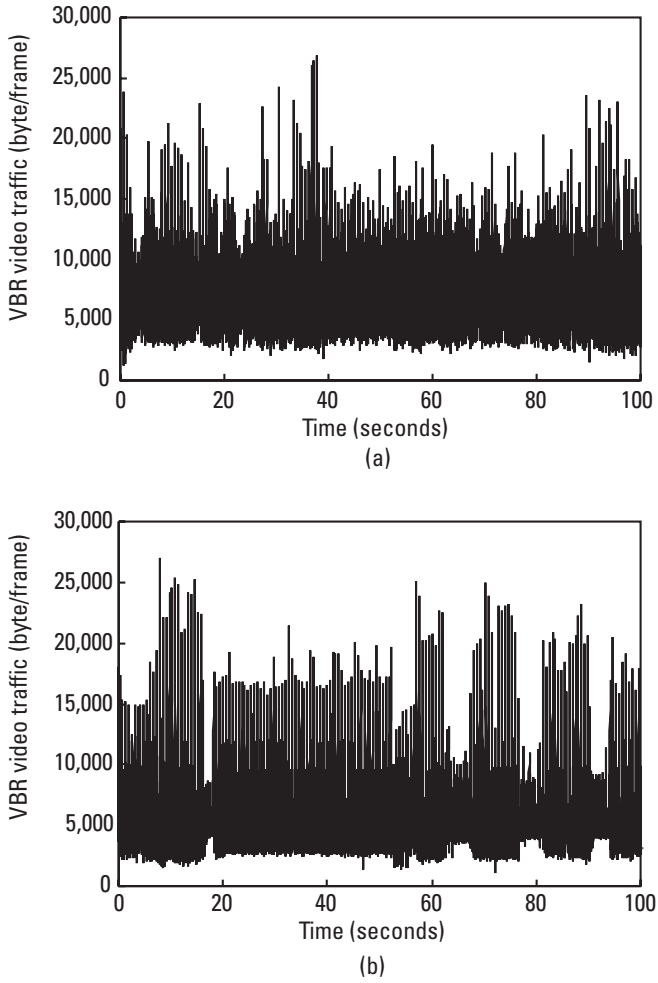


Figure 5.7 VBR video traces: (a) *vbrvideo1*; and (b) *vbrvideo2*.

The analyses of the traces show that TCP, WWW, and VBR video are statistically self-similar by nature. Self-similar processes are often used for traffic modeling of packet networks. In the next section we focus on self-similar processes and their properties.

5.4.2 Self-Similar Processes

We demonstrate that Internet traffic and real-time traffic are self-similar, that none of the commonly used models is appropriate to capture its behavior. First, let us define self-similarity. Traffic processes are said to be self-similar if they look

qualitatively the same irrespective of the time scale from which we look at them. In the case of Internet traffic or VBR video, self-similarity is manifested in the absence of a natural length of a burst; at every time scale ranging from a few milliseconds to minutes and hours, bursts consists of bursty subperiods separated by less bursty subperiods [14]. Some authors use the name fractals to refer to processes with self-similar properties. Below we give mathematical and statistical properties of self-similar processes. Overall conclusions for IP traffic so far are based on intuition (i.e., observed plots on different time scales look intuitively very "similar" to one another). Besides this intuitive property, fundamental properties of self-similar stochastic processes are:

- *Long range dependence* (LRD) and long-tailed distribution;
- Slowly decaying variance.

Let X be a wide-sense stationary process in the discrete time domain, defined as $X = \{x_t;\ t = 0, 1, 2\ldots\}$. The process has a constant mean value $\mu = E\{x_t\}$, variance $\sigma^2 = E\{(x_t - \mu)^2\} < \infty$ and an autocorrelation function $r(k) = E\{(x_t - \mu)\ (x_{t+k} - \mu)\}$, $k = 0, 1, 2,\ldots$.

Traffic processes that are used in teletraffic literature to model voice traffic are exclusively *short-range dependent* (SRD)—that is, they exhibit autocorrelations $r(k)$ that decay exponentially fast [16]:

$$r(k) \sim a^{|k|}, \text{ as } |k| \to \infty \tag{5.1}$$

where $0 < a < 1$ is a constant. Here and henceforth, "$\sim$" denotes that the expressions on both sides are asymptotically proportional to each other. However, the data and multimedia traffic turned out to differ drastically from voice traffic. Statistically, temporal high variability (or burstiness) in traffic processes is captured by long-range dependence (i.e., autocorrelations that exhibit a power low decay). Considering the Internet packet traffic, autocorrelation is slow decaying in the following form:

$$r(k) \sim |k|^{-\beta}, \text{ as } |k| \to \infty \tag{5.2}$$

where $0 < \beta < 1$. Autocorrelation function of a self-similar sequence tends to be the same on different time scales. Analytically, m-accumulated time sequence $X^{(m)}$ is defined by

$$X^{(m)} = \left\{X_k^{(m)}; k = 1,2,\ldots\right\} \tag{5.3}$$

where $X^{(m)}$ is a sequence of samples (e.g., traffic in bytes) generated by summing sample blocks with size m of the original sequence:

$$X_k^{(m)} = \frac{1}{m} \cdot \sum_{i=km-m+1}^{k \cdot m} X_i \tag{5.4}$$

If the time sequence is self-similar, the autocorrelation function of the aggregated process $X^{(m)}$ is equal to the autocorrelation function of the original process for all values of m. One may say that self-similar processes have the same autocorrelation function on all time scales. For nonself-similar processes, it is valid that

$$r^{(m)}(k) \to 0 \text{ as } m \to \infty \text{ for } k = 0,1,2\ldots \tag{5.5}$$

Because self-similar processes are defined by their first and second moment (mean and variance), we may find in the literature the following phrase: "exactly second-order self-similar process" [14]. The process X is called exactly second-order self-similar if the corresponding aggregated processes $X^{(m)}$ have the same correlation structure as X; that is,

$$r^m(k) = r(k);\ k = 0,1,2\ldots;\ m = 1,2,\ldots \tag{5.6}$$

So, a process is self-similar if aggregated processes are identical with X at least with respect to their mean values and variances (second-order statistical property). The nature of such a process is described by the self-similarity parameter $H = 1 - \beta/2$.

However, the last relation usually is not exact for real-time series. If $r^{(m)}(k)$ agrees asymptotically (i.e., for large m and large k) with the correlation $r(k)$ of X, then X is called an asymptotically second-order self-similar process. An example of the asymptotically self-similar process is *fractional AutoRegressive Integrated Moving-Average* (fARIMA) [14].

5.4.2.1 The *H* (Hurst) Parameter

An attractive property of the self-similar processes for modeling the time series of IP traffic is the degree of self-similarity, which is expressed with a single parameter. Considering the relation (5.2), the parameter expresses the speed of decay of the series autocorrelation function.

But initially, the H parameter and self-similar processes are not introduced for the analyses of telecommunication traffic data. Hurst first discovered this property by investigating the amount of storage required in the Great Lakes of the Nile river basin [17]. He found that the expected value of the quality $R(n)/S(n)$ asymptotically followed a power law:

$$E\left[R(n)/S(n)\right] \approx cn^H,\ n \to \infty \tag{5.7}$$

where c is a positive constant, $R(n)$ is the adjusted range of the samples observed (in our case they are traffic samples expressed in bits), $S(n)$ is the sample standard deviation, and H is the Hurst parameter with range $0.5 < H < 1$.

If we denote with X_i the sequence of the samples, then the rescaled adjusted range $R(n)/S(n)$ may be calculate by using [15]

$$\frac{R(n)}{S(n)} = \frac{\max(0, W_1, W_2, \ldots, W_n) - \min(0, W_1, W_2, \ldots, W_n)}{S(n)} \tag{5.8}$$

where

$$W_k = \sum_{i=1}^{n} iX - k\overline{X(n)}, \quad k = 1,2,\ldots,n \tag{5.9}$$

In order to calculate the H parameter, we need to calculate $R(n)/S(n)$ for different values of n. Then, we need to plot a diagram where $\log(\mathrm{E}[R(n)/S(n)])$ is plotted on the y-axis and $\log(n)$ is plotted on the x-axis. We calculate the H parameter using linear regression for the estimation of the parameter:

$$H = \frac{\log\left(E\left[R(n)/S(n)\right]\right)}{\log(n)} \tag{5.10}$$

The above-presented method for estimation of the H parameter is called the R/S method. In addition to R/S analysis, other methods can be used to estimate H such as variance time and periodogram analysis. The value of H is in range (0.5,1). For *independent identically distributed* (i.i.d.) processes, the H parameter approaches 0.5, while for computer traffic it approaches 1.

The variance time method relies on the slowly decaying variance of a self-similar series. The variance of $X^{(m)}$ is plotted against the aggregation factor m on a log-log plot, and the H parameter is given by $H = 1 - \beta/2$. The periodogram method uses the slope of the power spectrum of the series as frequency approaches zero. On a log-log plot, a periodogram is a straight line with slope $\beta - 1 = 1 - 2H$ close to the origin. However, there are also other methods for calculation of the H parameter, but they are less frequently used.

5.4.3 Statistical Analysis of Nonreal-Time Traffic

We first analyze Internet nonreal-time traffic traces to obtain their statistical properties and to check an assumption on their self-similar behavior such as: slow decaying autocorrelation function, long-range dependence, and/or slow decaying variance.

In Figure 5.8 we show normalized autocorrelations (correlation coefficients) for TCP traces *tcptrace1* and *tcptrace2*. We calculated the correlation coefficient by using a time scale of 10 ms [i.e., each sample of the trace is the accumulated traffic in 10-ms time intervals (time intervals are consecutive and nonoverlapping)]. One may notice slow decay of autocorrelation coefficients with an increase of the number of lags, which are used for the calculation of the autocorrelation. In this case each lag is a time period of 10 ms.

Analysis confirms the long-range dependence of TCP traffic, which is proved by the existence of long tails in autocorrelation functions of traces. The

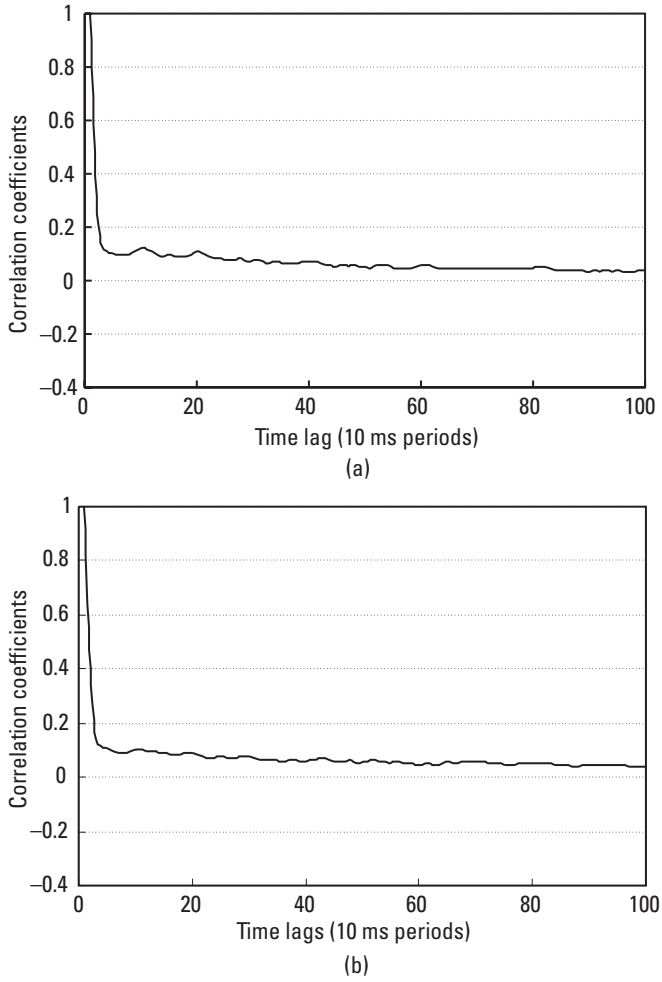


Figure 5.8 Normalized autocorrelation coefficients of TCP traces: (a) correlation coefficients of *tcptrace1*; and (b) correlation coefficients of *tcptrace2*.

same conclusion holds for different TCP traffic intensities: at lower traffic intensity in Figure 5.8(b), and at higher traffic intensity in Figure 5.8(a).

Furthermore, we analyze WWW traffic traces in Figure 5.9. We show correlation coefficients for two traces of aggregated WWW traffic, *wwwtrace1* and *wwwtrace2*, extracted from the TCP sequences *tcptrace1* and *tcptrace2*. We used the same time scale as for TCP traces.

One may notice a periodical component in the WWW traces, which decays for a larger number of lags. We have not noticed such a component in the aggregate TCP traffic. One simple explanation for this phenomenon is the

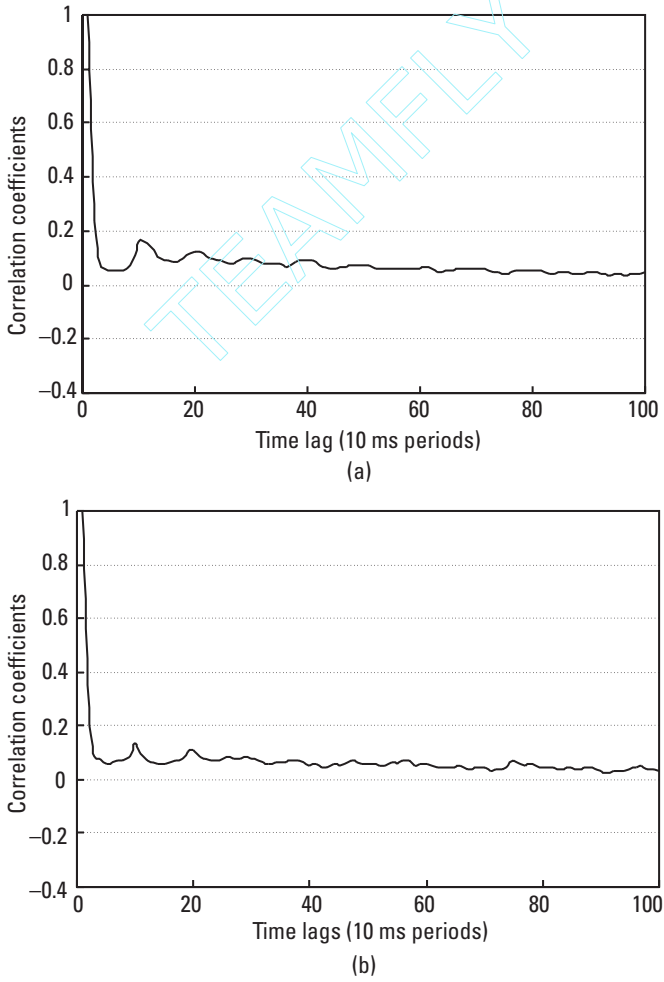


Figure 5.9 Normalized autocorrelation function of WWW traces: (a) correlation coefficients of *wwwtrace1*; and (b) correlation coefficients of *wwwtrace2*.

nature of the WWW traffic. Each WWW session contains active time periods, when user clicks on a link and downloads a page with text and figures, and passive time periods of perception/absorption of the information/contents (reading the text, looking at figures, listening to audio). Active and passive user periods alternate during a single communication. The periodicity of autocorrelation decreases with multiplexing larger number of WWW flows on the link.

Besides the analysis of autocorrelation for aggregate TCP and WWW traffic, we further analyze correlation coefficients of two single WWW traces, each with a duration of 10 minutes, randomly chosen from the aggregate WWW traffic. We may notice fast decay of the autocorrelation function in Figure 5.10(a), something that is a property of SRD processes. In contrast, Figure 5.10(b) shows a slow decaying autocorrelation function of the other WWW connection and a periodical component with period of 9 seconds. One conclusion is that there is no single conclusion from the analyses of WWW connections.

It is difficult to design a network for such a traffic "portfolio." The question is what causes Web-traffic self-similarity. We will refer to this question later in this chapter.

5.4.4 Statistical Analysis of Real-Time Services

We analyzed the best-effort traffic. However, in a network with multiple traffic classes we also need to analyze real-time services, such as IP telephony and audio and video streaming. For voice services we usually allocate a fixed amount of resources, although we may also exploit statistical multiplexing by using IP telephony. Voice conversation is sensitive to delay. Therefore, we should limit packet delay. The best solution to provide low delay to telephony streams is to differentiate IP telephony traffic from the best-effort traffic, as we discussed previously in this chapter.

However, there are mechanisms to classify IP packets from different flows, for example, by using the source and destination addresses and paths for the protocols. This should be done in Internet nodes. Also, mechanisms exist to separate packets from different flows into separate queues in a node (e.g., a router). It is convenient to use a priority scheme to separate delay-sensitive voice traffic from other flows, such as Web traffic. This can be accomplished by applying a DiffServ model in network nodes. In addition, MPLS should be applied in the network domain to support IP telephony.

The problem over IP telephony arises when users are not on the same network (e.g., one user is in America and the other one is in Europe). On the way between the end users, voice packets pass through several hops, which may belong to different network operators. Telephony, by itself, is sensitive because of two main reasons:

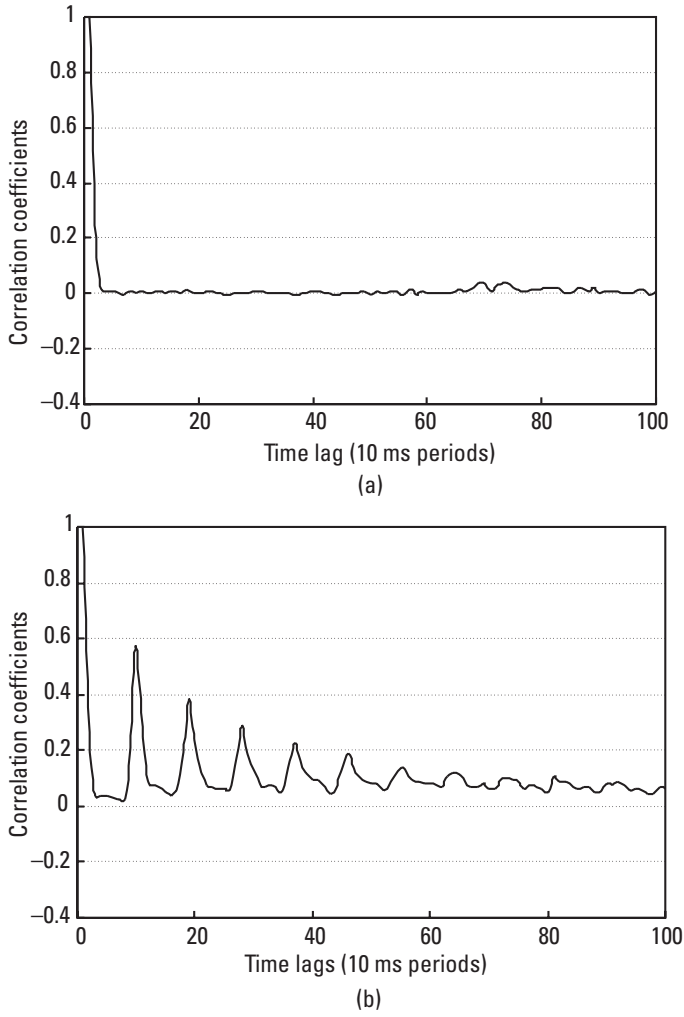


Figure 5.10 Normalized autocorrelation function for individual WWW connections with different traffic intensity: (a) lower intensity; and (b) higher intensity.

1. Telephony (except telegraphy) is the oldest telecommunications service, and it is still the basic service in telecommunications. In a period longer than a century users got accustomed to a certain quality of telephony. They would not tolerate any noticeable degradation on the quality that they are used to.
2. It is interactive conversational real-time service, which is delay-sensitive. ITU-T has specified in the Recommendation G.114 [6] the maximum tolerable delay for one-way voice connection.

Also, there is constraint on packet loss for IP telephony traffic [18]. It is the operator's choice how to design the network considering the IP telephony. Some value ranges for UMTS bearer service attributes may be found in [18].

Wireless IP networks should include a variety of services. Voice service is the basic service that defines a network as a telecommunications network, either wired or wireless. We may refer to all other services as additional services. In 2G mobile networks, such as GSM, additional services are called supplementary services, while services that consider the transmission of voice as an audio, or video telephony, are named teleservices. This nomenclature is compatible with the ISDN concept. One may investigate all possible services at a given moment, but the job will always be unfinished because while analyzing one service set, another service is already considered, recommended, or implemented somewhere. Therefore, we should limit discussions to the most representative services that have influence on the network behavior. If we classify telephony as a basic service, then we may classify video as the most demanding service (e.g., video streaming). Considering video communication, the most common standardized video coding scheme is MPEG. MPEG-1 is used for local video storage retrieval (e.g., on CD-ROM, hard disk); MPEG-2 is convenient for high-bandwidth broadcast (e.g., digital TV) or retrieval video (e.g., DVD); and MPEG-4 is defined for mobile and error-prone environments. (For the sake of completeness, we should mention that there are additional MPEG implementations, such as MPEG-7 and MPEG-21, but they are oriented more to content-based retrieval.)

We choose to analyze VBR video traffic as the third characteristic service—besides telephony and WWW. We show that transmission of video information has similar properties as aggregate TCP or WWW traffic. Similarly to TCP and WWW traces, we consider video traces, which are shown in Figure 5.7. They are bursty over a wide range of time scales. Figure 5.11 shows autocorrelation functions of video sequences *vbrvideo1* and *vbrvideo2*, obtained from MPEG coded movies.

For the calculation of the autocorrelation we use lags equal to interframe periods, which are 40 ms for *vbrvideo1* (frame rate of 25 Hz), and 33.3 ms for *vbrvideo2* (frame rate of 30 Hz).

We used traces, each with 10^4 samples, to obtain the autocorrelation functions given in Figure 5.11. Analyzing the autocorrelation functions of VBR video traces, one may conclude slow-decay. It points to the self-similar nature of the traffic.

Also, strong periodicity is present in the autocorrelation functions of VBR video traces. This is due to the periodical property of MPEG video coding (i.e., the existence of longer referent video frames and smaller spatially or time-dependent frames (with reduced redundancy). Video frames are grouped into *group of pictures* (GoP), which are also periodic. So, we have a different

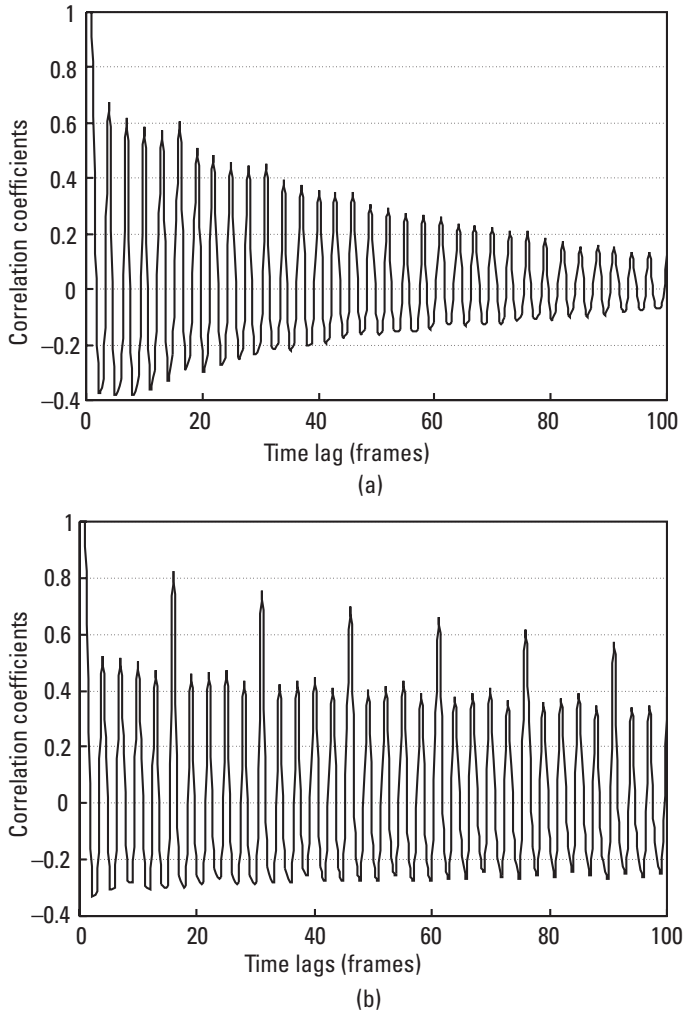


Figure 5.11 Autocorrelation of VBR video traces: (a) correlation coefficients of *vbrvideo1*; and (b) correlation coefficients of *vbrvideo2*.

explanation for self-similarity present in VBR video traffic from the explanation we had for WWW traffic (where it is mainly due to the user behavior).

5.4.5 Genesis of IP-Traffic Self-Similarity

Because self-similarity is believed to have a significant impact on network performance, understanding the causes of self-similarity may be crucial. In [19] it is shown that traffic due to WWW transfers can be self-similar when demand is high. We also showed that a single WWW connection could be LRD as well as

SRD depending upon user behavior. To analyze the heavy-tailed property of the Web, one may do several analyses: on transmission times, silent period tails, distributions of active and silent periods. Transmission times are heavy-tailed due to the character of user time needed for absorption of obtained information. Because active periods are more heavy-tailed than silent periods, it is believed that file sizes in the Web (which determine active periods) are likely the primary cause for Web-traffic self-similarity. One conclusion is that human behavior is the main cause for such traffic characteristic. If so, changes in protocols and document display are not likely to fundamentally remove self-similarity from the Web, although they will influence that characteristic.

The third possible reason for self-similarity of WWW traffic may be found in TCP congestion avoidance mechanisms. The authors of [20] showed that TCP flows have a chaotic nature.

Our analysis of single WWW connections showed that Web traffic self-similarity depends upon user activity, considering separate WWW connections. At lower user activity, which results in lower mean bit rate, we capture behavior similar to SRD processes, which are usually modeled with Poisson processes. In [21] the authors show that aggregate WWW traffic at a lower network load (without losses in the buffers) is also well modeled with the Poisson process. Higher intensity WWW connections show self-similar behavior, LRD, and the periodical autocorrelation function. This periodical component we also found in aggregate TCP traffic, which is also self-similar by nature.

In Table 5.3 we present statistical properties of the following analyzed traffic traces: mean bit rate, peak rate, P/M ratio (peak to mean), and CoV (covariance). Internet traces are analyzed on a smaller time scale (10 seconds) and on longer time scale (1 hour). Considering the statistics, single WWW connections, *singlewww1* and *singlewww2*, have the highest covariance and the highest peak-to-mean ratio. This is due to the on-off (active-passive) character of single connections, with longer off periods than on periods in the aggregate TCP or WWW traffic. Aggregation of traffic lowers the covariance and burstiness. In addition, an increase of traffic intensity causes a decrease of these statistical parameters. The same conclusion holds on different observation time periods (i.e., *tcptrace1* and *tcptrace2*, or *wwwtrace1* and *wwwtrace2*) but also at different traffic intensity (i.e., *singlewww1* and *singlewww2*).

Considering mean bit rates, there is no significant differences on different time intervals (1 hour and 10 minutes). This conclusion usually holds in cases where observation periods are in the same time scale as the connection duration.

We defined the Hurst parameter as a measure of self-similarity. Table 5.4 provides values of H parameter for the analyzed traffic sequences. Calculations of H are performed using R/S methodology (which is described in Section 5.4.2). Single WWW connections have a smaller H parameter. Aggregate traffic sequences have higher H values, closer to 1 then to 0.5 (H = 0.5 for i.i.d.

Table 5.3
Statistical Parameters of the Traffic Sequences

Sequence	Mean Rate (Mbps)	Peak Rate (Mbps)	P/M	CoV
tcptrace1 (1 hour)	2.103	10.184	4.84	0.741
tcptrace1 (10 minutes)	2.174	9.344	4.30	0.713
tcptrace2 (1 hour)	1.007	10.576	10.50	1.255
tcptrace2 (10 minutes)	0.812	9.344	11.51	1.373
wwwtrace1 (1 hour)	0.338	6.552	19.38	1.887
wwwtrace1 (10 minutes)	0.410	5.936	14.47	1.685
wwwtrace2 (1 hour)	0.096	8.192	84.86	3.695
wwwtrace2 (10 minutes)	0.086	6.554	76.29	3.550
singlewww1 (10 minutes)	0.00467	2.458	526.16	14.616
singlewww2 (10 minutes)	0.02414	3.686	152.69	8.529
vbrvideo1	1,412	6.212	4.40	0.575
vbrvideo2	1,564	8.605	5.50	0.785

Table 5.4
H (Hurst) Parameter for Different Traffic Traces

Sequence	*H* (Hurst) Parameter
tcptrace1	0.97
tcptrace2	0.93
wwwtrace1	0.88
wwwtrace2	0.85
singlewww1	0.68
singlewww2	0.77
vbrvideo1	0.98
vbrvideo2	0.95

processes). One may conclude that self-similarity in Internet traffic increases with the aggregation of traffic flows. The smallest H in Table 5.4 is for single connections, while H is highest for aggregate TCP traffic. Within the same time scale, sequences with higher bit rates have higher self-similarity (higher H values). So,

self-similarity of IP traffic increases with the aggregation and intensity of the traffic. In the opposite case, with lower traffic intensity and single connections, traffic is less self-similar.

Video sequences have H close to 1. Hence, VBR video traffic is also self-similar. The H parameter for video traffic is usually in the range of 0.8 to 1.

For the purpose of complete traffic characterization we should also consider marginal distributions, or, in other words, histograms of the sequences. We show histograms of the sequences *tcptrace1* and *vbrvideo1* in Figures 5.12 and 5.13, respectively. For both histograms we may notice slow decay of the histograms towards larger sample sizes. For the TCP traces, each sample is the

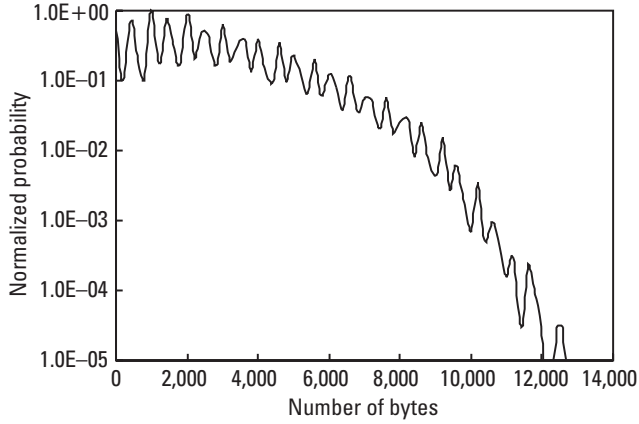


Figure 5.12 Histogram of TCP sequence *tcptrace1*.

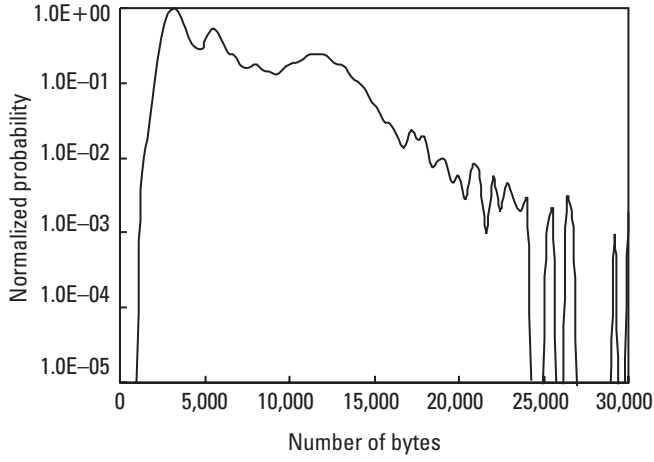


Figure 5.13 Histogram of VBR video sequence *vbrvideo1*.

accumulated traffic in 10-ms intervals. For the *vbrvideo1* trace, we create the histogram by using frame sizes as samples, with 40 ms between adjacent samples. We may notice that the histogram of the VBR trace starts at a value greater than zero because there is no zero bytes video frame size. However, in the TCP case we may have idle periods of 10 ms. WWW traffic has a similar histogram as TCP, but shifted to lower values (Figure 5.14), because WWW is part of the aggregate TCP traffic. In the case of TCP traffic, we notice peaks at sizes that are multiples of around 500 bytes. This is due to typical segment sizes in TCP.

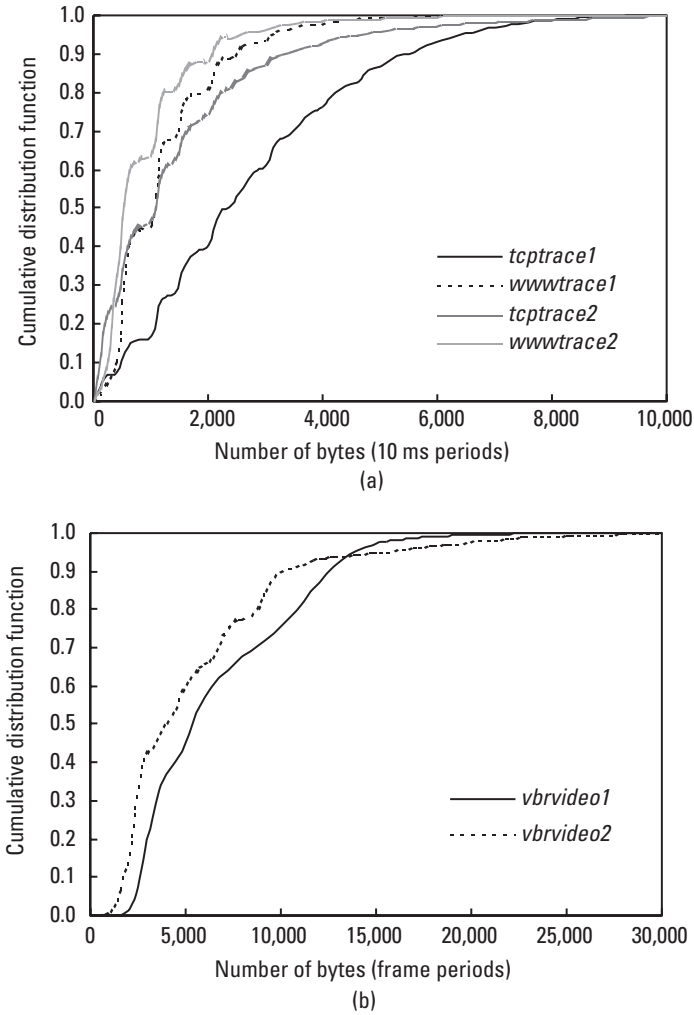


Figure 5.14 Cumulative distribution functions of the sequences used in the analysis: (a) TCP and WWW traces; and (b) VBR video traces.

In Figure 5.15 we show daily, weekly and monthly traffic intensity measured on a link that connects an ISP with the worldwide Internet [22]. In the figures, therefore, we observe aggregated traffic of many user sessions. The ISP has approximately 10,000 subscribers and up to 1,000 simultaneous 56-Kbps dial-up connections as well as a dozen 2-Mbps leased lines to connect several private LANs to the Internet. First, on daily basis we may notice the highest user activity late at night. This is because most of the users are browsing during the evenings, when they are at home. This will change in a wireless network environment—peak hours should shift to daytime period. This is strongly dependent upon the charging scheme that is applied in the network. If we look at weekly and monthly statistics, we may find that there is no noticeable traffic fluctuation between different days or even weeks. A possible explanation is the dominance of dial-up connections in the total traffic that creates a situation similar to circuit-switched services, due to the limited maximum bit rate of the users (56 Kbps). Of course, traffic intensity should increase by adding more users to the network.

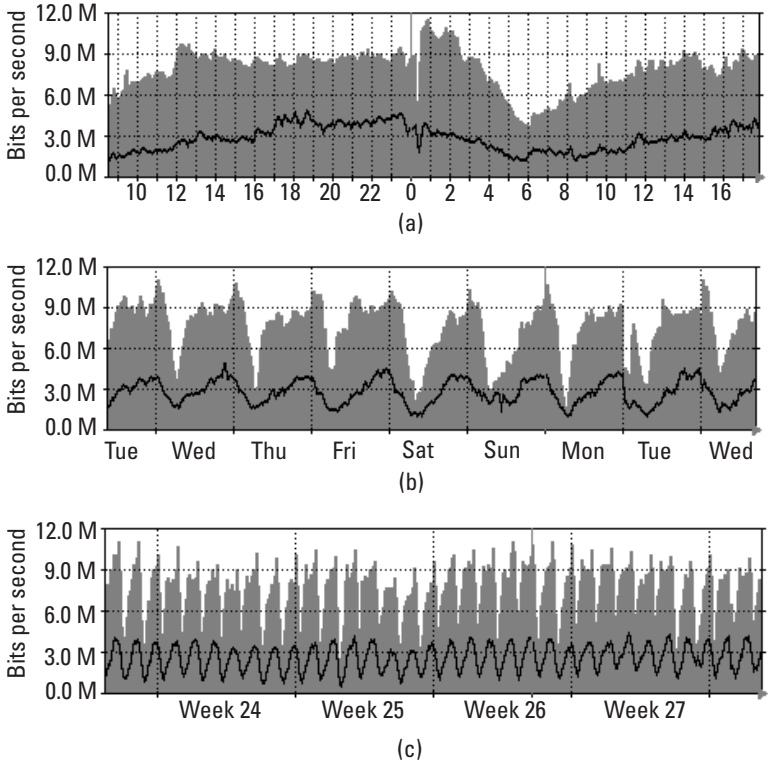


Figure 5.15 Internet traffic measurements: (a) daily, (b) weekly, and (c) monthly.

We may predict resources for the backbone IP network due to small user access rates. So, the main problem that need to be solved is dimensioning of the access network, especially when it is wireless.

5.5 Discussion

We performed statistical analysis of the traffic in the current Internet as well as VBR video traffic. In the aggregate Internet traffic today, the dominant traffic type is TCP traffic, because the Internet is based on best-effort service. Second place belongs to UDP traffic. Its share in the aggregate Internet traffic should increase by adding more real-time services, such as IP telephony, VBR video services, and streaming audio.

Using as metrics requirements on QoS, bandwidth and real-time communication, we classified current and foreseen services into two main classes: class-A, with QoS guarantees on one or more QoS parameters, and class-B, without QoS guarantees, compatible with best-effort traffic in Internet today. Network nodes should use priority to differentiate between the packets belonging to the two classes. IP packets from class-A are always served before any class-B packets. The higher priority class is further divided into three subclasses: subclass-A1 with highest QoS demands, subclass-A2 with high QoS demands (but lower than the previous one), and subclass-A3 for providing improved best-effort service, but without stringent guarantees on the QoS.

We need traffic characterization to provide desired QoS support. From the analysis of traffic traces, we may conclude that TCP/WWW traffic and VBR video traffic are self-similar processes. Because of this, it is impossible to provide quantities on the QoS. Therefore, for services with highest QoS requirements (e.g., IP telephony and video streaming), the most efficient solution is higher priority than the rest of the traffic. It is task of the Admission Control algorithm to determine whether a user request will be admitted to the network, or be negotiated or rejected.

References

[1] NOKIA White Paper, *IP-Radio Access Network, Nokia's Vision for an All-IP Based Architecture for Radio Access Networks*, http://www.nokia.com/networks/.

[2] ERICSSON AB, *3G*, http://www.ericsson.se/3g/.

[3] Thompson, K., G. J. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, November/December 1997.

[4] Claffy, K. C., "Internet Traffic Characterization," Ph.D. dissertaion, University of California, San Diego, CA, 1994.

[5] Feldmann, A., et al., "The Changing Nature of Network Traffic: Scaling Phenomena," *ACM SIGCOMM*, Vol. 28, No. 2, April 1998.

[6] ITU-T, *Transmission Systems and Media, General Recommendation on the Transmission Quality for an Entire International Telephone Connection: One-Way Transmission Time*, Recommendation G.114, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, March 1993.

[7] Zografski, Z., and T. Janevski, "Analyses of MPEG-4 Video Streams Processing in Computer Clusters Based on Multihop ATM Networks," *Fifth International Symposium on Computers and Communications 2000*, Antibes, France, July 4–7, 2000.

[8] Bolot, J.-C., and T. Turletti, "Experience with Control Mechanisms for Packet Video in the Internet," *ACM SIGCOMM*, Vol. 28, No. 1, January 1998.

[9] The Internet Traffic Archive, http://ita.ee.lbl.gov/index.html.

[10] Paxson, V., and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Trans. on Networking*, June 1995, pp. 226–244.

[11] Paxson, V., and S. Floyd, "Why We Don't Know How To Simulate the Internet," *Proc. Winter Simulation Conference*, Atlanta, December 1997.

[12] Paxson, V., "End-to-End Internet Packet Dynamics," *ACM SIGCOMM*, Vol. 27, No. 4, October 1996.

[13] Willinger, W., and V. Paxson, "Where Mathematics Meets the Internet," *Notes of the American Mathematical Society*, Vol. 45, No. 8, August 1998, pp. 961–970.

[14] Leland, W. E., et al., "On the Self-Similar Nature of Ethernet Traffic," *ACM SIGCOMM 1993*, San Francisco, CA, September 1993.

[15] Izquierdo, M. R., and D. S. Reeves, *A Survey of Statistical Models for Variable Bit-Rate Compressed Video*, Center for Advanced Computing and Communications Technical Report 97/10, June 1997.

[16] Erramilli, A., O. Narayan, and W. Willinger, "Experimental Queuing Analyses with Long-Range Dependent Packet Traffic," *IEEE/ACM Trans. on Networking*, Vol. 4, No. 2, April 1996.

[17] Hurst, H. E., "Long Term Storage Capacity of Reservoirs," *Trans. Amer. Soc. of Civil Engineering*, Vol. 116, 1951, pp. 770–799.

[18] 3GPP TS 23.107, *QoS Concept and Architecture (Release 5)*, 3GPP Technical Specification, V5.3.0, 01-2002.

[19] Crovella, M. E., and A. Bestavros, *Explaining World Wide Web Traffic Self-Similarity*, Computer Science Department, Boston University, Technical Report TR-95-015, October 12, 1995.

[20] Veres, A., and M. Boda, "The Chaotic Nature of TCP Congestion Control," *INFOCOM'00*, Tel Aviv, Israel, March 26–30, 2000.

[21] Morris, R., "Variance of Aggregated Web Traffic," *INFOCOM'00*, Tel Aviv, Israel, March 26–30, 2000.

[22] Firfov, O., T. Janevski, and B. Spasenovski, "Modeling the Internet—State of the Art," *ETAI 2000*, Ohrid, Macedonia, September 21–23, 2000.

6

Architecture for Mobile IP Networks with Multiple Traffic Classes

6.1 Introduction

Future mobile networks should support heterogeneous traffic including voice, video, audio, data, and multimedia. QoS provisioning for different traffic types in cellular networks is essential for their implementation as commercial networks. Mobile users are moving among cells in the network, thus changing the access point (i.e., the base stations engaged in the calls). The network should provide transparent and seamless handovers between the adjacent cells for different application types (i.e., services). On the other hand, network operators are interested in having efficient and maximum possible utilization of the resources. This is a trade between the quality of service on one side, and utilization on the other side. The final target is satisfied users and the highest possible revenue for the service providers or network operator.

Various events in a mobile system are happening on different time scales. For example, the connection duration of a call is measured with tens of seconds or minutes, while packet transmission or serving is measured in milliseconds or even microseconds, depending upon wireless link capacity. Furthermore, user movement occurs on different time intervals, measured in seconds, minutes, or hours. This heterogeneity of events needs integration of many different elements in single network architecture. Also, we need to integrate several time scales considering the network design. In a mobile system one should consider the core network (wired part) and the wireless cellular access network (radio network), as well as their mutual functionality. The base stations are wireless access points of the network, which are interconnected via the core network with other nodes,

such as base station controllers (in 2G mobile networks) or the radio network controllers (in 3G UMTS UTRAN [1]), switching/routing nodes, and databases. The most sensitive part is the wireless access network, for two main reasons:

- Capacity in the radio access network is limited because of the limited frequency spectrum that may be used over some geographical area.
- Wireless links have higher error ratio (BER) than wired links, due to fading, multipath, interference.

In this chapter we define general wireless network architecture based on IP and provide a general description and modeling of network nodes, traffic sources, and user mobility. The approach is consistent with 2G network architectures and 3G UMTS proposals by ETSI, but is not limited to them.

6.2 Architecture of Wireless IP Networks with Integrated Services

We define an open, packet-based network architecture that enables integration of voice, data, and multimedia for wireless mobile networks.

Mobile networks of second generation successfully provide voice service with desired QoS (GoS), as well as some supplementary services (SMS, voice mail, call waiting, conference call, closed user group, caller/calling line identification, and call forwarding). The emerging 2G+ and 3G mobile networks are being introduced with useful capability for data and higher-speed services. At the same time, wireless local area networks, such as WLAN, IEEE 802.11, HIPERLAN from ETSI, are evolving. Also, high-speed wireless networks for home environment (e.g., IEEE 802.16) as well as personal wireless communication (Bluetooth and PANs) are being developed [2]. Thus, there are multiple wireless networks with multiple functions and different interfaces. The future mobile network should have an architecture where multiple networks are able to function in such a way that interfaces are transparent to users and services (Figure 6.1). That will be a paradigm change. One thing is clear, however, and that is the technology for the integration of this multiservice and multinetwork environment. It is the IP technology, because it is the best integration technology for different types of services and it is spread worldwide. Of course, there should be the possibility of having non-IP segments in the IP connection paths. In the existing generation of mobile networks 2G, in 2G+ and 3G, we have a circuit-switched air interface for real-time services, such as voice, although 3G provides packet modes within physical channels for latency-insensitive data. To provide maximum statistical multiplexing (i.e., higher utilization of the wireless

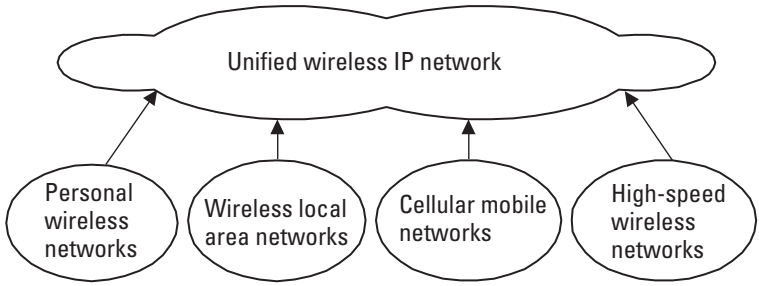


Figure 6.1 Unified wireless IP network.

link) while maintaining heterogeneous services over the same link, we should use native IP over the wireless link that uses IP switching concepts and technologies over the air for all services.

True IP wireless capability has much greater availability to resources and to traffic and service needs.

6.2.1 Network Architecture

We address a native wireless IP network. The architecture of such a network is shown in Figure 6.2. The access network consists of routers/switches, which are interconnected via a core network. End routers have wireless access capability. Of course, the access process and user terminal/network node communication, as well as node/node communication, should be transparent to mobile users.

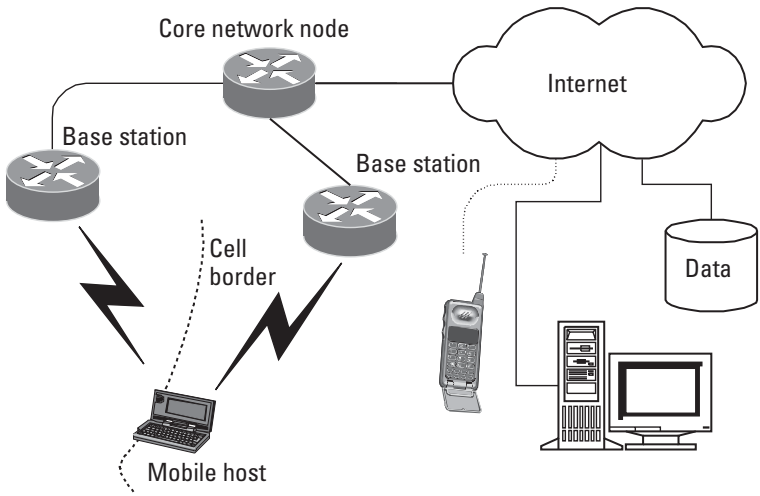


Figure 6.2 Wireless IP network architecture.

We refer to end routers as base stations. They should have integrated functions for mobility management, radio resources management, location management, and routing/switching of IP packets by using appropriate mechanisms for admission control, traffic control (e.g., classification, scheduling), and mobility control (micromobility and macromobility). Part of these functions may be given to a network node aimed to control the radio network. It is called the base station controller in 2G, and the radio network controller in 3G UMTS networks. Also, switching and routing are usually given to specialized nodes for that purpose. It is the mobile switching center in 2G mobile networks, while in 3G we have circuit-switching nodes (e.g., for voice) and packet-switching nodes (e.g., for nonreal-time data).

Each base station serves users in a particular geographic area called cell. The radio network consists of many cells. A base station may have one or more cells. In the latter case it is called sectorized cell. Usually at a modeling phase, we assume a hexagonal or circle form of the cells. In practice, the form can be anything, depending upon the topography and type of the area (urban or rural). All users have access to the resources in the serving cell. Usually, by definition, a serving cell is the best cell considering downlink signal strength and quality of the link (BER level). To increase capacity, the network operator usually sets the network to make possible allocation of resources from a worse cell, if minimum requirements on the link quality are satisfied. A user demanding a service from the network may be admitted to the network or rejected depending upon the amount of free resources and admission control algorithm.

Users are mobile and they move from one location within a single cell to another or they move among cells, depending upon the type of activity and services that users are using. One user may have one or multiple streams. For analysis purposes we assume that communication connections of different users are independent events. The aggregate traffic consists of flows to/from different mobile users, which may be at different locations in the cell. We have a similar situation for each traffic type. The amount of traffic per class is changing with birth events (i.e., new connections or incoming handovers) and death events (i.e., connection terminations or handovers to adjacent cells). For mobility management in a native IP network there exist several protocols for micromobility (within a network domain) and for macromobility (between domains).

6.2.2 Integrated Simulation Architecture

In order to perform analysis of a mobile IP network, we need to model the system on different time scales. We should capture different events, such as packet generation, connection initiation or termination, and handovers. Simulation models and architectures for packet and wireless networks may be found in [3–5]. In this case we should consider a hierarchical approach, where we

should integrate events on different time scales and from different system aspects. Also, various network elements (core, wireless) have divergent characteristics. In the following sections, we define a conceptual model of wireless IP network architecture. Typical elements that should be modeled include the following:

- *Network nodes:* core routers/switches, base stations, and mobile terminals;
- *Traffic sources:* call connections from different applications and aggregate network traffic;
- *Network links:* fixed and wireless.

We define each of the elements above by its characteristic architecture and parameters according to its role in the network scenario as well as its traffic characteristics.

6.3 Conceptual Model of Network Nodes

If the capacity is hard blocked (i.e., limited by the amount of physical network resources) and it is not dependent upon the network load and the environment, then we have hard capacity in the network. All circuit-switched systems are characterized by the hard capacity. It is suitable, however, for packet-switched systems, when the amount of available network resources is fixed to the amount of hardware. Then, we have a time-invariant network capacity. The utilization of the resources depends upon the traffic types and allocation scheme (e.g., one channel per call, two channels per call, allocation of single channel or multiple channels). When we have hard capacity, if all channels are busy, the call is rejected. This is called *hard blocking*. By blocking we will assume hard blocking, unless specified elsewhere. In this case we may use the Erlang-B formula for network dimensioning (Chapter 4).

A conceptual model of a network router is shown in Figure 6.3. The basic elements of a node are:

- Classifier (a class selector);
- Buffers for each class/subclass;
- Packet scheduler;
- Admission control module (exists in base stations only, or in a designated node for that purpose).

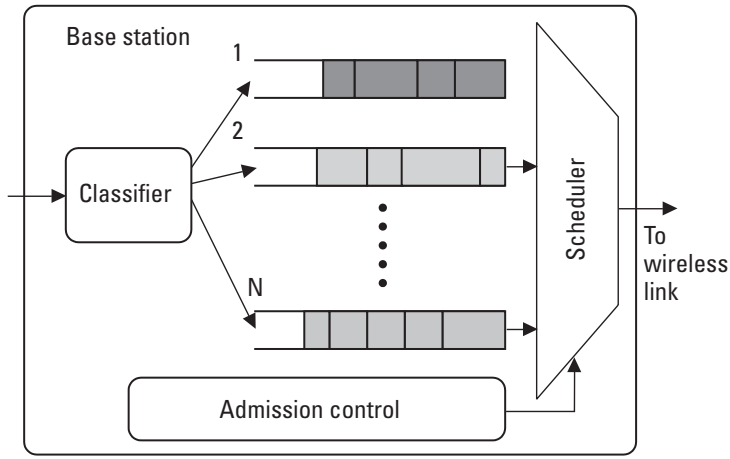


Figure 6.3 Conceptual model of a router in wireless access network.

The class selector performs classification of IP packets. These are 1:*N* elements (i.e., one aggregate flow at the input) and *N* logically separated traffic flows at the output. Each IP packet contains information about its class in the ToS byte (ToS field uses only 6 bits) or DS field, in IPv4 and IPv6, respectively. Packets from different classes are queued into separated logical buffers.

During connection setup, the network and the user terminal negotiate on QoS level. To control the behavior of the flows, base stations should monitor the traffic. According to the monitoring results, packets from nonconformant flows are dropped. Here, we should emphasize the difference between packet dropping due to nonconformance of the flow and packet-dropping due to overflow of the buffers. Packet dropping in the buffers may be realized in different ways by buffer management algorithms. One way is deterministic dropping of all packets over the buffer threshold (drop tail). Another way is to drop packets randomly when the buffer crosses a given load threshold. The latter case may be used to avoid loss of packets in consecutive series that may affect flows within a same class unevenly. In our analysis we assume that all flows are behaving in accordance with the network agreement. Then, losses occur due to buffer overflow only.

By using scheduling algorithms (we will refer to them later in this chapter), network nodes are serving IP packets and routing them to the next node on the path, or to mobile terminals (if the node is base station). An admission control module exists at base stations only. It is suited to perform the scheduling at two hierarchical levels:

- Scheduling of packets from different traffic classes;
- Scheduling of packets from individual flows within a same traffic class.

This separation of the scheduling should provide QoS support in the wireless environment, where we have mobility of users and location-dependent, time-variable bit errors due to the specifics of radio propagation in a real environment.

6.3.1 Scheduling Schemes

In Internet nodes the dominant scheduling discipline is *first-come first-serve* (FCFS) [which is often referred to as *first-in first-out* (FIFO)]. However, this scheduling scheme does not provide isolation of different traffic flows in the network when the flows have various bandwidth requirements and are bursty by nature [6]. For support of real-time applications we should use *priority queuing* (PQ). If we need fair distribution of the bandwidth, we shall use *weighted fair queuing* (WFQ)-like scheduling mechanism. We address each of them in the following sections.

6.3.1.1 FCFS

FCFS scheduling policy is based on a simple principle: The packet that first arrives in the FCFS queue will be first served. FCFS scheduling does not, by itself, provide sufficient isolation between flows. With such a policy, misbehaving aggressive flows can easily starve compliant flows.

The simplicity of FCFS introduces problems for real-time services (e.g., real-time video and audio). When a packet passes several hops, each node adds delay due to buffering (we neglect delay due to signal processing). In a case of aggressive flows (i.e., burst of packets and high data rate), other flows may suffer high delays. So, packets from one flow may significantly impact on delay and losses of packets from other flows, by filling the buffers with its own packets.

6.3.1.2 PQ

PQ is the basis for a class of queue scheduling algorithms that are designed to provide a simple method of supporting differentiated service classes. In PQ packets are first classified by the system and then placed into different priority queues. Packets are scheduled from a head of a given queue only if all queues of higher priority are empty. Within each of the priority queues, packets are scheduled in FIFO order.

Call blocking may occur when admission control is applied. The admission control is necessary to maintain desired QoS to the offered services. It is especially important in the case of real-time services. Call blocking, either new call or handover, is the result of insufficient network resources for serving all user requests. We define the blocking probability as the ratio of the number of blocked calls and total number of call attempts. Thus, the handover blocking

probability of a cell is equal to the ratio of the number of rejected handovers and the total number of handover attempts to the cell. The ratio of the number of dropped calls and the number of all established calls provides the call dropping probability. It is directly related to the handover blocking probability (Chapter 7). In a scenario with multiple traffic classes and services, the bandwidth is shared among the classes. In such a scenario we have performance parameters for each traffic type.

To avoid bandwidth starvation for lower priority service classes, we may use rate controlled PQ. In this case, PQ allows packets from the high priority queue to be scheduled before packets in lower priority queues if the amount of traffic in high priority queue stays under the user-defined threshold. For example, we may allocate 10% of the link bandwidth for IP telephony (high priority traffic) and use the rest 90% of the bandwidth for nonreal-time services. However, it is much easier to provide needed resources for a well-defined application, such as IP telephony, where we know the packet size, traffic volume, and traffic behavior, than it is to provide resources for other applications, such as interactive video, where we have too many traffic parameters.

6.3.1.3 WFQ

The WFQ scheme provides fair scheduling of the packets from different flows in the buffers. It allows isolation of flows by using weights that prevents monopolization of the bandwidth by some flows. WFQ supports flows with different bandwidth requirements by giving each flow a weight that assigns it a different percentage of the link bandwidth.

WFQ supports fair distribution of the bandwidth for variable-length packets by approximating a *generalized processor sharing* (GPS) system [7]. While GPS is a theoretical scheduler, WFQ serves the packets from the flows in round robin cycles, bit by bit. Different logical channels may get different bandwidth allocations of the total link capacity, by using weighted cyclic channel serving [*weighted round robin* (WRR)]. The term "weighted" is used to stress that some channels may carry more than 1 bit in a round. While WRR can be supported in a TDMA network, it cannot be supported by statistically multiplexed network. Hence, WFQ approximates this theoretical scheduling discipline by calculating and assigning a finish time to each packet. In this calculation, WFQ uses bit rate of the link, number of queues (i.e., flows), assigned weights, and the length of each packet in each of the queues. The WFQ scheduler then forwards the packet with the earliest finish time (it is actually the order number of the packet) from among all of the queued packets.

To define WFQ analytically, for each flow i, let us assign weight r_i. If we denote with $B(t_1, t_2)$ the set of all active flows for the time interval $[t_1, t_2]$, when there is no change in the number of flows on the link, the capacity W_i allocated to the flow i should satisfy the following relation:

$$\forall i, j \in B(t_1, t_2), \left| \frac{W_i(t_1, t_2)}{r_i} - \frac{W_j(t_1, t_2)}{r_j} \right| = 0 \tag{6.1}$$

WFQ is a fluid algorithm, which is found to be successful for scheduling in wired packet networks. It provides minimum bandwidth guarantees for each service class or flow. On the other hand, WFQ has high computational complexity, especially when attempting to support a large number of flows on a high-speed link. We may find several different modifications to WFQ. For example, class-based WFQ assigns packets to queues based on user-defined packet classification (e.g., by using IP ToS bits). Afterwards, packets can receive prioritized service based upon user-configured weights assigned to different queues.

In the case of wireless packet networks, however, WFQ fails to provide isolation of different flows. To adapt fair queuing to wireless networks, modifications are needed in the scheduling mechanism.

6.3.1.4 Wireless Scheduling

We consider a wireless IP network architecture. Each base station schedules the packets in uplink and downlink direction. In the downlink direction, logical queues are mapped onto physical buffers in the base station. In the uplink direction the base station maintains a logical queue of all packets that need to be sent, while each mobile terminal queues the packets into its own physical buffers. Also, it is usually assumed that neighboring cells transmit on different logical channels. The characteristics of the wireless channel that influence the scheduling at the air interface, according to [8], are the following:

- The wireless channel capacity is dynamically varying.
- Channel errors are location-dependent and bursty by nature.
- There is a contention on the link among multiple mobile terminals.
- Mobile terminals do not have a notion on the global link state (i.e., they do not know which other terminals have packets to transmit).
- The scheduling must take care of the both directions on the wireless link, uplink and downlink.
- Mobile terminals are often constrained in terms of battery power.

Fluid fair queuing models, such as WFQ, provide fairness among the flows in an error-free environment (i.e., full separation between the flows). Minimum guarantees provided for a flow are unaffected by the behavior of other flows.

To adapt WFQ-like algorithms to wireless IP networks, we have to address two main issues:

- Influence of location-dependent errors, due to mobility of the users and radio propagation characteristics;
- Compensation model for the flows that perceive errors.

Whether compensation can possibly be applied depends upon the type of service (e.g., it is not appropriate for real-time services, only for nonreal-time services).

Wireless fair queuing is important for the wireless link because it handles the flows much better than simple best-effort service (i.e., FCFS). Wireless resources are scarce, and therefore should be utilized to the maximum. Adapted WFQ to a wireless cellular environment within a single traffic class should provide fair and efficient usage of the wireless link bandwidth. We analyze wireless scheduling in more detail in Chapter 11.

6.4 Simulation Architecture for Performance Analysis

For simulation analysis we use the general network architecture shown in Figure 6.2. Network nodes are routers that are capable of processing IP packets. Simulation models should provide analysis of the traffic at a call-level and a packet-level. In the former case, one should specify parameters considering the mobility of the users and network topology, while analysis should produce results on new call and handover blocking probabilities, as well as average number of handovers. Simulation analysis is used to determine or balance QoS offered to users as well as the utilization of network resources. While doing the analysis on a call-level, the information on a packet level is hidden. For the performance analysis on packet-level, we usually use traffic tracing methodology. In this case, a simulation tool traces a single flow from its source to the destination. Internet traffic is asymmetrical (i.e., higher traffic volume is expected towards the mobile terminals in the downlink direction compared to the uplink). Therefore, the downlink direction is more sensitive considering the QoS. The peers of the communication link may be far away from each other. In such cases IP packets pass through multiple hops before they reach their destination. So, IP packets in the downlink direction may have significant delay or delay variation, even before they are scheduled for the wireless link transmission. Also, handover events may cause packet losses in the downlink direction (we refer to handovers in Chapter 10). In the uplink direction, packets originated at the mobile terminals are routed through the serving base station.

According to the previous discussions, for the packet-level analysis we should use tracing of packets through multiple hops (i.e., each packet goes through sequence of routers) (Figure 6.4). In the downlink direction, the destination router is always a base station. For the uplink direction, the base station is the first node on the path (we do not consider ad hoc networks). We assume that all packets follow the same path within the access network domain. Of course, IP packets may have different paths until they enter the observed wireless network domain. Thus, we accept that rerouting (change of the route) occurs only at handover initialization.

We may trace flows from different services. For example, we can trace a video flow in downlink direction (asymmetrical communication), voice conversation (symmetrical communication, uplink and downlink), and so on. To model real network behavior in the simulation, we should multiplex cross-traffic (background traffic) with the observed flow(s). For example, it may be aggregated Internet traffic. We describe traffic models in the next section.

Within the simulation analysis, background traffic multiplexed at a network node input sinks on the same node output. Network nodes perform classification of IP packets and serve the packets with the specified scheduling algorithm.

For flows with variable bit rate some form of traffic shaping is needed to smooth the traffic. Uniform distribution of the traffic maximizes "trunking" gain in the network. An example of such method is token-bucket algorithm [9]. In this method, tokens arrive in the bucket at a rate equal to the admitted bandwidth to that flow.

6.5 Wireless Link Model

Wireless links differ fundamentally from the wired ones. Loss characteristics of wireless medium are time-varying and bursty by nature. One way to model the bit error on the wireless link is by applying a uniform error model, where errors

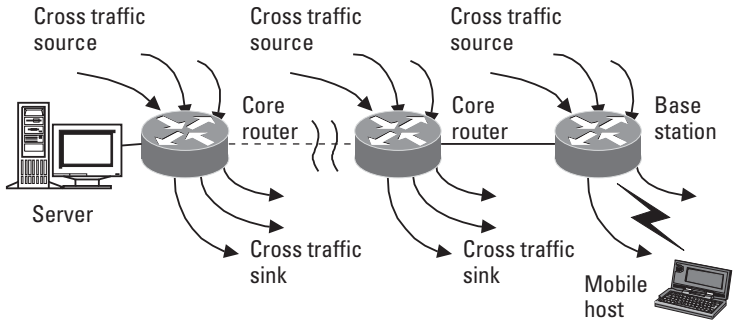


Figure 6.4 Traffic tracing in mobile IP network.

occur continuously in time with some probability. However, the loss characteristics of the wireless channels have been empirically observed to be bursty due to various fading effects [10]. One of the most used models for time-varying wireless link errors is the two-state Markov model.

The Markov error model has two states: error state and error-free state, each having its own distribution. When a channel is in error-state, any IP packets sent would be either lost or corrupted. In the error-free state all packets are successfully transmitted over the wireless link. One should know that this characteristic of the wireless link is associated with a single user, not with all active users in the cell. In other words, each user has its own Markov error model (i.e., some users may be experiencing an error state at a given time interval, while others may have error-free transmission). This effect is a result of location dependence of errors as well as mobility the users. In the Markov model the length of stay in each state can be expressed in terms of transitional probabilities, as shown in Figure 6.5.

We label the error state with E, and the error-free state with F. Let us denote with L_E and L_F mean lengths of error and error-free state, respectively. If the length of each of the states is geometrically distributed, then the transition probability from error to error-free state P_{EF}, and the transition probability in the reverse direction P_{FE}, can be expressed by

$$P_{EF} = \frac{1}{L_E} \tag{6.2}$$

$$P_{FE} = \frac{1}{L_F} \tag{6.3}$$

The transitions between states in the Markov model are memoryless. If we determine distribution for the lengths of the states, then one may calculate the length of staying in a state. For that purpose we need the state leaving probability (e.g., P_{EF} is the leaving probability for the error-state). Thus, if we denote with x a number uniformly distributed in the interval (0, 1), then the length L of staying in a state with leaving probability P is given by

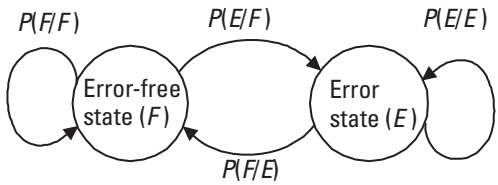


Figure 6.5 Two-state Markov error model.

$$L = \frac{\ln(x)}{\ln(1-P)} \tag{6.4}$$

If P is leaving probability of a state, then $(1 - P)$ is staying probability for that state. Real measurements show that $P_{EF} >> P_{FE}$. For example, measurements of errors on the wireless link, given in [10], show $P_{EF} = 0.3820$, $P_{FE} = 0.0060$, while measurements of errors in a GSM network, given in [11], show $P_{EF} = 0.1491$, $P_{FE} = 0.0087$. We may find in the literature some modifications of the two-state Markov model to better fit real measurements [10]. However, this model is basic and widely used for modeling the errors on the wireless channels.

6.6 Traffic Modeling

For resource planning and dimensioning of networks with multiple traffic classes, we need traffic modeling. At the modeling phase, we need to describe more accurately those parameters that are of interest for the analysis. In that sense, it is not so important to make an exact model of the traffic, but it is more important to model all traffic parameters that influence network performances.

In this section we define traffic models for the wireless IP networks with multiple traffic types. We use the classification of the traffic made in Chapter 5. According to the previous discussions, we separate modeling into two levels: call-level and packet-level. We define traffic models for each traffic class. To analyze the performances by simulation approach, we also need to model the background traffic.

6.6.1 Call-Level Traffic Modeling

Basic parameters for call modeling are call arrival process and call duration. Teletraffic theory for circuit-switched networks, given in Chapter 4, is very successful in dimensioning of traditional telecommunication networks. The Erlang loss formula is still widely used in network dimensioning. Also, it was empirically shown that the Poisson process is appropriate for modeling the call arrivals considering telephony. Traditional teletraffic theory uses the Poisson process for modeling the call arrivals:

$$P(X = k) = \frac{(\lambda \Delta t)^k}{k!} e^{-\lambda \Delta t}, k \geq 0 \tag{6.5}$$

In the above relation, λ is call arrival rate, while $X = k$ is number of call arrivals in time interval Δt. Then, time T between consecutive call arrivals is modeled with exponential distribution:

$$P(T \le t) = 1 - e^{-\lambda t} \tag{6.6}$$

Processes like the Poisson process, which are described with a single parameter (it is arrival rate λ for Poisson), are very important for network dimensioning. Furthermore, it is proven that a superposition of Poisson processes also gives a Poisson process, with arrival rate equal to the sum of the arrival rates of all processes in the superposition. This result allows the Poisson process to be used for trunk dimensioning.

According to the empirical results repeated in many cases [12–14], the moments of initiation of Internet sessions by individual users are also well described by the Poisson process. This may be explained by the nature of human behavior (i.e., each connection starts upon the user decision for it). The same behavior is found in telephone networks and also on the Internet. Compared to the packet-level, call-level analyses are on higher time scales (seconds, minutes, hours). Each communication connection includes transmitting and receiving many packets between the end peers of the communication. Although calls may be modeled with Poisson process, it does not have much impact on the average capacity results. While telephony call duration is well modeled by the exponential distribution, Internet connections are characterized with longer correlation of call/session durations. For each single real-time call we should choose a certain distribution to model the call duration. For modeling real-time call duration we usually use exponential distribution:

$$f(x) = de^{-dx} \tag{6.7}$$

where $T = 1/d$ is the mean call duration. We may use the Poisson process for call arrivals in both cases, either for real-time or nonreal-time services. But, while duration of real-time services (particularly conversational services such as class-A1 traffic) is well suited into exponential distribution, call duration of nonreal-time services shows self-similar behavior. According to [13], Internet connection sizes or durations are well described by using the lognormal distributional family; that is, the distribution of the logarithm of packet sizes (durations) is well approximated with a Gaussian distribution [15].

6.6.2 Packet-Level Traffic Modeling

During an established Internet connection many packets with different sizes are sent and received. In the previous chapter we characterized today's Internet traffic, as well as VBR video traffic, as self-similar by nature. Thus, one should model the self-similarity to provide analytical tools for network analysis and dimensioning. Self-similarity is higher in the aggregate background traffic than in the individual connections.

There are several approaches for modeling the traffic on the packet level. All of them are based on comparison of empirical results and available mathematical models with similar statistical characteristics. In the following sections we define models for each traffic class.

In the high-priority class we need to model IP telephony traffic (subclass-A1). We have a different case with sources with variable bit rate and with real-time requirements (subclass-A2) or nonreal-time services (subclass-A3 and class-B). For modeling self-similar VBR flows, we may use *Markov modulated Poisson processes* (MMPPs), *autoregressive* (AR) processes, Pareto models, and *fractional Brownian motion* (FBM). But, according to [16] the choice of the applied traffic model is not dependent only upon the traffic type of the source, but also upon the characteristics of the system elements such as buffer sizes. Small buffers cannot capture longer autocorrelations and vice versa [16, 17]. There is no unique description of the Internet traffic due to the great heterogeneity of network topologies, protocols, and applications. However, the analysis of buffer utilization in the system nodes upon the Hurst parameter shows that buffer utilization decreases with an increase of the H parameter [18]. Due to the unavailability of appropriate models for a wide range of VBR applications, which have strong self-similarity (i.e., H parameter close to unity), traffic traces are often used for simulation analysis of the system under VBR traffic. If we use traffic traces with higher self-similarity, then we should have at least the same or better performances for traffic with lower self-similarity (i.e., lower H).

For modeling the best effort, we use the definition of the TCP mechanism shown in Chapter 3. The choice of TCP as a typical protocol in the current Internet is justified by the traffic characterization in Chapter 5. TCP traffic should be modeled separately in each direction, uplink and downlink, because mobile terminals are usually clients that demand a service from a server on the core Internet. Data packets are sent on the downlink: At the slow start of TCP, the typical packet size is 1,500 bytes [19], while during the communication it is around 500 or 1,000 bytes in most cases. Acknowledgments and synchronization packets are sent on the uplink from the mobile terminal. The latter are generated at the phase of initiation of a TCP call.

To perform simulation traffic analysis, we also need to model the background traffic on the link. According to the analysis of the histograms of the TCP traces from real measurements (Figure 5.12), we may notice the distribution of the packet length, and according to the analysis results of TCP traces, given in [20], packet lengths may be grouped into three groups: [0,180), [80,180), and [180,$\propto$) bytes. Then, one may use a histogram model for the background TCP traffic (i.e., packet lengths may be generated using the histogram of empirical analysis of the background traffic).

6.6.2.1 IP Telephony Model

Past voice service was mainly based on circuit-switched technology. However, the development of the computer industry and the low cost of communication devices (palm-top devices, communicators, mobile phones, lap-top computers) moved telecommunications beyond just voice service. Within such a scenario, voice will be just one of the many services offered to the end user. It will remain the most used one and the oldest one (except telegraphy). On the other hand, it is almost certain that cellular access networks are going to be based purely on IP, which allows network transparency and statistical multiplexing of different service types. The question is how to design cellular access networks based on IP that will provide desired QoS for voice service.

We assume that voice over IP traffic is differentiated from data traffic, which is based on TCP. If IP telephony traffic is mixed with TCP traffic, which is long-range dependent, then it will add unmanageable packet delays and packet loss. In Chapter 5 we proposed classification of IP traffic into two main classes [21]: class-A, for traffic with QoS constraints, and class-B, for best-effort traffic. Subclass-A1 should be used for IP telephony due to low delay.

Today, mechanisms exist to differentiate traffic, such as differentiated services models. We assume that IP telephony is differentiated from other traffic on the wireless link, and it is not mixed with TCP traffic. Packets from IP telephony are buffered into separate buffers (of course, there are also other mechanisms to bound packet delay or loss). However, we use a priority scheme to differentiate IP voice traffic from the rest.

Single Source Properties

As for traffic models, voice connections arrive according to a Poisson process. Once a connection (or call) is established, the voice source is modeled as two-state Markov chain with one state representing the talk spurt (ON) and the other state representing the silent period (OFF), as shown in Figure 6.6. A simple ON-OFF model accurately models the behavior of a single voice source. During ON (talk) periods the source is transmitting IP packets. Most encoding schemes have fixed bit rate and fixed packetization delay. During OFF (silence)

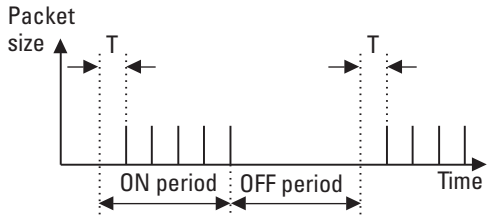


Figure 6.6 Characteristics of a single source.

periods the source sends no packets. We assume that ON and OFF periods are exponentially distributed, which is well analyzed in [22]. The voice sources can be viewed as two-state birth-death processes with birth rate α_{on} (arrival rate for on periods) and death rate α_{off} (ending rate for on periods). Then, $1/\alpha_{on}$ and $1/\alpha_{off}$ are average durations of talk period and silent period of a voice source, respectively. The typical ratio between talk periods and silent periods is 1/2, where the average spurt duration is in the range from several hundreds milliseconds to several seconds.

During talk spurts (ON periods), the model produces a stream of fixed size packets with fixed interarrival times T. Because of the exponentially distributed talk spurts and subsequent OFF periods, the emission of packets can be regarded as a Poisson process.

The Superposition of Independent Voice Sources

The superposition of the voice sources can also be viewed as a birth-death process, where the total incoming rate is the sum of incoming rates of individual sources. A convenient model in teletraffic theory for a superposition of many ON-OFF voice sources is the MMPP. For voice sources with talk spurts and silent periods (without packets on link), it is more convenient to use the special case of MMPP—that is, *Interrupted Poisson Process* (IPP), which is a special case of the Cox process with two phases (refer to Section 4.6.2). When the process is in state j, that means j sources are on. In Figure 6.7 we show the transition-state diagram for superposition of N active voice sources.

6.6.2.2 Packet Traffic Model

The dominant type of traffic on the Internet today is WWW traffic. Therefore, we present a traffic model for WWW flows, which are the dominant nonreal-time traffic. In the packet-generating mode, one browsing session consists of a sequence of packet calls. Packets call correspondents to download from a WWW document (e.g., text page with figures). As we discussed in Chapter 5, after downloading a particular WWW document, the user spends some time for absorption of the information by reading, watching, or hearing. We will refer to this time interval as reading time [23]. The generic model for nonreal-time traffic is shown in Figure 6.8.

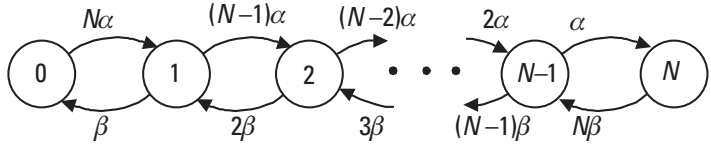


Figure 6.7 Superposition of N voice sources.

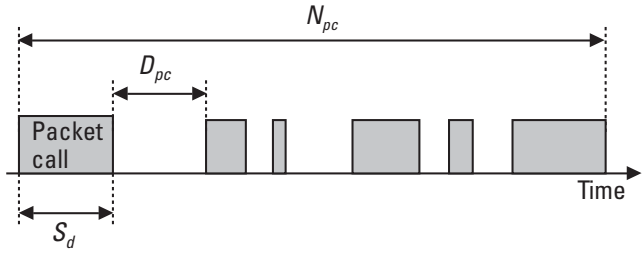


Figure 6.8 Generic model for nonreal-time traffic (e.g., WWW/TCP traffic).

Thus, one WWW session consists of a sequence of packet calls. A user may initiate a packet call by requesting an information entity. During the packet call several packets may be generated. One may say that a packet call is a burst of packets. Hence, for modeling WWW traffic, we can consider the following processes:

1. Session arrival process;
2. Number of packets per session N_{PC};
3. Reading time between packet calls D_{PC};
4. Size of a packet call S_d.

We already agreed to use the Poisson process as an arrival process for nonreal-time traffic; it is also used for real-time traffic. The number of packets per session is well modeled by using geometrical distribution with mean $\mu_{N_{PC}}$.

Also, we may use geometrical distribution for modeling the reading time between two consecutive packet call requests D_{PC} with mean $\mu_{D_{PC}}$. Reading time starts when the user receives completely the last packet of the packet call. It ends when the user makes a request for the next packet call.

For modeling the size of a packet call, Pareto distribution may be used due to its characteristic of having long tails as packet call sizes have. The classical Pareto distribution with shape parameter α and location parameter k has the *probability density function* (pdf)

$$f_x(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}}, x \geq k \tag{6.8}$$

and corresponding cumulative distribution function

$$F_X(x) = 1 - \left(\frac{k}{x}\right)^\alpha, x \geq k \tag{6.9}$$

If $\alpha \leq 2$, then the distribution has infinite variance, and if $\alpha \leq 1$, then it has infinite mean. The Pareto distribution is also referred to as the power-law distribution, double-exponential distribution, and the hyperbolic distribution [12]. Pareto is the only distribution that is invariant under truncation from below. That is, for classical Pareto distribution, for $y \geq x_0$, we have

$$P\left[X > y | X > x_0\right] = P\left[X > y\right] \tag{6.10}$$

Hence, the conditional distribution is also a Pareto distribution with the same shape parameter , but new location parameter $k' = x_0$.

The mean value of Pareto distribution is

$$\mu = \int_{\infty}^{+\infty} xf(x)dx = \int_{k}^{+\infty} xf(x)dx = \frac{\alpha \cdot k}{\alpha - 1}, \ \alpha > 1 \tag{6.11}$$

In reality, however, we cannot have infinite packet call sizes, so we limit the maximum packet call size, which we denote with m. This way we get a cutoff Pareto distribution with mean

$$\mu = \int_{\infty}^{+\infty} xf(x)dx = \int_{k}^{m} xf(x)dx = \frac{\alpha k - m\left(\frac{k}{m}\right)^{\alpha}}{\alpha - 1}, \ \alpha > 1 \tag{6.12}$$

In [24] the k parameter is mapped onto the H (Hurst) parameter as follows:

$$k = 3 - 2H \tag{6.13}$$

Because the range of $H \in [0.5, 1)$, it follows that $k \in (1, 2]$. For WWW traffic we use as a default value $k = 1.1$, what leads to $H = (3 - k)/2 = 0.95$. However, the H parameter of a single WWW session may vary over the whole range [0.5, 1). Therefore, we cannot dimension a network with stringent QoS guarantees for nonreal-time services such as WWW, but we may provide minimum guarantees, if the user demands them. So, we need information about the average packet call sizes over many WWW sessions during the highest network load. Then, we may guarantee the user minimum QoS for WWW services (subclass-A3), or the network may reject the WWW call with QoS demand, and instead offer a best-effort service to the user (class-B). Of course, different pricing schemes should be applied for each traffic class and service.

A summary of typical distributions for modeling WWW traffic is given in Table 6.1. The average session arrival intensity depends on the number of users.

Table 6.1
Random Processes Used for Modeling WWW Traffic

Process (WWW Traffic)	Random Process
Session arrivals	Poisson
Packet call size	Pareto with cutoff
Reading time	Geometric
Number of packets per session	Geometric

Packet call size has the average value of 25 KB [23]. The packet call size can vary between 4.5 KB and 2 MB. The typical mean reading time value is 5 seconds, and there is an average of five packet calls per session. Of course, these values may be different in various situations.

Pareto is not the only process that can be used for modeling self-similar traffic, but it is the simplest one. When we perform traffic modeling, we tend to use the simplest possible models that are dependent upon only a few parameters: (1) the Poisson process for modeling the arrivals, (2) exponential distribution for modeling SRD processes such as voice call durations, and (3) Pareto distribution for LRD processes such as packet call sizes. The first two processes are described by using a single parameter (i.e., mean arrival rate and mean call duration, respectively), while Pareto depends upon two parameters, k and α.

6.7 Mobility Modeling

To perform analysis in a mobile environment, we also need to model mobility of the users. This is important for the analysis of traffic parameters such as call arrivals and handovers. An overview of some of existing mobility models may be found in [25]. There are different models for capturing user mobility, such as the fluid model, the Markov model, and user tracking models.

For example, the fluid model captures the traffic flow as a flow of a fluid. It is appropriate for describing macroscopic user movement. In its simplest form the model formulates the amount of traffic that flows out of some geographical area to be proportional with the population density, the length of the boundary of the area, and the speed of the mobiles. Some authors use the Markov model [26] and the queuing theory in mobility modeling. These models describe individual movement behavior of users. There are specified probabilities for the subscriber to stay within the cell or region or move out of it. We may model mobility by using M/G/m queuing discipline assuming Poisson arrival process of the users and independence of the user cell residence time due to a cell.

The gravity model, for instance, is used to model human behavior in national and international models, and traffic intensity is proportional to the "attractivity" of the regions involved in the movement of the users. In that sense, the factor of proportionality can be specified to have inverse square dependence with the distance between the zones of interest. On the other hand, the mobility traces model records the actual movement of the subscribers. Several mobility user classes are introduced in [27].

In [28] user movement is described using on-off time intervals in the cell with uniformly distributed speeds of the mobiles and their direction, which can be changed at the beginning of each moving time interval. There are different ad hoc methods for mobility modeling because there is no standardized one. However, every mobility model is created for further use in teletraffic modeling.

6.7.1 Macromobility Model

Here we define a general model that considers the tracking of mobile users. To account for user behavior, [29] suggests a generalized model, one that we can use in various mobile scenarios (urban, highways) with a range of variation in the speed and directions of the mobiles in stochastic time intervals.

The mobility model is based on mobile users tracking within the cell [30, 31]. For modeling purposes it is assumed that cells have hexagonal form with side a, and subscribers are uniformly distributed within the cell. In reality, the form of the cell is everything but hexagonal. In our model we approximate hexagonal cell with a circle, with radius R:

$$R^2\pi = \frac{3\sqrt{3}}{2}a^2 \tag{6.14}$$

where a is the size of the hexagonal side.

The position of the user at the initiation of a call is defined with radius r, where r is the distance from the center of the cell (it is the position of the base station in case of omni cell). Considering Figure 6.9 we obtain $dP = 2\pi rdr$ and $dN/N = dP/P$, where N is the number of subscribers in the cell and P is the area of the cell, so that pdf for the user density in the cell is given with

$$f_r(r) = \begin{cases} 2r / R^2, & 0 \le r \le R \\ 0, & r > R \end{cases} \tag{6.15}$$

The direction of user movement within a cell is defined with angle θ, uniformly distributed. So, the probability distribution function for the direction of user movement after call initiation is $f_\theta(\theta) = 1/2\pi$, $0 \le \theta < 2\pi$.

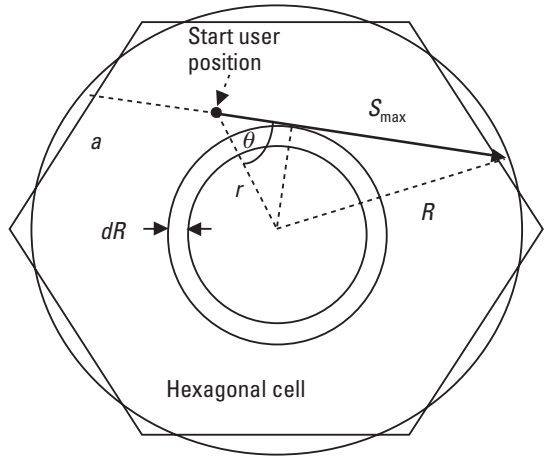


Figure 6.9 Evaluation of the distribution of the subscribers in a cell and user movement definition.

In our model we suppose that the direction and the speed of the mobiles remain constant within one cell; these are allowed to change at handover to another cell. The initial velocity of the mobile stations is assumed to be a random variable with Gaussian probability density function truncated at $v = 0$ km/hr. For this case we introduce a factor k, so

$$k\int_0^{+\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}dx = 1 \tag{6.16}$$

where m is the average speed of the mobiles in the cell. To determine the pdf for the mobiles speed, k should be evaluated. After some operations, we get

$$k = \frac{1}{\frac{1}{2} + \frac{1}{\sqrt{\pi}}\int_0^{\frac{m}{\sqrt{2}\sigma}} e^{-x^2}dx} \tag{6.17}$$

So, for the pdf of the velocity we may write

$$f_v(v) = \begin{cases} k\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(v-m)^2}{2\sigma^2}}, v \geq 0 \\ 0, v < 0 \end{cases} \tag{6.18}$$

In [29] it is shown, using simulation methodology, that cell residence time follows generalized gamma distribution.

In this mobility model the following assumption are made:

- Subscribers are uniformly distributed within a cell.
- The initial location of the subscriber is defined with radius r from the center of the cell.
- Angles for the direction of the movement are uniformly distributed.
- Mobiles are allowed to move in any direction from the starting point.
- Velocity of the mobiles is constant within a cell.
- Initial velocity of the mobiles is assumed to be Gaussian pdf, truncated at 0 km/hr.
- Calls from different users are independent.
- Equilibrium of handovers is assumed.

Let us define with (r, θ) the initial position of the user in a cell and the direction of the movement at the call setup. We can derive the maximum time that subscriber will stay in the current cell. Using trigonometry we derive a relation for the maximum length of the user trajectory in the cell:

$$S_{\max} = \sqrt{R^2 - r^2 \sin^2 \theta} + r\cos\theta, \theta \in [0, 2\pi) \tag{6.19}$$

With the given initial velocity of the user V, the maximum time the user can spend in current cell is given by

$$T_{\max} = S_{\max} / V \tag{6.20}$$

where $S_{\max}$ is calculated by using (6.19). If t is a relative time from the beginning of the call, and T_c is call duration, then one may obtain channel-holding time T_{ch} in a case of a new call:

$$T_{ch} = \begin{cases} T_c, t < T_{\max} \\ T_{\max}, t \geq T_{\max} \end{cases} \tag{6.21}$$

In a case with a handover call to the cell, let us denote with t_h the time interval until the moment of handover. Then, for T_{ch} we obtain

$$T_{ch} = \begin{cases} T_c - t_h, t - t_h < T_{\max} \\ T_{\max}, t - t_h \geq T_{\max} \end{cases} \tag{6.22}$$

The direction and speed of the mobiles is constant within a cell, but they are allowed to change at a handover to a neighboring cell. This mobility model can be used in urban areas. However, it can be easily extended to a highway scenario by limiting the changes in mobile speed and direction at cell borders.

6.7.2 Micromobility Model

In wireless IP networks, we expect also to have slow-moving users (e.g., in the office). For indoor office users we may apply a two-state Markov mobility model as shown in Figure 6.10.

According to this model, there are two states of the mobile: *M* (mobility) state, and *S* (stationary) state. The mobile moves while in state *M*, and there is no movement in state *S*. Usually, the velocity is assumed to be constant while mobile is in *M*-state [1].

Mobility of the users, however, depends upon the environment (rural, suburban, urban). For different environments we use different cell sizes: picocells, microcells, and macrocells. Also, the available maximum bit rate for a given user depends upon the mobility. Table 6.2 shows the relation between mobility classes defined for IMT-2000 and cell sizes, environment, and bit rate. One may notice that higher mobility results in lower bit rates and larger cells and vice versa. Additionally, wireless LAN can be added to provide many times higher data rates for low mobility users.

6.8 Performance Parameters

In order to analyze the quality of different services, we need to define the QoS parameters of wireless IP networks. We create two groups of parameters: call-level and packet-level performance parameters, according to our definition of traffic analysis on both levels.

6.8.1 QoS Parameters on Call-Level

In circuit-switched networks new calls are blocked if there are no available channels when the call is initiated. In packet mode, the most common way of treating users is not to block them, but to queue them. However, for services with

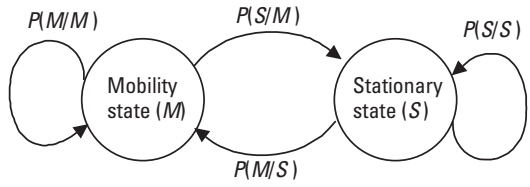


Figure 6.10 Mobility model for modeling indoor user movement.

Table 6.2
Mobility, Cell Types, Bit Rates, and Environments in IMT-2000

Cell Type	Cell Size	Mobility (km/hr)	Bit Rate	Environment
Picocell	Several tens of meters	<10 (low)	2 Mbps	Indoor (households, offices, building floors)
Microcell	Several hundreds of meters			Urban (hot spots, inner city area, airports)
		<120 (medium)	384 Kbps	
Macrocell	Several kilometers			Suburban
		<500 (high)	144 Kbps	Rural
Satellite	Hundreds of kilometers	<1,000 (highest)		Everything else

QoS requirements such as real-time services, admission control needs to be deployed in the access network. So, there will exist a ratio of blocked packet users. Blocking of users means clearing their calls from the system (i.e., they are not put in a queue). In wireless networks we have two types of calls: new calls and handovers from the adjacent cells. The following parameters should be considered on call-level in wireless IP networks:

- Mean cell residence time during a single session/connection;
- Handover intensity (incoming and outgoing handover intensities are equal in equilibrium);
- Average number of handovers per call;
- New call blocking probability;
- Handover call blocking probability;
- Call dropping probability.

The mean connection cell residence time is always smaller than call duration, or equal to it when no handovers occur during the connection.

Handover intensity is the average number of handovers in a cell. It is directly related to mean connection cell residence time (i.e., handover intensity is inverse proportional to it). The last two parameters are related to user mobility. Thus, smaller cells and higher mobility of users are decreasing the mean connection cell residence time (i.e., increasing the handover intensity in the cell). Also, higher handover intensity leads to higher average number of handovers per call.

Call blocking may occur when admission control is applied. The admission control is necessary to maintain desired QoS to the offered services. It is especially important in the case of real-time services. Call blocking, either new call or handover, is the result of insufficient network resources for serving all user requests. We define the blocking probability as the ratio of the number of blocked calls and total number of call attempts. Thus, the handover blocking probability of a cell is equal to the ratio of the number of rejected handovers and the total number of handover attempts to the cell. The ratio of the number of dropped calls and the number of all established calls provides the call dropping probability. It is directly related to the handover blocking probability (Chapter 7). In a scenario with multiple traffic classes and services, the bandwidth is shared among the classes. In such a scenario we have performance parameters for each traffic type.

6.8.2 QoS Parameters on Packet-Level

Let us first define the meaning of *packet flow.* We need an explicit definition of a flow to define QoS parameters on the packet-level. By definition, a flow is a continuing communication between two network entities. It may be a one-way or two-way, symmetrical or asymmetrical communication, which results in sending and receiving IP packets. A flow starts with the transmission of the first packet from a particular connection, and it ends after a longer period of inactivity. In 2G+ mobile systems (such as GPRS) and in 3G systems (such as UMTS), *Packet Data Protocol* (PDP) contexts are defined. Each PDP context exists during the state of packet transmission or reception or in "standby" state (following shortly after the packet communication) [32]. In the following we define the network performance parameters: mean packet losses, mean packet delay, delay variation, and throughput. In multiple access techniques, such as CDMA and its specific types (WCDMA targeted for UMTS and cdma2000), we cannot do performance analysis without considering the bit error ratio and SER [9].

Each IP packet transmitted on the network may be dropped by one of the network nodes due to limited buffer space or high delay (i.e., loss occurs). We define a packet loss L_j as a ratio of total length (in bytes or bits) of all packet losses and total length of all generated packets within a flow *j*:

$$L_j = \frac{\sum_{i \in X_{loss}} l_i^{(j)}}{\sum_{i \in X_{total}} l_i^{(j)}} \tag{6.23}$$

Buffering in the network nodes adds delay and also adds delay variation due to the different load at the nodes on the packets' path. If we denote with

N_{nodes} the number of nodes on the flow path, and with $D^{(n)}$ we denote the delay due to buffering at node n, then the mean packet delay of the flow j may be calculated by using the following:

$$D_j = \frac{1}{N_j} \sum_{i=1}^{N_{nodes}} \sum_{k=1}^{N_j} D_{jk}^{(i)} \tag{6.24}$$

where N_j is the number of all packets from flow j.

Delay variation DV (or jitter) occurs due to different delay of the packets, which is a consequence of the bursty traffic load in network nodes. We may define delay variation as follows:

$$DV_j = \frac{1}{N_j} \sum_{i=1}^{N_{nodes}} \sum_{k=1}^{N_j} \left(D_{jk}^{(i)} - D_j \right)^2 \tag{6.25}$$

For every traffic type (audio, video, and data) we define throughput as a ratio of the number of received bytes (from all packets) and total time of flow duration. We measure the throughput in bits per second (or kilobits per second or megabits per second). We also may define an effective throughput as a ratio of the number of transmitted bytes and the number of all generated bytes. If no losses occur, then the effective throughput will be equal to one. In all other cases, when losses occur, effective throughput is less than one.

6.8.3 Capacity

Related to the network capacity is traffic intensity (Chapter 4). According to [33], the traffic intensity as defined as follows:

> *Definition of traffic intensity:* Traffic intensity in a pool of resources is the number of busy resources at a given instant of time.

The pool of resources may be a group of servers, such as trunk lines. We usually use mean traffic intensity as given by (4.58). Here we may define the capacity:

> *Definition of network capacity:* Network capacity is the maximum traffic intensity that can be carried by network resources under given constraints on call-level and packet-level (if appropriate) performance parameters.

Capacity of telecommunications systems is usually measured in unit Erlang, the same used for measuring traffic intensity (Chapter 4). It follows directly from the above definitions. A line (e.g., channel) can carry one Erlang at

most. An Erlang unit is defined over a time period of 1 hour. The unit is dimensionless. The total carried traffic in a given time period is the traffic volume, and it is measured in Erlang-hours (Eh). The traffic intensity is also measured in Erlangs. Analytically, it may be also defined as

$$\text{Traffic intensity} = \frac{\lambda}{\mu} \tag{6.26}$$

where λ and μ are call arrival rate and call departure rate, respectively.

In the circuit-switched networks, such as 2G mobile networks, we have a fixed amount of network resources (i.e., channels). The physical capacity of the system is defined as the total number of channels in a system (a cell or a switch) that can be allocated to users. However, we have two channel considerations: a physical channel and a logical channel. A physical channel is one frequency in FDMA systems, or one time slot in TDMA systems. Several logical channels may be mapped onto a physical channel. Typical examples of logical channels are control channels defined in 2G, 2G+, and 3G mobile systems, in the radio interface.

Packet switching is included for the first time in 2G+ mobile systems. There it coexists with the circuit-switched network domain. Packet-switched networks are characterized by statistical multiplexing of data from different users onto the medium [i.e., sharing the resources among all users in that domain (e.g., in a cell)]. Therefore, we define the capacity of a system in a packet-switched network as the total data rate (in bits per second) that is supported by the network. So, the capacity of the cell is the total bit rate of the wireless link in that cell. However, the mapping between the network resources and bit rates is dependent upon the access technique. We have three major access techniques for wireless networks today: FDMA, TDMA, and CDMA. There are many variations of these access techniques as well as their combinations. For example, the GSM access network is a combination of FDMA and TDMA; WCDMA-FDD is a combination of FDMA and CDMA; and WCDMA-TDD is a combination of the all three techniques. Thus, we always have a limited spectrum for the mobile system, we may have a limited number of time slots when using the TDMA, or we may have a limited number of codes that can be dedicated to the users when using CDMA.

Considering the discussion above, we may define two types of capacities in the cellular access network: hard capacity and soft capacity.

6.8.3.1 Hard Capacity

If the capacity is hard blocked (i.e., limited by the amount of physical network resources) and it is not dependent upon the network load and the environment, then we have hard capacity in the network. All circuit-switched systems are

characterized by the hard capacity. It is suitable, however, for packet-switched systems, when the amount of available network resources is fixed to the amount of hardware. Then, we have a time-invariant network capacity. The utilization of the resources depends upon the traffic types and allocation scheme (e.g., one channel per call, two channels per call, allocation of single channel or multiple channels). When we have hard capacity, if all channels are busy, the call is rejected. This is called *hard blocking*. By blocking we will assume hard blocking, unless specified elsewhere. In this case we may use the Erlang-B formula for network dimensioning (Chapter 4).

6.8.3.2 Soft Capacity

If the capacity is limited by the amount of interference in the radio access network, besides the available hardware, then we have a soft capacity. In such a case, there is no fixed value for the maximum capacity of the system. Our system is a cell surrounded by its neighboring cells. So, the less interference that comes from the neighboring cells, the more channels (resources) that are available in the observed cell. Low traffic load in neighboring cells adds capacity. We have a soft capacity that is suitable for bursty traffic, because capacity can be "borrowed" from the neighbors. We have soft capacity in CDMA (e.g., WCDMA). We might even have soft capacity in GSM, if the radio network capacity is limited by the amount of interference instead of the number of time slots. GSM was created with the voice service as the main service, which needs a smaller amount of resources per connection. Therefore, time slots are more suitable for GSM and primarily voice-oriented mobile networks.

We define soft blocking as blocking due to no available resources considering the soft capacity. One may refer to the soft capacity as an upgrade of the hard capacity. So, if we denote with C_{Hard} and C_{Soft} Erlang capacity of a cell with hard and soft blocking, respectively, then we may write

$$\text{Relative soft capacity} = \frac{C_{Soft} - C_{Hard}}{C_{Hard}} \tag{6.27}$$

Soft capacity and soft blocking are characteristic for power-based admission control, as we find for CDMA systems.

6.9 Discussion

In this chapter we defined an open architecture for wireless IP networks with multiple traffic classes, as well as a simulation environment for QoS analysis of wireless access networks. We modeled two main differences between wireless and wired networks (i.e., errors due to the wireless channel and user mobility).

We proposed an integrated simulation architecture that provides traffic tracking on different levels: call-level and packet-level. For the purpose of network analysis, we defined general models for different elements: network nodes, traffic sources, and links (wired and wireless). For every element we defined a conceptual model. For instance, network nodes should perform traffic differentiation [i.e., classification of incoming aggregate traffic into different streams according to the traffic class (for traffic classification refer to Chapter 5)]. Base stations, which are wireless access points of the mobile network, should differentiate each traffic flow (not the aggregate traffic per class) due to location management during an ongoing connection.

Wireless channels are characterized with a location-dependent and time-varying bit error ratio. This property of wireless channels may be modeled by using the two-state Markov error model.

Also, we modeled the mobility of the user. It may be divide into macromobility (car, trains, bicycles or outdoor pedestrians) and micromobility (indoor or office) modeling.

Finally, we defined QoS parameters for measuring the quality of the service. There are two types of QoS parameters: call-associated parameters (new call blocking probability, handover blocking probability, and mean bit rate) and packet-associated parameters (packet losses, packet delay, jitter, and throughput). Also, we distinguished between hard and soft blocking (and associated capacities).

References

[1] 3GPP TS 25.401, *UTRAN Overall Description (Release 5)*, V5.1.0, 3GPP Proposed Technical Report, 09-2001.

[2] Berezdivin, R., R. Breineg, and R. Topp, "Next-Generation Wireless Communications Concepts and Technologies," *IEEE Communications Magazine*, Vol. 40, No. 3, March 2002.

[3] Zografski, Z., and T. Janevski, "Simulation Models and Methodologies for ATM Switches," *International Conference on Parallel and Distributed Computer and Networks '98*, Brisbane, Australia, December 1998.

[4] Takai, M., et al., "Impact of Channel Model on Simulation of Large Scale Wireless Networks," *ACM/IEEE MSWiM'99*, Seattle, WA, August 1999.

[5] Liu, W., et al., "Parallel Simulation Environment for Mobile Wireless Networks," *Proc. Winter Simulation Conference*, Coronado, CA, 1996.

[6] Guerin, R., et al., "Scalable QoS Provision Through Buffer Management," *ACM SIGCOMM*, Vancouver, B.C., September 1998.

[7] Parekh, A. B., and R. Galager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," *IEEE/ACM Transactions on Networking*, Vol. I, No. 3, June 1993.

[8] Nangopal, T., S. Lu, and V. Bharghavan, "A Unified Architecture for the Design and Evaluation of Wireless Fair Queuing Algorithms," *ACM Mobicom'99*, Seattle, WA, August 1999.

[9] 3GPP TS 23.107, *QoS Concept and Architecture (Release 5)*, V.5.3.0, 01-2002.

[10] Nguyen, G. T., et al., "A Trace-Based Approach for Modeling Wireless Channel Behavior," *Proc. Winter Simulation Conference*, Coronado, CA, December 1996.

[11] Konrad, A., et al., *A Markov-Based Channel Model Algorithm for Wireless Networks*, Report No. UCB/CSD-01-1142, EECS, University of California, Berkeley, May 2001.

[12] Paxson, V., and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Trans. on Networking*, June 1995, pp. 226–244.

[13] Paxson, V., and S. Floyd, "Why We Don't Know How to Simulate the Internet," *Proc. Winter Simulation Conference*, Atlanta, GA, December 1997.

[14] Feldmann, A., et al., "The Changing Nature of Network Traffic: Scaling Phenomena," *ACM SIGCOMM*, Vol. 28, No. 2, April 1998.

[15] Paxson, V., "Empirically Derived Analytic Models of Wide-Area TCP Connections," *IEEE/ACM Trans. on Networking*, Vol. 2, No. 4, August 1994, pp. 316–336.

[16] Huebner, F., D. Liu, and J. M. Fernandez, "Queuing Performance Comparison of Traffic Models for Internet Traffic," *GLOBECOM'98*, Sydney, Australia, November 8–12, 1998, pp. 1931–1936.

[17] Erramilli, A., O. Narayan, and W. Willinger, "Experimental Queuing Analyses with Long-Range Dependent Packet Traffic," *IEEE/ACM Trans. on Networking*, Vol. 4, No. 2, April 1996.

[18] Sahinoglu, Z., and S. Tekinay, "On Multimedia Networks: Self-Similar Traffic and Network Performance," *Communications Magazine*, January 1999.

[19] Anderlind, E., "Resource Allocation in Multi-Service Wireless Access Networks," Ph.D. dissertation, Royal Institute of Technology, Stockholm, Sweden, October 1997.

[20] You, C., and K. Chandra, "Time Series Models for Internet Data Traffic," *24th Conference on Local Computer Networks, LCN'99*, October 1999.

[21] Janevski, T., and B. Spasenovski, "QoS Provisioning for Wireless IP Networks with Multiple Classes Through Flexible Fair Queuing," *GLOBECOM 2000*, San Francisco, CA, November 27–December 1, 2000.

[22] Ahlgren, B., et al., "Dimensioning Links for IP Telephony," *IPTEL 2001 SICS*, CNA Laboratory, Sweden.

[23] 3GPP TR 25.881, *Improvement of RRM Across RNS and RNS/BSS (Release 5)*, V5.0.0, 3GPP Proposed Technical Report, 12-2001.

[24] Huebner, F., D. Liu, and J. M. Fernandez, "Queuing Performance Comparison of Traffic Models for Internet Traffic," *GLOBECOM'98*, Sydney, Australia, November 8–12, 1998, pp. 1931–1936.

[25] Lam, D., D. C. Cox, and J. Widom, "Teletraffic Modeling for Personal Communications Services," *IEEE Communications Magazine*, Vol. 35, No. 2, February 1997.

[26] Lin, Y.-B., "Modeling Techniques for Large-Scale PCS Networks," *IEEE Communications Magazine*, Vol. 35, No. 2, February 1997.

[27] Markoulidakis, J. G., et al., "Mobility Modeling in Third-Generation Mobile Telecommunications Systems," *IEEE Personal Communications*, Vol. 4, No. 4, August 1997.

[28] Cheung, B. H., and V. C. M. Leung, "Network Configurations for Seamless Support of CDMA Soft Handoffs Between Cell Clusters," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 7, September 1997, pp. 1276–1288.

[29] Zonoozi, M. M., and P. Dassanayake, "User Mobility Modeling and Characterisation of Mobility Patterns," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 7, September 1997, pp. 1239–1252.

[30] Gavrilovska, L., and T. Janevski, "Modelling Techniques for Mobile Communications Systems," *GLOBECOM'98*, Sydney, Australia, November 1998, pp. 2623–2628.

[31] Gavrilovska, L., and T. Janevski, "Modelling Techniques in Cellular Networks," *APCC/ICCS'98*, Singapore, November 1998, pp. 556–560.

[32] Bettstetter, C., H. J. Vogel, and J. Eberspacher, "GSM Phase 2+, General Packet Radio Service GPRS: Architecture, Protocols and Air Interface," *IEEE Communication Surveys*, Vol. 2, No. 3, Third Quarter 1999, http://www.comsoc.org/pubs/surveys/.

[33] ITU, *Traffic Intensity Unit*, ITU-T Recommendation B.18, 1993.

7

Analytical Analysis of Multimedia Mobile Networks

7.1 Introduction

Mobile networks provide additional multimedia services, besides the traditional voice service. The introduction of multimedia services to mobile networks creates new problems for the design of such systems. We need to provide certain QoS for the offered multimedia services. In multimedia mobile IP networks one may expect different bandwidth demands by different traffic classes. Also, different classes have different traffic parameters, such as new call intensity and call/session duration.

For appropriate planning of multimedia mobile networks, we need to derive an analytical framework for the expected traffic. Usually, we use an analytical approach for the time period during a day with the highest traffic load in the network or in a particular link (in circuit-switched networks we refer to this time period as the busy hour).

In cellular networks we usually measure quality of service with new call blocking probability and call dropping probability. Call blocking probability refers to blocking of new calls in a cell. Dropping probability refers to forced termination of an already-established call due to no availability of resources in any of the neighboring cells at a handover. To be able to use the analytical description of the resource allocation and call blocking/dropping, it is more convenient to divide the available bandwidth into logical channels. Allocation of the logical channels may be done statically [*fixed channel allocation* (FCA)] or dynamically [*dynamic channel allocation* (DCA)]. In this chapter we assume fixed allocation of logical channels in the wireless access network.

In the following sections we present analytical traffic analysis of a cellular network with single and multiple traffic classes.

7.2 Analysis of Mobile Networks with Single Traffic Class

In fixed networks, the user's attachment point to the network is fixed during longer time periods (i.e., months or years), but in mobile networks, users tend to be on the move and therefore frequently change the wireless access point (e.g., the base station). Thus, a user is allowed to move during the duration of a call (or a session), thus performing handovers while moving through different cells. The mobility of the users (we described some mobility models in Chapter 6) is the reason for a different approach in dimensioning mobile networks than traditional fixed networks [1].

For example, in traditional telecommunications networks we usually analyze two main parameters: call duration and call arrival rate (or call interarrival time), but in mobile networks, mobile terminals may allocate and release logical channels in the same or different cells several times during the call. In a case of multimedia traffic, the bandwidth requirement by a specific connection varies during the time. This is different from the classical mobile communication networks where each call occupies or releases a fixed number of logical channel(s) or bandwidth portion. Hence, in packet mobile networks, a bandwidth portion of the wireless link (in a cell) may be allocated by a new or incoming handover call. Also, a portion of the cell capacity may be released with the termination of the call or outgoing handover (to an adjacent cell).

Traffic parameters depend on the cell sizes and the user's mobility. In this chapter we present analysis on a call-level. Therefore, we use call-level traffic parameters according to their definitions in Chapter 6. Using the declared definitions of the traffic parameters, we derive an analytical model for single-class mobile networks.

7.2.1 Analytical Modeling

We need to learn how the traffic parameters are related to different cell sizes and the mobility of users. For that purpose we develop an analytical model. Traffic parameters for single-class wireless cellular networks were introduced in Section 4.7. Traffic parameters are related to mobility parameters, such as velocity, initial position in the cell, and direction of the movement, which are modeled as random variables in Chapter 6. In order to perform traffic analysis in mobile networks, we need to analyze the following traffic and performance parameters: average channel holding time T_{ch}, handover intensity, new call blocking probability P_{bn}, handover call blocking probability P_{Fh} (P_{bn} and P_{Fh} parameters define

the grade of service), average number of handovers per call, and call dropping probability P_D.

Handovers are typical events in a wireless environment. When a handover call is blocked, however, the ongoing connection is dropped. Call dropping reduces the average call duration. Thus, we may define an effective call duration $1/\mu_e$, where

$$\mu_e = \mu_c + P_B \mu_h \tag{7.1}$$

where μ_c and μ_h are call termination rate and handover departure rate, as defined in Chapter 4.

Consider a system consisting of two cells, which are exchanging handovers between each other. Denote with $p(n_1, n_2)$ the joint probability distribution when n_1 channels are busy in the first cell and n_2 channels are busy in the second cell. The Markov state diagram for this scenario is given in Figure 7.1. Assuming C_1 channels for cell 1 and C_2 channels for cell 2, for $n_i = \{0, 1, \ldots, C_i\}$, $i = 1, 2$:

$$\begin{aligned}
&\lambda_{n2} p(n_1, n_2 - 1) + \lambda_{n1} p(n_1 - 1, n_2) + (n_2 + 1)\mu_{h2} p(n_1 - 1, n_2 + 1) \\
&+(n_2 + 1)\mu_{c2} p(n_1, n_2 + 1) + (n_1 + 1)\mu_{c1} p(n_1 + 1, n_2) \\
&+(n_1 + 1)\mu_{h1} p(n_1 + 1, n_2 - 1) \\
&= [n_1 \mu_{c1} + n_2 \mu_{c2} + n_1 \mu_{h1} + n_2 \mu_{h2} + \lambda_{n1} + \lambda_{n2}] p(n_1, n_2)
\end{aligned} \tag{7.2}$$

We get a system of $(C_1 + 1)(C_2 + 1)$ equations. By solving these equations we will obtain the joint probability distribution. For the equilibrium case, when there are no handovers, state distribution is given with

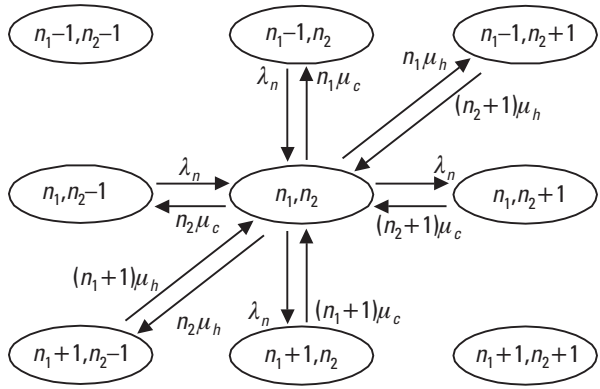


Figure 7.1 Markov state diagram for two-cell scenario.

$$p(n_1, n_2) = p_0 \left(\frac{\rho_1^{n_1}}{n_1!} \right) \left(\frac{\rho_2^{n_2}}{n_2!} \right) \tag{7.3}$$

But when handovers exist, the above approach is not efficient.

The average number of handovers per call is a significant parameter for obtaining the performances of the mobile system. We proceed with a more tractable analytical approach. If $P[H = j]$ denotes the probability that there are j handovers per call, then the average number of handovers per call is

$$E[H] = \sum_{j=1}^{+\infty} jP[H = j] \tag{7.4}$$

Let P_n be the probability that an accepted new call will require at least one handover before completion, let P_h be the probability that a call after a successful handover will require at least one more handover before completion, and let P_{Fh} denote the probability that a handover attempt will fail. Then we can calculate $P[H = j]$:

$$\begin{aligned} &P[H = 0] = (1 - P_n) + P_n P_{Fh} \\ &P[H = 1] = P_n (1 - P_{Fh})(1 - P_h + P_h P_{Fh}) \\ &\ldots \\ &P[H = j] = P_n (1 - P_{Fh}) \left[P_h (1 - P_{Fh}) \right]^{j-1} (1 - P_h + P_h P_{Fh}) \end{aligned} \tag{7.5}$$

Replacing (7.5) into (7.4), we obtain the following expression for the average number of handovers per call $E[H]$:

$$E[H] = \frac{P_n (1 - P_{Fh})}{1 - P_h (1 - P_{Fh})} \tag{7.6}$$

The probabilities P_n and P_h can be calculated using the following relations:

$$P_n = P[T_c > T_n] = \int_0^{+\infty} e^{-\mu_c t} f_{T_n}(t) dt \tag{7.7}$$

$$P_h = P[T_c > T_h] = \int_0^{+\infty} e^{-\mu_c t} f_{T_h}(t) dt \tag{7.8}$$

where $f_{T_n}(t)$ and $f_{T_h}(t)$ are the probability distribution functions for new call cell residence time T_n and handover call cell residence time T_h, respectively.

Call dropping probability is the probability that a nonblocked call will be forced to terminate due to lack of free resources in the target cell at a handover. We may calculate the probability $P_D[H]$ that a call performs H handovers before it is dropped:

$$P_D[H] = 1 - (1 - P_{Fh})^H \tag{7.9}$$

Then, the call dropping probability is

$$P_D = 1 - \sum_{i=0}^{\infty} (1 - P_{Fh})^i P[H = i] \tag{7.10}$$

where $P[H = i]$ is the probability that the call has i handovers during its duration. In our approach we assumed that call duration is exponentially distributed with a mean $T_c = 1/\mu$ [μ is given by (4.119)]. Furthermore, if we assume that channel holding time is also exponentially distributed with a mean T_{ch} (i.e., number of handovers per call follows the Poisson process), we may write

$$P[H] = \int_0^{\infty} P[H|x] \mu e^{-\mu x} dx = \int_0^{\infty} \frac{\left(\frac{x}{T_{ch}}\right)^H}{H!} e^{-\frac{x}{T_{ch}}} \mu e^{-\mu x} dx \tag{7.11}$$

where $P[H \mid t]$ is the probability that the call has H handovers when the holding time is t. If we insert (7.11) into (7.10), we obtain

$$P_D = 1 - \sum_{H=0}^{\infty} (1 - P_{Fh})^H \int_0^{\infty} \frac{\left(\frac{x}{T_{ch}}\right)^H}{H!} e^{-\frac{x}{T_{ch}}} \mu e^{-\mu x} dx \tag{7.12}$$

Furthermore, if we change the order of summing and integration, and utilize the expansion of the exponential function, we get

$$P_D = 1 - \int_0^{\infty} e^{\left(\frac{1 - P_{Fh}}{T_{ch}} x\right)} e^{-\frac{x}{T_{ch}}} \mu e^{-\mu x} dx \tag{7.13}$$

Finally, from (7.13) we obtain the relation for the call dropping probability P_D:

$$P_D = 1 - \frac{\mu T_{ch}}{\mu T_{ch} + P_{Fh}} = \frac{P_{Fh}}{\mu T_{ch} + P_{Fh}} \tag{7.14}$$

The last equation shows the relation between the call dropping probability P_D and handover blocking probability P_{Fb}. Mutual dependence of these two parameters is fundamental for traffic analysis of mobile networks and admission control procedures.

In the following section we extend the traffic theory for mobile networks with a single traffic type to a scenario with multiple traffic types.

7.3 Analysis of Multimedia Mobile Networks with Deterministic Resource Reservation

Multimedia mobile networks will include different traffic types, thus leading to different demands for network resources. We want to obtain the efficiency limit of such a multimedia mobile network. First, let us define our assumptions in the analysis.

We assume that no handover or new calls are queued. If any call is blocked, it is cleared from the system. Furthermore, we will use an assumption that capacity C of a cell (i.e., wireless access link) is divided into a set of logical channels i, $i = 1, 2, ..., N$. Also, we assume that resource allocation is deterministic (i.e., changing of allocated resources within a single cell is not allowed).

The calls from different users are independent. This assumption holds in the cases where the number of users in a cell is many times greater than the available number of logical channels in that cell, and it usually holds in telecommunication mobile networks. Let us assume that the network has defined K different traffic classes. We assume that all call arrivals (for all classes) are according to the Poisson process. Similarly, call duration is exponentially distributed for all traffic classes. Hence, the call arrival process for a traffic type k is the Poisson process with rate λ_k, and call duration time is exponentially distributed with mean $1/\mu_k$. All calls from a same class k require the same amount of bandwidth denoted as b_k. In a case of no free resources in the serving cell at the call initiation, the call will be rejected (i.e., blocked). We assume that the number of users in the cell is large enough, so the number of ongoing connections does not vary new call and handover intensity.

Utilization of resources may be calculated as a relation between average time of resource's allocation and cell capacity.

Considering the QoS, an important quality measure for a mobile user is the probability of a successfully established call getting blocked at handover (i.e., call-dropping probability). For mobile users the dropping of an ongoing call is highly undesirable [2]. Therefore, we usually bound call-dropping probability by reserving a part of the cell capacity only for handover calls. The drawback is that such an approach increases new call blocking probability. Call dropping is directly associated with the handover blocking probability. If the system does

not allow or use reservation of resources for handover, then handover blocking probability equals new call blocking probability (i.e., $P_{Bn} = P_{Fh}$). To bound call-dropping or handover-blocking probability, we need to reserve channels (or bandwidth) for handovers in advance.

The average number of handovers per call may be calculated using (7.6), applied to the specific traffic type, or we may calculate it by dividing the average call duration time T_c and average channel holding time T_{ch} (within the cell):

$$E[H]_k = H_k = \frac{T_{c,k}}{T_{ch,k}} = \frac{1}{\mu_k T_{ch,k}} \tag{7.15}$$

To obtain the efficiency of the wireless networks, we should first consider a system where resources for a handover call must be reserved in advance in all cells that a mobile will visit during the call. This is the limiting case for handover mechanisms, because it provides zero handover blocking probability. Every other strategy should provide higher utilization of resources than this strategy. To provide the limiting case, we also assume that a mobile does not visit the same cell more than once throughout the call (although it may happen in reality). If a call of type k has H_k handovers, then the number of visited cells is $H_k + 1$. If we assume handover arrivals as a Poisson process during the call duration, then the average number of handovers may be calculated as $H_k = T_{c,k}/T_{ch,k}$, where $T_{c,k}$ and $T_{ch,k}$ are mean call duration and channel holding time for a traffic type k, respectively. In our analysis we assume that resources are deterministically reserved in advance in all the cells that the mobile will visit during the call. If the average bandwidth requirement per call of traffic type k is b_k, then the average total bandwidth reserved for handovers during the entire duration of the call must be

$$B_k = b_k \left(1 + \frac{T_{c,k}}{T_{ch,k}}\right) \tag{7.16}$$

If the average number of users per cell is N, and $\rho_k = \lambda_k/\mu_k$ is the offered traffic per mobile user of call type k, the average total used bandwidth in the system is given by

$$\bar{b} = N \sum_{k=1}^{K} b_k \rho_k \tag{7.17}$$

Using (7.16) and (7.17), we may obtain the total bandwidth reserved for handovers in the system as

$$\overline{B} = N\sum_{k=1}^{K} B_k \rho_k = N\sum_{k=1}^{K} b_k \rho_k \left(1 + \frac{T_{c,k}}{T_{ch,k}}\right) \tag{7.18}$$

We may define utilization u of resources in the cellular mobile network as

$$u = \frac{\overline{b}}{\overline{b} + \overline{B}} = \frac{\sum_{j=1}^{K} b_j \rho_j}{\sum_{i=1}^{K} b_i \rho_i \left(2 + \frac{T_{c,i}}{T_{ch,i}}\right)} \tag{7.19}$$

Because channel holding time is inverse-proportional to the mobility of users, from (7.19) we may conclude that utilization is also inversely proportional to the handover intensity. Thus, cellular networks with smaller cells or higher user mobility will have lower utilization of the resources. In a trivial case, when $K = 1$ traffic type, (7.19) transforms into

$$u[K=1] = \frac{\mu T_{ch}}{2\mu T_{ch} + 1} \tag{7.20}$$

Furthermore, let us analyze the case with $K = 2$ traffic types (i.e., we consider a multiclass network). The utilization of wireless resources will be

$$u[K=2] = \frac{\rho_1 b_1 + \rho_2 b_2}{\rho_1 b_1 \left(2 + \frac{T_{c,1}}{T_{ch,1}}\right) + \rho_2 b_2 \left(2 + \frac{T_{c,2}}{T_{ch,2}}\right)} \tag{7.21}$$

After some algebra, from the last relation we obtain

$$u[K=2] = \frac{\overline{b}/N}{\frac{\lambda_1}{\mu_1^2} b_1 \left(2\mu_1 + \frac{1}{T_{ch,1}}\right) + \frac{\lambda_2}{\mu_2^2} b_2 \left(2\mu_2 + \frac{1}{T_{ch,2}}\right)} \tag{7.22}$$

Different traffic types have different traffic characteristics. Calls from one traffic type can have shorter duration, but higher intensity (e.g., phone calls). On the other hand, other traffic types can have longer call duration, but less frequent calls (e.g., multimedia calls). Without losing generality, we may define parameters as functions of one variable x (i.e., $\mu_1 = \mu/x$, $\lambda_1 = \lambda/x^2$, $b_1 = bx$, $T_{ch,1}$

$= T_{ch}$) for the first traffic type, and $\mu_2 = \mu x$, $\lambda_2 = \lambda x^2$, $b_2 = b/x$, $T_{ch,2} = T_{ch}$ for the second traffic type. According to such definitions of the parameters, we can control call duration, call intensity, and bandwidth requirements by changing the parameter x. If we introduce these definitions of the traffic parameters in (7.22), we may write

$$u[K=2] = \frac{\mu T_{ch}}{2\mu T_{ch} + \frac{1}{2}\left(x + \frac{1}{x}\right)} \tag{7.23}$$

The last relation significantly differs from (7.20), except in the trivial case $x = 1$ when the analysis is transiting to the scenario with a single traffic type in the network. It is interesting to calculate utilization for the limiting cases (i.e., $x \to 0$ and $x \to \infty$):

$$\lim_{x\to\infty}\left(u[K=2]\right) = \lim_{x\to 0}\left(u[K=2]\right) = 0 \tag{7.24}$$

This simple example with two traffic types in the network shows that utilization of the wireless resources decreases when we increase the diversity in traffic characteristics of different traffic types. Of course, this discussion considers the boundary case, when we reserve resources for handovers to reduce handover blocking probability to zero.

In FDMA/TDMA systems (e.g., GSM) we need to perform frequency planning of the wireless radio network. For that purpose, we usually group frequency carriers into groups of frequencies called clusters. Let δ denote the cluster size, and C_T denote the amount of capacity that base stations can support within a cell. Then, capacity density is $C = C_T/P\delta$, where P is the coverage area of the cell. We usually use hexagonal model of the cells, although a cell in a real network can have various forms. The hexagonal cell is further approximated with a circle. Therefore, we may write $P \sim R^2$, where R is the cell radius. Using the hexagonal form of the cells (with side a) and considering (6.14), we may calculate capacity density using the following equation:

$$C = \frac{C_T}{\delta \frac{3\sqrt{3}}{2} a^2} \tag{7.25}$$

The above equation shows the dependence between the capacity density C and cell radius R. By decreasing the cell size, we increase the physical capacity of the network; and this is a scenario of microcell and picocell networks. On the other hand, decreasing the cell size increases the average number of handovers

per call (channel holding time T_{ch} decreases). In such a case, using deterministic advance reservation of resources for handovers, we need to reserve a larger amount of bandwidth. Hence, a higher handover ratio leads to lower utilization of the wireless resources. Thus, we have two contradictory demands in cellular wireless networks: to increase the capacity we need to reduce the cell size, but at the same time smaller cells result in reduced utilization of the resources.

From the analysis of the deterministic theoretical reservation scheme, we can draw the following main conclusions:

- Utilization of wireless resources decreases as cell size decreases;
- Utilization of wireless resources decreases as the diversity in traffic parameters of different traffic types increases (e.g., some multimedia calls have long holding time and large bandwidth demands, and others have shorter holding time and smaller bandwidth requirements).

7.4 Analysis of Multimedia Mobile Networks with Statistical Local Admission Control

The most used strategies for admission control are based purely on local information. They decide to accept/reject a call based on the information of available resources in the cell where the new call or handover requests the admission, without any information on availability of resources in neighboring cells. These algorithms are simplest for implementation, and therefore, they are most frequently used in today's mobile systems.

In this section we analyze statistical local admission control in mobile networks with multiple traffic types. We use the assumption of Poisson arrivals for both new calls and handovers. Our approach in this part is consistent with the analysis in Section 7.3. Because users are more sensitive to dropping of an already established call, we usually use handover-blocking probability as a hard constraint in the admission control algorithms. Our tendency is to determine the strategy that ensures maximum efficiency (i.e., the highest utilization of resources upon given constraints on the QoS parameters, for example, new call blocking and call dropping probability).

Let the system have K traffic types. Total incoming traffic in a cell is a sum of new call arrivals and incoming handovers. Let M be the number of neighboring cells for the observed one. If we use hexagonal cell model, then each cell will have six neighbors (one per side). The number of neighbors, however, depends on the network configuration, and it is not necessarily limited to six (in real networks, the cellular operator may define as neighbors cells that

do not have a joint border). Let us denote the call acceptance rate of traffic type k with $a_k(c)$, where c is number of allocated logical channels (i.e., bandwidth). Furthermore, let $h_k(c)$ denote the incoming handover rate when there are$c$ busy logical channels in the cell. Under an assumption of equilibrium of incoming and outgoing handovers for a given cell, we may write the following equation:

$$a_k(c) + h_k(c) = \mu_k c + \frac{1}{T_{ch,k}} c \tag{7.26}$$

The number of busy logical channels is a time-dependent variable. However, the average cumulative traffic per class does not change drastically over shorter intervals, in the range of seconds or minutes. In the opposite case, dimensioning and planning of the network resources would be an almost impossible task. In traditional telecommunications networks, dimensioning is performed for the busiest hour of the day. This approach considers that, if we provide services with desired QoS during the hour with highest traffic intensity (i.e., busy hour), then we may expect that at least the same or better QoS will be provided during intervals with lower traffic intensity. In multimedia mobile networks we have different bandwidth requirements for different traffic types (i.e., classes). That is the main difference between multimedia and traditional mobile networks. Hence, in a multimedia environment we also should plan the network considering the time period with the highest traffic volume (e.g., this time period may be defined over periods of 30 minutes, or 1 hour). We will refer to this time period in a multimedia network as the *busiest traffic period* (BTP).

We tend to analyze macro-behavior of a multimedia mobile network (i.e., we consider call-level parameters only). The total number of busy logical channels is the sum of logical channels that are allocated to all K traffic types:

$$c = \sum_{k=1}^{K} c_k \tag{7.27}$$

where c_k is the average number of allocated channels during the BTP of traffic type k.

We may calculate incoming handover rate for traffic type k by using the following relation:

$$h_k(t) = \sum_{i=1}^{M} h_{k,i}(t) \tag{7.28}$$

where $h_{k,i}$ is handover intensity from neighboring cell i, $i = 1, 2, ..., M$, to the observed cell. If $a_i(t)$ denotes number of active users at the moment t in the cell i, then incoming handover intensity from the neighboring cell i can be calculated by

$$h_{k,i}(t) = a_{k,i}(t)\frac{1}{M}\frac{1}{T_{ch,k}}P_{h,k} \tag{7.29}$$

where $P_{h,k}$ is the probability that an active call will perform at least one more handover before it terminates, as given by

$$P_{h,k} = \int_0^\infty P_k[t \geq x]\frac{1}{T_{ch,k}}e^{-\frac{1}{T_{ch,k}}x}dx \tag{7.30}$$

where $P_k[t \geq x]$ is the probability that a call of traffic type k will remain active at least x time. By solving the integral in (7.30), we get

$$P_{h,k} = \int_0^\infty \frac{1}{T_{ch,k}}e^{-\mu_k x}e^{-\frac{1}{T_{ch,k}}x}dx = \frac{1}{1+\mu_k T_{ch,k}} \tag{7.31}$$

By inserting (7.31) into (7.29), and then in (7.28), we obtain the following equation for incoming handover intensity for traffic type k:

$$h_k(t) = \sum_{i=1}^{M} a_{k,i}(t)\frac{1}{M}\frac{1}{T_{ch,k}}\frac{1}{(1+\mu_k T_{ch,k})} \tag{7.32}$$

Under an assumption of uniformly distributed traffic in all neighboring cells, we may consider that all cells have equal average number of busy logical channels. In such a case, the average number of incoming handovers in longer time interval T will be

$$\begin{aligned} h_k(c_k) &= \frac{1}{T}\int_0^T h_k(t)dt = \\ &\frac{1}{T_{ch,k}(1+\mu_k T_{ch,k})}\frac{1}{T}\int_0^T a_k(t)dt = \frac{1}{T_{ch,k}(1+\mu_k T_{ch,k})}c_k \end{aligned} \tag{7.33}$$

where c_k is the average number of busy logical channels (i.e., amount of bandwidth) in longer time interval.

If the system is in balance at c_k allocated logical channels to traffic class k, for all classes $k = 1, 2, ..., K$, as given by (7.26), then we can exploit the relation (7.33) to calculate the call acceptance rate as a function of c_k:

$$a_k(c_k) = \left(\mu_k + \frac{1}{T_{ch,k}}\right)c_k - \left[\frac{1}{T_{ch,k}\left(1+\mu_k T_{ch,k}\right)}\right]c_k \tag{7.34}$$

From the last equations, after some algebra, we get

$$a_k(c_k) = \mu_k c_k \frac{2+\mu_k T_{ch,k}}{1+\mu_k T_{ch,k}} \tag{7.35}$$

To determine the efficiency of network resources in a given cell under given constraints on call-dropping probability, we need to define a new call acceptance rate at a different number of busy logical channels in the cell. If we assume that the network is balanced over a longer time interval for each traffic type $k = 1, 2, ..., K$, then we may deterministically allocate the available bandwidth in advance. At the network design we always specify traffic demands in advance, or we dynamically allocate the resources based on real-time traffic measurements

In the following sections we obtain the efficiency and perform optimization of a mobile network upon given QoS constraints. For the tractability of the analysis, we consider mobile networks with a single traffic type.

7.4.1 Efficiency of the Mobile Network

In this section we consider a system of a single traffic type (i.e., $K = 1$). We provide analytical analysis of the statistical local admission control that provides the highest system efficiency upon given constraints on handover blocking probability.

According to [3] we should distinguish between short and long behavior of the system. On a short term the incoming handover rate is independent of the state of the cell (i.e., how many channels are busy), while the call termination rate and outgoing handover rate are proportional to the number of ongoing calls in the cell. On the other hand, efficiency of the mobile networks is related to long-term behavior of the system (e.g., BTP). Let c^*be the long-term average of allocated channels in a cell. By using (7.35), the system is in balance at c^* if the rate of admitted new calls in the cell is

$$a(c^*) = \mu_c c^* \frac{2+\mu_c T_{ch}}{1+\mu_c T_{ch}} \tag{7.36}$$

The $a(c)$ function is nonincreasing, because as the number of allocated channels (slots) c increases less bandwidth is left for additional new calls. Also, the new call arrival rate is independent of c, because we assumed that the number of users is many times greater than the number of channels in the cell. Therefore, the system is balanced at c^* allocated channels. The stringent statistical admission control that results in the smallest handover blocking probability is defined by the boundary case [3]:

$$a(c) = \begin{cases} a(c^*), & x < c^* \\ 0, & x \geq c^* \end{cases} \tag{7.37}$$

Defined by the relation above, we may classify this method into fractional guard policies, where the threshold for handovers is placed at the mean occupancy.

At this point, it is convenient to define the threshold for handovers. By a definition, the handover threshold is equal to highest number of channels in the cell that can be allocated to both new calls and handovers. Also, we can use as a parameter the difference between the cell capacity and the threshold, $g = C - Threshold$, which is referred to as the number of guard channels. They can be used only for incoming handovers, while the remaining $C - g$ channels may be used for both new calls and handovers [4].

The problem that we consider is the optimization of the mobile network at given capacity C and constraint on the handover blocking probability P_{Fh}. At a given number of channels in the cell C, we are looking for highest c^* that will satisfy the given constraint. At the same time it will provide the maximum efficiency of the system, because the highest c^* will result in highest traffic load in the cell (i.e., highest efficiency).

In fact, c^* is the threshold for handovers. Instead of c^*, we can use the number of guard channels $g^* = C - \lfloor c^* \rfloor$ where $\lfloor x \rfloor$ denotes the largest integer that is less than or equal to x. The number of busy channels in the cell is well described by a birth-death process (i.e., Markov chain) as shown in Figure 7.2.

The state changes in the Markov chain are given by

$$\lambda = \lambda_n + \lambda_h = a(c^*) + h(c^*) \tag{7.38}$$

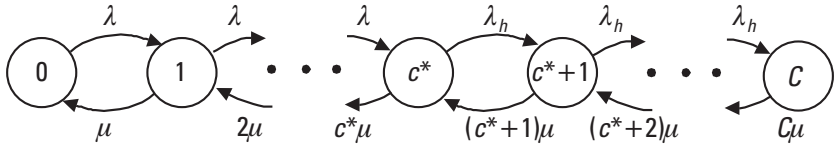


Figure 7.2 Markov chain of a mobile system with single traffic class.

$$\lambda_h = h(c^*) \tag{7.39}$$

where the new call rate $a(c^*)$ and handover rate $h(c^*)$ are given by (7.35) and (7.33), respectively. State departure rate, μ, can be calculated according to (4.119)—that is, $\mu = \mu_c + \mu_h$, where μ_c and μ_h are the call completion rate (in the cell) and the call handover rate (to the neighboring cells), respectively.

By solving the Markov state diagram by using birth-death processes [5], we can calculate call-dropping probability. For that purpose, we need to obtain the handover-blocking probability. Blocking of a handover happens when all logical channels at the target cell are busy (or the number of idle channels is less than the bandwidth requirements for that call). Hence, handover-blocking probability equals the probability that the system is in state *C*, as shown in Figure 7.2.

The total offered traffic (in Erlangs) to a cell is $A = \lambda/\mu$, while $A_2 = \lambda_h/\mu$ is the handover traffic in Erlangs. From the Markov chain we obtain the steady-state probabilities:

$$P(j) = \begin{cases} \dfrac{A^j}{j!} P(0), & 0 \le j \le c^* \\ \dfrac{A^{c^*} A_2^{j-c^*}}{j!} P(0), & j \ge (c^*+1) \end{cases} \tag{7.40}$$

Using (7.40) we can calculate the probability $P(0)$—that is, the probability that there are no allocated channels in the cell. Then, we can obtain the probability that the system (i.e., the cell) has allocated j channels by

$$P(j) = \begin{cases} \dfrac{A^j/j!}{\sum_{i=0}^{c^*} \dfrac{A^i}{i!} + \sum_{i=c^*+1}^{C} \dfrac{A^{c^*} A_2^{i-c^*}}{i!}}, & 0 \le j \le c^* \\ \dfrac{A^{c^*} A_2^{i-c^*}/j!}{\sum_{i=0}^{c^*} \dfrac{A^i}{i!} + \sum_{i=c^*+1}^{C} \dfrac{A^{c^*} A_2^{i-c^*}}{i!}}, & j \ge (c^*+1) \end{cases} \tag{7.41}$$

Handover-blocking probability is equal to the probability that all logical channels in the cell are busy. Therefore, it is given by

$$P_{Fh} = \frac{\dfrac{A^{c^*} A_2^{C-c^*}}{C!}}{\sum_{i=0}^{c^*} \dfrac{A^i}{i!} + \sum_{i=c^*+1}^{C} \dfrac{A^{c^*} A_2^{i-c^*}}{i!}} \tag{7.42}$$

If we use the above relation to calculate P_{Fb}, we can calculate the call-dropping probability P_D by using (7.14). New call blocking probability is

$$P_{Bn} = \sum_{j=c^*}^{C} P(j) = \frac{\sum_{j=c^*}^{C} \frac{A^{c^*} A_2^{j-c^*}}{j!}}{\sum_{i=0}^{c^*} \frac{A^i}{i!} + \sum_{i=c^*+1}^{C} \frac{A^{c^*} A_2^{i-c^*}}{i!}} \tag{7.43}$$

In a special case, when $A = A_2$, (7.43) becomes the Erlang-B formula, which is widely used in the dimensioning of telecommunication networks (refer to Chapter 4). The explanation of this phenomenon is simple. If we do not consider reservations for handovers, we get one Poisson arrival process equal to the sum of new call and handover intensities, which is served by the channel pool of the cell.

From (7.43) one may conclude that at a given throughput A, new call-blocking probability increases with decreasing of c^*. At the same time call-dropping probability decreases. So, we have two opposing requirements on new calls and handovers.

We may evaluate the relation between the handover blocking probability and number of handovers per call. Using (7.33), (7.36), (7.38), and (7.39), we define a new parameter θ as follows:

$$\theta = \frac{A_2}{A} = \frac{\lambda_h}{\mu}\frac{\mu}{\lambda} = \frac{\lambda_h}{\lambda_n + \lambda_h} = \frac{h(c^*)}{a(c^*) + h(c^*)} = \frac{1}{(1+\mu_c T_{ch})^2} = \frac{1}{(1+1/H)^2} = \frac{H^2}{(1+H)^2} \tag{7.44}$$

where $H = 1/(\mu_c T_{ch}) = T_c / T_{ch}$ is average number of handovers per call. If we include θ into the relation for calculation of handover blocking probability, we get

$$P_{Fh} = \frac{\frac{A^C}{C!}\theta^{C-c^*}}{\sum_{i=0}^{c^*} \frac{A^i}{i!} + \sum_{i=c^*+1}^{C} \frac{\theta^{i-c^*} A^i}{i!}} \tag{7.45}$$

We get that $\theta \to 0$ when $H \to 0$ (i.e., there are no handovers), thus giving zero handover-blocking probability. In such a case, the threshold c^* approaches

C, because there are no handovers to the cell. For $\theta \to 1$ (i.e., $H \to \infty$) the threshold c^* approaches a value determined by the Erlang system, because in that case almost all arrival calls would be handovers.

7.4.2 Optimization of Mobile Networks

Another problem is optimization of the network, which we define as the determination of an optimal number of channels in a cell upon given constraints on new call blocking probability $P_{Bn}(C, g)$ and handover blocking probability $P_{Fh}(C, g)$. We consider integer numbers of cell capacity C and guard channels g. Then, the optimization problem is the following:

$$\text{Minimize } C \text{ such that} \begin{cases} P_{Bn}(C, g) \le P_{b0} \\ P_{Fh}(C, g) \le P_{h0} \end{cases} \tag{7.46}$$

When we have no guard channels, then both probabilities equal the Erlang-B formula (i.e., $P_{Bn}(C, g) = P_{Fh}(C, g) = E_B(A, C)$, where A is the offered traffic to the cell). The number of guard channels always satisfies the trivial condition $g \le C$. Also, $C > 0$, $g \ge 0$, always holds. Therefore, we will analyze the optimization problem considering the first quadrant of the (C, g) plane shown in Figure 7.3.

Let C' and C'' be the number of channels in the cell, such that $P_{Bn}(C', 0) = P_{b0}$, $P_{Fh}(C'', 0) = P_{h0}$. It is easy to show that the following relations are true [6]:

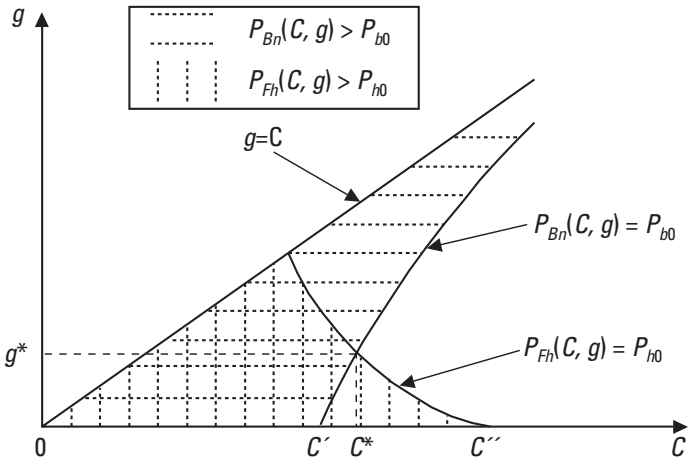


Figure 7.3 Optimization of a mobile network.

- $P_{Bn}(C, g)$ is an increasing function of g (for a fixed C): $P_{Bn}(C, g) > P_{Bn}(C, g-1)$.
- $P_{Bn}(C, g)$ is a decreasing function of C (for a fixed g): $P_{Bn}(C, g) < P_{Bn}(C-1, g)$.
- $P_{Bn}(C, g)$ is an increasing function of C and g, that is, $P_{Bn}(C, g) > P_{Bn}(C-1, g-1)$.
- $P_{Fh}(C, g)$ is a decreasing function of g, that is, $P_{Fh}(C, g) < P_{Fh}(C, g-1)$.
- $P_{Fh}(C, g)$ is a decreasing function of C, that is, $P_{Fh}(C, g) < P_{Fh}(C-1, g)$.
- $P_{Fh}(C, g)$ is a decreasing function of C and g, that is, $P_{Fh}(C, g) < P_{Fh}(C-1, g-1)$.

We can consider two different cases considering the values C' and C'':

1. $C' \geq C''$, in this case $P_{b0} \geq P_{h0}$, and therefore the active constraint is $P_{Bn}(C, g) \leq P_{b0}$. The optimal number of channels is the minimum number of channels $C = C^*$, such that $P_{Bn}(C, g) \leq P_{b0}$. Obviously, in this case guard channels are not needed; thus, $g^* = 0$. From Figure 7.3 we can conclude that $C^* = C'$.
2. $C' < C''$, this is the case shown in Figure 7.3. In this case optimization should satisfy both constraints, $P_{Bn}(C, g) \leq P_{b0}$ and $P_{Fh}(C, g) \leq P_{h0}$. The smallest number of channels that satisfies the constraints is $C = C^*$, for which $P_{Bn}(C^*, g^*) = P_{Fh}(C^*, g^*)$. To obtain the optimization pair (C^*, g^*), we can use a binary search algorithm.

Optimization is performed for the BTP, which may vary. For example, BTP of traditional voice service is during the working hours (e.g., between 1 p.m. and 2 p.m.), while browsing the Internet can have a BTP in late evening (e.g., around 12 a.m.). However, these characteristics can vary in different geographical areas. If a network is launched into commercial operation, then real-traffic measurements will be used for dimensioning and design of the system. In the rest of the day we may expect lower blocking probabilities than during BTP. If the system is overloaded, then call blocking probabilities continuously increase due to the unbalanced system. If such an overload of the network continues during longer time intervals (e.g., 0.5 hour or 1 hour), then the system is not well dimensioned. We perform dimensioning of mobile networks by using analytical models or simulations under given traffic parameters and constraints on the QoS. For initial dimensioning of the mobile network, we can use predictions for the traffic parameters based on a theoretical approach or values taken by measurements in existing networks.

7.5 Traffic Loss Analysis in Multiclass Mobile Networks

In this section we provide a generalization of the traffic theory in mobile networks in multiclass environment. We consider several traffic classes with different traffic parameters, such as call arrival process, call duration, and bandwidth requirements, offered to a group of logical channels. This approach considers mainly real-time services, where we allocate certain bandwidth, which must be divided into logical channels. We may, however, consider nonreal-time services as well, if they require some QoS guarantees (e.g., bandwidth).

In a multiclass environment we need to restrict the number of simultaneous calls for each traffic class. Thus we define the following class limitations for calls of class k by the following relations:

$$0 \le i_k \le c_k \le C, \quad k = 1,2,\ldots,K \tag{7.47}$$

$$\sum_{k=1}^{K} c_k > C \tag{7.48}$$

where i_k is number of simultaneous calls of traffic class k, c_k is the limit in number of channels that can be allocated to that class at the same time, C is the total number of channels in the cell, and K is the number of traffic classes. If (7.48) is not satisfied, then we get separate groups corresponding to K independent one-dimensional Markov chains.

7.5.1 Application of Multidimensional Erlang-B Formula in Mobile Networks

In this approach we consider a group of logical channels in a cell. We assume that all calls (new calls and handovers, from all traffic classes) are well modeled with the Poisson process. Let $\lambda_{n,k}$ and $\lambda_{h,k}$ be new call arrival rate and incoming handover rate of traffic class k. Also, we assume exponential distribution of call duration for each traffic class. Let $\mu_{c,k}$ and $\mu_{h,k}$ be call completion rate and outgoing handover rate of traffic class k. An important assumption in this case is that all traffic classes require the same number of logical channels per call (e.g., one channel per call). We have the following restrictions:

$$\begin{aligned} &0 \le i_k \le c_k \le C, \quad k = 1,2,\ldots,K \\ &0 \le \sum_{k=1}^{K} i_k \le C \end{aligned} \tag{7.49}$$

where i_k is number of calls from class k.

The total arrival process is a superposition of Poisson processes from different traffic classes. Thus, it is also a Poisson process with total arrival rate in the cell

$$\lambda = \sum_{k=1}^{K} \left(\lambda_{n,k} + \lambda_{h,k} \right) \tag{7.50}$$

The total (channel) holding time is hyper-exponentially distributed as given by

$$\begin{aligned} f(t) &= \sum_{j=1}^{K} \frac{\left(\lambda_{n,j} + \lambda_{h,j} \right)}{\sum_{i=1}^{K} \left(\lambda_{n,i} + \lambda_{h,i} \right)} \left(\mu_{c,j} + \mu_{h,j} \right) e^{-\left(\mu_{c,j} + \mu_{h,j} \right) t} \\ &= \sum_{j=1}^{K} \frac{\lambda_j}{\lambda} \mu_j e^{-\mu_j t} \end{aligned} \tag{7.51}$$

where $\lambda_k = \lambda_{n,k} + \lambda_{h,k}$ and $\mu_k = \mu_{c,k} + \mu_{h,k}$ are call arrival rate and call departure rate from traffic class k, respectively. The mean value of the total holding time is

$$T_{ch,total} = \sum_{j=1}^{K} \frac{\lambda_j}{\lambda} \frac{1}{\mu_j} = \frac{\sum_{j=1}^{K} \lambda_j / \mu_j}{\lambda} = \frac{\sum_{j=1}^{K} A_j}{\lambda} = \frac{A}{\lambda} \tag{7.52}$$

where A_j is offered traffic from class j, while A is the total offered traffic to the cell. An example of a Markov state diagram for $K = 2$ is shown in Figure 7.4.

Let $p(n_1, n_2, ..., n_K)$ denote the state probability of the system with n_1 ongoing calls from class 1, n_2 calls from class 2, ..., n_K ongoing calls from class K. Due to the independence of the calls from different traffic classes, we obtain

$$\begin{aligned} p(n_1 n_2, ..., n_K) &= p(n_1) p(n_2) ... p(n_K) \\ &= Q \frac{A_1^{n_1}}{n_1!} \frac{A_2^{n_2}}{n_2!} ... \frac{A_K^{n_K}}{n_K!} \end{aligned} \tag{7.53}$$

where Q is normalization constant. By the binomial expansion of Poisson processes, we can obtain the normalization constant:

$$p(n_1 + n_2 + ... + n_K = n) = Q \frac{\left(A_1 + A_2 + ... + A_K \right)^n}{n!} = Q \frac{A^n}{n!} \tag{7.54}$$

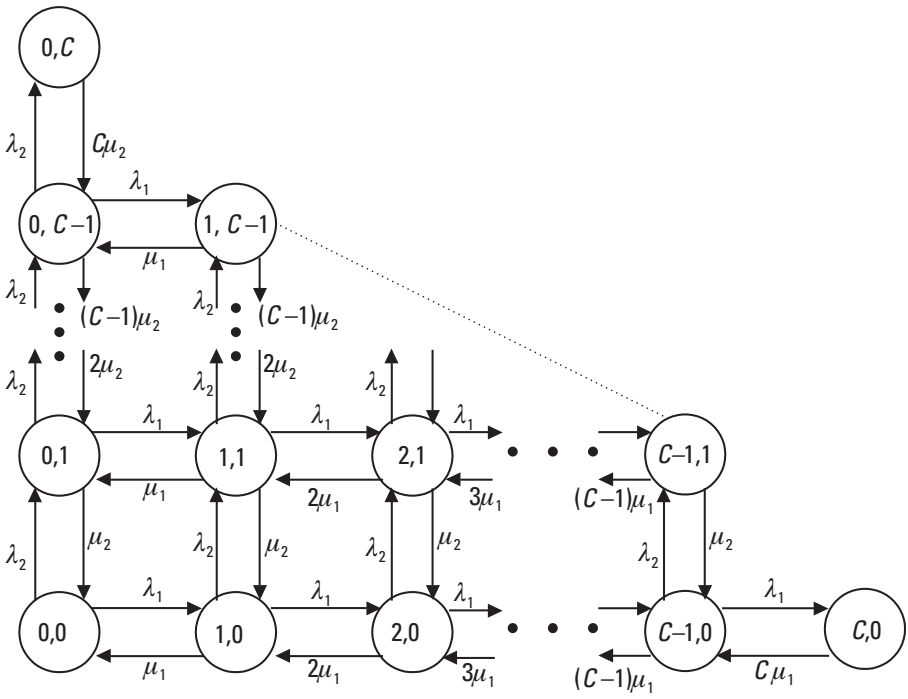


Figure 7.4 Two-dimensional Markov state diagram for two traffic classes.

$$Q = 1 / \left[\sum_{n=0}^{C} \frac{\left(A_1 + A_2 + \ldots + A_K \right)^n}{n!} \right] \tag{7.55}$$

From (7.54) we get a recursive relation for the state probabilities as follows:

$$p(n) = \frac{A}{n} p(n-1) \tag{7.56}$$

It is more convenient to define relative state probabilities $q(n)$ instead of absolute state probabilities $p(n)$, that is,

$$p(n) = \frac{q(n)}{Q(C)}, n = 0,1,2,\ldots,C \tag{7.57}$$

where

$$Q(C) = \sum_{j=0}^{C} q(j) \tag{7.58}$$

In this case we may chose $q(0) = 1$, and then use the recursive equation (7.56) to obtain $q(j)$, $j = 1, 2, ..., C$, as given by

$$q(j) = \frac{A}{j} q(j-1) = \frac{\sum_{i=1}^{K} A_i}{j} q(j-1),\ q(0) = 1 \tag{7.59}$$

Relative state probabilities provide easy normalization of the absolute state probabilities (e.g., when we truncate the system due to some restrictions).

If we do not define guard channels for handovers, then the blocking probability of the cell with C logical channels can be calculated by

$$P_B = \sum_{\forall\left(\sum_{j=1}^{K} n_j = C\right)} p(n_1, n_2, ..., n_K) = p(C) \tag{7.60}$$

If we introduce guard channels for handover calls in a cell, then we can calculate handover and new call blocking probability by using (7.42) and (7.43), respectively.

7.5.2 Multirate Traffic Analysis

In the previous section we used the assumption of equal bandwidth requirements from all traffic classes. However, 3G and future generations of mobile networks should support different traffic types with different bandwidth requirements at the same time. Thus, a voice call may require one logical channel (e.g., slot), while multimedia streaming may require several channels simultaneously. We consider resource sharing by all K traffic classes. Let b_j be the number of channels required by a call from traffic class j. We get additional limitations:

$$b_j n_j \le c_j \le C,\ j = 1,2...,K \tag{7.61}$$

where n_j is the number of active calls of class j, while c_j is limitation of that class (i.e., maximum number of channels in a cell that can be allocated). However, all traffic classes may occupy maximum C channels, which is the cell capacity:

$$\sum_{j=1}^{K} b_j n_j \le C \tag{7.62}$$

In Chapter 5 we classified the traffic in wireless IP networks into two main classes: class-A for services with QoS guarantees, and class-B for best-effort

service. Class-A traffic should be serviced with priority over class-B traffic. Therefore, class-B streams are invisible to class-A. Each class may be divided into subclasses and traffic subtypes. All call types allocate a certain number of channels at call start and keep them until call termination.

7.5.2.1 Aggregation Method

In multiclass multirate networks we need to solve a multidimensional Markov diagram, where dimension is equal to the number of classes. At high dimensions, the size of the state space will explode and we become unable to evaluate the system by calculating the individual state probabilities. We eliminate this problem by aggregation of the states.

If we assume Poisson arrival processes for all traffic classes, then we may use a modification of (7.59) to obtain the relative state probabilities [7, 8]; that is,

$$q(j) = \frac{\sum_{i=1}^{K} A_i b_i q(j - b_i)}{j}, \quad q(0) = 1 \tag{7.63}$$

If no channels are reserved for handovers and there are no restrictions (i.e., $c_j = C, j = 1, 2, \ldots, K$), we can calculate the blocking probability of k traffic class by

$$P_{Bn,k} = P_{Fh,k} = \sum_{j=C-b_k+1}^{C} p(j) \tag{7.64}$$

If we define g guard channels for handovers to the cell (which will be shared among all traffic classes), then blocking probabilities of each traffic type can be calculated by

$$P_{Fh,k} = \sum_{j=C-b_k+1}^{C} p(j) \tag{7.65}$$

$$P_{Bn,k} = \sum_{j=C-g-b_k+1}^{C} p(j) \tag{7.66}$$

Numerical Example

Let there be $K = 2$ traffic classes. The traffic parameters for each class are specified in Table 7.1.

Table 7.1
Traffic Parameters of Two Traffic Classes in a Cell

Traffic Class 1	Traffic Class 2
$\lambda_{n,1} = 1.5$ calls/time unit	$\lambda_{n,2} = 0.9$ calls/time unit
$\lambda_{h,1} = 0.5$ calls/time unit	$\lambda_{h,2} = 0.1$ calls/time unit
$\mu_{c,1} = 2.5$ time unit^{-1}	$\mu_{c,2} = 0.9$ time unit^{-1}
$\mu_{h,1} = 1.5$ time unit^{-1}	$\mu_{h,2} = 0.1$ time unit^{-1}
$A_1 = \lambda_1/\mu_1 = 0.5$ Erlang	$A_2 = \lambda_2/\mu_2 = 1$ Erlang
$b_1 = 1$ channel/call	$b_2 = 2$ channels/call
$n_1 = C = 6$ (no restrictions)	$n_2 = C = 6$ (no restrictions)

The relative state probabilities can be calculated by using (7.63):

$q(0) = 1$

$q(1) = [A_1 b_1 q(1 - b_1) + A_2 b_2 q(1 - b_2)]/1 = A_1 b_1 q(0) = 1/2$

$q(2) = [A_1 b_1 q(2 - b_1) + A_2 b_2 q(2 - b_2)]/2 = [A_1 b_1 q(1) + A_2 b_2 q(0)]/2 = 9/8$

$q(3) = [A_1 b_1 q(3 - b_1) + A_2 b_2 q(3 - b_2)]/3 = [A_1 b_1 q(2) + A_2 b_2 q(1)]/3 = 25/48$

$q(4) = [A_1 b_1 q(4 - b_1) + A_2 b_2 q(4 - b_2)]/4 = [A_1 b_1 q(3) + A_2 b_2 q(2)]/4 = 241/384$

$q(5) = [A_1 b_1 q(5 - b_1) + A_2 b_2 q(5 - b_2)]/5 = [A_1 b_1 q(4) + A_2 b_2 q(3)]/5 = 1{,}041/3{,}840$

$q(6) = [A_1 b_1 q(6 - b_1) + A_2 b_2 q(6 - b_2)]/6 = [A_1 b_1 q(5) + A_2 b_2 q(4)]/6 = 10{,}681/46{,}080$

$$Q(6) = \sum_{j=0}^{6} q(j) = 197{,}053/46{,}080$$

Now, we can calculate absolute state probabilities according to (7.57):

$p(0) = q(0)/Q(6) = 0.23385$

$p(1) = q(1)/Q(6) = 0.11692$

$p(2) = q(2)/Q(6) = 0.26308$

$p(3) = q(3)/Q(6) = 0.12179$

$p(4) = q(4)/Q(6) = 0.14676$

$$p(5) = q(5)/Q(6) = 0.06339$$
$$p(6) = q(6)/Q(6) = 0.05420$$

Using (7.64) we obtain blocking probabilities for both traffic classes:

$$P_{Bn,1} = P_{Fh,1} = p(6) = 0.0542 = 5.42\%$$
$$P_{Bn,2} = P_{Fh,2} = p(5) + p(6) = 0.06339 + 0.0542 = 0.11759 = 11.76\%$$

The reader should note that the above values of traffic parameters and obtained blocking probabilities are given for the example's purpose (i.e., they are not for a real-network scenario).

So far, in the analysis of multirate mobile networks, we assumed state-independent Poisson processes (e.g., the number of users is many times greater than the number of channels for each traffic class), and no restrictions on the number of channels for all traffic classes. If we introduce restrictions for each class i to c_i channels in a cell, as defined by (7.61), and allow state-dependent Poisson processes, we get a generalization of the previous approach of the aggregation of the states, denoted as a convolution algorithm [7].

7.5.2.2 Convolution Algorithm

The convolution algorithm for mobile networks is described in the following steps:

1. Calculate the state probabilities of each traffic class, as if it is the only one in the system:

$$\underline{P_k} = \{p_k(0), p_k(1), \ldots, p_k(C)\}, k = 1,2,\ldots,K \tag{7.67}$$

2. For each traffic class k, calculate the aggregation state probabilities by successive convolutions of the state probabilities of all traffic classes, as defined in step 1, excepting the class k:

$$\underline{Q_{C|k}} = \underline{P_1} * \underline{P_2} * \ldots * \underline{P_{k-1}} * \underline{P_{k+1}} * \ldots * \underline{P_{K-1}} * \underline{P_K},\ k = 1,2,\ldots,K \tag{7.68}$$

where the convolution is defined as

$$\underline{P_i} * \underline{P_j} = \left\{ p_i(0)p_j(0), \sum_{m=0}^{1} p_i(m)p_j(1-m), \ldots, \sum_{m=0}^{t=\min\{C, c_i+c_j\}} p_i(m)p_j(t-m) \right. \tag{7.69}$$

3. Calculate the traffic parameters of traffic class k. So far, we used one value for call congestion, time congestion, and traffic congestion, because they are equal in a case of state-independent Poisson processes. In this case, we have to calculate each type of congestion separately, because of the state-dependent arrivals (in a general case). This is done during the convolution between the aggregation probabilities $Q_{C|k}$ and state probabilities of traffic class k:

$$\underline{Q_C^{(k)}} = \underline{Q_{C|k}} * \underline{P_k} = \left\{Q_C^{(k)}(0), Q_C^{(k)}(1), \ldots, Q_C^{(k)}(C)\right\}$$
$$Q_C^{(k)}(j) = \sum_{m \in S_k} Q_{C|k}(j-m)\, p_k(m) = \sum_{m=0}^{j} p_k(m|j) \tag{7.70}$$

where S_k = {all pairs (j, m); $m \le j \le C$ and $m \ge (c_k - b_k)$, or $j > (C - b_k)$}, while $p_k(m|j)$ is the probability that m channels in the cell are allocated to k traffic class at the total number of occupied channels j. The normalization constant is

$$Q^{(k)} = \sum_{j=0}^{C} Q_C^{(k)}(j) \tag{7.71}$$

We can calculate call congestion by using (4.67):

$$B_{c,k} = \frac{\sum_{(j,m) \in S_k} \lambda_k(m)\, p_k(m|j)}{\sum_{j=0}^{c_k} \sum_{m=0}^{j} \lambda_k(m)\, p_k(m|j)} \tag{7.72}$$

where $\lambda_k(x)$ is the arrival rate of k traffic class at x allocated channels to that class.

Time congestion of k class can be calculated by using its definition (4.68), as

$$B_{t,k} = \frac{\sum_{j \in S_k} Q_C^{(k)}(j)}{Q^{(k)}} \tag{7.73}$$

Finally, carried traffic by k traffic class can be calculated according to relation (4.69):

$$Y_k = \sum_{i=0}^{c_k} \sum_{j=0}^{i} j p_k(j|i) \tag{7.74}$$

Then, we can calculate traffic congestion by using (4.71) as

$$B_{T,k} = \frac{A_k - Y_k}{A_k} \tag{7.75}$$

where A_k is the offered traffic to the cell by k class.

According to [7], the calculation time of the convolution algorithm is linear in the number of traffic classes K and quadratic in the number of logical channels in the cell C. To illustrate this procedure, we give an example of numerical analysis.

Numerical Example

We use the same parameters from Table 7.1. We analyze the blocking probabilities with no restrictions on the number of channels to be able to compare the results from the convolution algorithm to those of the aggregation algorithm from the previous section.

Table 7.2 shows the convolution algorithm for two traffic classes.

Using the results from Table 7.2, we calculate time congestion of each traffic class by using (7.73) and global state probabilities:

$$B_{t,1} = p_{12}(6) = 0.0542 = 5.42\%$$
$$B_{t,2} = p_{12}(5) + p_{12}(6) = 0.1176 = 11.76\%$$

From (7.72), we find out that call congestion equals the time congestion, because $\lambda_k(x) = \lambda_k$ for all x, $k = 1, 2$ (we assumed state-independent Poisson processes). Thus,

Table 7.2
The Convolution Algorithm for Two Traffic Classes

State i	$p_1(i)$	$p_2(i)$	$q_{12}(i)$	$p_{12}(i)$
0	0.6065	0.3750	0.2274	0.2338
1	0.3033	0	0.1137	0.1169
2	0.0758	0.3750	0.2559	0.2631
3	0.0126	0	0.1185	0.1218
4	0.0016	0.1875	0.1427	0.1468
5	0.0002	0	0.0617	0.0634
6	0.0000	0.0625	0.0527	0.0542
$\sum$	1	1	0.9726	1

$$B_{c,1} = B_{t,1} = 5.42\%$$
$$B_{c,2} = B_{t,2} = 11.76\%$$

For traffic congestion, from (7.74) we obtain

$$Y_1 = \sum_{i=0}^{6}\sum_{j=0}^{i} j p_1(j|i) = \sum_{i=0}^{6}\sum_{j=0}^{i} p_2(j)(i-j)\,p_1(i-j) = 0.4599\ [\text{Erlang}]$$

$$B_{T,1} = \frac{A_1 - Y_1}{A_1} = 0.0802 = 8.02\%$$

$$Y_2 = \sum_{i=0}^{6}\sum_{j=0}^{i} j p_2(j|i) = \sum_{i=0}^{6}\sum_{j=0}^{i} p_1(j)(i-j)\,p_2(i-j) = 1.716\ [\text{Erlang}]$$

$$B_{T,2} = \frac{A_2 - Y_2}{A_2} = 0.1418 = 14.18\%$$

One should note that, in the above example, the carried traffic Y_2 and the offered traffic $A_2 = 2$ [Erlang] are evaluated considering the number of occupied channels per call of traffic class 2 ($b_2 = 2$ channels/call). The value $A_2 = 1$ [Erlang], given in Table 7.1, is calculated by using the call arrival and call departure intensity (in that case, two channels are considered as one logical channel for traffic class 2).

Time congestion is usually used as a congestion measure. It is exactly the same as blocking probability calculated by the aggregation algorithm in the previous section. Time congestion equals call and traffic congestion in the case of independent Poisson arrivals and equal bandwidth requirements of all traffic classes. But, in the case of different bandwidth requirements of each class and/or class limitations, time, call, and traffic congestion of the same class may differ. The aggregation algorithm is only able to calculate the time congestion for a traffic class. On the other hand, the convolution algorithm calculates all three types of congestion. Also, it allows class restrictions (e.g., by specifying $c_k < C$ for some traffic class) and different types of arrivals (e.g., dependent Poisson arrivals, such as Engset streams). Thus, we may consider the convolution algorithm as a generalized one for dimensioning multiclass multirate networks.

7.6 Traffic Analysis of CDMA Networks

In previous sections we analyzed single class and multiple class networks with hard capacity. However, some technologies in 2G and 3G mobile networks have

soft capacity (refer to definitions of hard and soft capacity in Section 6.8.3). We want to calculate maximum traffic intensity that can be supported with a given blocking probability. The traffic intensity can be measured in Erlangs as follows:

$$\text{Traffic intensity}\left[\text{Erlang}\right] = \frac{\text{Call arrival rate}\left[\text{calls/second}\right]}{\text{Call departure rate}\left[\text{calls/second}\right]} \quad (7.76)$$

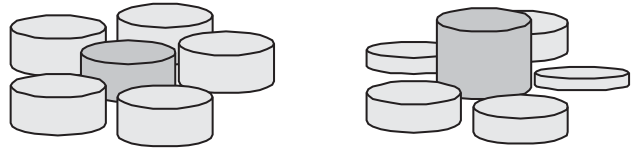
If the capacity is hard blocked (i.e., limited by the amount of hardware), than the capacity could be obtained from the Erlang-B formula. But, if the maximum capacity is limited by the amount of interference in the air interface, it is by definition soft capacity, since there is no single fixed value for the maximum capacity. For a soft capacity limited system, the capacity cannot be calculated directly from the Erlang-B formula, because it would give too pessimistic results. In the case of soft capacity, the total channel pool is larger than just the average number of channels per cell, since the adjacent cells share part of the same interface. Therefore, more traffic can be offered with the same blocking probability than in the hard blocking case.

The soft capacity can be explained as follows: The less interference that comes from the neighboring cells, the more channels there are available in the middle cell, as shown in Figure 7.5.

With a low number of channels per cell (i.e., for high bit rate real-time data users), the average traffic load must be quite low to guarantee low blocking probability, leaving extra capacity available in the neighboring cells, which can be borrowed in the middle cell, and therefore, interference sharing gives soft capacity. The soft capacity is important for high bit rate real-time applications, such as video streaming.

7.6.1 Capacity Analysis of CDMA Network

We now provide the theoretical background for the calculation of the capacity in a CDMA network. The number of simultaneous connections that may exist



Soft capacity of the middle cell when all cells are equally loaded

Soft capacity of the middle cell when less interference is coming from the adjacent cells

Figure 7.5 Soft capacity in a CDMA network.

with an acceptable level of interference defines the capacity of the cell. We consider CDMA's uplink and downlink capacity separately.

7.6.1.1 Uplink Capacity

Uplink capacity of a CDMA network is limited by the amount of interference caused by mobile terminals that can be tolerated in a given cell. Interference in the uplink consists of superposition of the signals from mobile terminals at the base station. Therefore, power control needs to be applied at mobile terminals. Such power control tends to minimize all users' received signal powers at the base station, while maintaining satisfactory wireless link performance. The main interference comes from the same cell users, but much interference also arrives at the base station from the users in adjacent cells (i.e., users attached to the neighboring base stations). In the following calculation of uplink capacity we assume identical users and ideal power control.

We suppose that a receiver for each user at the base station can operate at a bit-energy to noise-density level of E_b/N_0. Calculation of the soft capacity requires knowledge about the uplink load factor and its calculation. Therefore, we derive the uplink load equation in this section. Energy per bit from a user j can be calculated by

$$\left(E_b\right)_j = G_j P_j / W \tag{7.77}$$

where G_j is the processing gain of user j, P_j is the received signal power from user j, and W is the entire spread-spectrum bandwidth rate per processing gain. Total noise density is calculated by

$$N_0 = I_{total}/W \tag{7.78}$$

where I_{total} is the total received wideband power including the thermal noise in the base station. We may consider that total interference includes own-cell interference I_{own}, other-cell interference I_{other}, and thermal noise I_0:

$$I_{total} = I_{own} + I_{other} + I_0 \tag{7.79}$$

So, energy per bit divided by the noise spectral density (i.e., E_b/N_0), or in other words, carrier to interference ratio C/I, from a user j is defined as

$$\left(E_b/N_0\right)_j = \left(C/I\right)_j = G_j \frac{P_j}{I_{total} - P_j} \tag{7.80}$$

We may write (7.80) in the following form:

$$\left(E_b/N_0\right)_j = \frac{W}{\nu_j R_j} \cdot \frac{P_j}{I_{total} - P_j} \tag{7.81}$$

where W is the chip rate, v_j is the activity factor of user j [9], and R_j is the bit rate of user j. From the last equation, after some algebra we obtain P_j as follows:

$$P_j = \frac{1}{1 + \frac{W}{\left(E_b/N_0\right)_j \cdot R_j \cdot \nu_j}} I_{total} = L_j I_{total} \tag{7.82}$$

In the above relation we introduced a factor L_j for user j, defined as $L_j = P_j/I_{total}$. We refer to this factor as to a load factor for user j. The total received interference is a sum of the thermal noise I_0, the other-cell interference I_{other}, and the received powers from all N users in the same cell:

$$I_{total} = \sum_{j=1}^{N} P_j + I_{other} + I_0 = \sum_{j=1}^{N} L_j \cdot I_{total} + I_{other} + I_0 \tag{7.83}$$

We define a load-factor parameter as the ratio of the own-cell + other-cell interference [i.e., (I_{own} + I_{other})] and the total noise in the cell I_{totab} as given by

$$\begin{aligned} \eta_{UL} &= \frac{I_{own} + I_{other}}{I_{total}} = \frac{I_{own}}{I_{total}}\left(1 + \frac{I_{other}}{I_{own}}\right) \\ &= \frac{\sum_{j=1}^{N} L_j I_{total}}{I_{total}}(1+i) = (1+i)\sum_{j=1}^{N} L_j \end{aligned} \tag{7.84}$$

In the last equation we introduced additional parameter i, which takes into account the interference from the other (neighboring) cells. It is defined as

$$i = \frac{I_{other}}{I_{own}} \tag{7.85}$$

If we replace (7.82) into (7.84) we get the following relation for the load-factor:

$$\eta_{UL} = (1+i)\sum_{J=1}^{N} \frac{1}{1 + \frac{W}{\left(E_b/N_0\right)_j \cdot R_j \cdot \nu_j}} \tag{7.86}$$

Furthermore, we define a parameter called noise-rise as the ratio between the total received wideband power and the thermal noise power. Then, using (7.83) we obtain

$$\begin{aligned} Noise\ rise &= \frac{I_{total}}{I_0} = \frac{I_{total}}{I_{total} - I_{own} - I_{other}} \\ &= \frac{1}{1-\left(I_{own} + I_{other}\right)/I_{total}} = \frac{1}{1-\eta_{UL}} \end{aligned} \tag{7.87}$$

The load factor shows the noise rise over the thermal noise due to interference. When η_{UL} becomes close to 1, the corresponding noise rise approaches infinity and the system has reached its pole capacity. Using the last equation, we obtain the so-called pole equation:

$$P_i = \frac{1}{1+\dfrac{G_i}{\left(E_b/N_0\right)_i}} \frac{I_0}{\left(1-\eta_{UL}\right)} \tag{7.88}$$

The pole equation (7.88) shows that for a given spreading processing gain (spreading factor) a critical number of users exist. The required minimum distance that should be maintained from the critical loading of the network is specified by a maximum allowed uplink load factor η_{UL}.

The analytical framework for the uplink load, presented in this section, is commonly used to make a semi-analytical prediction of the average capacity of a CDMA cell and planning noise rise in the dimensioning process. The reason for showing uplink load factor first is the fact that the soft capacity algorithm uses uplink load factor assumption for calculating the number of channels in a cell for certain type of service (i.e., bit rate).

7.6.1.2 Downlink Capacity

In the downlink direction we use a similar approach as in the uplink. An additional new parameter is the orthogonality factor in the downlink, which is denoted as α_j. CDMA employs orthogonal codes in the downlink to separate users. If there is no multipath propagation, then the orthogonality remains when the mobile receives the signal from the base station. In the no-multipath case $\alpha_j = 1$. Multipath always exists, however; thus orthogonality is usually between 0.4 and 0.9. In the downlink, the other-to-own-cell interference ratio depends on the user location and therefore is different for each user j. Therefore, we denote it as i_j. Similar to the uplink load factor, we define a downlink load factor on a similar principle as follows:

$$\eta_{DL} = \sum_{j=1}^{N} \nu_j \cdot \frac{\left(E_b/N_0\right)_j}{W/R_j} \cdot \left[\left(1-\alpha_j\right) + i_j\right] \tag{7.89}$$

where $-10 \cdot \log_{10}(1-\eta_{DL})$ is equal to the noise rise over thermal noise due to the multiple access interference. The load factor can be approximated by its average value across the cell as follows:

$$\overline{\eta_{DL}} = \sum_{j=1}^{N} \nu_j \cdot \frac{\left(E_b/N_0\right)_j}{W/R_j}\left[\left(1-\bar{\alpha}\right) + \bar{i}\right] \tag{7.90}$$

7.6.1.3 Traffic Analysis

Because the traffic analysis in the downlink includes different values of the parameter *i*, depending on the user location, too many assumptions have to be made, thus leading to no basic conclusion. Therefore, we conduct traffic analysis in the uplink only, through analysis of the uplink load factor and noise rise. In this analysis we consider 3G WCDMA mobile networks. Different services are included in the analysis, defined by the data rate and signal-to-noise ratio E_b/N_0. Instead of showing the number of users *N*, the total data throughput per cell of all simultaneous users is shown.

Examples of uplink load factor and noise rise are shown in Figures 7.6 and 7.7, respectively. The figures are obtained using the E_b/N_0 for different

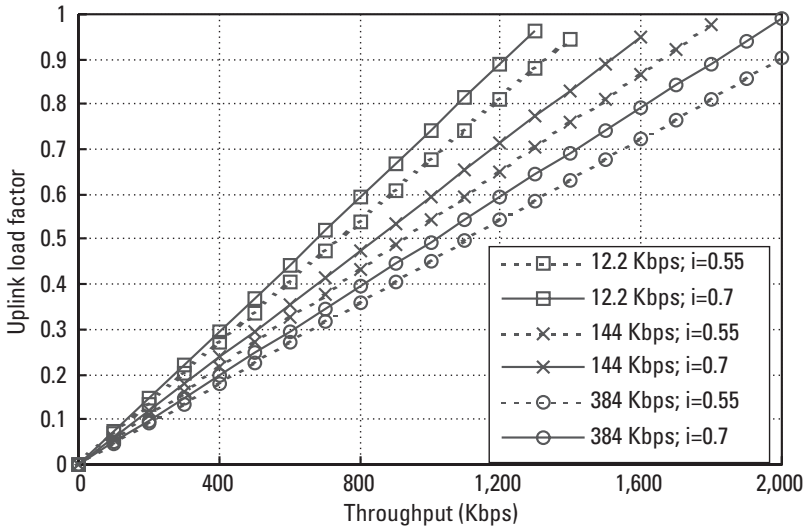


Figure 7.6 Uplink load factor versus uplink data throughput for different *i*.

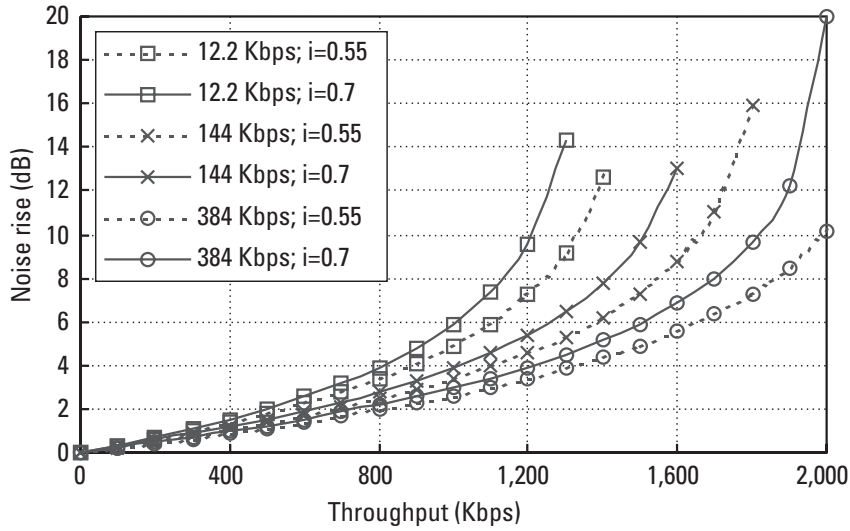


Figure 7.7 Uplink noise rise versus uplink data throughput for different *i*.

services from Table 7.3. We consider different data rates (i.e., services): 12.2 Kbps (for voice service), 144 Kbps and 384 Kbps (for data services). The uplink parameters are analyzed for different values of *i*. For example, from Figures 7.6 and 7.7 we get that, for 144-Kbps service data rate, noise rise of 4.6 dB corresponds to a 65% load factor, and noise rise of 9.6 dB corresponds to 89% load factor. For the same service, a throughput of 800 Kbps can be supported with 2.8-dB noise rise, and 1,500 Kbps with 9.6-dB noise rise.

Uplink load factor increases as data rate decreases. It is highest for voice service. This is because voice service has the highest E_b/N_0, as well as the highest throughput. The latter is directly related to the number of supported users N. The voice service can support more users than services with higher bit rates. Figure 7.6 shows that as interference from the adjacent cells increases, load factor increases as well, because the interference affects the loading in a certain cell.

Table 7.3
E_b/N_0 for Different Services in WCDMA Network

Service	E_b/N_0 [dB]
12.2-Kbps voice service	4
144-Kbps real-time data	1.5
384-Kbps nonreal-time data	1

The voice service also has the highest noise rise, as shown in Figure 7.7. For a certain throughput it can serve the largest number of users, causing a rise in interference. The noise rise is higher for higher values of the i parameter, because in such cases more interference is coming from the neighboring cells, thus influencing the uplink noise rise.

7.6.2 Calculation of the Soft Capacity

In the soft capacity calculation algorithm, shown in this section, we assume that we have the same number of subscribers in all cells. Also, we assume that call arrivals (start of the connections) and call departures are Poisson processes.

We defined soft capacity in Chapter 6 (6.27). Hence, soft capacity is defined as the increase of Erlang capacity by applying soft blocking instead of hard blocking with the same maximum number of channels per cell on average.

We can approximate uplink soft capacity by using a calculation of the total interference at the base station. The total interference includes both own-cell and other-cell interference. Therefore, the total channel pool that is offered to the users in the cell can be obtained by multiplying the number of channels per cell in the equally loaded case by $1 + i$:

$$\begin{aligned} i + 1 &= \frac{I_{other}}{I_{own}} + 1 = \frac{P_S / I_{own}}{P_S / \left(I_{other} + I_{own}\right)} \\ &= \frac{\left(C/I\right)_{\text{isolated cell}}}{\left(C/I\right)_{\text{multicell}}} = \frac{\text{isolated cell capacity}}{\text{multicell capacity}} \end{aligned} \tag{7.91}$$

where P_s is the receiving power of the signal from the mobile terminal at the base station.

We can use the Erlang-B formula for estimation of the soft capacity by applying it to the extended channel pool. To be able to apply the Erlang-B formula, we assume that there is a single traffic type (i.e., bit rate) at a time. The soft capacity calculation algorithm is summarized as follows:

1. Using the equation for the uplink load factor (7.86), calculate the number of channels per cell, N, in the equally loaded case.
2. Calculate the maximum offered traffic from the Erlang-B formula in order to obtain the Erlang capacity in the hard blocking case.
3. Multiply the number of channels with hard blocking by $1 + i$ to obtain the total number of channels with soft blocking.
4. Calculate the maximum offered traffic from the Erlang-B formula in the soft blocking case.

5. Divide the Erlang capacity by $1 + i$ in order to obtain the Erlang capacity in the soft blocking case.
6. Calculate the soft capacity using (6.27).

We should mention once more that the above procedure gives only an estimation of the soft capacity, because it is based on the assumptions of equally loaded cells and a single traffic type in the cell. It can be used, however, in dimensioning and planning of WCDMA networks. We illustrate this algorithm via numerical examples in the following section.

7.6.3 Numerical Analysis

For the purpose of giving numerical examples, we implemented the soft capacity calculation algorithm described in the previous section. It works on a recursive basis. In numerical examples we use the parameters given in Table 7.4 [10].

We show soft capacity calculations for different values of the factor i (i.e., $i = 0.3$, 0.55, and 0.7) in Figures 7.8 through 7.10, respectively. The calculations are performed for different user data rates, ranging from 12 Kbps up to 512 Kbps, and for different blocking probabilities, as listed in Table 7.4.

The results of the calculations show that soft capacity is a function of bit rate for real-time connections and is increasing with the service bit rate. The real-time connections with high bit rate get more soft capacity than the ones with low bit rate. It can be noted that the soft capacity value decreases as the

Table 7.4
Assumptions in the Soft Capacity Calculations

Bit rates	Voice: 12.2 Kbps
	Real-time data: 16–512 Kbps
Voice activity factor ()	Voice: 67%
	Data: 100%
Signal-to-noise ratio (E_b/N_0)	Voice: 4 dB
	Data 16–32 Kbps: 3 dB
	Data 64 Kbps: 2 dB
	Data 144–512 Kbps: 1.5 dB
Other-to-own-cell interference (i)	0.3; 0.55; 0.7
Noise rise	3 dB
Blocking probability	1%; 2%; 3%; 5%; 7%; 10%

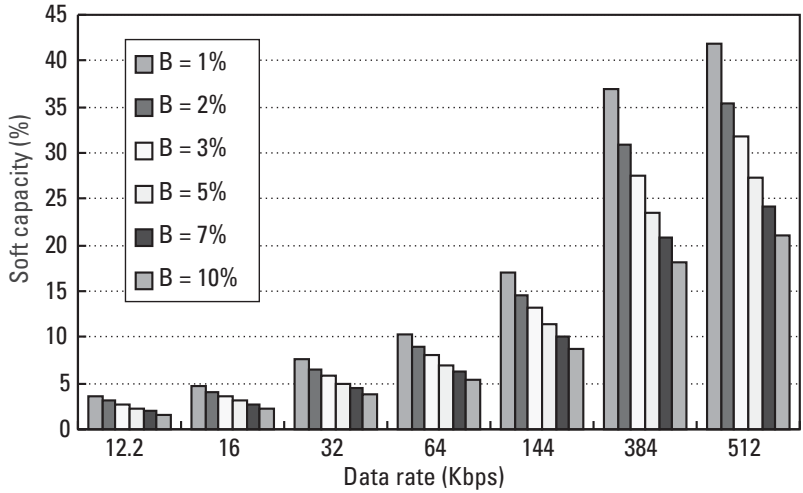


Figure 7.8 Soft capacity versus service data rate for $i = 0.3$.

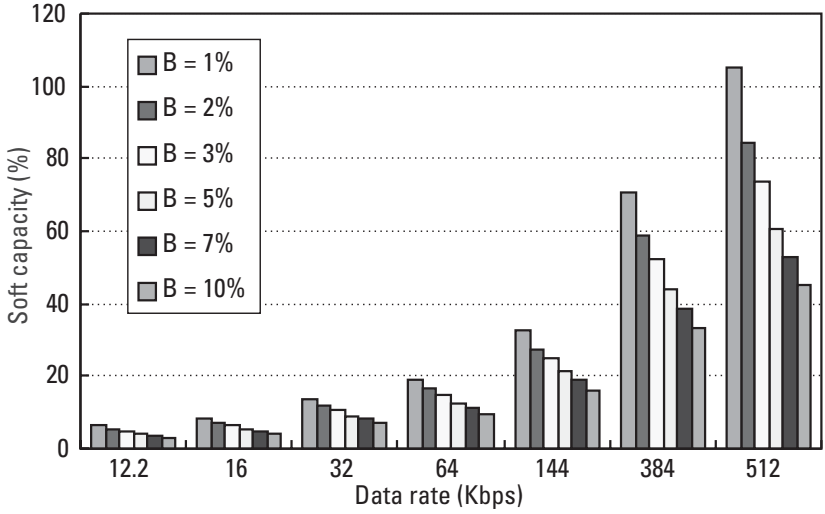


Figure 7.9 Soft capacity versus service data rate for $i = 0.55$.

blocking probability increases. This is because more traffic can be served in the neighboring cells affecting the air interface capacity in the middle one, thus leading to less soft capacity. The same conclusions are valid in all three cases (i.e., for different values of the i parameter).

One may notice that soft capacity goes above 100% for $i = 0.55$ and $i = 0.7$ in Figures 7.9 and 7.10, when we use 512-Kbps service's data rate and a blocking probability of 1% in the former and 1% and 2% in the latter cases.

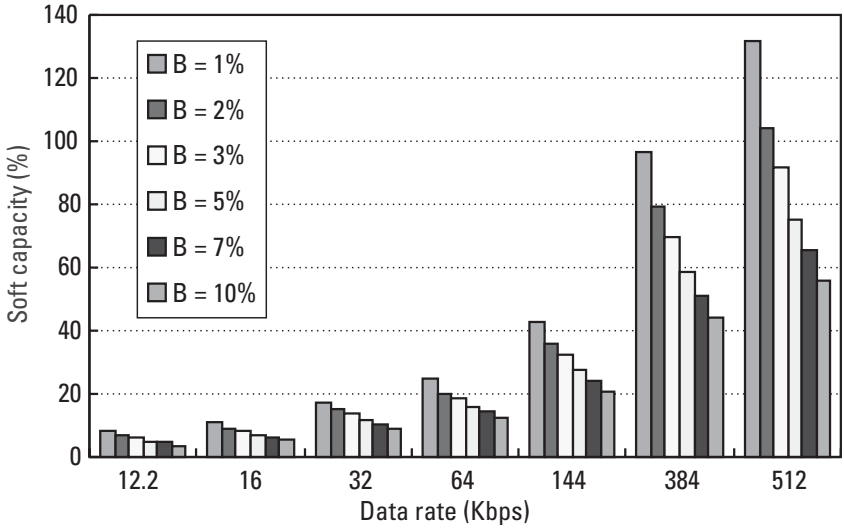


Figure 7.10 Soft Capacity versus service data rate for $i = 0.7$.

The main reason for such results is the feature of small trunk groups (e.g., logical channels, slots) to have lower trunk efficiency than larger trunk groups, and therefore soft capacity is higher. Soft capacity is very useful for high bit rate services, which cause lower trunk efficiency.

Furthermore, from the figures we may conclude that soft capacity increases as the other-to-own-cell interference factor (i) increases. A higher i value means that more capacity can be borrowed from the neighboring cells, thus leading to higher soft capacity.

In this analysis we showed the total data throughput per cell, instead of showing the number of users. The number of users, however, can be obtained by dividing the total throughput with service data rate (we assumed single service data rate at a time). For example, if we consider 144-Kbps data rate, then total throughput of 1,440 Kbps will correspond to 1,440 Kbps / 144 Kbps = 10 users.

Thus, we obtained soft capacities for different blocking probabilities by using the iterative calculations and exploiting the Erlang-B formula. These results may be used for dimensioning the WCDMA wireless links or in the admission control algorithms.

7.7 Discussion

In this chapter we created an analytical framework for analysis of mobile networks with single traffic type (e.g., voice traffic) and with multiple traffic types (i.e., multimedia networks). We can use analytical and/or simulation techniques

for dimensioning and design of the mobile networks. There are two main requirements at the design phase of a telecommunications network:

1. High quality of service, which satisfies the users;
2. Maximum utilization of network resources, which increases the revenue of the mobile operator.

Both requirements are opposite, and therefore dimensioning and design of mobile networks should balance them.

Traffic analysis and network dimensioning differs between circuit-switched mobile networks such as 2G, and heterogeneous wireless networks such as 3G. While in the former case we can use the Erlang-B formula adapted to mobile environment (i.e., considering handovers), in 3G mobile networks we have a multimedia environment and different wireless access techniques (e.g., combinations of FDMA, TDMA, and CDMA). Multimedia networks are supposed to support different services with different traffic parameters and bandwidth requirements. For their analysis we can exploit the multidimensional Erlang-B formula, the aggregation method, or the more generalized convolution algorithm. Additionally, CDMA networks use soft capacity as well (e.g., IS-95 in 2G, and WCDMA and cdma2000 in 3G), and it should be considered in the traffic analysis. We can also use the Erlang-B formula for iterative calculation of soft capacity in CDMA networks.

The purpose of traffic analysis is to model network behavior under different traffic conditions. Usually, mobile networks are dimensioned at given constraints on QoS parameters, such as new call blocking probability, handover-blocking probability, call-dropping probability, and average holding time. Then, network optimization is needed to provide the highest utilization of the resources under the given QoS constraints.

In a real network, QoS is supported by admission control as well as congestion control (Section 8.7.6). While admission control admits/rejects call/handover requests, congestion control is targeted to the modification of allocated resources of the ongoing connections. The latter approach is supported in 3G mobile networks via PDP context modification. PDP context is modified when requested QoS falls outside the limits that were granted at the PDP context activation, due to network congestion (e.g., data bursts) or because of changes in the application needed (e.g., AMR voice codec). Although such QoS modifications of the ongoing connections are not analytically tractable, one should be aware of the issue.

References

[1] Lam, D., D. C. Cox, and J. Widom, "Teletraffic Modeling for Personal Communications Services," *IEEE Communications Magazine*, Vol. 35, No. 2, February 1997.

[2] Hong, D., and S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures," *IEEE Trans. on Vehicular Technology*, Vol. VT-35, No. 3, August 1986.

[3] Valko, A., and A. Campbell, "An Efficiency Limit of Cellular Mobile Systems," *Computer Communication Journal*, Special Issue on Recent Advances in Mobile Communications Networks, 1999.

[4] Ramjee, R., R. Nagarajan, and D. Towsley, "On Optimal Call Admission Control in Cellular Networks," *Wireless Networks Journal*, Vol. 3, No. 1, 1997.

[5] Kleinrock, L., *Queuing Systems, Vol. I: Theory*, New York: John Wiley & Sons, 1975.

[6] Haring, G., et al., "Loss Formulas and Their Application to Optimization for Cellular Networks," *IEEE Trans. on Vehicular Technology*, Vol. 50, No. 3, May 2001.

[7] Iversen, V. B., *Teletraffic Engineering Handbook*, ITU-D SG 2/16 & ITC Draft, June 2001.

[8] Hlavacs, H., et al., "Modeling Resource Management for Multi-Class Traffic in Mobile Cellular Networks," *35th Hawaii International Conference on System Sciences*, 2002.

[9] Viterbi, A. J., *CDMA: Principles and Spread Spectrum Communication*, Reading, MA: Addison-Wesley, 1995.

[10] Holma, H., and A. Toskala, *WCDMA for UMTS*, New York: John Wiley & Sons, 2001.

8

Admission Control with QoS Support in Wireless IP Networks

8.1 Introduction

Multimedia wireless IP networks should provide transparent communication between the mobile terminal and the wired IP network. Our main goal is QoS support of the multimedia applications (voice, interactive applications, audio, video, and data) in mobile hosts. In such a scenario one needs appropriate classification of the traffic into classes. However, one should not separate the traffic into too many classes because that causes lower bandwidth utilization. Furthermore, wireless IP networks should be easy for implementation. Therefore, we classify IP traffic into two main traffic classes (Chapter 5): class A (for applications with specified QoS) and class B (for those without any QoS guarantees). Usually, the simplest solution is the best one. Due to the diversity of the applications and their different QoS demands, we further divide the class A into three subclasses: A1, for *constant bit rate* (CBR) applications; A2, for *variable bit rate* (VBR) applications; and A3, for best-effort applications with minimum service guarantees (we will denote it as BEmin).

QoS parameters used in the evaluation of the admission control on a call level basis are new call blocking probability and forced call termination probability (here, termination is a result of the handover blocking only). But, due to the multimedia nature of the traffic that is expected in wireless IP networks and its classification in classes and subclasses, we should also define QoS parameters on a packet level.

Call admission control (CAC) in the first and second generation of commercial mobile networks is considered on a call-level basis [1, 2]. In [2] the

authors analyze algorithms for the minimization of new call blocking probability under given constraints on the call dropping probability, as well as algorithms for the minimization of both probabilities and network resources planning. Also, call-level QoS parameters are usually found in admission control in the case of a network with multiple traffic classes [3, 4]. Authors of [4] consider admission control in a network with multiple traffic classes, where traffic is classified into narrowband calls and calls with higher bandwidth demands. They propose adaptive distributed QoS admission control over all neighboring cells. In all these approaches, channel reservations in the wireless link for handover calls (*guard channels*) are used, because blocking of a handover call is more sensitive to the users compared to new call blocking. On the other hand, [5, 6] propose a strategy for admission control in multimedia mobile networks with variable resource reservation for handover calls. But, in all strategies for admission control in wireless networks, the problems are considered only on one level, either a call-level or a packet-level, for both cases, with one or multiple traffic classes.

In this chapter we define an efficient algorithm for admission control in wireless IP networks with QoS support for different traffic classes, using the proposed classification of IP traffic.

8.2 System Model

First, we consider QoS support on a call-level. As usual, we assume that users are more sensitive to the dropping of an already-established call. Using this assumption, one can conclude that call-dropping probability should be lower than new call blocking probability. Then, one should specify the grade of service, which is defined as a maximum allowed call dropping probability in the network. One way to trade between these two parameters is to limit the call dropping probability by reducing the number of channels (e.g., time slots in TDMA) that can be allocated to the new calls of the high-priority class A. At the same time, the new call blocking probability will increase.

Different A class connections may demand different bandwidth shares. We group the traffic into mini-classes, based on the amount of bandwidth that is demanded by applications. So, a mini-class consists of all flows that belong to the same traffic class or subclass and have equal bandwidth demands per call. Traffic class B is compatible with today's best-effort traffic on the Internet. We propose unconditional acceptance of B-calls either new or handover, independently of the bandwidth occupied by A flows. Packets of B class flows should be sent to the wireless link only when there will be no A packets in the queue. Both A3 packets and B packets are served in FCFS manner, but A3 packets have priority compared to all B packets. Actually, A3 traffic is targeted to nonreal-time

applications. However, most of these applications are interactive and they do have demands with regard to the packet delay. Therefore, one should consider A3 packet's delay as a QoS parameter. The question is how to combine different QoS demands in the admission control algorithm.

If one uses simple resource reservation by giving priority to A1 and A2 traffic, it may lead to a monopolization of the bandwidth by them as well as high delays of A3 packets. This problem intensifies with an increase of the intensity of A1 or A2 traffic. Subclass A3 is assumed to offer users a higher quality of the same services that are supported by B class. Because of this, there is a need to provide appropriate QoS support for A3 traffic. To avoid high delays of A3 packets, one should reserve a specified share of the wireless link bandwidth only for this subclass. The aim is to define minimum guaranteed QoS support for A3 flows. For that purpose we include packet delay of A3 packets as a parameter in the admission control. Thus, we have three main QoS parameters for the creation of an admission strategy in wireless IP networks.

Resource reservation depends on the wireless access technology in the network (FDMA and TDMA). However, CDMA access technology (e.g., used for UMTS and cdma2000) needs admission control based on interference power (Section 8.7). Here we refer to the admission problem in general, and we will not enforce some specific access technology. We only consider the conditions that should be met by the wireless medium.

In second generation cellular networks, separate frequency carriers are used for each direction, uplink (mobile to network) and downlink (network to mobile). In wireless local networks one carrier is used in both directions (CSMA/CA principle). A possible solution for wireless IP networks is statistical multiplexing of time slots in both directions, because multimedia wireless networks may have asymmetrical traffic in uplink and downlink. However, uplink and downlink may be separated (in different bands), and then we should apply admission control in one or both directions depending upon the application request. For example, if A1 flow is admitted in the network, then it should receive its demanded bandwidth share (e.g., number of time slots in a TDMA frame). As we mentioned above, all A3 packets are served in a FCFS manner, but one can also use some implementation of a WFQ-like scheme for isolation among the flows. In contrast to A, we do not consider B class in the admission control procedure. This kind of admission control provides utilization of the bandwidth that is left by A class. So, B packets are always accepted in the network, but they have the lowest priority in the wireless link. According to this policy, B traffic is assumed to use everything that is left after servicing A packets. On the other hand, the network rejects A1 and A2 calls when there are not enough available resources in the cell. A3 flows are always accepted in the network in the same way as B packets, but they have minimum guaranteed QoS support.

8.3 Hybrid Admission Control

Because dropping of established calls is considered to be more offensive to users, many algorithms for admission control in wireless networks give priority to handover calls compared to originating (new) calls in the network. For that purpose, a number of channels in a cell are usually reserved for handover calls. In such algorithms, the network accepts new calls only if available bandwidth, which will be left after accepting the new call, is below a given threshold. These policies are often called guard policies.

We consider two main traffic classes A and B, but we provide QoS support only to A class, which is divided into three subclasses. Each of these subclasses is assumed to consist of one or more mini-classes. Therefore, we define thresholds for each mini-class. Because the thresholds are applicable only to A class, we will further refer to them as A thresholds. Handover calls are always accepted, while new calls are accepted only if the available bandwidth is below the correspondent A threshold of that particular mini-class. By giving priority to handover calls over new calls, one will experience a higher level of new call blocking probability. However, it is a design issue to project the needed resources for servicing certain number of users with the given grade of service.

Classic grade of service parameters in cellular networks are call blocking and call dropping probabilities. In most of the literature these parameters are basic, and they are considered for admission control only in wireless networks. In a heterogeneous environment, however, one should include QoS parameter(s) that will also consider the traffic with smaller QoS demands (e.g., Internet browsing) compared to that of the traditional voice calls. Therefore, we consider an additional grade of service parameter: packet delay of A3 traffic. Similar to the bounding of new call blocking and call dropping probabilities, one should bound A3 packet delay to achieve desired QoS support for that traffic subclass. Nonreal-time flows are mainly weak in packet delay due to the interactive nature of most of these applications.

If one considers all QoS parameters given above, the admission control algorithm should determine the A thresholds based on optimization of new call blocking probability, with given GoS for call dropping probability of A1 and A2 calls, and constraint on average packet delay of A3 traffic. We further define an algorithm for admission control in wireless IP network.

8.3.1 Hybrid Admission Control Algorithm

Wireless IP networks are assumed to provide heterogeneous multimedia services with different QoS demands. The traffic is hybrid. The admission control should be hybrid, too. Therefore, here we define an admission control algorithm called *hybrid admission control* (HAC).

Let us assume that the number of available logical channels in a given cell is N. One logical channel corresponds to the minimum bandwidth that one can allocate to an A1 or A2 call (e.g., one logical channel can be a single TDMA time slot, but also it can be 1/3 time slots or 2 time slots). We denote with b_i the number of logical channels that are demanded by a flow from mini-class i ($i = 1, 2, ..., N$). Throughout the text, A_i is used to denote the A threshold of a mini-class i. The basic scheme of the algorithm is shown in Figure 8.1.

In the HAC algorithm we give traditional priority to handovers over the new calls. At this point we will assume that handover calls are always accepted if there are enough free nonreserved resources in the cell (it will deteriorate a little when we include in HAC the average packet delay of A3 packets). A new call from mini-class i is accepted only if the sum of reserved resources in the cell and demanded bandwidth by that call is lower than A_i. In Figure 8.1, m_i denotes the number of admitted flows from mini-class i. A handover request of mini-class i is accepted in the target cell if the sum of occupied or reserved bandwidth by the other active traffic flows in the cell and requested bandwidth for that call is

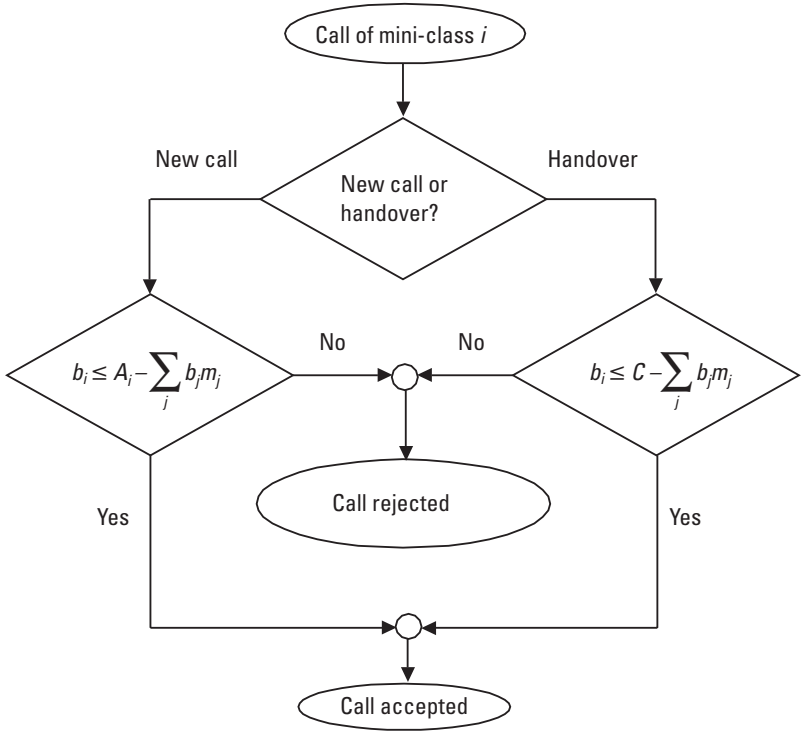


Figure 8.1 Hybrid admission control scheme in wireless IP networks with multiple traffic classes.

lower or equal to min(C, L_i), where C and L_i are the wireless link capacity and A1/A2 allowed bandwidth, respectively.

The input QoS parameters in the HAC are the call-dropping probability of A1/A2 traffic and the average packet delay of A3 traffic. With given initial input, the algorithm minimizes new call blocking probability for A1 and A2 flows. So far, we assumed different thresholds for every mini-class. However, it is possible to declare a single threshold for all of them. This requires less computation processing in the network nodes, but it decreases bandwidth utilization of the wireless link.

In the next section we present the analytical frame for analysis of the hybrid admission control algorithm.

8.4 Analytical Frame of HAC

We assume that A1 and A2 calls arrive according to the Poisson process with a given average arrival rate λ_i for mini-class i, and an average call duration distributed by exponential law with mean $1/\mu_{t,i}$. Average channel holding time is assumed to be exponentially distributed with mean $1/\mu_{h,i}$. Then, one can calculate the probability that an established call in a cell will make a handover by using the following equation:

$$P_{h,i} = \frac{\mu_{h,i}}{\mu_{t,i} + \mu_{h,i}} \tag{8.1}$$

The probability that a call will end in the current cell is calculated as $1 - P_{h,i}$. The average cell residence time for a call of mini-class i is given by

$$T_{ch,i} = \frac{1}{\mu_{t,i} + \mu_{h,i}} \tag{8.2}$$

In this analytical analysis we assume equilibrium between incoming and outgoing handovers to/from a single cell. For the purpose of traffic analysis we define a traffic model. The model for A1/A2 call analysis is shown in Figure 8.2.

Using the given traffic model for a single call, one can derive call dropping probability of mini-class i by using the following relation:

$$P_{F,i} = \sum_{j=0}^{\infty} P_i^{j+1} \left(1 - P_{Fh,i}\right)^j P_{Fh,i} \tag{8.3}$$

From (8.3), with some algebra, we get

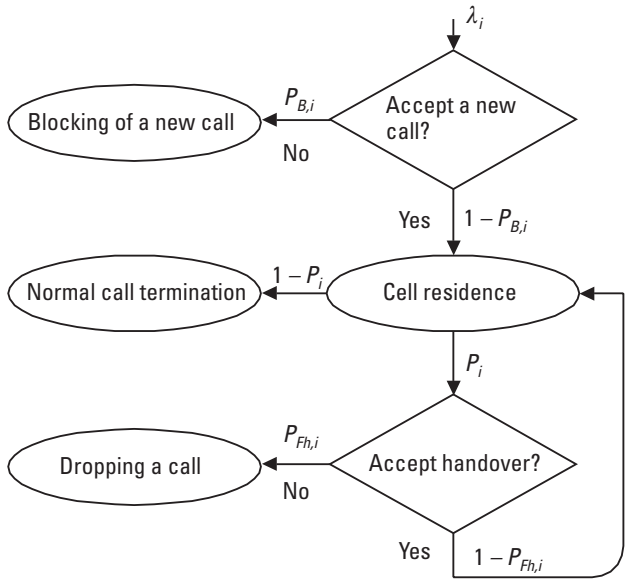


Figure 8.2 Traffic model for a call of A1/A2 subclasses.

$$P_{F,i} = \frac{P_i P_{Fh,i}}{1 - P_i\left(1 - P_{Fh,i}\right)} \tag{8.4}$$

The practical analytical method for traffic modeling on a call level is the Markov model, where each steady state in the model is defined by the number of flows from each mini-class. We denote new call intensity with λ_i, call termination intensity with μ_i, handover intensity with h_i, number of requested channels per call per mini-class with c_i, and highest number of simultaneously occupied or reserved logical channels by mini-class i with n_i. Then, one can model the wireless link as it is shown in Figure 8.3.

Let us denote the system state with a pair (i, j), where i and j are the number of flows from the first and second mini-class, respectively. Using the equilibrium assumption for all calls, we can analyze the wireless link by using a two-state Markov diagram. Then, the following conditions should be satisfied:

$$n_1 \le C,\ n_2 \le C,\ n_1 + n_2 > C \tag{8.5}$$

Also, there is a common limitation in the amount of allocated resources in a given cell with wireless link capacity C:

$$c_1 x_1 + c_2 x_2 \le C \tag{8.6}$$

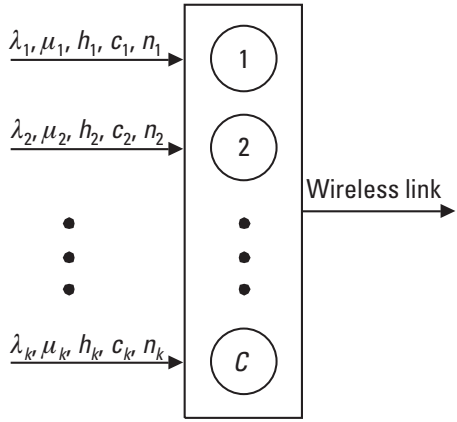


Figure 8.3 Traffic model of a wireless link in a multiclass environment.

where x_i is the number of flows from mini-class i.

For presentation purposes, we give an appropriate example of the analysis of the wireless link. In this example we set capacity of the wireless link to $C = 12$ logical channels, while the number of channels that can be allocated to a single call is $c_1 = 3$ and $c_2 = 4$ logical channels for the first and second mini-class, respectively. In this example let us assume that we have already determined optimal thresholds for each mini-class: $A_1 = 9$ logical channels (corresponds to three calls) and $A_2 = 8$ logical channels (corresponds to two calls). Then, each mini-class is allowed to have up to $\lfloor C / c_i \rfloor$ connections in a cell at the same time. So, at the same time a cell can serve up to four connections of mini-class 1, or up to three connections of mini-class 2. One should note that the settings given in the example are arbitrarily chosen, so that we can draw the Markov diagram, and they should not be considered as a practical implementation.

The Markov chain model for this particular example is shown in Figure 8.4, where $\lambda_i(x_i)$ is the incoming call rate, $\mu_i(x_i)$ is call termination rate (call ending or a handover) of mini-class i, and x_i is number of flows of mini-class i.

To define an optimal threshold per mini-class, which will not necessarily be an integer number [as is the case in [2] with their *integral guard channel policy* (IGP)], we use the principle of a partial fractional guard policy. Our target is to provide general optimal thresholds (e.g., the threshold may be 86.77 logical channels; it should not be necessarily rounded to an integer value 86 or 87). Hence, when the number of occupied logical channels in the cell is $j > \lfloor A_i / c_i \rfloor$, then all new calls are rejected from the network. In such a case we have network congestion. One can queue all incoming calls, but that approach is rarely used in wireless networks.

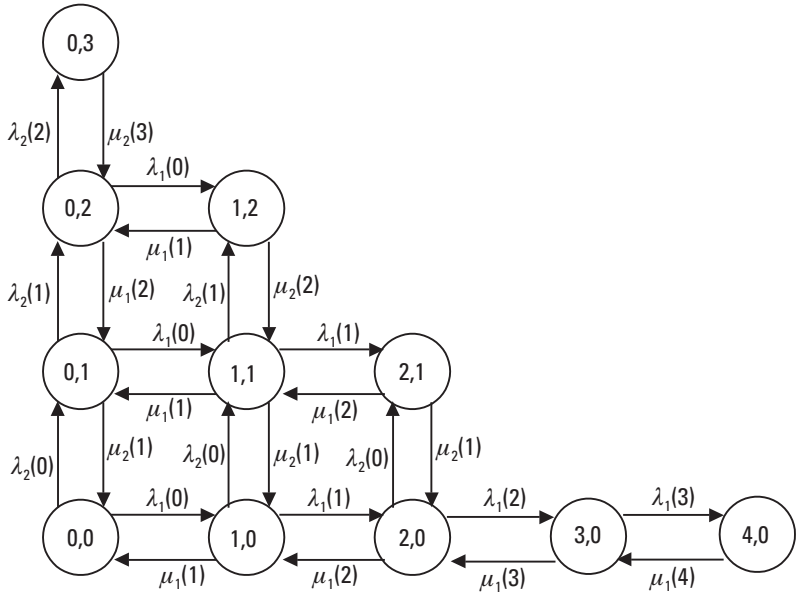


Figure 8.4 Two-dimensional Markov chain for describing bandwidth occupancy in the wireless link (two mini-classes).

Fractional guard policy is more general than IGP, which is defined and analyzed for the traditional case with one traffic class [2]. One can extend such an approach to a multimedia environment. Then, the new call intensity is given by

$$\lambda_i(j) = \begin{cases} \lambda_{n,i} + \lambda_{h,i}, 0 \le j < \lfloor A_i/c_i \rfloor \\ \beta_i \lambda_{n,i} + \lambda_{h,i}, \; j = \lfloor A_i/c_i \rfloor \\ \lambda_{h,i}, \lfloor A_i/c_i \rfloor < j \le \lfloor C/c_i \rfloor \end{cases} \tag{8.7}$$

where $\lambda_{n,i}$ and $\lambda_{h,i}$ are new and handover call intensities of mini-class i, respectively, and $\beta_i = A_i/c_i - \lfloor A_i/c_i \rfloor = fract(A_i/c_i)$ is fractional part of the threshold A_i. Call termination rate is given by

$$\mu_i(j) = j(\mu_{t,i} + \mu_{h,i}) \tag{8.8}$$

where $1/\mu_{t,i}$ and $1/\mu_{h,i}$ are the average time intervals between two consecutive call terminations and average time between two consecutive handovers of mini-class i, respectively; and $1/(\mu_{t,i} + \mu_{h,i})$ is average cell residence time.

We have to check the local balance of the system under assumed equilibrium in a cell. The steady state diagram is given in Figure 8.5. From Figure 8.5, going towards the clock direction, one gets

$$P(i,j)\lambda_2(j)P(i,j+1)\lambda_1(i)P(i+1,j+1)\mu_2(j+1)P(i+1,j)\mu_2(i+1) \tag{8.9}$$

If one is going in the opposite direction, the traffic flux is given by

$$P(i,j)\lambda_1(i)P(i+1,j)\lambda_2(j)P(i+1,j+1)\mu_1(i+1)P(i,j+1)\mu_2(j+1) \tag{8.10}$$

It is easy to notice that (8.9) and (8.10) are identical, which means the local balance in the system is fulfilled. Then, one can calculate the probability that a wireless link is in a particular state by using the steady state probabilities as follows:

$$P(i+1,j) = P(i,j)\frac{\lambda_1(i,j)}{\mu_1(i+1,j)} \tag{8.11}$$

Now, one can express all steady state probabilities through the initial state of the system $P(0,0)$, which corresponds to a cell with no traffic load, as is given by

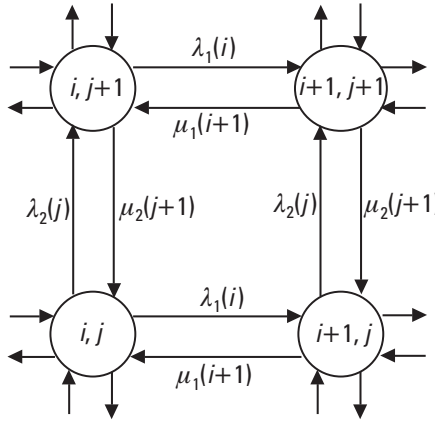


Figure 8.5 General Markov chain model.

$$P(i,j)=\frac{\lambda_1(0)\lambda_1(1)\ldots\lambda_1(i-1)\lambda_2(0)\lambda_2(1)\ldots\lambda_2(j-1)}{\mu_1(1)\mu_1(2)\ldots\mu_1(i)\mu_2(1)\mu_2(2)\ldots\mu_2(j)}\cdot P(0,0) \tag{8.12}$$

Using the above relation and normalized conditions (the sum of all steady state probabilities is equal to one), one can calculate the steady state probabilities $P(i, j)$ and then determine loss probabilities in the access network. The above analytical analysis is performed for the given example with two mini-classes.

Furthermore, one can extend the analytical framework to more then two classes. For the case of multiple traffic classes in a wireless network, we may define the probability that n logical channels in the cell are occupied by using the relation

$$P_A(n)=\sum_{\forall\left(\sum_{j=1}^{K} n_j b_j = n\right)} P(n_1,n_2,\ldots,n_K) \tag{8.13}$$

where $P(n_1, n_{2,} \ldots, n_K)$ is probability that there are n_1 flows from mini-class 1, n_2 flows from mini-class 2, ..., n_K flows from mini-class K. Then, one can calculate the new call blocking probability of the ith mini-class:

$$P_{B,i}=\sum_{n=A_i-b_i+1}^{C} P_A(n) \tag{8.14}$$

where C is bandwidth of the wireless link, and b_i is the demanded bandwidth per connection from ith mini-class (if the number of allocated time slots per a call is c_i, then $b_i = c_i$).

The handover blocking probability of A calls can be calculated by

$$P_{Fh,i}=\sum_{n=C-b_i+1}^{C} P_A(n) \tag{8.15}$$

If we apply Markov diagrams in a wireless network with multiple traffic classes and higher capacity per cell (as is expected), then one will get a huge system of equations to solve (there may be several hundred, up to thousands, of equations). This is time-consuming and processing-demanding. To avoid having to solve the multidimensional Markov diagram model, it is usually practiced to split K-dimensional Markov diagram into K one-dimensional Markov chains [4]. A common Markov chain model is shown in Figure 8.6, $i = 1, \ldots, K$, where K is the number of mini-classes.

From Figure 8.6 one can calculate the optimal threshold of mini-class i as $S_{i1} + \beta_i$, where S_{i2} is calculated by

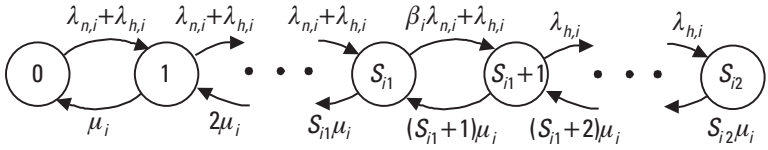


Figure 8.6 One-dimensional Markov chain for a wireless link with multiple traffic classes.

$$S_{i2} = \left\lfloor \frac{C - \sum_{j=1, j\neq i}^{K} c_j m_j}{c_i} \right\rfloor \tag{8.16}$$

where $c_j m_j$, $j = 1, \ldots, K$, is the average number of allocated resources per mini-class. Then, new call blocking and call dropping probabilities can be calculated as follows:

$$P_{B,i} = (1-\beta_i) P(S_{i1}) + \sum_{k=S_{i1}+1}^{S_{i2}} P(k)$$

$$= \frac{(1-\beta_i) \cdot \dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{S_{i1}}}{S_{i1}} + \sum_{k=S_{i1}+1}^{S_{i2}} \dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{S_{i1}} \left(\dfrac{\beta_i\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)\left(\dfrac{\lambda_{h,i}}{\mu_i}\right)^{k-(S_{i1}+1)}}{k!}}{\sum_{k=0}^{S_1} \dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{k}}{k!} + \sum_{k=S_1+1}^{S_{i2}} \dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{S_{i1}} \left(\dfrac{\beta_i\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)\left(\dfrac{\lambda_{h,i}}{\mu_i}\right)^{k-(S_{i1}+1)}}{k!}} \tag{8.17}$$

$$P_{Fh,i} = P(S_{i2}) = \frac{\dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{S_{i1}} \left(\dfrac{\beta_i\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)\left(\dfrac{\lambda_{h,i}}{\mu_i}\right)^{S_{i2}-(S_{i1}+1)}}{S_{i2}!}}{\sum_{k=0}^{S_1} \dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{k}}{k!} + \sum_{k=S_1+1}^{S_{i2}} \dfrac{\left(\dfrac{\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)^{S_{i1}} \left(\dfrac{\beta_i\lambda_{n,i}+\lambda_{h,i}}{\mu_i}\right)\left(\dfrac{\lambda_{h,i}}{\mu_i}\right)^{k-(S_{i1}+1)}}{k!}} \tag{8.18}$$

One can calculate the total incoming call intensity in the cell, denoted as Λ_i, by using the following relation:

$$\Lambda_i = \lambda_i\left(1 - P_{B,i}\right) + \lambda_{h,i}\left(1 - P_{Fh,i}\right) \tag{8.19}$$

where Λ_i is the intensity of the calls accepted in a cell. Handover intensity from a cell to its adjacent cells is given by

$$\lambda_{h,i} = P_i\Lambda_i = P_i\left[\lambda_i\left(1 - P_{B,i}\right) + \lambda_{h,i}\left(1 - P_{Fh,i}\right)\right] \tag{8.20}$$

where P_i is the probability that a given call will perform a handover before it will terminate. From (8.20) one cannot directly determine new call blocking probability and handover blocking probability. For that purpose, we should use iterative calculations, where initial values for $P_{B,i}$ and $P_{Fh,i}$ are set to zero. Then, one does iterations until both probabilities converge.

So far, we have analyzed QoS parameters of A1 and A2 subclasses, but have not referred to A3 traffic at all. But, although A3 flows have lower priority compared to A1 and A2 traffic, A3 average packet delay cannot be analyzed separately. Simply, this is a consequence of the fact that A3 flows use the remaining resources after servicing A1 and A2 flows. For the simplicity of the analysis one may assume that A3 packets arrive at wireless link buffers by a Poisson process, although this is not exactly the case (the reader should refer to the IP traffic characteristics in Chapter 5). One can use buffering of A3 packets in base stations according to the FCFS scheme, so packets that enter into the wireless link buffer first are transmitted first. The total A3 packet delay is a sum of waiting time in the buffer and transmitting time over the wireless link. According to the discussion above, one can model A3 traffic in the base station as a queue with a varying service rate. The service rate can be anything between zero and cell capacity *C*. The admission control algorithm is used to allocate a specific number of logical channels (bandwidth) for each call. Below, we discuss admission control for each subclass in A class.

A call of the A1 subclass, created primarily for real-time services with near to constant bit rate, will receive a fixed number of logical channels at the call admission in a cell. A2 is dedicated mainly to real-time flows with variable bit rate, so each call should be allowed to request the changing of its current allocated network resources. The resource allocation for A2 traffic can be either static or dynamic. One usually uses traffic shaping to smooth the burstiness of VBR traffic flows. In that case, base stations should monitor the flows and mark nonconformant packets with lower priority labels (e.g., by a token bucket). These marked packets should have same service level as B class traffic. But in both cases the bandwidth used by A1 and A2 traffic can be

viewed as near constant in the analysis of A3 or B applications since the connection duration of A1 and A2 flows is much longer than the packet service time. However, there is no bandwidth allocation for A3 (BEmin) flows, but this subclass has a priority over B packets in the base stations. One has to adjust admission control for A3 flows to be able to get their guaranteed QoS support.

Then, we can use a single server queue for A3 packets at the base station with a service rate equal to the difference between wireless link capacity and allocated bandwidth to A1 and A2 connections. If we assume the exponentially distributed packet interarrival time and the exponential distribution of packet length, then we can use the M/M/1 or M/G/1 queuing model for the analysis of A3 traffic at the base stations (for delay analysis in priority queuing, refer to Section 4.6.4). But, if all bandwidth is occupied by A1 or A2 connections, all A3 packets will be waiting in the queue (Figure 8.7). To avoid infinite delays during high network loads, one reserves a part of the bandwidth for A3 traffic only (one or more logical channels). Basically, it should be smaller part of the wireless link resources, which depends upon prediction of traffic load per class in the network. For that purpose, we introduce another threshold L_{A12}, which defines the maximum capacity allowed to A1 and A2 connections ($C - L_{A12}$ is bandwidth reserved only for A3 traffic).

Let $E[D_i]$ denote the average packet delay of A3 packets at the base station when there are i logical channels occupied by A1 and A2 flows. Then, one can calculate average packet delay by using the following:

$$E[D] = \sum_{i=0}^{C} E[D_i] P_A(i) \tag{8.21}$$

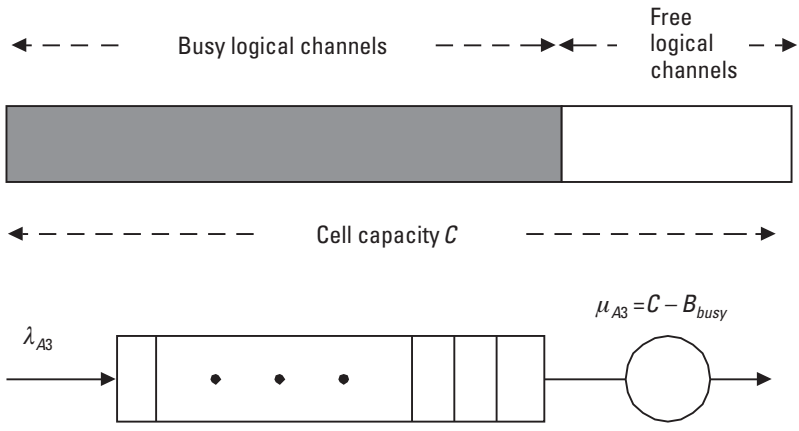


Figure 8.7 A3 packets servicing at the base station.

To satisfy grade of service, given at the network dimensioning process, we need to determine optimal A thresholds in the HAC algorithm for the admission control in the wireless network.

The thresholds are initially set at the network design phase, and later they are evaluated by using real traffic measurements. In both cases stated above, however, we need an algorithm to determine the optimal A thresholds under given constraints on call dropping probabilities of A1 and A2 classes, and average packet delay of A3. Such an algorithm should lead to the minimization of new call blocking probability while satisfying the previous two constraints.

8.5 Optimal Thresholds in HAC Algorithm

Now we determine the optimal A thresholds by minimizing the new call blocking probability. The main problem arises from various bandwidth demands of different traffic subclasses and the mini-classes within them.

Let us briefly discuss the dependence of thresholds upon given QoS parameters of A traffic. We first consider a single-class network scenario. If there is only one mini-class in the network, then moving the threshold up causes an increase of call dropping probability and a decrease of new call blocking probability, and the opposite way as well. The behavior of the average packet delay of A3 traffic is expected to be similar to that of the call dropping probability. This is not always the case because it also depends on new call and handover intensities in the network. In a multiclass wireless network one can determine one threshold or multiple thresholds. With only one A threshold, one can solve the problem of an optimal threshold by a binary search. However, the problem becomes more complex when there is more then one A threshold.

Here we propose a general procedure for obtaining multiple optimal thresholds under a given traffic classification. The steps of the procedure are outlined as follows:

1. Set call dropping probability $P_{F,i}$ and new call blocking probability $P_{B,i}$ for each mini-class i to their given maximum. Also, set average packet delay of A3 traffic to the given maximum $E[D]_{max}$.
2. Calculate the optimal threshold of mini-class i when all other thresholds are set to their maximum by using binary search algorithm: $A_j = C$ ($\lfloor C/c_j \rfloor$ calls) for $j \neq i$. Use the obtained threshold in the rest of this algorithm as initial values for the optimal thresholds search. Repeat this step for each mini-class i. Here, let us denote with $P_{Bopt,i}$ the new call blocking probability of mini-class i at optimal A_i threshold.
3. Repeat steps 4, 5, and 6 for all combinations of resource allocation per mini-class.

4. Calculate $P_{B,i}$, $P_{F,i}$ (using finite number of iterations) for A1 and A2 traffic, and $E[D]$ for A3 traffic.
5. If given conditions for the QoS parameters are satisfied (i.e., $P_{F,i} < P_{Fmax,i}$ and $E[D] < E[D]_{max}$), then if $P_{B,i} < P_{Bopt,i}$ then $P_{B,i} = P_{Bopt,i}$.
6. If $\{P_{Bi} > P_{Bi,threshold}$ and $(P_{Fi} > P_{Fi,threshold}$ or $E[D] > E[D]_{max})\}$, then go to step 7.
7. If it is not possible to determine an optimal A threshold, then it means that there are not enough resources in the wireless network for the given traffic demands or that initial constraints are too strict for at least one QoS parameter.

Exact determination of optimal thresholds necessitates the solving of the K-dimensional Markov chain model, a process that requires huge calculations. One will not want to perform this processing in real-time at the base station, due to the limited processing power of the base station and its multifunctionality in a wireless IP network. However, traffic intensity is not uniformly distributed during the day; the traffic volume changes with the time of the day. The measurements from traditional circuit networks, as well as from packet networks such as the Internet [7], show the existence of a traffic pattern during a typical weekday. We denote *main traffic volume* the time interval during the day with the highest traffic intensity. For example, in traditional circuit-switched telecommunication networks, the traffic is higher during working days compared to holidays. The peak traffic hour is usually somewhere between 12 p.m. and 3 p.m., which is geographically dependent. On the other side, the Internet may have a peak traffic hour in other periods of the day (e.g., in [7] peak traffic hour is between 12 a.m. and 1 a.m.). Because of the overwhelming processing necessary for the calculation of optimal thresholds, one can schedule this calculation at during periods of lower traffic load in the network (i.e., late at night). Base stations should be able to measure the traffic load in the access network. Then, it is possible to calculate different sets of optimal thresholds for different periods during the day. One can use the obtained optimal thresholds during the low network load until the next update. Operators determine the update rate by using traffic measurements and its structure (A1, A2, A3, or B flows). Each base station should have information of the status of each subscriber that resides within its cell(s). Such information is necessary for the admission control of A1 and A2 calls, after paging at the call initiation. On the other hand, wired nodes in the network do not need to have information on a per-flow basis. It is enough for them to have information on class/subclass bases. Wired nodes perform differentiation of the packets according to their classification (routing and location management in wireless IP networks are described in Chapter 10).

8.6 Analysis of the Admission Control in Wireless Networks

Here, we present a performance analysis of the hybrid admission control in a multiclass environment in a wireless IP network. We do experiments with different simulation scenarios by using the hybrid simulation environment evaluated in Chapter 6.

In these experiments we observe the following QoS parameters: new call blocking probability and call dropping probability of A1 and A2 subclasses, and average packet delay of A3. First, we perform analysis of A3 packet delay for different values of A threshold. In this experiment we use a single threshold for new calls of A1 and A2 subclasses. It is assumed that the base station allocates a single logical channel per call, and it is not changed during the connection duration. The following input settings are used in the experiment: cell size is set to 1 km, average velocity of the users is 50 km/hr, bit rate of the wireless link is 2 Mbps (this value is arbitrarily chosen), and A3 packets arrive with a rate of 30 packets/second with average packet length 1,000 bytes, exponentially distributed. We set a new call rate to 3 calls/hour/user. The average number of users per cell is 1,000, while the average call duration is set to 100 seconds. In the following experiments we reserve one logical channel for A3 traffic only. The capacity of a cell is set to $C = 100$ logical channels.

We analyze A3 packet delay versus A3 packet intensity for a different number of reserved logical channels for handover calls of A1 and A2. The results are shown in Figure 8.8. We conclude that the average delay of A3 packets is higher at a higher intensity of new calls, because higher traffic load occupies more of the bandwidth resources and leaves less bandwidth for servicing the A3 traffic. By increasing the number of reserved channels for A1 and A2 handovers,

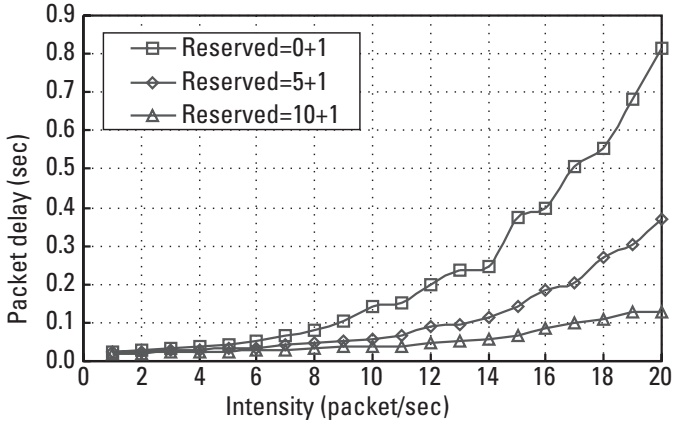


Figure 8.8 Average delay of A3 packets as a function of new call intensity in a cell for different A thresholds.

we notice a decrease of the average A3 packet delay. The main reason for this is the smaller number of admitted connections in the access network when we have more reserved bandwidth for handovers. It is a consequence of a higher number of rejected new calls at lower A thresholds. But, at the same time it means more channels for servicing A3 traffic. This conclusion is confirmed by the results in Figure 8.9, where we show new call blocking probability versus reserved logical channels for handover calls.

The average delay of A3 packets decreases while new call blocking probability increases. For lower handover intensities (e.g., 2 calls/hour/user) we do not detect blocking of a new call, and therefore, the average A3 packet delay is a constant for varying A thresholds. In Figure 8.10 we show simulation results for average packet delay as a function of the number of reserved channels for A1 or A2. As one can expect, the results show an exponential decrease of the average A3 packet delay with an increase of the number of reserved channels. Thus, lower threshold (more bandwidth is reserved for handovers only) leads to smaller average packet delay because fewer logical channels are being allocated to new A1 or A2 calls. However, a decrease of A threshold causes an increase of new call blocking probability.

Next, we show the QoS parameter behavior in a wireless network with multiple classes. For presentation purposes we consider network analysis for two scenarios: first with two mini-classes and then with three mini-classes. In the scenario with two mini-classes, the average number of arrival calls is set to 0.1 call/second, average call duration is 250 seconds, while average cell residence time of an ongoing call is 100 seconds. One can calculate that there should be

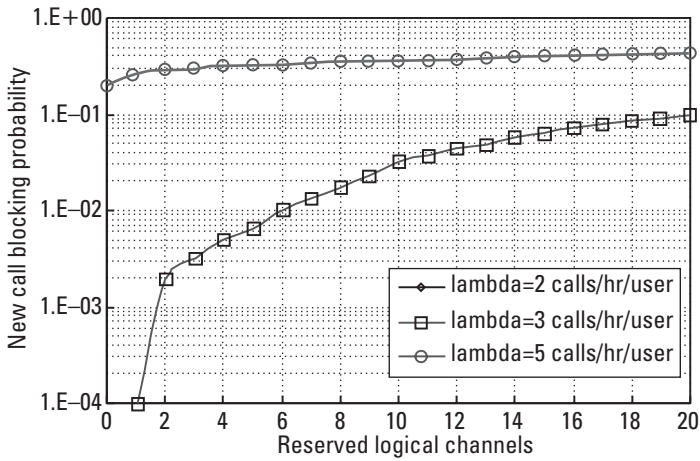


Figure 8.9 New call blocking probability for A1 and A2 subclasses versus reserved logical channels.

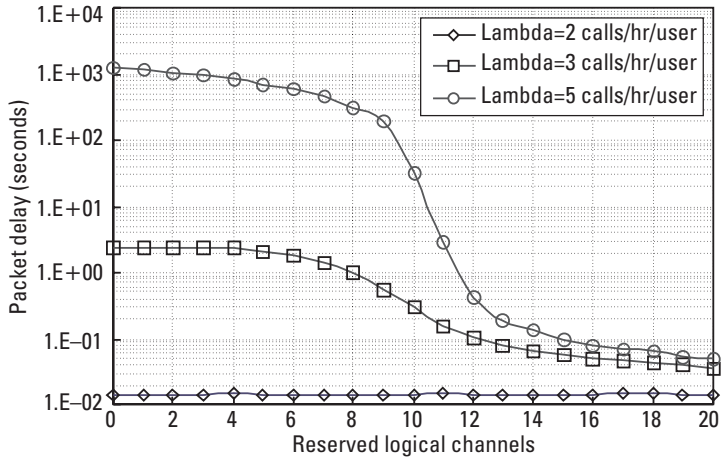


Figure 8.10 A3 packet delay for different number of logical channels reserved for handovers.

2.5 handovers per call of each mini-class. The only difference between the two scenarios is the number of allocated logical channels per call: $c_1 = 1$ channel/call, $c_2 = 2$ channel/call. For the first mini-class we allocate one logical channel per call, while two logical channels are allocated per call for the second mini-class. Here, we change A threshold simultaneously for both mini-classes. The results from simulation runs are shown in Figures 8.11 to 8.13. One can notice

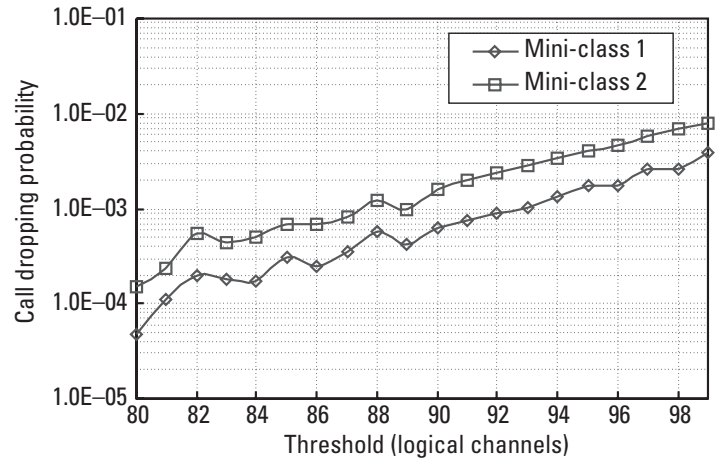


Figure 8.11 Call dropping probability as a function of the A threshold (a scenario with two mini-classes).

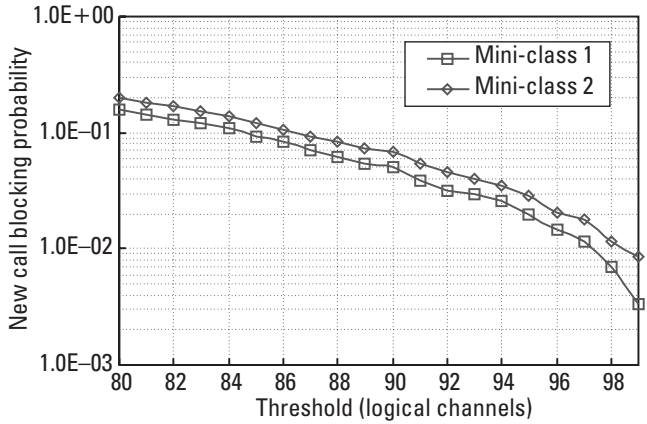


Figure 8.12 New call blocking probability versus varying threshold for a scenario with two mini-classes.

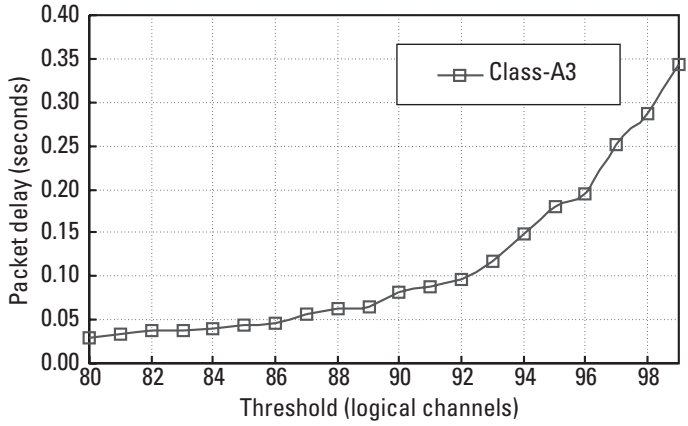


Figure 8.13 Average A3 packet delay versus varying threshold for a scenario with two mini-classes.

that both mini-classes have similar behavior considering new call blocking and call dropping probabilities (Figures 8.11 and 8.12). However, the blocking probabilities are higher for the second one.

This is because calls from the second mini-class, when compared to calls from the first mini-class, require more logical channels per call. So, calls of the second mini-class cause larger segmentation of the wireless link bandwidth and lead to lower bandwidth utilization and higher call losses, either new or handover calls.

The average packet delay of A3 in this experiment is given in Figure 8.13. It shows an exponential increase with an increase of the threshold. The

explanation for the behavior of the average packet delay is the same as the one given above. The reader should notice that in this experiment we used one A threshold for both mini-classes.

For the scenario with three mini-classes, we use the following input data: new call intensities are $\lambda_1 = 0.15$ calls/second, $\lambda_2 = 0.05$ call/second, $\lambda_3 = 0.01$ call/second; average call durations are $1/\mu_1 = 100$ seconds; $1/\mu_2 = 250$ seconds, $1/\mu_3 = 500$ seconds; average cell residence intervals are $1/h_1 = 50$ seconds, $1/h_2 = 50$ seconds, $1/h_3 = 200$ seconds; while allocated bandwidth shares are $c_1 = 1$ channel/call, $c_2 = 2$ channel/call, $c_3 = 5$ channel/call.

With the purpose of analyzing different admission control conditions in wireless IP networks, we choose to restrict bandwidth reservation for handovers of the third mini-class (i.e., its threshold is fixed at the cell capacity C). The other two mini-classes have the same varying threshold. The obtained results are given in Figures 8.14 and 8.15. Using these results, one can notice that an increase of the threshold of the first two mini-classes results in a decrease of new call blocking probability of A1 and A2 subclasses and an increase of forced call termination probability. Unlike the first two mini-classes, one notices an increase of all QoS parameters for the third mini-class. We can explain this behavior by the fact that the network accepts more new connections by increasing the threshold of the first two. However, this results in less available bandwidth for new calls and handovers of the third mini-class.

So far, we have observed the most important scenarios through the given examples above. One can continue with the analysis by adding more mini-classes. However, the results show the advantages of an applied hybrid admission control in wireless IP networks with heterogeneous traffic.

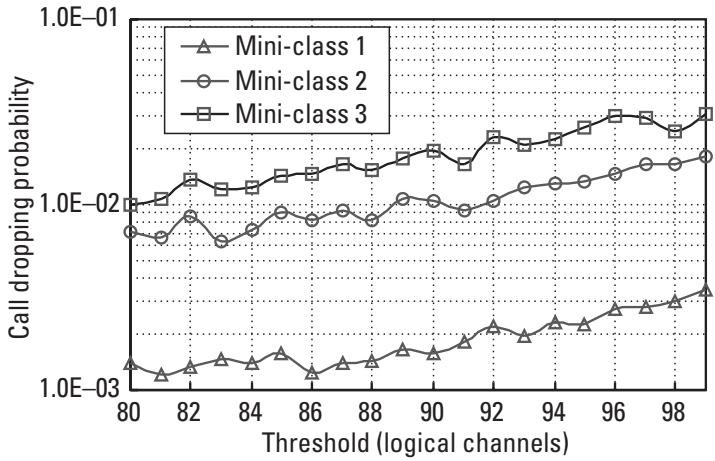


Figure 8.14 Call dropping probability for a scenario with three mini-classes.

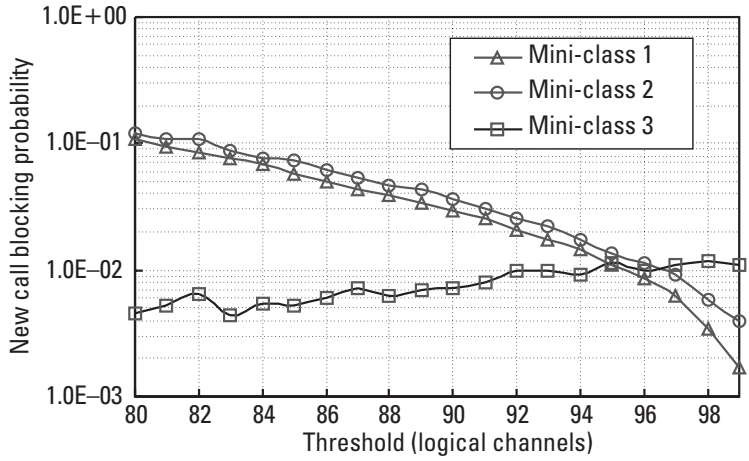


Figure 8.15 New call blocking probability for a scenario with three mini-classes.

In the following section we consider admission control in CDMA networks, due to the specific characteristics of soft handovers and soft capacity.

8.7 Admission Control in Wireless CDMA Networks

In CDMA networks the quality of ongoing connections will decline if cell interference is allowed to increase, due to the soft capacity. Therefore, we need some admission control to limit amount of interference in the system. Admission control needs to check that admission of a new connection will not sacrifice the planned coverage area or QoS of the ongoing connections. In 3G networks, such as UMTS, admission control is located at the radio network controller, where the load information from several cells can be obtained. Because many applications may request asymmetrical bandwidths in uplink and downlink, the admission control should estimate the load increase that the new connection will cause separately for uplink and downlink—that is, the admission control decision is made independently for each direction (e.g., in WCDMA-FDD or cdma2000).

In FDMA/TDMA-based mobile networks, we have prespecified the capacity per cell (i.e., hard capacity). But CDMA has no hard limit on the capacity, which makes admission control a more complex soft capacity management issue. Several admission control schemes for CDMA networks have been suggested. Generally, these admission control schemes for CDMA can be classified into the following groups:

1. *Signal-to-interference ratio* (SIR)-based admission control;

2. Load-based admission control;
3. Power-based (i.e., interference-based) admission control.

There are, however, different classifications of admission control schemes for CDMA systems. For example, another possible classification is into two types [8]: One is based on the number of users [9], and the other is based on interference level [10, 11].

8.7.1 SIR-Based Admission Control

SIR-based admission control policy is introduced in [9]. The admission control is made on an individual basis comparing mobile's SIR to a given threshold value at the base station. It refers to the uplink direction.

For cellular systems, including CDMA, radio propagation is influenced by three independent factors: path loss with distance, log-normal shadowing, and multipath fading. The average received field from a mobile at distance r from a base station can be modeled as

$$\Gamma(r) = \frac{1}{r^{\alpha}} 10^{\frac{\xi}{10}} \tag{8.22}$$

where ξ is a random variable (expressed in decibels) that has normal distribution with zero mean and standard deviation of σ, which is independent of distance and ranges 5~12 dB with a typical value of 8 dB. Typical values for α in a cellular environment are 2.7~4.

Let us denote with $P_i(h, k)$ the power received by the base station in cell k from a mobile i, which is transmitting to its servicing base station of its home cell h. Then, total received power at cell k is

$$\begin{aligned} I(k) &= \sum_{h=1}^{K} \sum_{i=1}^{n_k} P_i(h,k) + I_0 \\ &= Pn_k + P \sum_{h=1,\ h\neq k}^{K} \sum_{i=1}^{n_k} \left(\frac{r_{i,h}}{r_{i,k}} \right)^{\alpha} 10^{\frac{\xi_{i,k} - \xi_{i,h}}{10}} + I_0 \end{aligned} \tag{8.23}$$

where K is the number of cells in the CDMA system, $r_{i,h}$ is the distance between the mobile i and the base station of its home cell h, $r_{i,k}$ is the distance between the mobile i and the base station of cell k, and P is the power level received by a mobile's home cell base station. The first term in the above relation is the power generated by the users in the home cell k and who use the power P; the second

term is generated by the users in other cells and with log-normal shadowing effect; and the last term is the thermal background noise.

SIR at a given base station k is a random variable SIR_k, which is dependent upon three stochastic processes: radio propagation, traffic variation, and mobile distribution. Also, aggregate congestion at the local cell or other cells influences the SIR. The SIR value at the cell k can be expressed as

$$SIR_k = \frac{P}{I(k) - P} = \frac{1}{n_k - 1 + \sum_{h=1,\ h \neq k}^{K} \sum_{i=1}^{n_k} \left(\frac{r_{i,h}}{r_{i,k}} \right)^{\alpha} 10^{\frac{\xi_{i,k} - \xi_{i,h}}{10}} + \frac{I_0}{P}} \tag{8.24}$$

We introduce the SIR threshold at the base station, denoted as $SIR_{threshold}$, as a design parameter in SIR-based admission control. Overall, we can distinguish two types of SIR-based call admission control [9]. The first algorithm considers measurements of SIR only at the local base station. In such a case, the amount of available resources (i.e., residual capacity) locally in the cell k can be calculated using SIR_k as follows:

$$0 \leq A_k \leq \left(\frac{1}{SIR_{threshold}} - \frac{1}{SIR_k} \right) \tag{8.25}$$

The call is accepted if there is enough residual capacity A_k in the cell, otherwise it is rejected. The second algorithm considers SIR measurements of the adjacent cells besides the local measurements. In such a case, the residual capacity at the cell k is estimated according to the following relation:

$$0 \leq A_{k,j} \leq \frac{1}{\beta_{k,j}} \left(\frac{1}{SIR_{threshold}} - \frac{1}{SIR_j} \right),\ j = 1,2,\ldots,K \tag{8.26}$$

where $\beta_{k,j}$ is estimate of interference coupling between the adjacent cells ($\beta_{k,j} = 1$ for $j = k$). Then, the maximum residual capacity at cell k that satisfies the conditions of the home cell and adjacent cell is calculated by

$$A_k = \min\left[A_{k,1}, A_{k,2}, \ldots A_{k,K} \right] \tag{8.27}$$

8.7.2 Load-Based Admission Control

Another way of performing admission control is using directly the load factors in uplink and downlink. In such a case, a new call is admitted in uplink if

$$\eta_{UL} + \Delta\eta < \eta_{UL-threshold} \tag{8.28}$$

Similarly, a new call is admitted in the downlink if

$$\eta_{DL} + \Delta\eta < \eta_{DL-threshold} \tag{8.29}$$

The load factor of the new user $\Delta\eta$ can be calculated using (8.30). It is obtained as load factor L_j for a single user j from (7.82). Hence,

$$\Delta\eta = L_j = \frac{1}{1 + \dfrac{W}{\left(E_b / N_0\right)_j \cdot R_j \cdot \nu_j}} \tag{8.30}$$

where W is the chip rate, R_j is the bit rate of the new user j, ν_j is the assumed factor of the new connection, and $(E_b/N_0)_j$ is the uplink carrier-to-interference ratio for that user.

8.7.3 Power-Based Admission Control

In the downlink we have to consider the total transmitted power from the base station to mobile users. In the uplink we have to consider the total interference level at the base station from all users with ongoing connections. Hence, we consider uplink and downlink directions separately.

Let us first consider the uplink. We should have a predefined threshold value for maximum allowed interference. Methods for estimation of interference increase due to an admission of a new connection are different in different algorithms. A new user is admitted by the uplink admission control if the new total interference value is lower than the threshold value $I_{threshold}$:

$$I_{new\_total} < I_{threshold} \tag{8.31}$$

where $I_{new\_total} = I_{old\_total} + \Delta I$, and ΔI is estimated increase in the interference power caused by a new user. The threshold $I_{threshold}$ should be set by radio network planning.

In Chapter 7 we introduced load factor η, a measure of network congestion in the cell. It is also used in admission control for estimation of interference increase. Using (7.84) for the uplink load factor, we obtain

$$\eta_{UL} = \frac{I_{total} - I_0}{I_{total}} = 1 - \frac{S / I_{total}}{S / I_0} = 1 - \frac{SIR_{loaded}}{SIR_{empty}} \tag{8.32}$$

where S is the received power at the base station of a given user. From the above relation, we obtain the following:

$$I_{total} = \frac{I_0}{1-\eta_{UL}} \tag{8.33}$$

The last relation can be used for estimation of the interference increase ΔI caused by the admission of a new user.

There are two main methods for estimation of ΔI: the derivative method and the integral method. In the derivative method the total interference power increase is derivative of the old uplink interference power with respect to the uplink load factor. Using (8.33) we obtain

$$\frac{dI_{total}}{d\eta_{UL}} = \frac{I_0}{(1-\eta_{UL})^2} = \frac{I_{total}}{1-\eta_{UL}} \tag{8.34}$$

Then, the estimation of the total uplink interference increase is

$$\Delta I_{total} \approx \frac{I_{total}}{1-\eta_{UL}} \Delta\eta_{UL} \tag{8.35}$$

The integral method gives the following relation:

$$\begin{aligned} \Delta I_{total} &= \int_{\eta_{UL}}^{\eta_{UL}+\Delta\eta} \frac{I_0}{(1-\eta)^2} d\eta \\ &= \frac{I_0}{1-\eta_{UL}} \frac{1}{1-\eta_{UL}-\Delta\eta} \Delta\eta = \frac{I_{total}\Delta\eta}{1-\eta_{UL}-\Delta\eta} \end{aligned} \tag{8.36}$$

In the relations above, $\Delta\eta$ is the estimated increase in uplink load factor η_{UL} due to a new user, which is given by (8.30).

In the downlink we can use a similar approach for interference-based admission control. In this case, we should consider the total transmitted power from the base station. Hence, a new user is admitted in downlink if

$$P_{new\_total} < P_{threshold} \tag{8.37}$$

where P_{new_total} is the new total downlink transmission power including the power increase in the downlink due to a new user, while $P_{threshold}$ is the maximum allowed total transmission power in the downlink, which should be set by radio

network planning. Power transmission to every user depends on the distance of that user from the base station, and it is determined by the open loop power control scheme.

8.7.4 Power Control

Generally, the uplink open loop power control sets the initial power of the mobile terminal by using broadcasted (on control channels) cell/system parameters as input. In the downlink, open loop power control sets the initial powers of downlink channels using downlink measurement reports from mobile terminals. In UMTS, for long-term quality control of the radio channel, outer loop power control is used, which uses inputs from quality estimates of the transport channel [12]. The outer loop in UMTS includes Node B and RNC. It aims to control the target level SIR_{target} of the inner loop. For that purpose RNC measures the block error rate (BLER) and sets SIR_{target} in order to match the desired BLER [13]. The inner loop power control is used between the mobile terminal and Node B for uplink and for downlink. It sets the powers of the uplink and downlink dedicated physical channels, respectively. The term *open loop* refers to power control algorithms that use quality estimates of channels to set the transmit power, and it is mainly applied with common channels (e.g., random access channels). On the other hand, the term *closed loop* refers to power control that uses feedback from receiving station to directly set the power levels at the transmitting station for both the data channel and the corresponding control channel (e.g., uplink inner loop power control in the FDD mode is a closed loop process).

8.7.5 Performance Measures for CDMA Systems

We will consider the following performance measures for CDMA system [14]: call blocking probability, outage probability, and call removal (i.e., dropping) probability.

Blocking probability in *uplink* (UL) and *downlink* (DL) is defined by

$$P_b^{UL} = P\left[\left(I_{total} + \Delta I\right) > I_{threshold}\right] \tag{8.38}$$

$$P_b^{DL} = P\left[\left(P_{total} + \Delta P\right) > P_{threshold}\right] \tag{8.39}$$

where ΔI and ΔP are interference (at the base station) and transmitted power (from the base station) increase due to a new user in uplink and downlink, respectively.

Outage probability is defined as, current SIR_{total} at base station does not satisfy the specified $SIR_{threshold}$:

$$P_{outage} = P\left[SIR_{total} < SIR_{threshold}\right] \tag{8.40}$$

Removal probability is the probability that an ongoing call is dropped because the system does not meet the specified SIR (e.g., due to a congestion). In this manner, it is sometimes more appropriate to use as a performance measure the loss probability of communication quality [15]. We refer to this parameter as *quality loss*, and it is defined as the probability that the system does not meet the specified threshold(s)—that is, interference, SIR, or load threshold in uplink, and maximum transmitted power or load threshold in downlink.

8.7.6 Congestion Control

Congestion is defined as a situation where QoS requirements cannot be met. Possible reasons for congestion to occur are user mobility or channel variations, or traffic fluctuations due to burstiness of some connections. Indication of congestion is higher total interference at the base station than maximum level $I_{max} = I_{threshold}$ for the uplink, and a total transmitted power above some maximum power level $P_{max} = P_{threshold}$ for the downlink. There are several actions that can be taken by the congestion control (i.e., load control) [13]:

1. Lowering data rates of the nonreal-time ongoing connections beginning with services with lowest priority;
2. Handover to another carrier (e.g., in WCDMA) or to another network (e.g., to a GSM network, if possible);
3. Dropping connections (i.e., bearers).

The congestion is considered resolved when $I_{total} < I_{min}$ for the uplink case, and when $P_{total} < P_{min}$ for the downlink case, where $I_{min} < I_{max}$ and $P_{min} < P_{max}$ to avoid the *ping-pong effect*. After the congestion has successfully been resolved, a change in load (i.e., a decrease in load due to dropped calls or mobility of users) might allow the increasing of data rates again.

8.7.7 Hybrid Admission Control Algorithm for Multiclass CDMA Networks

In wireless CDMA networks with multiple traffic types, we can define different threshold values for different traffic classes. In the following section we extend the admission control policies in CDMA networks from a single traffic type to multiple traffic types for each CDMA call admission control policy.

In the case of SIR-based admission control policy we can define different SIR values for calls belonging to different classes. Hence, we have different $SIR_{threshold,j}$, $j = 1, 2, \ldots, n_c$, where n_c is number of different traffic classes.

Load-based CAC algorithms are focused to keep the individual cell load (and hence the network load) below some specified value. Most proposed CAC algorithms for WCDMA networks are load based [13]. With the aim to allow multiple traffic classes (e.g., by using prioritization of services) we can define different maximum cell load levels for different services j (i.e., $\eta_{UL\text{-}threshold,j}$ and $\eta_{DLthreshold,j}$ for uplink and downlink, respectively).

Using a similar approach as given above, in the power-based CAC policy we can define different maximum interference levels $I_{threshold,j}$ in the uplink, and different maximum transmitted power $P_{threshold,j}$ in the downlink, for different classes j.

Additionally, it could be distinguished between new calls and soft handovers, resulting in two maximal thresholds for each class, direction (i.e., uplink and downlink), and cell. For example, in load-based CAC we will have two maximum load levels ($\eta_{threshold\_new\ call,j}$, $\eta_{threshold\_\ handover,j}$) for each direction uplink and downlink and each class j in the cell. Because dropping of an ongoing call is more offensive to users than blocking a new one, we should use values $\eta_{threshold\_\ handover,j} \geq \eta_{threshold\_\ new\ call,j}$.

In order to analyze admission control based on teletraffic theory, we can transform the interference level, power level, or load into an equivalent number of logical channels (refer to the example in Section 7.6). Then, we can easy extend the application of the HAC concept into a CDMA environment.

8.8 Discussion

In this chapter we analyzed admission control with QoS support in wireless IP networks with multiple traffic classes. In such a heterogeneous environment, the network needs suitable admission control [16]. Different traffic types have different QoS constraints. For instance, real-time services have higher QoS demands and they need particular guarantees on the allocated bandwidth during the connection. On the other hand, nonreal-time services and applications are more flexible to QoS support.

To adapt various QoS requirements in the network, we proposed a new type of admission control called hybrid admission control, in which we integrated call-level and packet-level QoS parameters. New call blocking probability and call dropping probability are considered as QoS parameters of A1 and A2 subclasses, while average packet delay is a parameter of A3. The algorithm bounds call dropping probability of A1 and A2 subclasses and the average packet delay of A3, while at the same time minimizing new call blocking probability of A1 and A2. B class, however, is not considered in admission this control algorithm. B packets are serviced only when all A packets from queues are transmitted over the wireless link.

The analytical and simulation analyses showed two main compromises that have to be made in the HAC algorithm: (1) that between new call blocking probability and call dropping probability of A1 and A2, and (2) that between new call blocking probability and average delay of A3. Constraints on QoS parameters are given at the phase of network design, but they can change later because of network policy or traffic behavior. If it is not possible to determine the optimal thresholds, then the network has too few resources for the given QoS demands or the initial constraints are too strict for one or more parameters.

The hybrid admission control can be extended to CDMA networks, which are characterized by soft capacity. In a CDMA network, however, we can use different thresholds for new calls and handovers for different traffic classes. The threshold can refer to interference, transmitted power, or cell load, which is dependent upon the admission control scheme applied in the network.

References

[1] Hong, D., and S. S. Rappaport, "Traffic Model and Performance Analyzes for Cellular Mobile Radio Telephone Systems with Prioritized and Nonproritized Handoff Procedures," *IEEE Trans. on Vehicular Technology*, Vol. VT-35, No. 3, August 1986, pp. 77–92.

[2] Ramjee, R., R. Nagarajan, and D. Towsley, "On Optimal Call Admission Control in Cellular Networks," *Wireless Networks Journal*, Vol. 3, No. 1, 1997, pp. 29–41.

[3] Naghshineh, M., and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 4, May 1996, pp. 711–717.

[4] Misic, J., S. T. Chanson, and F. S. Lai, "Quality of Service Management for Wireless Networks with Heterogeneous Traffic," *Globecom'98*, Sydney, Australia, November 1998, pp. 1406–1412.

[5] Oliver, M., "Admission Control Strategy Based on Variable Reservation for a Wireless Multi-Media Networks," *First International Symposium on Wireless Personal Multimedia Communications WPMC'98*, Yokosuka, Japan, November 1998, pp. 338–343.

[6] Oliver, M., and J. Paradelis, "Variable Channel Reservation Mechanism for Wireless Networks with Mixed Types of Mobility Platforms," *48th Annual Vehicular Technology Conference VTC'98*, Ottawa (Ontario), Canada, May 1998, pp. 1259–1263.

[7] Firfov O., T. Janevski, and B. Spasenovski, "Modeling the Internet – State of the Art," *ETAI 2000*, Ohrid, Macedonia, September 21–23, 2000, pp. TI27–TI32.

[8] Ishikawa, Y., and N. Umeda, "Capacity Design and Performance of Call Admission Control in Cellular CDMA Systems," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, October 1997.

[9] Liu, Z., and M. Zarki, "SIR Based Call Admission Control for DS-CDMA Cellular System," *IEEE Journal on Selected Areas in Communications*, Vol. 12, 1994, pp. 638–644.

[10] Holma, H., and J. Laakso, "Uplink Admission Control and Soft Capacity with MUD in CDMA," *IEEE Vehicular Technology Conference*, Vol. 1, September 1998, pp. 431–435.

[11] Viterbi, A. M., and A. J. Viterbi, "Erlang Capacity of a Power Controlled CDMA System," *IEEE Journal on Selected Areas in Communications*, Vol. 11, August 1993, pp. 892–900.

[12] 3GPP TS 25.401, *UTRAN Overall Description (Release 5)*, V.5.1.0, September 2001.

[13] Winter, T., (ed.), *Identification of Relevant Parameters for Traffic Modeling and Interference Estimation*, Information Report No. IST-2000-28088-MOMENTUM-D21-PUB, Information Society Technologies (IST), November 2001.

[14] Kim, K., and Y. Han, "A Call Admission Control with Thresholds for Multi-Rate Traffic in CDMA Systems," *VTC 2000-Spring*, Tokyo, May 2000.

[15] Ishikawa, Y., and N. Umeda, "Capacity Design and Performance of Call Admission Control in Cellular CDMA Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, October 1997, pp. 1627–1635.

[16] Janevski, T., and B. Spasenovski, "QoS Analyses of Multimedia Traffic in Heterogeneous Wireless IP Networks," *ICPWC 2000*, Hyderabad, India, December 17–20, 2000.

9

Performance Analysis of Cellular IP Networks

9.1 Introduction

Cellular networks and the Internet are converging. This convergence challenges the QoS provisioning in such cellular IP networks. The future cellular Internet will include many portable devices connected to the global network. In order to achieve higher bandwidth for the users, the cell size will have to be limited. That leads to the creation of microcellular, or even picocellular environments, where the users move frequently among cells [1–4]. In this chapter we address problems that arise from the integration of mobile networks and the Internet, which are mainly due to user mobility. We analyze the impact of handovers on different traffic types, such as CBR, VBR, as well as best-effort traffic.

Also, we analyze the differentiation of the traffic types in a wireless environment. In a multiclass network, services can be offered to users based on SLAs, as defined in Chapter 3. Mechanisms for service differentiation may be based on different parameters, such as delay, capacity or bandwidth, price, and prioritization [5–8]. Resource allocation in IP networks is analyzed in [9], while resource management in wireless networks is given in [10]. In this chapter we show analysis of services in a cellular IP environment with different user mobility, different load in the access network, and different BER in the wireless link [11–14].

The performance analysis is made using the mobile/cellular IP network architecture given in Chapter 6. Through the analysis we show the behavior of different traffic types, which are defined according to the classification of the IP

traffic proposed in Chapter 5, under the given mobile network's characteristics: user mobility and bit error ratio in the wireless link.

9.2 Service Differentiation in Cellular Packet Networks

Future cellular networks should incorporate different traffic types, such as CBR, VBR, and best effort. Constant bit rate traffic is defined by its peak rate (which is also the mean rate) and it requires a constant data rate during the entire connection. However, even CBR traffic experiences jitter (packet delay variance) due to statistical multiplexing of flows at the network nodes. The description of the variable rate traffic is more complex. A VBR flow experiences rate variations during the communication. Best-effort flows utilize the bandwidth that is left after servicing the traffic with QoS guarantees (refer to Chapter 5). In our analysis we assume that all calls have passed the admission control.

Class differentiation in the wired Internet provides relative guarantees (i.e., performance of the higher-level class should be better than that of lower-level classes). By introducing wireless access to the Internet, we face several additional problems. First, we need to manage handovers, which are related to single calls, and not to aggregate traffic in the network. Second, the wireless link is characterized with a higher bit error ratio, which is dependent upon the users' location. The mobile hosts have a random position in the cell, and each one will experience a different error ratio. In our analysis we assume one flow per mobile host. Hence, if the number of users in the cell is many times greater than the maximum number of simultaneous connections, then we may consider that call arrivals are independent events. Furthermore, it is not appropriate to apply the PHB concept in the same manner as in wired networks. In most cellular networks only the last hop (or the first) is wireless (i.e., the link between the base station and the mobile terminal). Thus, service differentiation (i.e., per-hop behavior) should be used in the wired links of the core network, while in the wireless access network we need to consider each flow separately. Such network setup requires a bandwidth broker (i.e., admission controller) that will control the allocation of resources by using the base stations in its domain.

Various differentiation models are proposed, analyzed, or implemented, such as strict differentiation and capacity differentiation. For example, in a strict prioritization scheme, packets from the highest backlogged class are serviced first. Capacity differentiation allocates resources between classes so that the higher class has more bandwidth than the lower class (in this case WFQ can be used to manage the link bandwidth [15]). In a delay differentiation scheme, the higher-level class should receive less average and/or peak delay than the lower-level class.

In the capacity differentiation model higher classes have more bandwidth (i.e., a higher rate) and packet buffers than lower classes, relative to their long-term expected load. Analytically, relative capacity differentiation is defined by

$$\frac{r_i}{r_j} = \frac{q_i}{q_j} \tag{9.1}$$

Here, differentiation is defined through relative capacity differentiation coefficients q_i, where $0 < q_1 < q_2 < \ldots < q_N$, and N is the number of classes. Rate r_i dedicated to specific class i is always higher than rates given to lower classes $j < i$, $i \neq 1$. On the other hand, if absolute differentiation is applied (proportional bandwidth share of each class), then we may write

$$\frac{r_i}{B} = \frac{w_i}{\sum_{j=1}^{N} w_j} \tag{9.2}$$

where B is the total bandwidth of the wireless link, while w_i, $i = 1, \ldots, N$, are weight coefficients. But, in the case of the delay, only the relative differentiation is applicable. Let us denote the average queuing delay of class i with $\bar{d}_i$. Then, according to the delay differentiation model, for all pairs of classes (i, j) should be satisfied:

$$\frac{\bar{d}_i}{\bar{d}_j} = \frac{\Delta_i}{\Delta_j} \tag{9.3}$$

where Δ_i, $i = 1, 2, \ldots, N$, are referred to as delay differentiation parameters. These parameters are ordered as $\Delta_1 > \Delta_2 \ldots > \Delta_N > 0$, because higher classes should receive better service, which means lower average delay.

In contrast to the wired networks where packet losses occur due to congestion of the queues at the nodes, losses in a cellular environment may also occur due to the error-prone wireless link. Therefore, in this chapter we provide analysis of the influence of the wireless bit errors in two general cases of service differentiation: complete partitioning (e.g., circuit-switched cellular network) and complete sharing of the resources (e.g., wireless LAN). There are different proposals for service differentiation in existing wireless LANs, such as IEEE 802.11 [16], and in 3G mobile networks [17] where we have different resource scarcity for the uplink and downlink.

Packet loss, however, should be considered as a possible differentiation parameter in a wireless IP network. We have stated previously that base stations

should manage single flows rather than the aggregate traffic. Our aim is to distinguish error-free flows from erroneous traffic. It can be accomplished by introducing a weight coefficient for each flow at the base station. A possible solution to this problem is given in Chapter 11.

On the other hand, handovers may also introduce packet losses. In the following section we define the handover problem in cellular IP networks, and then we will continue with performance analysis.

9.3 Handover in Cellular Networks

Handover is a main characteristic of mobile networks. Its influence on QoS is proportional with its intensity. Small cells and higher user mobility increase the handover intensity, and hence more significantly influence the QoS. Therefore, one of the main goals in cellular IP networks is the design of efficient handover mechanisms.

In 2G mobile systems, the BSC initiates and controls the handovers. Handover initiation is based on periodical measurements of the received signal strength and link quality, which are recorded by the mobile terminal and then transmitted to the BSC via the base station(s). Sometimes, the BSC makes handover decisions based upon the given control algorithm for the radio resources, such as hierarchical cell structure (e.g., umbrella, macro, micro, picocells) and overlaid/underlaid cell. Thus, in 2G, the BSC determines the target cell and initiates the handover. Third generation mobile networks (and beyond) should support higher data rates per single user, as well as provide multimedia support. It requires different bandwidth allocation to calls from different traffic types, and additional maintenance of the allocated bandwidth at the handovers.

9.3.1 Handover in Cellular Packet Networks

In packet-based networks we need to reroute the connection from one cell to another (i.e., base station) at the handover. The main goal is to provide transparent rerouting of the ongoing call. Globally, considering the macro-mobility, Mobile IP solves this problem. As we discussed in Chapter 3, Mobile IP uses two agents, an HA and an FA. At each handover, the FA sends a control message to the HA to inform about the new location of the mobile terminal. However, the HA can be far away from the foreign network, and thus the FA, so Mobile IP is not efficient for micromobility management (i.e., local handovers). Therefore, we need an efficient and transparent handover algorithm to account for the micromobility.

When designing local handover schemes, we should consider several goals that are crucial for QoS support:

- Lowest possible degradation of the ongoing traffic at the handover (lower packet losses or duplicate packets);
- Efficient rerouting of the packet flow, which is especially important for real-time communications;
- Handover latency (i.e., the time between the initialization and end of the handover) should be kept small;
- Maximum reduction of the signaling traffic in the wireless interface due to scarce radio resources;
- Support for multiclass cellular packet networks.

Transport delay in the wireless network is also an important parameter with regard to handovers. Usually, transport bearer delay is in the range of several tens of seconds [18]. For example, the air interface transmission time interval in UMTS can be 10, 20, 40, or 80 ms. Time delay due to processing of data at network nodes and mobile terminals is in the range of several milliseconds. But, if we also consider queuing delays at the network nodes, then the total transport delay might be several hundreds of milliseconds [19].

9.3.2 Handover Mechanisms

There do exist several design approaches for handovers that solve some of the problems listed in the previous section. According to their design, handover mechanisms can be classified into the following groups:

- Hard handover;
- Soft handover;
- Predictive multicast handover;
- Chained handover.

All of the above handover mechanisms are in the horizontal layer. We may, however, find proposals for vertical handover schemes [20]. In the vertical handover approach, network resources are organized in vertical layers, and the mobile terminal should have a different radio adapter for each layer. Such an approach is based on detection of beacon signals by the mobile terminal, and thus, it results in longer handover delay (i.e., latency).

9.3.2.1 Hard Handover

The hard handover is the simplest mechanism [4], but it is also the fastest handover mechanism. The crossover node simply reroutes the traffic through a new

path towards the new base station, while the connection between the source and crossover node remains unchanged. The discovery of the crossover node is described in Chapter 10. So far, we assume that the crossover node is already determined for a given connection. The drawback of this algorithm is the packet loss: packets that reach the crossover node during the handover latency are lost, because they will be routed towards the old base station and the mobile terminal will be attached to the new one.

9.3.2.2 Soft Handover (Simultaneous Multiple Bindings)

By a definition, during soft handover there are two simultaneous active wireless links. Data and control information are carried over both at the same time. In a soft handover the uplink data from each access line is multiplexed, and the downlink data is multicast to each access line. Soft handover is an effective way to increase the capacity reliability and coverage range of CDMA networks (e.g., IS-95, WCDMA, and cdma2000). In 2G and 3G networks soft handover is extended by a macro-diversity [21]. In a macro-diversity state the mobile terminal is allowed to communicate with multiple base stations simultaneously (i.e., send and receive data during the handover). Thus, in such a soft handover scheme the mobile terminal communicates through a set of base stations called the active-set, while neighboring base stations form the so-called neighboring-set. When SIR of a neighboring base station becomes larger than a predefined threshold, which is referred to as the ADD-threshold, the neighboring base station is added to the active-set. Similarly, when the SIR becomes lower than a predefined DROP-threshold, then the base station is taken out of the active-set. Usually, there is limit on the number of base stations in the active-set. The base station with the highest signal strength is referred to as a primary base station [21], and all other cells from the active-set are classed as nonprimary. If the signal strength of the base station exceeds that of the primary base station (with applied threshold to avoid the ping-pong effect), handover is initiated. The main objective is to transmit on the downlink from the primary cell, thus reducing the interference caused by multiple transmissions in a soft handover mode. A second objective is to achieve fast site selection without network intervention, thus maintaining the advantage of the soft handover.

Thus, in a soft handover scheme a route to the new base station is established before the old route is terminated. There are also modifications of this mechanism, such as semi-soft handover [22], which we described in Chapter 3.

9.3.2.3 Predicted Handover

Predicted handover has the goal of minimizing the impact of disruption in the loss and delay of sensitive real-time flows. The idea behind this approach is to predict the possible targets for a handover of a given call. Such targets are neighboring cells of the mobile's current cell. This multicast procedure differs from

the soft handover, because neighboring base stations buffer the last packets addressed to the mobile stations (e.g., t seconds worth of data), while only the current base station transmits packets to the mobile. Thus, if the mobile terminal is handed to a neighboring base station, the data received in the last t seconds is available at the base station, and disruption of the flow is significantly reduced. This policy requires maintenance of multiple connection paths to the mobile terminal and its neighbors. The multicast-based intradomain handover [23] (i.e., Deadalus project), described in Chapter 3, belongs to this type of handover.

9.3.2.4 Chained Handover

Chained handover is a scheme found in 2G cellular networks. These mobility management schemes are designed to manage handovers in connection-oriented networks. Second generation cellular networks link connections through switches; that is, they eliminate the link towards the old base station. Similar procedures may be applied in 3G networks or in wireless LAN, if most of the user movements will be within the local domain. In such a scenario, extending the connection from the old base station to the new base station should extend the path at each handover. In this manner, the consecutive attachment points of the mobile terminal, as it moves, are *chained*. Thus, it is called the chained handover [24].

This handover mechanism reduces the losses to zero. The disadvantage of this type of handover is the occurrence of loops when the mobile visits one cell more than once during the connection duration. Also, it adds extra delay due to the new links, which may be considered insignificant if large bandwidth is available in the wired core network. However, rerouting is necessary when the mobile leaves the current domain or when the chain becomes too long. If we have centralized switches (e.g., in 2G and 3G mobile systems), we eliminate the loops and the hops between the switch and base stations.

9.3.3 Analysis of Packet Losses at Handover

Packet loss is a possible consequence of a handover in a wireless IP network. We assume that hard handover is applied in the network, which is the simplest and the fastest one. So, by analyzing the hard handover we can obtain the maximum packet losses. In all other schemes, packet loss should be lower, but other parameters such as packet delay would be worse.

We analyze losses at different traffic and mobility conditions that would eventually appear in a cellular packet-based network. The focus is on packet loss caused by the handovers, and therefore we assume an ideal mobile interface (i.e., without bit errors in the radio part). The hard handover avoids explicit signaling messages and buffering or forwarding the packets. The trade-off for its

simplicity is the packet loss. All packets that are routed to the mobile host during the handover latency are considered lost.

We want to obtain the boundaries for handover packet losses. So, we assume that the whole bandwidth of the wireless link in the cell is used by one connection. Also, we assume that the target cell admits the handover call. Then, packet loss due to the handover can be calculated using the following approximation:

$$l_h \approx b\left(t_p + t_q\right) \tag{9.4}$$

where l_h is total number of lost bits at one handover event, and b is the data rate of the connection ($b \leq B$, where B is the total wireless link bandwidth). Furthermore, t_p is the total round-trip propagation time between the crossover node and the mobile terminal (including the wired hops and the wireless hop in the path), and t_q is the queuing time at the intermediate nodes between the crossover node and the old base station, including the end nodes. Now, we will assume that the crossover node knows in advance which packet will be lost first. Before this packet is sent to the mobile host, the mobile host performs a handover and immediately generates a route update packet, which is sent to the crossover node. Let us assume that we observe a CBR flow. One may expect that there will be queuing at the intermediate nodes at higher network load. To obtain the losses we need to know the queuing time of the marked packet in all nodes towards the old base station, including the two end nodes, and the packet propagation time from the crossover node to the old base station. Then, packet losses due to the handover for a CBR flow at data rate b can be calculated by

$$l_h \approx b\left(2\sum_{i=1}^{N+1} t_{p_i} + \sum_{i=1}^{N+2} t_{q_i}\right) \tag{9.5}$$

where N is the number of intermediate nodes between the crossover node and the old base station, and t_{pi} and t_{qi} are link propagation time and queuing time from/at node i.

In the handover analyses, we use a mobility model that is defined in Chapter 6. User movement in a cell is random and can be considered independent from other users' movement. Thus, we may consider handover events independently. To calculate packet losses due to handovers we use the average handover intensity, which is not dependent on the connection duration. Let λ_h be the handover intensity and l_h be the average bit losses per handover in the cellular network. Then, the handover bit loss rate L_h (lost bits per time unit) can be calculated as

$$L_h = \lambda_h l_h \tag{9.6}$$

The maximum loss rate is obtained at maximum throughput in the cell, which is the wireless link bandwidth. Such throughput may be realized only by a single CBR flow in the cell, which occupies the whole bandwidth B. For the example, let us consider a single CBR flow in the cell at rate $b = B = 2$ Mbps (i.e., equal to the wireless link bandwidth), and let us assume that the average total sum of the round-trip propagation time and first packet queuing time is 25 ms. Furthermore, if we set handover intensity to $\lambda_h = 6$ handovers/minute (handover intensity should be higher in microcellular and picocellular environments), then we get $l_h = 2$ Mbps * 25 ms = 50 Kb, and an average handover packet loss rate $L_h = 5$ Kbps, that is, 0.25% of the flow data rate (i.e., bit loss ratio due to handovers). Moreover, the useful throughput is usually significantly less than the link data rate due to the statistical multiplexing, so the percentage of packet losses due to the handovers would be higher. We have to point out that these losses occur in the core network (i.e., on the wired links). The loss ratio in the range of 10^{-3} to 10^{-2} is not negligible, and it is highly undesirable. Also, to this loss ratio we should add bit error ratio in the wireless link; hence, the total bit loss/error ratio will be higher.

9.4 Network Model

We use the wireless IP architecture defined in Chapter 6. The network consists of wired nodes and wireless access points (i.e., base stations). The architecture of the mobile IP network is shown in Figure 6.2. We assume that the bandwidth in the cell is shared between all active users. Buffering at the network nodes is performed using an FCFS mechanism. Packets from different classes enter different queues at each node. In the wired nodes, the schedulers service the aggregate traffic for each class. Because radio resources are scarce, the base station differentiates every flow in the cell. Traffic of a higher class (i.e., class A) is serviced before class-B packets. So, base stations keep the information of all ongoing class-A connections in the cell. Thus, we can differentiate each class-A flow at the base station. So, at the base stations we differentiate aggregate class-A traffic from class-B traffic, and then apply per-flow differentiation within the higher class.

The mobility model from Chapter 6 is used to model user movement. The aggregate traffic of each class consists of flows from/to different users, the cell location of which is randomly chosen. However, the exception may be applied in the case of users in cars and trains, which certainly have predefined trajectories.

Each flow travels from the source to the destination through a sequence of network nodes (e.g., routers). Only the last hop (in downlink) or the first hop

(in uplink) is a wireless link. An exception to such a scenario is ad hoc wireless networking, which is not considered here. We assume that each base station serves one cell. Every base station performs packet scheduling and admission control for the wireless medium.

In this analysis we consider the micromobility. Handover can be mobile initiated, or network initiated and mobile assisted. We assume mobile-initiated handovers in the network, which could be preferred in future wireless IP networks.

9.5 Simulation Analysis in Wireless IP Networks

In this section we show the results of several experiments on performance analysis of wireless IP networks by using the simulation techniques. We focus our analysis to two main characteristics of wireless networks: handover and bit errors in the wireless links. In all experiments we use the following settings: average processing and propagation time in the wireless link is set to 13 ms, delay in the wired link is 2 ms, what gives at least a 30-ms round-trip delay (when there is no congestion at the network nodes). We performed the analysis at different user velocity and different network load. The average length of IP packets is 1,000 bytes. Most of the experiments are performed for small cells (e.g., cell radius r = 0.1 km) and average user velocity of 50 km/hr. Traditionally, the scheduling discipline that is applied in most routers on the network is FCFS. But, to be able to perform capacity differentiation of the traffic, we use a WFQ-like scheduling mechanism. Therefore, we used both FCFS and WFQ in the analysis. Also, we used different wireless link data rates.

Performance analysis is done via four different types of simulation experiments of wireless IP networks. In the first three experiments we analyze traffic degradation due to user mobility (i.e., handovers) for CBR, VBR, and best-effort traffic. In the fourth experiment we analyze the influence of wireless bit errors on the performance.

9.5.1 Handover Loss Analysis for CBR Flows

In the first experiment we consider performances of the CBR flows at the handovers in wireless IP networks. Let us first consider a CBR flow that occupies the whole bandwidth in the cell. Simulations are made at various user mobility and different cell sizes. Packet loss ratio of the CBR flow is shown in Figure 9.1, while cumulative losses (in kilobytes) are shown in Figure 9.2. Due to the constant rate of the flow, the loss ratio due to handovers increases linearly with the user velocity, while it decreases as cell size increases. This is due to higher handover intensity at higher mobility of users and smaller cells.

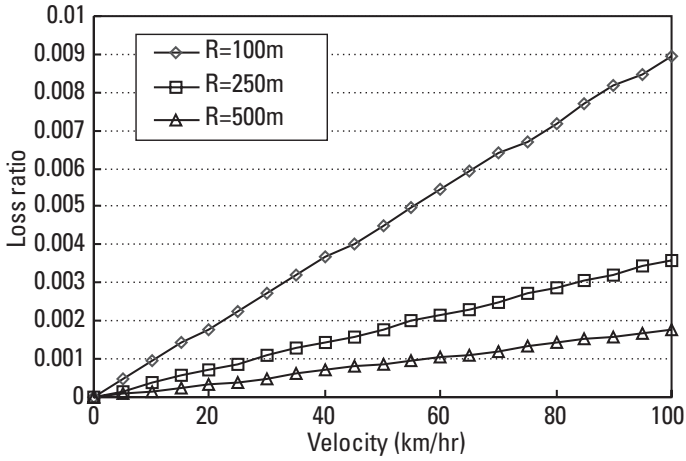


Figure 9.1 Loss ratio of a CBR flow that occupies the whole wireless link bandwidth at different mobility parameters.

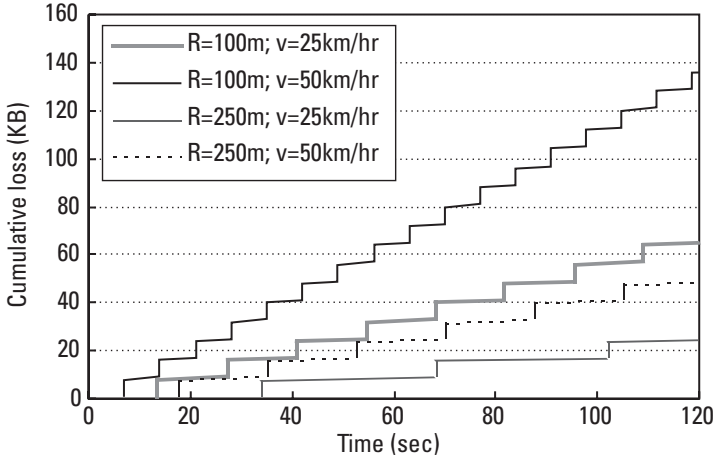


Figure 9.2 Aggregate packet losses of a 100% CBR flow.

The relation between the loss ratio and user mobility, shown in Figure 9.1, is linear because there is no background traffic in this simulation (the whole wireless link is allocated to a single CBR flow). Aggregate packet loss at 100% (of the link bandwidth) CBR flow is shown in Figure 9.2. Therefore, we have almost equal losses at every handover, thus the curve has the form of a set of steps.

Next, we consider a CBR flow that occupies 20% (i.e., 400 Kbps) of the total link bandwidth in the cell, which is 2 Mbps (Figure 9.3). In this case the CBR flow is multiplexed with background traffic with high correlation. The

total background traffic load is 70%; thus, the total load is 90% of the link bandwidth. The throughput of the CBR flow significantly changes by increasing the number of the background flows in the same cell. Figure 9.3(a), which is

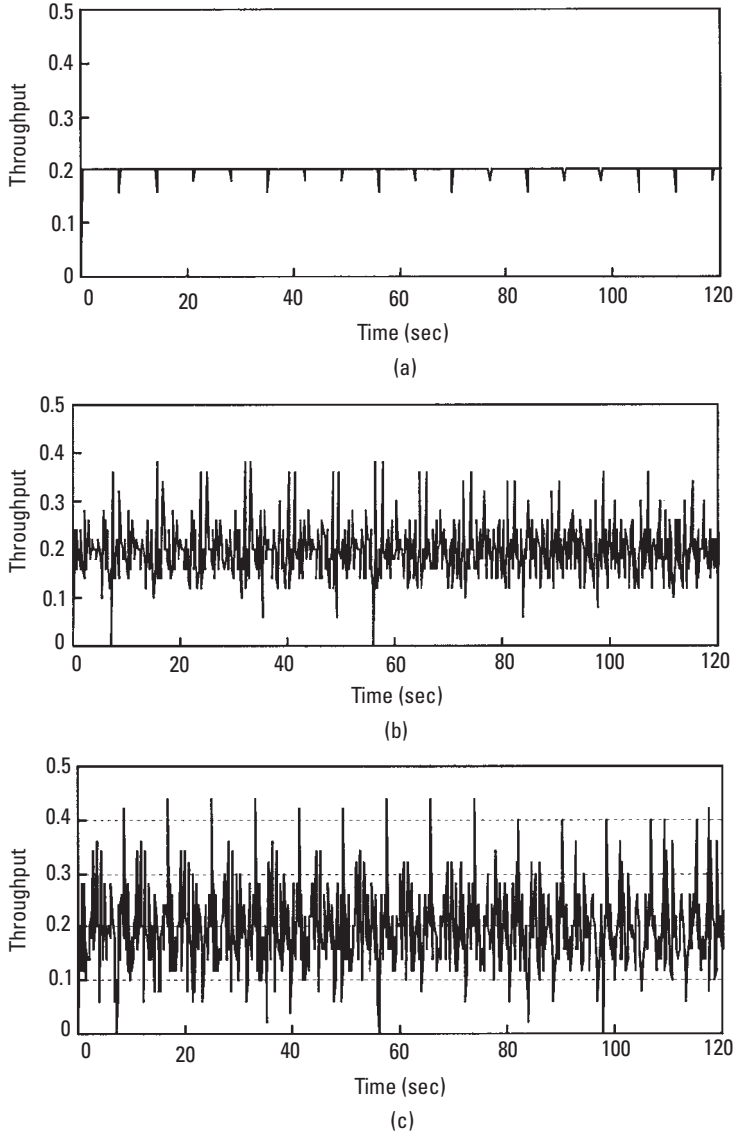


Figure 9.3 Throughput of a CBR flow at different traffic conditions: (a) single 20% CBR flow without background traffic; (b) four background flows at 90% total link load; and (c) eight background flows at 90% total link load.

created by analysis of the single CBR flow, shows that the throughput has low peaks only at handovers. Figures 9.3(b, c) show that degradation of the CBR throughput increases with the number of background flows on the link. Due to high correlation of the flows, packets from different streams enter the buffers at the network nodes at the same time, thus increasing the congestion and burstiness of the traffic.

Figure 9.4 gives the cumulative loss at handovers in the case of 20% CBR flow. The analysis shows that losses due to handovers become bursty at high correlation in the background traffic. Thus, at different handovers we observe various losses. Such results can be explained by the correlation between the background flows, even in the case of WFQ scheduling at the networks nodes. Therefore, to provide QoS guarantees on loss and delay of the CBR flow, we should isolate these flows by appropriate scheduling at the base stations. We also performed the analysis with random correlation between the background flows. In this case the curve of cumulative loss due to handovers is smoother. Such behavior can be explained by the constant interpacket time of the CBR flow and low correlation of the background traffic. From Figure 9.5 we can derive the same conclusion at a different user mobility.

Loss due to handovers of a 400-Kbps CBR stream at different network loads is shown in Figure 9.6. From these results, one may conclude that a higher network load results in higher loss at handovers. This is due to packet buffering at nodes along the communication path between the crossover node and the old base station, thus resulting in shorter interpacket time distance of the CBR flow.

Finally, we finish the analysis of the CBR traffic type by exploring scenarios with a different number of hops between the crossover node and the old base

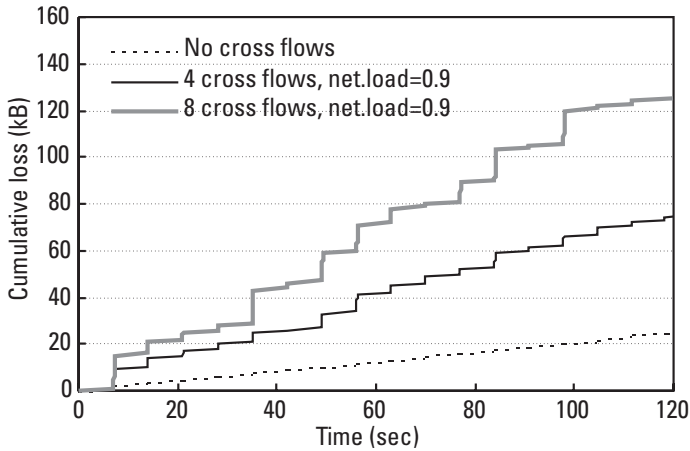


Figure 9.4 Cumulative losses of 20% CBR flow at different traffic conditions.

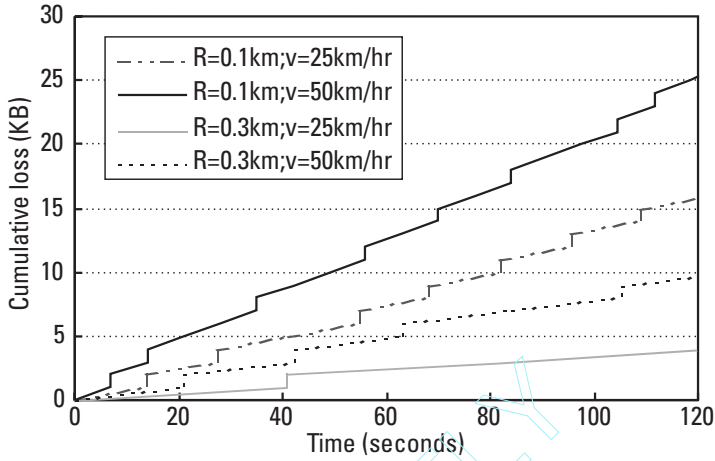


Figure 9.5 Cumulative losses of 20% CBR flow at different user mobility.

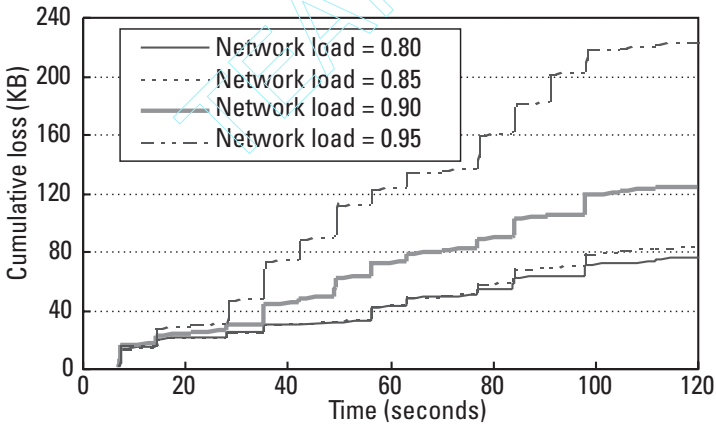


Figure 9.6 Losses due to handovers of 400-Kbps CBR flow at different network load.

station, as given in Figure 9.7. The results show that increasing the number of hops increases the packet loss due to handovers, because each additional hop adds delay to the packets; thus, more packets are in the queues targeted to the old base station when the handover is initiated.

9.5.2 Handover Loss Analysis for VBR Flows

To conduct our analysis of loss at handovers for VBR flows, we use traces from video sequences, statistical characteristics for which are given in Chapter 5. In contrast to CBR traffic, VBR flows have higher burstiness.

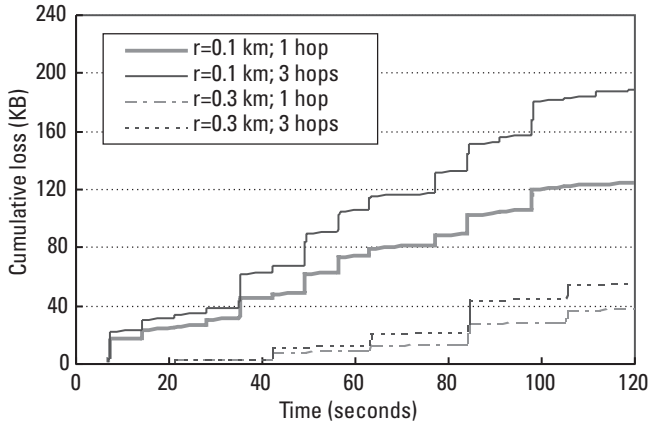


Figure 9.7 Cumulative packet losses for a different number of hops between the crossover node and the old base station.

The analysis results of VBR flows at handovers are shown in Figures 9.8 and 9.9, where we used the video trace *vbrvideo1*. Two different scheduling schemes are used in this experiment: FCFS and WFQ are used for obtaining results shown in Figures 9.8 and 9.9, respectively. In this simulation the wireless link bandwidth is set to 2 Mbps, while the average data rate of the video stream is 1.4 Mbps—that is, 70% of the link bandwidth. We assume that both wired and wireless links have equal bandwidth. The total network load is set to 90%, while 20% is background traffic. The results show that higher mobility does not necessarily mean higher losses due to handovers, although higher mobility of users and small cells increase the handover intensity. This phenomenon can be explained by

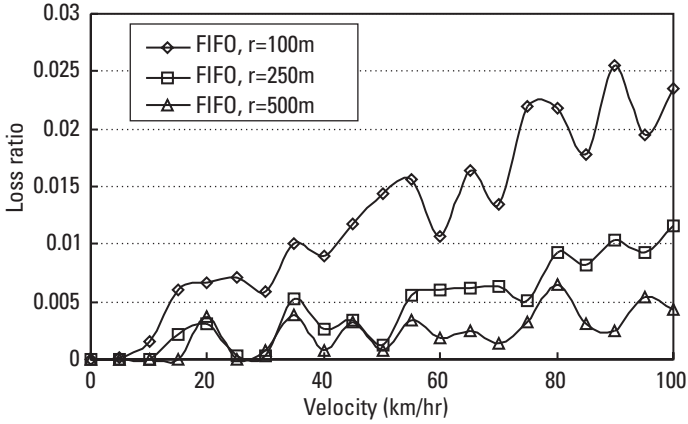


Figure 9.8 Loss ratio of a VBR flow versus mobility when FCFS is applied at the network nodes.

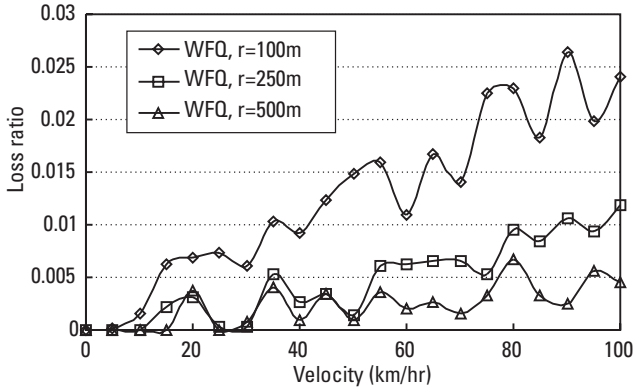


Figure 9.9 Loss ratio of a VBR flow versus mobility when WFQ is applied at the network nodes.

the stochastic nature of handover events and high burstiness of the VBR flow. Usually, burst periods in the VBR flow cause higher delay due to higher congestion at the buffers (e.g., at the base stations). Thus, if the handover latency time period overlaps with the bursty period of the video stream, then we get higher losses at the handover, and vice-versa. So, in some cases we may get lower loss due to handover at higher mobility than at lower mobility, as it can be seen from Figures 9.8 and 9.9. In the case of VBR flows, losses due to handovers also increase with the mobility, but this increase is driven by the flow's burstiness.

We use the same VBR stream for simulations on 20-Mbps wireless link (e.g., wireless LAN or broadband RAN). The simulation results are shown in Figures 9.10 and 9.11. In this case the VBR stream occupies only 7% of the link

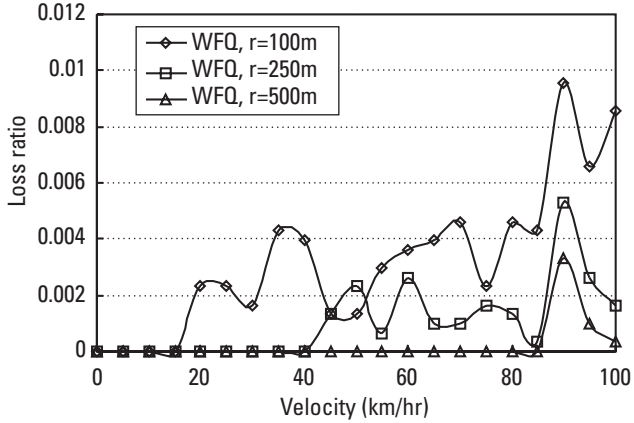


Figure 9.10 Loss ratio due to handovers of a VBR flow at 20-Mbps link bandwidth, and FCFS scheduling at the network nodes.

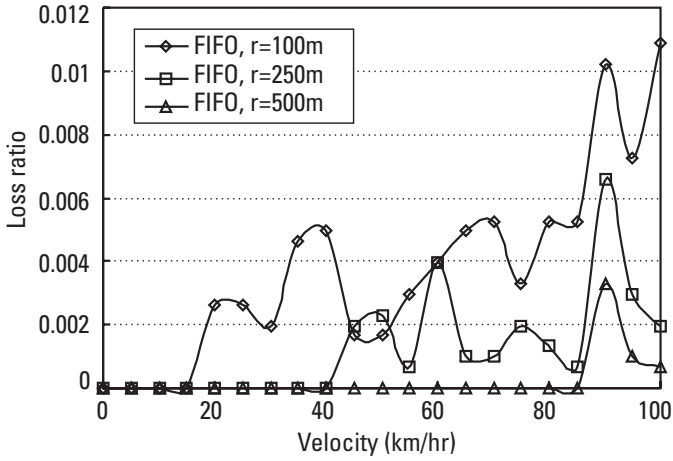


Figure 9.11 Loss ratio due to handovers of a VBR flow at 20-Mbps link bandwidth, and WFQ scheduling at the network nodes.

bandwidth. It is obvious, however, that the results show the same behavior as in the case of 70% VBR flow. But, now the background traffic fills the buffers at nodes along the path, thus reducing the burstiness of the VBR flow. That is why the packet loss ratio due to handovers is lower when the bandwidth share of the flow is smaller.

From the results given in Figures 9.12 and 9.13 one may conclude there is not much difference in the case when WFQ is applied instead of FCFS. Analysis

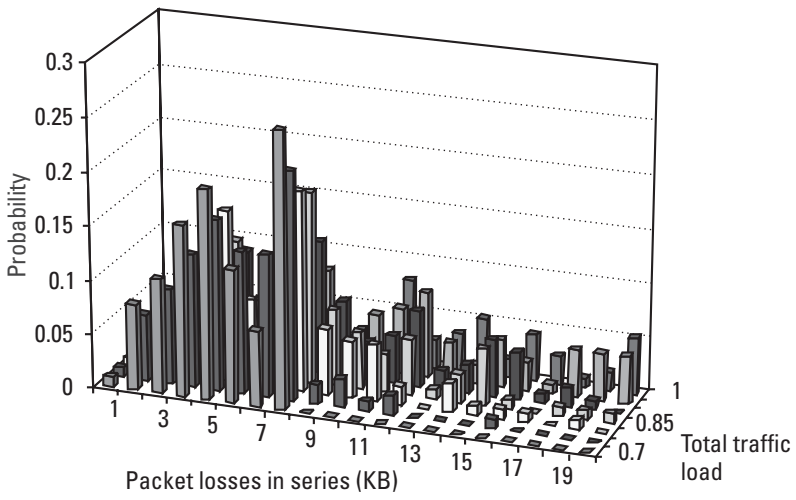


Figure 9.12 Probability distribution function of packet losses in series at handovers: 2-Mbps wireless link bandwidth, FCFS scheduling.

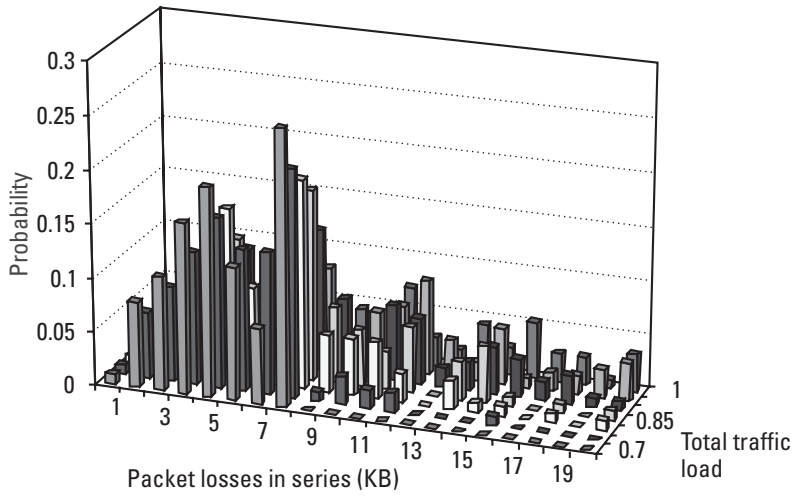


Figure 9.13 Probability distribution function of packet losses in series at handovers: 2-Mbps wireless link bandwidth, WFQ scheduling.

of the losses in series is performed at different network load, from 0.7 (70%) to 1.0 (100%). In the case when a VBR stream occupies a larger part of the bandwidth, a long series of lost packets can be located at a higher network load.

When the VBR flow share is just a small part of the bandwidth, we notice smaller consecutive losses at handovers, as shown in Figures 9.14 and 9.15, with

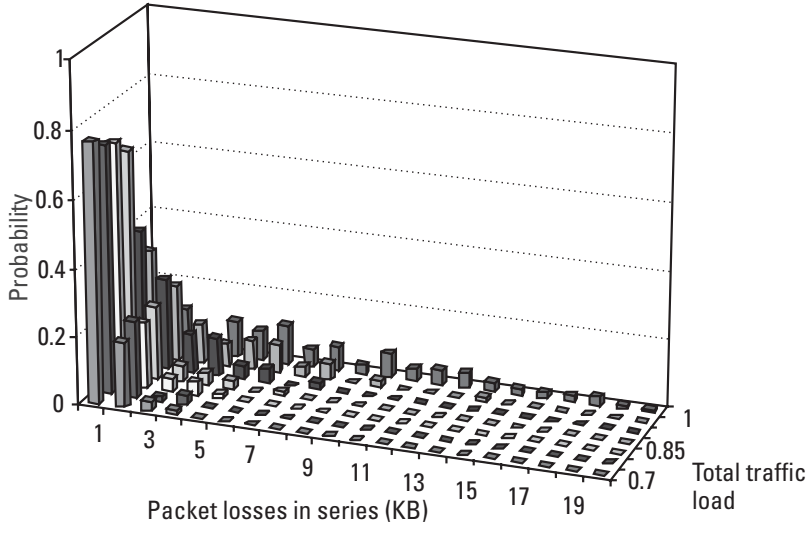


Figure 9.14 Probability distribution function of packet losses in series at handovers: 20-Mbps wireless link bandwidth, FCFS scheduling.

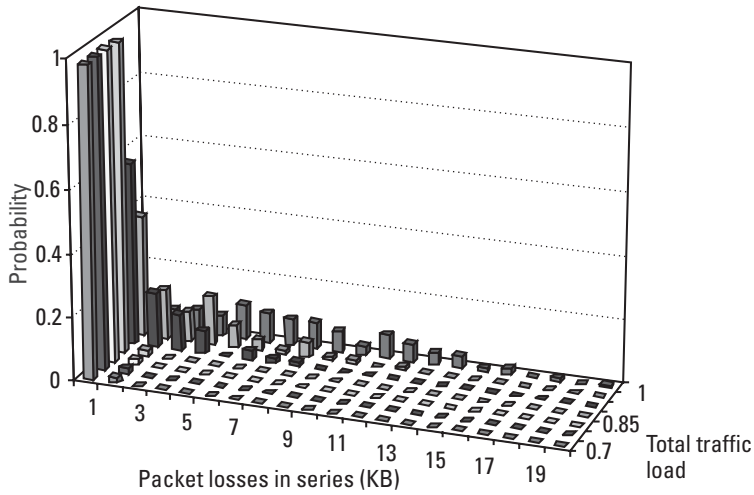


Figure 9.15 Probability distribution function of packet losses in series at handovers: 20-Mbps wireless link bandwidth, WFQ scheduling.

applied FCFS and WFQ scheduling, respectively. This is due to the background traffic that reduces the burstiness of the flow. Hence, scheduling discipline influences the losses. As one may expect, WFQ scheduling results in lower packet loss than FCFS in the case when the flow is multiplexed with other (background) flows.

Packet loss of the VBR flow as a function of time is shown in Figure 9.16. The simulations are performed at different network loads. We notice that a

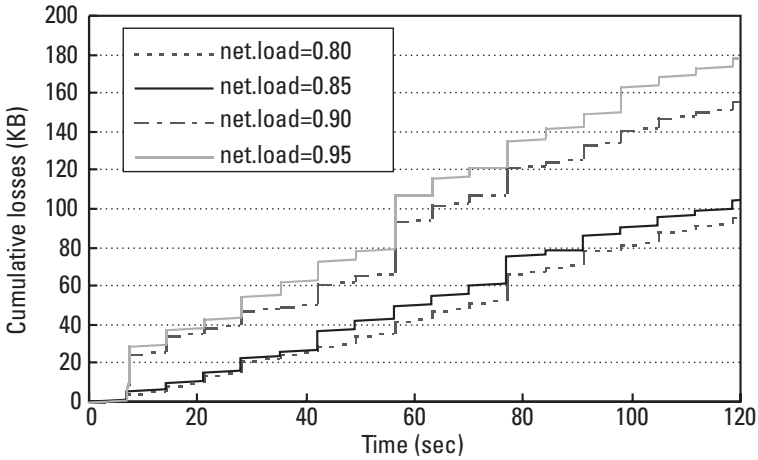


Figure 9.16 Packet loss at handovers of a VBR flow at different traffic load, and 2 Mbps wireless link bandwidth.

higher network load increases losses as well, due to longer queuing time at the network nodes.

In the case of soft handover we may have losses or duplicate packets. We can reduce packet losses in the soft handover scheme by semi-soft handover [22], which we described in Chapter 3. Typical semi-soft delay is 100 ms. Without losing generality, in our simulations we use single hop between the crossover node and the base stations. In this case we analyze the losses under two different differentiation mechanisms: priority mechanism and WFQ. But even in the case when priority is given to VBR packets over the background traffic, as shown in Figure 9.17(b), we notice the delay peak at each handover due to the additional semi-soft delay. If we compare packet delay of the hard handover, shown in Figure 9.17(a), to packet delay of the semi-soft handover, shown in Figure 9.17(c), one may notice a higher packet delay at handovers in the latter case. In this example, average packet delay of the VBR flow is 51.31 ms when using semi-soft handover, while the delay is 43.62 ms when hard handover is applied (mobility parameters are $r = 0.1$ km, and $v = 50$ km/hr, while total traffic load is 90%).

9.5.3 Handover Loss Analysis for Best-Effort Flows

Today's Internet is based on best-effort service. Most of the best-effort applications are TCP based, as we discussed in Chapter 5. TCP itself is characterized by the congestion avoidance mechanism (refer to Chapter 3). But, the protocol assumes that all losses occur due to congestion. Thus, handover losses may trigger the congestion avoidance mechanism. To analyze TCP performance we use a simulation experiment with a FTP flow (FTP is based on TCP). We attached the FTP source at the crossover node, although it can be far away from the mobile's home network. FTP is going in downlink (which will be the case in most situation), while ACKs are sent in uplink. We set one hop between the crossover node and each of the base stations, the old one and the new one. In the analysis we use the hard handover mechanism. On the other side, we use the Tahoe version of the TCP protocol. We assume wireless link without bit errors, thus all losses are only due to handovers.

Figure 9.18 shows the sequence numbers of TCP segments routed to the mobile in the downlink, and ACKs that are sent by the mobile to the FTP source in the uplink. We use 100-ms round-trip time of the TCP connection. The TCP packet size is 1,000 bytes. In the simulations, the mobile terminal initiates the handover at 6.24 seconds from the start of the connection. The route-update packet sent by the mobile terminal reaches the crossover node at 6.25 seconds. During the handover five consecutive packets of the TCP flow are lost. After the handover latency, the packets continue to arrive at the mobile terminal. For each received packet, after the handover, the TCP receiver at the

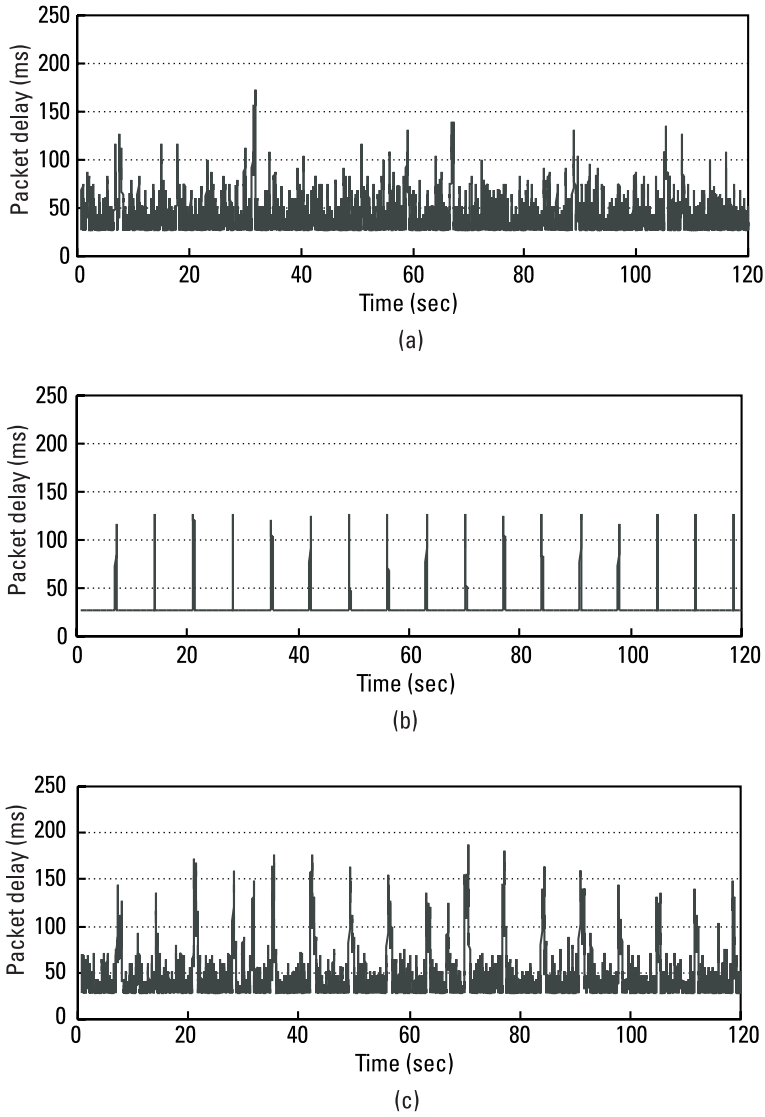


Figure 9.17 Packet delay of a VBR flow with different handover mechanisms: (a) hard handover, WFQ scheduling; (b) semi-soft handover, priority differentiation for the VBR flow; and (c) semi-soft handover, WFQ scheduling.

mobile sends a duplicate ACK to the FTP source (the horizontal line in Figure 9.18). On the sender's side (the FTP source), three duplicate ACKs in a row activate the congestion avoidance mechanism and the sender starts with retransmission of the lost packets. When we use TCP Tahoe, the source waits

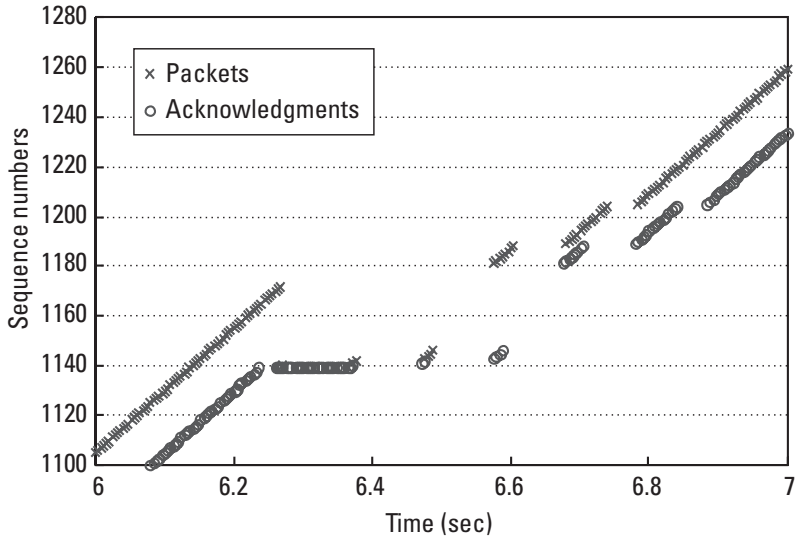


Figure 9.18 Sequence numbers and ACKs of a TCP flow in downlink at the handover.

for an ACK for the retransmitted packet before it continues with retransmissions. Upon receipt of a positive ACK from the mobile, the FTP sender increases the congestion window and continues with the next packets. The full TCP rate is regained at 6.78 seconds (i.e., after 0.54 second), as shown in Figure 9.18. The reason for such behavior is that TCP reacts to losses as if they were the result of network congestion. Behavior of TCP Reno at the handover is even worse than that of TCP Tahoe, because multiple losses within a single congestion window push the TCP Reno at the sender into timeout followed by a slow start.

In this experiment we assumed FTP flow in the downlink direction. In the opposite case, when the TCP is used to carry data from the mobile terminal to the far-end receiver, handover packet loss affects the acknowledgments. This is a trivial case, because missed ACKs does not interrupt the flow significantly. The next ACKs, if there is no congestion, will acknowledge the packet for which the ACK was lost. In the uplink direction, handover does not cause packet losses; thus, there will be no throughput degradation of the TCP flow.

The problem with TCP in mobile networks can be solved in two ways: (1) by adaptation of the TCP to the mobile environment [25–27], or (2) by creation of an efficient handover algorithm that will be transparent to the data flow, and, without losses or duplicate packets. According to the discussion above, handovers generate more problems to TCP flows in the downlink than in the uplink direction.

9.5.4 Performance Analysis of Different Traffic Types Under Location-Dependent Bit Errors

The wireless link is characterized by nonnegligible BER due to fading and shadowing. Wireless bit errors are related to the location in the cell; thus, users at different locations experience a different level of BER.

In a multiclass environment, according to the classification that we made in Chapter 5, we have various requirements on the QoS. Real-time services, such as CBR and VBR streams, require higher QoS (i.e., lower loss ratio and lower delay). Retransmission of lost or corrupted packets is not appropriate for real-time communication because of the unacceptable delays. On the other hand, losses in a nonreal-time flow, such as a best-effort flow, are recovered by retransmissions of the lost packets. But, we have different classes within nonreal-time services. We grouped the nonreal traffic in two groups: traffic with QoS requirements (e.g., Internet browsing), and traffic without any QoS guarantees (e.g., e-mail). The first traffic type is BEmin from class-A, while the second is class-B traffic. However, if we assume that bit errors rarely occur, then we may apply the same mechanisms for retransmission of the lost packets for both BEmin subclass of class-A and class-B traffic. Our tendency is to provide short-term and long-term fair scheduling of the flows under location-dependent bit errors in the wireless link.

For the purpose of analysis of wireless bit errors, we predefine the time interval of noticeable bit errors in the wireless channels for a given user. We use a VBR flow on 2-Mbps link bandwidth. To create a realistic scenario we multiplexed three flows on the link: one of each type CBR, VBR, and best effort. We simulate a 40% bit error ratio for the VBR flow in the time interval between 25 and 35 seconds from the simulation start. The other two flows are error-free. Out of the error-interval for the VBR flow, all traffic is error-free. The throughputs of all flows in the cell are shown in Figure 9.19.

If we assume that the MAC layer performs detection of the channel state considering the bit error ratio, then when MAC detects bit errors in the wireless link, VBR flow will not send packets. In that case, during the erroneous period of the VBR flow, its allocated bandwidth is used by the best-effort flow. But, if the VBR flow is real-time communication, then there will no possibility for compensation of the lost bandwidth due to bit errors in wireless channel. CBR flow does not have any changes on the throughput because it is error-free during the simulation, thus keeping its bandwidth allocated by the admission control at the connection start.

In Figure 9.20 we show the throughput of all three flows, using the same settings as in the previous simulation, but in this case we applied capacity isolation among the flows (i.e., complete partitioning) instead of the complete sharing. This differentiation policy causes a part of the wireless link bandwidth to be wasted due to the error-state of the VBR flow. On the other hand, the VBR flow

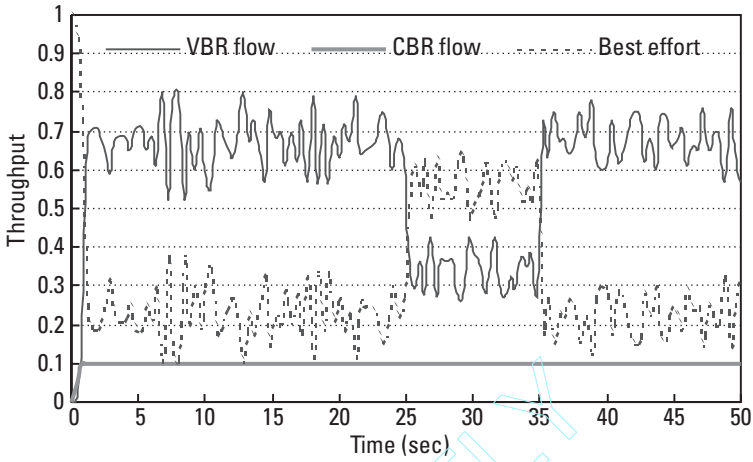


Figure 9.19 Influence of bit errors in the wireless link on a VBR flow (*vbrvideo1*) with complete sharing of the resources.

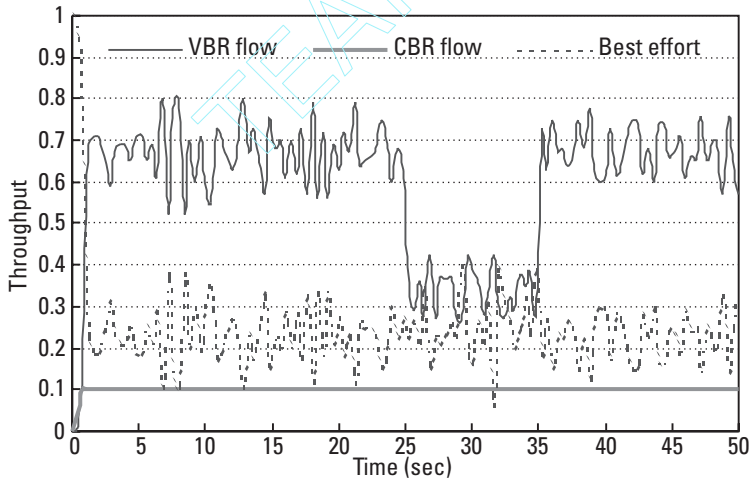


Figure 9.20 Influence of bit errors in the wireless link on a VBR flow (*vbrvideo1*) with complete partitioning of the resources.

is degraded due to the bit errors. Hence, capacity isolation as a way of flow differentiation leads to inefficient utilization of the wireless resources under the influence of location-dependent bit errors.

The analysis of an error-state of the CBR flow will lead to the same discussion as for the VBR flow. In the case of several flows belonging to a same class/subclass, most of the offered solutions [28, 29] propose a compensation principle: graceful service compensation for the lagging flows (that have lost

bandwidth due to wireless bit errors) and graceful service degradation for the leading flows (that have received more bandwidth due to bit errors in other flows on the same link). Such a compensation approach is helpful when we have a single traffic class in the network and nonreal-time communication. However, when we have real-time traffic and interactive communication, a compensation mechanism would not be beneficial. The main reason for this conclusion is that when we communicate in real time, lost information due to bit errors in the wireless link cannot be compensated because they will be out-of-date if transmitted at eventual compensation (this is similar to the discussion about retransmission of lost packets from a real-time flow). Second, in the error-free state, the throughput of an admitted real-time flow is enough for transmission of all information data, thus no compensation is needed.

A compensation method for the bit errors in the wireless link can be efficient in the case of traffic that has no strict QoS requirements, such as best-effort traffic. But, as we mentioned several times before, best-effort traffic is based on the TCP protocol. TCP is characterized by mechanisms (e.g., congestion avoidance mechanism) that are inert to fast changes of the bandwidth such as gaining additional bandwidth when another flow is in error-state and vice versa.

The above discussion leads to the need for the creation of an algorithm that will provide flexible scheduling of different traffic types under location-dependent bit errors in the wireless link. Such an algorithm is described in Chapter 11.

9.6 Discussion

QoS provisioning is crucial for the proper functioning of wireless cellular IP networks. In this chapter we conducted QoS analysis considering the two most significant features of mobile networks: handovers and bit errors in the wireless channel.

We performed handover analysis in wireless IP networks for different traffic types, such as CBR, VBR, and best effort. From the analysis, we concluded that higher user mobility, smaller cells, and higher traffic load in the cell cause higher loss due to handovers. This is due to the increased handover intensity, as well as the longer waiting time in the buffers at higher load. Through simulations, we showed that, while packet losses at handovers linearly increase in the case of a CBR flow, for a VBR flow they depend upon the burstiness of the flow at the handover events. Thus, for VBR flows, we may find lower packet losses due to handovers at higher user mobility than at lower mobility. Furthermore, consecutive packet losses have a negative influence on the ongoing traffic, causing significant performance degradation.

We compared hard and semi-soft handover through simulation analysis. It was shown that hard handover experiences a higher level of packet losses than

semi-soft handover, but the latter type adds additional delay, which is not desirable for real-time communication. Depending on the application type, the delay might be compensated by buffering at the receiving end (e.g., video/audio streaming). Also, packet losses can be recovered by retransmissions when it is possible (e.g., nonreal-time services).

Handover analysis with CBR flows showed dependence between packet losses and correlation of the background flows in the same cell. Burstiness of losses at handover increases as we increase the number of the flows multiplexed on the link, even at the same traffic load.

For analysis of the best-effort traffic we performed simulations with TCP flows using the hard handover. Simulations showed that packet losses at handovers cause activation of the TCP congestion avoidance mechanism, which is not necessary in such cases. This results from the fact that TCP was initially created for the wired Internet where packet losses occur only due to a congestion at the network nodes. Therefore, the throughput of TCP flows is being significantly degraded. Possible solutions are the modification of the TCP or the creation of an appropriate handover algorithm and using the classical TCP. Of course, an efficient handover scheme will actually improve not only the TCP performance, but also the QoS for the CBR and VBR traffic.

The second QoS issue that was analyzed in this chapter is the influence of bit errors in the wireless channel. Through simulations we observed the interaction among the flows when one of them experiences bit errors (we chose a VBR flow to be in error-state during a predefined time interval, because VBR is class-A traffic and has a time-varying bit rate). The analysis showed that complete partitioning of the resources leads to inefficient utilization of the wireless resources. On the other hand, complete sharing allows a flow that is in error-state to give its bandwidth to best-effort flows on the link during that state. Also, we considered that the compensation between leading and lagging flows is not applicable to real-time applications (e.g., voice over IP, multimedia streaming). The analysis showed the need for a flexible scheduling algorithm for the wireless segment that will provide QoS support to flows under the influence of bit errors in the channel, and at the same time will provide efficient and flexible resource utilization.

References

[1] Bakker, J. D., and R. Prasad, "Handoff in a Virtual Cellular Network," *Vehicular Technology Conference (VTC'99)*, Amsterdam, September 1999.

[2] Hadjiefthymiades, S., et al., "Mobility Management in an IP-Based Wireless ATM Network," *ACTS Mobile Summit 98*, Rhodos, Greece, May 1998.

[3] Toh, C.-K., *Wireless ATM and Ad-Hoc Networks: Protocols and Architectures*, Norwell, MA: Kluwer Academic Publishers, 1997.

[4] Karagiannis, G., and G. Heijenk, *Mobile IP – State of the Art Report*, Open Report, Ericsson, 13-07-1999.

[5] Dovrolis, C., D. Stiliadis, and P. Ramanathan, "Proportional Differentiated Services: Delay Differentiation and Packet Scheduling," *Proc. of ACM SIGCOMM'99,* Boston, MA, August 1999.

[6] Aiello, W. A., et al., "Competitive Queue Policies for Differentiated Services," *Infocom 2000*, Tel Aviv, Israel, March 2000.

[7] Nandagopal, T., et al., "Delay Differentiation and Adaptation in Core Stateless Networks," *Infocom 2000*, Tel Aviv, Israel, March 2000.

[8] Semret, N., et al., "Peering and Provisioning of Differentiated Internet Services," *Infocom 2000*, Tel Aviv, Israel, March 2000, 26–30.

[9] Floyd, S., and V. Jacobson, "Link-Sharing and Resource Management Models for Packet Networks," *IEEE/ACM Trans. on Networking*, Vol. 3, No. 4, August 1995.

[10] Zander, J., "Radio Resource Management in Future Wireless Networks—Requirements and Limitations," *Communication Magazine*, 1997.

[11] Janevski, T., and B. Spasenovski, "Performance Issues of Handoffs in Wireless IP Networks with Heterogeneous Traffic," *Wireless Access Systems—WAS 2000*, San Francisco, CA, December 4–6, 2000.

[12] Janevski, T., and B. Spasenovski, "QoS Analyses for VBR Differentiated Services in Cellular IP Networks," *Proc. of 10th Virginia Tech Symposium on Wireless and Personal Communications*, Blacksburg, VA, June 4–16, 2000.

[13] Janevski, T., and B. Spasenovski, "QoS Analyses of a Handoff in Cellular IP Networks," *CECS 2000*, Sofia, Bulgaria, May 18–19, 2000.

[14] Janevski, T., and B. Spasenovski, "Packet Loss in Cellular IP Networks," *ETAI 2000*, Skopje, Macedonia, September 21–23, 2000.

[15] Parekh, A. K., and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," *IEEE/ACM Trans. on Networking*, Vol. 1, June 1993, pp. 344–357.

[16] Veres, A., A. T. Campbell, and M. Barry, "Supporting Service Differentation in Wireless Packet Networks Using Distributed Control," *IEEE Journal on Selected Areas in Communication*, Vol. 19, No. 10, October 2001.

[17] Siris, V. A., B. Briscoe, and D. Songhurst, "Service Differentiation in Third Generation Mobile Networks," *3rd International Workshop on Quality on Future Internet Services (QofIS'02)*, Zurich, Switzerland, October 16–18, 2002.

[18] 3GPP TR 25.936, *Handovers for Real-Time Services from PS Domain (Release 4)*, V4.0.1, December 2001.

[19] Hernandez, E. J., and M. C. Chuah, "Transport Delays for UMTS VoIP," *WCNC'00*, Chicago, IL, September 2000.

[20] Stemm, M., and R. H. Katz, "Vertical Handoffs in Wireless Overlay Networks," *ACM Mobile Networking (MONET)*, Special Issue on Mobile Networking in the Internet, Summer 1998.

[21] 3GPP TR 25.841, *DSCH Power Control Improvement in Soft Handover*, V4.1.0, March 2001.

[22] Valko, A. G., et al., "On the Analysis of Cellular IP Access Networks," *IFIP Sixth International Workshop on Protocols for High Speed Networks (PfHSN'99)*, Salem, MA, August 1999.

[23] Seshan, S., H. Balakrishnan, and R. H. Katz, "Handoffs in Cellular Wireless Networks: The Deadalus Implementation and Experience," *Kluwer International Journal on Wireless Communications Systems*, 1996.

[24] Ramjee, R., "Supporting Connection Mobility in Wireless Networks," Ph.D. dissertation, University of Massachusetts, May 1997.

[25] Tamura, Y., Y. Tobe, and H. Tokuda, "EFR: A Retransmit Scheme for TCP in Wireless LANs," *IEEE Annual Conference on Local Computer Network (LCN'98)*, Boston, MA, October 1998, pp. 2–11.

[26] Calveras, A., and J. Paradells, "TCP/IP over Wireless Links: Performance Evaluation," *48th Annual Vehicular Technology Conference VTC '98*, Ottawa (Ontario), Canada, May 1998.

[27] Balakrishnan, H., S. Seshan, and R. H. Katz, "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," *ACM Mobicom'95*, Berkeley, CA, November 14–15, 1995.

[28] Lu, S., T. Nandagopal, and V. Bharghavan, "A Wireless Fair Service Algorithm for Packet Cellular Networks," *Mobicom'98*, Dallas, TX, October 1998, pp. 10–20.

[29] Bharghavan, V., and R. Srikant, "Fair Queueing in Wireless Packet Networks," *ACM SIGCOMM*, Vol. 27, No. 4, October 1997.

10

Handover Agents for QoS Support

10.1 Introduction

In the previous chapter we analyzed the performance of the existing types of handovers in wireless packet networks and learned of their disadvantages. The main problem during handover is the routing of the packets from the network to the mobile terminal. Packet losses occur at the handovers, but they should be avoided whenever possible because it causes QoS degradation. Communication in the reverse direction (i.e., the uplink), from the mobile terminal to the network, is less critical because the mobile terminal communicates through its current base station.

In this chapter we propose a mechanism that should improve handovers in cellular IP networks considering the QoS [1, 2]. Mobile IP is already standardized for providing global mobility (i.e., macromobility). This is a technique in which data is forwarded from the mobile's home network to a visited network, by using a home agent and a foreign agent (as we discussed in Chapter 3). The concept of Mobile IP is an imitation of the HLR-VLR concept in mobile networks such as GSM. Mobile IP is not adequate for handling micromobility—which requires fast handovers for real-time communication—and therefore, several different solutions for micromobility have been proposed, such as Cellular IP and HAWAII. A micromobility solution based on the Mobile IP protocol would introduce high delays and possible packet losses at the network nodes (if HA is far away from the mobile's current domain). Thus, such a micromobility concept would result in unacceptable performance for real-time communication (e.g., interactive services, voice service), as well as for best-effort services (e.g., throughput degradation of TCP flows).

Furthermore, FutureG (e.g., 4G mobile systems) should include heterogeneous access technologies. While 3G initiatives are based on packet-switched wide-area cellular networks, the future generation(s) mobile networks will include networks from 2G and 3G cellular networks to wireless LANs (e.g., IEEE 802.11 and HIPERLAN) and Bluetooth-based WPANs, as we discussed is Chapter 2. As a result, there will be truly IP-based access by the mobile users. For example, in a FutureG network a mobile user should be allowed to perform a handover during a real-time conversation from a wide-area cellular network to a wireless LAN or WPAN, as it moves from an outdoor environment into an office [3]. Therefore, we need to define a unified handover mechanism that will be applicable to multiclass heterogeneous access networks.

To avoid such problems considering the micromobility, we propose introducing additional modules at the network nodes, which we will denote as handover agents. These modules are software-based and should process handovers within the domain. Considering the macromobility (i.e., interdomain handovers), we propose using the Mobile IP protocol, which is the de facto standard for global mobility. The handover agents provide in-order and no-loss packet delivery during the handover in both directions, to and from the mobile. We describe the handover agent algorithm in detail in the following sections.

10.2 Handover Agent Algorithm for Wireless IP Networks

From the analysis in the previous chapter, we may classify the disadvantages of handovers into the following categories:

- *Packet loss:* highest in the case of hard handover, lower with soft handover;
- *Packet reordering:* typical for the soft handover scheme;
- *Packet delay:* highest at the chaining handover, but semi-soft handover may also introduce significant delay;
- *Additional signaling and/or buffering:* multicast-based algorithm requires buffering at each neighboring station, which consumes buffer space and processing resources.

10.2.1 Who May Initiate a Handover?

In most of the schemes considered so far, the handover is mobile initiated. In 2G cellular networks and 3G phase 1, handover is initiated by the network and assisted by the mobile terminal. Mobile-initiated handover is due to the transfer of the Ethernet principal from a wired to a wireless environment, such

as in a wireless LAN. If we want to support mobile networks with multiple traffic classes and create commercial cellular networks (not local computer networks), then it is difficult to provide guarantees if the mobile terminals control the handover process (e.g., choosing the target cell). For example, if there are several candidates for a target cell at the handover, the mobile will choose the destination cell without prior knowledge about the traffic conditions in the network (i.e., in the target cell and the neighboring cells). On the other hand, the mobile terminal can receive traffic information from the network via the serving base station. There are two problems with such an approach. First, maintaining information of the traffic conditions in the network would require additional memory space at the mobile terminal and signaling on the wireless link, as well as additional complexity of mobile terminals that should be cheap enough. Second, the mobile terminal can violate its rights, and thus the operator would not be able to provide desired QoS for different traffic classes.

Thus, in the case of class-A handover initiation by the mobile, the problem will be the admission control (we analyzed the admission control mechanism for wireless IP networks in Chapter 7 and we proposed a hybrid admission algorithm in Chapter 8). Therefore, that network should control the handover for class-A traffic. It may be mobile initiated, but the network should make a decision whether to perform the handover or not. Of course, the mobile terminal should be allowed to initiate handover for class-B connections, because there is no QoS support for that traffic class.

The most appropriate way to conduct handover control is to apply it at the nodes that are closest to the wireless interface—that is, the base stations. In 2G and 3G mobile systems the control and management of base stations is given to a centralized node (e.g., base station controller, radio network controller). In the handover agent algorithm we propose handover initiation by the network, but we give the control to the base stations. The centralized control would result in additional signaling traffic and transmission costs (transmission has its highest costs in a telecommunications network).

10.2.2 Handover Types on a Link Layer

Considering the access technology in the wireless interface only, there are two possible types of handovers: hard and soft handover. So far, we have three basic wireless access technologies: FDMA, TDMA, and CDMA. Usually, implemented or proposed wireless access technologies are based on their combinations (e.g., GSM radio access is TDMA/FDMA, UTRA-FDD is FDMA/CDMA, and UTRA-TDD is FDMA/TDMA/CDMA). We may apply hard handover in all cases. However, the soft handover is applicable in radio access technologies that include CDMA-based techniques.

Although our attention is towards micromobility support, for the sake of completeness we will refer to possible problems of the soft handover on a link layer in wireless IP (i.e., all-IP) networks. In 2G CDMA networks, such as the IS-95 system, a centralized *selection and distribution unit* (SDU) is responsible for data delivery in the forward direction. The SDU distributes streams of the same data over layer-2 circuits (layer 2 is in reference to the OSI model) to multiple base stations that belong to the active-set of the soft handover. Each base station then relays the data to the mobile terminal. The mobile's radio system synchronizes radio channel frames with the base stations and combines the signals received from different base stations to obtain a single copy of the received data. In the reverse direction, the mobile terminal ensures data synchronization (i.e., matching layer-2 frames) for the copies of the same data sent to multiple base stations. The SDU then selects one of the frames as the final copy of the data in uplink.

We are interested in all-IP wireless networks. In that case base stations should use IP protocols for data transport as well as signaling (e.g., routing of the traffic, performing IP-layer mobility management, or QoS management). In such environment, however, soft handover is not so straightforward. The first problem is loss of data content synchronization. Even though the CDMA radio interface can synchronize layer-2 frames, it cannot guarantee on its own that the matching frames from different base stations will carry copies of the same data. For example, packets may be lost on their way due to congestion. Also, frames of the same data may arrive at the mobile terminal at different times due to random transport delays (e.g., different congestion at different nodes, different propagation time). There are few efforts to provide IP-layer synchronization of the data for soft handover in wireless IP networks [4]. So, we may assume that with the current IP mobility approach, soft handover in an all-IP wireless network may lead to packet loss or duplicate packets.

Therefore, we need location and mobility management for an access domain that may comprise one or multiple access technologies (e.g., UMTS, wireless LAN, and WPAN), a typical scenario in a FutureG network. Also, it should support QoS requirements by different traffic classes (i.e., class-A and class-B). For that purpose, we define a handover agent scheme for intradomain mobility management.

10.2.3 Handover Agents

To explain the handover agent algorithm, at this point we assume that the crossover node is discovered (the discovery of the crossover node will be explained later).

The proposed handover scheme is based on establishing handover agents at network nodes within a domain. We use a two-level architecture. The first

level (i.e., phase) involves the corresponding host and the gateway. The Mobile IP protocol is used in this level to handle the macromobility. The second level involves the gateway and the mobile terminal, where the handover agent mechanism is used to manage the micromobility.

We will explain the functioning of the handover agent scheme using the time diagram shown in Figure 10.1. We assume that the mobile terminal communicates to only one base station at a time. The base stations send periodic packets to the mobile terminals, which we call beacons or paging messages (we refer to this mobility function as paging). A beacon is a signaling packet in a wireless LAN, while paging messages may be found in 2G and 3G mobile systems. A mobile terminal performs periodic measurements of the beacon signal strengths from the base stations, and then the mobile sends a measurement report to the base station to inform it about the possible targets for a handover. In a case of class-B connection, the mobile is also allowed to initiate a handover. When the base station decides to perform a handover (based on the report on received signal strength by the mobile as well as bit error ratio in the wireless channel), it activates the HA, which starts to scan all incoming packets to

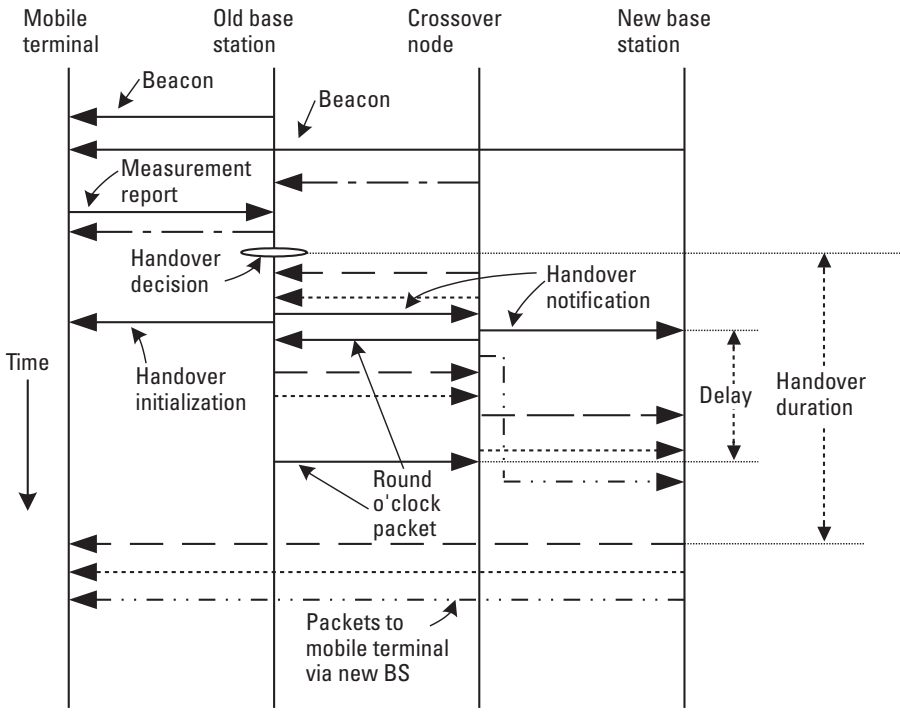


Figure 10.1 Time diagram of the handover agent scheme.

the base station towards the mobile terminal involved in the handover. At the same time, the old base station sends a message to the mobile terminal to order handover execution (i.e., transfer of the mobile from the old to the new wireless access point). The old base station tunnels a handover-notification message to the new base station to change the route of the packets towards mobile's new location. After receiving the message for handover initiation, the mobile terminal starts to listen to the new base station. Also, the mobile terminal sends all packets in the uplink through the new base station. To be sure that the handover-notification packet will reach the crossover node before the first data packet from the mobile terminal via the new base station, we give priority to the signaling messages over the data messages. It should not affect the quality, because the wired part of the network should have higher link capacity (it is easy to upgrade the capacity of wired links) than the wireless part. After receiving the handover-notification packet, the crossover node activates a handover agent, which sends a new signaling packet towards the old base station. We refer to this packet as the "round o'clock" packet, because it travels a round-trip between the crossover node and the old base station. The crossover node changes the routing information for the mobile terminal (i.e., the old route is deleted and the new route, to the new base station, is created). All packets addressed to the mobile terminal, which will reach the crossover node after the handover-notification packet, are buffered at the crossover node.

Until the reception of the round o'clock packet, the handover agent at the old base station automatically starts to forward all packets addressed to the mobile terminal back to the crossover node in the reverse direction. These packets were routed to the old base station in the time interval between initiation of the handover and the time when the crossover node receives the handover-initiation packet. The purpose of the round o'clock packet is to inform the old base station that there are no more packets to be forwarded to the new one. Thus, the old base station can delete the routing information for the mobile terminal. After receiving the route-update packet, the handover agent at the old base station forwards the round o'clock packet back to the crossover node, and it deletes the mobile's old routing information on its way.

All packets that are rerouted from the old base station towards the crossover node are further forwarded to the new base station without any waiting time. After receiving the round o'clock packet, the handover agent at the crossover node starts to forward the packets, which were buffered at the crossover node during the round-trip of the round o'clock packet, to the new base station. That ends the task of the handover agent at the crossover node for the given connection. Because a node can be a crossover node and a base station at the same time, we propose implementation of handover agents at all nodes of the wireless access network.

10.3 Routing in the Wireless Access Network

In the previous section we explained the handover scheme based on handover agents at the network nodes. Now, we need to define necessary functionalities for mobility support in a wireless IP network (i.e., domain): routing of the IP packets from/to the mobile terminals and location control.

A conceptual model of the wireless IP network is given in Figure 10.2. The network is connected to the global Internet via a so-called gateway node. In the gateway node we should have an HA and an FA, which are defined by the Mobile IP protocol [5]. So, we use Mobile IP to control the movement of the mobile between different wireless IP networks. The packets that should be routed to the mobile terminal have the address of the gateway as a destination address (i.e., care-of address). The Mobile IP is inefficient due to the triangle routing between the HA, the FA, and the corresponding node that sends the packets to the mobile. We can solve such problems by temporarily memorizing the IP address of the FA (of the mobile's current network) at the source. This problem is solved, however, in Mobile IPv6, and hence the FA is omitted. Within the wireless IP network, the gateway forwards the packets addressed to the mobile terminal using the unique IP address of the mobile. The mobile terminal address has no significance inside the wireless IP network. So, any unique IP addresses can be used to identify mobile terminals within the access network. Also, the network nodes maintain a logical connection tree topology over a possibly mesh wireless IP network infrastructure. The base stations are leaves of the

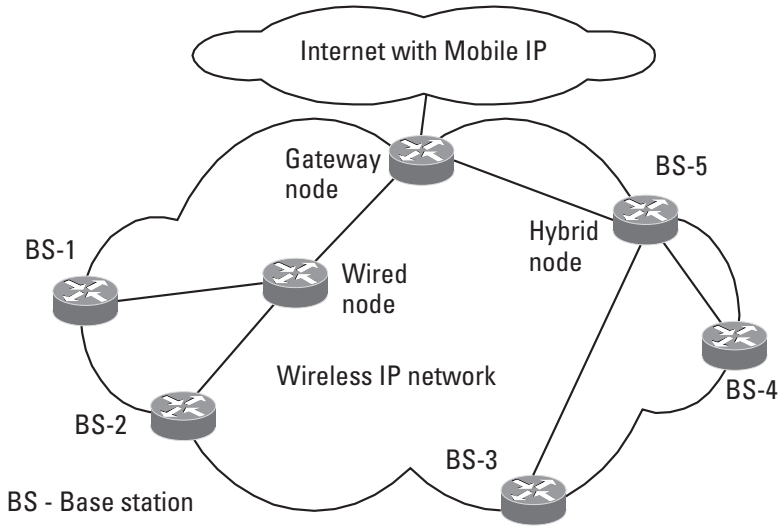


Figure 10.2 Conceptual topology of a wireless IP network.

tree, and there are also wired and hybrid nodes as well as a root node (i.e., gateway), as shown in Figure 10.2.

The packet transmitted to the mobile terminal uses the downlink routing algorithm within the wireless IP network. The algorithm is illustrated in Figure 10.3. It is targeted to suit a multiclass environment. Therefore, the base stations perform admission control of class-A flows to provide the desired QoS guarantees (an admission control algorithm for multiclass wireless IP networks is given in Chapter 8). Using this approach, before each class-A connection we need to establish a communication between the gateway node and the mobile's

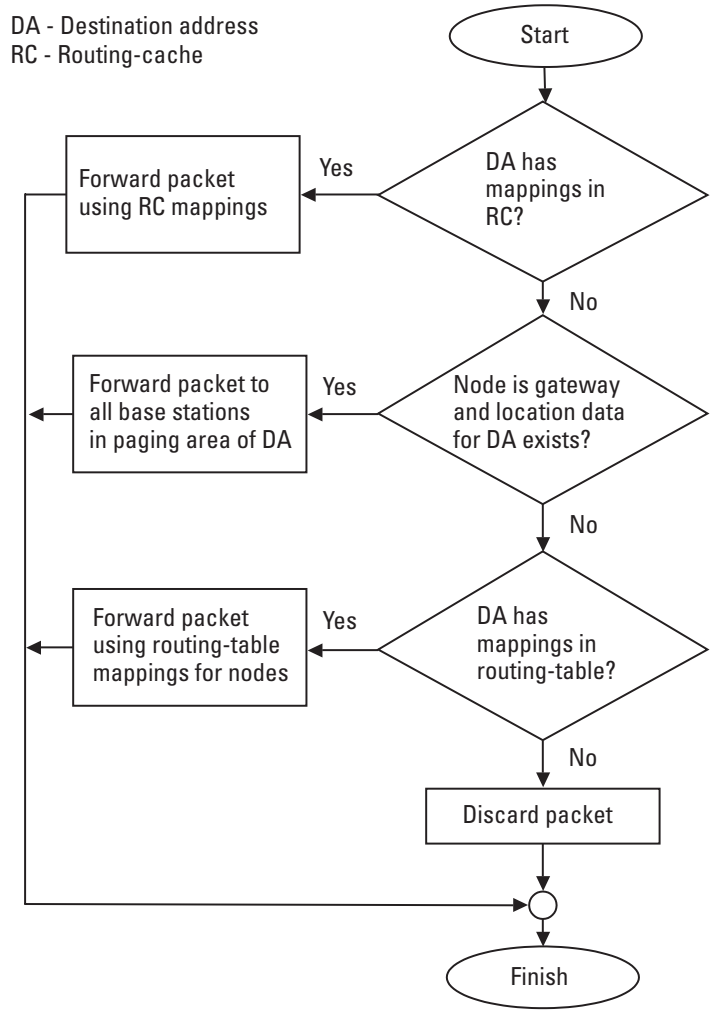


Figure 10.3 Downlink routing algorithm in a wireless IP network.

base station. Using the location control, if the mobile is attached to the network, the gateway has information about the paging/location area in which it resides. This concept is similar to the location control in today's cellular networks. To locate the mobile, the gateway node sends a paging message to all base stations in the current paging area of the mobile terminal. These beacons are routed by using fixed mappings in the routing-tables of intermediate network nodes. Such mapping are created or deleted when a base station is added to the access network or an existing base station is taken out (or is out of order at the moment), respectively. After receiving the paging message, the mobile terminal sends an acknowledgment to the gateway via the serving base station.

After locating the mobile, the current base station performs an admission control. The result of the admission control (accepted/rejected call) is sent to the gateway node from the base station as an *admission control packet*. The admission control packet contains information whether a class-A call is granted or not. If the call request is rejected, the gateway sends a notification to the far-end sender (i.e., the source). If the call is accepted, then the admission control packet is used to update or create routing information for the mobile on the way between the base station and the gateway. The created routing path is used for routing all packets that are addressed to the mobile terminal, until a handover is initiated. To store temporary routing information, each network node maintains a routing cache. So, there are two different types of routing information at each node in the wireless IP network:

- *Routing-table*, which maintains semi-permanent routing information that is referred to the routers in the access network;
- *Routing-cache*, which maintains routing information for mobile terminals.

Routing information in caches may be further classified into two groups:

- *Soft route mapping*, which expires after a certain timeout if it is not refreshed—this should be used for class-B connections;
- *Semi-soft route mapping*, which is explicitly deleted by a signaling packet at the handover—this should be used for class-A connections.

Packets addressed to the mobile host are routed on a hop-by-hop basis, using the mappings from the routing-cache or the routing-table.

In the reverse direction, packets transmitted by the mobile are routed via the gateway using the same routing information. Uplink routing is shown in

Figures 10.4 through 10.6. Rerouting of packets by using the handover agent handover scheme is the same for both traffic classes, A and B. We further distinguish between uplink routing of class-A and class-B traffic.

A class-A flow connection is initiated or terminated by IP-layer signaling messages (i.e., *connection-start* and *connection-end* signaling). Therefore, each node must maintain a classifier that will sort packets according to their class and type of information (e.g., data packet, handover signaling, class-A connection setup signaling). So, connection-start signaling packets are used to initially create semi-soft mappings in the routing-caches, as shown in Figure 10.5. Connection-end signaling transfers the semi-soft state of the routing information into a soft state, thus allowing simultaneous B-class flow(s) from the same user to continue with the communication. With this approach we eliminate paging multicast of downlink B-class packets if they continue to arrive at the mobile terminal after the termination of the class-A connection.

According to our discussion in Chapter 8, the class-B flow does not go through the admission control. All users attached to the network are allowed to initiate a class-B connection. In a case of a class-B flow towards the mobile terminal, the first packet of the flow is transferred to all base station in the paging

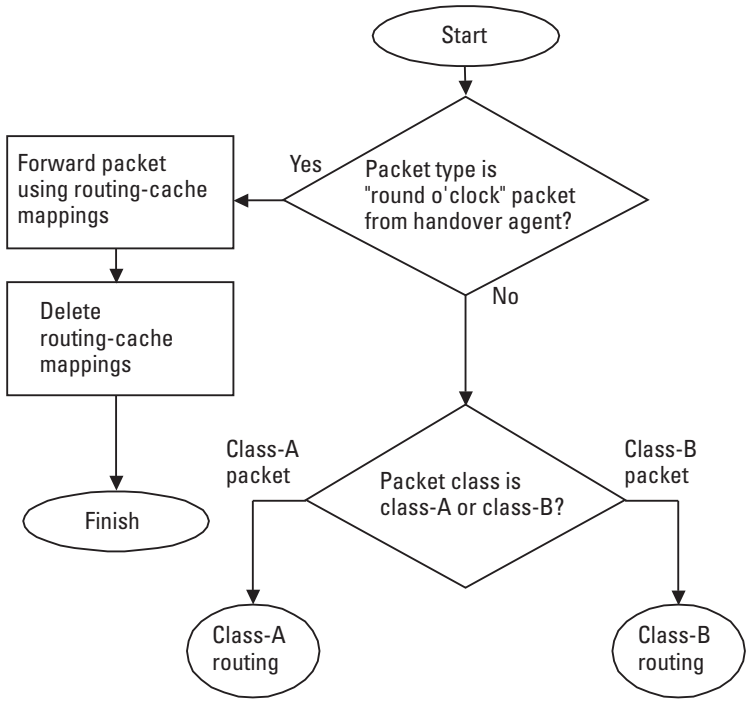


Figure 10.4 Uplink routing algorithm in a wireless IP network.

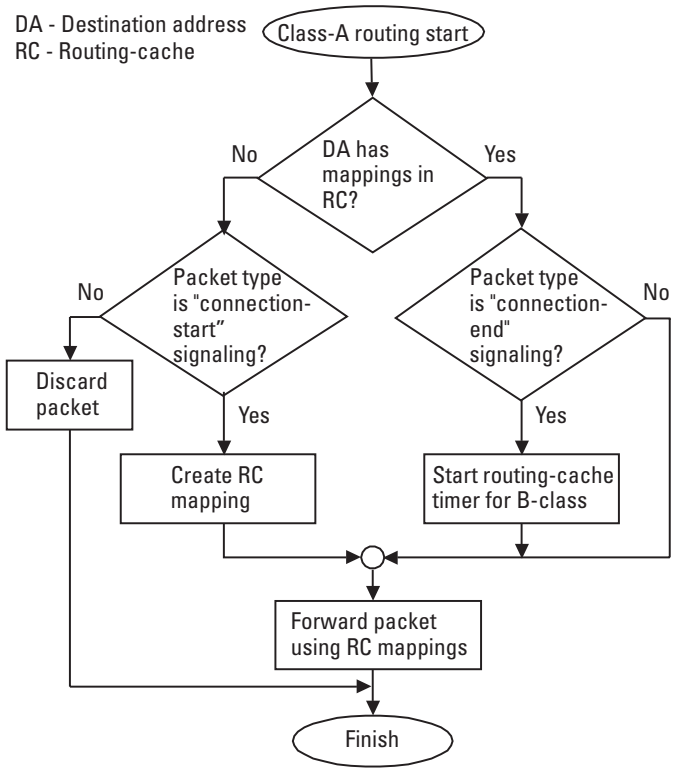


Figure 10.5 Class-A uplink routing algorithm.

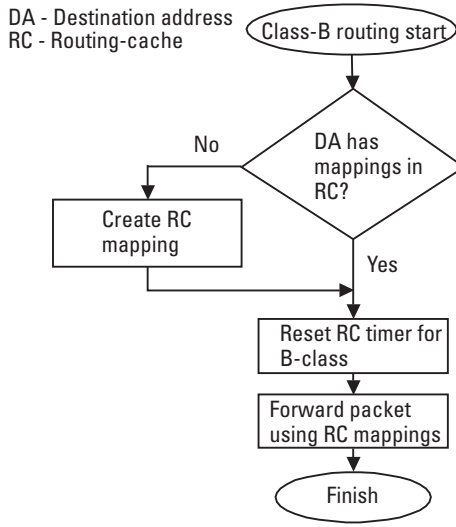


Figure 10.6 Class-B uplink routing algorithm.

area of the mobile, according to the location information that is present at the gateway. After receiving the first B-packet, the mobile terminal sends a packet towards the gateway that should create routing information in all intermediate nodes. Also, when there is no packet from the mobile terminal, it can send periodical paging-update packets to refresh the routing and/or paging information [6]. Such an approach will result in additional signaling traffic in the wireless link, which is not desirable. In our approach, route-updates are sent to the gateway only during the connection duration. In the reverse direction (i.e., in uplink), B-packets that are sent by the mobile are routed hop-by-hop to the gateway. These packets create or update the routing information at all intermediate nodes between the mobile's base station and the gateway, as shown in Figure 10.6.

There is a difference between class-A and class-B flows considering the routing information at the network nodes. In a class-A flow the routing information is kept at all intermediate nodes during the entire call duration; in a class-B flow it expires if it is not updated by a data packet from/to the mobile or by a route-update packet. If there is no data or route-update packets, then routing mappings for the given connection are cleared at all nodes in the wireless IP network after the timeout expires.

10.4 Location Control and Paging

We define location control by grouping the calls into two groups: class-A and class-B. Class-A calls must go through the admission control, but this is not the case with class-B calls. So, each mobile that is attached to the network (i.e., the gateway has location information for the mobile) is allowed to receive or transmit class-B packets. Considering the class-B traffic, all attached users are always available, in a similar manner to that in a local IP network. Hence, a user can receive an e-mail or location-based information, or download files while having a phone conversation (for voice calls we should use class-A). According to this discussion, considering class-A traffic, users can be idle or busy. When a mobile maintains an active class-A connection, it is in busy state. In that state, there are semi-soft routing-cache mappings at all intermediate nodes between the mobile's base station and the gateway. Considering class-B traffic, users are always available, but they may have two different states: idle or on-line. In the idle state there are no routing-cache mappings (soft or semi-soft) for the mobile at the network nodes. In the on-line state there are soft routing-cache mappings. Each class-B packet triggers a transition from idle to on-line state. Also, each class-A call to the mobile triggers a transition from any other state, idle or on-line, to busy state. Termination of a class-A connection results in the transition from busy to on-line state, because in that case the uplink routing algorithm for

class-A traffic causes the starting of the route timer for B-class (i.e., the semi-soft route is transformed into a soft route). When the route timer expires, the routing information is cleared from the caches and the mobile transitions into idle state. According to the previous discussion, we may define a state-model of a mobile terminal in a wireless IP network, as shown in Figure 10.7.

Let us now consider location management. Two possible types of location management are location registration and paging [7]. Management load for them is exclusive. If paging is executed every time over the entire service area of the network (i.e., in all cells), then location registration is not needed. If location registration (i.e., location update) is executed every time a mobile terminal crosses a cell border, then paging is not needed because the system will always have the information of the mobile's current cell. In 2G cellular networks a combination of these two types of location management is implemented (i.e., the entire service area is divided into several location areas where each location area includes many cells). Some of the new killer services, however, will be location-based (refer to Chapter 2). Therefore, we need to maintain location information per user to provide the advanced location services. Thus, the location management should be location-registration oriented or be combined with paging by using small paging areas. The size of such paging areas should be several cells, not many.

Thus, we divide the service area of a wireless IP network into location/paging areas, similarly to today's cellular networks. Cells from one *paging area* (PA) form a multicast group at the gateway. So, each beacon carries a *PA identifier* (PAI). This is similar to the *location area identifier* (LAI) in current cellular networks. Mobile terminals listen to beacons. When a mobile detects a different PAI in the beacon than its current one, it performs location update by sending a message with the new PAI addressed to the gateway.

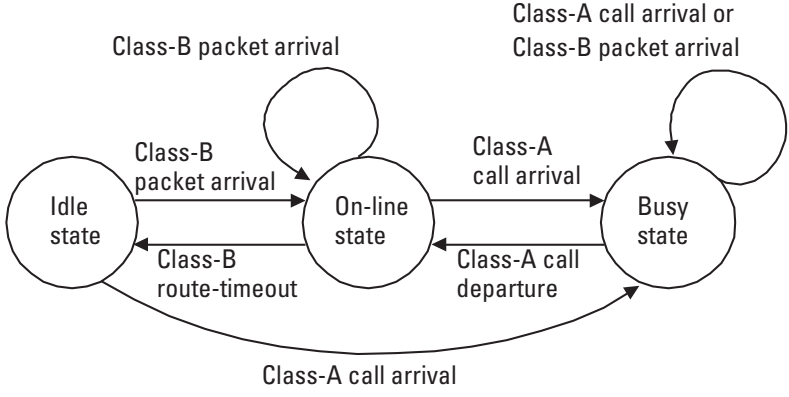


Figure 10.7 State-model of a wireless IP mobile terminal.

When a new call from class-A or a B-packet arrives in the network, the gateway looks up the state of the mobile terminal: attached/detached. If the terminal is attached to the network, then the gateway checks the cache memories considering the mobile. If there are route mappings in the cache, then these mappings are used to route packets all the way from the gateway towards the mobile's base station. In case of a class-A call, it is necessary to send a signaling packet to initiate admission control at the base station. All class-B packets are directly routed from the gateway towards the mobile terminal.

If the mobile terminal is in idle state (i.e., there are no route mappings in the caches) and new B-packet or class-A call request arrives, then it is multicast into the wireless paging area. Each base station decapsulates the packet and forwards it. The mobile responds to its base station with an acknowledgment packet. Then, the base station forwards the packet to the gateway (in case of class-B traffic) or sends an admission control signaling message for the call acceptance/rejection to the gateway. These packets are also used to create routing-cache mappings at all intermediate nodes. The mobile state changes from idle state to online state (if the packet type is class-B) or to busy state (if the packet type is class-A).

The user can detach from the network by turning off the mobile. But when the user leaves the coverage area of the network domain or when the mobile terminal goes off due to a low battery, then the gateway cannot be informed at the time of the event. In that case, when a packet or a call arrives for that user, the gateway multicasts the packet in the paging area according to the latest location update from the mobile terminal. If it does not get an answer, then paging is done over the whole network area. If there is still no answer, the user is detached from the network and the gateway informs the mobile's HA that the user is unreachable. One may choose to apply periodic location updates by the mobile even in idle state, similar to a circuit-switched cellular network. Of course, such updates should be in longer time intervals (e.g., several hours).

10.5 Discovery of the Crossover Node

To be able to perform a handover, we need to determine the crossover node between the old and the new base station. Again, we discuss separately the determination of the crossover node for class-A and for class-B flows.

10.5.1 Crossover Node Discovery for B Flows

In the handover agent algorithm the old base station initiates the handover. The handover decision is based on the measurements of the signal strengths taken by the mobile. Because we usually use a hexagonal cell form, the mobile transmits

to the base station the list of six neighbors with the strongest signal strength as well as the signal strength of the serving cell. The base station initiates the handover when one or more cells has better signal strength and/or quality at the mobile's location (with appropriate hysteresis to avoid the ping-pong effect at the handover). The base station sends a packet to the mobile terminal for handover initiation. After receiving the packet, the mobile terminal hands over to the new base station. At the same time, the old base station sends a handover-notification packet addressed to the new base station. The packet uses the mobile's old route on the way to the crossover node (the new route does not exists at this moment). Then, the crossover node will be the node where this packet has to be routed onto a different link than the old route. After receiving the rerouting-packet, the crossover node activates its handover agent. At the same time, the rerouting-packet is being forwarded to the new base station, and on its way it is used to create routing mapping in caches. Thus, a new route is created between the crossover node and the new base station. In the reverse direction, the mobile terminal sends the packets through the new base station. Each packet in uplink is routed towards the gateway via the crossover node using the route mappings at the intermediate nodes. If there are no route mappings, then routing of the packets from the mobile to the gateway node is performed hop-by-hop using the semi-permanent information in routing-tables.

10.5.2 Crossover Node Discovery for A Flows

Class-A flows have specific guarantees on the QoS, and thus, they require explicit signaling before every handover. To obtain smaller handover latency, we need to reduce signaling as much as possible. One solution is to introduce a centralized server (at some of the network nodes) that will maintain information regarding the resource occupancy in cells under its control. A second possible solution is to provide exchange of messages between the current base station of the mobile and the target base station. A third solution has each base station maintaining information about the traffic conditions at each of its neighboring cells (they are limited in number).

The first solution may be efficient, but the problem is in additional delay of the handover due to signaling between the control server and the current base station. The second solution is inefficient because of two reasons: First, it adds long delay to the handover latency, and second, it might happen that the target base station is unable to accept the handover, thus another neighboring base station should be probed. The third solution requires each base station to inform its neighbors about every traffic change within the cell. In this case, the network operator should define neighbors for every base station. Thus, a base station should store information about traffic condition in each of its neighboring cells. This solution provides minimum delay at the handover and allows using the

same crossover node discovery algorithm for all traffic classes. Therefore, one may find the last solution as the most efficient, because it provides the lowest handover delay and most unified handover scheme for all traffic types. In that case the admission control for the target cell is performed at the old base station, and thus the crossover node discovery is the same for class-A as for class-B.

10.6 Performance Analysis of the Handover Agent Scheme

We may distinguish among four different aspects when analyzing the handovers and micro-mobility: a traffic aspect, a routing aspect, a radio propagation aspect, and a performance aspect.

The traffic aspect is related to the transparency of the handover to the ongoing traffic in both the old cell and the new cell. When a handover is performed, the connection occupies resources in the new cell, thus influencing the traffic in that cell. In a multiclass environment, the highest QoS degradation will be experience by the class-B traffic and lower or no-degradation should be maintained for class-A traffic.

The routing aspect is related to rerouting of the connection at a handover event. The rerouting is related to establishing a path between the crossover node and the new base station in the wired part of the network.

The radio propagation aspect is related to the cell coverage, and thus it is dependent upon the interference, fading, and shadowing.

The performance aspect is related to the QoS requirements from the traffic flows. A given flow may have strict QoS demands that are not allowed to be degraded at the handover, or, it may have "softer" QoS requirements that are negotiable (i.e., it can accept certain QoS degradation at the handover). When QoS requirements are strict and the target cell cannot satisfy them because it will deteriorate the ongoing connections in that cell, the handover should be rejected by the cell.

In the previous chapter we analyzed some handover schemes and their dependence upon the mobility parameters (e.g., cell radius, velocity, and so forth), traffic load, type of traffic, and number of hops between the old base station and the crossover node. Primary attention, however, was given to the downlink direction (to the mobile terminal) as more critical (statistically, the end users have larger IP traffic volume in downlink than in uplink—that is, the traffic is not symmetrical).

Different traffic types (i.e., CBR, VBR and best-effort flows) exhibit different behavior at handover. In Chapter 9 we analyzed the packet loss and packet delay for hard handover. Soft and semi-soft handover schemes reduce the handover losses, but they are not eliminated. Also, soft handovers can cause packet reordering.

The handover agent scheme introduces additional packet delay due to the round trip of the packets after the handover initialization (i.e., between the crossover node and the old base station). However, this delay affects only the packets that will reach the crossover node in the time period between the handover initialization and time when the round o'clock packet reaches the crossover node again. Hence, added delay to the buffered packets is less or equal to the round-trip time between the crossover node and the old base station.

To analyze the characteristics of the proposed handover scheme, we perform comparison analysis between the handover agent scheme, and the hard and the semi-soft handover. With hard handover, all packets that will pass the crossover node before the route-update packet arrives from the new base station are considered lost. With soft handover, packets are sent to the mobile terminal through both the old and the new base station. The mobile, however, can receive packets from only one base station at a time. Thus, if the packets travel longer from the crossover node to the old base station than to the new one, we can have packet losses at the handover. In the opposite case, we can have duplicate packets at the mobile terminal. When the semi-soft handover scheme is applied in the network, the mobile terminal sends a rerouting packet to the crossover node through the new base station and then continues to listen at the old base station during a certain semi-soft delay time. After the semi-soft delay, which can be anywhere between the round-trip time between the mobile terminal and the crossover node and the route timeout, the mobile hands over to the new cell. But, the semi-soft delay might be too long or too short, depending on the traffic load during the handover and on the network topology. So, it may result in packet losses or in duplicate packets.

In the following, we present results of simulations, which are performed assuming 2-Mbps wireless links (i.e., cell bandwidth). The crossover node is directly connected to the old and to the new base station (i.e., there is only one hop between them). We performed simulations with different traffic types.

Figure 10.8 shows the throughput of a CBR flow at higher handover intensity (we maintain the mobility parameters to obtain very high handover intensity, which is near 1 handover/sec in these simulations). We used different values for the semi-soft delay (in the case of semi-soft handover): 20 ms (comparable with the round-trip time between the crossover node and the old base station) and 100 ms [6]. The throughput is normalized, and the throughput is calculated by considering all packets from the beginning of the simulation time. The handover agent scheme has the highest throughput because it does not have any packet losses. However, the throughput of the handover agent is slightly lower than 1 due to added packet delay at the handovers. The hard handover has significantly lower throughput because of the packet losses. The throughput with the semi-soft handover may get close to the throughput with the handover agent scheme only in the case when semi-soft delay, which is deterministically

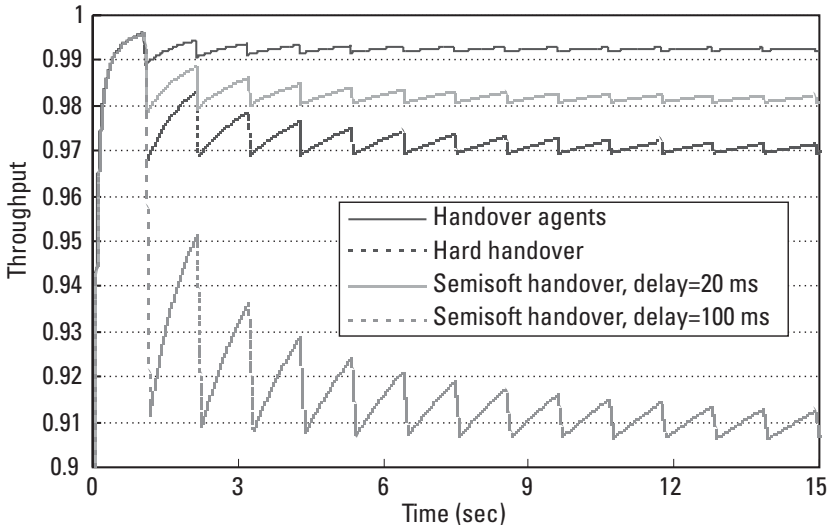


Figure 10.8 Normalized throughput of a CBR flow for different handover schemes.

specified, equals the delay difference between the crossover node and the old base station, and between the crossover node and the new base station. Because the delay depends upon the traffic load in the network and upon the network topology, the deterministic semi-soft delay cannot provide protection from packet losses or duplicate packets at handovers.

Figure 10.9 shows the throughput of the CBR flow under the same traffic conditions as in the previous simulation, but with different mobility parameters, such as cell sizes and mobility of the users. The results show that the advantages of the handover agent over other handover schemes become more significant at higher mobility of the users and smaller cells (i.e., higher handover intensity).

To analyze the gain of the handover agent scheme we also performed analysis with a VBR flow. We already showed that VBR traffic has different behavior at handovers than CBR traffic. In the simulations we used the VBR flow *vbrvideo1*. Figure 10.10 shows the normalized throughput of the VBR flow when different handover schemes are applied in the network (i.e., the handover agent scheme and hard handover). The curves are obtained by calculation of the throughput on 400-ms intervals. The lowest curve is related to the throughput gain when the handover agent scheme is used instead of the hard handover. One may notice that the throughput gain varies from handover to handover. This is due to the variable bit rate of the observed flow. It results in different packet losses at different handovers (we showed this feature of the VBR traffic in Chapter 9).

The handover agent scheme eliminates the packet losses due to handovers. The trade-off is the additional delay of the packets that arrive at the crossover

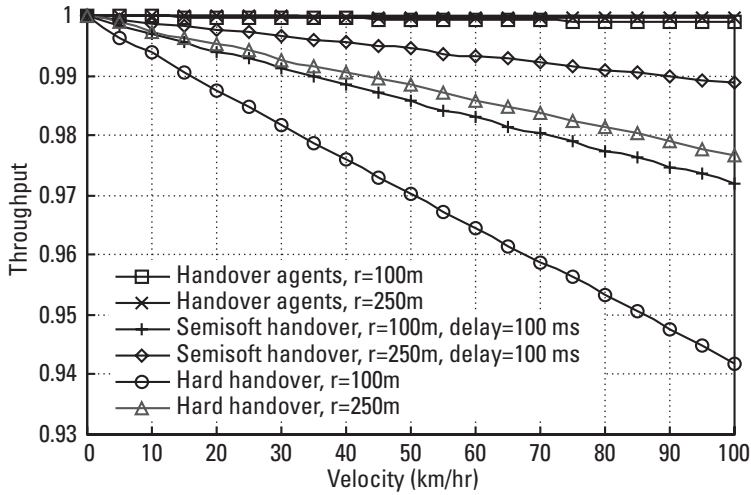


Figure 10.9 Throughput of a CBR flow versus mobility of the users for different handover schemes.

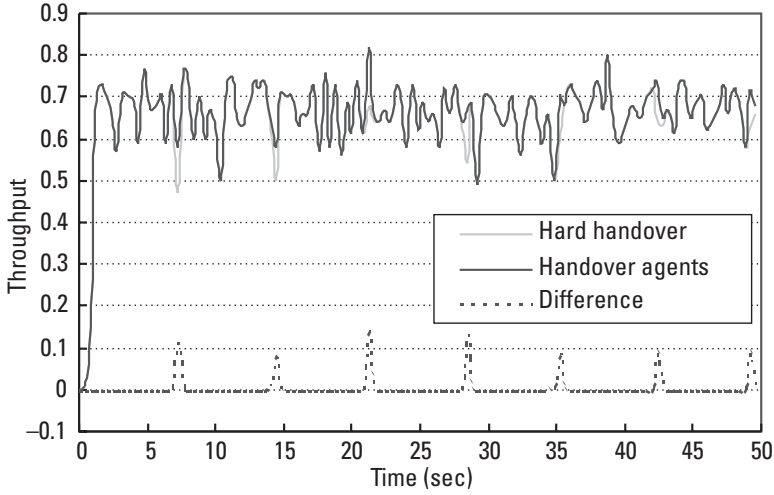


Figure 10.10 Comparison of the normalized throughput of a VBR flow with different handover schemes.

node during the handover. The handover agent scheme causes small additional traffic load in the core network (i.e., the wired links) only during the handover; but the scheme does not increase the traffic volume in the wireless link, which is scarcer than in the wired ones. Furthermore, the wired links have deterministic

quality (very low BER), while wireless links are also characterized by the non-negligible bit errors in the wireless channel.

To analyze the additional delay of packets during a handover with the handover agent scheme, we conducted an experiment with the hard handover algorithm (which is the fastest one, without any buffering of the packets for handover purposes), and then we repeated the experiment with applied handover agents. Again, we use a network topology with a single hop between the crossover node and the base stations. The propagation and processing time in the wired link is set to 8 ms, while wireless access time is set to 20 ms (these values are arbitrarily chosen). The measurement of the packet delay is performed at two different network nodes: the crossover node and the mobile terminal. The results of the simulations, given in Figure 10.11, show that there is no noticeable difference between probability distribution functions of the packet delay with hard handover and the handover agent scheme. Therefore, one may conclude that added packet delay by the handover agent scheme is negligible from the aspect of average packet delay; thus, it is worth accepting such additional delay to reduce the packet loss at handovers to zero and eliminate duplicate packets at the mobile terminal. The uplink communication is trivial, because the mobile communicates with one base station at a time; thus, it sends the packet through its current base station where it is then routed to the gateway and further to Internet by using the Mobile IP protocol for handling the global mobility.

We analyzed only CBR and VBR flows in this section. Best-effort traffic, which is mainly based on the TCP/IP protocol stack, is sensitive to packet loss,

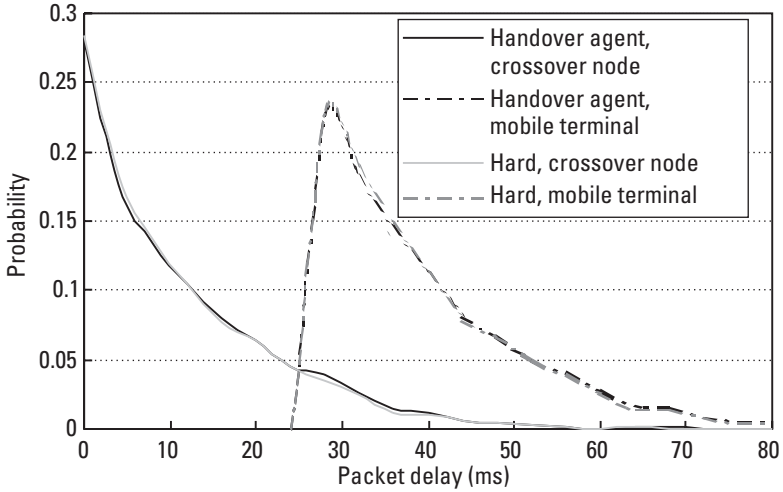


Figure 10.11 Probability distribution function of the handover delay for the handover agent scheme and hard handover at different network nodes.

as we showed in Chapter 9. Hence, if there are no packet losses and there is no significant delay in the communication link, then there is no change in the behavior of the best-effort flows.

10.7 Discussion

In this chapter we proposed a handover algorithm that provides a solution to the problems that occur at the handovers in wireless IP networks [1, 2]. The proposed algorithm is based on establishing handover agents at the nodes of the wireless network. Hence, we denoted it as the handover agent scheme. The handover agents provide efficient packet rerouting during the handover between the old and the new base station. During the handover latency, all packets that arrive at the crossover node are buffered until the round o'clock packet, which is sent by the handover agent at the crossover node towards the old base station, is received back at the crossover node. Thus, the handover agent scheme eliminates the packet losses as well as duplicate packets due to handovers. So, there is no need for creation of a robust application for the mobile terminals that will handle the losses or duplicate packets at handovers. The trade-off for the lossless handovers is the added delay to the packets of the flow only during the handover duration, which was shown to be negligible.

To support micromobility, we proposed a routing algorithm for the wireless access network, which is based on temporary routing-cache and semi-permanent routing-tables at each network node within the domain. The admission control packet of the class-A flow triggers the creation of mappings in the routing-caches, which last for the entire duration of the connection (they are deleted when the connection is terminated). Also, mappings in the routing-cache are created or updated (if they already exist) by every class-B packet from mobile terminal. In this case route mappings expire after a certain timeout unless a packet from the mobile updates them. Network nodes maintain a logical tree topology over possibly a mesh network. The rerouting of the connection is located at the crossover node, which is the closest node to the new base station of all nodes on the old route. This way we get an efficient routing and packet delivery in sequence.

Through a simulation analysis we compared the handover agent scheme with the hard and the semi-soft handover for different traffic types, and we confirmed the gain that we get when we are using the handover agent scheme. The gain in constant for CBR flows, while for VBR flows it is dependent upon the burstiness of the flow during the handover.

If we compare the handover agent scheme with other approaches to the micromobility issue, such as Cellular IP [6] (Section 3.5.2.1), HAWAII (Section 3.5.2.2), hierarchical Mobile IPv6 (MIPv6) [8], or fast handovers for

MIPv6 [9], we may find many similarities considering the management of best-effort traffic. For example, comparison with Cellular IP, HAWAII, and hierarchical Mobile IP shows that their handover performance depends only on the position of crossover routing point. Furthermore, fast handovers for MIPv6 are based on establishing a routing path between the old access router (i.e., base station) and the new one, to enable the mobile node to send and receive IP packets during the handover. Hence, this algorithm has similarities with handover agent with regard to the access network. However, all micromobility proposals lack the QoS support. The advantage of handover agent is that it provides a conceptual solution for QoS support in a wireless IP environment, based on traffic differentiation.

We summarize the advantages of the handover agent scheme into the following elements:

- Provides QoS support in a multiclass wireless IP network;
- Eliminates packet losses due to handovers;
- Avoids duplicate packets at the mobile terminal;
- Provides efficient rerouting of the connection at handovers;
- Eliminates the need for the creation of a robust application in mobile terminals to deal with QoS problems due to mobility.

References

[1] Janevski, T., and B. Spasenovski, "A Novel QoS Scheme for Handoffs in Wireless IP Networks," *IEEE Wireless and Communications Networking Conference—WCNC 2000*, Chicago, IL, September 23–28, 2000.

[2] Janevski, T., and B. Spasenovski, "QoS Improvement on Handovers in Wireless IP Networks," *Wireless 2000 Conference*, Calgary, Alberta, Canada, July 10–12, 2000.

[3] Misra, A., et al., "IDMP-Based Fast Handoffs and Paging in IP-Based 4G Mobile Networks," *IEEE Communication Magazine*, Vol. 40, No. 3, March 2002, pp. 138–145.

[4] Zhang, T., and P. Agrawal, "IP-Based Base Stations and Soft Handoff in All-IP Wireless Networks," *IEEE Personal Communication*, October 2001, Vol. 8, No. 5, pp. 24–30.

[5] Perkins, C., (ed.), *IP Mobility Support*, RFC2002, proposed standard, IETF Mobile IP Working Group, October 1996.

[6] Valko, A. G., et al., "On the Analysis of Cellular IP Access Networks," *IFIP Sixth International Workshop on Protocols for High Speed Networks (PfHSN'99)*, Salem, MA, August 1999.

[7] Yumiba, H., K. Imai, and M. Yabusaki, "IP-Based IMT Network Platform," *IEEE Personal Communication*, Vol. 8, No. 5, October 2001, pp. 18–23.

[8] Soliman, H., et al., "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)," IETF draft, draft-ietf-mobileip-hmipv6-07.txt, October 2002.

[9] Koodli, R., (ed.), "Fast Handovers for Mobile IPv6," IETF draft, draft-ietf-mobileip-fast-mipv6-05.txt, September 2002.

11

QoS Provisioning in Wireless IP Networks Through Class-Based Queuing

11.1 Introduction

Time-varying transmission quality in wireless links introduces problems for QoS support to different traffic classes. The question is how to provide the guaranteed data rate when there is a higher bit error ratio in the wireless channel (we use the notion of *channel* in the sense of a connection to a single user; it does not mean that it is a circuit-switched channel). On the other hand, we cannot predict the behavior of the wireless interface in a wider area because of the users' mobility. Also, one may expect wireless bit errors to occur in bursts.

To provide solutions to such problems, this chapter defines a scheduler for wireless IP networks that should be used at wireless access points (i.e., base stations). Effort-limited scheduling for a wireless environment is proposed in [1]. This is obtained by extension of WFQ via dynamic weight adjustments. To provide fairness among the flows, the algorithm introduces factor coefficients that are used to adjust the throughputs of the flows at a higher bit error ratio. Through such factor coefficients, network operators are given the possibility of controlling the QoS level at error occurrence. But, this scheme does not provide QoS support in a multiclass environment, which is expected in future wireless multimedia networks.

Packet scheduling in a case of bursty traffic is analyzed in [2]. It follows from that analysis that WFQ can provide service differentiation in cellular Internet only in specific network conditions. The performance of WFQ is satisfactory only at higher traffic loads. Also, it is proven that propagation time,

existing TCP connections, and user distribution have little influence on the performances of the WFQ scheme. Hence, appropriate modification of WFQ may be helpful for packet scheduling in a wireless IP network.

Fair queuing of multiclass traffic for a hybrid wireless/wired network is proposed in [3]. In particular, scheduling is considered on a MAC layer in an ATM network. Different traffic classes are distinguished from one another by using priorities. For example, real-time data uses a wireless fair queuing model. On the other hand, a weighted round robin scheduler processes nonreal-time data. Best-effort flows are serviced using the FIFO (i.e., FCFS) mechanism. The drawback of such a scheme, however, is the lack of a mechanism for support of real-time flow's throughput under location-dependent bit errors in the wireless channel.

In some approaches to the problem of bit errors in the wireless link, authors use a compensation method—that is, compensation of the flows that experienced bit errors using the bandwidth of the flows that received more bandwidth (i.e., higher QoS) during the error-state of other flows. This solution can be found in different proposals [4–8]. But, the question is whether the compensation method is applicable to real-time flows.

The main goal in this chapter is to analyze and define of a scheduling algorithm for wireless IP networks that support multiclass traffic. The scheduler development is guided by the following requests:

1. When all flows are error-free, the throughput of the scheduler must be the same as with applied WFQ within every traffic class with QoS support (i.e., within class-A).
2. The capacity loss of a specific flow in error-state should be dependent on traffic class.
3. Flows within the same class experiencing equal error rates should experience equal capacity loss.
4. Network administrators should be involved only in setting the bounds for guaranteed services.
5. Real-time flows should be adjusted to their error-free throughput in real time (if possible), and there should be no compensation on channel errors during error-free periods.
6. Nonreal-time flows within class-A may be allowed to use some compensation model.
7. Scheduling for best-effort traffic from class-B should be as simple as possible.

The scheduler is built in two steps:

- Differentiation between class-A (guaranteed) services and class-B (best-effort) services, as well as between different subclasses within class-A is based on priority.
- Differentiation of the flows within a subclass of class-A is based on the modification of weights of the flows for real-time traffic, and wireless fair scheduling (e.g., compensation) for nonreal-time traffic.

In Section 11.3 we provide an overview of existing scheduling disciplines for wireless networks. Then we propose a scheduling mechanism for multiclass wireless IP networks.

11.2 Wireless Network and Channel Model

The network architecture is given in Chapter 6. The network consists of interconnected routers. The routers that are used as wireless access points are referred to as base stations. We assume that every base station serves a single cell in the network. A flow is said to be active if it has packets queued at the network nodes; otherwise, it is referred to as a passive flow. All active flows in a cell share the same wireless link. We usually assume that there is one flow per active user.

Mobile hosts do not have information about the global state in the wireless link in terms of how many and which other mobile terminals have packets to transmit. Also, they are constrained by battery power and processing power. Hence, base stations should perform scheduling in both the uplink and downlink. Every mobile terminal in a cell communicates with a base station; thus, there is only one wireless hop in each direction. It is assumed that the scheduler in the base station views the traffic as a set of flows to the users. Users can be fast-moving mobile hosts that often make changes in the link state. Therefore, the wireless scheduler should be flexible enough to follow the channel behavior. The error state is tied with single users (i.e., it is location-dependent). The flows of different users are assumed to be independent.

Let us make some main assumptions about the wireless channel model. A wireless channel refers to bandwidth allocated to a single connection, which may be fixed or varying in time. Due to different factors—such as fading, shadowing, and multipath—the total capacity of a wireless link, as well as the capacity of wireless channels, is dynamically varying. Due to the random position of mobile hosts within a cell, errors are location-dependent. They are also bursty in nature due to different time scales on which changes occur in a user's position and in packet transmission delay (e.g., a user with a velocity of 50 km/hr travels 0.29m during time intervals of 21 ms, which is the time needed to transmit 1,000-byte IP packet over a 384-Kbps link). All traffic with QoS support must go through

the admission control phase, while for best-effort traffic multiple mobile terminals collide over the bandwidth.

We assume that the admission control module in the base station has admitted all active flows by assigning traffic class and a bandwidth share. For differentiation of the classes we can use ToS and DS bits in IPv4 and IPv6, respectively.

11.3 Design of Wireless Scheduling Algorithms

Because of the scarcity of wireless resources, the large user population, and burstiness of the traffic, it is necessary to apply aggressive admission control to fully utilize wireless resources. We already have discussed that future wireless networks will include multiple traffic types. In a multiclass environment different services have different QoS requirements. Also, within the same traffic class we should provide fairness between different flows because wireless media can exhibit high, variable error rates that affect network users.

In wireline networks, fluid fair queuing (i.e., WFQ) has long been a concept for providing bounded delay channel access and fairness among packets flows over a shared unidirectional link. WFQ provides full separation between flows. The minimum guarantees are unaffected by the behavior of other flows. Fluid-fair, however, assumes that the channel is error-free, or at least that errors are not location-dependent (i.e., all backlogged flows have the ability to transmit at a given time, or none of the flows can). Adapting fair queuing to the wireless environment is not a trivial task because of the unique problems in the wireless channels, such as location-dependent bursty errors, channel contention (e.g., in wireless LAN, best-effort traffic in 3G network), as well as joint scheduling of uplink and downlink flows.

There are several existing proposals for wireless fair queuing. The basic goal of wireless fair queuing is to emulate WFQ when all flows perceive error-free channels, but to swap channel allocation between flows that perceive channel error and flows that perceive a clean channel. The main differences between different wireless fair queuing algorithms are [7]:

- How the swapping takes place;
- Between which flows the swapping takes place;
- How the compensation model is structured.

11.3.1 Wireline and Wireless Fluid Fair Queuing

Let us first describe the fair queuing in a case of no-channel errors, and consider why such an approach fails to provide fair service when the environment is

error-prone. Weighted fair queuing model WFQ allows any flow i to be granted channel capacity over a given time interval $[t_1, t_2]$ so it minimizes (6.1) from Chapter 6. In WFQ each packet is associated with a start tag and a finish tag, which correspond to the virtual time at which the first bit of the packet and the last bit of the packet are served by that mechanism. Let $B(t)$ denote the set of backlogged flows at time t. If we denote with $A_{i,k}$ the arrival time of the kth packet of the ith flow, and $S_{i,k}$ and $F_{i,k}$ are start time and finish time for that packet, respectively, then we may write

$$S_{i,k} = \max\left\{V\left(A_{i,k}\right); F_{i,k-1}\right\} \tag{11.1}$$

where $V(t)$ is the virtual time at time t, which denotes the current round of service. Thus, the packets are sorted according to the minimum eligible finish time. The finish time is computed from the start time by adding the time needed to send a packet of size L_p:

$$F_{i,k} = S_{i,k} + \frac{L_p}{r_i} \tag{11.2}$$

where r_i is the rate of the flow i. If we denote with $C(t)$ the link capacity at time t, which is dynamically varying, we can obtain the progression of the virtual time by using the following:

$$\frac{dV(t)}{dt} = \frac{C(t)}{\sum_{i \in B(t)} r_i} \tag{11.3}$$

Often, approximations of WFQ are used, such as WRR and *start-time fair queuing* (STFQ) that do not need to compute dV/dt given by (11.3).

However, WFQ provides two important guarantees: a bounded delay and associated minimum throughput of the flow. In WFQ the flow cannot reclaim time from another flow that used its empty channel time (when the first flow had no packets to transmit). However, in a wireless environment a flow may be backlogged, but unable to transmit due to channel errors.

We will show how the WFQ behaves in a wireless environment through a simple example. Let flows f_1 and f_2 be two flows that share a wireless channel, and let both have equal weights. So, when both flows are error-free, each of them should receive $W_1 = W_2 = 0.5$ channel allocation. Let us consider time window [0,1]. We assume that flow f_1 is error-free over the entire time window. But, let us suppose that flow f_2 perceives channel error in the time interval [0,0.5]. Then, in the interval [0,0.5] WFQ will allocate all bandwidth to flow f_1, because f_2 perceives channel errors. In the interval [0.5,1] both flows are error-

free, and WFQ allocates half of channel capacity to each of them. Finally, over the considered time window, flow f_1 gets average channel allocation $W_1 = (1 + 0.5)/2 = 0.75$, while flow f_2 gets $W_2 = (0 + 0.5)/2 = 0.25$. So, the first flow receives 0.25 more channel allocation than the fair share of 0.5, while the second flow receives 0.25 less than its error-free channel share.

The question is whether, in a case of error-prone channel, the backlogged flow should be compensated for the lost capacity in the future. In other words, should the channel loss and empty queues be treated in the same way or differently? Most of the wireless fair queuing algorithms apply a compensation model for flows that perceive channel error during some time intervals. However, compensation of the flows should be limited to avoid degradation of other flows. So, there is a trade-off between separation and compensation of the flows.

11.3.2 WFQ Algorithms

There are several different approaches for wireless fair queuing. One should note, however, that all of them are based on compensation (i.e., lead and lag model—or credit and debit model) and are created for nonreal-time communication such as best-effort traffic. Almost all of these algorithms are created for wireless LANs (e.g., IEEE 802.11). All of them are modifications and adaptations of WFQ or its approximation algorithms (e.g., WRR) to wireless networks.

In this section we describe the most well-known wireless fair scheduling algorithms. At this point, it is convenient to define certain terms—such as lagging flow, leading flow, backlogged flow—that are used in the descriptions of the algorithms.

A flow is said to be leading if it has received channel allocation in excess of its error-free service. A flow is lagging if it has received less channel allocation than its error-free service. A flow is backlogged if it has packets to transmit over the channel.

Idealized Wireless Fair Queuing

Idealized wireless fair queuing (IWFQ) uses WFQ for the error-free service [6]. Both start and finish tags are assigned according to the WFQ. The service tag for a flow is set to the finish tag of its head-of-line packet. IWFQ selects the flow with a minimum service tag among all backlogged flows that are error-free. The lead of the leading flow is the difference between its real service tag and its service tag in an error-free channel. However, the service tag is not allowed to increase/decrease by more/less than a predefined bound. IWFQ always allocates the slot (channel time) to the error-free flow with the lowest tag until it either perceives an error channel or its finish tag becomes greater than that of some other flow with an error-free channel. IWFQ was the first algorithm to propose adaptation of WFQ to a wireless environment [9]. It provides long-term fairness

and bounded delay channel access. The possible drawback is that lagging flows can capture the channel, and starve out other flows. Hence, IWFQ does not support graceful degradation of service.

Wireless Packet Scheduling

The *wireless packet scheduling* (WPS) packet scheduler uses WRR with spreading as its error-free service [10]. WRR with spreading is identical to the schedule generated by WFQ if all flows are backlogged. WPS generates a *frame* of slot allocation from the WRR-spreading algorithm and provides fairness by swapping time allocations between mobile terminals experiencing error bursts and currently error-free terminals. The compensation is two-fold. WPS first tries to swap slots within a frame. If this fails, then it maintains the difference between the real service and the fair service for the flow by changing the effective weight in each frame based on the result of the previous frame. Hence, it attempts to provide graceful trading of the bandwidth between the leading and the lagging flows. This way it provides bounded delay channel access and long-term fairness, and at the same time it prevents the total channel capture by using the effective weights.

Channel-Condition Independent Packet Fair Queuing

In *channel-condition independent packet fair queuing* (CIF-Q), for error-free service STFQ is used [5, 10]. As we already stated, STFQ is an approximation of WFQ that does not require dV/dt computation by setting the virtual time $V(t)$ to the start tag of the transmitting packet. Each flow has a lag, which is defined as the difference between the error-free service and the real perceived service. If the lag is positive, than the flow is lagging; while in the opposite case it is a leading flow. This scheduling mechanism provides a graceful linear degradation for leading flows. For that purpose CIF-Q introduces a parameter α, which is a probability that a leading flow will retain its allocated slot, while $1 - \alpha$ is the probability that it will relinquish the slot to the lagging flows. CIF-Q can provide short-term and long-term fairness and bounded delay channel access.

Server-Based Fairness Approach

Server-based fairness approach (SBFA) reserves part of the bandwidth for compensation of the lagging flows via so-called virtual compensation flow [11]. It conceptually differs from other wireless fair scheduling algorithms. When a backlogged flow is allocated channel time, but it cannot transmit due to channel errors, then it requests service time (e.g., a slot) in the compensation flow. When a compensation flow is allocated a slot, it gives the slot to the flow to which its head-of-line request belongs. If there are no slots for compensation, then the bandwidth of the compensation flow is shared among all flows. SBFA does not monitor the lead of the leading flows. Hence, leading flows do not give up their

lead. This algorithm provides long-term fairness, but not short-term fairness or worst-case delay bounds. A lagging flow may request compensation slots until it receives its error-free fair service. However, SBFA is bounded by the reserved bandwidth for the virtual compensation flow. If this portion of the link bandwidth is less than the lags of all backlogged flows over some time interval, then long-term fairness cannot be guaranteed.

Wireless Fair Service

The *wireless fair service* (WFS) scheduling algorithm [12] uses WFQ scheduling for error-free wireless link. It allocates to each flow two parameters: a rate weight r_i and delay weight φ_i for a flow i. The start tag is computed using the rate weight: $S_{i,k} = \left\{ V\left(A_{i,k}\right), S_{i,k-1} + \frac{L_{i,k-1}}{r_i} \right\}$. The finish tag is computed using the delay tag: $F_{i,k} = S_{i,k} + L_{i,k}/\varphi_i$. Using the delay and bandwidth weights allows for delay-bandwidth decoupling. If a backlogged flow perceives channel errors, its lag is increased only if there is a backlogged error-free flow that increases its lead. Each flow is bounded by per-flow parameters—that is, a lead bound $l_i^{\max}$ and a lag bound $b_i^{\max}$. A leading flow with a current lead l_i relinquishes $l_i/l_i^{\max}$ of its allocated service time. A lagging flow with a current lag b_i receives a fraction $b_i / \sum_{j \in B} b_j$ of all relinquished slots by leading flows, where B is the set of backlogged flows. This way, WFS provides fair compensation among the lagging flows. Degradation of leading flows is graceful, and a fraction of the bandwidth relinquished by the leading flows decreases exponentially. The WFS algorithm provides both short-term and long-term fairness, as well as delay and throughput bounds.

Channel State Dependent Packet Scheduling

Channel state dependent packet scheduling (CSDPS) uses a WFQ-like scheduling discipline for error-free service (e.g., WFQ and WRR). This algorithm does not provide compensation between lagging and leading flows. CSPDS does not measure lead and lag of flows, and therefore it is simple for implementation. When service time is allocated to a flow that perceives channel error, then that flow is skipped and the service time is given to the next eligible flow in the WRR cycle. Thus, it may happen that a leading flow increases its lead. Because there is no compensation, this mechanism does not provide short-term and long-term fairness. However, it provides throughput guarantees to error-free channels. Also, if all flows are backlogged with equal probability, lagging flows can reduce their lag over the long term.

Discussion on Design Approaches for Wireless Fair Scheduling

Considering the described algorithms, we may distinguish among three design issues in wireless fair scheduling algorithms [7]: (1) error-free service algorithm,

(2) lead-lag model, and (3) compensation algorithm. For error-free service WFQ is used, or its modifications WRR with spreading and STFQ. There are two possibilities for the lead-lag model: (2a) lagging flow is compensated irrespective of whether its lost service time was used by an error-free flow (e.g., IWFQ, CIF-Q, SBFA); and (2b) lagging flow is compensated only if another flow that took its slot is prepared to relinquish a slot in the future (e.g., WPS, WFS). Considering the compensation between lagging and leading flows, in general, there are three approaches: (3a) no compensation—the flow perceiving channel error is skipped (e.g., CSPDS); (3b) swapping service time (i.e., slots) between the leading and the lagging flows (e.g., IWFQ, WFS, CIF-Q); and (3c) reservation of bandwidth for compensation (e.g., SBFQ).

All of the algorithms are created on the basis that the channel state is known. So, the scheduler should have information about the channel state for each backlogged flow. The key idea is the monitoring of the wireless channel for each flow and then making predictions about the future channel state. Errors are usually bursty in nature and correlated in successive time intervals. But they are usually uncorrelated over longer time intervals, thus making channel prediction possible using the Markov state model, even using a simple one-step prediction by the two-state Markov model [4, 7] (Section 6.5).

11.3.3 Service Differentiation Applied to Existing Systems

In this section we give examples of particular proposals for service differentiation in existing or standardized mobile packet-based networks, such as IEEE 802.11 wireless LAN and 3G mobile networks.

Service Differentiation in IEEE 802.11 Wireless LAN

Wireless LANs provide superior bandwidth compared to any existing cellular technology. The state-of-the-art standard in this area is IEEE 802.11b, which provides data rates up to 11 Mbps using the 2.4-GHz frequency band (there are also higher speed alternatives, such as IEEE 802.11a and IEEE 802.11g). However, it lacks QoS support—that is, it does not have implemented mechanisms for service differentiation.

For example, service differentiation may be based on modification of function of the IEEE 802.11 network, which was initially created to support best-effort traffic. IEEE 802.11 networks have two basic functions on the MAC layer: *point coordination function* (PCF) and *distributed coordination function* (DCF). PCF is intended to support real-time services by polling mobile terminals in its service area. DCF is created for best-effort traffic by using the CSMA/CA protocol. In the DCF mode, a terminal must sense the medium before sending a packet. The sensing time must be long enough to avoid collision between different mobile terminals, and this time is referred to as *distributed interframe space* (DIFS). If a mobile terminal detects a signal, it backs off a

random time interval within a specified *contention window* (CW). The 802.11 standard specifies alternation between PCF and DCF intervals, although PCF may be not supported by some wireless card interfaces. Support of both PCF and DCF may lead to inefficient usage of wireless resource. Therefore, some authors [13] propose an extension of DCF to provide service differentiation. One way to accomplish such a task is to create a *DiffServ-enabled MAC*, where packets are differentiated by DS field in the IP packet's header. Specifying different CW sizes for different services provides support to different classes in this algorithm. Packets with a smaller CW value are more likely to be transmitted first; that is, high-class service can get better service than lower-class service. To provide absolute QoS guarantees, one needs an accurate estimation of traffic parameters in the cell. For such purposes, one may find it suitable to use a *virtual MAC* (VMAC) that simulates real MAC behavior and thus provides, in advance, traffic information needed for admission control.

Currently, there are efforts to provide higher QoS support through an extension to the IEEE 802.11 standard called IEEE 802.11e [14]. With the aim to provide service differentiation, a new access mechanism is selected called *enhanced DCF* (EDCF). EDCF combines two differentiation techniques. First, the contention window can be set differently for different priority classes, similar to the approach presented above. For further differentiation, different interframe space can be used for different classes [instead of DIFS, we will have *arbitration interframe space* (AIFS)]. In the latter case higher-priority classes will have smaller AIFS.

Service Differentiation in 3G CDMA-Based Mobile Networks

Several 3G mobile standards are CDMA-based, such as UMTS and cdma2000. Therefore, we consider an example of service differentiation in a CDMA network. In such networks, resource allocation to users is mainly controlled by SIR and spreading control. One approach [15] is to use adaptive power control based on fixed target SIR, in conjunction with variable spreading control to adjust bandwidth offered to a user in a particular frame. In such an environment, class-based scheduling can be provided by introducing additional parameter *elasticity* (besides the bandwidth requirements), which refers to how the rate will decrease in a period of congestion. In the uplink, the mobiles can reduce its rate upon congestion according to the elasticity. In the downlink, the limiting factors are path loss and total base station transmitted power to users. Therefore, in the downlink case elasticity must be considered together with the path loss the corresponding mobile terminal sees from base station. To provide multiclass communication from a single mobile terminal, each class should be assigned a different code. Also, base stations control the scheduling in the wireless channel. While downlink scheduling is trivial because the base station has a complete

knowledge about the traffic, uplink scheduling requires signaling information from mobile terminals to base stations.

The above approach in CDMA mobile networks can be extended by allocation of resources proportionally to weights, thus leading to fair allocation [16]. With such an approach, naturally one should take into account the difference in resource scarcity for the uplink and downlink. First, let us consider service differentiation in the uplink. We assume that each mobile user has associated weight that corresponds to its service class. In 3G UMTS's WCDMA, transmission occurs in fixed-frame sizes with minimal duration of 10 ms, and the rate may change only between frames (it is fixed within a single frame). Let us denote with $r_i = R_i \nu_i$ the transmission rate of the user i (R_i is the bit rate, and ν_i is the activity factor), and with $SIR_i = (E_b/N_0)_i$ the signal-to-interference ratio of user i. If we assume a large number of users in a cell (e.g., low-rate service), then the assumption $(W/r_i SIR_i) >> 1$ is valid. In this case, using (7.86) we obtain

$$\sum_{i=1}^{N} r_i SIR_i = \frac{\eta_{UL}}{(1+i)} W = \eta_{UL}' W \tag{11.4}$$

where W is the chip rate (e.g., W = 3.84 Mcps for WCDMA) and η_{UL} is the uplink load factor. With the aim of achieving fair resource allocation, wireless channels should be allocated in proportional weights [16], as given by

$$r_i SIR_i = \frac{w_i}{\sum_j w_j} \eta_{UL}' W \tag{11.5}$$

Assuming that the user can potentially control both the transmission rate in the uplink and the SIR, we can use the above relation to calculate the needed SIR_i for fixed rate requirements r_i (e.g., CBR service), or to provide a given *frame error ratio* (FER) for user i (i.e., fixed SIR_i) by applying rate adaptation (i.e., by varying r_i).

In the downlink the limiting factors are the base station's total transmission power and multipath fading. Because of multipath fading, the received signal quality at mobile terminals will fluctuate. Therefore, it is convenient to use average power levels in the downlink and then calculate the transmission rate. The average power for user i can be written as

$$\overline{P_i} = \frac{w_i}{\sum_j w_j} \eta_{DL} P \tag{11.6}$$

where η_{DL} is the downlink load factor (Section 7.6.1.2), and P is the total transmission power of the base station. Because of the multipath, users at different

locations in the cell experience different path loss and interference. Therefore, one may find it suitable to use average values on these parameters with the aim of avoiding dependence of service differentiation upon the mobile's location. Then transmission rates in the downlink can be calculated by

$$r_i = \frac{w_i}{\sum_j w_j} \frac{W}{SIR_i \bar{I}\,\bar{L}} \eta_{DL} P \tag{11.7}$$

where $\bar{I}$ and $\bar{L}$ are average values of the interference and the path loss in the cell, respectively.

11.4 Wireless Class-Based Flexible Queuing

The *wireless class-based flexible queuing* (WCBFQ) algorithm is a scheduling scheme created to support multiple traffic classes in wireless IP networks [i.e., real-time flows, CBR, VBR, as well as best-effort traffic (Web, FTP, and so forth)]. It should be applied at wireless access points.

Our tendency in creating this scheduling algorithm was to take into consideration the high BER in the wireless environment. BER is flow-specific due to the different location of single users and the different states of the air interface. Location-dependent errors are more likely to be expected than uniformly distributed errors over the whole bandwidth of the cell. In such conditions we have to satisfy guaranteed services when they are experiencing high error rate by increasing their share of the bandwidth. On the other hand, it is not desirable to allow flows in the error state to decrease significantly the performances of the entire wireless link. The WCBFQ scheduler model is shown in Figure 11.1.

11.4.1 Class Differentiation

The base station assigns the traffic flow a channel according to a hierarchy of priorities. The first differentiation of the traffic is into two main classes: class-A with bandwidth guarantees, and class-B for best-effort traffic. A class selector (Figure 11.1) separates arriving packets into different queues for every class. According to the discussion in Chapter 5, class-A is divided into CBR subclass, VBR subclass, and BEmin. CBR subclass should be used for real-time applications that have strict demands on network delay, such as voice over IP. This is high-priority class. The flows belonging to the CBR subclass will be first served until the buffer for this class is emptied. VBR is intended for real-time applications with time-varying rate, such as video streams. Because video usually has

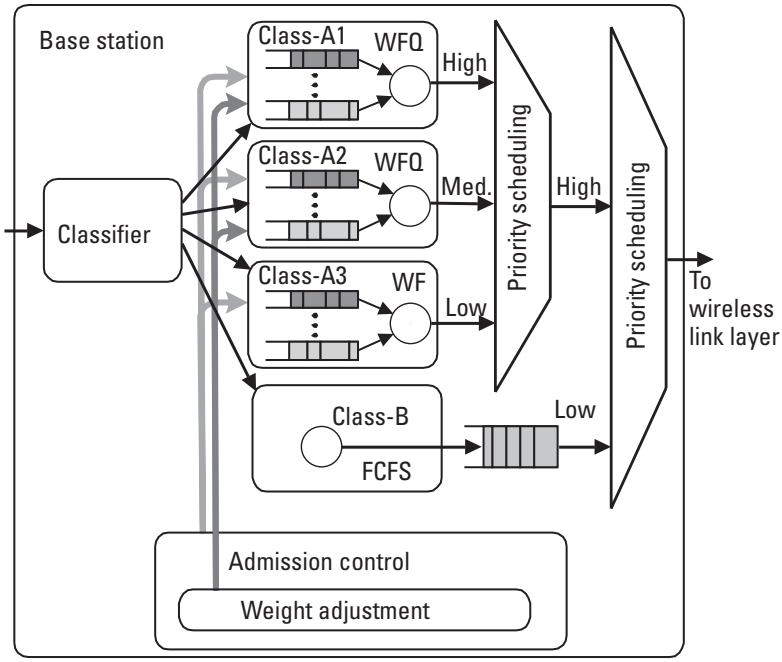


Figure 11.1 Model of WCBFQ scheduler.

higher bandwidth demands than voice, it is given lower priority to this subclass compared with CBR. That is a consequence of the characteristics of video information, where information is referred to a limited number of video frames per second that are less deterministic than traffic such as voice (Chapter 5). Also, video flows require many times greater bandwidth than voice-oriented services. Video communication is usually one-way (e.g., video streaming), although it can be bidirectional (e.g., video telephony). In the latter case one may decide to apply CBR subclass instead of VBR. Due to such characteristics of VBR sources, we give lower priority to VBR subclass than to CBR. But, to avoid monopolization of the bandwidth by the CBR flows, we should limit the maximal capacity that can be allocated to them. This can be accomplished by an admission control mechanism. The last subclass of class-A is dedicated to users who want to have some QoS guarantees (they should pay more for their services than class-B users).

Let us use B for a bandwidth of the wireless link. The weights assigned to flows in a subclass j are w_{ji}, $i = 1, \ldots, N$, where N is the number of active flows

on the link. We define the throughput of each flow, normalized on the link bandwidth admitted for that subclass (RT: relative throughput):

$$RT_{ji}(t) = \frac{w_{ji}(t)}{\sum_{j=1}^{N_C}\sum_{i=1}^{N_f} w_{ji}(t)} \tag{11.8}$$

When the wireless path is error-free, the flow should get bandwidth share $b_{ji}(t)$:

$$b_{ji}(t) = RT_{ji}(t) * B = \frac{w_{ji}(t)}{\sum_{j=1}^{N_C}\sum_{i=1}^{N_f} w_{ji}(t)} B \tag{11.9}$$

The above relations refer to a situation when we are using absolute weights for all flows from all classes over the entire bandwidth of the wireless link. However, we may also apply weights relatively within each class that uses fair-like queuing.

We assume that the base station has knowledge of the channel state (e.g., by monitoring or prediction), as well as which mobiles attend to send uplink data. Since location-dependent error is a specific of the wireless interface, [3] suggests queuing the packets during the error period. But this is not appropriate for traffic with strict delay requirements, such as voice traffic. In our scheduler there is no queuing of the packets during error state, but also there is no compensation on errors for real-time flows because it is redundant.

Maximum delay for a CBR flow i without errors is denoted as $D_{CBR}^{\max}$, and it is given by

$$D_{CBR,i}^{\max} = \frac{L_{p,\max}}{B} + \frac{L_{p,\max}}{B}\frac{\sum_{j\in F_{CBR}}^{N_{CBR}} w_j}{w_i} + \Delta t_p \tag{11.10}$$

where N_{CBR} is number of CBR flows, maximum packet length is $L_{p,\max}$, and F_{CBR} is the set of all CBR flows. The last term Δt_p includes all delays due to processing, such as framing, segmentation, encoding, spreading, rate matching, and multiplexing. Usually, however, queuing delay in packet networks is higher than processing delay in order of magnitude, due to the statistical multiplexing of data.

Because the CBR subclass has the highest priority, CBR packets use all of link bandwidth B until they are all served. The maximum delay corresponds to the situation when the packet of a flow is the last on the list of the active CBR

flows. Total buffer space for CBR flows can be calculated using (11.11), where L_{CBR} is the maximum length of CBR packets and N_{CBR} is the number of CBR flows:

$$Q_{CBR} = L_{CBR} N_{CBR} \tag{11.11}$$

When all CBR queues are emptied, the scheduler will start serving VBR flows. The bandwidth that is left for VBR flows can be calculated by (11.12).

$$B_{VBR} = B - \sum_{i \in CBR} b_i \tag{11.12}$$

Considering (11.11), the buffer requirement for the flows of the VBR subclass of class-A is calculated as follows:

$$Q_{VBR} = q_{burst} + \frac{L_{p\max} N_{CBR}}{B} r_{VBR} \tag{11.13}$$

In the calculation of buffer space for VBR flows, the bursty nature of the VBR traffic (e.g., video) should be taken into account. The additional length of the VBR queue, which is aimed to capture burstiness of VBR flow, is denoted as q_{burst}. If maximum burst duration is t_{burst} with peak rate of the flow r_{peak} and admitted rate r_{VBR}, then it can be calculated using

$$q_{burst} = t_{burst} \left(r_{peak} - r_{VBR} \right) \tag{11.14}$$

Because VBR flows are serviced with a lower priority than CBR traffic, the additional delay due to higher-level traffic must be considered. The worst-case delay of VBR flow includes delay due to serving higher-level A1 packets, and delay for serving packets from other VBR flows. Using the effective throughput of VBR traffic, we may calculate the worst-case delay by the following equation:

$$D_{VBR,i}^{\max} = \frac{N_{CBR} L_{p,\max}}{B_{VBR}} + \frac{L_{p,\max}}{B_{VBR}} + \frac{\sum_{j \in F_{VBR}} w_j}{w_i} \frac{L_{p,\max}}{B_{VBR}} + \Delta t_p \tag{11.15}$$

The third subclass, called *best-effort with minimum guarantees* (BEmin), is targeted to nonreal-time traffic with minimal QoS guarantees. Therefore, we use a fair scheduling mechanism for this subclass, such as WFQ or WRR, together with admission control to provide the minimal QoS support. These flows are serviced with lowest priority from all subclasses within class-A.

Therefore, the packets of this subclass have to wait until CBR and VBR queues are drained out. Also, a packet might wait for all other BEmin flows to be served. Therefore, the A3 traffic subclass requires the following buffer space:

$$Q_{BE\ \min} = \frac{L_{p\max} N_{CBR}}{B} r_{BE\ \min} + \frac{\sum_{i \in F_{VBR}} Q_{VBR_i}}{\sum_{j \in F_{VBR}} r_{VBR_j}} r_{BE\ \min} \tag{11.16}$$

Each of the classes, class-A and class-B, are scheduled in different queues. Modification of the WFQ is applied for class-A traffic. Class-B flows get the remaining part of the bandwidth after class-A flows are serviced. Most class-B flows are based on the TCP protocol. TCP adjusts to the available bandwidth by managing its congestion window, and in longer time intervals TCP flows get equal bandwidth shares of the link. However, some application may start several simultaneous TCP connections to get a larger share of the bandwidth. Hence, TCP gets as it can, but best-effort can suffer from some other aggressive flows that are established between peers based on some other protocol or agent module. Therefore, if one needs minimal QoS guarantees, then the A3 subclass for best-effort traffic should be used. Otherwise, the option is class-B, which does not offer any QoS guarantees. All class-B packets are serviced according to the FCFS principle.

11.4.2 Scheduling in an Error State

Now, we will introduce the error state in the wireless link. Different policies should be applied on different classes while the channel is in error state. We assume that error rate is measured by MAC level or is predicted, so error rate per flow is a time-dependent function $E_{ji}(t)$, for every flow i within a class j. This measurement assumes fast link-level acknowledgment.

According to the WCBFQ algorithm, when a CBR flow is experiencing errors, its weight will be increased in order to get its effective share of the bandwidth as it is in error-free state. The weight adjustment should be done only during noticeable flow error rate. To avoid frequent flip-flops to and out of error mode, we introduce hysteresis thresholds: *high error threshold* (HET) and *low error threshold* (LET), which are in the range from 0 to 1 (e.g., 1 corresponds to 100% error rate, and 0 corresponds to error-free state), and always $HET > LET$. Only when $E_{ji}(t) > HET$ will the flow transit from error-free to error mode in the scheduler. The flow will return to error-free mode after being in the error mode when $E_{ji}(t) < LET$. This is done to avoid the ping-pong effect and unnecessary computation. After crossing the *HET*, the weight of the erroneous CBR flow is adjusted according to the following relation:

$$w_i^{eff}(t)\left[1 - E_i(t)\right] = w_i, i \in F_{CBR} \tag{11.17}$$

where $w_i^{eff}(t)$ is the adjusted effective weight of the flow i when it is in error mode with error ratio $E_i(t)<1$. Weight adjustment of a CBR flow while it is in error state is possible only when the following condition is satisfied:

$$B \geq \sum_{i \in F_{CBR}} b_i + \sum_{j \in F_{VBR}} b_j + \sum_{k \in F_{BE\,\min}} b_k \tag{11.18}$$

In the above relation are given guaranteed bandwidth shares of class-A flows: CBR, VBR, and BEmin, in error-free state.

To compensate for the increase in weight of a CBR flow, first, the bandwidth share will be taken from the class-B flows. If it is not enough, it will be taken from BEmin flows—but BEmin minimum bandwidth guarantees should remain. If it is not enough, the next step is to decrease the weights of the VBR flows, but they should have at any time the admitted rate at the call admission phase. If it is not enough (e.g., the network is highly loaded), then the scheduler will not be able to adjust entirely the weight of the CBR flow in error state.

Adjustment of weights causes degradation of the other flows by decreasing their throughputs. But when the error rate is high, the affected flow can significantly decrease throughput of the other flows especially if it occupies a larger amount of the bandwidth. To avoid such a situation, the increase of the $w_i^{eff}(t)$ should be less than a predefined limit $L_i w_i$, where $L_i>1$. For example, a typical value for voice service based on CBR traffic type will be $L_i = 2$, which corresponds to a 50% error ratio in the wireless channel. We distinguish two regions considering the error rate E_i: (1) $1/(1 - E_i)<\min\{L_i; 1 + B_{free}/(Bw_i)\}$, which we refer to as an *adjusting region* (or outcome region [1]); and (2) $1/(1 - E_i)\geq\min\{L_i; 1 + B_{free}/(Bw_i)\}$, which we refer to as an *effort region*. In the effort region we may be limited by the limit factor L_i for flow i or by the amount of nonreserved resources. According to the discussion above, the adjusted effective weight for a CBR erroneous flow will be

$$w_i^{eff} = \min\left(\frac{w_i}{1-E_i}; L_i w_i; w_i + \frac{B - B_{admitted}}{B}\right) \tag{11.19}$$

Using the adjusted weight, we obtain the following throughput in the adjusting region:

$$b_i^{eff} = \frac{w_i^{eff}(1-E_i)}{\sum_{j \in F_{CBR}} w_j} B = \frac{\frac{w_i}{1-E_i}(1-E_i)}{\sum_{j \in F_{CBR}} w_j} B = \frac{w_i}{\sum_{j \in F_{CBR}} w_j} B = b_i \tag{11.20}$$

The above relation shows that this algorithm adjusts the flow's throughput exactly to its value in error-free state. However, the limit-factor L_i is necessary to

limit the adjustment so that flows with high error rates cannot degrade the performance of the whole link.

In reality, the CBR class should be dedicated to voice over IP. Voice service demands lower bit rates, so each connection will usually occupy a small share of the bandwidth. For example, for a wireless link rate of 2 Mbps and a voice data rate in a cellular environment of 10 Kbps, each voice connection occupies less than 1% of the total link bandwidth.

When a VBR flow is in error state, WCBFQ reacts in the same manner as for CBR, but coefficients are adjusted with lower limit-factors than coefficient adjustment of CBR flows because of higher data rates. But VBR traffic is served with lower priority than CBR. The guaranteed data rates are agreed at the admission control (Chapter 8). For example, at a new CBR-call request, admission control should consider initially agreed throughputs of VBR flows (i.e., it should not consider the modified VBR weights).

When BEmin flows are in error state, WCBFQ does not react with weight adjustment because BEmin subclass does not request real-time services and does not have strict QoS guarantees per flow (there are only minimum guarantees on the delay of the aggregate traffic). Fair scheduling of flows within a subclass of class-A is provided by the WFQ mechanism.

BEmin flows suffer when a CBR flow or a VBR flow is in error mode. These flows are also serviced by WFQ within the subclass-A3 in an error-free environment, or its approximations such as WRR. For BEmin flows (i.e., subclass-A3), WCBFQ uses some of the wireless fair algorithms described in Section 11.3. The choice of the algorithm is a matter of the design approach. In other words, the designer of the algorithm should make the choice considering the importance of the following issues: fairness, complexity, and costs. So, the simplest solution for scheduling A3 flows will be CSDPS, but considering the fairness one may choose to apply WFS [7].

We may calculate the A3 flow's throughput by using the two-state Markov error model (Section 6.5). The Markov model is used to describe the error-free and error states of a wireless flow. The transition matrix of the Markov model is given by

$$\begin{aligned} P &= \begin{bmatrix} P(0/0) & P(1/0) \\ P(0/1) & P(1/1) \end{bmatrix} \\ &= \begin{bmatrix} 1-\lambda_{1/0} & \lambda_{1/0} \\ \lambda_{0/1} & 1-\lambda_{0/1} \end{bmatrix} \end{aligned} \tag{11.21}$$

where $\lambda_{1/0}$ is state-transition probability from error-free to error state, while $\lambda_{0/1}$ is state transition probability in the reverse direction. Assuming steady state, we

can calculate error and error-free state probabilities using the Markov model, as given by (11.22) and (11.23), respectively:

$$\pi_1 = \frac{\lambda_{1/0}}{\lambda_{1/0} + \lambda_{0/1}} \tag{11.22}$$

$$\pi_0 = \frac{\lambda_{0/1}}{\lambda_{1/0} + \lambda_{0/1}} \tag{11.23}$$

If we apply a compensation method, then we can provide fairness among the A3 flows. The simplest wireless fair queuing algorithm is CSDPS, which provides WFQ or WRR scheduling with skipping of flows that are in error-state in each round. For the case of CSPDS, assuming that error periods of different flows are not overlapping, and using the Markov model for wireless channel state with average error rate E_i in the error state, the effective throughput of the flow i can be calculated by

$$\begin{aligned} b_i^{eff} &= \pi_0 b_i + \pi_1 b_i (1 - E_i) \\ &+ \pi_1 \sum_{j \neq i,\, j \in F_{BE}} \left(b_j E_j \frac{w_i^{BE\,\min}}{\sum_{k \neq j} w_k^{BE\,\min}} \right) \end{aligned} \tag{11.24}$$

where $w_i^{BE\min}$ are weight coefficients of the WFQ (or WRR) applied within BEmin traffic class. Because this traffic class is targeted to best-effort traffic without strict QoS requirements (only minimal considering the minimum rate), one may find as the most appropriate design solution to apply equal sharing of the BEmin bandwidth by all flows within this class [i.e., $w_i^{BE\,\min} = B_{BE\,\min}/(N_{BE\,\min} \cdot B)$, where $N_{BE\,\min}$ is number of ongoing subclass-A3 flows in the cell, and $B_{BE\,\min}$ is the bandwidth for servicing these flows]. However, minimal QoS guarantees should be provided by the admission control (a design approach is given in Chapter 8), because BEmin belongs to class-A. Then, for error-free wireless link for BEmin flows, we can calculate available bandwidth per flow using the following relation:

$$b_i^{BE\,\min} = B_{BE\,\min} / N_{BE} = b_{BE\,\min} \tag{11.25}$$

In the above relation $B_{BE\,\min}$ is the bandwidth left for A3 flows after servicing the higher-level traffic classes, which have admitted data rates and allowed adjustment of their weights in the case of errors in the wireless channel, that is:

$$B_{BE\,\min} = B - \sum_{i \in F_{CBR}} b_i - \sum_{j \in F_{VBR}} b_j - \sum_{k \in F_{adjustments}} b_k \tag{11.26}$$

If all flows experience the same average error rate in the long term (i.e., $E_i = E$ for all i in the cell), then from (11.24) the effective bandwidth for all BEmin flows will be equal to the bandwidth as if all flows were in the error-free state (i.e., $b_i^{eff} = b_i$ for every flow i). So, in such cases, even the CSDPS can provide long-term fairness between BEmin flows. If we want to provide short-term fairness of the flows, we may use the WFS algorithm instead of CSDPS, but with increased complexity of the system and additional delay due to the later compensation.

Finally, class-B traffic has no QoS guarantees. Because it does not operate within the constraints of fair queuing, no weights have to be calculated. Hence, a simple FCFS scheduler should naturally serve class-B packets.

Priorities of different traffic classes in WCBFQ, as well as the queuing discipline for each class, are summarized in Table 11.1.

11.4.3 Characteristics of WCBFQ

The choice of the limits for weight adjustment of CBR flows is left to network administrators. Typical values of the limits L_i should be 2 or higher for flows that occupy the smaller part of the bandwidth, and less for flows that highly utilize the link resources. Of course, in every case, guaranteed services that are error-free should get the minimum guaranteed data rate.

A CBR flow carrying voice will not cause high degradation of the wireless link performance, but this is not the case with video content. Video streams usually occupy a larger amount of the bandwidth and they may produce higher performance oscillation in the wireless link. For best-effort flows we may apply any of the existing schedulers created for a wireless LAN environment.

Table 11.1
Priorities and Queuing Disciplines in WCBFQ Algorithm

Traffic Class	Priority	Subclass	Priority	Queuing Discipline
A	High	A1	High	Flexible WFQ
		A2	Medium	Flexible WFQ
		A3	Low	WFS, CSDPS, WPS
B	Low	—	—	FCFS

When does a flow enter an error state? The scheduler at the base station with TDD access technology services packets in both the uplink and downlink. In a multiple access technology, different schedulers may be applied in different directions. The flow transits into an error state if the average number of time slots or frames with detected errors divided by the total number of allocated time slot/frames to that flow is over the predefined error threshold. For example, if $HET = 0.2$, and if errors are detected in two or more time slots out of 10 consecutive slots allocated to that flow, then the flow transits into an error state and the scheduler applies modification of the weights for A1 and A2 flows. In this way we overcome the problem that arises from the scheduling algorithm created for wireless networks with best-effort traffic where only the compensation method between leading and lagging flows is used in different implementations [10]. Compensation methods refer only to the location-dependence of bit errors in the wireless link, but they do not capture the requirements from real-time flows. Wireless errors usually occur in bursts, because of the inertia of signal propagation in a cellular network, as well as the inertia of users' movement in time intervals comparable to the time needed for processing of an individual IP packet (e.g., several milliseconds). By using the WCBFQ algorithm, we address both issues: the location-dependence of wireless bit errors and the multiclass environment.

11.5 Simulation Analysis

For simulation analysis of the WCBFQ algorithm we performed several experiments. In all simulations we used wireless link bandwidth of 2 Mbps. Each active user competes for a transmission over the wireless link. Simulations are performed using real-time flows (video traces), CBR flows, and nonreal-time FTP traffic. For the simplicity of the analyses, we use average packet length of 1,000 bytes.

We performed three experiments to evaluate the WCBFQ algorithm. The first simulates multiplexed traffic consisting of a CBR flow that occupies 10% of the link bandwidth, a VBR video stream with admitted rate of 1.4 Mbps, and an FTP flow that gets the rest of the bandwidth capacity (Figure 11.2). Error rate is introduced in the CBR flow only, in the interval between 20 and 30 seconds of the simulation time. The simulation is run for error rates of 0%, 25%, and 50%. The throughputs of the flows for 50% error rate on the CBR flow are shown in Figure 11.3. WCBFQ reacts by increasing the bandwidth share of the affected CBR flow and keeping constant its throughput because there is enough not-admitted bandwidth that allows complete modification of the weight of the CBR flow during the error state. If we make a comparison with the error-free state for all flows given in Figure 11.2, it is

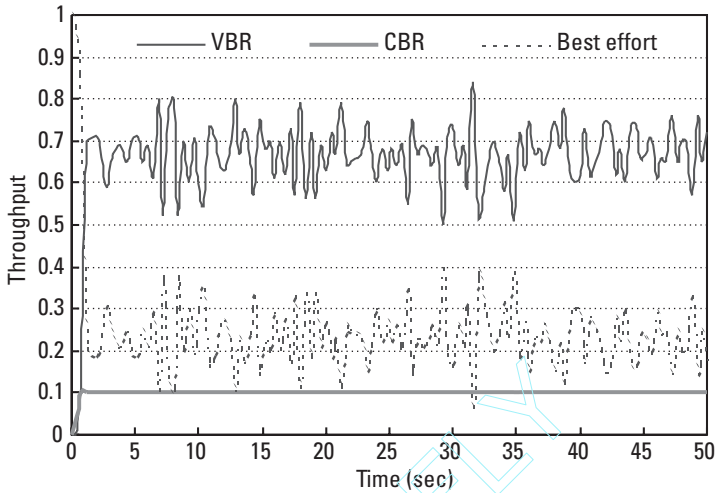


Figure 11.2 Throughputs when all flows are in error-free state.

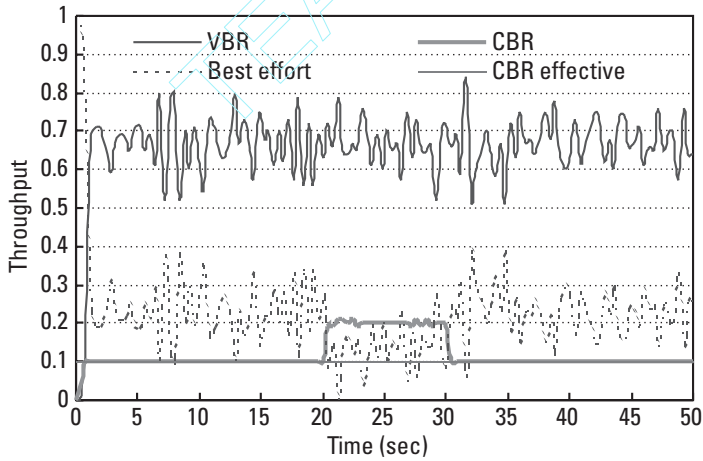


Figure 11.3 Throughputs of flows when CBR is affected by 50% error rate in predefined time period.

noticeable that the FTP flow suffers the most, while VBR has almost identical throughput except on the peak rates. If we analyze the delay of the VBR packet (Figure 11.4), an increase in the packet delay while the CBR flow is in error state it is easily noticed. This can be explained by the priority of CBR over VBR; so by increasing the bandwidth share of the CBR flow, VBR packets have to wait longer in the queue (i.e., until CBR packets are all served). This

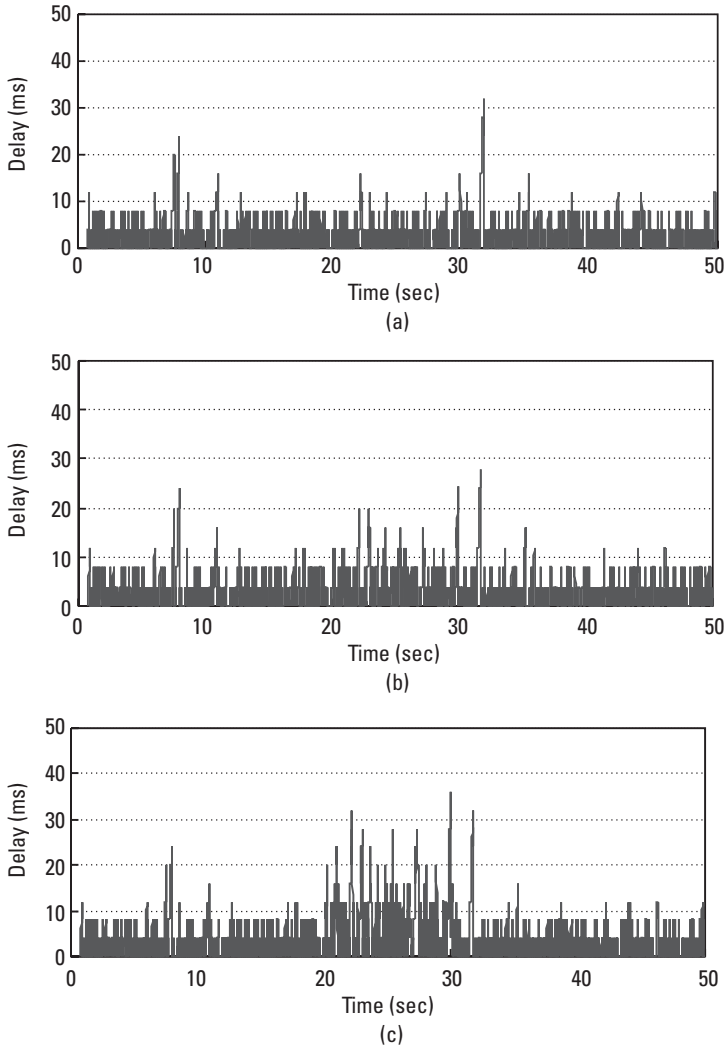


Figure 11.4 Delay of the VBR packets for different error ratio on CBR flow: (a) 0% error rate; (b) 25% error rate on CBR flow; and (c) 50% error rate on CBR flow.

discussion is confirmed by Figure 11.5, where probability distribution functions of VBR packet delay for different error ratio on the CBR flow are given.

In the second experiment we used one CBR flow and two FTP flows, as shown in Figure 11.6. The error rate is applied on CBR in the same time interval as in the first experiment. In error-free state every FTP flow has half of the remaining bandwidth, or 45%, and CBR occupies 10% of bandwidth. After transiting to error state, WCBFQ performs weight adjustment, raising the CBR

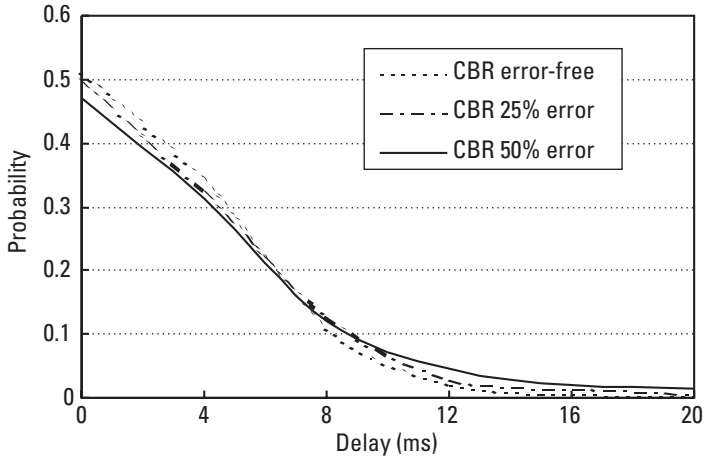


Figure 11.5 Probability distribution function of packet delay for different error rates on CBR flow.

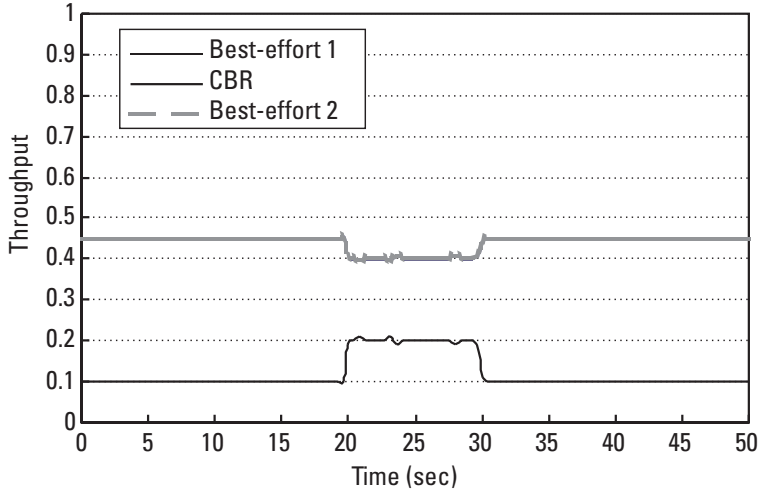


Figure 11.6 Throughputs of the flows when CBR flow is experiencing 50% error ratio in a predefined time period.

share of bandwidth up to 20%, while FTP flows are equally decreased down to 40%.

In the last experiment we used only FTP flows from A3-subclass, as shown in Figure 11.7. We show a time sequence of the available throughputs for the two FTP flows where error periods of both flows alternate. This situation should be considered only as an example in which error periods of the flows are not

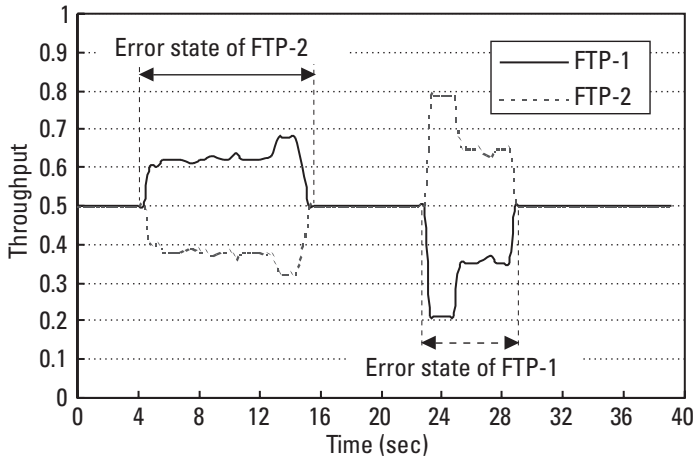


Figure 11.7 Available throughputs of two FTP flows from A3-subclass with applied WCBFQ when wireless scheduling of A3 flows is done by CSDPS.

overlapping. According to the Markov error model, over a long time scale each of the flows within a cell has an equal probability of entering/leaving the error state. In this example the simplest wireless fair scheduling is used—that is, CSDPS. The bandwidth share that is released by the flow in error state is shared among all other BEmin flows. Because there are only two FTP flows in this experiment, all the released bandwidth from the erroneous flow is taken by the other FTP flow, which is error-free. However, in a real network scenario we may expect many users within a single cell; thus, the probability that all users are in error state will be close to zero. We consider only the available bandwidth for each of the flows. However, the achievable data rate of the flow is dependent upon the transport protocol (e.g., TCP) and how it adapts the data rate to the bandwidth fluctuations.

11.6 Discussion

In this chapter we proposed a scheduling algorithm for wireless IP networks [17–19]. The main motivation for creation of such an algorithm was efficient scheduling under location-dependent and bursty wireless bit errors in a multiclass environment, where traffic is defined according to the classifications made in Chapter 5.

From the aspect of packet scheduling in a wireless environment, most of the algorithms consider a single traffic class (i.e., best-effort traffic) and use the compensation method—that is, giving the bandwidth (e.g., time slots and frames) to other flows during the error state and compensation of the bandwidth

during the error-free periods. The compensation method fits well for scheduling under location-dependent errors, but it does not consider the inertia of the error state as well as requirements from the real-time flows.

With the proposed WCBFQ algorithm we capture the behavior of different traffic classes/subclasses and their QoS requirements. Thus, CBR flows (i.e., subclass-A1), which are mainly targeted to voice over IP, do not require compensation of the lost service time or bandwidth, because real-time communication cannot use and does not need an additional bandwidth during the error-free period. Therefore, WCBFQ compensates CBR flows in real-time by maintaining unchanged effective throughput, of course, when there is enough bandwidth that is not dedicated to other class-A flows either CBR, VBR, or BEmin. On the other hand, VBR services, which also may be real-time traffic, have different traffic demands (e.g., video VBR services). They have guaranteed average bit rate agreed at the call admission phase that cannot be degraded by other flows. In a case of error state of a VBR flow, WCBFQ modifies the weight of the flow only when there is enough nonreserved bandwidth in the cell. However, adjustments of CBR flows' weights have higher priority over adjustments of VBR flows' weights, because VBR traffic is bursty in nature and thus should be flexible enough to adapt to certain bandwidth fluctuations within the range between its minimum guaranteed and peak data rate. WCBFQ does not modify weights of BEmin flows in error state because this subclass is targeted to nonreal-time services and provides only minimum service guarantees, which are more related to the aggregated BEmin traffic than to individual flows. This subclass has the lowest priority within class-A. It is the designer's choice whether to apply short-term wireless fair algorithm for BEmin flows, such as WPS, or to use a simpler solution, such as CSDPS.

Class-B traffic has lower priority than class-A, and therefore, class-B uses the remaining part of the bandwidth after servicing class-A flows. We propose a simple FCFS scheduler since class-B does not provide QoS guarantees.

Finally, we may conclude that WCBFQ provides flexible support to different traffic classes in a wireless IP environment considering the requirements for the QoS and real-time service under the influence of location-dependent and bursty bit errors in the wireless link.

References

[1] Eckhardi, D. A., and P. Steenkiste, "Effort-Limited Fair (ELF) Scheduling for Wireless Networks," *INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[2] Jiang, Z., L. F. Chang, and N. K. Shankaranarayanan, "Providing Multiple Service Classes for Bursty Data Traffic in Cellular Network," *INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[3] Moorman, J., and J. Lockwood, "Multiclass Priority Fair Queuing for Hybrid Wired/Wireless Quality of Service Support," *IEEE Mobicom/WowMom*, Seattle, WA, August 1999.

[4] Gomez, J., A. T. Campbell, and H. Morikawa, "A System Approach to Prediction, Compensation and Adaptation in Wireless Networks," *First ACM/IEEE International Workshop on Wireless and Mobile Multimedia (WoWMo'98)*, Dallas, TX, October 1998.

[5] Eugene, T. S., I. Stoica, and H. Zhang, "Packet Fair Queuing Algorithms for Wireless Networks with Location-Dependent Errors," *INFOCOM 1998*.

[6] Lu, S., V. Bharghavan, and R. Srikant, "Fair Scheduling in Wireless Packet Networks," *ACM Sigcomm '97*, Cannes, France, September 1997.

[7] Nandagopal, T., S. Lu, and V. Bharghavan, "A Unified Architecture for the Design and Evaluation of Wireless Fair Queueing Algorithms," *ACM/Baltzer Wireless Networks Journal*, Vol. 8, No. 2–3, January 2002.

[8] Lu, S., T. Nandagopal, and V. Bharghavan, "Design and Analysis of an Algorithm for Fair Service in Error-Prone Wireless Channels," *ACM/Baltzer Wireless Networks Journal*, Vol. 6, No. 4, 2000.

[9] Bharghavan, V., S. Lu, and T. Nandagopal, "Fair Queueing in Wireless Networks: Issues and Approaches," *IEEE Personal Communications Magazine*, Vol. 6, No. 1, February 1999.

[10] Nandagopal, T., S. Lu, and V. Bharghavan, "A Unified Architecture for the Design and Evaluation of Wireless Fair Queueing Algorithms," *ACM Mobicom '99*, Seattle, WA, August 1999.

[11] Ramanathan, P., and P. Agrawal, "Adapting Packet Fair Queuing Algorithms to Wireless Networks," *ACM Mobicom'98*, Dallas, TX, October 1998.

[12] Lu, S., T. Nandagopal, and V. Bharghavan, "A Wireless Fair Service Algorithm for Packet Cellular Networks," *ACM Mobicom'98*, Dallas, TX, October 1998.

[13] Veres, A., A. T. Campbell, and M. Barry, "Supporting Service Differentation in Wireless Packet Networks Using Distributed Control," *IEEE Journal on Selected Areas in Communication*, Vol. 19, No. 10, October 2001.

[14] Lindgren, A., A. Almquist, and O. Schelen, "Evaluation of Quality of Service Schemes for IEEE 802.11 Wireless LANs," *IEEE Conference on Local Computer Networks (LCN 2001)*, November 2001.

[15] Guo, Y., and H. Chaskar, "Class-Based Quality of Service over Air Interface in 4G Mobile Networks," *IEEE Communications Magazine*, Vol. 40, No. 3, March 2002.

[16] Siris, V. A., B. Briscoe, and D. Songhurst, "Service Differentiation in Third Generation Mobile Networks," *3rd International Workshop on Quality on Future Internet Services (QofIS'02)*, Zurich, Switzerland, October 16–18, 2002.

[17] Janevski, T., and B. Spasenovski, "QoS Provisioning for Wireless IP Networks with Multiple Classes through Flexible Fair Queuing," *GLOBECOM 2000*, San Francisco, CA, November 27–December 1, 2000.

[18] Janevski, T., and B. Spasenovski, "Flexible Fair Scheduling for Wireless IP Networks with Heterogeneous Traffic," *Personal and Indoor Mobile Radio Communications PIMRC 2000*, London, England, September 18–22, 2000.

[19] Janevski, T., and B. Spasenovski, "Flexible Fair Queuing for Wireless Packet Networks," *Wireless 2000 Conference*, Calgary, Alberta, Canada, July 10–12, 2000.

12

Conclusions

In this book we addressed wireless IP networks, which we defined as all-IP networks end-to-end. The evolution of both mobile networks and the Internet has come to the point of their convergence. Future generation mobile systems are expected to include heterogeneous wireless access networks (3G, WLAN, WPAN) with multiple traffic classes. Such a scenario requires traffic classification—and hence appropriate dimensioning and admission control—efficient mobility, and location management. However, there are several key characteristics of wireless networks and IP networks that complicate matters. On the wireless networks side, the key characteristics are:

- Mobility of the users;
- Bit errors in the wireless channels;
- Scarce wireless resources.

On the IP network side, the key problems are:

- Lack of QoS support;
- Lack of data synchronization.

In this book we addressed the above issues in wireless IP networks considering the existing approaches, as well as giving design proposals for each of them. The following section provides a summary of the book's content.

Highlights

Existing implemented and proposed mobile systems are described in Chapter 2, starting from 2G (e.g., GSM), via 2G+ (e.g., GPRS), towards 3G mobile systems (e.g., UMTS and cdma2000). Chapter 3 provides information on IP protocols and existing solutions and problems for the introduction of mobility and QoS support to the Internet.

Chapter 4 gives the traditional teletraffic theory for wired and wireless networks, which is based on the famous Erlang loss formula for dimensioning circuit-switched networks. We extended the traffic scenario from single to multiple traffic types, which resulted in the multidimensional Erlang loss formula. Furthermore, we introduced the basis for traffic modeling and analysis in mobile environments and provided fundamental principles for the design of telecommunications networks.

In order to be able to provide QoS support in wireless IP networks, one first needs to classify IP traffic. Most of the IP traffic is WWW-based. Statistical analyses of TCP, WWW, and VBR video traces from real measurements are conducted in Chapter 5. The analyses showed that TCP and WWW traffic, as well as VBR video, are self-similar. Based on the analysis and the discussion, we proposed classification of IP traffic into two main classes: class-A for traffic with QoS guarantees, and class-B for traffic without guarantees (i.e., best-effort traffic). Class-A is further divided into three subclasses: A1 for CBR flows with highest QoS requirements, A2 for VBR flows with strict QoS requirements, and A3 for best-effort traffic with minimal QoS guarantees. These classes and subclasses can be mapped onto the four traffic classes defined by IMT-2000: conversational, streaming, interactive, and background class; but are not limited to them.

We proposed architecture for wireless IP networks with multiple classes in Chapter 6. Also, we proposed an integrated simulation architecture based on two-level analysis: call-level and packet-level. We defined a conceptual model of a network node, as well as traffic models for real-time (e.g., Poisson model) and nonreal-time traffic (e.g., Pareto model). The characteristics of mobile networks, such as mobility and wireless bit errors, are modeled by using Markov models. Chapter 6 also provides definitions of QoS parameters and capacity definitions in wireless IP networks.

The analytical framework for traffic analysis, as well as the dimensioning and optimization of wireless networks, is given in Chapter 7. We first created an analytical model for wireless cellular networks with a single traffic class, and then we extended the analysis to a multiclass environment where different classes have different call intensities, different bandwidth demands, and different call durations. We considered the optimization problem at given constraints on new call and handover blocking probabilities. Also, we performed analysis of local deterministic handover reservations in neighboring cells. From the analysis we concluded that utilization of the resources in a wireless multimedia network

decreases with the cell size and with an increase of the diversity between different traffic types. Finally, we provided traffic analysis and dimensioning of multiple access wireless networks such as CDMA.

Admission control needed for QoS support was the subject of Chapter 8. We proposed a HAC algorithm, which minimizes call blocking probabilities for A1 and A2 traffic subclasses under given constraints on call dropping probabilities of A1 and A2 and average delay of A3 packets. So, we integrated the requirements on call-level and packet-level from different traffic classes. Class-B traffic, however, is not considered by the admission control. In this chapter we also considered admission control in wireless CDMA networks.

In Chapter 9 we showed results of simulation QoS analysis that considered the mobility of users and wireless bit errors. The analyses were performed for different traffic types: CBR, VBR, and best effort. It was showed that smaller cells and higher mobility of users increase packet losses at hard handovers. The behavior of VBR flows at handovers was shown to be dependent upon the burstiness of the particular flow. TCP flows are degraded at handovers due to losses that result from the activation of the TCP congestion-avoidance mechanism. Higher correlation of the background traffic influences higher losses due to handovers. We made simulation analysis under the influence of location-dependent errors. It was shown that complete partitioning of the resources results in inefficient resource utilization. Also, the compensation method is not appropriate for real-time flows.

Micromobility management is crucial for wireless IP networks. In Chapter 10 we proposed a handover algorithm that avoids packet loss and duplicate packets at the handovers. It is based on establishing handover agents at the network nodes in the access domain, while Mobile IP handles the macromobility. We proposed uplink and downlink routing algorithms based on maintenance of routing-cache and routing-tables at the network nodes for the mobile node's location and semi-permanent routing information, respectively. Also, we defined a location management scheme for wireless IP networks.

Scheduling in wireless IP networks was discussed in Chapter 11. We proposed a scheduling algorithm for multiclass wireless IP networks called wireless class-based flexible queuing, which is flexible to different traffic demands from different traffic classes. It provides real-time compensation for A1 and A2 flows, where A1 traffic is given higher priority for compensation than A2. Because subclass-A3 is targeted to nonreal-time traffic, we proposed servicing these packets with a lower priority than subclasses A1 and A2, but minimal bandwidth guarantees are provided by some of the wireless fair algorithms (e.g., CSDPS and WFS), which are adaptations of WFQ to the wireless environment. However, class-B traffic is serviced using the FCFS scheduler because this traffic class is defined for traffic without any QoS guarantees (identical to today's best-effort traffic in the Internet).

The reader may use the material provided in this book for traffic dimensioning, analysis, and optimization, as well as for the design of wireless IP networks.

About the Author

Toni Janevski received a Dipl. Ing., M.Sc., and Ph.D. in electrical engineering from the University "Sv. Kiril i Metodij" of Skopje, Macedonia, in 1996, 1999, and 2001, respectively. Since 2001, he has been an assistant professor on the Faculty of Electrical Engineering at the University "Sv. Kiril i Metodij" of Skopje, where he teaches undergraduate courses in switching and traffic theory and in telecommunications networks, and graduate courses in wireless multimedia networks and in the design, modeling, and analysis of telecommunications networks. He is also an adjunct assistant professor at the Military Academy in Skopje, where he is teaching courses on multiplex systems. From 1996 to 1999, Dr. Janevski worked for the Macedonian GSM 900 mobile operator Mobimak on the cell planning and dimensioning of cellular networks, as well as on traffic and performance analyses. In 2001 he conducted research in optical communication at the IBM T. J. Watson Research Center in New York. His research interests include mobile and multimedia networking, IP technology, traffic theory, quality of service, dimensioning, and optimization of wireless networks. He is a member of the IEEE Communication Society.

Index