論文 / 著書情報
Article / Book Information

題目(和文)	セルラ移動通信システムにおけるトラヒック制御方式のモデル解析
Title(English)	Model analysis of traffic control schemes in cellular mobile communication systems
著者(和文)	吉野秀明
Author(English)	Hideaki Yoshino
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:乙第4018号, 授与年月日:2010年3月31日, 学位の種別:論文博士, 審査員:
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:乙第4018号, Conferred date:2010/3/31, Degree Type:Thesis doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Model Analysis of Traffic Control Schemes in Cellular Mobile Communication Systems

by

Hideaki YOSHINO

Submitted to the
Department of Mathematical and Computing Sciences,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Science

March, 2010

Acknowledgement

First and foremost, I wish to express my deepest gratitude to Professor Yukio Takahashi at the Tokyo Institute of Technology, for leading me toward the completion of this thesis. My work could never have been completed without his kind guidance and encouragement.

I am also deeply indebted to Professors Masakazu Kojima, Shigeru Mase, Naoto Miyoshi, Hidetoshi Shimodaira, at the Tokyo Institute of Technology, as well as Professor Toshihisa Ozawa at the Komazawa University, for giving me insightful comments and questions on my research. Their feedback enabled me to think about my research from various different perspectives, and to improve the overall quality of this thesis.

This research was conducted at NTT Laboratories. I am greatly indebted to former supervisors at NTT Laboratories, Professor Konosuke Kawashima at the Tokyo University of Agriculture and Technology, Professor Hisao Yamamoto at the Tokyo City University, Professor Akiya Inoue at the Chiba Institute of Technology, Professor Tsuyoshi Katayama at the Chubu Teletraffic Engineering Laboratory, Professor Yoshitaka Takahashi at the Waseda University, and the late Professor Hisayoshi Inamori at the Shonan Institute of Technology.

I would like to express my appreciation to Dr. Kou Miyake, Director of NTT Information Sharing Laboratory Group, and Dr. Kazuhiko Kushima, Director of NTT Service Integration Laboratories for giving me the opportunity to develop this thesis.

Special thanks are due to Dr.Hiroyuki Funakoshi, Dr. Takanori Hayashi, Dr. Ryoichi Kawahara, Dr. Arata Koike, Dr. Akira Takahashi, and to other colleagues at NTT Laboratories for their helpful discussions in my daily research.

This thesis is dedicated to my wife Yoshiyo, my daughter Makana, my son Keitoku, and my parents. Without their continuous support and encouragement, the completion of this work would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the past ten years, demand for mobile communication services, such as those of cellular telephones, has expanded at a very high rate. In Japan, the number of cellular telephone subscribers increased significantly and exceeded that of fixed analog telephone subscribers in March 2000, and reached 107 million in March 2009. The cellular telephone system became one of the most important social information infrastructures, and came to occupy the main part in the information and communication industry, which plays a leading role of an engine for economic growth.

Mobile communication systems cover technologies such as cellular communication systems, satellite communication systems, personal handy-phone systems (PHS), and wireless LAN. This thesis focuses primarily on the cellular mobile communication system, which consists of the cellular telephone system and PHS. Hereafter, we simply call it the *cellular system*.

In addition to the market growth of the cellular system, forms of communication are changing from voice to Internet, such as e-mail transmission and web browsing via mobile terminals. This change will lead to increased demand for mobile multimedia communication.

Overall, the massive increase in traffic volume and the introduction of multimedia applications have resulted in pressing technological problems, which require much more sophisticated traffic engineering solutions. In particular, traffic control that can handle variable and multimedia traffic is essential in operating cellular systems efficiently and robustly. Traffic control is defined as a mechanism that efficiently allocates limited system resources to communication requests satisfying quality of service (QoS) requirements under variable traffic conditions. A typical resource in cellular systems is radio frequency. Other examples of resources are switching nodes and communication links (paths) in transmission networks. Because communication requests occur stochastically, traffic control that takes into account probabilistic behaviors of the request-generation process is significant. In cellular systems, a bias of requests, caused by mobile terminal concentration at a specific time or place, occasionally occurs.

There are traffic control problems inherent in cellular systems compared to fixed-wire telephone communication systems.

One typical problem is resource competition in sharing the radio access channel. While a wired access line is dedicated to a single user in fixed-wire systems, the radio channel in cellular systems is shared by many users. Particularly, sharing radio channels according to a random access scheme causes a throughput decrease due to the collision of the signals. Therefore, traffic controls for random access, which make effective use of the channel by suppressing the throughput decrease, are important.

Another problem is congestion at the origination-side. In fixed networks, such as current telephone switching networks, an overload may occur at the termination-side when traffic concentrates on a specific number or region because of an event or disaster. In cellular systems, in addition to this congestion, overload may occasionally occur at the origination-side caused by mobile terminal concentration at a specific time or space.[1] For example, when the public transportation system stops somewhere during commuting hours, many mobile users in a specific cell of the radio access network may originate calls to their offices at the same time. Consequently, congestion control in radio access networks as well as core networks is an important issue in cellular systems. This issue is also essential from the viewpoint of effective use of the precious resource of radio frequencies.

Traffic control problems inherent in cellular systems exist not only in the radio access layer but also in the network control layer, which consists of a signaling network and databases. Mobility management is a typical problem in the network control layer. If the number of cellular users increases, the infrastructures of the network control layer may require changes in order to support the increased signaling load. In addition to the user location and call setup signals, adding personal and service mobility to future global environment could trigger an order-of-magnitude increase in signaling traffic. Therefore, a sophisticated mobility management scheme in the global distributed database environment, which prevents huge signaling loads, is important.

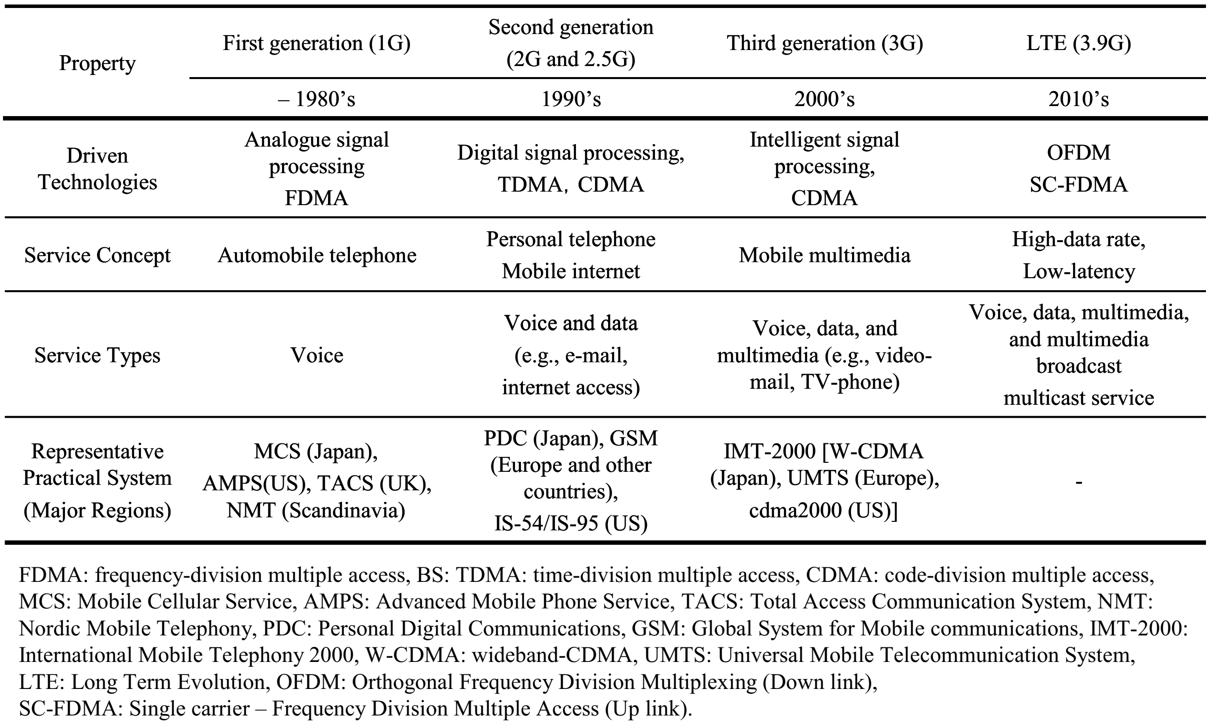
In addition to the above traffic control problems peculiar to cellular systems, realization of multimedia services on an integrated communications network is a shared challenge issue in both wireless and wired networks.

On the basis of the above mentioned background and problems, we focus on traffic control in cellular systems. The aim of this study is to design and manage efficient cellular systems satisfying QoS requirements under variable traffic conditions. We specifically study model analysis methods of traffic control schemes, which enable quantitative performance evaluations of cellular systems.

The reminder of this chapter is organized as follows. In Sect. 1.2, we summarize the evolution of cellular systems and their architecture, which consists of a three-layer structure. In Sect. 1.3, we classify general traffic control schemes in communication systems. Then, we categorize traffic control problems by the three-layer structure of cellular system architecture and show the problems of current control schemes and performance evaluation methods in Sect. 1.4. We summarize the previous work related to traffic control in cellular systems in Sect. 1.5. Finally, we show the outline of this thesis in Sect. 1.6.

[1] Note that also in fixed networks, overload occurs at origination-side caused by concentration of dialing in a short interval. To prevent this overload, subscriber switches have origination restriction functions. The origination-side overload in cellular systems differs from that in fixed networks in that the number of subscribers in a cell validate with time.

Table 1.1: Evolution of cellular systems



Property	First generation (1G)	Second generation (2G and 2.5G)	Third generation (3G)	LTE (3.9G)
	– 1980's	1990's	2000's	2010's
Driven Technologies	Analogue signal processing FDMA	Digital signal processing, TDMA, CDMA	Intelligent signal processing, CDMA	OFDM SC-FDMA
Service Concept	Automobile telephone	Personal telephone Mobile internet	Mobile multimedia	High-data rate, Low-latency
Service Types	Voice	Voice and data (e.g., e-mail, internet access)	Voice, data, and multimedia (e.g., video-mail, TV-phone)	Voice, data, multimedia, and multimedia broadcast multicast service
Representative Practical System (Major Regions)	MCS (Japan), AMPS(US), TACS (UK), NMT (Scandinavia)	PDC (Japan), GSM (Europe and other countries), IS-54/IS-95 (US)	IMT-2000 [W-CDMA (Japan), UMTS (Europe), cdma2000 (US)]	-

FDMA: frequency-division multiple access, BS: TDMA: time-division multiple access, CDMA: code-division multiple access, MCS: Mobile Cellular Service, AMPS: Advanced Mobile Phone Service, TACS: Total Access Communication System, NMT: Nordic Mobile Telephony, PDC: Personal Digital Communications, GSM: Global System for Mobile communications, IMT-2000: International Mobile Telephony 2000, W-CDMA: wideband-CDMA, UMTS: Universal Mobile Telecommunication System, LTE: Long Term Evolution, OFDM: Orthogonal Frequency Division Multiplexing (Down link),
SC-FDMA: Single carrier – Frequency Division Multiple Access (Up link).

1.2 Evolution of cellular systems and their architecture

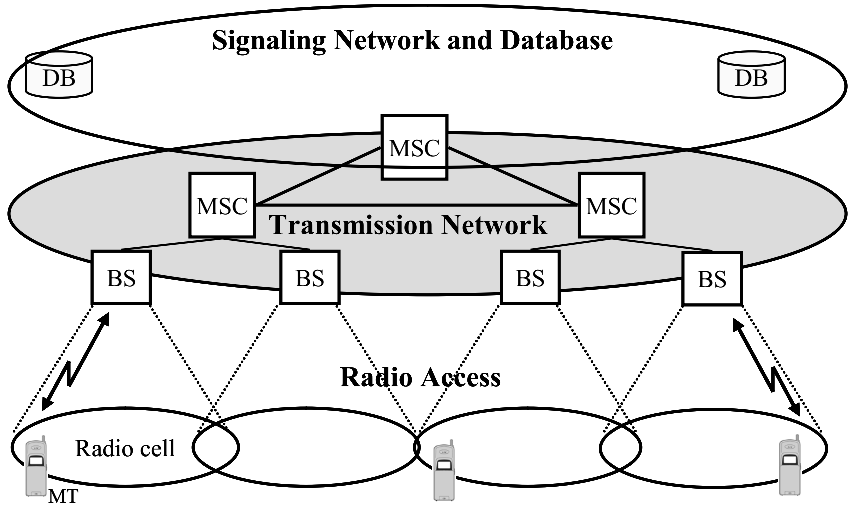
1.2.1 Evolution of cellular systems

The evolution of cellular systems, driven technologies, service concepts, service types, and representative practical systems are shown in Table 1.1.

First generation (1G) cellular systems appeared in the 1980s. The first commercial cellular system was launched as a land mobile telephone system by NTT in Tokyo, Japan in 1979. It provided only classical analog voice service. Mobile terminals were somewhat larger than current ones, and at first, all were designed for permanent installation in vehicles.

The second generation (2G and 2.5G) in the 1990s introduced digital signal processing technology, additional modem-based data services, as well as voice services. Representative practical systems are Personal Digital Communications (PDC) in Japan, Global System for Mobile communication (GSM) in Europe, and IS-54/IS-95 in the United States and Canada. During this generation, the size of mobile terminals decreased dramatically to conveniently portable size, i.e., 100-200 g hand-held devices. This change was possible through technological improvements, such as more advanced batteries and more energy-efficient electronics, but also was largely related to the higher density of cellular sites caused by increasing traffic demands.

The third generation (3G) cellular systems appeared in the first years of this decade.



MT: Mobile terminal, BS: Base Station, MSC: Mobile Switching Center, DB: Database

Figure 1.1: Cellular system architecture

The International Telecommunication Union (ITU) played a leading role in global standardization of 3G and set the standard called International Mobile Telephony-2000 (IMT-2000), which was created to coordinate different initiatives for 3G cellular systems, e.g., Universal Mobile Telecommunication System (UMTS) in Europe and cdma2000 in the United States and Canada. The first commercial systems based on the Wideband Code Division Multiple Access (W-CDMA) technology started in 2001 in Japan. The 3G realized mobile multimedia services through Internet connectivity and packet-switched networks besides the traditional circuit-switched ones, with data rates ranging from 144 Kb/s for fast moving users to 2 Mb/s for slow moving users.

Long term evolution (LTE) is the last step toward the 4th generation of radio technologies designed to increase the capacity and speed of cellular systems. In March 2009, the Third Generation Partnership Project (3GPP) created the Release 8 specification including the LTE and Evolved Packet Core (EPC) set of standards. The fundamental targets of LTE are to offer higher data rates and low delay – peak rates of 300 Mb/s and a radio-network delay of less than 5 ms. As well as high-speed data transfer services, real-time voice service will be provided by voice over IP (VoIP) technology. Further evolution toward LTE-advanced and IMT-advanced (4G) was initiated and beginning to take shape within 3GPP and ITU, respectively.

1.2.2 Cellular system architecture

Before we discuss the traffic controls of cellular systems, it will be useful to summarize the network architecture and its main features.

An example of a network architecture of cellular systems is shown in Fig. 1.1. The main components of cellular systems can be categorized into a three-layer structure, i.e., 1) radio access, 2) transmission network, and 3) signaling network and database.

At the radio-access layer, the service area is divided into a number of different cells. It is the basic idea behind cellular systems. A base station (BS) in each cell handles all

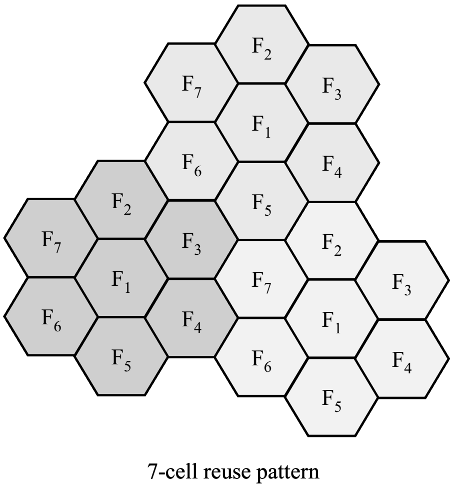


Figure 1.2: Cellular frequency reuse concept

requests on the shared radio channels made by mobile terminals (MTs). Compared to a single large zone system, like a satellite communication system, the distinguishing feature of the cellular system is that a radio channel may be used simultaneously in a number of physically separate cells. This feature, spatial reuse of frequencies, enables the increase in system capacity. The concept of frequency reuse is illustrated in Fig. 1.2.

The transmission network is composed of mobile switching centers (MSCs) and communication links between them. The MSC connects multiple BSs under its control and performs all the switching functions in its particular area.

The signaling network and database transmits, processes, and stores control signals such as location information and connection setup.

1.3 Classification of traffic controls in communication systems

Before we discuss traffic controls in each layer of the cellular system architecture, we summarize traffic control schemes in communication systems. As described in Section 1.1, traffic control is defined as a mechanism that efficiently allocates limited system resources to communication requests satisfying quality of service (QoS) requirements under variable traffic conditions.

From the perspective that "traffic" means *flow* of communication requests, traffic flow control can be classified broadly into the following three categories (see Fig. 1.3):

1. Volume control
   Volume control is a restrictive control which regulates offered load (flow volume) to the system in a certain level, and prevents ineffective resource holding and restrains performance degradation. Volume control can be further classified into two categories, i.e., total flow volume control and individual flow volume control. The control target of the total flow volume control is all the traffic flow of the system. Origination restriction control, in case of traffic congestion, and connection admis-

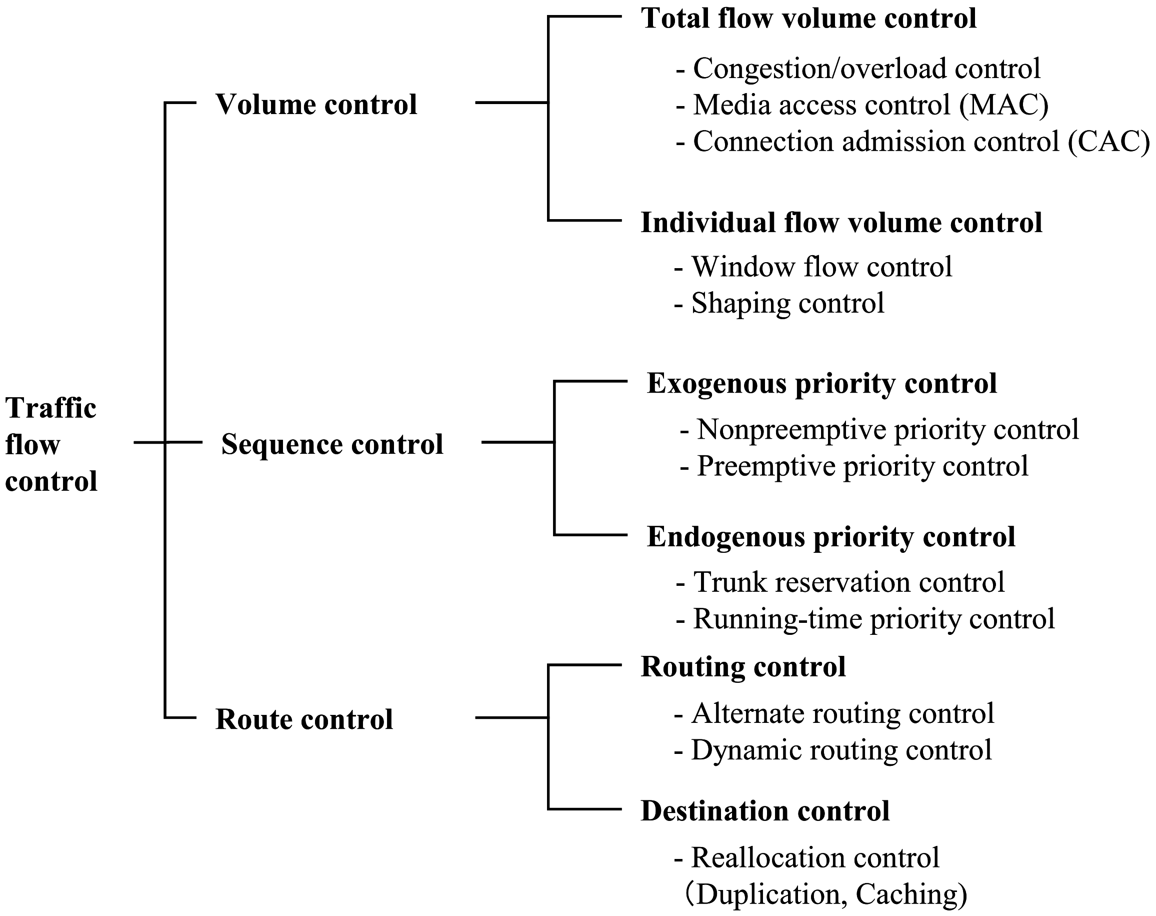

Traffic flow control
Volume control
Total flow volume control
- Congestion/overload control
- Media access control (MAC)
- Connection admission control (CAC)
Individual flow volume control
- Window flow control
- Shaping control
Sequence control
Exogenous priority control
- Nonpreemptive priority control
- Preemptive priority control
Endogenous priority control
- Trunk reservation control
- Running-time priority control
Route control
Routing control
- Alternate routing control
- Dynamic routing control
Destination control
- Reallocation control
(Duplication, Caching)

Figure 1.3: Classification of traffic control schemes

sion control, in case of normal traffic, are such examples. On the other hand, TCP window flow control is a typical example of individual flow volume control.

2. Sequence control
   Sequence control is traffic control that changes the sequence of transmission or processing of flow in the system. Sequence control can be further classified into two categories, i.e., exogenous priority control, such as priority queueing, and endogenous priority control, such as state dependant trunk reservation control.

3. Route control
   Route control is an expansive control that alters the transfer route according to the traffic load in the system. This control also includes destination control which reallocates content using duplication or caching servers.

The various traffic controls classified above are applied to each layer of the cellular system architecture. We summarize traffic control mechanisms in each layer in the next section.

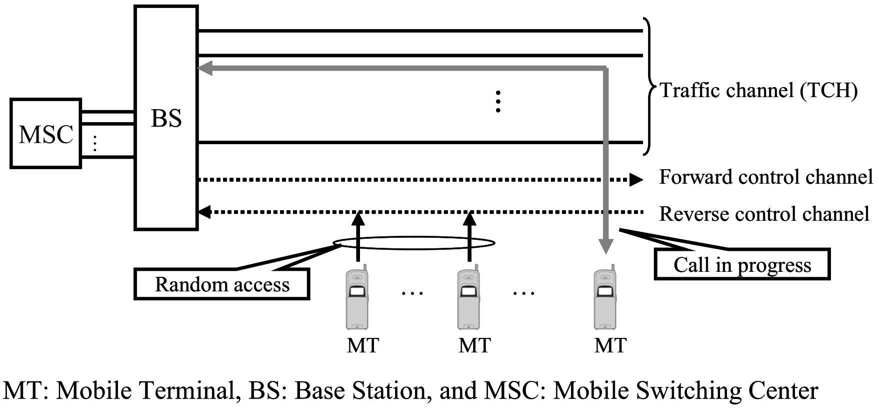


Figure 1.4: Radio channel structure of cellular systems

1.4 Traffic controls in cellular system architecture

According to the categorization described in Sect. 1.3, traffic controls applied to each layer of the cellular system architecture are summarized in Table 1.2. With reference to this table, we describe the traffic control schemes in each layer in the following subsections.

1.4.1 Traffic controls in radio access network

Resource competition in sharing the radio channel is a peculiar problem in this radio access layer. We first describe the radio channel concept and multiple access techniques briefly. The word *channel* refers to a system resource allocated between a given MT and a BS enabling the user to communicate with the network with tolerable interference from other users. The most common types of channels adopted for cellular systems are frequency channels, time slots within frequency bands, and distinct codes. These three different ways of multiple access techniques are termed, frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA), respectively. A radio channel structure of a typical cellular system is illustrated in Fig. 1.4. The radio channel consists of a traffic channel and two types of control channels, forward and reverse. The forward control channel is used for transmission of information, such as a paging signal for a called mobile user, from the BS to the MTs. The reverse control channel is used when the MT sends control signals, such as the call setup signal and the location registration signal, to the BS. After completion of call setup, the traffic channel is assigned and dedicated to the user. In that sense, the traffic channel is the same as the wired access line in fixed-wire systems.

Volume control schemes applied in the radio access layer are as follows:

Media access control (MAC) for reverse control channel

To obtain permission to transmit, a user must first access the BS indicating a desire to use the system. The media access control (MAC) provides this first addressing and channel access control mechanisms on the reverse control channel.

Table 1.2: Classification of traffic control in cellular systems

Traffic controls / Components	Volume control		Sequence control		Route control	
	Total flow volume control	Individual flow volume control	Exogenous priority control	Endogenous priority control	Routing control	Destination control
Signaling network and database	- Congestion / overload control	Transport layer control for individual flow - Window flow control - Shaping control	- Nonpreemptive priority control - Preemptive priority control Chap. IV	- Trunk reservation control - Running-time priority control (e.g., round robin) Chap. V	- Alternate routing control - Dynamic routing control	- Content reallocation control （Duplication, Caching) Chap. VI
Transmission network	- Congestion / overload control - Connection admission control (CAC)		- Nonpreemptive priority control - Preemptive priority control	- Trunk reservation control - Running-time priority control (e.g., round robin)	- Alternate routing control - Dynamic routing control	
Radio access	- Congestion / overload control Chap. II - Connection admission control (CAC) - Media access control (MAC) Chap. III		- Nonpreemptive priority control - Preemptive priority control	- Trunk reservation control for handoff calls - Running-time priority control (e.g., round robin)		

Appendix Channel assignment with autonomous distributed control

The MAC makes it possible for several MTs to share the reverse control channel. Therefore, the MAC is consequently based on the random access scheme and can be classified as total flow volume control. Sharing radio channels according to the random access scheme causes a throughput decrease due to the collision of the signals. Therefore, traffic controls for random access are important to use the channel effectively by suppressing the throughput decrease.

Congestion control

Congestion at the radio access layer occurs according to a spatial and time concentration of the mobile users' traffic. The establishment of a congestion control method, which is scalable and adaptable for handling increasing traffic loads, is a pressing need. A situation in which many mobile users simultaneously request calls, for example, a large-scale earthquake or new-year calls, increases the collision of the call setup signals and leads to a rapid decrease in throughput of the reverse control channel. The decrease in throughput of the reverse control channel causes a high number of call blockings at call setup phase, which leads to inefficient utilization of traffic channels. Therefore, congestion control for random access is essential from the viewpoint of effective use of the precious resource of radio frequencies.

In cellular system environments, the BS measures the traffic and determines congestion control parameters. The traffic information that can be measured at the BS is limited. For example, the BS has no information on the total number of users and the number of backlogged users waiting for retransmission of signals. Under this restricted condition, modeling of congestion control and establishing a method of setting control parameters is a key issue.

Connection admission control

Connection admission control (CAC) is an important feature of CDMA systems, which further distinguishes them from FDMA and TDMA systems. In FDMA and TDMA systems, the number of calls simultaneously carried in the cell is absolutely specified by the number of frequency channels and time slots, respectively. On the other hand, with soft overload [1] in CDMA, it is possible to perform a trade off between transmission speech quality and the number of voice calls carried in the cell. That is, CAC is necessary to provide improved blocking performance for new calls at the expense of some slight degradation in transmission quality. This CAC mechanism is also categorized as the total flow volume control.

Sequence control schemes applied in the radio access layer are as follows:

Priority control at BS queues

Resource allocation in cellular systems designed for packet transmission, such as 3G cellular systems, or those being proposed for LTE and 4G cellular systems, requires appropriate priority scheduling control of packet transmission at the BS. The purpose of priority control is to satisfy the appropriate QoS objectives such as packet loss probability and packet delay for each class of services. For example, packets for continuous-time services, such as VoIP, have priority to be transmitted over those for non-real-time data services. This priority control is categorized as the sequence control.

Trunk reservation control for handoff calls

Handoff is essential to cellular systems. Handoff is required for an MT, which is involved in an on-going call and moves from one cell to an adjacent one. A new traffic channel must be allocated to the handoff call without interrupting the call. From the viewpoint of users, forced call termination of an on-going call caused by handoff blocking is less desirable than blocking a new call. Therefore, handoff blocking probability should be less than the blocking probability of a new call request. To give priority to the handoff call over new call, trunk reservation control is effective, which reserves some channels for handoff calls. This trunk reservation control is a typical endogenous priority control classified in sequence control.

1.4.2 Traffic controls in transmission network

Aside from the radio access layer, the transmission network infrastructure also plays an important role in determining the overall performance of cellular systems. So congestion control in transmission networks is also important as in radio access networks. Connection admission control (CAC) is also necessary for a transmission network, if we use a connection-oriented and packet-based network, such as ATM and all-IP network, as transmission networks.

In addition to these traffic volume controls, this subsection focuses on traffic control for efficient transmission of multimedia on a transmission network layer. Cellular systems are expected to support traffic generated from a wide range of multimedia services, such as text, image, voice, and video. These media have very diverse bandwidth and QoS requirements. Traffic controls will be required to provide performance guarantees, while ensuring sufficient bandwidth is allocated for each media type according to its specific QoS requirements. The design and evaluation of simple controls to meet these requirements in an evolving network represents a significant challenge.

Traffic controls play an important role in transmitting different classes of traffic efficiently on a transmission network. The controls are classified into two levels: packet level control and call level control. Packet level control indicates the control in the information transfer phase after connection is established, and call level control indicates the control in the connection establishment phase. The following two level traffic controls are important.

Packet level priority control

In integrated packet transmission networks, packets for delay-sensitive services such as voice packets, usually have priority at switching nodes over other delay-insensitive media packets such as data packets. For such priority control networks, it is essential to evaluate overall end-to-end delay characteristics for each media service and to develop traffic dimensioning methods for system resources by taking into account priority control effects.

Such systems can be modeled as queueing networks with priority to obtain performance measures. Most of these priority network models are not in a class that can be analyzed exactly. In practice, non-Poissonian arrival and general service time distribution are often encountered in communication systems. The nature of bursting packet arrival processes and the constant packet length in the packet transmission network in cellular systems are examples.

Call level reservation control

We will now leave packet level control and turn to call level control. The question we have to ask here is how to solve the following problems that arise when multimedia traffic with different bandwidths joins a transmission network.

1. Imbalance of blocking probabilities: multiclass traffic with different bandwidths encounters different blocking probabilities at the connection setup phase. For example, a service requiring a wide bandwidth is more frequently blocked than one requiring a narrow bandwidth.

2. Network instability: the end-to-end blocking probabilities show bistable behavior in nonhierarchical networks. That is, for certain loads the network operates in two states: a low network blocking state and a congested state. This bistable behavior causes a high blocking rate during overload.

For each problem, it has been reported that reservation controls are efficient, but there is no control mechanism for combining the above two problems.

1.4.3 Traffic controls in signaling network and database

Mobility has a critical impact on the design and engineering of cellular systems. User location and tracking are major activities related to mobility support, and the efficiency with which the related procedures are performed has a direct bearing on network costs as well as the QoS experienced by users. To achieve this objective, the cellular system maintains user information such as location and account information in the database (DB) on the transmission network. The system must query and update this information when users make either call setup or location registration requests. The management method of this mobility support information is called *mobility management*. Previous studies [2, 3, 4] indicate that increasing the number of mobile users and developing personal communication in the future could trigger an order-of-magnitude increase in signaling traffic for both querying and updating user information. Thus, mobility management that can reduce the signaling traffic load is essential for designing future cellular systems.

Content reallocation control

A distributed database architecture rather than a centralized one is usually applied for mobility management. In a distributed database environment, content reallocation control among databases using duplicating and/or caching of stored information is effective for reducing traffic load. Many mobility management schemes that apply the load balancing control of signaling traffic have been proposed. However, efficient conditions of each mobility management scheme have not yet been sufficiently clarified. In particular, insufficient comparisons have been made using a unified performance measure that is free of assumptions as to the mobility model or database architecture. Clarifying the scope of each control scheme under a unified performance measure is a key issue.

Due to the complexity of the problem, most of the performance comparison studies of mobility management schemes are based on simulation models. A principal problem with simulation comparison is the lack of common context and scenarios within each scheme. Thus, more unified realistic quantitative studies are necessary.

1.5 Previous work related to traffic control in cellular systems

1.5.1 Congestion control for random access schemes

A conventional congestion control in radio access networks is called the *group regulation* scheme, which is used in current mobile systems, such as the digital cellular mobile radio system (PDC) [5] and the personal handyphone system (PHS) in Japan [6]. In this scheme, all mobile terminals are initially classified into eight groups. The traffic offered by MTs is controlled by prohibiting one or more terminal groups from transmitting call-origination signals and/or location registration signals. In this scheme, eight regulation levels, regulation steps of 12.5%, are made by adjusting the number of groups restricted at the same time.

This scheme is simple and effective for current mobile systems, but has two drawbacks for future mobile communication environments. First, the scheme has little scalability in response to an increase in terminals, since the number of groups is fixed at eight, hence the regulation rates are fixed at multiples of 12.5%. Second, it is not adaptive to the bias of terminal groups in a radio cell. That is, since the number of terminals of each group in a congested cell is not likely to be balanced, over- or under-regulation may occur by restricting transmissions by the major or minor group, respectively.

Several studies related to congestion or overload control in random access protocols have been reported [7, 8, 9, 10]. Lam et al. [7] formulated a Markovian decision model for dynamic control of unstable slotted ALOHA systems and found optimal decision rules. Using this Markovian decision model, Okada [8] found a global optimal decision rule for slotted ALOHA systems that maximizes channel throughput. These optimal control methods based on a Markovian decision model cannot be applied to mobile communication systems because the methods assume that all terminals have exact information on the total number of terminals and those that are backlogged at any time, which is impractical in mobile communication environments.

Kawabata et al. [9] proposed a congestion control method for fixed wireless access systems. This method is based on controlling the stand-by time before signal transmission. Kayama et al. [10] proposed an adaptive random access control method for the p-persistent idle-signal casting multiple access (ICMA) [11] protocol. This method uses two control parameters: permission probability q of signal transmission and probability p, which is used if terminals must wait for the end of another's transmission. These parameters are set from 0.0 to 1.0 in 0.2 steps, for example, and stored in a control table in base stations.

In Chap. 2 of this thesis, we generalized this method to increase its adaptability to traffic variation and applicability to other random access protocols.

As shown in Sect. 1.2, cellular systems are evolving toward an all-IP network where high-speed data transfer services and real-time voice over IP service (VoIP) are provided over shared channels, such as long term evolution (LTE) and IMT-advanced. In this new paradigm, congestion control based only on the consumption of radio resources or traffic load is insufficient because there are no dedicated traffic channels and transmission quality such as speech quality of VoIP depends on the number of connected users. In this direction some work has proposed adaptive congestion control based on service quality [12, 13]. Rodrigues et al. [13] propose a QoS-driven adaptive congestion control framework that provides QoS guarantees to VoIP service flows in mixed traffic scenarios for wireless cellular networks.

To realize congestion control based on service quality, it is desirable to develop objective quality assessment methods that accurately estimate users' perceptual quality. In our recent studies, we propose objective quality assessment methods for VoIP [14, 15, 16, 17, 18, 19] and for multimedia services [20, 21, 22]. These methods are practically used for designing and monitoring quality of services on IP networks. Traffic control based on these objective quality assessment methods would be greater challenges for further study.

1.5.2 Traffic analysis method for random access schemes

Several analysis methods have been proposed for traffic design of random access schemes.

Among these, *S-G analysis* is a general method used in the traffic analysis of various packet access protocols. The method was devised by Abramson [23] to evaluate the performance of the ALOHA system at the University of Hawaii. This analysis relates the throughput S as a function of channel-offered traffic G. Since the principle of the S-G analysis is simple, the technique has been applied to static evaluations of many random access schemes. The S-G analysis, however, cannot treat the dynamic characteristics of the system, i.e., behaviors taking into account the system's instability. This is because the channel-offered traffic G is assumed to be a homogeneous Poisson process including retransmission traffic.

Kleinrock and Lam proposed an analytical method called *Markov analysis* that can solve the stability problem in the slotted ALOHA protocol. The Markov analysis formulates a Markovian model taking the number of users in the retransmission phase as a label of state, and obtains the stationary state probability distribution through its transition probabilities. The Markov analysis is a desirable technique for the performance evaluation of random access protocols since it can study the dynamic behavior of the system. However, we usually encounter great difficulty in applying the Markov analysis to complicated protocols that are modeled as multidimensional Markov chains with a vast number of states. Calculation of the state transition probabilities of such Markov chains is difficult, and solving the corresponding set of simultaneous equations is not feasible.

By contrast, *equilibrium point analysis* [24, 25, 26] and *diffusion approximation analysis* [27, 28, 29, 30] are known as analytical methods that can evaluate the stability problems of complicated protocols. Equilibrium point analysis evaluates system performance by considering only the equilibrium of the stochastic process. Because of its ease of analysis feature, it is applied to the evaluation of various packet access protocols [25].

Diffusion approximation analysis is a method that approximates the state transition of a system by a diffusion process. It can provide a more accurate performance evaluation than the equilibrium point analysis based on fluid approximation. However, the diffusion approximation analysis has a problem in that it becomes complex with an increase in the number of states. Consequently, it is used only in the evaluation of packet access protocols such as ALOHA [27, 28, 30] and CSMA/CD [29], which are relatively easy to analyze. There has been no case reported where the method is applied to traffic analysis of complex random access protocols.

We propose a traffic analysis method based on a combination of S-G analysis and diffusion approximation in Chap. 3 of this thesis. By the use of diffusion approximation analysis, not only static characteristics of the system as derived by S-G analysis, but also dynamic characteristics considering stochastic fluctuation of system states can be evaluated for complex random access schemes such as the access inhibition control.

In recent years, significant research efforts have been made for the analyses of the medium access control (MAC) in the IEEE 802.11 standard for wireless local area networks (WLANs). Bianchi [31] applied the equilibrium point analysis for evaluating the IEEE 802.11 MAC protocol. Kumar et al. [32] simplified and generalized the analysis of Bianchi, and clarified a condition for the uniqueness of the equilibrium point. Zhai and Fang [33] applied the Markov analysis for evaluating the IEEE 802.11 MAC protocol. Ghaboosi et al. [34] proposed an analytical framework based on parallel space-time Markov chain to simultaneously model the backoff and post-backoff procedures, in addition to the transmission queue status.

A new approach in the analysis of random access schemes is game theoretic modeling of wireless MAC in which each terminal makes individual decisions regarding their power level or transmission probabilities. Power control non-cooperative games have been studied in Refs. [35, 36]. Random access non-cooperative games have been studied in Refs. [37, 38]. References [39, 40] have studied game theoretic models under both power control and random access. These analyses provide insights into the defence mechanisms against denial of service attacks at the MAC layer in wireless networks.

Interdisciplinary research between teletraffic and security, such as our recent studies [41, 42], would be active in the future.

1.5.3 Priority control in packet transmission networks

Integrated packet transmission networks can be modeled as queueing networks with priority. Most priority queueing network models are not in the class of product form (or BCMP) networks [43], hence several approximation approaches have been used. The most popular approximation approach is called the virtual server method or reduced occupancy approximation [44]. Closed queueing networks with preemptive priority are treated in Refs. [44, 45, 46, 47] using this method, and nonpreemptive cases are considered in Refs. [46, 47, 48]. These studies have shown that the virtual server method captures important properties of priority disciplines in queueing networks. However, this approach is restricted to exponential priority networks, i.e., both interarrival and service time distributions are assumed to be exponential. In practice, non-Poissonian external arrival processes and nonexponential service-time distributions are often encountered in communication systems. The nature of bursting packet arrival processes and the constant packet length in multimedia packet communication systems are examples.

The above restrictions have lead researchers to search for extensions and approximations. A useful way to analyze the steady-state performance of open queueing networks with non-Poissonian external arrival processes and nonexponential service-time distributions is the decomposition approximation method, first proposed by Reiser and Kobayashi [49] and subsequently extended by Kuehn [50], Whitt [51], and many others (see the references in [52]). The main idea is to approximately analyze the individual queues separately after approximately characterizing the arrival processes to each queue using a few parameters (usually two, one to represent the rate and another to represent the variability).

The decomposition method of Whitt, called a queueing network analyzer (QNA), has been refined by various authors. Harrison and Nguyen proposed the QNET method [53], based on reflective Brownian motion processes. Although their method is more exact, an efficient solution only exists for models with just two stations. Dai et al. [54] proposed an alternative decomposition method, the sequential bottleneck method, in which an open queueing network is decomposed into a set of groups of queues. Their method outperforms QNA and QNET in most cases, but not always. Sadre et al. [55] extended Whitt's approach to open queueing networks of finite-buffer $PH/PH/1/K$ queues. Kim et al. [56] proposed a variation of QNA which takes account of correlations between different streams in a queueing network. It should be noted that all these decomposition methods address open queueing networks with single-class nonpriority queues. Whitt [52] developed methods for approximately characterizing the departure process of multiclass single-server queues. Caldentey [57] generalized the results of Whitt [52] related to the interference effect. Balcioglu et al. [58] proposed a three-parameter renewal approximation to analyze splitting and superposition of autocorrelated processes. Although the methods provide a basis for improving the decomposition method for analyzing multiclass queueing networks, the methods are restricted to the first-in-first-out (FIFO) discipline.

In Chap. 4, we present an approximation method for analyzing open queueing networks with two-class nonpreemptive priority, non-Poissonian renewal arrival processes, and nonexponential renewal service-time distributions. We propose a two-parameter virtual server method for priority queueing networks in which the service time distributions of the virtual servers are approximately characterized by the first and second moments. Using the virtual server method, we can obtain two single-class networks from a two-class network, and then by applying the decomposition method to each single-class network, we are able to calculate performance measures.

1.5.4 Traffic control for multiclass routing networks

Dynamic routing for the single-class case has been studied extensively in the last two decades, mostly for circuit-switched networks (see, for example, [59, 60, 61, 62] and references therein). For single-class nonhierarchical fully-connected networks with alternate routing, Nakagome and Mori [63] carried out approximation analyses and revealed the existence of network instability. That is, for certain loads, the network operates mainly in two states, low blocking and congested. This bistable behavior causes a high blocking rate during overload. Later, Krupp [64] and Akinpelu [65] showed that bistable behavior can be stabilized by trunk reservation: reserving a small number of network trunks for direct-routed calls. By preferring direct-routed calls rather than alternate-routed calls, we not only reduce the overall resource usage but also limit the bistable behavior. The effect of trunk reservation for more general nonsymmetric networks can be found in Refs. [65, 66].

Performance analysis of loss networks, such as telephone networks with dynamic routing, is fundamentally difficult because dynamic routing destroys the product form solutions. It is, therefore, of interest to develop computational procedures that accurately approximate blocking probabilities for loss networks with dynamic routing. One such method is the *reduced load approximation method* (also referred to as the Erlang fixed-point approximation) [62]. This approximation assumes that blocking is independent from link to link, giving rise to a set of fixed-point equations whose solution supplies approximations for blocking. The analytical method of Krupp [64] and Akinpelu [65] for sequential alternate routing with trunk reservation is also based on this approximation. Kelly [67] gave a generalized reduced load approximation that can be adapted to essentially any dynamic routing scheme. Chung et al. [68] examined the accuracy and the computational requirements of the approximation procedure for a particular routing scheme, namely, least-loaded routing. Recently Wong et al. [69] proposed a novel blocking probability estimation method called *overflow priority classification approximation* for single-class overflow loss networks with alternate routing. Raskutti et al. [70] extended the method for sequential alternate routing with trunk reservation.

Kelly [71] and Chung and Ross [72] extended the reduced load approximation for multiclass networks with state-dependent routing. It should be noted that these works have not considered the effect of reservation control for direct-routed calls and wideband calls. Lee et al. [73] investigated an approximation method for a multiclass alternate routing network with trunk reservation for only direct-routed calls. Greenberg and Srikant [74] proposed computational techniques for multiclass sequential routing networks with state-dependent admission control and general topology. Medhi and Sukiman [75] gave a comparative study of multiclass dynamic routing schemes with reservation control by simulation.

In Chap. 5, we first show that there is an imbalance of individual end-to-end blocking probabilities and network instability in multiclass fully connected networks with sequential alternate routing, where the class of calls is differentiated by the required bandwidth and the offered load. To prevent imbalance and suppress instability, a link capacity allocation control, which is a combination of reservation control for wideband calls and for direct-routed calls, is introduced. Then, we propose an approximation method to evaluate end-to-end blocking probabilities based on the reduced load approximation.

The link capacity allocation control is effective not only for realizing fairness among multiclass traffic but also for providing *resilience* [76] of networks. Network node and link failure are still frequent events on the Internet [77]. Our recently studies focus on traffic measurement [78], network reliability estimation [79, 80, 81], and network topology design [82, 83, 84] which are important steps to ensure the network resilience in addition to the traffic control discussed in Chap. 5.

1.5.5 Evaluation of mobility management schemes

Current mobility management schemes are based on two types of databases, the home location register (HLR) and the visitor location register (VLR). This HLR-VLR architecture has been established as an industry standard in the global system for mobile telecommunication (GSM) in Europe [85] and the IS-41 recommendations for North America [86]. Signaling and database systems in the HLR-VLR architecture have potential performance problems due to the anticipated high terminal mobility and traffic demands in future wireless networks.

Three kinds of network database architectures are closely related to mobility management schemes.

The first is centralized; all user information is stored in a central database. This architecture is simple and has the advantage of easy operation and control if it is applied to a system that provides comparatively small-scale service in a limited area. However, it has some disadvantages, such as decreased serviceability and reliability because of concentrated signal processing, particularly when it is applied to a cellular system with a global environment.

The second architecture is a hierarchical database structure [87, 88, 89]. Each leaf-level database serves a location registration area and stores user information. Each database in the higher network levels stores path pointers to the next lower-level database that stores user information or has a pointer to a lower-level database. Obana et al. [87] have proposed and evaluated the applicability of an open system interconnection (OSI) directory as a hierarchical database architecture for universal personal telecommunication (UPT) services. This hierarchical architecture helps to reduce the total network signaling traffic when most calls received by users and most registration area crossings are geographically localized [88]. Because of the tree structure, however, the total number of database queries (and hence, call setup time) can potentially increase [88, 89]. Moreover, the performance of this architecture strongly depends on the design of the tree structure and mobility characteristics of the users.

The third is a distributed database architecture composed of fully distributed databases connected by signaling links. This architecture is commonly used in current digital cellular systems. In Chap. 6, we are concerned with mobility management schemes with distributed database architecture taking into account future trends for communication network architecture based on a distributed processing environment (DPE) [90].

Many claims of the advantage of one mobility management scheme over another have been presented, such as a scheme with replication [91] and caching [92]. Very few attempts, however, have been made at comparative studies of mobility management schemes. Leung and Levy [93] proposed and analyzed full and partial data replication schemes, and compared them with the centralized database scheme. Although several mobility management schemes were surveyed in Ref. [94], it is as yet unclear to what extent and in what circumstances each scheme increases the capacity of a system.

In Chap. 6, we categorize the various mobility management schemes with load balancing controls into four types. Then, we propose a comparative evaluation model and a unified performance measure. Finally, we clarify the efficient condition of each type of scheme.

1.6 Outline of the thesis

We are now in a position to outline this thesis that consists of six chapters. The relationships between the traffic controls and problems that we will treat in the subsequent chapters are also mapped in Table 1.2.

Traffic control problems in the radio access network, especially on the reverse control channel, are discussed in Chaps. 2 and 3. Chapter 2 deals with a congestion control in the reverse control channel, and Chap. 3 with random access problem there. Sequence and route control problems in the transmission network are dealt in Chaps. 4 and 5. Chapter 4 discusses the priority control for multimedia packet networks. Chapter 5 discusses composition control of trunk reservation and routing control for multimedia connection-oriented networks. Chapter 6 discusses the mobility management problem with content reallocation control. In addition to the discussions in these chapters, we deal with a channel assignment problem with autonomous distributed control in the Appendix. An outline of each chapter is as follows:

Chapter 2

A congestion control scheme for the control channel in the radio access network is proposed in Chap. 2 [95]. The main contribution is the proposal of an adaptive and scalable congestion control model and a method for setting its parameter that maintains maximum throughput even under overloaded conditions.

The main features of the proposed scheme are scalability for handling increasing numbers of MTs and adaptability for coping with drastic changes in traffic load. These are achieved by controlling the traffic load adaptively to maintain maximum throughput even under overloaded conditions. Procedures for measuring and estimating offered traffic and a method of setting control thresholds that maximize the average throughput are analytically derived, and an algorithm that is easy to implement is described.

This control scheme is applied to both the slotted ALOHA and idle-signal casting multiple access with collision detection (ICMA/CD) protocols. For each protocol, control parameters are analytically derived. Then, stationary throughput characteristics are numerically evaluated. The preferred range of the control parameter is also clarified under the condition that enables adaptive control and limits the amount of processing at terminals. We simulate its transient characteristics with three types of time-variant input models. The results indicate that the proposed control achieves nearly an optimal throughput even in the overload state for all input models.

Chapter 3

A traffic design method that enables comparative evaluations of various random access schemes is proposed in Chap. 3 [96]. The random access scheme in cellular systems differs from the simple random access schemes in *access inhibition control* using the forward control channel and in *retransmission control* restricting the retransmission interval and the maximum number of retransmissions.

The access inhibition control is used to prevent collision between the response signal from a called user and the call setup signal from calling users. Under this control, the BS transmits an inhibit signal (*busy signal*) immediately after the transmission of the incoming call setup signal through the forward control channel, which inhibits the transmission of call setup signals from all users in the cell. Retransmission control is used to reduce the call setup delay. Under this control, the retransmission interval for the call setup signal is set to a short interval, and the maximum number of retransmissions is limited.

The proposed method can clarify the effects of the access inhibition control, the retransmission interval, and the upper limit of the number of retransmissions on the performance.

We construct a system model considering the transitions among the states of mobile terminals, e.g., signal transmission and retransmission waiting. We then propose a traffic analysis method based on a combination of two approximation methods, i.e., S-G analysis and diffusion approximation. The S-G analysis relates the throughput S as a function of channel-offered traffic G. By using the diffusion approximation analysis, not only the static characteristics of the system as derived by S-G analysis, but also the dynamic characteristics considering the stochastic fluctuation of the system states can be evaluated. The performance measures that can be derived with the proposed method are the throughput, the mean number of active stations, the mean connection delay, the control failure rate, and the mean number of retransmissions.

Using the proposed method, the performances of ALOHA- and ICMA-type access schemes with access inhibition and retransmission control are evaluated. Through comparisons with the simulation, it is verified that the proposed approximation method gives satisfactorily accurate estimates under practical conditions.

Chapter 4

The priority control at the packet level in the transmission network is discussed in Chap. 4. Here an approximate method for queueing networks with nonpreemptive priority is proposed [97, 98]. It can be used to evaluate the end-to-end delay and throughput in the transmission network of multimedia cellular systems. The method is composed of three steps outlined below. These steps are based on the virtual server method for priority networks and the decomposition method for queueing networks with non-Poissonian external arrival processes and nonexponential service-time distributions. (1) Separation of a two-class queueing network into two single-class queueing networks, replacing each priority server in the original network with two virtual servers, one serving high priority customers, and one for low priority customers. (2) Analysis of the flow rates and variability parameters of internal arrival processes for each network. (3) Calculation of traffic characteristics for each node in the networks by regarding the node as a $GI/G/1$ queue.

The virtual server method has been validated for priority networks with Poisson arrivals and exponential service times, but not for priority networks with general interarrival and service time distributions. Hence, we propose a two-parameter virtual server method for renewal-queueing networks in which the service time distributions of the virtual servers are approximately characterized by the first and second moments, that is, the mean and variance of the service time distributions.

The proposed method takes into account the variability of the arrival processes and service times for each node in the network, and hence a general class of renewal queueing network models can be analyzed. The accuracy of the method is validated by comparisons with exact results for a tandem network model and simulation results for the basic component models of complex networks. Also, we apply our approximation method to an end-to-end delay analysis for a packet-switching transmission network for voice and data. The results indicate that the accuracy of the approximation method is sufficient for practical use.

Chapter 5

A link capacity allocation control for multiclass alternate routing networks is proposed in Chap. 5 [99]. While an evaluation method for the packet level priority control in transmission networks is discussed in Chap. 4, a control scheme at the call level is discussed in this chapter.

The control scheme proposed here is a combination of bandwidth reservation control for wideband calls and direct-routed calls. We also formulate a method for approximating the end-to-end blocking probabilities for a multiclass nonhierarchical routing network with the proposed control scheme. The method is based on solving two nonlinear equations expressed with the variables: offered load of alternate-routed calls to each link and link blocking probabilities for direct- and alternate-routed calls.

In this chapter, we first demonstrate that there is an imbalance of individual end-to-end blocking probabilities and network instability in multiclass nonhierarchical networks with alternate routing, where heterogeneous traffic classes are differentiated by the required bandwidth and the offered load. To prevent imbalance and suppress instability, a link capacity allocation control is introduced. Then, we propose an approximation method to evaluate end-to-end blocking probabilities under the control. In addition, the setting of allocation control parameters to suppress and to ensure maximum useful network throughput under various traffic conditions is discussed.

Chapter 6

The mobility management problem with content reallocation control in the signaling network and database layer in cellular systems is discussed in Chap. 6 [100]. Especially we focus on a comparative study for mobility management schemes with load balancing controls. We propose a comparative evaluation model and a unified performance measure, which is the number of signals at call setup and location registration. This performance measure is important in determining the QoS because it closely relates to call setup time and network efficiency.

We categorize the various mobility management schemes with load balancing controls into five types with respect to the control functions such as replicating and caching functions of user information. We then express the performance measure for each type in terms of four parameters: call setup rate, location registration rate, the probability that a user is in his/her home area, and the probability that the calling and called users are both in the same area.

One of these types with both replicating and caching functions is effective in reducing the number of accesses in the network control layer, and hence, the call setup time. This scheme is especially efficient when the probability that a user is in his/her home area is relatively small and/or the call setup rate is relatively high compared to the location registration rate.

Appendix

For efficient utilization of radio frequencies, a channel reuse scheme that is consistent with the objectives of increasing capacity and minimizing interference is required. Two main approaches have been followed for this purpose: *fixed channel assignment* (FCA) and *dynamic channel assignment* (DCA). Between the extremes of FCA and DCA lie various schemes such as flexible channel assignment, channel borrowing schemes, and hybrid channel assignment [101]. Among these, Furuya and Akaiwa [102] have proposed a method called *channel segregation* that combines the advantages of FCA and DCA. Channels are assigned to calls on the basis of a priority list that is updated according to the radio frequency interference detected by each BS that uses a learning function.

Traffic design of the channel assignment scheme with channel segregation is dealt with in the Appendix [103]. We propose an analytical algorithm that gives the upper limit of call blocking probabilities, and can be used for traffic dimensioning of the number of BS channels. We formulate an N-dimensional Markov model of the system and derive the product form solution of the equilibrium state probabilities. Using these state probabilities, we express the upper limit of call blocking probabilities for each BS.

The amount of calculation required for our analysis is greatly reduced by assuming that the relevant system state is not the number of connections of every cell, but the largest number of simultaneous connections among the cells with the same channel selection pattern. The improvement effect of the call blocking rate by increasing the number of BS channels is evaluated using this algorithm. Moreover, we derive traffic design and an administration method based on this channel-dimensioning algorithm.

Chapter 2

Congestion control for random access schemes

2.1 Introduction

This chapter presents a congestion control scheme for the control channel in the radio access layer of cellular systems.

As shown in Chap. 1, market demand for cellular services has expanded at an explosive rate in the past several years. In addition, the popularity of communication is changing from voice communication to Internet communication, such as e-mail transmission and web browsing via cellular terminals. This change will lead to increased demand for mobile multimedia communication. Overall, these trends may trigger a massive increase in the traffic volume in cellular systems.

Thus, traffic control that can handle huge traffic loads is essential in designing future cellular systems that satisfy QoS requirements and use system resources efficiently. In particular, scalable congestion control is an important issue for contending with the explosive change in traffic and for maximizing system resource efficiency.

Cellular systems experience a peculiar type of congestion that fixed-wire communication networks do not. In fixed-wire networks, such as existing telephone switching networks, an overload ordinarily occurs at a network termination when traffic concentrates on a specific telephone-number or region because of an event or disaster. In cellular systems, in addition to this congestion, overload occasionally occurs at a network origination caused by mobile terminal (MT) concentration at a specific time and place. For example, when a public transportation system stops somewhere during commuting hours, many mobile users in a specific cell at the radio access layer may originate calls to their offices at the same time. Consequently, congestion control in the radio access layer as well as core networks is an important issue in cellular systems. This issue is also essential from the viewpoint of effective use of the precious resource of radio frequencies.

In this chapter, we propose a congestion control scheme that can be applied to various random access protocols in cellular systems. Its features are scalability for handling increasing numbers of MTs and adaptability for coping with drastic changes in traffic load. These are achieved by controlling the traffic load adaptively to maintain maximum throughput even under overloaded conditions.

The reminder of this chapter is organized as follows. In Sect. 2.2, we propose a congestion control scheme, which consists of a procedure for measuring and estimating input traffic, a method of setting control thresholds, and an adaptive controlling algorithm. Then, we apply this scheme to both the slotted ALOHA and ICMA/CD (idle-signal casting multiple access with collision detection) protocols [11], derive the control parameters, and clarify the stationary characteristics in Sect. 2.3. We evaluate transient characteristics by simulation in Sect. 2.4, and draw some conclusions and discuss topics for future study in Sect. 2.5.

2.2 Congestion control scheme

In this section, we describe our congestion control scheme. The main advantage of the proposed congestion control is in its high average throughput even during an overload. This is achieved by regulating signal transmissions from MTs through *permission rate* adapted in accordance with the channel-offered traffic. The permission rate is the probability that each MT is allowed to transmit a signal to the random access channel.

Here we deal with a channel with a random access protocol satisfying the following conditions.

(1) Slotted uplink channel: The reverse channel for random accesses is slotted. This is a practical assumption because the random access channels in cellular systems are usually slotted.

(2) Noise-free channel: The reverse and forward (broadcast) channels are noise-free channels such that a signal is received incorrectly if and only if a channel collision occurs.

(3) Independent input: The input traffic is generated from independent signal attempts, i.e., total numbers of signal attempts (both newly arrived and previously collided retransmitted signals) by all MTs in subsequent time slots are mutually independent random variables. However, their distributions, which are sometimes assumed to be Poissonian, may gradually vary with time t, and we define their mean, which we call *the offered traffic*, at time t as $\lambda(t)$.

(4) Permission-rate-base regulation: MTs obey the instruction for the permission rate that is broadcasted from the BS.

2.2.1 Estimation of $\lambda(t)$

First, we consider the problem of measuring and estimating the offered traffic on a random access channel. When a collision occurs in a slot, the BS cannot recognize the number of signals transmitted in the slot. In this sense, the BS cannot know the offered traffic directly, but to regulate the signal transmissions from MTs, the BS has to estimate the offered traffic. To cope with this problem, we propose measuring and estimating the *unsuccessful transmission rate* instead of the offered traffic.

Let X be a random variable that represents the number of signals in a slot. Then the BS can recognize the following three states of the slot: $\{X = 0\} = \{\text{the slot is idle}\}$,

$\{X = 1\} = \{$only one signal was transmitted successfully$\}$, and $\{X \geq 2\} = \{$two or more signals were transmitted and they collided$\}$.

The unsuccessful transmission rate is the ratio of the number of collided slots to that of busy slots[1] defined by:

$$\begin{aligned} r &\equiv \frac{\Pr\{\text{Collision}\}}{\Pr\{\text{Success or Collision}\}} \\ &= \frac{\Pr\{X \geq 2\}}{\Pr\{X = 1\} + \Pr\{X \geq 2\}} = 1 - \frac{\Pr\{X = 1\}}{1 - \Pr\{X = 0\}}. \end{aligned} \tag{2.1}$$

This quantity, as a function of the offered traffic $x = \lambda(t)$, can be calculated for each protocol specified under some natural conditions. For example, assuming a Poisson process for the input traffic, the unsuccessful transmission rate of the slotted ALOHA channel is given by

$$r(x) = 1 - \frac{hxe^{-hx}}{1 - e^{-hx}}, \tag{2.2}$$

where h and x represent the slot length and the offered load, respectively. From l'Hospital's law, it can be easily seen that the unsuccessful transmission rate satisfies the following properties:

$$\lim_{x \to 0} r(x) = 0, \ \lim_{x \to \infty} r(x) = 1, \quad \text{and} \quad r'(x) > 0. \tag{2.3}$$

It follows that $r(x)$ is a monotonically increasing function in x bounded in the interval $[0, 1]$. Therefore, the offered traffic $x = \lambda(t)$ can be easily estimated from the unsuccessful transmission rate r.

The implementation of the estimation procedure is as follows. Let m slots be the length of a measurement interval. For the ith measurement interval, let

N_i^I be the number of idle slots,
N_i^S be the number of successful slots, and
N_i^C be the number of collided slots.

Then the unsuccessful transmission rate r_i for the ith measurement interval can be estimated by

$$r_i = \frac{N_i^C}{N_i^S + N_i^C} = 1 - \frac{N_i^S}{m - N_i^I}. \tag{2.4}$$

It follows that the BS is only required to count the number of idle slots, N_i^I, and the number of successful slots, N_i^S. This leads to easy implementation. In addition to this periodic measurement, sliding-window-based measurement may be possible.

[1] Note that the unsuccessful transmission rate does not stand for the unsuccessful call rate measured at a terminal, but stands for the ratio of the number of collided slots to that of busy slots excluding idle slots measured at the BS.

In the case of simple protocols such as the slotted ALOHA, instead of the unsuccessful transmission rate, a measure that is easier to understand, such as the ratio of the number of idle slots, can be used. Here, the unsuccessful transmission rate is defined above assuming that the measure would be applied to more complex protocols such as ICMA/CD, as will be shown in Sect. 2.3.

Hereafter, we abbreviate $r(t)$ to r, which is a time-variant measure for estimating $\lambda(t)$.

2.2.2 Outline of procedure

The outline of the congestion control scheme proposed in this paper is described as follows.

Step 1: The base station (BS) measures and estimates the offered traffic $\lambda(t)$ by measuring the unsuccessful transmission rate r_i and decides the power n of the permission rate α^n so that the offered traffic under regulation stabilizes in the interval $[\alpha\lambda^*, \lambda^*)$. Here, as will be shown later, λ^* is the target traffic for a given α that yields a high average throughput when the offered traffic is within the above interval.

Step 2: The BS broadcasts the regulated level in terms of the power n through a forward channel (down link) to the MTs.

Step 3: When an MT originates a signal, the signal transmission is permitted with probability α^n and not permitted with probability $1 - \alpha^n$ (i.e., regulated).

The key idea of the scheme is that the permission rate is chosen so as to keep the average throughput under the regulation high in Step 1. In the subsequent subsections, we give more details about the concrete procedure for determining the permission rate properly.

2.2.3 Target traffic

The target traffic λ^* that yields a high throughput is derived as follows. For a given offered traffic x, let $S(x)$ be the throughput that is defined as the average number of correctly received signals per signal transmission time. If we assume that the offered traffic extends from $\alpha\lambda$ to λ uniformly, then the average throughput[2] is given by

$$f(\lambda) = \frac{1}{(1-\alpha)\lambda} \int_{\alpha\lambda}^{\lambda} S(x)dx. \tag{2.5}$$

Then, we take the target traffic λ^* as λ, which maximizes $f(\lambda)$[3], namely

$$\lambda^* = \{\lambda \mid maximize\ f(\lambda)\}. \tag{2.6}$$

The target traffic λ^* can be numerically derived for a given random access protocol and a given permission base rate α, as will be shown in the next section. The throughput $S(x)$ of a slotted ALOHA random access protocol is illustrated in Fig. 2.1. The target traffic λ^* is equivalent to the traffic λ that maximizes the shaded area of this figure for a given α. The proposed scheme controls the offered traffic in the target interval $[\alpha\lambda^*, \lambda^*)$ by regulating the offered traffic in the domain of $x > \lambda^*$.

[2] Hereafter, we abbreviate the time-variant offered traffic $\lambda(t)$ to λ.

[3] Here, the target interval is expediently defined by $[\alpha\lambda^*, \lambda^*)$, where λ^* maximizes Eq. 2.5. This definition means that $\lambda(t)$ is assumed to be uniformly distributed on the target interval, which could not be a realistic assumption. An alternative interval can be defined as $[\sqrt{\alpha}\lambda^{**}, \lambda^{**}/\sqrt{\alpha})$, where λ^{**} is λ, which maximizes $S(\lambda)$.

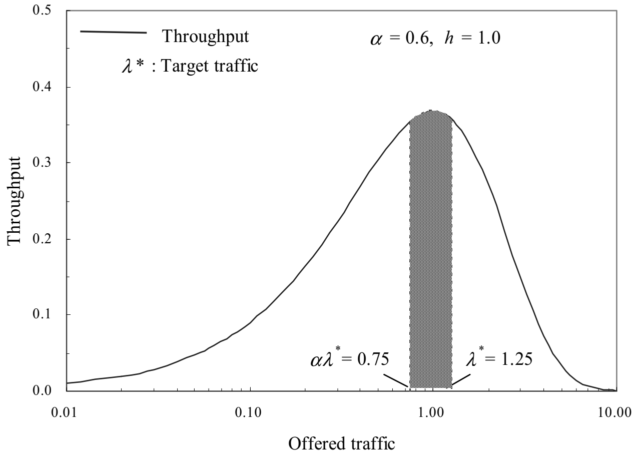


Figure 2.1: Throughput of slotted ALOHA and target traffic

2.2.4 Control thresholds

Next, we derive control thresholds for the estimated unsuccessful transmission rate. Control thresholds $R^{(k)}$s are set up so that the regulated traffic is within the target interval $[\alpha\lambda^*, \lambda^*)$. Figure 2.2 shows the unsuccessful transmission rate and the regulated traffic for the slotted ALOHA with $\alpha = 0.7$ and $h = 1.0$ as functions of offered traffic. The solid line is the unsuccessful transmission rate and the dotted line is the regulated traffic after transmission regulation. In Fig. 2.2, the control thresholds $R^{(k)}$s and the permission rates α^n are also designated.

As shown in this figure, the regulated traffic is controlled within the interval $[\alpha\lambda^*, \lambda^*)$ if the permission rate is set to $\alpha^k (k = 1, 2, ...)$ when the offered traffic is within the interval $[\alpha^{1-k}\lambda^*, \alpha^{-k}\lambda^*)$. The control threshold when the offered traffic is at the lower bound $\alpha^{1-k}\lambda^*$ of the interval $[\alpha^{1-k}\lambda^*, \alpha^{-k}\lambda^*)$ is given by

$$R^{(k)} = r(\alpha^{1-k}\lambda^*), \tag{2.7}$$

where $r(\cdot)$ is the unsuccessful transmission rate for a given random access protocol derived by analysis, and is given by Eq. 2.2 in the case of the slotted ALOHA.

2.2.5 Algorithm

Let d_i be the permission rate in the ith measurement interval, and r_i be the estimated unsuccessful transmission rate in the interval. Based on these values, the BS calculates d_{i+1} in the next interval by

$$\begin{aligned} &\text{if } R^{(k)} \leq r_i < R^{(k+1)}, \text{ then } d_{i+1} = \min(1, \alpha^k d_i), \\ &\qquad k = ..., -2, -1, 0, 1, 2, ..., i = 1, 2, ..., \end{aligned} \tag{2.8}$$

where $d_1 = 1, R^{(k)} = r(\alpha^{1-k}\lambda^*), k = ..., -2, -1, 0, 1, 2, ...$ Note that $R^{(k)}$ is also defined by Eq. 2.7 for $k \leq 0$. Here, the conditions $k < 0$ and $d_i \leq \alpha$ mean to cancel a regulation

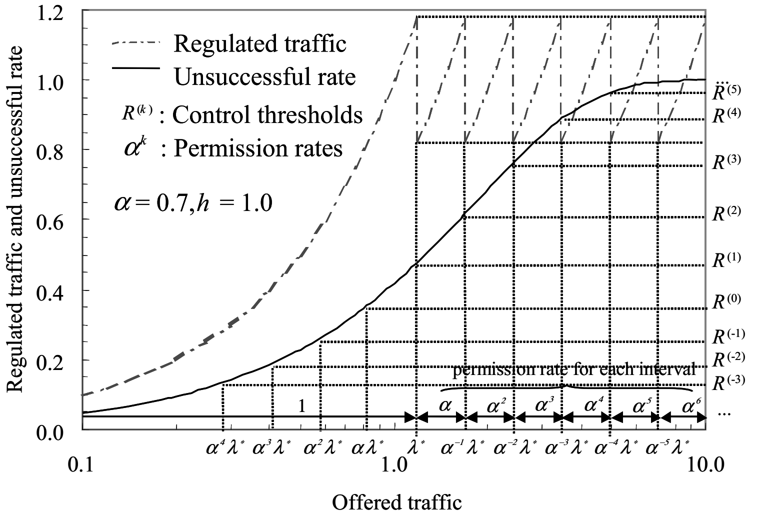


Figure 2.2: Regulated traffic and unsuccessful rate

because the offered traffic is going to return to the stable state from the overloaded state anyway, otherwise, it would become an over regulation. Additionally, the reasons for needing min in the above equation are (1) to cancel the regulation and set $d_{i+1} = 1$ when the offered traffic is less than λ^* and (2) to prevent the permission rate from becoming greater than 1 because of the estimation error of r_i.

The above algorithm calculates the permission rate d_i, and the rate is broadcast by the BS. On the other hand, if α is initially set up in the MTs, it is sufficient for the BS to broadcast the power of α^n. In that case, the above algorithm reduces to

$$\text{if } R^{(k)} \le r_i < R^{(k+1)} \text{ then } n_{i+1} = \max(0, n_i + k),$$
$$k = ..., -2, -1, 0, 1, 2, ..., i = 1, 2, ..., \tag{2.9}$$

where n_i represents the power number of the ith measurement interval and $n_1 = 0$. The BS broadcasts n_{i+1} decided by the above algorithm and the MTs calculate $\alpha^{n_{i+1}}$.

2.3 Applications and stationary characteristics

In this section, we apply the proposed congestion control scheme to both the slotted ALOHA and the ICMA/CD protocols [11] and analyze control parameters and stationary characteristics.

2.3.1 Slotted ALOHA

The slotted ALOHA is one of the most fundamental and important random access protocols. For example, it is used in the Physical Random Access Channel (PRACH) in third-generation mobile communication systems, e.g., IMT-2000 [104]. The basic characteristics of the slotted ALOHA —the throughput $S(x)$ and the unsuccessful transmission rate $r(x)$ for a given offered traffic x, assuming a Poisson process for the input traffic —

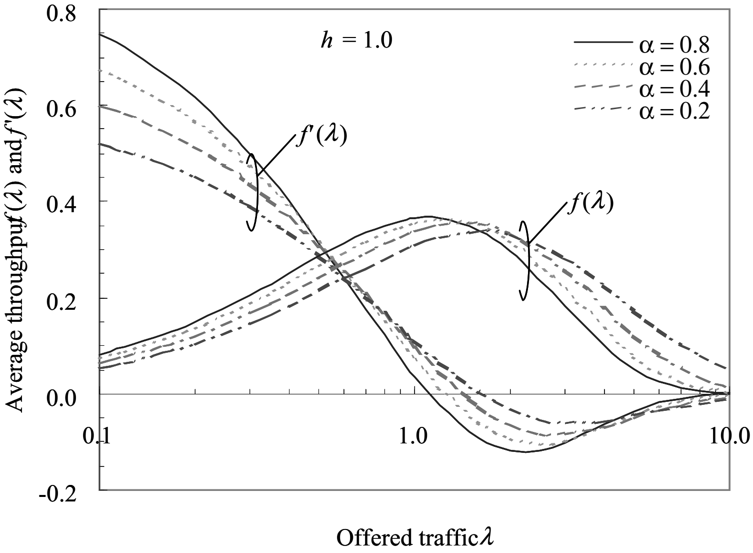


Figure 2.3: Average throughput $f(\lambda)$ and $f'(\lambda)$

are given by

$$S(x) = hxe^{-hx} \quad \text{and} \tag{2.10}$$

$$r(x) = 1 - \frac{hxe^{-hx}}{1-e^{-hx}}. \tag{2.11}$$

Target traffic

The average throughput $f(\lambda)$ where the offered traffic extends from $\alpha\lambda$ to λ is given by

$$\begin{aligned} f(\lambda) &= \frac{1}{(1-\alpha)\lambda}\int_{\alpha\lambda}^{\lambda} S(x)dx \\ &= \frac{(1+\alpha h\lambda)e^{-\alpha h\lambda} - (1+h\lambda)e^{-h\lambda}}{(1-\alpha)h\lambda}. \end{aligned} \tag{2.12}$$

Upon differentiating $f(\lambda)$, we have

$$f'(\lambda) = \frac{(1+h\lambda+h^2\lambda^2)e^{-h\lambda} - (1+\alpha h\lambda+\alpha^2h^2\lambda^2)e^{-\alpha h\lambda}}{(1-\alpha)h\lambda^2}. \tag{2.13}$$

It follows that the target traffic λ^* that maximizes $f(\lambda)$ can be derived by numerically solving the equation $f'(\lambda) = 0$, or equivalently :

$$e^{-(1-\alpha)h\lambda^*} = \frac{1+\alpha h\lambda^* + \alpha^2h^2\lambda^{*2}}{1+h\lambda^*+h^2\lambda^{*2}}. \tag{2.14}$$

Numerical examples of $f(\lambda)$ and $f'(\lambda)$ for $h = 1.0$, and $\alpha = 0.2$,0.4,0.6, and 0.8 are shown in Fig. 2.3. The intersection points of $f'(\lambda)$ and the x-axis represent the target traffic λ^*. The target traffic λ^* as a function of α for $h = 1.0$ is plotted in Fig. 2.4. It is a

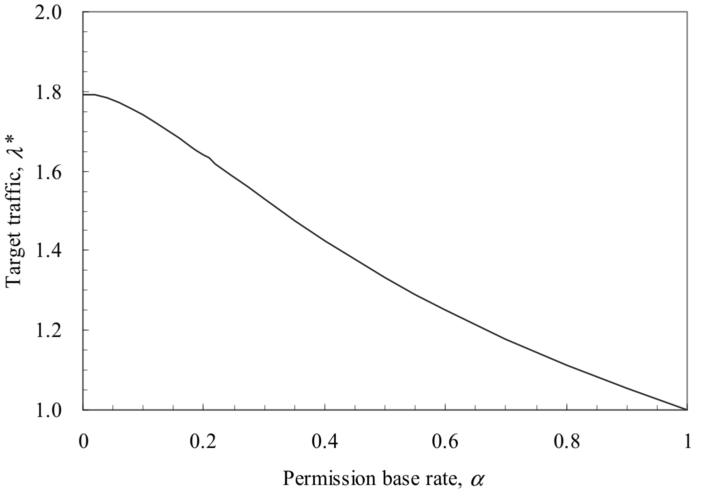


Figure 2.4: Target traffic for slotted ALOHA

monotonically decreasing function of α with minimum and maximum values

$$\lambda^*_{min} = 1/h \ (\alpha = 1), \ \lambda^*_{max} = x^*/h \ (\alpha = 0),$$

where $x^* = 1.793$ is a root of the following equation derived from setting $\alpha = 0$ and $f'(\lambda) = 0$ in Eq. 2.14.

$$1 + x + x^2 = e^x, \quad x > 0. \tag{2.15}$$

Control thresholds

As described in Subsect. 2.2.4, the control thresholds for the slotted ALOHA are given by

$$R^{(k)} = r(\alpha^{1-k}\lambda^*) = 1 - \frac{h\alpha^{1-k}\lambda^* e^{-h\alpha^{1-k}\lambda^*}}{1 - e^{-h\alpha^{1-k}\lambda^*}}. \tag{2.16}$$

The thresholds $R^{(k)}(k = -6, -5, \ldots, 8)$ for $h = 1.0$ as functions of α are plotted in Fig. 2.5. For $\alpha < 0.6$, the control thresholds $R^{(k)}$ for $k \geq 3$ are close to 1. In this case, it might be difficult to control the time-variant overloaded state adaptively. On the other hand, for $\alpha \to 1$, the optimal and adaptive control can be achieved. That is, the case of $\alpha = 1$ corresponds to the optimal scheme in which the permission rate is completely adapted to the estimated offered load. Precisely, the scheme always controls the offered traffic as $\lambda^* = \lambda^*_{min} = 1/h$ in order to achieve the maximum throughput, $S_{max}(1/h) = 1/e$. In this case, the threshold becomes a constant,

$$R^{(k)} = r(1/h) = \frac{e-2}{e-1} = 0.418. \tag{2.17}$$

Though the optimal control is achieved by setting $\alpha \to 1$, the power number n_i of the permission rate becomes large, and it leads to the power calculation and processing at

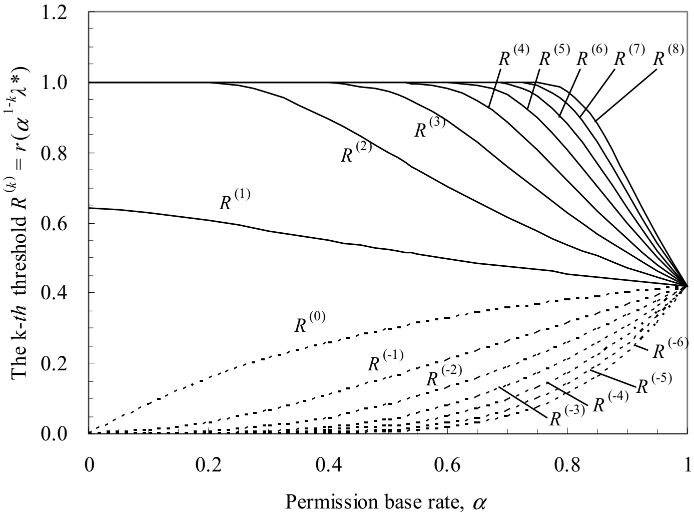


Figure 2.5: Control thresholds for slotted ALOHA

the MT taking a lot of time. For example, in order to set the permission rate to 10% for $\alpha = 0.99$, 230 power calculations are necessary since $n_i > \log(0.1)/\log(0.99) = 229.1$. Consequently, α should be set in the range $0.7 \leq \alpha \leq 0.9$, which enables reasonable adaptive control while limiting the amount of processing at the MT.

Throughput characteristics

The statistical throughput characteristics for $\alpha = 0.7$ and $h = 1.0$ are shown in Fig. 2.6. In this figure, the dotted and solid lines represent the throughput without and with control, respectively. The statistical throughput without control is given by Eq. 2.10. That with proposed control is also given by Eq. 2.10 except for that when the offered traffic is within the interval $[\alpha^{1-k}\lambda^*, \alpha^{-k}\lambda^*)(k = 1, 2, ...)$, the regulated traffic is controlled to the interval $[\alpha\lambda^*, \lambda^*)$. The permission rate in each interval is also shown on the x-axis. This figure shows that the proposed control scheme achieves high throughput by regulating transmission adaptively corresponding to the offered traffic in the range where it exceeds $1/h = 1$. Because this figure only shows the statistical characteristics, the transient behavior when the offered traffic varies dynamically in time is verified and discussed in the next section.

2.3.2 ICMA/CD

In this subsection, we apply the proposed congestion control scheme to the ICMA/CD protocol and analyze control parameters and stationary characteristics. The ICMA/CD protocol is the basis for ICMA/PE (ICMA with Partial Echo), the random access protocol used in the PDC (Personal Digital Cellular) system [5] in Japan. This is a centrally controlled multiple access protocol in which the BS broadcasts idle/busy information about the reverse channel states instead of MTs sensing the carrier transmitted from other MTs as in the CSMA (carrier sense multiple access) protocol.

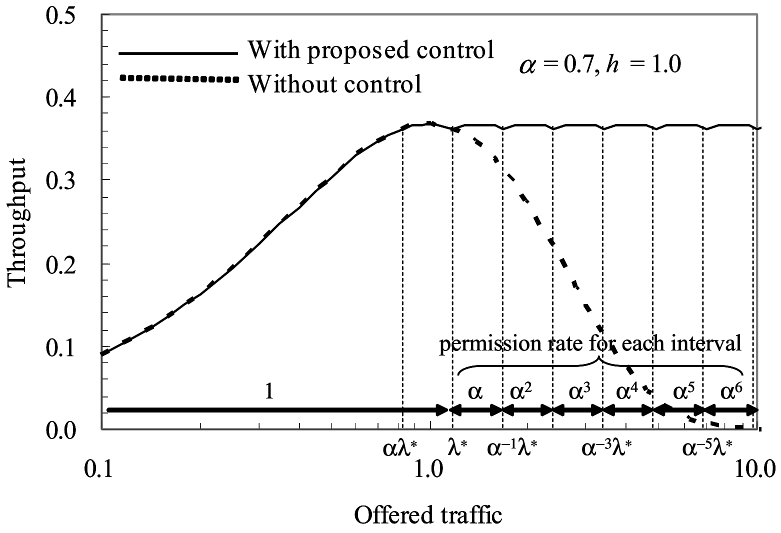


Figure 2.6: Statistical throughput for slotted ALOHA

Outline of protocol

The slotted non-persistent ICMA/CD protocol proposed in [11] is used as an example here. The time charts of successful and unsuccessful signal transmissions in this protocol are illustrated in Fig. 2.7. In this protocol, both reverse and forward channels are slotted with the transmission time of the idle/busy signal. The BS broadcasts a series of *busy signals* (B in Fig. 2.7) on the forward channel while receiving reverse channel signals. If the BS detects an reverse channel signal collision, it changes the busy signal to a *stop signal* (S in Fig. 2.7), and the MTs receiving stop signals abort their reverse channel signal transmissions. The station transmits a series of *idle signals* (I in Fig. 2.7) and announces that the reverse channel is idle unless it is sending busy or stop signals. The *reverse channel signal recognition delay D* (see Fig. 2.7) is the time interval between the start of signal transmission and the changeover point of the control signal from idle to busy in the forward control channel. In the sequence, we assume that the reverse channel signal transmission time for successful and unsuccessful transmission are fixed values T and γ, respectively.

Throughput without control

Throughput of the non-persistent ICMA/CD protocol is expressed as follows as defined in [11],

$$S(x) = \frac{P_s T}{\bar{T}_i + \bar{T}_p}, \tag{2.18}$$

where P_s, $\bar{T}_i$, and $\bar{T}_p$ represent the probability of successful transmission, the average inhibited period, and the average permitted period, respectively. In the following, we

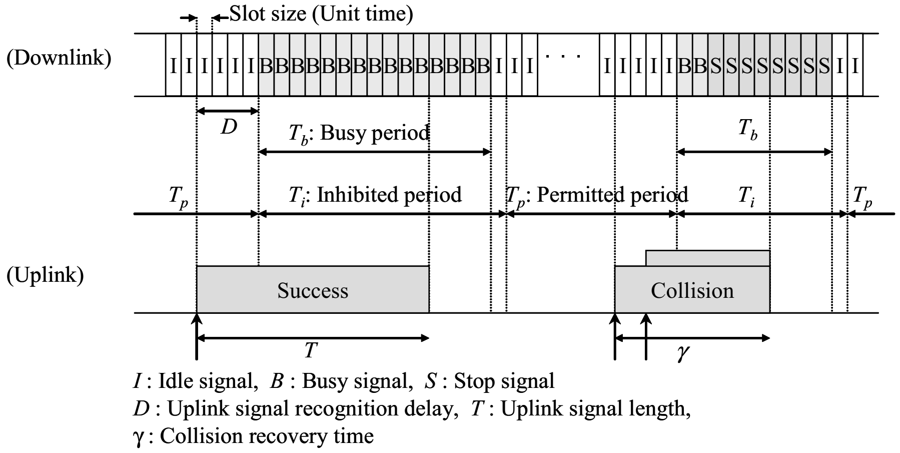

Slot size (Unit time)
(Downlink)
I I I I I I BBBBBBBBBBBBBBB I I I · · · I I I I I BBSSSSSSSSS I I
D
T_b: Busy period
T_b
T_p
T_i: Inhibited period
T_p: Permitted period
T_i
T_p
(Uplink)
Success
Collision
T
γ
I : Idle signal, B : Busy signal, S : Stop signal
D : Uplink signal recognition delay, T : Uplink signal length,
γ : Collision recovery time

Figure 2.7: Time charts of downlink and uplink control channels of ICMA/CD

derive throughput expressed in terms of offered traffic x since Eq. 2.2 in [11] is not correct. The probability of successful transmission is given by

$$P_s = \frac{q_1}{1-q_0} q_0^D, \tag{2.19}$$

where q_i represents the probability that i signals originate in a slot, $q_0 = e^{-x}$, and $q_1 = xe^{-x}$, assuming a Poisson process for the input traffic. The inhibited period is $T+1$ on successful transmission, $\gamma+1$ when more than one packet originates in a slot, and $\gamma+1+i$ when more than one packet originates over $i+1$ slots $(1 \leq i \leq D)$. The average inhibited period $\bar{T}_i$ and the average permitted period $\bar{T}_p$ are given by

$$\begin{aligned}
\bar{T}_i &= P_s(T+1) + \frac{1-q_0-q_1}{1-q_0} q_0^D(\gamma+1) + (1-q_0)\sum_{i=0}^{D-1} q_0^i(\gamma+1+D-i) \\
&= \frac{q_1}{1-q_0} q_0^D(T-\gamma) + q_0^D(\gamma+1) + (1-q_0)\sum_{i=0}^{D-1} q_0^i(\gamma+1+D-i) \\
&= 1+D+\gamma+\frac{\{q_0+(T-\gamma)q_1\}q_0^D - q_0}{1-q_0} \quad \text{and}
\end{aligned} \tag{2.20}$$

$$\bar{T}_p = \frac{q_0}{1-q_0} + D, \tag{2.21}$$

respectively. Substituting Eqs. 2.19–2.21 into Eq. 2.18, we have the throughput in terms of offered traffic x,

$$\begin{aligned}
S(x) &= \frac{q_0^D q_1 T}{\{q_0+(T-\gamma)q_1\}q_0^D+(1+2D+\gamma)(1-q_0)} \\
&= \frac{Txe^{-(1+D)x}}{\{1+(T-\gamma)x\}e^{-(1+D)x}+(1+2D+\gamma)(1-e^{-x})}.
\end{aligned} \tag{2.22}$$

Unsuccessful transmission rate

Next, we derive the unsuccessful transmission rate of the ICMA/CD protocol. As shown in Eq. 2.1, the unsuccessful transmission rate in terms of offered traffic x is defined by

$$r(x) = \frac{\Pr\{\text{Collision}\}}{\Pr\{\text{Success or Collision}\}} = 1 - \frac{\Pr\{\text{Success}\}}{\Pr\{\text{busy or stop}\}}. \tag{2.23}$$

where Pr{Success} involves throughput itself given by Eq. 2.22. The denominator on the right hand side of Eq. 2.23, Pr{busy or stop} indicates the probability that busy or stop signals are transmitted in the forward channel. Let the busy period be the interval with which these signals are transmitted in series. Then the length of this period is just one slot shorter than the inhibited period. Therefore, the average busy period $\bar{T}_b$ is given by

$$\bar{T}_b = \bar{T}_i - 1. \tag{2.24}$$

Thus, we have

$$\Pr\{\text{busy or stop}\} = \frac{\bar{T}_b}{\bar{T}_i + \bar{T}_p} = \frac{\bar{T}_i - 1}{\bar{T}_i + \bar{T}_p}. \tag{2.25}$$

From Eqs. 2.22, 2.23, and 2.25, we have the unsuccessful transmission rate,

$$\begin{aligned} r(x) &= 1 - \frac{P_s T}{\bar{T}_i - 1} \\ &= 1 - \frac{q_0^D q_1 T}{\{q_0 + (T-\gamma)q_1\}q_0^D + (1+D+\gamma)(1-q_0) - 1} \\ &= 1 - \frac{Txe^{-(1+D)x}}{\{1+(T-\gamma)x\}e^{-(1+D)x} + (1+D+\gamma)(1-e^{-x}) - 1}. \end{aligned} \tag{2.26}$$

Target traffic

The average throughput $f(\lambda)$ where the offered traffic extends from $\alpha\lambda$ to λ is given by

$$f(\lambda) = \frac{1}{(1-\alpha)\lambda} \int_{\alpha\lambda}^{\lambda} \frac{Txe^{-(1+D)x}}{\{1+(T-\gamma)x\}e^{-(1+D)x} + (1+2D+\gamma)(1-e^{-x})} dx. \tag{2.27}$$

In contrast to the slotted ALOHA case, the integral part of the above equation cannot be expressed explicitly; however, λ^* that maximizes $f(\lambda)$ can be derived numerically. The target traffic $\lambda^* T$ normalized by signal transmission time T as a function of α for $T = 20, D = 4$, and $\gamma = 4,8,12,20$ is plotted in Fig. 2.8. Note that the case of $\gamma = T = 20$ represents ICMA protocol without collision detection. Similar to the slotted ALOHA case, the target traffic $\lambda^* T$ is a monotonically decreasing function of α,

Control thresholds

From Eq. 2.26, the control thresholds $R^{(k)}$ are given by

$$R^{(k)} = 1 - \frac{T\alpha^{1-k}\lambda^* e^{-(1+D)\alpha^{1-k}\lambda^*}}{\{1+(T-\gamma)\alpha^{1-k}\lambda^*\}e^{-(1+D)\alpha^{1-k}\lambda^*} + (1+D+\gamma)(1-e^{-\alpha^{1-k}\lambda^*}) - 1} \tag{2.28}$$

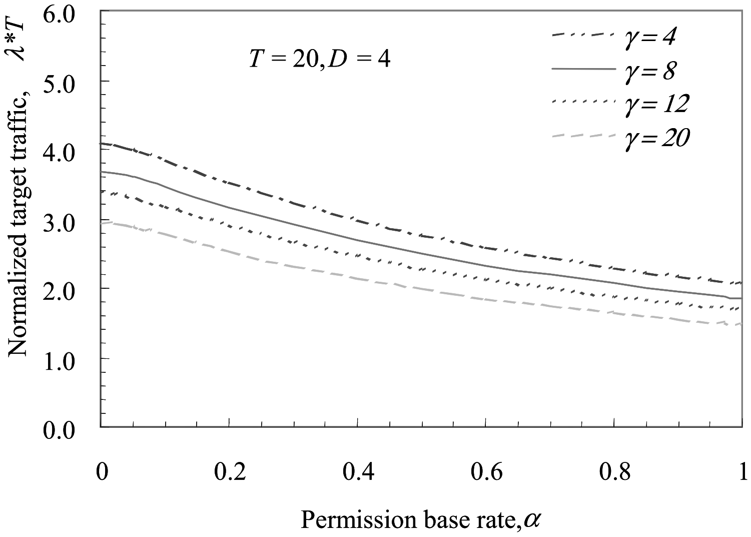


Figure 2.8: Target traffic for ICMA/CD

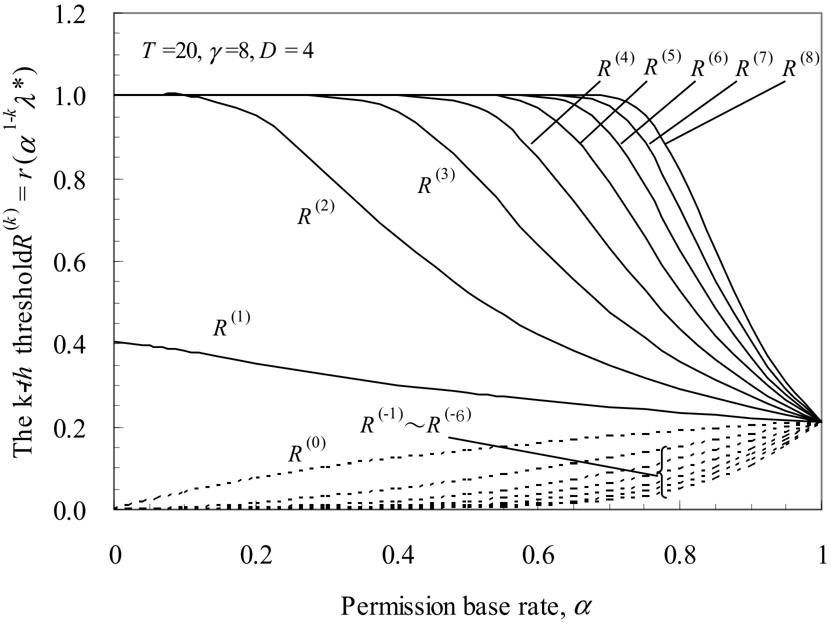


Figure 2.9: Control thresholds for ICMA/CD

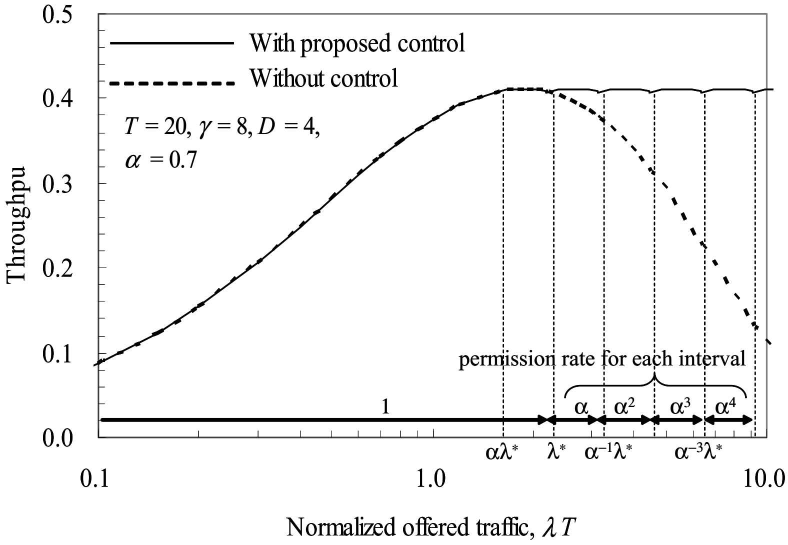


Figure 2.10: Statistical throughput for ICMA/CD

The thresholds, $R^{(k)} (k = -6, -5, \ldots, 8)$ for $T = 20$, $\gamma = 8$, and $D = 4$ as functions of the permission base rate α are plotted in Fig. 2.9. As $\alpha \to 1$, the thresholds become a constant for all k, as follows,

$$R^{(k)} = r(\lambda_1), \tag{2.29}$$

where λ_1 indicates the offered traffic that maximizes throughput, that is, it satisfies $S'(\lambda_1) = 0$. Hence, it is given by the root of the following equation:

$$1 - (1 + 2D + \gamma)\{1 - D\lambda_1 - (1 - \lambda_1 - D\lambda_1)e^{\lambda_1}\}e^{D\lambda_1} = 0. \tag{2.30}$$

Note that λ_1 and $R^{(k)}$ for $\alpha \to 1$ are independent of T. From Fig. 2.9, as in the case of slotted ALOHA, α should be set in the range $0.7 \leq \alpha \leq 0.9$, which enables reasonable adaptive control while limiting the amount of processing at the MT.

Throughput characteristics

The statistical throughput characteristics for $\alpha = 0.7$,$T = 20$,$\gamma = 8$, and $D = 4$ are shown in Fig. 2.10. In this figure, the dotted and solid lines represent the throughput without and with control, respectively. The permission rate in each interval is also shown on the x-axis. Similar to the slotted ALOHA case, it can be seen that the proposed control scheme achieves a high throughput by regulating transmission adaptively.

2.4 Transient characteristics

Since one of the main features of the proposed scheme is adaptability to drastic changes in offered traffic, we simulated transient characteristics of the proposed scheme for time-

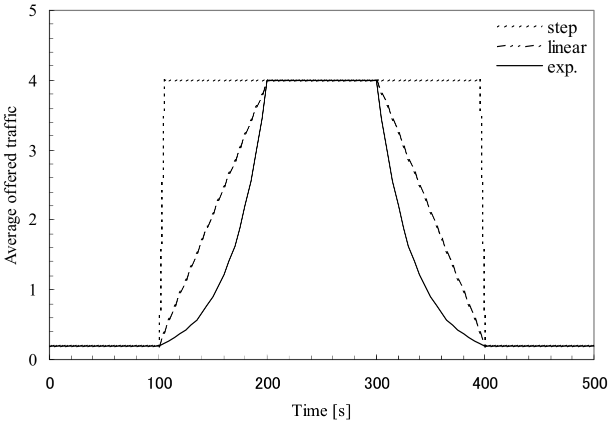


Figure 2.11: Time-variant input models of offered traffic

variant-offered traffic.

2.4.1 Simulation conditions

We used slotted ALOHA with slot length, $h = 5$ *ms* as a random access protocol. This protocol and slot length are used in the reverse control channel of the personal handy phone system (PHS) [5] in Japan. Time-variant input models of average offered traffic are illustrated in Fig. 2.11, where the vertical axis represents the average offered traffic normalized by the slot length. This means that random access signals are generated according to a Poisson process with rate $\lambda(t)/h$ as shown in Fig. 2.11. We evaluate three types of input variation models where the average offered traffic for normal and overloaded states are set to 0.2 and 4.0 *signal/slot*, respectively.

(a) Step type: The average offered traffic rapidly becomes overloaded according to a step function at time 100 *s*. After continuing in the overloaded state, it drops rapidly to the normal load at time 400 *s*.

(b) Linear type: The average offered traffic increases linearly from 100 *s* to 200 *s*. After continuing in the overloaded state until 300 *s*, it decreases linearly to the normal traffic until 400 *s*.

(c) Exponential type: The average offered traffic increases exponentially from 100 *s* to 200 *s*. After continuing in the overloaded state until 300 *s*, it decreases exponentially to the normal traffic volume until 400 *s*.

In addition, we use the following parameters in the simulations.

- $\alpha = 0.8$ (in this case, the target traffic normalized by the slot length is given by $\lambda^* h = 1.1111$),
- measurement interval: $m = 1000$ *slots* ($= 5$ *s*).

2.4.2 Simulation results

Simulation results for the transient characteristics of the average throughput in the measurement interval for the above three input models, (a)–(c), are shown in Figs. 2.12(a)–(c), respectively. In these figures, solid and broken lines, respectively, represent the characteristics with and without the proposed control. We can see from these figures that lack of control degrades throughput during an overload for each input model. For instance, the average throughput is restricted below 0.1 when the average offered traffic exceeds 1.0. On the other hand, even in the overloaded state for all input models, the proposed control achieves nearly theoretically optimal throughput given by

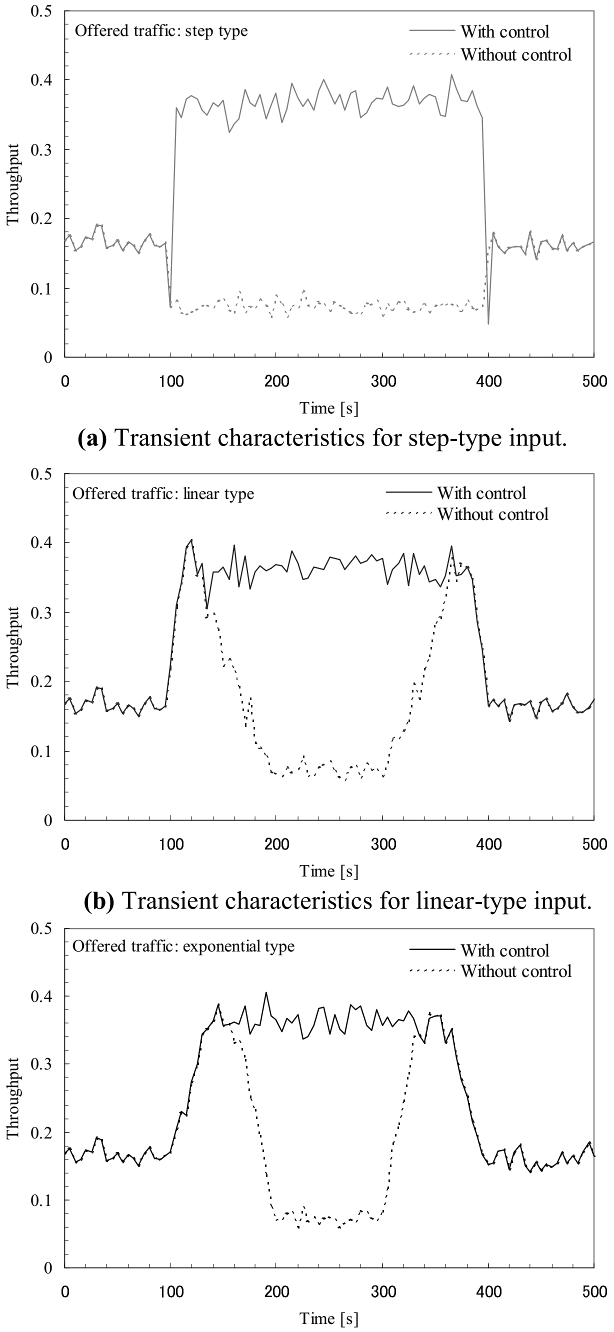
$$f(\lambda^*) = \frac{(1+\alpha h\lambda^*)e^{-\alpha h\lambda^*} - (1+h\lambda^*)e^{-h\lambda^*}}{(1-\alpha)h^2\lambda^*} = 0.3671 \tag{2.31}$$

In addition, we can see that adaptive control is achieved by the proposed scheme when there is a drastic variation in offered traffic. There are three points of instantaneous surges of throughput degradations when the offered traffic steeply crosses 1.0, that is, at 100 s and 400 s in Fig. 2.12 (a) and 140 s in Fig. 2.12 (b). These surges are caused by a control delay in the offered traffic variation, i.e., steep increases of the offered traffic at 100 s in Fig. 2.12 (a) and 140 s in Fig. 2.12 (b), and sharp decline of the offered traffic at 400 s in Fig. 2.12 (a). These degradations continue, however, only for one measurement interval 5 s in these simulations, and after that interval, nearly optimal throughput is achieved in all cases. To prevent these surges, it might be effective to set the measurement interval shorter or to implement some method of forecasting offered traffic trends. The problems in setting an appropriate measurement interval and developing forecasting mechanisms are for further study.

2.5 Conclusion

In this chapter, we proposed a congestion control scheme for random access channels in cellular systems. Its main features are scalability for handling increasing numbers of MTs and adaptability for coping with drastic changes in traffic load. These are achieved by controlling the traffic load adaptively so that the offered traffic under regulation settles down at the interval $[\alpha\lambda^*, \lambda^*)$, where λ^* is the target traffic for a given permission base rate α that maximizes the average throughput when the offered traffic is within the above interval. A method of measuring and estimating channel-offered traffic using the unsuccessful transmission rate was also proposed. Based on the unsuccessful transmission rate, control thresholds that maximize the average throughput were analytically derived, and the adaptive controlling algorithm was formulated. The measurement method can be achieved using only two counters for the idle and successful slots. Additionally, the algorithm has a simple form and requires no threshold table. From these properties, we can conclude that the proposed scheme is easy to implement and practical.

We applied the scheme to both the slotted ALOHA and ICMA/CD protocols. For each protocol, three control parameters —the unsuccessful transmission rate, target traffic, and control thresholds— were analytically derived. Then stationary throughput characteristics were numerically evaluated. We found that the scheme could achieve high throughput by regulating transmission adaptively depending on the offered traffic. The



(a) Transient characteristics for step-type input.

(b) Transient characteristics for linear-type input.

(c) Transient characteristics for exponential-type input.

Figure 2.12: Transient characteristics for time-variant input models

preferred range of α that enables adaptive control and restrains processing at MTs was also clarified.

The transient throughput characteristics with three types of time-variant input models were simulated. The simulation results indicated that the control scheme achieved nearly theoretically optimal throughput even during an overload for any input model. In addition, the adaptability of the control scheme was shown by its ability to handle drastic variations in traffic load.

Several issues remain for further study.

One of the most important is a detailed evaluation of the influence of a short retransmission interval. Throughout the chapter, the channel's offered traffic for both new and retransmitted signals was assumed to obey a Poisson process. In general, random access channels often exhibit unstable behaviors caused by retransmissions of signals with short interval. The unstable behavior, however, can be overcome by applying the proposed control scheme, because the offered traffic is regulated for both new and retransmitted signals. In a practical situation, it will be necessary to evaluate the influence of a short retransmission interval on the transient characteristics under regulation.

A second issue is an optimal control combining the random access channel and dedicated traffic channels. In the current cellular systems, there are two kinds of radio access channels, those are, a common control channel based on a random access protocol on which control signals (e.g., a connection setup signal) are transmitted, and a dedicated traffic channel for each MT on which user information is transmitted after connection setup. In this radio channel structure, the control scheme could be applied to the common control channel. For practical use in this situation, the optimal control taking into account the balance in the capacities of the common control and dedicated transmission channels would be important.

A third important issue is a control scheme for the multi-priority classes of access signals. This is an important problem from the viewpoint of guaranteeing urgent traffic priority during congestion. The scheme proposed in this chapter can probably be extended to solve these problems.

Chapter 3

Traffic design for random access schemes with access inhibition and retransmission control

3.1 Introduction

This chapter presents a traffic design method that enables comparative evaluations of various random access schemes for radio control channels with access inhibition and retransmission control. In cellular systems, a reverse control channel is shared by mobile terminals (MTs) on which control signals, such as call setup signals and response signals, are transmitted.

To prevent collisions among response signals and call setup signals from calling users and to improve the throughput of the reverse control channel, various random access schemes have been proposed [105, 106, 107]. Most of them are improved from simple well-known schemes, such as ALOHA and CSMA/CD. The main difference between simple random access schemes and sophisticated random access schemes is in the inclusion of *access inhibition control* using a forward control channel and *retransmission control* that restricts retransmission intervals and the number of retransmissions. The access inhibition control is used to prevent collisions among response signals from called users and call setup signals from calling users. Under this control, the base station (BS) transmits inhibition signals (*busy signals*) immediately after the transmission of an incoming call setup signal through the forward control channel. The busy signals inhibit transmissions of call setup signals from all users in the cell. Retransmission control is used to reduce call setup delays. Under this control, the retransmission interval for a call setup signal is set to a short interval, and the number of retransmissions is limited.

For traffic analyses of the channel throughput and the mean signal transfer delay in control channel access schemes and random access protocols, several approximation methods have been proposed.

Among those, the S-G analysis [108] is a method frequently used in traffic analyses of various random access protocols including ones having control channel access schemes with access inhibition control [109]. In this method, the signal generation process of call setup signals including ones for retransmissions is approximated as a Poisson process. This approximation makes the modeling and the analysis easy, but it hinders some detailed analyses, for example, it is difficult to make a performance evaluation with the retransmission interval as a parameter.

By contrast, the equilibrium point analysis [26, 25, 24] and the diffusion approximation analysis [27, 28, 29, 30] are methods that can evaluate the effect of the retransmission interval on the performance. The equilibrium point analysis evaluates the system performance considering the equilibrium point of the stochastic process based on the fluid approximation. Because of its ease of analysis, it is applied to evaluations of various random access protocols. But it captures the system performance only at the equilibrium point. The diffusion approximation analysis is a method that approximates the state transition of the system by a diffusion process. It can provide a more accurate performance evaluation than the equilibrium point analysis. However, it becomes more complex as the number of states gets larger. Consequently, as far as the author knows, at the time the original article of this chapter was written, it has been used only in the evaluation of simple random access protocols such as ALOHA and CSMA/CD, and there is no report which applies the method to the traffic analysis of a complex system such as the control channel access scheme with access inhibition control.

This chapter proposes an approximation method that integrates the S-G analysis and the diffusion approximation analysis for control channel access schemes in cellular systems. This method can be applied to various control channel access schemes such as those employing the ALOHA [105] and the ICMA (idle signal casting multiple access) [106] protocols as the random access protocol, taking account of the effects of the access inhibition control and the retransmission controls on the performance. The accuracy of approximation is examined and the traffic characteristics of these schemes are evaluated.

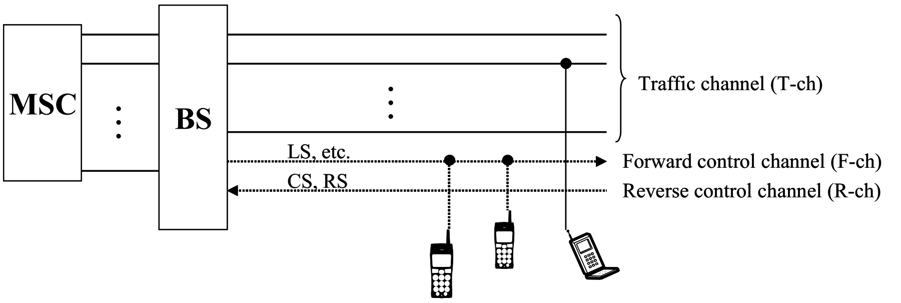
The reminder of this chapter is organized as follows. In the next section, the structure and the operation of control channel access schemes in cellular systems are described in detail. In Section 3.3, a traffic model is presented based on the state of the mobile user. In Section 3.4, our approximation method is proposed and applied to the cases where ALOHA and ICMA protocols are employed as the random access protocol. In Section 3.5, numerical results by the proposed method are compared with simulation results for the throughput and the average number of retransmissions in those schemes. In Section 3.6, the chapter is concluded with some comments.

3.2 Control channel access scheme

This section describes a cellular system employing a control channel considered here and presents the channel configuration, operations of MTs and BSs and signals sent from them.

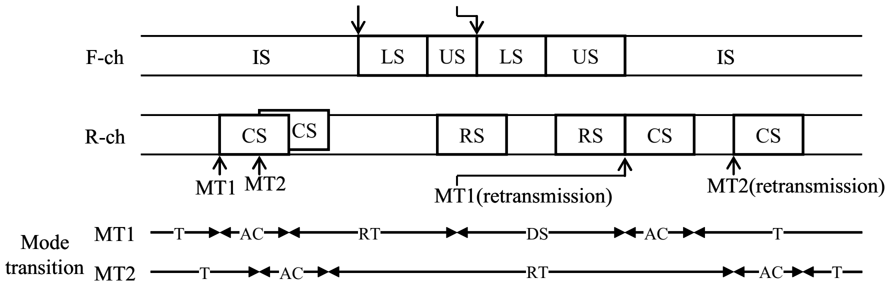
3.2.1 Structure of cellular systems

Figure 3.1 shows a typical radio channel structure in a cellular system with a control channel. Three kinds of channels are set between the BS and the MTs: the traffic channel (T-ch); the forward control channel (F-ch); and the reverse control channel (R-ch). The F-ch is used for transmission of control information from the BS to all MTs in the cell. For example, incoming call setup signals (LSs) for called mobile users are sent through


MSC
BS
Traffic channel (T-ch)
LS, etc.
CS, RS
Forward control channel (F-ch)
Reverse control channel (R-ch)

MSC: Mobile switching center, BS: Base station, MT: Mobile terminal
LS: Incoming call-setup signal, CS: Call-setup signal from calling MT, RS: Response signal to LS

Figure 3.1: Radio channel structure in cellular systems



LS: Incoming call-setup signal, CS: Call-setup signal from calling MT,
RS: Response signal to LS, IS: Idle signal, US: Busy signal

T mode: having no signal to be transmitted, DS mode: sensing the F-ch,
AC mode: transmitting a signal, RT mode: waiting for retransmission

Figure 3.2: Control channel operations for ALOHA access scheme

the F-ch. The R-ch is used when MTs send control signals to the BS. For example, call setup signals (CSs) from calling users in the cell, the response signals (RSs) to LSs and location registration signals are sent through the R-ch. Note that, after the completion of a call setup, the T-ch is assigned and dedicated to the user, and control channels are not used any more.

3.2.2 Access inhibition control

In order to suppress throughput decreases caused by collisions of signals on the R-ch, an access inhibition control is introduced. To avoid collisions between CSs, a random access scheme is applied in the access protocol of CS (called CS access protocol). Collisions between RSs are avoided by transmitting consecutive LSs during a certain interval. Collisions between CSs and RSs are avoided by inhibiting the transmission of CSs from MTs by transmitting busy signals (USs) through the F-ch. The traffic control involving the above functions is called *an access inhibition control.* The outline of operations of the control channel access scheme where ALOHA is used as the base CS access protocol (ALOHA access scheme) is shown in Fig. 3.2. Also that where ICMA is used as the base CS access protocol (ICMA access scheme) is shown in Fig. 3.3.

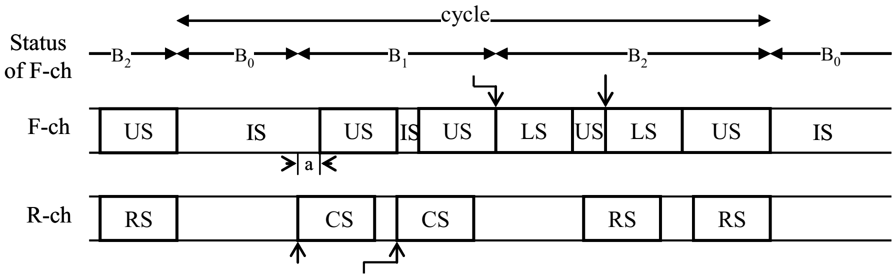


Figure 3.3: Control channel operations for ICMA access scheme

3.2.3 Operations of MTs and BS

Here we summarize operations of the MTs and the BS under the access inhibition protocol.

Operations of MTs

(1) An MT having a new CS to be sent senses the F-ch to check the possibility of signal transmission.

(2) If an idle signal (IS) is detected on the F-ch, the MT transmits the CS to the R-ch. If an LS or US is detected (access inhibition), the MT continues sensing of the F-ch until an IS is detected. If it detects an IS, it transmits the CS immediately (1-persistent protocol).

(3) When the CS collides with a signal from another MT, the MT waits for a certain time (retransmission interval) to retransmit the CS. The waiting time from the end of the previous signal transmission is subjecting to an exponential distribution. Then operation (2) is iterated. When the number of iterations exceeds the limit of retransmissions, the call is blocked and cleared. This operation (3) is called *retransmission control*.

(4) If an MT receives an LS through the F-ch, it sends out an RS immediately through the R-ch.

(5) After the completion of the transmission of the CS or the RS, a T-ch is assigned and dedicated to the MT. The transmission and reception of traffic information is then made through T-ch.

Operation of BS

The BS has a transmission buffer for LSs and operates as follows.

(1) When there is no need to send an LS or a US, the BS always sends an IS to the F-ch,

(2) When the BS receives an LS during transmission of ISs, the BS stops sending IS and sends an LS immediately.

(3) To avoid collisions among RSs, the BS sends LSs at a certain interval (LS transmission interval). The length of the LS transmission interval h_D is given so that $h_D \geq max(h_2, h_3)$, where h_2 is the LS length and h_3 is the RS length.

(4) To prevent transmission of CSs from other MTs during transmission of RSs, the BS sends USs following the LS, until the end of the RS reception. When the BS receives another LS during the transmission of USs, the BS stops sending US and sends the LS immediately. To prevent the collision between CSs and RS, we request $h_1 \leq h_2$ where h_1 is the CS length.

(5) In the ICMA access protocol, the BS detecting a CS replaces ISs with USs to prevent transmission of CSs from other MTs.

3.3 Traffic model

In this section, we construct a system model for the control channel access scheme described in the previous section considering transitions of the states of MTs. The following assumptions are made in the modeling.

1. The number of MTs is M.
2. Each MT generates CSs according to a Poisson process with rate λ_1 whenever it can generate CSs. LSs arrive at the BS according to a Poisson process with rate λ_2.
3. Retransmission intervals of CSs are subjecting to an exponential distribution with mean $1/\sigma$. Transmission delays of all signals are negligible compared with transmission times of signals.
4. The limit of the number of retransmissions is set as N.

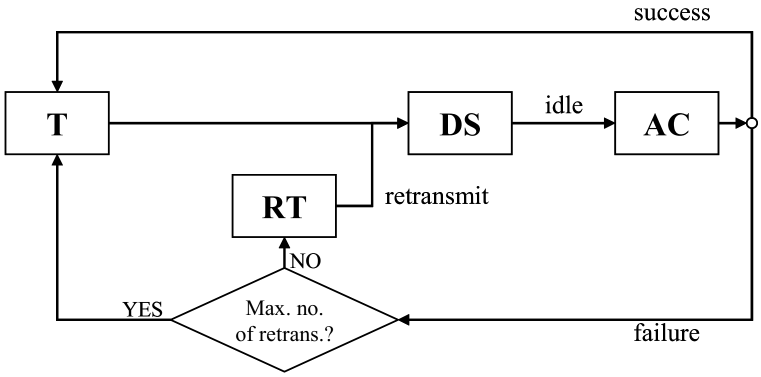
The state of an MT is classified into the following four modes (Figs. 3.2 and 3.3).

- T mode: having no signal to be transmitted.
- DS mode: sensing the F-ch.
- AC mode: transmitting a signal.
- RT mode: waiting for retransmission.

From the end of a signal transmission success to the end of speech is regarded as T mode.

Fig. 3.4 designates the state transition diagram of our model. To simplify the analysis, however, we modify the model as in Fig. 3.5, according to the method in [27], by dividing the T mode into two states (VT and TR modes). We assume that the sojourn times at the TR mode are subjecting to the same distribution as that at the RT mode. Then the TR mode can be merged with the RT mode to be represented as a single mode. Similar modifications have been used in models for analyzing the ALOHA and CSMA protocols [26, 25, 24, 27]. The mean sojourn time V of an MT in the VT mode is given by

$$E(V) \equiv 1/\lambda_v = 1/\lambda_1 - 1/\sigma. \tag{3.1}$$



T mode: having no signal to be transmitted, DS mode: sensing the F-ch, AC mode: transmitting a signal, RT mode: waiting for retransmission

Figure 3.4: State transition diagram of mobile terminal

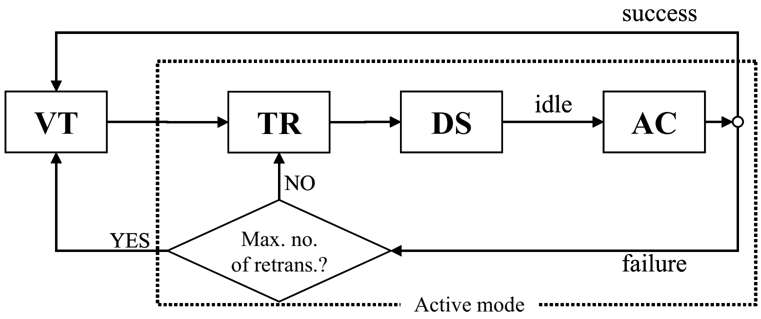


Figure 3.5: Modified state transition diagram of mobile terminal

3.4 Approximate analysis

All states other than the VT mode are called active modes. In this section, we propose an approximate method for obtaining the steady-state probabilities for the number of MTs in the active mode (number of active MTs), and derive approximate formulas for performance measures such as the throughput of the R-ch and the average number of retransmissions of a CS.

3.4.1 Analysis procedure

The approximate method proposed here is constructed in the following steps. In Step 1, the method uses the S-G analysis [108] which gives a relation between the conditional throughput $S(n)$ and the conditional mean generation rate of new signals and retransmitted signals $G(n)$ given the number of active MTs n. The relationship depends on the CS access protocol, and examples of this relationship for ALOHA and ICMA access schemes will be given in subsection 3.4.3.

Note that the conditional throughput $S(n)$ represents the average number of MTs per unit time which succeed in transmission and make transition to the VT mode. In Step 2, the conditional loss rate $L(n)$, the rate at which the number of retransmissions exceeds the limit N to result in call loss, is derived using the conditional throughput $S(n)$ and the generation rate $G(n)$. Note that the conditional loss rate $L(n)$ represents the average number of MTs per unit time which result in call loss and make transition to the VT mode. In Step 3, the diffusion parameters β_n and α_n are given using the conditional generation rate $G(n)$, throughput $S(n)$, and loss rate $L(n)$, as derived by the S-G analysis in Steps 1 and 2. In Step 4, using the diffusion approximation analysis, the stationary state probabilities π_n for the number of active MTs are derived. Finally in Step 5, using the stationary state probabilities π_n, the performance measures, such as the mean connection delay and the mean number of retransmissions, are derived.

The detailed description of the proposed approximation method is given in the following steps.

Step 1. For a given n, the number of active MTs, the conditional throughput $S(n)$ is determined by the S-G analysis from $G(n)$, the conditional mean generation rate of new signals and retransmitted signals at the TR mode. The number of MTs sensing the F-ch in the DS mode and the number of MTs in signal transmission in the AC mode are less in general compared with the number of MTs in the TR mode [110]. Consequently, the average number of MTs making transitions from the TR mode to the DS mode per unit time can be approximated as

$$G(n) \approx n\sigma. \tag{3.2}$$

It corresponds to the traffic G in the S-G analysis assuming an infinite number of MTs. In the following, the conditional throughput $S(n)$ is derived by the S-G analysis as a function of the rate $G(n)$.

Step 2. The relationships of the conditional throughput $S(n)$ to the collision probability $R(n)$ of CS and to the loss rate $L(n)$, the rate at which the number of retransmissions exceeds the limit N to result in call loss, are given as

$$\begin{aligned} R(n) &= 1 - \frac{S(n)}{G(n)}, & (3.3) \\ L(n) &= \frac{1 - R(n)}{1 - R(n)^{N+1}} R(n)^{N+1} G(n). & (3.4) \end{aligned}$$

Step 3. Using the transition probability $\lambda_a(n)$ from the VT mode to the active mode and the transition probability $\lambda_d(n)$ from the active mode to the VT mode, as determined by the S-G analysis, the diffusion parameters β_n and α_n are given by

$$\begin{aligned} \beta_n &= \lambda_a(n) - \lambda_d(n) \quad \text{and} & (3.5) \\ \alpha_n &= \lambda_a(n) + \lambda_d(n), & (3.6) \end{aligned}$$

where

$$\begin{aligned} \lambda_a(n) &= (M - n)\lambda_v \quad \text{and} & (3.7) \\ \lambda_d(n) &= S(n) + L(n). & (3.8) \end{aligned}$$

Step 4. Using the diffusion approximation analysis, the stationary state probabilities π_n $(n = 0, \cdots, M)$ for the number of active MTs are derived as (see Appendix 3.1)

$$\pi_1 = \begin{cases} \dfrac{h_0\pi_0}{\beta_1}\left[\dfrac{\alpha_1}{2\beta_1}\left\{\exp\left(\dfrac{2\beta_1}{\alpha_1}\right) - 1\right\} - 1\right], & \beta_1 \neq 0, \\ \dfrac{h_0\pi_0}{\alpha_1}, & \beta_1 = 0, \end{cases} \tag{3.9}$$

$$\pi_k = \begin{cases} \dfrac{\alpha_k\pi_0}{2\beta_k}(\theta_k - \theta_{k-1}), & \beta_k \neq 0, \\ \pi_0\theta_k, & \beta_k = 0, \end{cases} \quad k = 2, \cdots, M, \tag{3.10}$$

where

$$h_0^{-1} = 1/(M\lambda_v), \tag{3.11}$$

$$\theta_1 = \frac{h_0}{\beta_1}\left\{\exp\left(\frac{2\beta_1}{\alpha_1}\right) - 1\right\}, \quad \text{and} \tag{3.12}$$

$$\theta_{k+1} = \theta_k \exp\left(\frac{2\beta_{k+1}}{\alpha_{k+1}}\right), \quad k = 1, 2, \cdots, M-1. \tag{3.13}$$

Here π_0 in Eqs. 3.9 and 3.10 is determined from the normalization condition $\sum_{i=0}^{M}\pi_i = 1$.

Step 5. Using the stationary state probabilities π_n, the performance measures are derived as follows.

$$\text{throughput:} \qquad S = \sum_{k=0}^{M} S(k)\pi_k \tag{3.14}$$

$$\text{mean number of active MTs:} \qquad Q = \sum_{k=0}^{M} k\pi_k \tag{3.15}$$

$$\text{mean connection delay:} \qquad D = \frac{Q}{S} - \frac{1}{\sigma} \quad (N = \infty) \tag{3.16}$$

$$\text{control failure rate:} \qquad L = \sum_{k=0}^{M} \{R(k)\}^{N+1}\pi_k \tag{3.17}$$

$$\text{mean number of retransmissions:} \qquad R_t = \sum_{k=0}^{M} \left\{\frac{1 - R(k)^{N+1}}{1 - R(k)} R(k)\right\}\pi_k \tag{3.18}$$

3.4.2 Features of the approximate analysis

The approximation analysis described above has the following features.

(1) The signal generation process including retransmission (generation process of signals from MTs in the TR mode) is approximated by a Poisson process. This is the basic assumption for applying the S-G analysis in Step 1. This approximation is simple, and

the method can be applied to many random access schemes including ones with access inhibition control using the F-ch [109].

(2) By the use of the diffusion approximation analysis (steps 3 and 4), not only static characteristics derived by the S-G analysis, but also dynamic characteristics of the system, i.e., behaviors taking into account the system's instability [24], can be evaluated. Compared with the equilibrium point analysis [26, 25, 24, 27], the diffusion approximation analysis is able to analyze the dynamic characteristics considering the stochastic fluctuation around the equilibrium point.

(3) By the use of the basic return boundary and the discrete diffusion parameters in the diffusion approximation in step 4, the steady-state probabilities for the number of active MTs can be evaluated by a recurrence formula. This makes the analysis much easier (Appendix 3.1).

3.4.3 Examples of the analysis

Using the proposed analytical method, various access control schemes with different CS access protocols can be evaluated. In the analysis, the relationship between the conditional throughput $S(n)$ and the generation rate $G(n)$ in step 1 depends on the CS access protocol. In the following, as examples, this relationship is derived for ALOHA and ICMA access schemes considering the effect of CS inhibition control.

ALOHA access scheme (Fig. 3.2)

In contrast to the simple ALOHA protocol considered by Abramson et al. [23], this scheme has an access inhibition (busy) state due to LS. Therefore, the collision probability for CS depends on the state of the F-ch at the signal generation. The state of the F-ch can be divided into four modes:

A_1 : in the idle period,

A_2 : in the period of length h_1 immediately before a busy period,

A_3 : in the busy period, and

A_4 : in the period of length h_1 immediately after a busy period.

Let P_i be the probability that an arbitrarily chosen CS is generated in mode i, and call it the signal generation probability. Further let R_i be the probability that a CS generated in mode i collides with another CS, and call it the collision probability. By considering the range of collision, that is the range where a collision occurs if another signal is generated [109], these probabilities are given as follows.

A_1 : in the idle period,

$$P_1 = 1 - \sum_{i=2}^{4} P_i, \tag{3.19}$$

$$R_1 = 1 - \exp(-2h_1 G(n)). \tag{3.20}$$

A$_2$: in the period immediately before a busy period,

$$\begin{aligned} P_2 &= h_1 P_3/\bar{T}_B, && (3.21)\\ R_2 &= 1-[1-\exp(-h_1G(n))]\exp(-h_1G(n))/(h_1G(n)). && (3.22)\end{aligned}$$

A$_3$: in the busy period,

$$\begin{aligned} P_3 &= \bar{T}_B/(\bar{T}_B+1/\lambda_2), && (3.23)\\ R_3 &= 1-\int_0^\infty \exp(-(t+h_1)G(n))dF(t) &&\\ &= 1-\exp(-h_1G(n))F^*(G(n)). && (3.24)\end{aligned}$$

A$_4$: in the period immediately after a busy period,

$$\begin{aligned} P_4 &= h_1P_3/\bar{T}_B, && (3.25)\\ R_4 &= 1-\int_0^\infty\int_0^{h_1} \exp(-(t+h_1+u)G(n))/h_1 du dF(t) &&\\ &= 1-[1-\exp(-h_1G(n))]\exp(-h_1G(n))F^*(G(n))/(h_1G(n)). && (3.26)\end{aligned}$$

Here $\bar{T}_B$, $F(t)$ and $F^*(\cdot)$ represent, respectively, the mean, the distribution function and its Laplace-Stieltjes transform of the random variable T_B representing the busy period length due to LS; $\bar{T}_B$ and $F^*(\cdot)$ are given as follows (for the derivation, see Appendix 3.2).

$$\bar{T}_B = \frac{\rho_D+(1-\rho_D)q}{\lambda_2(1-\rho_D)(1-q)}, \quad \text{and} \tag{3.27}$$

$$F^*(s) = \frac{(1-q)\exp(-sh_3)F_0^*(s)}{1-qF_0^*(s)H^*(s)} \tag{3.28}$$

where

$$\begin{aligned} \rho_D &\equiv \lambda_2 h_D, && (3.29)\\ q &\equiv 1-\exp(-\lambda_2 h_3), && (3.30)\\ F_0^*(s) &\equiv \sum_{n=1}^{\infty}\frac{(n\rho_D)^{n-1}}{n!}\exp\{-(s+\lambda_2)h_Dn\}, \quad \text{and} && (3.31)\\ H^*(s) &\equiv \frac{\lambda_2[1-\exp\{-(s+\lambda_2)h_3\}]}{q(s+\lambda_2)}. && (3.32)\end{aligned}$$

As is seen in Eq. 3.31, for the rigorous calculation of $F^*(s)$, an infinite sum must be evaluated. To avoid it, we approximate the probability density function of T_B by a shifted exponential distribution,

$$f(x) = \gamma e^{-\gamma(x-d)}, \quad x \geq d, \tag{3.33}$$

where

$$d = \min\{T_B\} = h_D + h_3 \quad \text{and} \tag{3.34}$$

$$\gamma^{-1} = \bar{T}_B - d. \tag{3.35}$$

Then $F^*(G(n))$ contained in Eqs. 3.24 and 3.26 is given by

$$F^*(G(n)) = \frac{\exp(-G(n)d)}{G(n)/\gamma + 1}. \tag{3.36}$$

Thus, using P_i and R_i, the conditional throughput is given by

$$S(n) = \left(1 - \sum_{i=1}^{4} P_i R_i\right) G(n). \tag{3.37}$$

ICMA access scheme (Fig. 3.3)

In contrast to the ALOHA access scheme, this scheme inhibits access when a CS is detected. The detection delay from the transmission of a CS by an MT to the reception of an US by other MT is assumed to be constant a.

In this scheme, the state of the F-ch is divided into the following three states:

B$_0$: in the idle state,

B$_1$: in the busy state due to detection delay and call origination, and

B$_2$: in the busy state due to LS.

From the start of an idle state to the start of the next idle state is called a cycle. The throughput is determined by considering the number of CSs that succeed in a cycle.

We define transition probabilities among above states as follows.

$$\begin{aligned}
p_{ij} &= \{\text{transition probability from state B}_i \text{ to state B}_j\}, \\
&\qquad\qquad\qquad i = 0, 1,\ j = 0, 1, 2,\ i \neq j, \\
p_{11} &= \{\text{probability that state B}_1 \text{ continues} \\
&\qquad\qquad \text{at the end of US transmission due to a CS}\}, \quad \text{and} \\
p_{2j}^{(i)}(t) &= \{\text{transition probability from state B}_2 \text{ of length } t, \\
&\qquad\qquad \text{which follows state B}_i\text{, to state B}_j\}, i = 0, 1,\ j = 0, 1.
\end{aligned}$$

Then, they are given as

$$\begin{aligned}
p_{01} &= G(n)/\lambda_0, \qquad p_{02} = \lambda_2/\lambda_0, \\
p_{10} &= \exp(-h_1\lambda_0), \qquad p_{12} = 1 - \exp(-h_1\lambda_2), \\
p_{11} &= \exp(-h_1\lambda_2)\{1 - \exp(-h_1 G(n))\}, \\
p_{20}^{(0)}(t) &= \exp(-tG(n)), \\
p_{21}^{(0)}(t) &= 1 - \exp(-tG(n)), \\
p_{20}^{(1)}(t) &= \exp(-(h_1 + t)G(n)), \quad \text{and} \\
p_{21}^{(1)}(t) &= 1 - \exp\{-(h_1 + t)G(n)\},
\end{aligned} \tag{3.38}$$

where

$$\lambda_0 = G(n) + \lambda_2. \tag{3.39}$$

Let $C_{ij}(k)$ $(i, j = 1, 2,\ k = 1, 2, \cdots)$ be a cycle starting from B_i, which follows state B_0, containing k B_1's and ending in state B_j. Using the transition probabilities above, the probability, $P_{ij}(k, \vec{m}, \vec{t})$ $(i, j = 1, 2,\ k = 1, 2, \cdots, \vec{m} = (m_1, m_2, \cdots, m_k), \vec{t} = (t_1, t_2, \cdots, t_k))$ of a cycle $C_{ij}(k)$ being one such that the number of USs during the interval from the $(r-1)$th visit to B_2 to the rth visit to B_2 is equal to m_r and the sojourn time at the rth visit to B_2 is t_r, $r = 1, 2, \cdots, k$, is given as follows.

$$\begin{aligned}
P_{ij}(k, \vec{m}, \vec{t}) \;&=\; \begin{cases}
p_{01}p_{10}p_{12}^{k-1}\displaystyle\prod_{r=1}^{k-1}\{p_{11}^{m_r-1}p_{21}^{(1)}(t_r)\}p_{11}^{m_k-1}, & i = j = 1,\\
p_{01}p_{20}^{(1)}(t_k)p_{12}^{k}\displaystyle\prod_{r=1}^{k-1}\{p_{11}^{m_r-1}p_{21}^{(1)}(t_r)\}p_{11}^{m_k-1}, & i = 1,\ j = 2,\\
p_{02}p_{10}p_{21}^{(0)}(t_1)p_{12}^{k-1}\displaystyle\prod_{r=1}^{k-1}\{p_{11}^{m_r-1}p_{21}^{(1)}(t_{r+1})\}p_{11}^{m_k-1}, & i = 2,\ j = 1,\\
p_{02}p_{21}^{(0)}(t_1)p_{20}^{(1)}(t_{k+1})p_{12}^{k}\displaystyle\prod_{r=1}^{k-1}\{p_{11}^{m_r-1}p_{21}^{(1)}(t_{r+1})\}p_{11}^{m_k-1}, & i = j = 2,
\end{cases} \qquad (3.40)\\
k \;&=\; 1, 2, \cdots, \quad \text{and}\\
P_{22}(0, -, t_1) \;&=\; p_{02}p_{20}^{(0)}(t_1). \qquad (3.41)
\end{aligned}$$

Using those equations,

$$\begin{aligned}
L_{ij}(k) \;&=\; \mathrm{E}[\text{length of cycle } C_{ij}(k)] \quad \text{and} \qquad (3.42)\\
M_{ij}(k) \;&=\; \mathrm{E}[\text{number of successful CS in cycle } C_{ij}(k)] \qquad (3.43)
\end{aligned}$$

are derived. Then the conditional throughput $S(n)$ is derived as (for the derivation, see Appendix 3.3)

$$\begin{aligned}
S(n) \;&=\; \frac{h_1\sum_{k=1}^{\infty}[M_{11}(k) + M_{12}(k) + M_{22}(k) + M_{21}(k)]}{L_{22}(0) + \sum_{k=1}^{\infty}[L_{11}(k) + L_{12}(k) + L_{22}(k) + L_{21}(k)]}\\
&=\; \frac{\lambda_0^2\exp(-aG(n))}{1 - p_{11}} \qquad (3.44)\\
&\quad\times\frac{\{G(n)(1 - p_{11} + h_1G(n)p_{10}) + \lambda_2Y_2\}Y_3 + (\lambda_0 - \lambda_2C_1)Y_1(1 - \mathrm{e}^{-h_1\lambda_2})}{(G(n)^2 + \lambda_2^2 + \bar{T}_B\lambda_2\lambda_0^2)Y_3 + \lambda_0^2(\lambda_0 + \lambda_2C_1)\{h_1 + a + \bar{T}_B(1 - \mathrm{e}^{-h_1\lambda_2})\}},
\end{aligned}$$

where

$$C_1 \equiv \int_0^\infty \mathrm{e}^{-tG(n)} dF(t) = F^*(G(n)), \tag{3.45}$$

$$C_2 \equiv \int_0^\infty t\mathrm{e}^{-tG(n)} dF(t) = -\frac{d}{ds}F^*(s)\Big|_{s=G(n)}, \tag{3.46}$$

$$\begin{aligned} Y_1 \equiv\ & (1-p_{11})(h_1C_1+C_2)G(n)\exp(-h_1G(n)) \\ & +h_1G(n)p_{10}[1-C_1\exp(-h_1G(n)) \\ & -(h_1C_1+C_2)G(n)\exp(-(h_1+a)G(n))], \end{aligned} \tag{3.47}$$

$$Y_2 \equiv (1-p_{11})C_2G(n) + h_1G(n)p_{10}[1-C_1-C_2G(n)\exp(-aG(n))], \quad \text{and} \tag{3.48}$$

$$Y_3 \equiv C_1\exp(-h_1G(n)) + (1-C_1)p_{10}. \tag{3.49}$$

As in the case of the ALOHA access scheme, the probability density function of T_B is approximated by a shifted exponential distribution (Eq. 3.33). Then, C_1 is given by Eq. 3.36 and C_2 is given by

$$C_2 = \frac{(\bar{T}_B + G(n)d/\gamma)C_1}{G(n)/\gamma + 1}. \tag{3.50}$$

By substituting Eqs. 3.36 and 3.47 to 3.50 into Eq. 3.44, the conditional throughput $S(n)$ is obtained.

3.5 Numerical example

We apply our approximation method to ALOHA and ICMA access schemes, and compare the numerical results with simulation results to examine the accuracy of the method. We consider a system with $M = 100$, $h_1 = h_2 = h_3 = 1$, $a = 0.1$, $N = 3$, $\lambda_2 = 0.2$, and $\sigma = 0.08$. The parameters correspond, for example, to the signal length of 240 *ms*, the mean retransmission interval of 3 *s*, and a 2 *min* arrival interval of LS per mobile user. To examine the effect of being busy due to LS on the evaluation measure, the arrival rate λ_2 is set larger than in the actual traffic conditions.

Figure 3.6 shows the results for the average number of retransmissions when the arrival rate of CS varies. The figure indicates that the proposed approximation method is accurate enough in practical usages.

Figures 3.7 and 3.8 show results for the throughput and the mean number of retransmissions in the cases of $N = 2$ and 10 (other parameters remaining the same). It is seen that with the increase in the limit of retransmission, the total input traffic due to retransmissions is increased, thereby drastically decreasing the throughput. This property indicates that by limiting the number of retransmission, the instability of the channel [108], which is inherent to the random access protocol, can be suppressed.

Figure 3.9 shows the throughput performance when the retransmission rate σ varies. The throughput decreases with the increase in σ. This is due to the increase of the probability that signals once collided again collide. The numerical examples indicate that both the throughput and the mean number of retransmissions can be improved by the transmission of US by the BS due to CS.

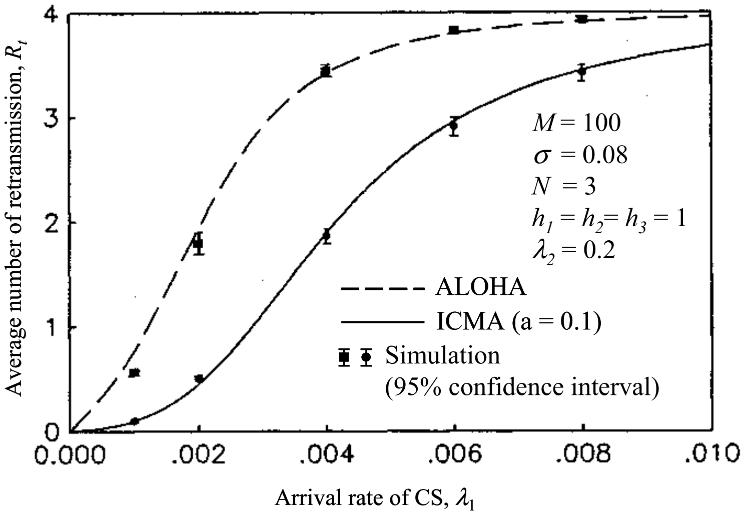


Figure 3.6: Comparisons with simulation results

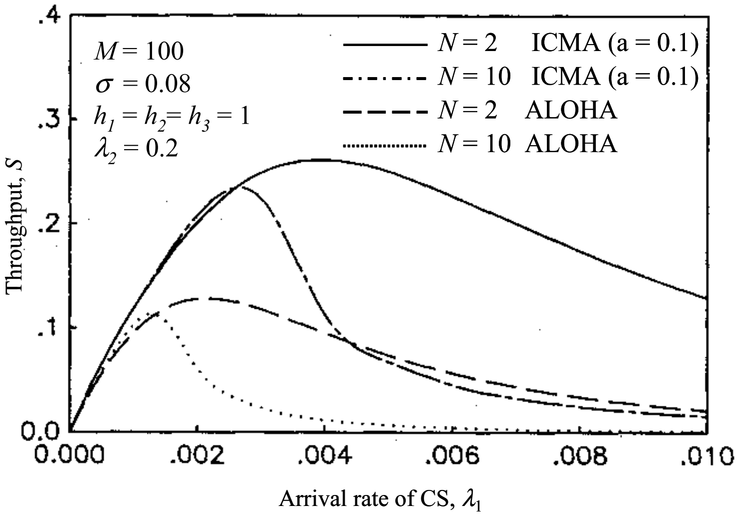


Figure 3.7: Throughput versus arrival rate of CS ($N = 2, 10$)

3.6 Conclusion

This chapter considered the performance evaluation of the control channel access scheme in cellular systems, and proposed an approximate method for the analysis that is a combination of the S-G analysis and the diffusion approximation analysis. Using the proposed method, the performances of ALOHA and ICMA access schemes are evaluated, and the effects of the access inhibition and retransmission control on the traffic characteristics are investigated. Through comparison with the results of simulation, the accuracy of approximation is examined and the effectiveness of the method is verified.

The approximation method proposed in this chapter is based on two approximations, the S-G analysis and the diffusion approximation analysis. Consequently, the range of applications of the proposed method will be limited by the conditions of these two methods. More definitely, the method should be used in cases where the following two conditions

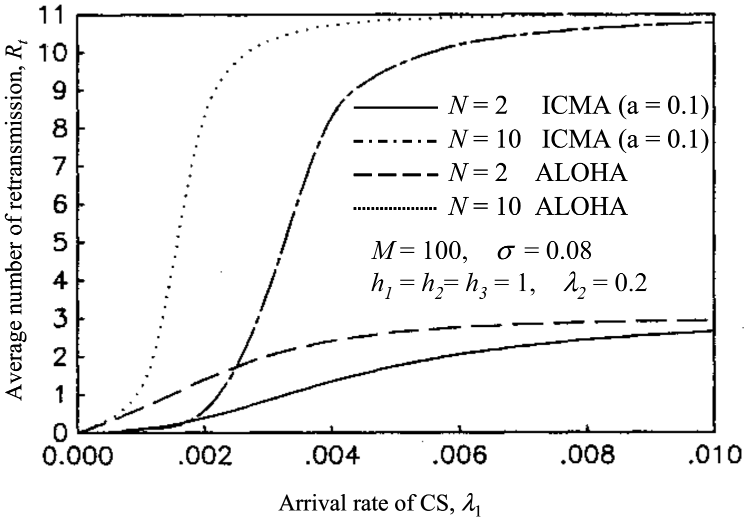


Figure 3.8: Average number of retransmission versus arrival rate of CS ($N = 2, 10$)

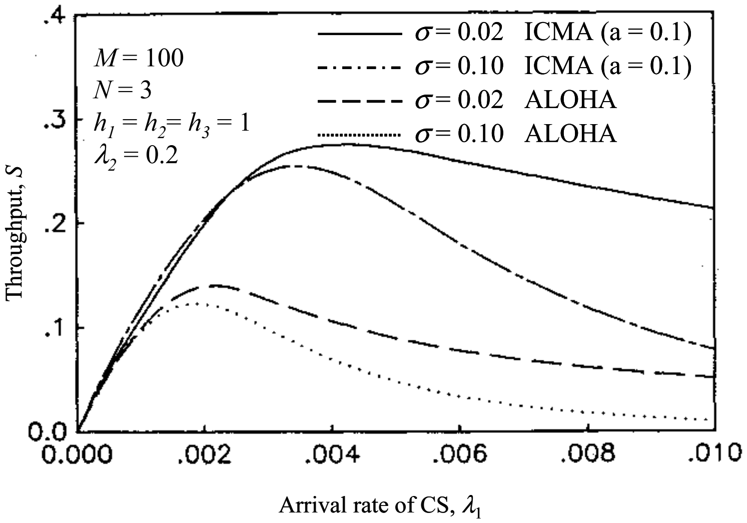


Figure 3.9: Throughput versus arrival rate of CS ($s = 0.02, 0.1$)

are satisfied.

(1) The number of MTs is large (inherited from the S-G analysis).

(2) Input load is relatively heavy (inherited from the diffusion approximation analysis).

It is left for further study to clarify the range of applications of the proposed analytical method.

Appendix 3.1 Derivation of Eqs. 3.9 and 3.10

Let $Q(t)$ be the number of active MTs at time t, $A(t)$ the cumulative number of MTs that made transitions from the VT mode to the active mode during interval $(0, t]$, and

$D(t)$ the cumulative number of MTs that made transitions from the active mode to the VT mode during that interval. Then

$$Q(t) = Q(0) + A(t) - D(t). \tag{3.51}$$

The discrete stochastic process $\{Q(t)\}$ is approximated by a diffusion process $\{X(t)\}$ with the probability density function

$$p(x,t)\,dx = Pr\{x \leq X(t) < x + dx\}. \tag{3.52}$$

The range of $\{Q(t)\}$ is the closed interval $[0, M]$, and for the approximation we have to introduce boundary conditions. In Refs. [27, 28, 29], reflecting boundaries are placed at the origin ($x = 0$) and $x = M$, and the ALOHA protocol is analyzed. However, in this chapter, at the origin, we use the elementary return boundary [30, 111], which is known to realize a high accuracy even for light traffic. A reflecting boundary is placed at $x = M$. The holding time distribution at the origin is approximated by an exponential distribution with mean $h_0^{-1} = 1/(M\lambda_v)$ [112]. Then $p(x,t)$ satisfies the following set of equations:

$$\begin{aligned}
&\frac{1}{2}\frac{\partial^2}{\partial x}\alpha(x)p(x,t) - \frac{\partial}{\partial x}\beta(x)p(x,t) = \frac{\partial p}{\partial t} - h_0\pi_0(t)\delta(x-1),\\
&\frac{d}{dt}\pi_0(t) + h_0\pi_0(t) = \frac{1}{2}\frac{\partial}{\partial x}\alpha(x)p(x,t) - \beta(x)p(x,t)\bigg|_{x=0}, \quad \text{and}\\
&0 = \frac{1}{2}\frac{\partial}{\partial x}\alpha(x)p(x,t) - \beta(x)p(x,t)\bigg|_{x=M},
\end{aligned} \tag{3.53}$$

where $\pi_0(t)$ denotes the holding probability at the origin; $\alpha(x)$ and $\beta(x)$ are diffusion parameters defined by

$$\begin{aligned}
\beta(x) &\equiv \lim_{\Delta t \to 0} \frac{\mathrm{E}[X(t+\Delta t) - X(t) \mid X(t) = x]}{\Delta t}, \quad \text{and}\\
\alpha(x) &\equiv \lim_{\Delta t \to 0} \frac{\mathrm{Var}[X(t+\Delta t) - X(t) \mid X(t) = x]}{\Delta t}.
\end{aligned}$$

The state space $[0, M]$ is divided into $I_0 = \{0\}$ and M intervals $I_n = (n-1, n]$, $n = 1, \cdots, M$. The diffusion parameters are assumed to be constants on each interval. The diffusion parameter in the interval I_n ($n = 1, \cdots, M$) is called the discretized diffusion parameters and are defined in the following manner.

Under the condition that the number of active MTs is equal to n, the increments of $A(t)$ and $D(t)$ in infinitesimal time interval $(t, t + \Delta t]$ are written as $\Delta A(t)$ and $\Delta D(t)$, respectively. Then,

$$Q(t + \Delta t) - Q(t) = \Delta A(t) - \Delta D(t). \tag{3.54}$$

Furthermore, we define

$$\begin{aligned}
\lambda_a(n) &\equiv \lim_{\Delta t \to 0} \frac{\mathrm{E}[\Delta A(t) \mid Q(t) = n]}{\Delta t}, \quad \text{and}\\
\lambda_d(n) &\equiv \lim_{\Delta t \to 0} \frac{\mathrm{E}[\Delta D(t) \mid Q(t) = n]}{\Delta t}.
\end{aligned}$$

Then $\lambda_a(n)$ is the transition rate from the VT mode (with $M-n$ terminals) to the active mode under the condition that the number of active MTs is n. Then, obviously,

$$\lambda_a(n) = (M-n)\lambda_v. \tag{3.55}$$

The transition rate from the active mode to the VT mode can be considered as the sum of transition rated to the VT mode with transmission success and to the VT mode due to call loss by exceeding the limit of retransmission. Consequently, using $S(n)$ and $L(n)$ derived by the S-G analysis in Sect. 3.4, $\lambda_d(n)$ is approximated as

$$\lambda_d(n) = S(n) + L(n). \tag{3.56}$$

Thus, the discretized diffusion parameter $\beta(x)$ is given by

$$\beta(x) \equiv \beta_n = \lambda_a(n) - \lambda_d(n), \quad n-1 < x \leq n. \tag{3.57}$$

By approximating the stochastic processes $\{A(t)\}$ and $\{D(t)\}$ by Poisson processes, $\alpha(x)$ is given by

$$\alpha(x) \equiv \alpha_n = \lambda_a(n) + \lambda_d(n), \quad n-1 < x \leq n. \tag{3.58}$$

Let $p_n(x) \equiv \lim_{t\to\infty} p(x,t)$ for $x \in I_n$. Solving Eq. 3.53 and using the discretization

$$\pi_n = \int_{n-1}^{n} p_n(x)dx, \quad n = 1, \cdots, M, \tag{3.59}$$

the state probability π_n is determined as in Eqs. 3.9 and 3.10 [30].

Appendix 3.2 Derivation of Eqs. 3.27 and 3.28

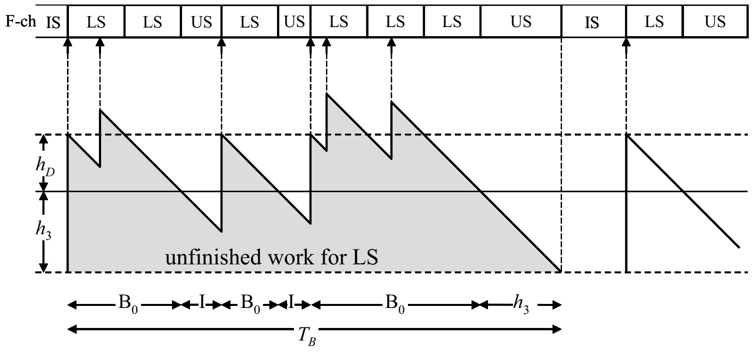


Figure 3.10: Busy period due to LS

As shown in Fig. 3.10, the busy period due to LS is composed of a busy period of $M/D/1$ queue with the LS signal transmitting interval h_D as the holding time and an idle period under the condition that the period length is h_3 or less. We denote the length of the former busy period as B_0 and the length of the latter idle time as I. Then B_0 and

I are subjecting to the following distributions [108].

$$\begin{aligned} F_0(t) &\equiv \Pr\{B_0 \le t\} = \sum_{n=1}^{[t/h_p]} \frac{(n\rho_D)^{n-1}}{n!} \exp(-n\rho_D), \quad \text{and} && (3.60) \\ H(t) &\equiv \Pr\{I \le t \mid I \le h_3\} = \frac{1-\exp(-\lambda_2 t)}{1-\exp(-\lambda_2 h_3)}. && (3.61) \end{aligned}$$

The number of idle periods N_I contained in a busy period follows the geometrical distribution

$$\Pr\{N_I = k\} = (1-q)q^k, \tag{3.62}$$

where

$$q \equiv \Pr\{I \le h_3\} = 1 - \exp(-\lambda_2 h_3). \tag{3.63}$$

Then, the length T_B of the busy period is given by

$$T_B = (N_I + 1)B_0 + N_I I + h_3. \tag{3.64}$$

We will write the Laplace-Stieltjes transform (LST) of a distribution $G(t)$ as $G^*(s)$. Then the LST of the distribution $F(t)$ of T_B is given as follows.

$$\begin{aligned} F^*(s) &\equiv \mathrm{E}[e^{-sT_B}] = \sum_{k=0}^{\infty} \mathrm{E}[e^{-sT_B} \mid N_I = k] \Pr\{N_I = k\} \\ &= \sum_{k=0}^{\infty} e^{-sh_3} [F_0^*(s)]^{k+1} [H^*(s)]^k \Pr\{N_I = k\} \qquad (3.65) \\ &= \frac{(1-q)\exp(-sh_3)F_0^*(s)}{1 - qF_0^*(s)H^*(s)}. \end{aligned}$$

Differentiating the above equation by s and letting $s = 0$, the mean of T_B (Eq. 3.27) is obtained.

Appendix 3.3 Derivation of Eq. 3.44

We put $G_1 = G(n)$ and $G_2 = \lambda_2$. Then from its definition (Eq. 3.42), $L_{ij}(k)$ is represented as follows.

$$\begin{aligned} L_{ij}(k) &= \int_0^{\infty} \cdots \int_0^{\infty} \sum_{m_1=1}^{\infty} \cdots \sum_{m_k=1}^{\infty} \left[G_i/\lambda_0^2 + (h_1 + a) \sum_{r=1}^{k} m_r + \sum_{r=1}^{k+(i-1)j-1} t_r \right] \\ &\quad \times P_{ij}(k, \vec{m}, \vec{t}) dF(t_1) \cdots dF(t_{k+(i-1)j-1}). \qquad (3.66) \end{aligned}$$

By substituting Eq. 3.40 and rearranging, we have

$$L_{ij}(k) = \begin{cases} G_1 a_0(k) b_1(k), & i = j = 1, \\ G_2 a_0(k)[b_2(k) + (1-p_{11})(\bar{T}_B - C_2)(1 - C_1 \mathrm{e}^{-h_1 G_1})], & i = j = 2, \\ (\mathrm{e}^{h_1 G_2} - 1) C_1 L_{ij}(k) & \\ \quad + G_i C_2 (1-p_{11}) \mathrm{e}^{-h_1 G_1} (1 - \mathrm{e}^{-h_1 G_2})(1 - C_1 \mathrm{e}^{-h_1 G_1}), & i \neq j, \end{cases} \tag{3.67}$$

$$k = 1, 2, \cdots, \quad \text{and}$$

$$L_{22}(k) = G_2(\lambda_2 C_1/\lambda_0^2 + C_2)/\lambda_0, \tag{3.68}$$

where

$$a_0(k) \equiv \frac{p_{10}(1 - \mathrm{e}^{-h_1 G_2})(1 - C_1 \mathrm{e}^{-h_1 G_1})^{k-2}}{\lambda_0 (1-p_{11})^{k+1}} \quad \text{and}$$

$$b_i(k) \equiv (1 - C_1 \mathrm{e}^{-h_1 G_1})\{(1-p_{11}) G_i/\lambda_0^2 + (h_1 + a)k\} + (k-1)(1-p_{11})(\bar{T}_B - C_2 \mathrm{e}^{-h_1 G_1}).$$

Next, we derive the mean number of successful CSs in a cycle $C_{ij}(k)$. The probability $p(n,m)$ that n among m CSs composing state B_1 succeed, depends on the state immediately before state B_1. Let us denote the (conditional) probability as $p_i(n,m)$ when the state immediately before B_1 is B_i, $i = 0, 2$. Then by Eq. 3.38, the probabilities satisfy the following recurrence relation.

$$p_i(n,m) = (1 - \mathrm{e}^{-h_1 G_1} - h_1 G_1 \mathrm{e}^{-(a+h_1)G_1}) p_i(n, m-1) + h_1 G_1 \mathrm{e}^{-(a+h_1)G_1} p_i(n-1, m-1). \tag{3.69}$$

The initial values are given as follows.

$$p_0(0,m) = (1 - \mathrm{e}^{-aG_1})(1 - \mathrm{e}^{-h_1 G_1} - h_1 G_1 \mathrm{e}^{-(a+h_1)G_1})^{m-1}, \quad m = 1, 2, \cdots,$$

$$p_0(1,1) = \mathrm{e}^{-aG_1},$$

$$p_2(0,m) = [1 - \mathrm{e}^{-(h_1+t)G_1} - (h_1+t) G_1 \mathrm{e}^{-(a+h_1+t)G_1}](1 - \mathrm{e}^{-h_1 G_1} - h_1 \mathrm{e}^{-(a+h_1)G_1})^{m-1}, \quad m = 1, 2, \cdots, \quad \text{and}$$

$$p_2(1,1) = (h_1 + t) G_1 \mathrm{e}^{-(a+h_1+t)G_1}.$$

Using these equations, the mean number of successes $N_i(m)$ $(i = 0, 2)$ under the condition that state B_1 is composed of m CSs are given as follows:

$$N_0(m) = \sum_{n=1}^{m} n p_0(n,m) = (m-1) h_1 G_1 \mathrm{e}^{-(a+h_1)G_1}/(1 - \mathrm{e}^{-h_1 G_1}) + \mathrm{e}^{-aG_1}, \tag{3.70}$$

and

$$N_2(m) = \sum_{n=1}^{m} n p_2(n,m) = (m-1) h_1 G_1 \mathrm{e}^{-(a+h_1)G_1}/(1 - \mathrm{e}^{-h_1 G_1}) + (h_1 + t) G_1 \mathrm{e}^{-(a+h_1+t)G_1}/(1 - \mathrm{e}^{-(h_1+t)G_1}). \tag{3.71}$$

Then $M_{ij}(k)$ is represented in the following manner.

$$M_{1j}(k) \;=\; \int_0^\infty \cdots \int_0^\infty \sum_{m_1=1}^\infty \cdots \sum_{m_k=1}^\infty \left[N_0(m_1) + \sum_{r=2}^{k} N_2(m_r) \right] P_{1j}(k, \vec{m}, \vec{t}) dF(t_1) \cdots dF(t_{k-1}) \tag{3.72}$$

and

$$M_{2j}(k) \;=\; \int_0^\infty \cdots \int_0^\infty \sum_{m_1=1}^\infty \cdots \sum_{m_k=1}^\infty \left[\sum_{r=1}^{k} N_2(m_r) \right] P_{2j}(k, \vec{m}, \vec{t}) dF(t_1) \cdots dF(t_{k+j-1}) \tag{3.73}$$

Substituting Eqs. 3.40, 3.70 and 3.71 into the above equations, we have the following result.

$$M_{ij}(k) \;=\; \begin{cases} G_1 a_0(k) \mathrm{e}^{-aG_1}[(1 - C_1 \mathrm{e}^{-h_1 G_1})(1 - p_{11} + h_1 G_1 \mathrm{e}^{-h_1 \lambda_0}) + (k-1) Y_1], & \\ & i = j = 1 \\ G_2 a_0(k) \mathrm{e}^{-aG_1}[(1 - C_1 \mathrm{e}^{-h_1 G_1}) Y_2 + (k-1)(1 - C_1) Y_1], & i = j = 2 \\ (\mathrm{e}^{h_1 G_2} - 1) C_1 M_{ii}(k), & i \neq j \end{cases}$$

$$k = 1, 2, \cdots, \tag{3.74}$$

where Y_1 and Y_2 are defined by Eqs. 3.47 and 3.48, respectively. Using Eqs. 3.67, 3.68, and 3.74, Eq. 3.44 is obtained.

Chapter 4

Priority control in packet transmission networks

4.1 Introduction

In this chapter, aside from the radio access layer in the previous chapters, the traffic control at the transmission network is discussed. The traffic control considered here is a packet level priority control at transmission nodes for different packet classes with different requirements on delays, while a call level control is considered in the next chapter. An approximate method is proposed for queueing networks with priorities, which can be applied to evaluate the end-to-end delay and throughput in the transmission layer of multimedia cellular systems.

Most of priority queueing network models are not in the class of product form (or BCMP) networks, and several approximation approaches have been used. One of the most popular approximation approaches is the *virtual server method* or reduced occupancy approximation [44, 45, 46, 47, 48]. Using this method, closed queueing networks with preemptive priority are treated in [44, 45, 46, 47], and nonpreemptive cases are considered in [46, 47, 48]. These studies have shown that the virtual server method can capture most of important properties of priority disciplines in queueing networks. However, these approaches are restricted to Markovian priority networks, i.e., both interarrival time and service time distributions are assumed to be exponential. In practice, non-Poissonian arrivals and general service time distributions are often encountered in communication systems. The bursting nature of packet arrival processes and the constant packet length in multimedia packet communication systems are examples.

In this chapter, we present an approximation method for analyzing open renewal queueing networks with two-class nonpreemptive priorities, non-Poissonian renewal arrivals, and nonexponential renewal service-times. The method is derived by combining the virtual server method with the *decomposition technique* used in the *queueing network analyzer* (QNA) [51]. The method does not use any iteration nor solution of complex equations, and it can approximately analyze large networks with small computational burden.

The remainder of this chapter is organized as follows. In Sect. 4.2, renewal queueing network models with nonpreemptive priorities treated in this chapter are described. The proposed approximation method is fully investigated in Sect. 4.3 as a two-parameter

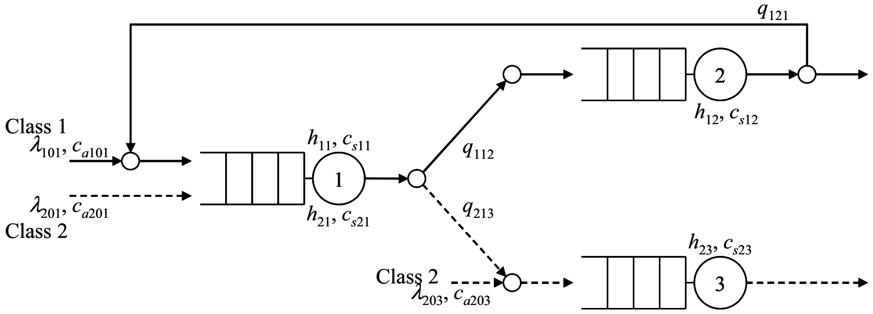


Figure 4.1: Priority queueing network model

approximation method. The validity of the proposed method is checked in Sect. 4.4 by comparing results by the method with exact solutions and simulation results for two-stage tandem network models and basic component models of complex networks. Additionally, in the section, an application of the proposed method to the end-to-end delay analysis of a packet-switching network for voice and data is presented. Section 4.5 draws some conclusions and discusses topics for future study.

4.2 Queueing network model

An example of the queueing network models treated in this chapter is depicted in Fig. 4.1. The following assumptions are made in the models.

(1) network form: The network is an open queueing network with N nodes (servers, switches) ($N \geq 1$). It may have arbitrary network configuration.

(2) class: There are two priority classes, class 1 (high priority) and class 2 (low priority).

(3) arrival: The arrival processes of customers (packets in the transmission network) are assumed to be mutually independent renewal processes. The interarrival times of class k ($k = 1, 2$) customers from outside of the network to node i are subjecting to a general distribution with mean λ_{k0i}^{-1} and coefficient of variation c_{ak0i} (suffix 0 stands for the outside of the network).

(4) service: Each node in the network consists of a nonpreemptive single server and unlimited waiting space. The service times of class k customers at node i are subjecting to a general distribution with mean h_{ki} and coefficient of variation c_{ski}. Service times are mutually independent and are also independent of arrival processes.

(5) routing: Each class of customers is routed independently according to the routing matrix Q_k=(q_{kij}), where q_{kij} denotes the probability of a class k customer being routed to node j after completing service at node i.

The total arrival rate λ_{ki} of class k customers to node i is obtained from the system of linear equations

$$\lambda_{ki} = \lambda_{k0i} + \sum_{j=1}^{N} \lambda_{kj} q_{kji}, \quad k = 1, 2; i = 1, 2, \cdots, N. \tag{4.1}$$

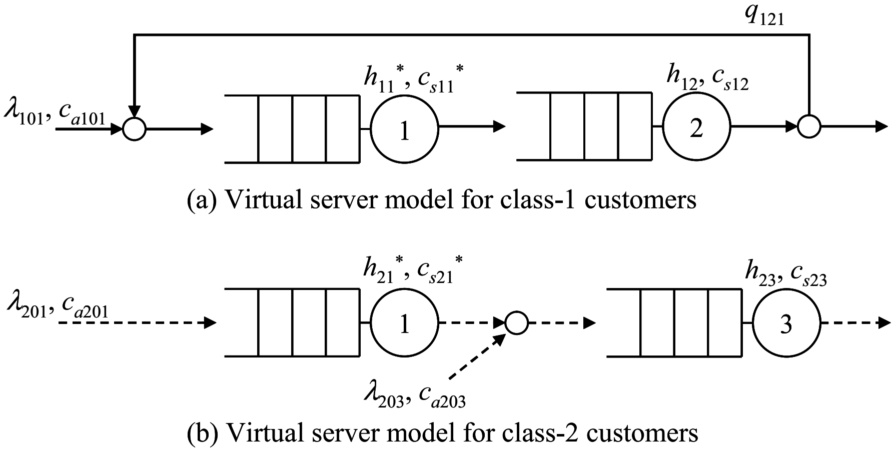

q_{121}
h_{11}^*, c_{s11}^*
h_{12}, c_{s12}
λ_{101}, c_{a101}
1
2
(a) Virtual server model for class-1 customers
h_{21}^*, c_{s21}^*
h_{23}, c_{s23}
λ_{201}, c_{a201}
1
3
λ_{203}, c_{a203}
(b) Virtual server model for class-2 customers

Figure 4.2: Virtual server models of the network model in Fig. 4.1

This set of equations is called the *traffic rate equations*, and its solution is obtained by inverting an $N \times N$ matrix for each class. Below, we assume the equations have a solution and the traffic intensities $\rho_{ki} = \lambda_{ki} h_{ki}$ $(k = 1, 2; i = 1, 2, \cdots, N)$ satisfy the equilibrium condition

$$\rho_{1i} + \rho_{2i} < 1, \quad i = 1, 2, \cdots, N. \tag{4.2}$$

4.3 Two-parameter approximation

4.3.1 Outline of the analysis

In this section, we describe an approximation method for analyzing the renewal priority queueing networks described above. The method is composed of the three steps outlined below. These steps are based on two recent results: the virtual server method for priority networks and the decomposition method for renewal networks.

(1) Separate a two-class queueing network into two single-class queueing networks, replacing each priority server in the original network with two virtual servers, one serving high priority customers, and one for low priority customers. For example, the separation of the priority network in Fig. 4.1 leads to the two single-class networks depicted in Fig. 4.2.

(2) Analyze the flow rates and variability parameters of internal arrival processes for each network.

(3) Calculate traffic characteristics for each node in the networks by regarding the node as a $GI/G/1$ queue.

Note that the virtual server method has been validated for Markovian priority networks, but not for renewal priority networks. Hence, we propose a two-parameter virtual server method for the renewal networks in which the service time distributions of the virtual servers are approximately characterized by the first and second moments. The single-class networks obtained by this method can be regarded as renewal queueing networks. Then we apply the decomposition method to the single-class networks and derive two-parameter approximations for performance characteristics of the renewal queueing networks.

4.3.2 Separation of priority queueing networks

Each node in the network with a nonpreemptive priority server can be regarded as a $GI_1, GI_2/G_1, G_2/1$ (nonpreemptive) model, if the arrival process for each class is approximated by a renewal process. In the $GI_1, GI_2/G_1, G_2/1$ (nonpreemptive) model, high priority and low priority customers affect each other, and the waiting times for them become mutually dependent.

The key idea of the virtual server method is to separate a priority server into two virtual servers, whose service time distributions are chosen so that the dependence is approximately taken into account. Each $GI_1, GI_2/G_1, G_2/1$ (nonpreemptive) node is separated into two single-server queues, i.e., $GI_1/G_1/1$ for class 1 customers and $GI_2/G_2/1$ for class 2 customers. Then, a two-class queueing network is transformed into two single-class networks, which can be analyzed more easily.

Here, the method of approximating the service time distributions at the virtual servers is important. For nonpreemptive Markovian networks, Kaufman [46] and Ikehara [48] have proposed virtual server models with constant reduced service rates. Schmitt [47] extended Kaufman's model to a virtual server with a state-dependent service rate.

Based on ideas of the decomposition method and the diffusion approximation, we approximate the service time distributions at the virtual servers taking account of not only the service rate but also the variability parameter. More definitely, we derive the mean service time h^*_{ki} and the coefficient of variation c^*_{ski} for the service time distribution in the $GI_k/G_k/1$ model from the following ideas [113].

(1) The mean service time in the $GI_k/G_k/1$ model is set equal to the mean completion time [114] of class k customers in the $GI_1, GI_2/G_1, G_2/1$ (nonpreemptive) model. The latter is approximately derived by further assuming the arrival process of class 1 customers is Poissonian.

(2) The mean waiting time in the $GI_k/G_k/1$ model is set equal to the mean waiting time of class k customers in the $GI_1, GI_2/G_1, G_2/1$ (nonpreemptive) model. The latter is approximately derived by the heavy-traffic limit theorem [115] for the diffusion approximation of the $GI_1, GI_2/G_1, G_2/1$ (nonpreemptive) model.

From (1) above, we have

$$h^*_{1i} = h_{1i}, \quad \text{and} \tag{4.3}$$

$$h^*_{2i} = \frac{h_{2i}}{1-\rho_{1i}}, \tag{4.4}$$

and from (2), we have

$$\begin{aligned}
\frac{\lambda_{1i} h_{1i}^{*2}(c_{a1i}^2 + c_{s1i}^{*2})}{2(1-\rho_{1i}^*)} &= \frac{h_{1i}\rho_{1i}(c_{a1i}^2 + c_{s1i}^2) + h_{2i}\rho_{2i}(1 + c_{s2i}^2)}{2(1-\rho_{1i})}, \quad \text{and} && (4.5)\\
\frac{\lambda_{2i} h_{2i}^{*2}(c_{a2i}^2 + c_{s2i}^{*2})}{2(1-\rho_{2i}^*)} &= \sum_{k=1}^{2} \frac{h_{ki}\rho_{ki}(c_{aki}^2 + c_{ski}^2)}{2(1-\rho_{1i}-\rho_{2i})(1-\rho_{1i})} + \frac{h_{2i}\rho_{1i}(c_{a2i}^2 - 1)}{2(1-\rho_{1i})}, && (4.6)
\end{aligned}$$

where $\rho_{ki}^* = \lambda_{ki} h_{ki}^*$, and c_{aki}^2 represents the squared coefficient of variation of interarrival times of class k customers at node i, which is obtained by solving the system of linear equations described in the next subsection. Substituting Eqs. 4.3 and 4.4 into Eqs. 4.5 and 4.6, respectively, the squared coefficients of variation of the service time distributions are derived as

$$\begin{aligned}
c_{s1i}^{*2} &= c_{s1i}^2 + \frac{h_{2i}\rho_{2i}}{h_{1i}\rho_{1i}}(1 + c_{s2i}^2), \quad \text{and} && (4.7)\\
c_{s2i}^{*2} &= c_{s2i}^2 + \frac{\rho_{1i}}{h_{2i}\rho_{2i}}\{h_{1i}(c_{a1i}^2 + c_{s1i}^2) + (1-\rho_{1i}-\rho_{2i})h_{2i}(c_{a2i}^2 - 1)\}. && (4.8)
\end{aligned}$$

Thus, we obtain two single-class queueing networks in which the service time distributions at node i are characterized by Eqs. 4.3 and 4.7 for class 1 customers and Eqs. 4.4 and 4.8 for class 2 customers.

4.3.3 Analysis of single-class renewal networks

In this section we analyze the renewal network for each class by decomposing it into a set of single nodes. The approximate procedure is based on the decomposition method used in the queueing network analyzer (QNA), which is composed of the approximations for three basic network operations: merging, splitting, and departure.

The key idea of the decomposition method in QNA is to characterize the arrival processes in the network by the arrival rates and the squared coefficients of variation. Obviously the arrival rates of customers of individual classes are not changed by the separation of the priority server into the virtual servers described above. Therefore, the arrival rates can be obtained by solving the traffic rate equations (Eq. 4.1). On the other hand, the squared coefficients of variation can be derived by considering the departure processes from the virtual servers as described below.

In QNA, the squared coefficient of variation of interdeparture times, c_{di}^2, from a single server node i is approximated as

$$c_{di}^2 = 1 + (1-\rho_i^2)(c_{ai}^2 - 1) + \rho_i^2\{\max(c_{si}^2, 0.2) - 1\}. \qquad (4.9)$$

We apply this formula to approximate the departure processes from the virtual servers. That is, substituting Eqs. 4.3, 4.4, 4.7, and 4.8 into the above equation, the squared coefficient of variation of the interdeparture times for class k customers at node i can be approximated as

$$\begin{aligned} c^2_{d1i} &= 1+(1-\rho^2_{1i})(c^2_{a1i}-1)+\rho^2_{1i}\left[max\{c^2_{s1i}+\frac{h_{2i}\rho_{2i}}{h_{1i}\rho_{1i}}(1+c^2_{s2i}),0.2\}-1\right], && (4.10)\\ \text{and} & && (4.11)\\ c^2_{d2i} &= \frac{\rho^2_{2i}}{(1-\rho_{1i})^2}max\{c^2_{s2i}+\frac{\rho_{1i}}{h_{2i}\rho_{2i}}\left[h_{1i}(c^2_{a1i}+c^2_{s1i})-(1-\rho_{1i}-\rho_{2i})h_{2i}\right],0.2\}\\ &\quad +\frac{1-\rho_{1i}-\rho_{2i}}{(1-\rho_{1i})^2}(1-\rho_{1i}-\rho_{2i}+\rho_{1i}\rho_{2i})c^2_{a2i}. && (4.12)\end{aligned}$$

For the other two network operations, merging and splitting, we can use the approximations in QNA directly, since the separation of the priority servers has no effect on these operations. The approximation formula for the squared coefficient of variation of superposed (merged) interarrival times for class k customers to node j, c^2_{akj}, is given by

$$c^2_{akj}=1-w_{kj}+w_{kj}\sum_{i=0}^{N}p_{kij}c^2_{akij}, \qquad (4.13)$$

where

$$\begin{aligned} p_{kij} &= \lambda_{ki}q_{kij}/\lambda_{kj},\\ p_{k0j} &= \lambda_{k0j}/\lambda_{kj}, \quad \text{and} && (4.14)\\ w_{kj} &= \left[1+4(1-\rho_{kj})^2\left\{(\sum_{i=0}^{N}p^2_{kij})^{-1}-1\right\}\right]^{-1}.\end{aligned}$$

The squared coefficient of variation for the split departure process from node i to node j is given by

$$c^2_{akij}=p_{kij}c^2_{dki}+1-p_{kij}. \qquad (4.15)$$

Combining Eqs. 4.10, 4.12, 4.13, and 4.15, we obtain the system of equations that leads the squared coefficients of variation for the arrival processes,

$$c^2_{akj}=a_{kj}+\sum_{i=1}^{N}c^2_{aki}b_{kij}, \quad k=1,2; j=1,2,\cdots,N, \qquad (4.16)$$

where

$$\begin{aligned} a_{kj} &= 1+w_{kj}\{p_{k0j}c^2_{k0j}-1+\sum_{i=1}^{N}p_{kij}(1-q_{kij}+q_{kij}\rho^{*2}_{ki}x_{ki})\}, && (4.17)\\ b_{1ij} &= w_{1j}p_{1ij}q_{1ij}(1-\rho^2_{1i}), && (4.18)\\ b_{2ij} &= w_{2j}p_{2ij}q_{2ij}\frac{1-\rho_{1i}-\rho_{2i}}{(1-\rho_{1i})^2}(1-\rho_{1i}+\rho_{2i}+\rho_{1i}\rho_{2i}), && (4.19)\\ x_{1i} &= \max\{c^2_{s1i}+\frac{h_{2i}\rho_{2i}}{h_{1i}\rho_{1i}}(1+c^2_{s2i}),0.2\}, \quad \text{and} && (4.20)\\ x_{2i} &= \max\{c^2_{s2i}+\frac{\rho_{1i}}{h_{2i}\rho_{2i}}\left[h_{1i}(c^2_{a1i}+c^2_{s1i})-(1-\rho_{1i}-\rho_{2i})h_{2i}\right],0.2\}. && (4.21)\end{aligned}$$

The set of equations represented by Eq. 4.16 is called the *traffic variability equations*. Solving the two systems of equations, the traffic rate equations (Eq. 4.1) and the traffic variability equations (Eq. 4.16) , for each class, we finally obtain two-parameters of arrival processes in the single-class networks.

4.3.4 Congestion measures

By the decomposition method described in the previous subsection, the arrival process to node i in the single-class network for class k customers is characterized by two-parameters, the arrival rate and the squared coefficient of variation. For the service time distribution, we have derived the mean service time and the squared coefficient of variation using the two-parameter virtual server method described in Sect. 4.3.2.

Using these four parameters, we can approximate congestion measures for node i in the single-class network by regarding the node as a standard $GI/G/1$ queue. We adopt the approximation formulas used in QNA for the $GI/G/1$ queue. Since we are focusing on a single node in the network of each class, we omit the subscripts k and i indexing the class and node throughout this subsection.

First, the mean waiting time is approximated as

$$E[W] = \frac{(c_a^2 + c_s^{*2})\rho^* h^*}{2(1-\rho^*)} \exp\left[-\frac{2(1-\rho^*)(1-\min(c_a^2,1))^2}{3\rho^*(c_a^2 + c_s^{*2})}\right]. \tag{4.22}$$

When $c_a^2 < 1$, the above formula is consistent with the approximation of Krämer and Langenbach-Belz [116].

Then the mean number of customers in the node can be obtained from Little's formula as

$$E[N_q] = \rho^* + \lambda E[W]. \tag{4.23}$$

The probability of delay is approximated as

$$\sigma = \rho^* + (c_a^2 - 1)\rho^*(1-\rho^*)g, \tag{4.24}$$

where

$$g = \begin{cases} \dfrac{c_a^2 + c_s^{*2}}{1 + \rho^*(c_s^{*2} - 1) + \rho^{*2}(4c_a^2 + c_s^{*2})}, & c_a^2 < 1, \\ \\ \dfrac{4\rho^*}{c_a^2 + \rho^{*2}(4c_a^2 + c_s^{*2})}, & c_a^2 \geq 1. \end{cases} \tag{4.25}$$

Equation 4.24 is the Krämer and Langenbach-Belz approximation for the probability of delay.

The variance of the waiting time is approximated as

$$Var[W] = (E[W])^2 \frac{c_D^2 + 1 - \sigma}{\sigma}, \tag{4.26}$$

where c_D^2 is the squared coefficient of variation of the conditional delay, given that the server is busy in the $GI/G/1$ queue, which is approximated in QNA as

$$c_D^2 = \frac{2\rho^* - 1 + 4(1-\rho^*)(2\min(c_s^{*2},1) + 1)}{3(c_s^{*2} + 1)}. \tag{4.27}$$

Table 4.1: Mean total sojourn time in a tandem network

							$E(T_1)$			$E(T_2)$		
case	λ_{101}	λ_{201}	h_{11}	h_{21}	h_{12}	h_{22}	exact	Schmitt	proposed	exact	Schmitt	proposed
1	0.1	0.4	1.0	1.0	1.0	1.0	3.084	3.110	3.116	4.212	4.206	4.280
2	0.1	0.4	1.0	0.5	1.0	0.4	2.408	2.440	2.445	1.680	1.635	1.647
3	0.2	0.4	1.0	0.5	1.0	0.5	2.709	2.750	2.755	2.349	2.246	2.300
4	0.1	0.4	0.5	1.0	0.5	1.0	1.970	1.895	1.895	3.618	3.624	3.638
5	0.2	0.4	0.5	1.0	0.5	1.0	2.073	1.998	2.004	3.980	3.992	4.040
6	0.3	0.4	0.5	1.0	0.5	1.0	2.187	2.118	2.128	4.442	4.464	4.560
7	0.4	0.4	0.5	1.0	0.5	1.0	2.315	2.248	2.270	5.041	5.076	5.280

The variance of the number of customers in the node, which is derived from modifying the $M/G/1$ formula for the $GI/G/1$ queue, is approximated as

$$Var[N_q] = (E[N_q])^2 Y_1 Y_2 / Y_3, \tag{4.28}$$

where

$$\begin{aligned} Y_1 &= \lambda E[W] + \rho^* + \rho^{*2} c_s^{*2} + \lambda^2 Var[W], \\ Y_2 &= (1 - \rho^* + \sigma)/\max\{1 - \sigma + \rho^*, 10^{-6}\}, \quad \text{and} \\ Y_3 &= \max\{\rho^* + \lambda E[W], 10^{-6}\}. \end{aligned} \tag{4.29}$$

The maximization is used in the above equations to avoid division by zero.

Similarly, we can apply the approximation formulas in QNA for network performance measures, for example, the formula for the expected total sojourn time for an arbitrary customer is given by

$$E[T] = \sum_{i=1}^{N} \left(\lambda_i / \sum_{j=1}^{N} \lambda_{0j} \right) (h_i + E[W_i]), \tag{4.30}$$

where $\sum_{j=1}^{N} \lambda_{0j}$ represents the throughput, which is defined as the total external arrival rate.

4.4 Numerical results and validation

In the following, we evaluate the approximation method by means of several test-bed queueing network models.

4.4.1 Two-stage tandem network model

First, we consider Markovian tandem models $M_1, M_2/M_1, M_2/1 \rightarrow \bullet/M_1, M_2/1$, which were analyzed by Schmitt [47] using the virtual server model with a state-dependant service rate. Comparisons among the exact solutions, Schmitt's results, and the results

Table 4.2: Input data for 2-stage tandem models

case	λ_{101}	λ_{201}	c^2_{a101}	c^2_{a201}	c^2_{s11}	c^2_{s21}
1	0.2	0.4	0.5	0.5	1.0	1.0
2	0.4	0.4	0.5	0.5	1.0	1.0
3	0.2	0.4	1.0	1.0	1.0	1.0
4	0.4	0.4	1.0	1.0	1.0	1.0
5	0.2	0.4	2.0	2.0	1.0	1.0
6	0.4	0.4	2.0	2.0	1.0	1.0
7	0.2	0.4	0.5	0.5	0.0	0.0
8	0.4	0.4	0.5	0.5	0.0	0.0
9	0.2	0.4	1.0	1.0	0.0	0.0
10	0.4	0.4	1.0	1.0	0.0	0.0
11	0.2	0.4	2.0	2.0	0.0	0.0
12	0.4	0.4	2.0	2.0	0.0	0.0

of the proposed method for the mean total sojourn times of individual classes are listed in Table 4.1. It can be seen that the results of the method are accurate, but slightly less accurate than Schmitt's results. Because the approximation provides exact results for the first node of this model, the errors are mainly caused by the approximation for the departure processes from the first node.

Next, we consider more general tandem models, $GI_1, GI_2/G_1, G_2/1 \to \bullet/M_1, M_2/1$, where service times of class k customers at node 1 are subjecting to a general distribution with mean h_{k1} and coefficient of variation c_{sk1}, and interarrival times of class k customers are subjecting to a general distribution with mean λ_{k01}^{-1} and coefficient of variation c_{ak01}. We assume that the mean service times for classes 1 and 2 customers at both nodes are common and equal to 1, i.e., $h_{11} = h_{21} = h_{12} = h_{22} = 1$. For the twelve cases shown in Table 4.2, we compare our results and simulation results in Table 4.3 for the mean waiting time of class k customers at node i, $E[W_{ki}]$, and the expected total sojourn time of class k customers, $E[T_k]$. In the simulation, we set the interarrival time distributions as the Erlang distribution for $c^2_{ak01} = 0.5$, and as the Hyperexponential distribution with balanced means for $c^2_{ak01} = 2.0$. We see that the accuracy is fairly good for the mean waiting times at node 1 for both priority customers. It is also observed that the approximation tends to overestimate the mean waiting times at node 2, especially for class 2 customers. This is primarily due to the accuracy of the approximation formula for the departure processes (Eq. 4.12).

4.4.2 Basic component model

The following queueing network topologies that are basic components found in any queueing network are examined:

(1) Splitting system (Fig. 4.3(a)),

(2) Merging system (Fig. 4.3(b)),

Table 4.3: Comparisons with simulation results for 2-stage tandem models

case	sim./approx.	$E[W_{11}]$	$E[W_{21}]$	$E[W_{12}]$	$E[W_{22}]$	$E[T_1]$	$E[T_2]$
1	simulation	0.602	1.266	0.593	1.461	3.205	4.721
	(95% confidence)	± 0.036	± 0.201	± 0.029	± 0.166	± 0.071	± 0.275
	approximation	0.609	1.240	0.677	1.600	3.286	4.841
2	simulation	1.038	4.693	1.077	5.625	4.112	12.31
	(95% confidence)	± 0.099	± 1.107	± 0.066	± 0.889	± 0.163	± 1.874
	approximation	1.090	4.700	1.300	7.200	4.383	13.88
3	simulation	0.743	1.875	0.701	1.898	3.436	5.790
	(95% confidence)	± 0.070	± 0.112	± 0.060	± 0.361	± 0.155	± 0.361
	approximation	0.750	1.870	0.770	2.110	3.520	5.987
4	simulation	1.330	6.324	1.269	6.752	4.616	15.08
	(95% confidence)	± 0.063	± 1.302	± 0.094	± 1.163	± 0.159	± 1.412
	approximation	1.330	6.670	1.440	8.980	4.773	17.64
5	simulation	0.881	2.552	0.850	2.401	3.755	6.941
	(95% confidence)	± 0.097	± 0.249	± 0.083	± 0.478	± 0.127	± 0.697
	approximation	0.875	2.940	0.890	3.110	3.765	8.050
6	simulation	1.641	10.37	1.497	8.395	5.144	20.76
	(95% confidence)	± 0.039	± 1.700	± 0.130	± 1.865	± 0.151	± 3.300
	approximation	1.670	10.30	1.720	12.60	5.387	24.89
7	simulation	0.263	0.407	0.519	1.079	2.786	3.489
	(95% confidence)	± 0.007	± 0.023	± 0.043	± 0.211	± 0.054	± 0.238
	approximation	0.239	0.314	0.633	1.200	2.872	3.519
8	simulation	0.429	1.512	0.938	4.208	3.367	7.730
	(95% confidence)	± 0.012	± 0.091	± 0.079	± 0.815	± 0.085	± 0.875
	approximation	0.423	1.370	1.140	4.720	3.559	8.083
9	simulation	0.370	0.912	0.602	1.308	2.954	4.223
	(95% confidence)	± 0.018	± 0.049	± 0.034	± 0.092	± 0.061	± 0.122
	approximation	0.375	0.937	0.755	1.790	3.130	4.725
10	simulation	0.651	3.358	1.050	4.972	3.698	10.33
	(95% confidence)	± 0.029	± 0.305	± 0.053	± 1.114	± 0.086	± 1.272
	approximation	0.667	3.330	1.330	6.670	4.000	12.00
11	simulation	0.440	1.581	0.697	2.013	3.145	5.603
	(95% confidence)	± 0.024	± 0.088	± 0.049	± 0.315	± 0.077	± 0.344
	approximation	0.500	2.000	0.875	2.790	3.375	6.790
12	simulation	0.864	6.409	1.269	6.915	4.141	15.33
	(95% confidence)	± 0.029	± 0.722	± 0114	± 1.209	± 0.134	± 1.880
	approximation	1.000	7.000	1.610	10.20	4.613	19.24

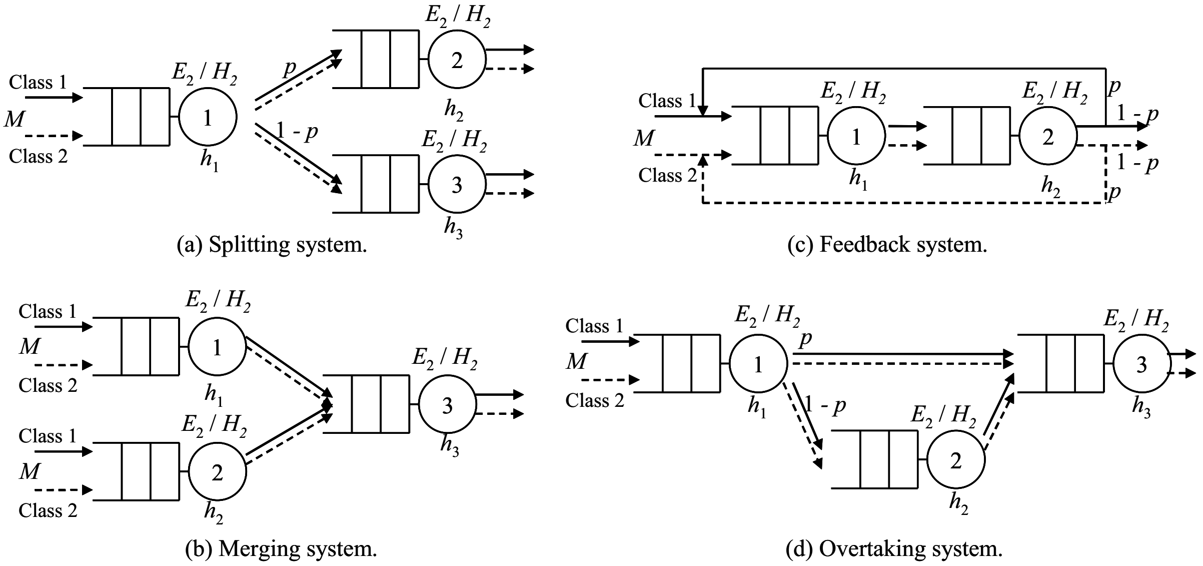


Figure 4.3: Basic component models

(3) Feedback system (Fig. 4.3(c)), and

(4) Overtaking system (Fig. 4.3(d)).

We consider the cases where the following conditions are satisfied.

- The external arrival process to node i is Poissonian with rate $\lambda_{k0i} = \lambda_i^{(k)}$ for class k customers.

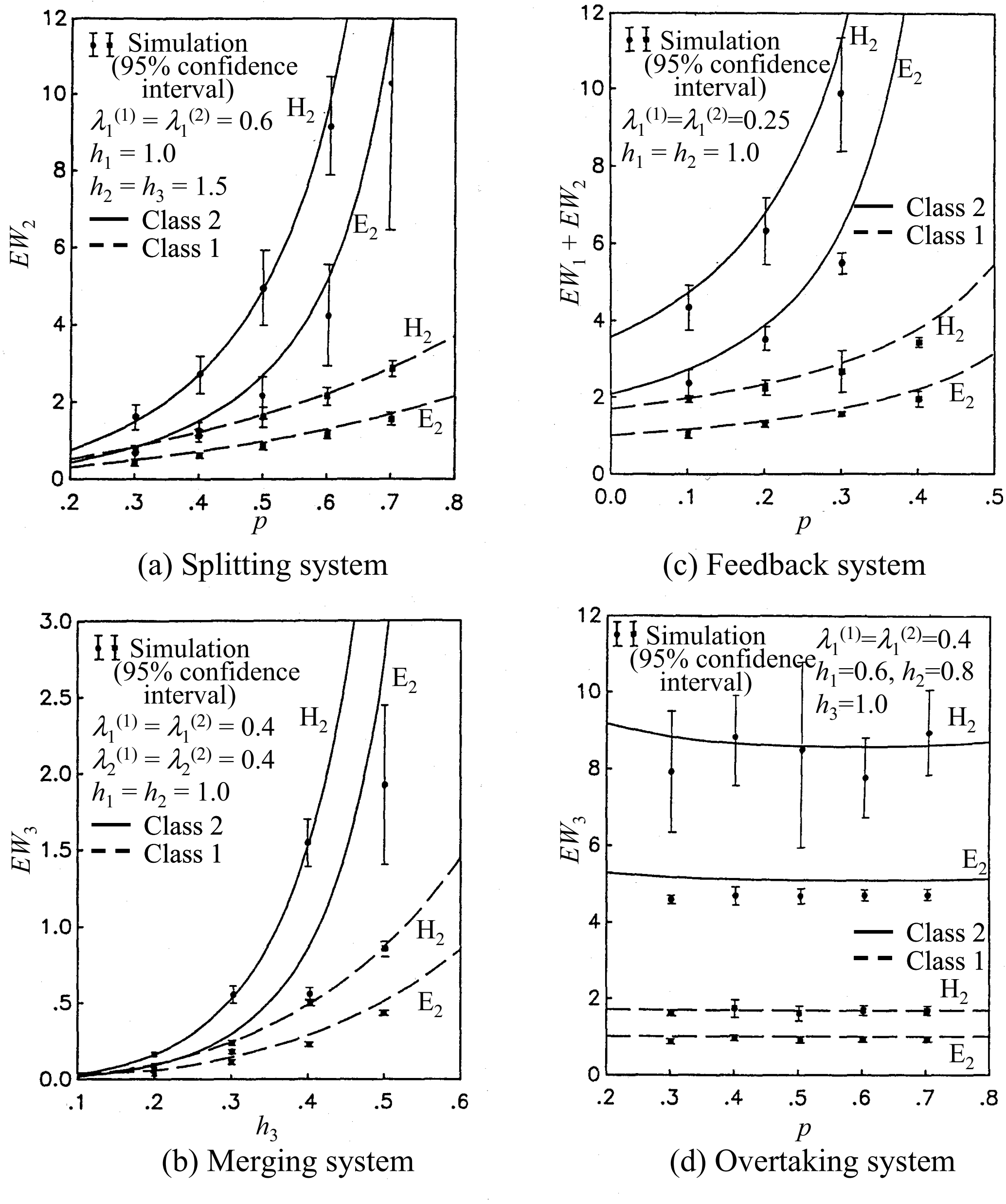
- Types of service time distributions are identical through all the nodes in these networks. The type considered here is an Erlang-2 (E_2) or a two-stage Hyperexponential (H_2). The service time distribution has mean h_i at node i for both class customers. The squared coefficient of variation of it is 0.5 when the type is E_2, while it is set equal to 1.5 when the type is H_2.

We validate the approximation for the mean waiting time and the variance of the waiting time in the following.

Mean waiting time

First, we consider a splitting system of three queues (Fig. 4.3(a)), where the external arrival process to node 1 is Poissonian with rate $\lambda_1^{(k)}$ for class k customers.

Comparisons with simulation results for the mean waiting time at node 2 are displayed in Fig. 4.4 (a). In this experiment, the splitting probability p of each class customer being routed from node 1 to node 2 is varied from 0.2 to 0.8 for $\lambda_1^{(1)} = \lambda_1^{(2)} = 0.4$ and $h_1 = 1.0$, $h_2 = h_3 = 1.5$. The approximation gives good results for both service time distributions, E_2 and H_2. This shows that the approximate procedure is accurate for the splitting system.



(a) Splitting system

(c) Feedback system

(b) Merging system

(d) Overtaking system

Figure 4.4: Mean waiting time for basic component models

Second, we examine a merging system of three queues (Fig. 4.3 (b)), where the external arrival processes to node 1 and node 2 are Poissonian with rates $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ for class k customers, respectively.

Comparisons with simulation results for the mean waiting time at node 3 are shown in Fig. 4.4 (b). In this case, the mean service time at node 3, h_3, is varied from 0.1 to 0.6 for $\lambda_i^{(k)} = 0.4$ for class k customers at node i, while $h_1 = h_2 = 1.0$. For E_2 service time distributions, the approximation seems to overestimate the mean waiting time for both class customers. This finding shows the limitation of approximating the mean waiting time for merging in the case of superposed customers flowing into the queue whose coefficient of variation of service time distribution is less than one.

Next, we examine a feedback system of two queues (Fig. 4.3 (c)), where the external arrival process to node 1 is Poissonian with rate $\lambda_1^{(k)}$ for class k customers.

Figure 4.4 (c) compares the approximation with simulation results for the mean waiting time at node 1 plus the mean waiting time at node 2. The feedback probability p at node 2 is common for both classes and is varied from 0.0 to 0.5 for $\lambda_1^{(1)} = \lambda_1^{(2)} = 0.25$ and $h_1 = h_2 = 1.0$. As before, the approximation seems to overestimate the mean waiting time for E_2 service time distributions. This is expected because the interarrival process to node 1 is the merging stream formed by the external arrival process and the departure process from node 2.

Finally, we consider an overtaking system of three queues (Fig. 4.3 (d)), where the external arrival process to node 1 is Poissonian with rate $\lambda_1^{(k)}$ for class k customers.

Comparisons with simulation results for the mean waiting time at node 3 are shown in Fig. 4.4 (d). As before, the common splitting probability p is varied from 0.2 to 0.8 for $\lambda_1^{(1)} = \lambda_1^{(2)} = 0.4$ and $h_1 = 0.6$, $h_2 = 0.8$, and $h_3 = 1.0$. Again, in this system, the approximation slightly overestimates the mean waiting time for class 2 customers, for E_2 service time distributions. The main reason for these discrepancies in this case is thought to be the same as that in the previous example of the merging system.

Variance of the waiting time

Next, we validate the approximation for the variance of the waiting time. Consider the four basic component models shown in Fig 4.3. The model assumptions for each component model are the same as before.

Table 4.4 shows comparisons with simulation results for the variance of waiting time when all the service time distributions are E_2, and Table 4.5 shows those when the service time distributions are H_2. For the E_2 service time distributions, the approximation overestimates the variance of the waiting time. This is because the approximation for the mean waiting time for the E_2 service time distributions gives overestimations (Eq. 4.26). On the other hand, the approximation gives good results for the H_2 service time distributions.

4.4.3 Packet switching network model

We apply the approximation method to an end-to-end delay analysis for a packet switching network for voice and data. Consider the tandem network model with three nodes depicted in Fig. 4.5. In this model, we set the following conditions.

Table 4.4: Variance of waiting time at node i, Var[W_i], when service time distributions are all E_2

		class 1		class 2	
system	parameter	approx.	sim. ±95%†	approx.	sim. ±95%
Splitting	$p = 0.4$	1.482*	1.230±0.331	8.326*	6.915±2.596
(Var[W_2])	$p = 0.6$	2.696*	2.413±0.476	57.10*	44.90±22.90
Merging	$h_3 = 0.3$	0.059	0.039±0.002	0.317	0.116±0.016
(Var[W_3])	$h_3 = 0.4$	0.156	0.112±0.008	1.847	0.791±0.149
Feedback	$p = 0.1$	1.580	1.281±0.168	11.31	8.086±2.577
(Var[W_1]+Var[W_2])	$p = 0.3$	2.333	2.121±0.121	44.71	37.12±6.971
Overtake	$p = 0.4$	1.353	1.114±0.114	46.43*	42.08±19.58
(Var[W_3])	$p = 0.6$	1.347	1.125±0.057	45.84*	41.96±19.69

†: 95% confidence interval, * means within 95% confidence interval

Table 4.5: Variance of waiting time at node i, Var[W_i], when service time distributions are all H_2

		class 1		class 2	
system	parameter	approx.	sim. ±95%†	approx.	sim. ±95%
Splitting	$p = 0.4$	5.746*	5.480±1.380	32.02*	25.71±10.53
(Var[W_2])	$p = 0.6$	10.62*	10.05±0.476	214.4*	150.6±65.60
Merging	$h_3 = 0.3$	0.226*	0.234±0.042	1.178*	1.053±0.256
(Var[W_3])	$h_3 = 0.4$	0.603*	0.614±0.074	6.632*	5.548±1.093
Feedback	$p = 0.1$	6.049*	6.389±1.645	39.54*	33.17±10.77
(Var[W_1]+Var[W_2])	$p = 0.3$	9.094*	7.856±1.803	161.3*	121.1±40.75
Overtake	$p = 0.4$	5.133*	5.523±1.632	145.2*	130.8±39.24
(Var[W_3])	$p = 0.6$	5.094	4.904±0.814	141.9	108.8±32.29

†: 95% confidence interval, * means within 95% confidence interval

(1) Packetized voice and data flow into the network. The packet stream from a single voice source is modeled by arrivals at fixed intervals of T_0 during talkspurt and no arrivals during silence (see Fig. 4.6). We assume that the voice packetization period (T_0) is 16 *ms*, the talkspurt and silence periods are independent and exponentially distributed with means of $\alpha^{-1} = 352$ *ms* and $\beta^{-1} = 650$ *ms*. This is the same voice-source model used in Ref. [117]. For packetized data, a Poissonian arrival process is assumed.

(2) The arrival processes to node 1 are composed of the aggregated voice arrival process resulting from the merging of N_1 voice sources and a data arrival process with the arrival rate λ_2. After service completions at node 1 and node 2, both voice and data packets depart from the network with probability p and go to the next node with probability $1 - p$. To nodes 2 and 3, voice packets from pN_1 sources and data packets with rate $p\lambda_2$ enter from outside.

(3) At each node, nonpreemptive priority is given to voice packets. The service time

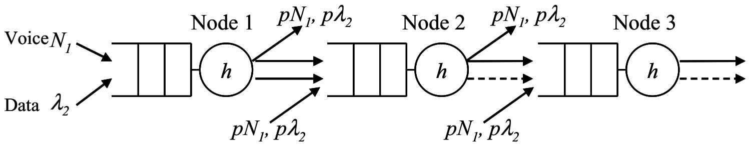


Figure 4.5: Packet switching network model

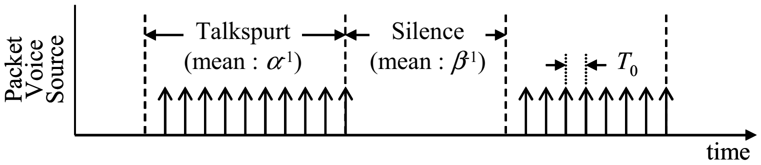


Figure 4.6: Packet arrival process from one voice source

(packet transmission time) is of fixed length h for both voice and data packets.

Then in our notations for the approximation model, parameters are set as follows.

$$\begin{aligned}
N_1 &= 3, \\
\lambda_{101} &= \frac{N_1\beta}{(\alpha+\beta)T_0}, \ \lambda_{102} = \lambda_{103} = \frac{pN_1\beta}{(\alpha+\beta)T_0}, \\
\lambda_{201} &= \lambda_2, \ \lambda_{202} = \lambda_{203} = p\lambda_2, \\
c^2_{a101} &= \left(\frac{\alpha}{\alpha+\beta}\right)^2 \left(\frac{2}{\alpha T_0} - 1\right), \\
c^2_{a2i} &= 1 \ \text{for } i = 1,2,3, \\
h_{ki} &= h, \ c_{ski} = 0 \ \text{for } \ k = 1,2; \ i = 1,2,3, \quad \text{and} \\
Q_k &= \begin{bmatrix} 0 & 1-p & 0 \\ 0 & 0 & 1-p \\ 0 & 0 & 0 \end{bmatrix} \ \text{for } \ k = 1,2,
\end{aligned}$$

where c^2_{a101} is the squared coefficient of variation for the arrival process from one voice source. Using c^2_{a101}, the squared coefficient of variation for the superposed voice arrival process c^2_{a1i} can be obtained by applying the approximation formula for merging (Eq. 4.13). The mean end-to-end delay (mean total sojourn time) and the mean sojourn time in node 1 versus the number of voice sources are shown in Fig. 4.7.

In this figure, it is assumed that $\lambda_2 = 0.6 \ ms^{-1}$, $h = 1/3 \ ms$, and $p = 0.7$. We see that the results are sufficiently accurate for both light and heavy traffic. It is also observed that the mean end-to-end delay is less than three times the mean sojourn time in node 1, although the traffic intensities for all nodes are the same. For example, for 100 voice sources, the mean end-to-end delay for voice packets ($E[T_1]$) is 3.04 ms, while the mean sojourn time in node 1 for voice packets ($E[W_1] + h$) is 1.18 ms. This indicates that the approximation method can capture the effect of the departure processes from the preceding node.

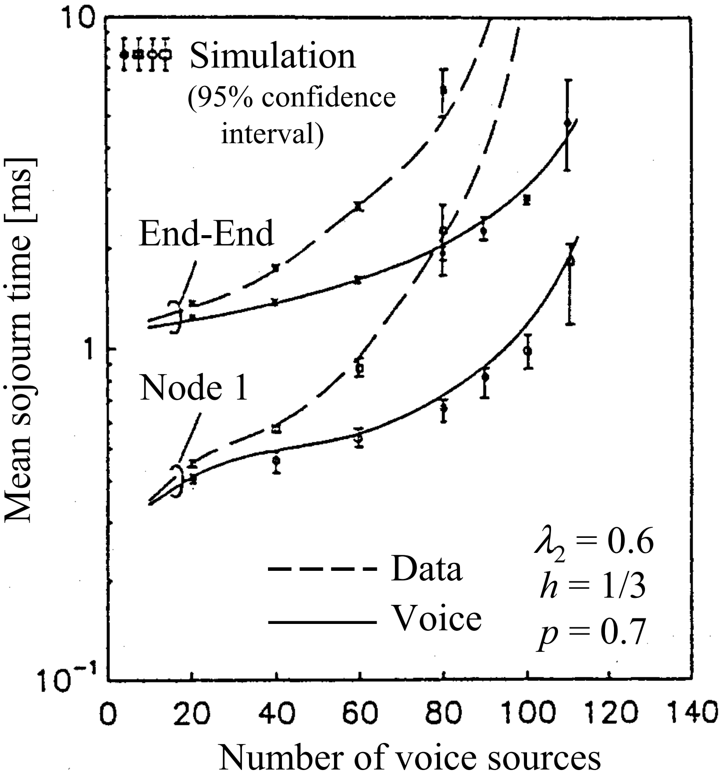


Figure 4.7: Mean sojourn time for voice and data packets

4.5 Conclusion

We have proposed a new approximation method for analysis of renewal queueing networks with nonpreemptive priorities. The method is based on the virtual server and decomposition methods. We have established a virtual server model characterizing the service time distribution by two parameters. The approximation method has been validated by the exact solutions to Schmitt's two-stage tandem models and by simulation of the basic component models of complex networks and of the packet switching network model for voice and data. The results indicate that the accuracy of the approximation method is sufficient for practical use.

Although only nonpreemptive priority is considered in this chapter, the basic idea is quite general and an extension to the preemptive case is possible. Expansion to more general models such as multiclass structures and closed queueing networks remains for further study.

Chapter 5

Link capacity allocation control for multiclass alternate routing networks

5.1 Introduction

This chapter also discusses traffic control at the transmission network. While packet level control was considered in the previous chapter,the traffic control considered here is call (connection) level reservation control for wideband and alternate-routed calls.

Dynamic routing for the single-class network has been studied extensively in the last two decades, mostly for circuit-switched networks (see, for example, [59, 60, 61, 62] and references therein). For single-class nonhierarchical networks with alternate routing, Nakagome and Mori [63] carried out approximation analyses and revealed the existence of network instability. That is, for certain loads the network operates in two states: a low network blocking state and a congested state. This bistable behavior causes a high blocking rate during overload. Later, Krupp [64] showed that the bistable behavior can be stabilized by reserving a small number of trunks for direct-routed calls. The effect of the trunk reservation for more general nonsymmetric networks can be found in Refs. [65, 66].

Performance analysis of loss networks such as telephone networks with dynamic routing is fundamentally difficult because dynamic routing destroys the product form solutions. It is, therefore, of interest to develop computational procedures that accurately approximate blocking probabilities for loss networks with dynamic routing. One such method is the *reduced load approximation method* (also referred to as the Erlang fixed-point approximation) [62]. This approximation assumes that blocking is independent from link to link, giving rise to a set of fixed-point equations whose solution supplies approximations for blocking. The analytical method of Krupp [64] and Akinpelu [65] for sequential alternate routing with trunk reservation is also based on this approximation. Kelly [67] gave a generalized reduced load approximation that can be adapted to essentially any dynamic routing scheme. Chung et al. [68] examined the accuracy and the computational requirements of the approximation procedure for a particular routing scheme, namely, least-loaded routing.

For multiclass alternate routing networks, not only the network instability but also the imbalance of blocking probabilities for different traffic classes occur. Since these unstable phenomena lead to ineffective use of network resources, traffic control which stabilize the network performance is required.

Performance characteristics and control scheme for multiclass alternate routing networks have not yet been sufficiently clarified. Kelly [71], Chung and Ross [72] extended the reduced load approximation for multiclass networks with state-dependent routing. These works have not considered the effect of reservation control for direct-routed calls and wideband calls. Greenberg and Srikant [74] proposed computational techniques for multiclass sequential routing networks with state-dependent admission control and general topology. Medhi and Sukiman [75] gave a comparative study of multiclass dynamic routing schemes with reservation control by simulation.

In this chapter, a link capacity allocation control scheme and its evaluation method for multiclass alternate routing networks are proposed. End-to-end blocking probabilities and bistable behavior are clarified by approximate analysis for various traffic mixes and allocation control parameters. In addition, the setting of allocation control parameters to suppress the instability and to ensure maximum useful network throughput under various traffic conditions is discussed.

The reminder of this chapter is organized as follows. In Sect. 5.2, a multiclass fully connected network model with sequential alternate routing treated in this chapter is described. A link capacity allocation control scheme, which is a combination of reservation control for wideband calls and for direct-routed calls, is proposed in Sect. 5.3. Then, we propose an approximation method to evaluate end-to-end blocking probabilities based on the reduced load approximation in Sect. 5.4. In Sect. 5.5, the approximation method is validated by means of comparisons with simulation results. In Sect. 5.6, some numerical results are given. First, it is shown that the imbalance of individual end-to-end blocking probabilities and network instability exist in a multiclass alternate routing network without the control. Then, the effects of the proposed control to prevent the imbalance and suppress the instability are demonstrated. Furthermore, the parameter setting problem of allocation capacities are explored. Finally, Sect. 5.7 gives some conclusions.

5.2 Network model

An example of the network model treated in this chapter is shown in Fig. 5.1. The model is assumed to be a fully connected, symmetric, and nonhierarchical network with N nodes. The link capacity of every one-way link is assumed to be identical. The offered load to each node-pair is also assumed to be the same. The routing procedure is sequential alternate routing, that is, every fresh call attempts the direct link first. If the link is busy, the call will attempt alternate paths up to m times. The calls overflowing the mth alternate path are lost and cleared. The number of links of the alternate path is assumed to be 2 (all the alternate paths with length 2 are allowable). We use the following notation:

K: number of traffic classes,

L: link capacity,

$w^{(k)}$: bandwidth to serve a kth class call,

$a_f^{(k)}$: offered load of the kth class direct-routed calls for each link,

u_l: total capacity in use in link l.

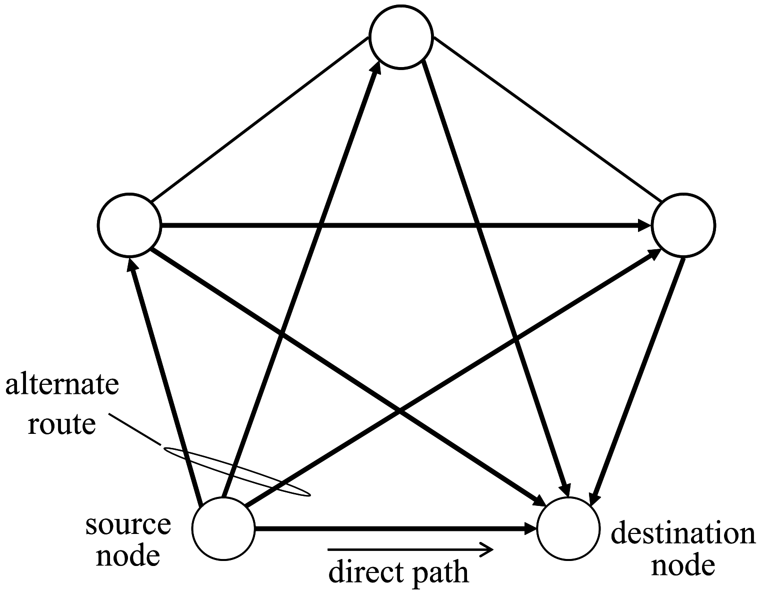


Figure 5.1: A network model example

5.3 Link capacity allocation control

Under the link capacity allocation control, the allowable capacity in each link is predetermined for direct- and alternate-routed calls of each class (see Fig. 5.2). The definition of the link capacity allocation control is as follows.

The kth class direct-routed call arriving upon link l is accepted iff $u_l \leq c_f^{(k)} - w^{(k)}$, similarly, the kth class alternate-routed call arriving upon link l is accepted iff $u_l \leq c_o^{(k)} - w^{(k)}$, where

$c_f^{(k)}$: allocated capacity for k-class direct-routed calls,

$c_o^{(k)}$: allocated capacity for k-class alternate-routed calls.

This control scheme can be considered as a generalization of trunk reservation control for both wideband calls and direct-routed calls. With this scheme, the following effects are expected:

(1) Balancing of blocking probabilities for different traffic classes by reducing the allowable capacity for narrowband calls. (The effect of trunk reservation for wideband calls).

(2) Eliminating instability due to alternate routing by reducing the allowable capacity for alternate-routed calls. (The effect of trunk reservation for direct-routed calls).

5.4 Approximation analysis

To simplify the analysis:

(Al) arrival processes of both direct- and alternate-routed calls are assumed to be Poissonian and call holding time is to be exponentially distributed, and

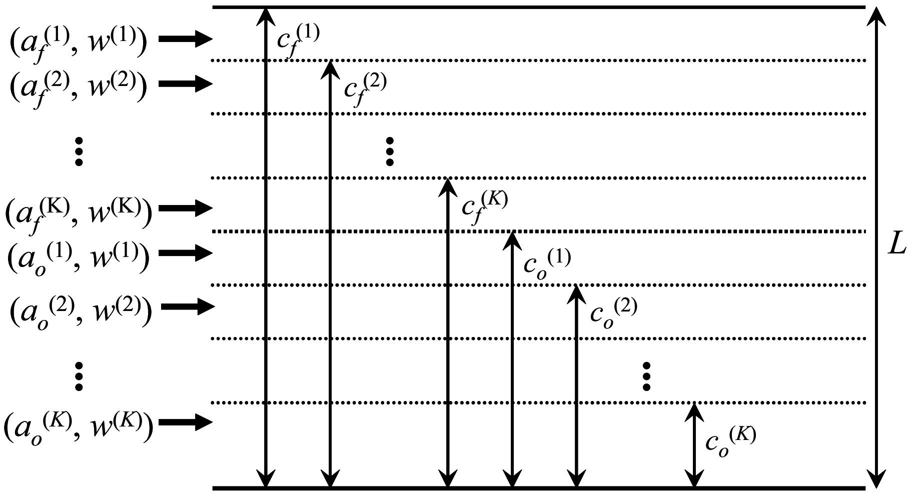


Figure 5.2: Link capacity allocation control

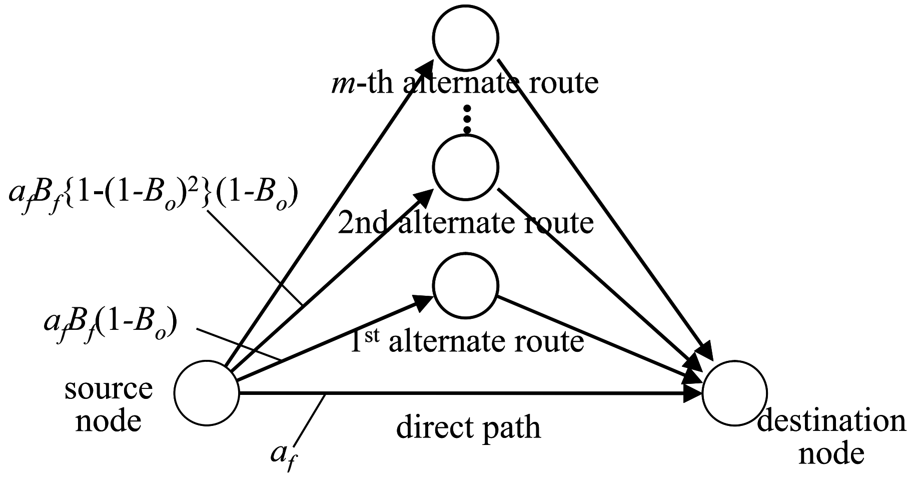


Figure 5.3: Offered load to each link

(A2) control times to connect and to disconnect calls are assumed to be very short compared to the call holding times and are negligible.

The end-to-end blocking probability, $B_E^{(k)}$ for class k calls can be found by solving two nonlinear equations expressed with the following variables simultaneously:

(1) $a_o^{(k)}$: offered load of class k alternate-routed calls to each link,

(2) $B_f^{(k)}$ ($B_o^{(k)}$): link blocking probability for class k direct- (alternate-) routed calls.

In the following we derive the equations on the above variables. The offered load of alternate-routed calls is considered to be the sum of 1st through mth alternately routed traffic as shown in Fig. 5.3. The offered load of the ith alternately routed traffic can be defined as the amount of traffic overflowing the $(i-1)$th alternate path and offered to the ith alternate path. Note that an overflowing call is offered to the link only if the other link of a two-link alternate path is free, i.e., with probability $1 - B_o^{(k)}$ for class k calls. The offered load of class k alternate-routed calls is thus given by

$$\begin{aligned}
a_o^{(k)} =& 2a_f^{(k)} B_f^{(k)} (1 - B_o^{(k)}) \\
& + 2a_f^{(k)} B_f^{(k)} \{1 - (1 - B_o^{(k)})^2\}(1 - B_o^{(k)}) \\
& + 2a_f^{(k)} B_f^{(k)} \{1 - (1 - B_o^{(k)})^2\}^2(1 - B_o^{(k)}) \\
& + \cdots \\
& + 2a_f^{(k)} B_f^{(k)} \{1 - (1 - B_o^{(k)})^2\}^{m-1}(1 - B_o^{(k)}) \\
=& \frac{2B_f^{(k)}}{1 - B_o^{(k)}} [1 - \{1 - (1 - B_o^{(k)})^2\}^m] a_f^{(k)}.
\end{aligned} \tag{5.1}$$

Because we assumed that the overflow traffic is also Poissonian, the link blocking probability can be derived using the heterogeneous state-dependent input loss model with trunk reservation. In the model, the exact analysis requires solving the system of balance equations of a huge state space. Therefore, we use the approximation based on the following equation proposed by Roberts [118]:

$$nQ(n) = \sum_{k=1}^{2K} a^{(k)} S_k(n - w^{(k)}) Q(n - w^{(k)}), \;\; n = 1, 2, \cdots, L, \tag{5.2}$$

where $Q(n)$ is the probability that n servers are busy (without loss of generality, the link capacity, L, is assumed to be an integer), and

$$S_k(n) = \begin{cases} w^{(k)}, & n < c^{(k)} - w^{(k)}, \\ 0, & otherwise, \end{cases} \tag{5.3}$$

$$a^{(k)} = \begin{cases} a_f^{(k)}, & k = 1, 2, \cdots, K, \\ a_o^{(k)}, & k = K+1, K+2, \cdots, 2K, \end{cases} \tag{5.4}$$

$$c^{(k)} = \begin{cases} c_f^{(k)}, & k = 1, 2, \cdots, K, \\ c_o^{(k)}, & k = K+1, K+2, \cdots, 2K, \end{cases} \quad \text{and} \tag{5.5}$$

$$w^{(K+k)} = w^{(k)}, \; k = 1, 2, \cdots, K. \tag{5.6}$$

This approximation is very good not only for the case of equal holding times [118], but also for the case of different holding times [119].

The distribution $Q(n)$ yields the link blocking probability for each class direct- and alternate-routed calls as follows.

$$\begin{aligned} B_f^{(k)} &= \sum_{n=c_f^{(k)}-w^{(k)}+1}^{L} Q(n), \ \ k=1,2,\cdots,K, \quad \text{and} \\ B_o^{(k)} &= \sum_{n=c_o^{(k)}-w^{(k)}+1}^{L} Q(n), \ \ k=1,2,\cdots,K. \end{aligned} \tag{5.7}$$

After solving Eqs. 5.1 and 5.7 simultaneously, the end-to-end blocking probability for class k calls is given by

$$B_E^{(k)} = B_f^{(k)}\{1-(1-B_o^{(k)})^2\}^m. \tag{5.8}$$

5.5 Validation

In this section, we evaluate the approximation by means of comparisons with simulation results under the following conditions,

$L = 6.144\ Mb/s$: link capacity,

$N = 7$: number of nodes,

$m = 5$: number of alternate routes,

$K = 2$: number of classes (wideband and narrowband calls)

$w^{(1)} = 512\ Kb/s$: bandwidth of wideband calls

$w^{(2)} = 32\ Kb/s$: bandwidth of narrowband calls

$A \equiv a_f^{(1)}w^{(1)} + a_f^{(2)}w^{(2)}\ erl * Mb/s$: total offered load

$r_a \equiv a_f^{(1)}w^{(1)}/a_f^{(2)}w^{(2)} = 1$: offered load ratio,

$c_f^{(1)} = L$, $c_o^{(1)} = L - R_d$, and

$c_f^{(2)} = L - (w^{(1)} - w^{(2)})$, $c_o^{(2)} = L - R_d - (w^{(1)} - w^{(2)})$: allocated link capacity,

where R_d represents the reservation capacity for direct-routed calls (see Sect. 5.6).

Comparisons with simulation results for the end-to-end blocking probabilities are displayed in Fig. 5.4. In this experiment, R_d was varied from 0.0 to L ($6.144\ Mb/s$) for $A = 4.2$, 4.4, and 5.0.

The results indicate that the approximation predicts the end-to-end blocking probabilities in multiclass alternate routing networks with reasonable accuracy. We can also see that the approximation captures the effect of control parameters on the network performance. The accuracy of the approximation relies on the following assumptions,

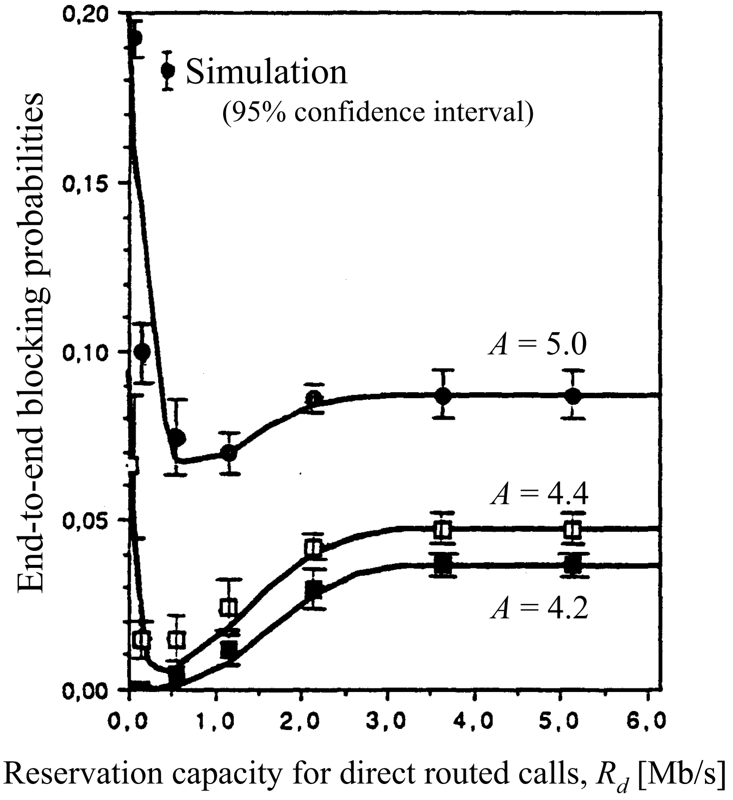


Figure 5.4: Comparisons with simulation results for the end-to-end blocking probabilities

(1) Poissonian assumption of the arrival processes of alternate-routed calls,

(2) independency assumption of call blocking between links in the network, and

(3) adoption of Roberts' approximation (Eq. 5.2).

5.6 Numerical results

In this section, we describe the numerical results and clarify the end-to-end blocking probabilities in a multiclass nonhierarchical alternate routing network for various traffic mixes and control parameters. The following conditions and notations are used in this section:

$K = 2$, $w^{(2)} = 32\ Kb/s$,

$N = 7$, $m = 5$ (7-node symmetric network),

$r_a \equiv a_f^{(1)} w^{(1)} / a_f^{(2)} w^{(2)}$: offered load ratio,

$r_w \equiv w^{(1)} / w^{(2)}$: bandwidth ratio,

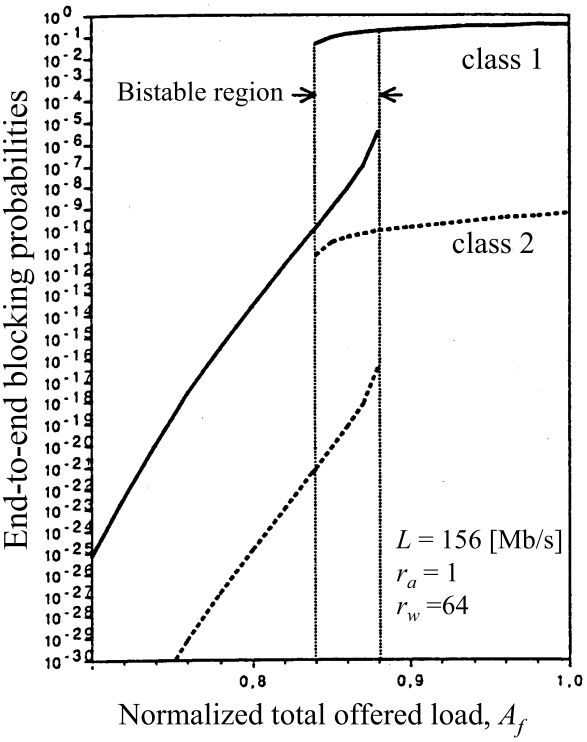


Figure 5.5: End-to-end blocking probabilities without controls

$A_f \equiv (a_f^{(1)} w^{(1)} + a_f^{(2)} w^{(2)})/L$: normalized total offered load.

5.6.1 Results without controls

To begin with, we describe the end-to-end blocking probabilities of a multiclass alternate routing network without controls. The results are obtained by the approximation described in Sect. 5.4 setting the control parameters for each class call as follows,

$$c_f^{(1)} = c_f^{(2)} = c_o^{(1)} = c_o^{(2)} = L \tag{5.9}$$

The end-to-end blocking probabilities for each class versus the normalized total offered load, A_f, with $L = 156\ Mb/s$, $r_a = 1$, and $r_w = 64$ are shown in Fig. 5.5. It is clearly shown that the wideband calls experience the higher call blocking. Furthermore, the bistable blocking behavior is observed as in the case of single class alternate routing networks [63, 64]. The bistable region of both classes occurs at the same value of offered load. This type of network instability causes a high blocking especially for wideband calls under heavy traffic conditions.

5.6.2 Equalizing blocking probabilities

We now apply the link capacity allocation control to balance the end-to-end blocking probabilities for different call classes. A control equalizing the blocking probabilities for each class is considered here. This control can be achieved by setting the control

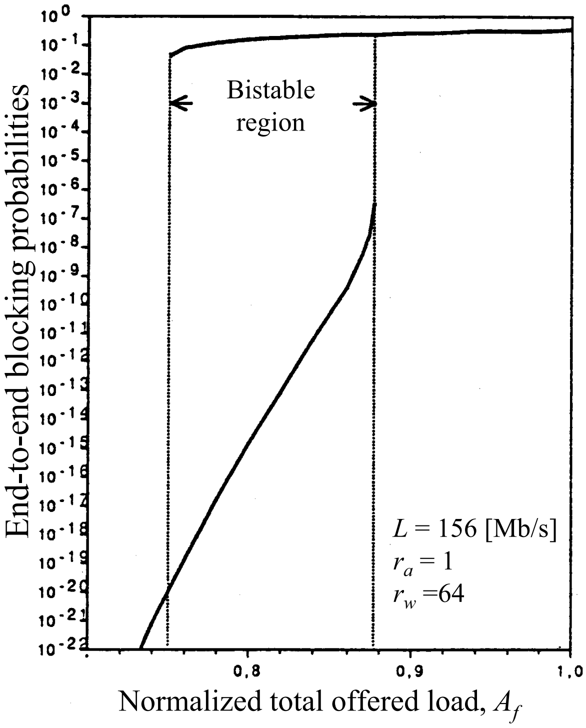


Figure 5.6: Equalized blocking probabilities

parameters as follows,

$$c_f^{(1)} = c_o^{(1)} = L, \;\; c_f^{(2)} = c_o^{(2)} = L - (w^{(1)} - w^{(2)}). \tag{5.10}$$

The equalized end-to-end blocking probabilities under the same traffic conditions of Fig. 5.5 are plotted in Fig. 5.6. It can be seen that the equalization of blocking probabilities for different call classes can be achieved by this control. However, the control does not improve the blocking performance for wideband calls. Even worse, the range of the bistable region becomes wider under the control.

5.6.3 Stabilizing blocking probabilities

We now examine the effect of trunk reservation for direct-routed calls on the performance of multiclass routing networks. Controls equalizing the blocking for each class and reserving link capacity for direct-routed calls are considered here. Specifically, we set

$$\begin{aligned} c_f^{(1)} &= L, \; c_o^{(1)} = L - R_d, \quad \text{and} \\ c_f^{(2)} &= L - (w^{(1)} - w^{(2)}), \; c_o^{(2)} = L - R_d - (w^{(1)} - w^{(2)}), \end{aligned} \tag{5.11}$$

where R_d represents a reservation capacity for direct-routed calls.

The end-to-end blocking probabilities with this control with $L = 156\ Mb/s$, $r_a = 1$, and $r_w = 64$ are shown in Fig. 5.7. In this figure, five curves are plotted with $R_d = nw^{(1)}$ ($n = 0$, 0.5, 1, 2, and 3). It is clearly shown that the instability behavior seen without controls disappears with the use of the reservation control for direct-routed calls. The upper bound of the bistable region is not sensitive to R_d. In the case that offered load is

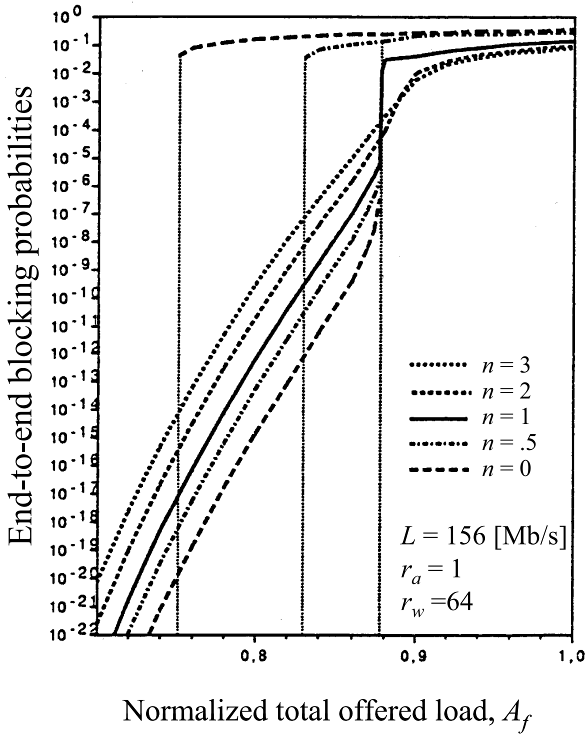


Figure 5.7: End-to-end blocking probabilities of reservation control for direct-routed calls

lower than the upper bound, the blocking probability increases as R_d increases. On the other hand, if the traffic is higher than the bound, the blocking probability decreases as R_d increases. Therefore, the reservation control for direct-routed calls turns out to yield a smooth and gradual increase in the end-to-end blocking with the traffic load.

The end-to-end blocking probabilities plotted as a function of the parameter r_a with $L = 156$ Mb/s, $r_w = 64$, $R_d = 6.144$ Mb/s, and $A_f = 0.9$, 0.92, 0.95, and 1.0 are shown in Fig. 5.8. We can see that the end-to-end blocking probability is insensitive to the offered load ratio (traffic mix) if B_E is between 10^{-2} to 10^{-1} which can be thought of as a typical range of QoS objective.

The end-to-end blocking probabilities plotted as a function of the parameter r_w with $L = 156$ Mb/s, $r_a = 1.0$, $R_d = 10$ Mb/s, and $A_f = 0.95$, 0.98, and 1.0 are shown in Fig. 5.9. Again, the network performance is insensitive to the bandwidth ratio if B_E is between 10^{-2} to 10^{-1}.

5.6.4 Control parameter setting

We now discuss the setting problem of link capacity allocation parameters that maximize a network's total carried traffic. Unfortunately, it is quite difficult to solve the optimization problem directly because the end-to-end blocking probabilities are nonlinear functions of offered load, and the total offered load (direct- and alternate-routed traffic) depends on the allocated capacity.

Here, we explore the setting problem of reservation capacities for direct-routed calls under the condition that the end-to-end QoS constraint is the same for all call classes. The end-to-end blocking probabilities versus the reservation capacity for direct-routed

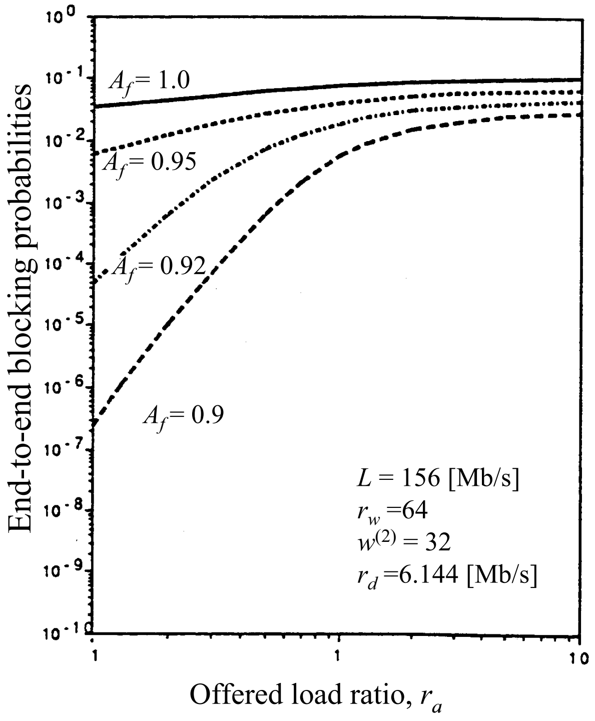


Figure 5.8: End-to-end blocking probabilities versus offered load ratio

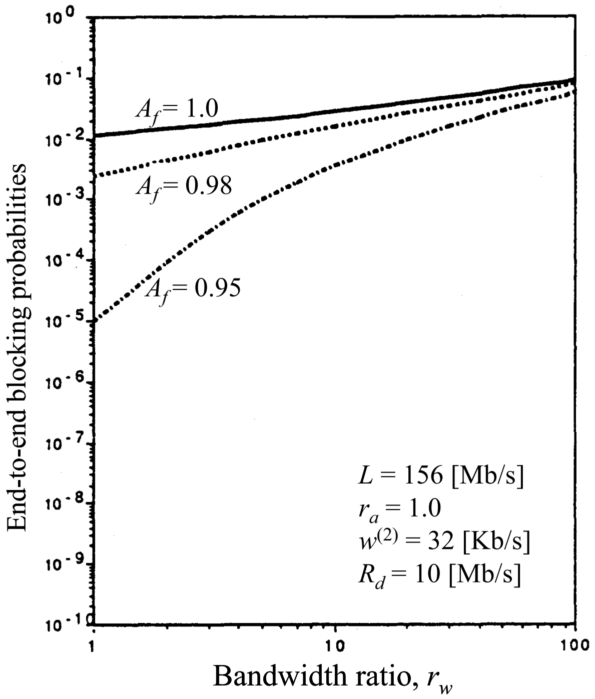


Figure 5.9: End-to-end blocking probabilities versus bandwidth ratio

calls, R_d, with $L = 156\ Mb/s$, $r_a = 1.0$, $r_w = 64$, and $A_f = 0.85$, 0.90, and 0.95 are shown in Fig. 5.10. Notice that the values of $R_d = 0$ and $R_d = L(156\ Mb/s)$ correspond to the end-to-end blocking probabilities of alternate and nonalternate routing, respectively.

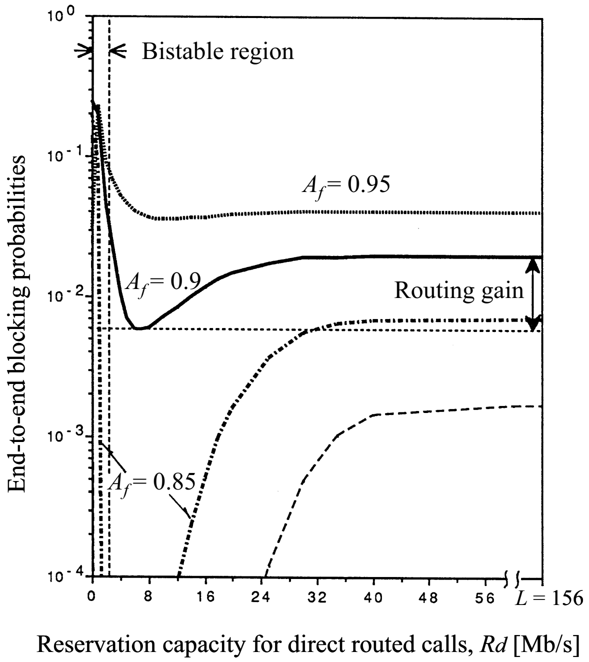


Figure 5.10: End-to-end blocking probabilities versus reservation capacities for direct-routed calls

It can be seen that there exists an optimal reservation parameter between $R_d = 0$ and $R_d = L$ for each offered load. The difference between the optimal value and the value of $R_d = L$ can be considered as routing gain. Therefore, R_d must be small enough so that the routing gain can be achieved and R_d must be large enough so that bistable behavior will not appear. A numerical experiment with various traffic parameters ($r_a = 0.1 \sim 10$, $r_w = 1 \sim 100$) shows that if R_d is set to be 2 or 3 times of $w^{(1)}$, no bistable behavior was found and the routing gain was achieved efficiently.

5.7 Conclusion

We have examined the end-to-end performance of a multiclass alternate routing network. The results obtained indicate that, without control, imbalance and instability of individual blocking probabilities occur for certain offered loads. We have applied a control method named link capacity allocation control to the network model to prevent the imbalance and suppress the instability. The effectiveness of this control has been demonstrated by approximate analysis of some typical network examples. We have also discussed the setting of control parameters to ensure maximum useful network throughput under various traffic conditions. We recommend that such a control be considered in the introduction of integrated services networks with alternate routing.

Chapter 6

Comparative evaluation for mobility management schemes with load balancing controls

6.1 Introduction

This chapter aims to provide a guideline for developing a mobility management scheme appropriate for future cellular systems for personal communication service. As defined in the Universal Personal Telecommunication (UPT) standard [120], the main concepts for providing future personal communication service (PCS) are *terminal mobility*, enabled by wireless access; *personal mobility*, based on personal number; and *service mobility*, provided by per-user service profiles on the network databases (DBs).

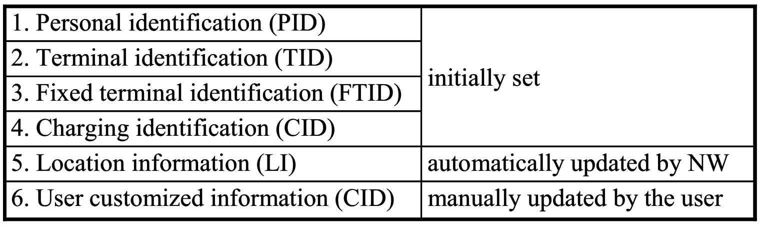
Some previous studies [2, 3] indicate that as the number of users increases, the network control layer such as Intelligent Network and Signaling System 7 infrastructures may require changes in order to support the increased signaling load. Moreover, a recent study [4] indicates that adding personal and service mobility to future networks could trigger an order-of-magnitude increase in signaling traffic.

Many mobility management schemes have been proposed for the reduction of signaling traffic. However, the appropriate conditions of each mobility management scheme have not yet been sufficiently clarified. In particular, insufficient comparisons have been made using a unified performance measure that is free of assumptions as to mobility model or DB architecture.

In this chapter, we categorize the various mobility management schemes into five types. We then clarify their appropriate conditions, using as a unified performance measure, the number of signals at connection setup and location registration. This characteristic closely relates to connection setup time and network efficiency. The number of accesses made at connection setup directly influences the response time for connection setup, as it requires real-time processing. Hence, it is an important measure in determining the Quality of Service (QoS).

The reminder of this chapter is organized as follows. In Sect. 6.2, we define user information and a reference network model for PCSs. Then, we categorize mobility management schemes into five types and illustrate the procedures for connection setup and location registration for each type of scheme in Sect. 6.3. We evaluate the characteristics

Table 6.1: Example of user information table



1. Personal identification (PID)	initially set
2. Terminal identification (TID)	
3. Fixed terminal identification (FTID)	
4. Charging identification (CID)	
5. Location information (LI)	automatically updated by NW
6. User customized information (CID)	manually updated by the user

of the five types of schemes and clarify the appropriate conditions of each scheme using a unified performance measure in Sect. 6.4. We draw some conclusions and discuss topics for future study in Sect. 6.5.

6.2 Reference network model

6.2.1 User information

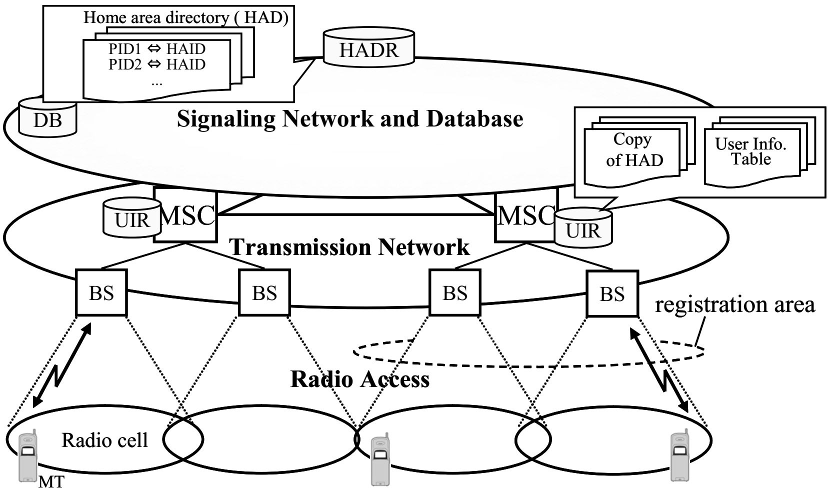
In cellular systems, terminals are not physically connected to a local switch, because they move freely throughout the service areas. Each user is given a unique and lifelong number, called the personal identification number (PID). For a fixed terminal, PID registration is usually performed by inserting an IC card containing the PID.

Three types of user information (UI) are necessary for personal communications. These are authentication information (AI); location information (LI); and user customized information, such as service-available time of day and service-available areas. In the UPT service definition, these three types of information are treated together as one collection of a UPT service profile [120].

An example of a UI table is shown in Table 6.1. Authentication information includes the PID, terminal identification (TID), fixed terminal identification (FTID), and charging identification (CID). These items should be established when the user subscribes, that is, they should be initially set. For multicell communication systems, the network must know the user's location. The current location area should be automatically registered in the network when the user moves from one area to another. User customized information should be updated by the user from his or her terminal.

6.2.2 Network model

Now, we will define a cellular system reference model, illustrated in Fig. 6.1. In this chapter, we are concerned with the mobility management schemes with the distributed DB architecture taking into account future trends for communication network architecture. The network has three layers: the access layer, the transport layer, and the network control layer. The access layer consists of base stations (BSs) for radio access, and the geographical area is partitioned into a number of cells, each of which is served by a BS. The transport layer consists of mobile switching centers (MSCs) and a fixed wired network, which provide switching and transmission functions. The network control layer consists of signaling network and databases, which provides signal transport by means of distributed



HADR: Home area directory register, HAID: Home area identification, UIR: User information register, MT: Mobile terminal, BS: Base Station, MSC: Mobile Switching Center, DB: Database

Figure 6.1: Reference network model for personal communication services

user information registers (UIRs).

A registration area (location area) is composed of an aggregation of cells. We assume that each registration area is served by a paired single MSC and UIR. (This assumption is used to simplify the following analysis, although in practice, each MSC and UIR serve several registration areas.) The home area directory register (HADR) on the network control layer contains a correspondence table giving the relationship between the home area identification (HAID) and PID of all users in the network. This directory should be centrally located as a master register. If a UIR does not have an HAID directory, it can be copied from the HADR. Once the HAID directory has been copied from the HADR to the UIR, HADR accesses are not necessary for subsequent connection setups.

6.3 Mobility management schemes

6.3.1 Categorization of mobility management schemes

Numerous mobility management schemes for cellular systems have been proposed in recent years. We categorize them into five types with respect to the control functions, i.e., replicating and caching functions of UI, which can be applied to advanced cellular systems with distributed DB architecture.

TYPE 1: Home Location Register (HLR)

This scheme is based on distributed DBs known as home location registers (HLRs), and is used for UPT service set 1 [121]. The information about each user is assigned to a DB when he/she subscribes to the cellular system. In this scheme, the UIRs correspond to the HLR. Connection setup requires accessing the UIRs for both the calling and the called users' home area (HA), to authenticate both users and find out the called user's current location. This scheme is simple, but obviously time consuming, especially for establishing a connection in a global service environment, if one of the users is far from his/her home location.

TYPE 2: UI Replication

This scheme is based on distributed UIRs with a replication function that enables copying of UI from the HA's UIR to the UIR in the visiting area (VA) that corresponds to the user's movement, in addition to the TYPE 1 function. This scheme is commonly known as HLR/VLR (visitor location register), which is used in existing cellular system standards such as GSM in Europe and PDC in Japan. Note that most HLR/VLR schemes for current cellular systems store only LI for visiting users in the VLR. But for future personal and service mobility, the TYPE 2 scheme defined here stores all UI that is needed at connection setup such as AI and service scripts in the UIR of the VA. The details of the connection setup procedure will be discussed in a subsequent section.

TYPE 3: TYPE 2 + LI Caching

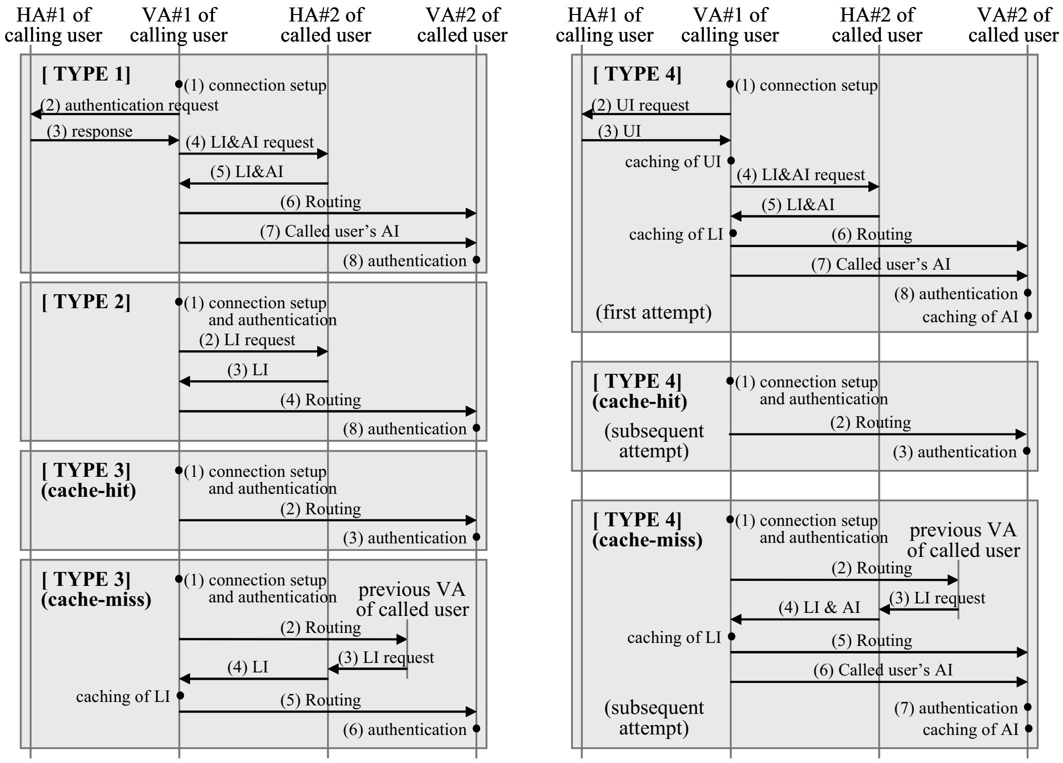
This is a scheme proposed in Refs. [92, 122] based on distributed UIRs, in addition to the TYPE 2 function, with a function that enables caching of the LI of called users in the originating UIR and reuse for subsequent calls from any user in that originating area. The cached information is stored as a called user LI list in the originating UIR. In a request for connection setup to a called user, a *cache hit* occurs if the called user is in the same registration area as cached in the list at a previous call. A *cache miss* occurs if the called user has moved from the listed location area to another area. This scheme is especially suitable for repeated calls from a certain registration area to a certain user who does not move very frequently.

TYPE 4: TYPE 1 + UI Caching

This is a scheme based on HLR, in addition to the TYPE 1 function, with a function that enables caching of the UI of both calling and called users in the originating UIR. TYPE 2 and TYPE 3 schemes require transmissions of UI at every location registration. On the other hand, these transmissions are not required for TYPE 4 scheme, because caching of UI is executed in the connection setup phase instead of the location registration phase. In this scheme, cache hit and cache miss for UI for called users occurs in the same manner as that of the TYPE 3 scheme. This scheme is also suitable for repeated calls from a certain registration area to a certain user who does not move very frequently.

CONV: Centralized DB

This is the conventional scheme, based on a centralized DB that records all UI. With this scheme, all requests for connection setup and location registration are served by only one central DB; hence, the UIRs and HADR are not needed. We use this scheme for comparison with the other types of schemes described above.



HA: Home area, VA: Visiting area, LI: Location information, UI: User information, AI: Authentication information

Figure 6.2: Example of connection setup procedures

6.3.2 Procedure of each scheme

We illustrate the procedures for connection setup for each type of scheme in Fig. 6.2 when the calling and called users are not in the same area and neither are in their HAs.

TYPE 1

The connection setup procedure for the TYPE 1 scheme is as follows.

(1) Connection setup is originated.

(2) The calling user's authentication is requested from the UIR in the calling user's home area (HA #1).

(3) The calling user's authentication is completed.

(4) The called user's LI and AI are requested from the UIR in the called user's home area (HA #2).

(5) The called user's LI and AI are sent to the originating UIR in the calling user's visiting area (VA #1).

(6) Routing is determined according to the LI.

(7) The calling user's AI is sent to the terminating UIR in the called user's visiting area (VA #2).

(8) The called user's authentication is executed, and then the connection is established.

TYPE 2

The connection setup procedure for the TYPE 2 scheme is as follows.

(1) Connection setup is originated and the calling user's authentication is carried out at the UIR in VA #1.

(2) The called user's LI is requested from the UIR in HA #2.

(3) The called user's LI is sent to the UIR in VA #1.

(4) Routing is determined according to the LI.

(5) The called user's authentication is executed, and then the connection is established.

TYPE 3

In the TYPE 3 scheme, the called users' LI is stored in the originating UIR as a called user LI list. This is done when a connection setup is requested. From then on, the LI of that user is stored in the list and is available for each subsequent call. The connection setup procedure for this scheme is different due to the cases of cache hit and cache miss.

The connection setup procedure for a cache hit is as follows.

(1) Connection setup is originated. The calling user's authentication and the called user's LI are referenced in the originating UIR in VA #1.

(2) Routing is determined according to the LI.

(3) The called user's authentication is executed, and then connection is established.

The connection setup procedure for a cache miss is as follows.

(1) Connection setup is originated. The calling user's authentication and the called user's LI are referenced in the originating UIR in VA #1.

(2) Routing is attempted to the previous VA, based on the LI.

(3) The current LI is requested from the UIR in the called user's HA #2 by the UIR in the previous VA.

(4) The current LI is transferred from the UIR in HA #2 to the originating UIR. The called user's LI in the list is updated.

(5) Routing is determined based on the received LI.

(6) The called user's authentication is executed, and then the connection is established.

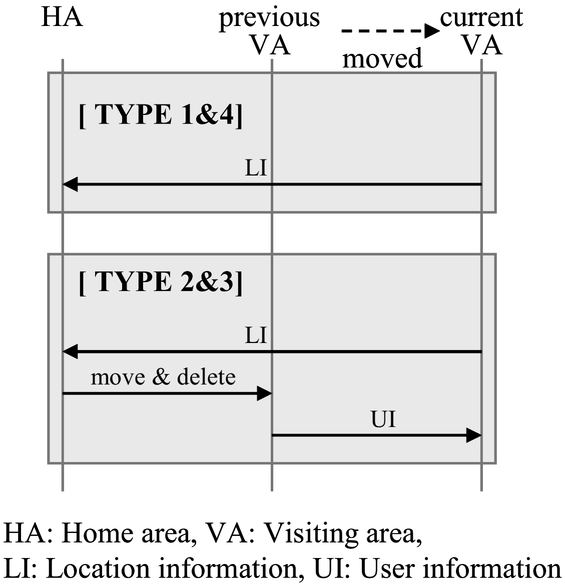

HA
previous VA
moved
current VA
[TYPE 1&4]
LI
[TYPE 2&3]
LI
move & delete
UI
HA: Home area, VA: Visiting area,
LI: Location information, UI: User information

Figure 6.3: Example of location registration procedures

TYPE 4

The procedure for the first connection setup attempt for the TYPE 4 scheme is mostly the same as that of the TYPE 1 scheme. The differences are the caching of the calling user's UI and the called user's LI in the UIR in VA #1, and the caching of the called user's AI in the UIR in VA #2. The procedures for subsequent connection setup attempts for the TYPE 4 scheme are also mostly the same as that of the TYPE 3 scheme. The differences are the updating of the called user's LI in the UIR in VA #1, and the caching of the called user's AI in the UIR in VA #2 for the cache miss case.

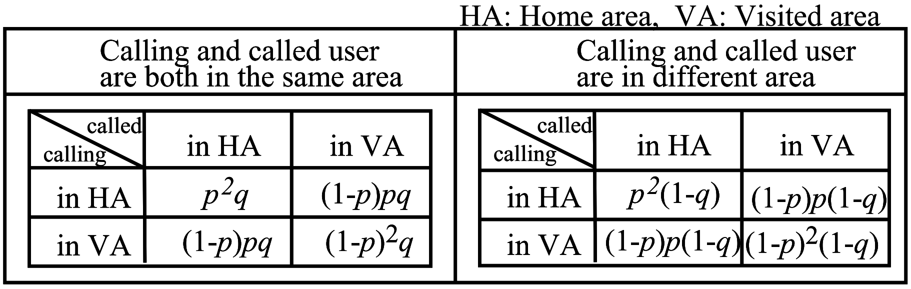
The location registration procedure for each scheme when the user's current and previous areas are not the HA is shown in Fig. 6.3. For the TYPE 1 and TYPE 4 schemes, when the user moves to a new location area, the LI update is sent to the UIR in the user's HA from the UIR in the current VA. For the TYPE 2 and TYPE 3 schemes, the path from the UIR in the previous VA to the UIR in the current VA indicates the copying of the LI according to the user's location registration, and the path from the UIR in the HA to the UIR in the previous VA indicates the signal that directs the referral and deletion of the user's information. Here, we adopt a UI transmission procedure that copies the UI from the UIR in the previous VA to the UIR in the current VA.

6.4 Characteristics of mobility management schemes

In this section we evaluate the characteristics of the five kinds of mobility management schemes described previously and clarify the appropriate conditions of each scheme. we use the number of accesses (signal forwarding frequencies) to the network control layer at connection setup and location registration as performance measures.

Signals in the network control layer can be classified into three categories: (1) signals

Table 6.2: Probabilities of connection setup patterns



HA: Home area, VA: Visited area

Calling and called user are both in the same area			Calling and called user are in different area		
calling \ called	in HA	in VA	calling \ called	in HA	in VA
in HA	p^2q	$(1-p)pq$	in HA	$p^2(1-q)$	$(1-p)p(1-q)$
in VA	$(1-p)pq$	$(1-p)^2q$	in VA	$(1-p)p(1-q)$	$(1-p)^2(1-q)$

at connection setup, (2) signals at location registration, and (3) signals related to operation and management. Since signals related to operation and management are backoffice signals and do not affect service quality, we do not take them into account.

6.4.1 Combination patterns of connection setup

The number of accesses in the network control layer at connection setup depends on the locations of the calling and called users. Eight patterns of connection setup are possible, depending on whether the calling and called user are in the same or different areas, and whether they are in their respective HA. (Here, to clarify the characteristics of mobility management schemes based on the network model in Sect. 6.2, we focus on communication between the users on this network. Interaction with other networks, such as a fixed-wired telephone network, is, therefore, not considered.)

6.4.2 Assumptions and notations

In evaluating the number of accesses, we use the following notation:

p : the probability that a user is in his/her HA,

q : the probability that the calling and called user are both in the same area,

λ : connection setup rate per unit time,

γ : location registration rate per unit time.

We assume that the above four parameters are mutually independent. Note that we can easily measure these four parameters at the points where user LI are stored (UIR) and at the MSC in both current and future cellular systems.

6.4.3 Number of accesses in network control layer

Applying parameters p and q to the eight connection setup patterns, we can represent the probabilities of their appearance as in Table 6.2. The numbers of accesses in the network control layer for each pattern of the five schemes are listed in Table 6.3 (see Fig. 6.2 for the pattern when the calling and called users are not in the same area and neither are in their HAs). Here, the number of accesses at connection setup for the TYPE 2, TYPE

Table 6.3: Number of accesses in network control layer

		Number of accesses at connection setup			
		Calling and called user are both in the same area		Calling and called user are in different area	
	called / calling	in HA	in VA	in HA	in VA
CONV @ = 1	in HA	4	4	5	5
	in VA	4	4	5	5
TYPE1 @ = 1	in HA	0	2	3	4
	in VA	2	4	5	6
TYPE2 @ = 3	in HA	0	0	3	3
	in VA	0	0	3	3
TYPE3 @ = 3	in HA	(0,0) / 0	(0,0) / 0	(3,4) / 1	(3,4) / 1
	in VA	(0,0) / 0	(0,0) / 0	(3,4) / 1	(3,4) / 1
TYPE4 @ = 1	in HA	(0,0) / 0	(2,0) / 0	(3,4) / 1	(4,5) / 1
	in VA	(2,0) / 0	(4,0) / 0	(5,4) / 1	(6,5) / 1

@:Number of accesses at location registration, HA:Home area,VA:Visited area
(first attempt, cache-miss)/cache-hit

3, and TYPE 4 schemes are derived by assuming the following search procedure in the originating UIR for the LI of called users.

(1) First, search the originating UIR. If the called user is in the originating area, the called user's information can be found in the originating UIR for the TYPE 2 and TYPE 3 schemes.

(2) Second, if the called user is not in the originating area,

- for TYPE 2, search the copy of HAD;
- for TYPE 3 and TYPE 4, search the cached list.

With this searching procedure, the number of accesses in the network control layer for each scheme can be reduced to zero when the calling and called user are both in the same area (for the TYPE 4 scheme, this can be achieved in the case of a subsequent connection setup attempt).

In Table 6.3, the upper left-hand figures in each column for TYPE 3 and TYPE 4 represent the number of accesses according to the first attempt or the cache miss at a connection setup, respectively. The lower right-hand figures in each column represent the number of accesses according to the cache hit. The TYPE 3 procedure for the first connection setup attempt for a certain called user is the same as that of the TYPE 2 scheme. The TYPE 4 procedure for the first connection setup attempt for a certain called user is the same as that of the TYPE 1 scheme. Consequently, the maximum of the upper left-hand figures in each column for TYPE 3 and TYPE 4 represent the upper bounds of the numbers of accesses in the network control layer. The lower right-hand figures in each column represent the lower bound of these numbers of accesses.

Next, we determine the mean number of accesses, N, in the network control layer per unit time for each scheme. This can be derived by using Tables 6.2 and 6.3 and multiplying the appearance probabilities and the numbers of accesses for each pattern, as follows.

$$\begin{aligned}
N_{conv} &= (5-q)\lambda + \gamma, \\
N_{conv} &= (5-q)\lambda + \gamma, \\
N_{TYPE1} &= (6-3p-pq-2q)\lambda + \gamma, \\
N_{TYPE2} &= 3(1-q)\lambda + 3\gamma, \\
\overline{N}_{TYPE3} &= 4(1-q)\lambda + 3\gamma, \quad \underline{N}_{TYPE3} = (1-q)\lambda + 3\gamma, \\
\overline{N}_{TYPE4} &= 2(3-p-pq-q)\lambda + \gamma, \quad \underline{N}_{TYPE4} = (1-q)\lambda + \gamma.
\end{aligned} \tag{6.1}$$

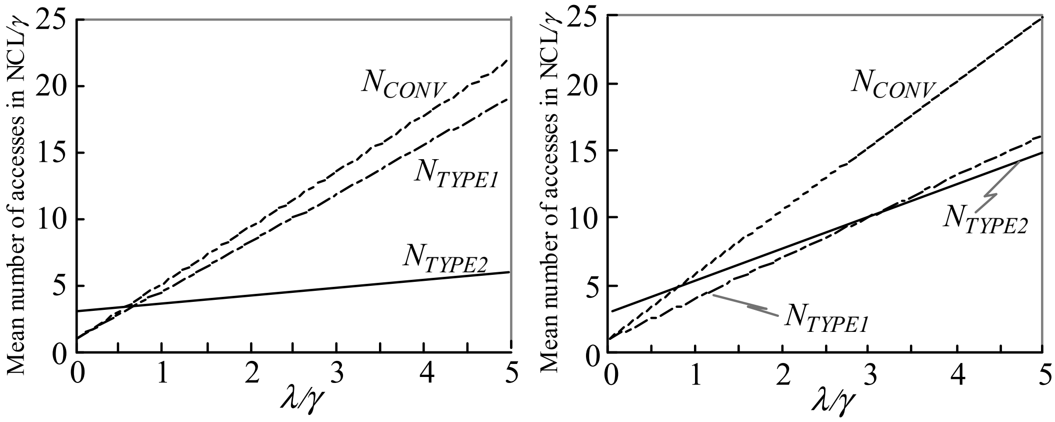
The left-hand terms represent the mean number of connection setup accesses and the right-hand terms represent those of location registration. $\overline{N}$ and $\underline{N}$ indicate the upper and lower bound for N, respectively.

6.4.4 Performance comparisons

Characteristics of the mean numbers of accesses in network control layer

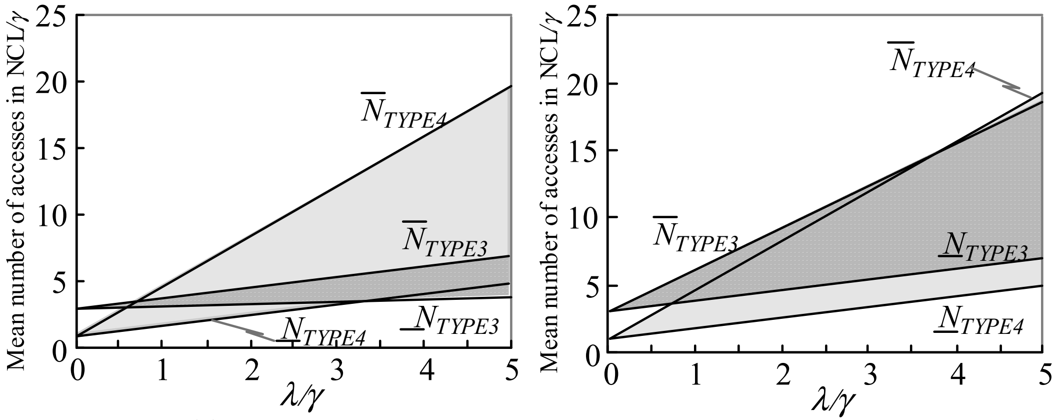
The mean numbers of accesses in the network control layer (normalized by γ) for the CONV, TYPE 1, and TYPE 2 schemes while changing λ/γ from 0.0 to 5.0 are compared in Fig. 6.4. Note that the number of accesses for location registration might be greater than that for connection setup, i.e., $\lambda/\gamma < 1$, in the present cellular systems, still we set the parameter range from 0.0 to 5.0, taking into account the growth of communication in the future. We will discuss this point further in the next subsection. The cases with small p and large q are shown in Fig. 6.4(a), and the cases with converse p and q are shown in Fig. 6.4(b).

We can see from these figures that for a small λ/γ (less than about 1.0), that is, when the connection setup rate is less than the location registration rate, the mean numbers of accesses for both the CONV and the TYPE 1 schemes can be kept low. This is because there is no ineffective processing for the UI transfer at location registration for the CONV and TYPE 1 schemes. In this case, however, the differences in mean value between CONV or TYPE 1 and TYPE 2 is relatively small. Furthermore, these figures indicate that as



(a) p=0.2, q=0.8 (b) p=0.8, q=0.2

Figure 6.4: Mean number of accesses in network control layer (TYPE1 and TYPE2)



(a) p=0.2, q=0.8 (b) p=0.8, q=0.2

Figure 6.5: Mean number of accesses in network control layer (TYPE3 and TYPE4)

λ/γ increases, the mean number of accesses for TYPE 2 increase more slowly than those of CONV and TYPE 1. This tendency is quite noticeable when p is low (Fig. 6.4(a)). This is the effect of the TYPE 2 scheme, in which the UI is transferred and stored in the VA's register. When p is high (Fig. 6.4(b)), we can see that the mean number of accesses for TYPE 1 and TYPE 2 are almost equal.

The mean number of accesses in the network control layer (normalized by γ) for the TYPE 3 and TYPE 4 schemes are compared in Fig. 6.5 under the same conditions as that of Fig. 6.4. In these figures, the shaded area represents the condition for TYPE 3 and TYPE 4 schemes bounded by $\overline{N}$ and $\underline{N}$. These figures indicate that the bound for TYPE 3 is tight, especially when p is low (Fig. 6.5(a)). Furthermore, we can see that the lower bounds are quite low and not sensitive to λ/γ under any conditions.

Appropriate condition for TYPE 1 and TYPE 2 schemes

We need to clarify the appropriate conditions for each type of scheme, because characteristics for each scheme depend on the parameters, p, q, λ, and γ, as shown in Figs. 6.4

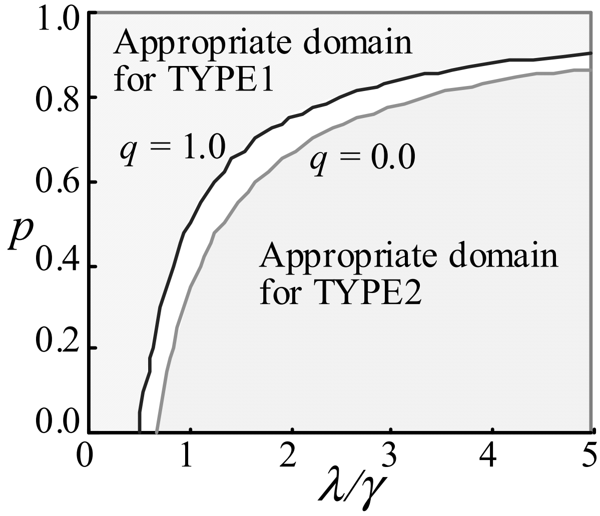


Figure 6.6: Appropriate domains for TYPE1 and TYPE2 schemes

and 6.5. In this subsection, we clarify the appropriate conditions for TYPE 1 and TYPE 2 schemes concerning these parameters from the standpoint of the mean number of accesses in the network control layer. Here, we do not consider the conditions for TYPE 3 and TYPE 4, because, as shown in Eq. 6.1, $\overline{N}_{TYPE3}$ and $\overline{N}_{TYPE4}$ are very close to N_{TYPE2} and N_{TYPE1}, respectively. From Eq. 6.1, we can derive the boundary of the appropriate conditions for the TYPE 1 and TYPE 2 schemes as follows:

$$p = 1 - 2\frac{\gamma}{(3+q)\lambda}. \tag{6.2}$$

The appropriate conditions on the $p-\lambda/\gamma$ plane for the TYPE 1 and TYPE 2 schemes for $q = 0.0$ and 1.0 are shown in Fig 6.6. This figure illustrates the following.

(1) Increasing the value of λ/γ widens the appropriate conditions for TYPE 2.

(2) The boundary is not sensitive to parameter q.

(3) The boundary is sensitive to parameter p, and TYPE 2 becomes effective under the condition where p is low.

In the present cellular systems, the number of accesses for location registration might be greater than that of connection setup (i.e., $\lambda/\gamma < 1$). For the future, we can expect both λ and γ to increase with the rise in social activities. Recent studies (see [2] and references therein) indicate that the introduction of terminal and personal mobility could trigger an order-of-magnitude increase in the query traffic (in proportion to λ and γ), and this increase would be even greater than that of the update traffic (in proportion to γ). Moreover, γ is a parameter that depends on a person's movement characteristics. On the other hand, λ is a parameter that depends on a person's communication characteristics, and the frequency of originating calls for each user shows a tendency to increase with the progress of the cellular system environment, such as mobile computing and globalization of communication. Therefore, we can predict that λ/γ will increase taking into account the growth and globalization of communication in the future. Furthermore, we can assume that in global environments, p will most likely be low, in which case both TYPE 2 and TYPE 3 schemes are effective. Therefore, we can say that TYPE 2 and TYPE

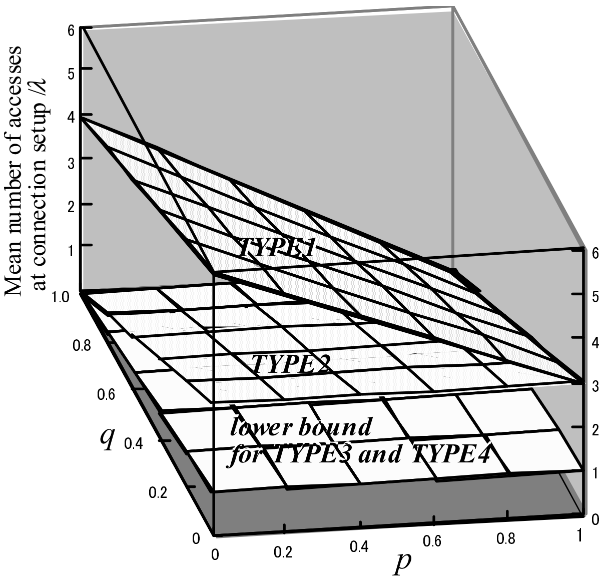


Figure 6.7: Mean number of accesses at connection setup

3 are flexible schemes that can also respond to the need for diversification of personal communication services in the future.

Characteristics of the mean number of accesses at connection setups

In personal communication services, connection setup time is an important QoS measure because it is actually experienced by the user. Here, we evaluate the mean number of accesses to the network control layer at a user's connection setup as a characteristic that is directly related to the connection setup time.

The first terms in the right-hand sides of Eq. 6.1 represent the mean number of accesses at connection setup for each scheme, M_{TYPE}. That is,

$$\begin{aligned}
M_{CONV} &= (5-q)\lambda, \\
M_{TYPE1} &= (6-3p-pq-2q)\lambda, \\
M_{TYPE2} &= 3(1-q)\lambda, \\
\overline{M}_{TYPE3} &= 4(1-q)\lambda,\ \underline{M}_{TYPE3} = (1-q)\lambda, \\
\overline{M}_{TYPE4} &= 2(3-p-pq-q)\lambda, \text{ and } \underline{M}_{TYPE4} = (1-q)\lambda,
\end{aligned} \tag{6.3}$$

where $\overline{M}$ and $\underline{M}$ represent the upper and lower bound of M, respectively. From the above equations we can easily obtain the following relations:

$$\begin{aligned}
&M_{CONV} \geq \overline{M}_{TYPE3} \geq M_{TYPE2} \geq \underline{M}_{TYPE3} = \underline{M}_{TYPE4}, \text{ and} \\
&\overline{M}_{TYPE4} \geq M_{TYPE1} \geq M_{TYPE2}.
\end{aligned} \tag{6.4}$$

Therefore, compared to the CONV and TYPE 1 schemes, the TYPE 2 and TYPE 3 schemes are always effective with respect to the mean number of accesses at connection setup. These characteristics in the form of p and q versus M_{TYPE} normalized by λ for each scheme are shown in Fig. 6.7. The figure indicates that M_{TYPE2}, $\underline{M}_{TYPE3}$, and $\underline{M}_{TYPE4}$

are not sensitive to both p and q. Moreover, we can see that M_{TYPE2} is suppressed to 1/2 and $\underline{M}_{TYPE3}$ to 1/6 or less compared with M_{TYPE1}.

6.5 Conclusion

In this chapter, we categorized and clarified the appropriate conditions of five kinds of mobility management schemes for advanced personal communication services in distributed environments. We used the number of signals at connection setup and location registration as the performance measure that is closely related to connection setup time and network efficiency. We found that two kinds of schemes (TYPE 2 and TYPE 3) that have replicating and caching functions of user information are effective in reducing both the number of accesses in the network control layer and hence the connection setup time. These schemes are especially efficient when the probability that a user is in his/her HA is relatively small and/or the connection setup rate is relatively high compared to the location registration rate. These situations are the ones most likely to occur in the advanced personal communication services for global environments.

Hence, we conclude that these two schemes can be candidates for future personal communication services. Since the input variables of the performance measures used in this chapter can be easily measured in current cellular systems, we can verify the results and design future cellular systems based on the actual traffic data on existing cellular systems.

Several issues remain for further study. One of the most important is detailed evaluation of connection setup times and DB sizes and dynamic traffic estimation that considers user population and movement models [123]. Evaluation of the variance of connection setup time is particularly, important, since the response time experienced by calling users may become variable when the TYPE 3 scheme is applied. Another issue is verification of per-user mobility management, in which information on each user is managed according to an appropriate scheme, such as TYPE 1 or TYPE 2 in this chapter, based on each user's mobility characteristics. Also, an extension of the TYPE 3 scheme to an agent-based architecture [124] may be valuable for further study. With an agent-based architecture, if personal IDs of users to whom frequently made calls are registered in a personal agent, it would be possible to transfer the location information of registered users to that agent from the registered users' agents automatically. This agent-based extension would increase non-real-time signaling traffic, but eliminate cache misses and reduce both the mean and the variance of the connection setup times.

Chapter 7

Conclusion

This thesis has discussed the traffic controls in cellular systems that efficiently allocate limited system resources to communication requests satisfying quality of service (QoS) requirements under variable traffic conditions. This thesis specifically proposed model analysis methods of traffic control schemes that enable quantitative performance evaluations of cellular systems.

First, we categorized traffic control problems by the three-layer structure of cellular system architecture, i.e., 1) radio access, 2) transmission network, and 3) signaling network and database.

Traffic control problems in the radio access layer focusing on congestion control and media access control (MAC) in the reverse control channel were analyzed in Chaps. 2 and 3. Our main contribution to traffic control in this layer is achieving stable and effective use of the radio resource by suppressing the throughput decrease caused by collisions of signals.

In Chap. 2, we proposed an adaptive and scalable congestion control scheme and a method of setting its parameter that maintains maximum throughput even under overloaded conditions. Scalability for handling increasing numbers of MTs and adaptability for coping with drastic changes in traffic load were achieved by controlling the traffic load adaptively to maintain maximum throughput even under overloaded conditions. Procedures for measuring and estimating offered traffic and a method of setting control thresholds that maximize the average throughput were analytically derived. The parameter setting method was proposed for the standardization of congestion control of the personal handy phone system [6].

In Chap. 3, we proposed an approximation method for MAC schemes that integrates the S-G analysis and the diffusion approximation analysis. This method was applied to various control channel access schemes and investigated the effects of access inhibition control and retransmission controls on performance. The approximation accuracy was examined and the traffic characteristics of these schemes were evaluated.

Another important problem addressed in this thesis was how to deal with multimedia traffic efficiently in a transmission network. Chapters 4 and 5 discussed sequence and route control for multimedia traffic with diverse bandwidth and QoS requirements in the transmission network. Traffic controls in transmitting different classes of traffic were classified into two levels: packet level control and call level control. For each level, packet level priority control and call level reservation control were analyzed in Chaps. 4 and 5, respectively.

In Chap. 4, an approximation analysis method of priority queuing network models with renewal arrival and general service time distribution was proposed. The accuracy of the proposed method was validated by exact and simulation results and clarified to be sufficient for practical use. The proposed method was applied to evaluate the end-to-end delay and throughput in the transmission layer for an integrated voice and data packet network, coping with the nature of bursting packet arrival processes and constant packet length in multimedia networks.

In Chap. 5, a link capacity allocation control scheme and its evaluation method for multiclass alternate routing networks were proposed. End-to-end blocking probabilities and bistable behavior were clarified by approximate analysis for various traffic mixes and allocation control parameters. This thesis further demonstrates that the proposed control scheme enables suppression of instability and ensures maximum useful network throughput.

In Chap. 6, this thesis further discussed the mobility management problem with contents reallocation control in signaling network and database layer. We categorized and clarified the appropriate conditions of five kinds of mobility management schemes for advanced personal communication services in distributed environments. We used the number of signals at connection setup and location registration as a performance measure that is closely related to connection setup time and network efficiency. We found that two kinds of schemes that have replicating and caching functions of user information are effective in reducing the number of accesses in the network control layer and hence the connection setup time. These schemes are especially efficient under the situations most likely to occur in advanced personal communication services for global environments.

Furthermore, work to which we contributed related to channel assignment problems with autonomous distributed control is summarized in the Appendix.

In summary, we investigated model analysis of traffic control schemes in cellular systems. The main contributions of our proposed approaches are the following:

- evaluating and comparing the performance of several alternatives in the initial design phase, and determining the best system structure from the viewpoint of traffic design;
- evaluating overall performance characteristics, verifying whether or not the QoS requirements are met, and establishing adequate control parameters; and
- developing dimensioning and management methods for system resources taking into account traffic control effects.

The evolution of both cellular systems and the Internet has reached the point of their convergence. Future generation cellular systems are expected to achieve higher speed data transfer services with multiple traffic classes. Such a scenario requires much more sophisticated traffic controls in operating cellular systems. We expect that the proposed approaches for model analysis of traffic controls will provide a solution for such a future scenario.

Appendix A

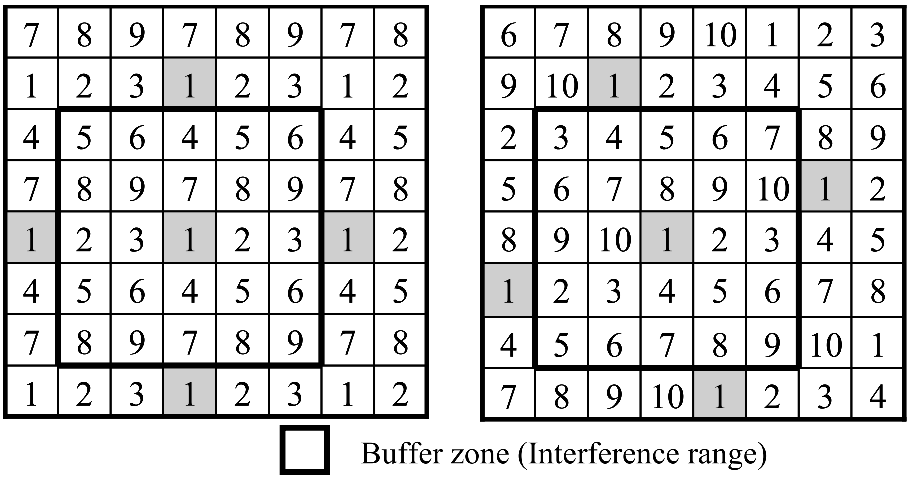
Traffic design for channel assignment scheme with autonomous distributed control

A.1 Introduction

In this Appendix, we treat multicell problems at the radio access network. Radio channels considered here are *traffic channels* rather than the control channels of Chaps. 2 and 3. We analyze a channel assignment scheme using autonomous distributed control in microcellular systems. The microcellular system consists of a very large number of base stations (BSs). The coverage area of a BS is limited to an area with about a 200 m radius. Very low power (such as 10 mW) and low antenna height of the BSs easily changes radio interference conditions. From these factors, both the *fixed channel assignment* (FCA) and *dynamic channel assignment* (DCA) [101] are difficult to apply to the microcellular systems.

To improve this situation for microcellular systems, distributed adaptive channel assignment schemes have been proposed [101], in which BSs assign channels autonomously by measuring whether the traffic channels are in use or not without communicating among other BSs.

In these autonomous distributed control schemes, it is difficult to optimally assign channels; however, various ideas have been proposed to enhance channel utilization. Some of these schemes use genetic algorithms or neural networks [125, 126], but it is difficult to implement these algorithms in actual systems. Some algorithms enable BSs to assign channels adaptively by learning if the previous channel assignments in each BS resulted in efficient transmission. A simple and effective adaptive channel assignment scheme called *channel segregation* was proposed by Furuya and Akaiwa [102]. In this scheme, BSs learn the ordering of channel selection by measuring the radio interference conditions. With such a learning effect, BSs will gradually tend to select certain channels, and the allocation pattern will converge to a suboptimum allocation pattern. The effectiveness of this scheme has been demonstrated by simulation [127]. In order to apply this scheme to actual systems, it is important to know how well the scheme assigns channels as well as how to design traffic, that is, how many BS channels are required to provide the required quality of service (QoS). Traffic design methods for this scheme have not yet been studied.



7	8	9	7	8	9	7	8
1	2	3	1	2	3	1	2
4	5	6	4	5	6	4	5
7	8	9	7	8	9	7	8
1	2	3	1	2	3	1	2
4	5	6	4	5	6	4	5
7	8	9	7	8	9	7	8
1	2	3	1	2	3	1	2

(a) 9-cell repeat pattern

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

(b) 10-cell repeat pattern

Figure A.1: Multicell model and channel allocation pattern

The main purposes of this Appendix are to construct an analytical model for channel dimensioning in the channel segregation scheme, and to describe traffic design using this analytical model.

The reminder of this Appendix is organized as follows. In Sect. A.2, we describe a multicell model of the channel segregation scheme and introduce a new performance measure, *channel reuse distance*, which represents the learning effect for channel allocation pattern. In Sect. A.3, under a converged allocation pattern, we formulate an N-dimensional Markov model of the system and derive the product form solution of the steady-state probabilities. Using these state probabilities, we express an upper bound for a call blocking probability at each BS. In Sect. A.4, we demonstrate by simulation that the channel reuse distance converges to a certain value because of the learning effect of the autonomous control. Then, we show numerical results of the upper bound of call blocking probabilities and compare them with the simulation results. We then discuss a channel dimensioning method based on the channel reuse distance and the upper bound of call blocking probabilities. Section A.5 summarizes this Appendix.

A.2 Multicell model and assumptions

A.2.1 Multicell model

The service area is divided into cells and the coverage area of a cell coincides with the coverage area of a BS. We choose a cell and label it cell #1. Channels used in cell #1 may not be used simultaneously in areas that are within cell #1's interference range. In this Appendix, we assume that this interference range is up to two cells from cell #1. The area comprising cell #1 and its interference range is called *buffer zone* of cell #1. As shown in Fig. A.1, we use a square cell model, so the buffer zone contains 25 cells.

A.2.2 Channel reuse distance

The channel allocation pattern is determined by the relative position of the same channels, and the most effective allocation requires nine channels within a buffer zone as shown in Fig. A.1(a). We call this allocation pattern the *9-cell repeat pattern*. The next-most effective pattern requires ten different cells within a buffer zone, which we call the 10-*cell repeat pattern*, as shown in Fig. A.1(b). We introduce a performance measure, *channel reuse distance*, defined as the shortest distance between the centers of cells in which the same channels are allocated. In the following, we represent the channel reuse distance in units of the length of the side of a cell. In the cases of Figs. A.1(a) and (b), the channel reuse distances are 3.0 and 3.16, respectively. We call a set of cells that use the same channel a *zoneclass*. The 9-cell repeat pattern contains nine zoneclasses.

A.2.3 Assumptions

In dynamic channel assignment scheme, channel allocation patterns vary a lot. The restriction conditions due to radio interference conditions are complicated so much. It is, therefore, difficult to describe its state transition equations and, furthermore, it is nearly impossible to solve those equations.

If we want to approach these problems by simulation, it requires a lot of calculation time because we have to take into account the inter-cell relationship for a large number of cells. Instead of those above approaches, we seek solutions approximately match the real characteristics and reside in safety side. We describe an approximation method which yields an upper bound for a call blocking probability. To construct an algorithm for evaluating the upper bound, we assume that the above multicell model satisfies the following three conditions besides the usual assumption of mutually independent Poisson arrivals and exponential holding times: (1) the system is in equilibrium, which means that the channel reuse distance converges to some value, and the channel allocation pattern can therefore be fixed to a pattern corresponding to this value, (2) the BSs belonging to the same zoneclass select channels according to the same ordering, and (3) the cells within the buffer zone of interest are free from interference from other cells outside of the buffer zone, that is, we are only concerned with the channels within the buffer zone.

A.3 Channel dimensioning method

A.3.1 N-dimensional Markov chain model

In this section, we propose an approximation method that yields an upper bound of the call blocking probability[1].

[1] We note that the upper bound of call blocking probability is obtained by focusing on a center cell and its buffer zone. According to the third assumption in Sect. A.2, we neglect the effect from outside the buffer zone. Under this assumption, the cells within buffer zone can use channels which may be unavailable in those cells because of channel interference in an actual situation. This leads to new interference between the cells and the center cell of the buffer zone. So there arises some possibility that the center cell cannot use channels under above assumption, while the center cell can use the channels in the actual situation. Therefore, the calculated blocking probability under above assumption is greater than that of an actual one.

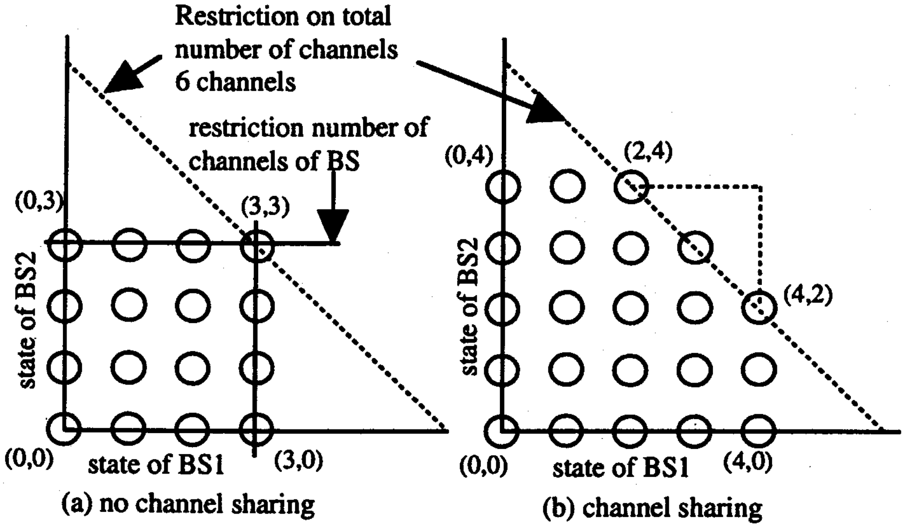


Figure A.2: Restrictions on number of channels

A base station in each cell can handle several channels simultaneously. Therefore, the system state is usually represented as the number of simultaneously used channels of every cell. The amount of calculation required for our analysis is reduced by assuming that the relevant system state is not the number of simultaneously used channels of every cell, but the largest number of simultaneously used channels among the cells with the same zoneclass.

Here, we have L different traffic channels in the whole system. In cell k, we have L_k channels simultaneously available for communications. We denote the number of zoneclasses as N and the set of cells in zoneclass n $(= 1, 2, \cdots, N)$ in the buffer zone as Ω_n. We also denote $\max_{k \in \Omega_n}\{L_k\}$ as $L^{(n)}$.

Traffic channels are shared by cells according to the interference conditions. We let $\boldsymbol{m}_k(t)$ be the number of simultaneously used channels in a cell k at time t, and $\boldsymbol{l}_n(t)$ the maximum of $\boldsymbol{m}_k(t)$'s over cells belonging to zoneclass n, that is,

$$\boldsymbol{l}_n(t) \equiv max_{k \in \Omega_n}\{\boldsymbol{m}_k(t)\}. \tag{A.1}$$

Then the interference can be expressed as

$$\sum_{n=1}^{N} \boldsymbol{l}_n(t) \leq L. \tag{A.2}$$

Note that, under our assumption of mutually independent Poisson arrivals and exponential holding times, if the interference condition Eq. A.2 does not exist, then $\{\boldsymbol{m}_k(t)\}$ behaves as the $M/M/S/S$ model, where $S = L_k$, with arrival rate λ_k and service rate μ_k. Furthermore, from our assumptions (2) and (3), when the interference condition Eq. A.2 does not exist, $\{\boldsymbol{l}_n(t)\}$ is a continuous time Markov chain on the state space $\{0, 1, 2, \ldots, L^{(n)}\}$, and Markov chains $\{\{\boldsymbol{l}_n(t)\}, n = 1, 2, \ldots, N\}$ are mutually independent. Therefore, even under the interference condition Eq. A.2, the vector-valued process $\{(\boldsymbol{l}_1(t), \boldsymbol{l}_2(t), \cdots \boldsymbol{l}_n(t))\}$

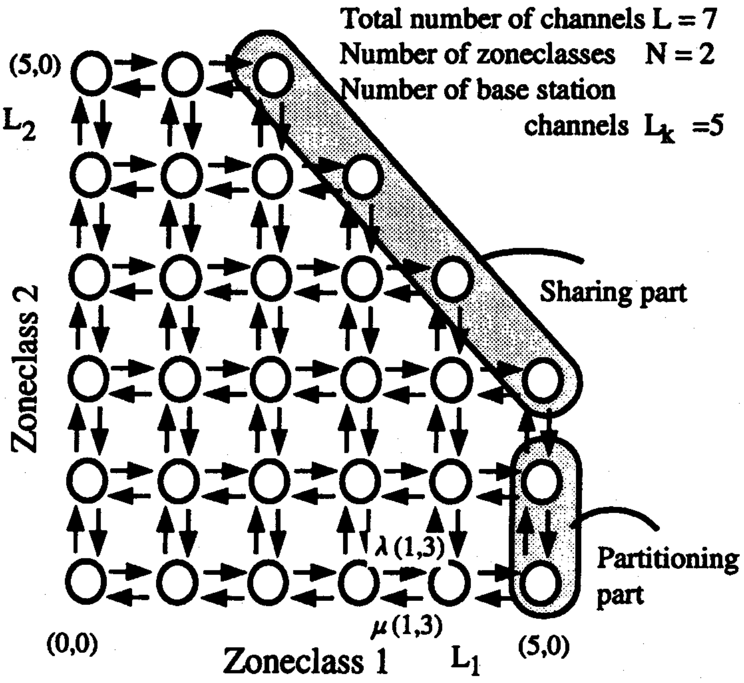

Total number of channels L = 7
Number of zoneclasses N = 2
Number of base station channels L_k =5
(5,0)
L_2
Zoneclass 2
Sharing part
Partitioning part
λ(1,3)
μ(1,3)
(0,0)
Zoneclass 1
L_1
(5,0)

Figure A.3: N-dimensional Markov chain model ($N = 2$)

is a Markov chain on the state space

$$\Omega = \left\{ (l_1, l_2, \cdots, l_n) \;\middle|\; l_n \leq L^{(n)},\ n = 1, 2, \cdots, N,\ \text{and}\ \sum_{n=1}^{N} l_n \leq L \right\}, \tag{A.3}$$

(see Fig. A.3). From our assumption (1), we assume that the Markov chain is in the steady state.

A.3.2 Product form solution

For the Markov chain $\{(\boldsymbol{l}_1(t), \boldsymbol{l}_2(t), \cdots \boldsymbol{l}_n(t))\}$, the local balance equations are easily derived and the steady-state probability $P(l_1, l_2, \cdots, l_n)$ is given by the product form solution

$$P(l_1, l_2, \cdots, l_n) = G^{-1}(\Omega) \prod_{n=1}^{N} P_n(l_n), \tag{A.4}$$

where $G(\Omega)$ is a normalization constant and $P_n(l_n)$ denotes a function of l_n representing the steady-state probability of $\{\boldsymbol{l}_n(t)\}$ when the interference condition Eq. A.2 does not exist. The function $P_n(l_n)$ is calculated in the following manner. Hereafter, in this subsection, we consider cells in zoneclass n under the assumption that the interference condition Eq. A.2 does not exist.

For the Markov chain $\{\boldsymbol{m}_k(t)\}$, the steady-state probability that l_n channels are oc-

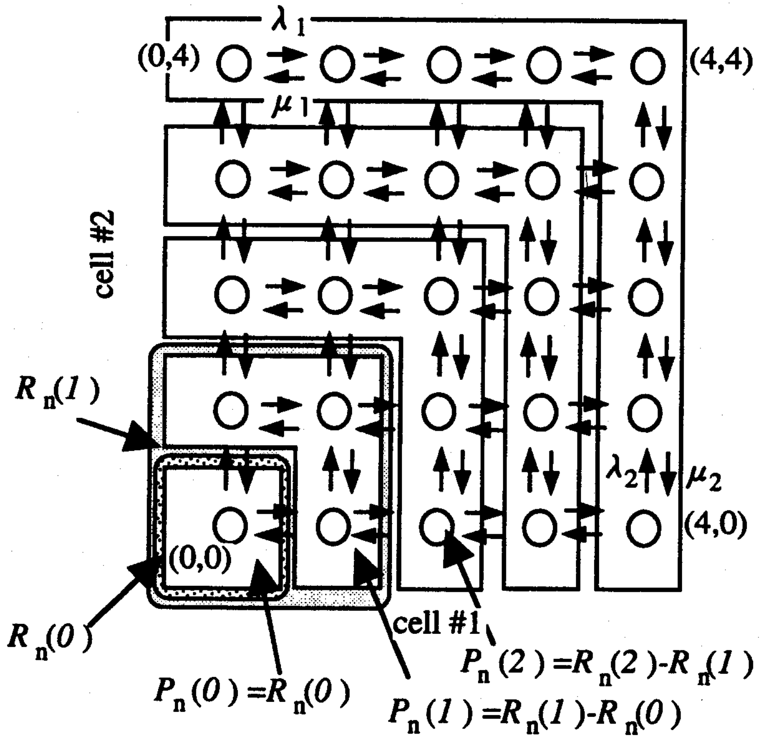


Figure A.4: Structure of state transition of a zoneclass

cupied in cell k is described by the $M/M/S/S$ model ($S = L_k$) as

$$q_k(l_n) = \begin{cases} \dfrac{a_k^{l_n}}{l_n!} \Bigg/ \sum\limits_{j=0}^{L_k} \dfrac{a_k^j}{j!}, & 0 \le l_n \le L_k, \\ 0, & l_n > L_k, \end{cases} \tag{A.5}$$

where $a_k \equiv \lambda_k/\mu_k$, the load at cell k. Then the relative portion that the number of channels occupied is equal to l_n or less is given by

$$Q_k(l_n) = \begin{cases} \sum\limits_{i=0}^{l_n} \dfrac{a_k^i}{i!} \Bigg/ \sum\limits_{j=0}^{L_k} \dfrac{a_k^j}{j!}, & 0 \le l_n \le L_k, \\ 1, & l_n \ge L_k. \end{cases} \tag{A.6}$$

From our assumption (2), the probability $R_n(l_n)$ that the number of channels occupied is equal to l_n or less in every cell of zoneclass n is given by

$$R_n(l_n) = \prod_{k \in \Omega_n} Q_k(l_n). \tag{A.7}$$

Since $P_n(l_n)$ is the steady-state probability that the maximum number of simultaneous used channels in zoneclass n is equal to l_n (Fig. A.4), it is given by

$$P_n(l_n) = \begin{cases} \prod\limits_{k \in \Omega_n} Q_k(0), & l_n = 0, \\ \prod\limits_{k \in \Omega_n} Q_k(l_n) - \prod\limits_{k \in \Omega_n} Q_k(l_n - 1), & l_n \ge 1. \end{cases} \tag{A.8}$$

A.3.3 Calculation of normalization constant

To calculate the normalization constant $G(\Omega)$ in Eq. A.4, we apply the convolution method proposed by Buzen [128]. First, we introduce the generation function of $\{P_n(l_n)\}$ and their product as

$$g(z) \equiv \prod_{n=1}^{N} g_n(z) = \prod_{n=1}^{N} \left[\sum_{l_n=0}^{\infty} P_n(l_n) z^{l_n} \right]. \tag{A.9}$$

We denote their partial products as

$$\begin{aligned} &\gamma_1(z) \equiv g_1(z), \text{ and} \\ &\gamma_i(z) \equiv \gamma_{i-1}(z) g_i(z), \quad i = 2, 3, \cdots, N, \end{aligned} \tag{A.10}$$

and we let $G(j, n)$ be the coefficient of $\gamma_n(z)$ of the order z^j. Then, we have the recurrence formula

$$\begin{aligned} G(j, n) = \sum_{i=0}^{\min(j, L^{(n)})} & G(j-i, n-1) P_n(i), \\ & n = 1, 2, \cdots, N, \; j = 0, 1, \cdots, \min(L, \textstyle\sum_{m=1}^{n} L^{(m)}), \end{aligned} \tag{A.11}$$

with initial conditions

$$\begin{aligned} &G(j, 0) = \begin{cases} 1, & j = 0, \\ 0, & j \neq 0, \end{cases} \quad \text{and} \\ &G(j, n) = 0, \; j = \min(L, \textstyle\sum_{m=1}^{n} L^{(m)}) + 1, \cdots, \min(L, \textstyle\sum_{m=1}^{n+1} L^{(m)}). \end{aligned}$$

Using these expressions, we can calculate the normalization constant as

$$G(\Omega) = \sum_{j=0}^{L} G(j, N). \tag{A.12}$$

A.3.4 Call blocking rate

Note that, in the buffer zone, zoneclass 1, which contains cell #1, consists of a single cell, namely cell #1. Then using the quantities obtained above we can calculate the call blocking rate B_1 of cell #1 as

$$B_1 = G^{-1}(\Omega) \left\{ G(L, N) + P_1(L^{(1)}) \sum_{j=0}^{L-L^{(1)}-1} G(j, N-1) \right\}. \tag{A.13}$$

As shown in Fig. A.3, B_1 is made up of two parts: one due to the lack of free channels in the whole system (sharing part) and the other due to the lack of BS channels in cell #1 (partitioning part). The first and second terms of the right side of Eq. A.13 represent the sharing and partitioning parts, respectively.

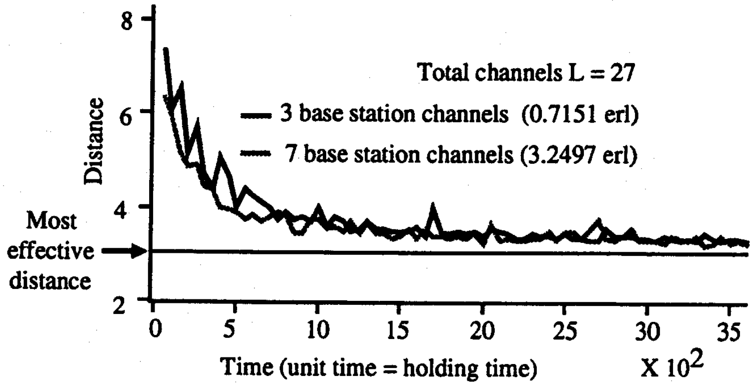


Figure A.5: Transient characteristics of the channel reuse distance

A.4 Numerical examples

A.4.1 Channel reuse distance

At first, we demonstrate by simulation that the channel reuse distance converges to a certain value. In the simulation experiment, a 16×16 square area model composed of 256 square cells are used. To eliminate the influence of border cells, the target of statistical data is chosen from the central 8×8 square area. Poisson arrivals and the exponential holding time distribution at the uniform rates for each cell are assumed. The initial order of channel selection is assumed to be identical for each cell. The priority function proposed in Ref. [129] is used as a channel selection algorithm at each BS.

Transient characteristics of the channel reuse distance for the total number of traffic channels, $L = 27$, and the number of BS channels, $L_k = 3$ and 7, are shown in Fig. A.5. Simulation results show that the channel allocation pattern varies with time from the start of simulation, but after some time, the channel reuse distance converges to a certain value and becomes stable. The converged channel reuse distance is about 3.3, only slightly greater than that of the 10-cell repeat pattern.

A.4.2 Call blocking rate

Next, using this channel reuse distance, we can calculate the call blocking rate by Eq. A.13. Note that there is no systematic pattern that corresponds to the simulation results with the channel reuse distance of 3.3. Thus, in the following numerical examples, we choose the 10-cell repeat pattern with the channel reuse distance of 3.16. Hence, the characteristics of call blocking probability in these examples may yield better values than we expected with the channel reuse distance of 3.3.

We set the total number of traffic channels to $L = 27$, the number of cells in the interference area to $K = 25$, and the number of zoneclasses to $N = 10$. In Fig. A.6, the call blocking rates versus the channel offered load calculated by Eq. A.13 are compared to the simulation results for $L_k = 3$, 4, and 7. The results indicate that our approximation and the simulation results closely match, which indicates the assumptions in Sect. A.2 work well. The call blocking rates with various numbers of BS channels when we set the offered load at a constant value (= 1.4 erl) are shown in Fig. A.7. If we increase

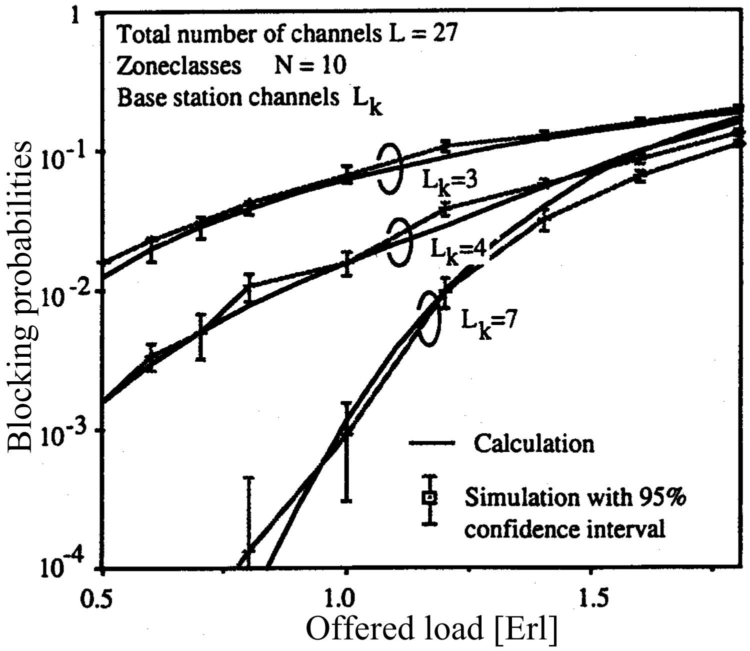


Figure A.6: Blocking rate versus offered load

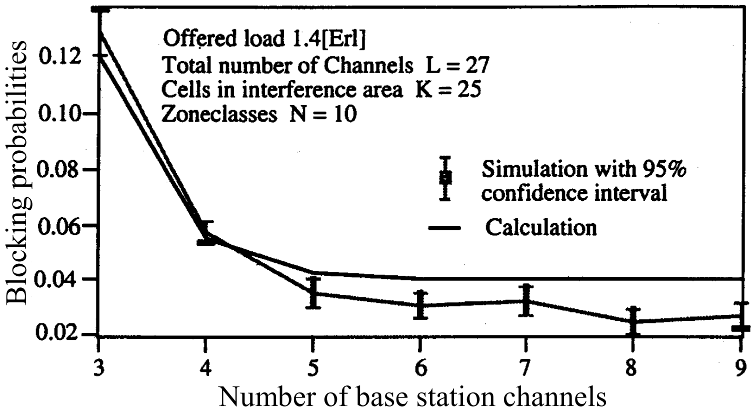


Figure A.7: Blocking rate versus the number of base station channels

the number of BS channels with a constant offered load, the call blocking probability converges to a constant value. This indicates that there is a limit to the radio-channel-sharing effect. Thus, we can find the smallest number of BS channels that gives the most effective radio-channel-sharing effect.

A.4.3 Traffic design for channel dimensioning

Iteratively solving the expression of B_1 for offered load, we obtain the relationship between the offered load for a required call blocking rate and the number of simultaneously available BS channels as shown in Fig. A.8. In calculating the values shown in this figure, we set the total number of traffic channels to $L = 200$ to show the performance of the system whose scale is similar to that of actual systems. Notice that the offered load for a fixed

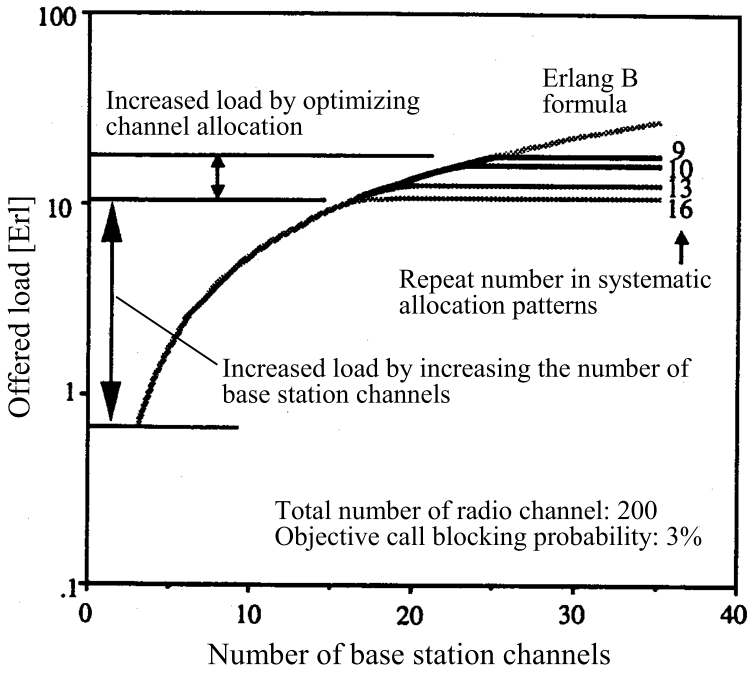


Figure A.8: Offered load versus the number of base station channels

call blocking rate saturates with an increase in the number of BS channels. That is, if the number of BS channels exceeds some threshold value, there is no radio-channel-sharing effect even if we add BS channels. Traffic design for channel dimensioning should thus be based on this relationship between the number of BS channels and the offered load taking into account the radio-channel-sharing effect.

A.5 Conclusion

In this Appendix, we analyzed a channel assignment scheme using autonomous distributed control in microcellular systems and derived an algorithm for evaluating the upper bound of the call blocking rate. Because our analysis assumes that a system's channel allocation converges to a fixed systematic pattern characterized by its channel reuse distance, we can easily analyze the randomly distributed complex allocation patterns of adaptive distributed control schemes. Next, because our analysis focuses on a single cell and its buffer zone, our results give an upper bound for the call blocking rate. Finally, the amount of calculation required for our analysis is greatly reduced by assuming that the relevant system state is not the number of used channels of every cell, but the greatest number of simultaneously used channels of every zoneclass.

The proposed analytical approach gives results that closely match the simulation results and can be used to evaluate the call blocking rate of systems under practical conditions such that the number of system channels exceeds 100.

Bibliography

[1] D. Everitt, “Traffic engineering of the radio interface for cellular mobile networks,” *Proceedings of the IEEE*, vol. 82, no. 9, pp. 1371–1382, 1994.

[2] P. E. Wirth, “Teletraffic implications of database architectures in mobile and personal communications,” *IEEE Communications Magazine*, vol. 33, no. 6, pp. 54–59, 1995.

[3] G. P. Pollini, K. S. Meier-Hellstern, and D. J. Goodman, “Signaling traffic volume generated by mobile and personal communications,” *IEEE Communications Magazine*, vol. 33, no. 6, pp. 60–65, 1995.

[4] C. Lo, R. S. Wolff, and R. C. Bernhardt, “An estimation of network database transaction volume to support universal personal communications services,” in *Proc. 1st Int. Conf. on Universal Personal Commun.*, (Dallas, Texas, US), pp. 236–241, 1992.

[5] RCR STD, “Personal digital cellular telecommunication system,” Tech. Rep. RCR STD-27, ARIB, 1997.

[6] RCR STD, “Personal handy phone system,” Tech. Rep. RCR STD-28, ARIB, 1993.

[7] S. S. Lam and L. Kleinrock, “Packet switching in a multi-access broadcast channel: dynamic control procedures,” *IEEE Trans. Commun.*, vol. COM-23, no. 9, pp. 891–904, 1975.

[8] H. Okada, “Optimal control of an ALOHA channel —throughput characteristics under input limit control,” *Trans. IEICE*, vol. J65-D, no. 1, pp. 96–103, 1982. (in Japanese).

[9] T. Kawabata, T. Miyazaki, H. Aoyagi, K. Murakami, Y. Asano, and Y. Yamazaki, “Performance of dynamic slot assignment for fixed wireless access systems,” *Technical report of IEICE*, vol. SSE200-89, pp. 43–48, 2000. (in Japanese).

[10] H. Kayama, T. Hattori, and H. Yoshida, “Adaptive control for random access traffic in mobile radio systems,” *IEEE Trans. Vehic. Technol.*, vol. VT-42, no. 1, pp. 87–93, 1993.

[11] A. Murase and K. Imamura, “Idle-signal casting multiple access with collision detection (ICMA-CD) for land mobile,” *IEEE Trans. Vehic. Technol.*, vol. VT-36, no. 2, pp. 45–50, 1987.

[12] S. K. Kasera, R. Ramjee, S. R. Thuel, and X. Wang, “Congestion control policies for IP based CDMA radio access networks,” *IEEE Trans. Mobile Comp.*, vol. 4, no. 4, pp. 349–362, 2005.

[13] E. B. Rodrigues and F. R. P. Cavalcanti, “QoS-driven adaptive congestion control for voice over IP in multiservice wireless cellular networks,” *IEEE Communications Magazine*, vol. 46, no. 1, pp. 100–107, 2008.

[14] A. Takahashi, A. Kurashima, and H. Yoshino, “Objective assessment methodology for estimating conversational quality in VoIP,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1984–1993, 2006.

[15] A. Takahashi, A. Kurashima, H. Aoki, and H. Yoshino, “Conversational quality estimation of IP telephony services,” in *Proc. World Telecommunications Congress (WTC2006)*, Session QT1, (Budapest, Hungary), 2006.

[16] A. Takahashi, H. Yoshino, and N. Kitawaki, “Quality assessment methodologies for IP-telephony services,” *Trans. IEICE*, vol. J88-B, no. 5, pp. 863–874, 2005. (in Japanese).

[17] A. Takahashi, A. Kurashima, C. Morioka, and H. Yoshino, “Objective quality assessment of wideband speech by an extension of ITU-T recommendation P.862,” in *Proc. 9th European Conference on Speech Communication and Technology (Eurospeech2005)*, (Lisbon, Portugal), 2005.

[18] A. Takahashi, A. Kurashima, and H. Yoshino, “Subjective quality index for compatibly evaluating narrowband and wideband speech,” in *Proc. Measurement of Speech and Audio Transmission Quality in Telecommunication Networks (MESAQIN2005)*, (Prague, Czech Republic), 2005.

[19] A. Takahashi, H. Yoshino, and N. Kitawaki, “Perceptual QoS assessment technologies for VoIP,” *IEEE Communications Magazine*, vol. 42, no. 7, pp. 28–34, 2004.

[20] T. Hayashi, A. Takahashi, and H. Yoshino, “State-of-the-art of QoS assessment methodologies for multimedia telecommunication services,” *Trans. IEICE*, vol. J91-A, no. 6, pp. 600–612, 2008. (in Japanese).

[21] T. Hayashi, K. Yamagishi, and H. Yoshino, “Perceptual QoS evaluation model for audiovisual communication services,” in *Proc. World Telecommunications Congress (WTC2006)*, Session QT1, (Budapest, Hungary), 2006.

[22] T. Kurita and H. Yoshino, “Quality assessment method for multiparty audiovisual communication services,” in *Proc. IEEE Globecom’05*, GC03.10, (St.Louis, US), 2005.

[23] N. Abramson, “The ALOHA system – another alternative for computer communications,” in *Proc. Fall Joint Comput. Conf. AFIPS Conf.*, vol. 37, pp. 281–285, 1970.

[24] A. Fukuda, "On the dynamic behavior of ALOHA-type systems," *Trans. IECE*, vol. J60-D, no. 8, pp. 649–650, 1977. (in Japanese).

[25] S. Tasaka, *Performance Analysis of Multiple Access Protocols.* The MIT Press Series in Computer Systems, 1986.

[26] A. Fukuda, "Equilibrium point analysis of ALOHA-type systems," *Trans. IECE*, vol. J61-B, no. 11, pp. 959–966, 1978. (in Japanese).

[27] H. Kobayashi, Y. Onozato, and D. Huynh, "An approximate method for design and analysis of an ALOHA system," *IEEE Trans. Commun.*, vol. COM-25, no. 1, pp. 148–157, 1977.

[28] K. Matsumoto and H. Takahashi, "Analysis of an ALOHA packet switching network," *Rev. of the Radio Research Labs.*, vol. 24, no. 127, pp. 191–203, 1978. (in Japanese).

[29] H. Miyahara, T. Matsumoto, and K. Takashima, "Transient analysis of carrier sense multiple access with collision detection via diffusion approximation," Tech. Rep. OS 26-8, IPSJ SIG Technical Report, 1985. (in Japanese).

[30] T. Kimura, "A unifying diffusion model for state-dependent queues," *Optimization*, vol. 18, no. 2, pp. 265–283, 1987.

[31] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Select. Areas Commun.*, vol. JSAC-18, no. 3, pp. 535–547, 2000.

[32] A. Kumar, E. Altman, D. Miorandi, and M. Goyal, "New insights from a fixed-point analysis of single cell IEEE 802.11 WLANs," *IEEE/ACM Trans. Networking*, vol. 15, no. 3, pp. 588–601, 2007.

[33] H. Zhai and Y. Fang, "Performance of wireless LANs based on IEEE 802.11 MAC protocols," in *Proc. 14th IEEE Personal, Indoor and Mobile Radio Communications (PIMRC 2003)*, vol. 3, pp. 2586–2590, 2003.

[34] K. Ghaboosi, B. Khalaj, Y. Xiao, and M. Latva-aho, "Modeling IEEE 802.11 DCF using parallel space-time Markov chain," *IEEE Trans. Vehic. Technol.*, vol. VT-57, no. 4, pp. 2404–2413, 2008.

[35] T. Alpcan, T. Basar, R. Srikant, and E. Altman, "CDMA uplink power control as a noncooperative game," *ACM Wireless Networks*, vol. 8, no. 6, pp. 659–670, 2002.

[36] M. Hayajneh, I. Khalil, and M. Awad, "Non-cooperative uplink power control game for CDMA wireless communications systems," in *Proc. IEEE Symposium on Computers and Communications (ISCC 2009)*, pp. 587–592, 2009.

[37] A. B. MacKenzie and S. B. Wicker, "Game theory and the design of self-configuring, adaptive wireless networks," *IEEE Communications Magazine*, vol. 39, no. 11, pp. 126–131, 2001.

[38] J. W. Lee, M. Chiang, and A. Calderbank, “Utility-optimal random-access control,” *IEEE Trans. Wireless Communications*, vol. 6, no. 7, pp. 2741–2751, 2007.

[39] Y. E. Sagduyu and A. Ephremides, “A game-theoretic analysis of denial of service attacks in wireless random access,” in *Proc. 5th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2007)*, pp. 1–10, 2007.

[40] Y. E. Sagduyu, R. Berry, and A. Ephremides, “MAC games for distributed wireless network security with incomplete information of selfish and malicious user types,” in *Proc. International Conference on Game Theory for Networks (GameNets ’09)*, pp. 130–139, 2009.

[41] S. Harada, R. Kawahara, T. Mori, N. Kamiyama, S. Hasegawa, and H. Yoshino, “A method of detecting network anomalies in cyclic traffic,” in *Proc. IEEE Globecom’08*, (New Orleans, US), 2008.

[42] N. Kamiyama, T. Mori, R. Kawahara, S. Harada, and H. Yoshino, “Extracting worm-infected hosts using white list,” in *Proc. Symposium on Applications and the Internet (SAINT2008)*, Session 4, (Turku, Finland), 2008.

[43] F. Baskett, K. M. Chandy, R. R. Muntz, and F. Palacios, “Open, closed and mixed networks of queues with different classes of customers,” *J. ACM*, vol. 22, pp. 248–260, 1975.

[44] M. Reiser, “Interactive modeling of computer systems,” *IBM Syst. J.*, vol. 15, pp. 283–294, 1976.

[45] K. C. Sevcik, “Priority scheduling disciplines in queueing network models of computer systems,” in *Proc. IFIP Information Processing ’77*, (North-Holland, Amsterdam), 1977.

[46] J. S. Kaufman, “Approximate analysis of priority scheduling disciplines in queueing network models of computer systems,” in *Proc. IEEE ICC ’82*, (Philadelphia, Pennsylvania, US), pp. 955–961, 1982.

[47] W. Schmitt, “Approximate analysis of Markovian queueing networks with priorities,” in *Proc. 10th Int. Teletraffic Congr.*, Session 1.3.3, (Montreal, Canada), 1982.

[48] S. Ikehara and M. Miyazaki, “Approximate analysis of queueing networks with nonpreemptive priority scheduling,” in *Proc. 11th Int. Teletraffic Congr.*, Session 3.4A.2, (Kyoto, Japan), 1985.

[49] M. Reiser and H. Kobayashi, “Accuracy of the diffusion approximation for some queueing systems,” *IBM Journal of Research and Development*, vol. 18, pp. 110–124, 1974.

[50] P. J. Kuehn, “Approximate analysis of general queueing networks by decomposition,” *IEEE Trans. Commun.*, vol. COM-27, no. 1, pp. 113–126, 1979.

[51] W. Whitt, "The queueing network analyzer," *Bell Syst. Tech. Journal*, vol. 62, no. 9, pp. 2779–2815, 1983.

[52] W. Whitt, "Towards better multi-class parametric-decomposition approximations for open queueing networks," *Annals of Operations Research*, vol. 48, pp. 221–248, 1994.

[53] J. M. Harrison and V. Nguyen, "The QNET method for two-moment analysis of open queueing networks," *Queueing Systems: Theory and Applications*, vol. 6, no. 1, pp. 1–32, 1990.

[54] J. G. Dai, V. Nguyen, and M. I. Reiman, "Sequential bottleneck decomposition: an approximation method for generalized Jackson networks," *Operations Research*, vol. 42, no. 1, pp. 119–136, 1994.

[55] R. Sadre, B. Haverkort, and A. Ost, "An efficient and accurate decomposition method for open finite and infinite buffer queueing networks," in *Proc. 3rd Int. Workshop on Numerical Solution of Markov Chains*, (Zaragoza, Spain), pp. 1–20, 1999.

[56] S. Kim, R. Muralidharan, and C. A. O'Cinneide, "Taking account of correlations between streams in queueing network approximations," *Queueing Syst. Theory Appl.*, vol. 49, no. 3-4, pp. 261–281, 2005.

[57] R. Caldentey, "Approximations for multi-class departure processes," *Queueing Syst. Theory Appl.*, vol. 38, no. 2, pp. 205–212, 2001.

[58] B. Balcioglu, D. L. Jagerman, and T. Altiok, "Merging and splitting autocorrelated arrival processes and impact on queueing performance," *Perform. Eval.*, vol. 65, no. 9, pp. 653–669, 2008.

[59] A. Girard, *Routing and Dimensioning in Circuit-Switched Networks.* MA: Addison-Wesley, Readings, 1990.

[60] F. P. Kelly, "Loss networks," *Annals of Applied Probability,* vol. 1, pp. 319–378, 1991.

[61] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks.* Berlin, Germany: Springer-Verlag, 1995.

[62] G. R. Ash, *Dynamic Routing in Telecommunication Networks.* McGraw-Hill, 1998.

[63] Y. Nakagome and H. Mori, "Flexible routing in the global communication network," in *Proc. 7th Int. Teletraffic Congr.*, Session 426.1-8, (Stockholm, Sweden), 1973.

[64] R. S. Krupp, "Stabilization of alternate routing network," in *Proc. IEEE ICC'82*, (Philadelphia, Pennsylvania, US), p. 31.2.1, 1982.

[65] J. M. Akinpelu, "The overload performance of engineered networks with nonhierarchical and hierarchical routing," in *Proc. 10th Int. Teletraffic Congr.*, (Montreal, Canada), p. Session 3.2.4, 1982.

[66] T. G. Yum and M. Schwartz, "Comparison of routing procedures for circuit-switched traffic in nonhierarchical networks," *IEEE Trans. Commun.*, vol. COM-35, no. 5, pp. 535–544, 1987.

[67] F. P. Kelly, "Routing and capacity allocation in networks with trunk reservation," *Math. of Operat. Res.*, vol. 15, no. 4, pp. 771–792, 1990.

[68] S. P. Chung, A. Kashper, and K. W. Ross, "Computing approximate blocking probabilities for large loss networks with state-dependent routing," *IEEE/ACM Trans. Networking*, vol. 1, no. 1, pp. 105–115, 1993.

[69] E. W. M. Wong, A. Zalesky, Z. Rosberg, and M. Zukerman, "A new method for approximating blocking probability in overflow loss networks," *Comput. Netw.*, vol. 51, no. 11, pp. 2958–2975, 2007.

[70] G. Raskutti, A. Zalesky, E. Wong, and M. Zukerman, "Blocking probability estimation for trunk reservation networks," in *Proc. IEEE ICC '07*, (Glasgow, Scotland), pp. 223–228, 2007.

[71] F. P. Kelly, "Blocking probabilities in large circuit switched networks," *Advances in Applied Probability*, vol. 18, pp. 473–505, 1986.

[72] S. P. Chung and K. W. Ross, "Reduced load approximations for multirate loss networks," *IEEE Trans. Commun.*, vol. COM-41, no. 8, pp. 1222–1231, 1993.

[73] D. Lee, J. Kim, and S. Bahk, "Performance analysis of alternate routing with trunk reservation in multirate switched networks," in *Proc. IEEE Globecom'98*, vol. 5, (Sydney, Australia), pp. 3047–3052, 1998.

[74] A. G. Greenberg and R. Srikant, "Computational techniques for accurate performance evaluation of multirate, multihop communication networks," *IEEE/ACM Trans. Networking*, vol. 5, no. 2, pp. 266–277, 1997.

[75] D. Medhi and I. Sukiman, "Multi-service dynamic QoS routing schemes with call admission control: A comparative study," *Journal of Network and Systems Management*, vol. 8, no. 2, pp. 157–190, 2000.

[76] P. Cholda, J. Tapolcai, T. Cinkler, K. Wajda, and A. Jajszczyk, "Quality of resilience as a network reliability characterization tool," *IEEE Network*, vol. 23, no. 2, pp. 11–19, 2009.

[77] G. Iannaccone, C. Chuah, R. Mortier, S. Bhattacharyya, and C. Diot, "Analysis of link failures in an IP backbone," in *Proc. 2nd ACM SIGCOMM Workshop on Internet measurement (IMW '02)*, (New York, NY, US), pp. 237–242, 2002.

[78] R. Kawahara, T. Mori, K. Ishibashi, N. Kamiyama, and H. Yoshino, "Packet sampling TCP flow rate estimation and performance degradation detection method," *IEICE Trans. Commun.*, vol. E91-B, no. 5, pp. 1309–1319, 2008.

[79] H. Funakoshi, H. Watanabe, and H. Yoshino, “A simple estimation method of network reliability with failure scale,” *Trans. IEICE*, vol. J88-B, no. 8, pp. 1444–1453, 2005. (in Japanese).

[80] H. Funakoshi, T. Matsukawa, H. Yoshino, and N. Komatsu, “The reliability analysis and administration of telecommunication networks with social influence,” *Trans. IEICE*, vol. J91-B, no. 2, pp. 151–158, 2008. (in Japanese).

[81] H. Funakoshi, T. Matsukawa, H. Yoshino, and S. Goto, “Analysis of repair time distribution for improving the maintainability of telecommunication networks,” *Trans. IEICE*, vol. J92-B, no. 7, pp. 1153–1163, 2009. (in Japanese).

[82] H. Fukuda, R. Amano, H. Miwa, N. Kamiyama, H. Hasegawa, and H. Yoshino, “A reliable and efficient path control method and a network re-design method,” in *Proc. 4th Sino-Japanese Optimization Meeting (SJOM 2008)*, (Tainan, Taiwan), p. 82, 2008.

[83] N. Kamiyama, R. Kawahara, and H. Yoshino, “Network topology design considering detour traffic caused by link failure,” in *Proc. 13th International Telecommunications Network Strategy and Planning Symposium (Networks2008)*, Session C3, (Budapest, Hungary), 2008.

[84] R. Kawahara, E. K. Lua, M. Uchida, S. Kamei, and H. Yoshino, “On the quality of triangle inequality violation aware routing overlay architecture,” in *Proc. IEEE INFOCOM’09*, (Rio de Janeiro, Brazil), pp. 2761–2765, 2009.

[85] M. Rahnema, “Overview of the GSM system and protocol architecture,” *IEEE Communications Magazine*, vol. 31, no. 4, pp. 92–100, 1993.

[86] EIA/TIA IS-41.1, “Cellular radiotelecommnications intersystem operations,” tech. rep., EIA/TIA, 1991.

[87] S. Obana, H. Horiuchi, T. Kato, and K. Suzuki, “Applicability of OSI directory to universal personal telecommunication (UPT) and its evaluation,” *Trans. IEICE*, vol. J74-B-I, no. 11, pp. 959–970, 1991. (in Japanese).

[88] J. Z. Wang, “A fully distributed location registration strategy for universal personal communication systems,” *IEEE J. Select. Areas Commun.*, vol. JSAC-11, no. 6, pp. 850–860, 1993.

[89] B. C. Kim, J. S. Choi, and C. K. Un, “A new distributed location management algorithm for broadband personal communication networks,” *IEEE Trans. Vehic. Technol.*, vol. VT-44, no. 3, pp. 516–524, 1995.

[90] F. Dupuy, G. Nilsson, and Y. Inoue, “The TINA consortium: toward networking telecommunications information services,” *IEEE Communications Magazine*, vol. 33, no. 11, pp. 78–83, 1995.

[91] B. R. Badrinath and T. Imielinski, “Replication and mobility,” in *Proc. Second Workshop on the Management of Replicated Data*, (Monterey, California, US), pp. 9–12, 1992.

[92] R. Jain, Y.-B. Lin, C. Lo, and S. Mohan, "A caching strategy to reduce network impacts of PCS," *IEEE J. Select. Areas Commun.*, vol. JSAC-12, no. 8, pp. 1434–1444, 1994.

[93] K. K. Leung and Y. Levy, "Global mobility management by replicated databases in personal communication networks," *IEEE J. Select. Areas Commun.*, vol. JSAC-15, no. 8, pp. 1582–1596, 1997.

[94] I. F. Akyildiz and S. M. Ho, "On location management for personal communications networks," *IEEE Communications Magazine*, vol. 34, no. 9, pp. 138–145, 1996.

[95] H. Yoshino, "An adaptive congestion control for random access channels in mobile communication systems," *IEICE Trans. Commun.*, vol. E86-B, no. 2, pp. 743–756, 2003.

[96] H. Yoshino and H. Inamori, "Traffic analysis for control channel access protocols of mobile communication systems," *Trans. IEICE*, vol. J72B-I, no. 12, pp. 1173–1183, 1989. (in Japanese).

[97] H. Yoshino, "An approximation method for queueing networks with nonpreemptive priority," *Trans. IEICE*, vol. E73-E, no. 3, pp. 386–394, 1990.

[98] H. Yoshino and T. Katayama, "Queueing network analysis system : TEDAS-Q," *Trans. IPSJ*, vol. 31, no. 4, pp. 624–633, 1990. (in Japanese).

[99] H. Yoshino and Y. Hoshiai, "End-to-end blocking in an integrated services network with link capacity allocation control," in *Proc. ITC Specialists Seminar*, (Krakow, Poland), pp. 581–590, 1991. (in Telecommunication Services for Developing Economies, J.Filipiak(Editor), Elsevier Science Publishers B.V. (North-Holland)).

[100] H. Yoshino, H. Yamamoto, and H. Matsue, "Mobility management schemes and their characteristics for advanced personal communication services in distributed environments," *IEICE Trans. Commun.*, vol. E81-B, no. 6, pp. 1162–1170, 1998.

[101] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey," *IEEE Personal Commun.*, vol. 6, pp. 10–31, 1996.

[102] Y. Furuya and Y. Akaiwa, "Channel segregation — a distributed adaptive channel allocation scheme for mobile communication systems," in *Proc. 2nd Nordic Seminar on Digital Mobile Radio Communications DMR II*, (Stockholm,Sweden), pp. 311–315, 1987.

[103] A. Koike and H. Yoshino, "Traffic design and administration for distributed adaptive channel assignment methods in microcellular systems," *IEICE Trans. Commun.*, vol. E78-B, no. 3, pp. 379–386, 1995.

[104] T. S. G. R. A. Networks, "Physical layer procedures (FDD)," Tech. Rep. 3G TS 25.214 V.3.2.0, 3rd Generation Partnership Project, 2000.

[105] N. Yoshikawa, S. Okasaka, and H. Komagata, "UHF land mobile telephone radio control techniques," *Trans. IECE*, vol. J61-B, no. 2, pp. 83–90, 1977. (in Japanese).

[106] T. Yoshikawa, E. Hagiwara, H. Komagata, and H. Mishima, "Link configuration for mobile satellite communication systems," in *Proc. IEEE ICC'85*, (Chicago,US), pp. 761–765, 1985.

[107] H. Yoshida and T. Mitsuichi, "The throughput characteristics in a maritime/aeronautical telephone control channel," in *Proc. IECE Gen. Conf.*, p. 2419, 1985. (in Japanese).

[108] L. Kleinrock, *Queueing Systems — Computer Applications*, vol. 2, ch. 5. John Wiley & Sons, 1976.

[109] H. Inamori, H. Yoshino, H. Komagata, and Y. Yasuda, "Performance evaluation for control channel access schemes in mobile satellite communication systems," *Trans. IEICE*, vol. E72, no. 1, pp. 43–54, 1989.

[110] I. Iida and Y. Yasuda, "Performance analysis of CSMA/CD protocol with backoff algorithms," *Trans. IECE*, vol. J66-B, no. 10, pp. 1247–1254, 1983. (in Japanese).

[111] Y. Takahashi, "An approximation for multiserver loss system with renewal batch-inputs and general service times(II) : Diffusion approximation," *Technical report of IEICE*, vol. SE83-91, 1983.

[112] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems.* London: Acad. Press, 1980.

[113] Y. Takahashi, "Mean-delay approximation for a single server priority queue with batch arrivals of two classes," *Trans. IEICE*, vol. E72, no. 1, pp. 29–36, 1989.

[114] N. K. Jaiswal, *Priority Queues.* New York: Academic Press, 1968.

[115] W. Whitt, "Refining diffusion approximations for queues," *Opns. Res. Letters*, vol. 1, pp. 165–169, 1982.

[116] W. Kraemer and M. Langenbach-Belz, "Approximate formulae for the delay in the queueing system GI/G/1," in *Proc. 8th Int. Teletraffic Congr.*, Session 23.5, (Melbourne, Australia), 1976.

[117] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. JSAC-4, no. 6, pp. 847–855, 1986.

[118] J. W. Roberts, "Teletraffic models for the telecom 1 integrated services network," in *Proc. 10th Int. Teletraffic Congr.*, Session 1.1.2, (Montreal, Canada), 1982.

[119] M. Pioro, J. Lubacz, and U. Koerner, "Traffic engineering problems in multiservice circuit switched networks," in *Proc. ITC Specialists Seminar*, Session 11.3, (Adelaide), 1989.

[120] ITU-T, “Principles of universal personal telecommunication (UPT),” ITU-T, Recommendation F.850, ITU-T, 1995.

[121] ITU-T, “Universal personal telecommunication (UPT) - service description (service set 1),” ITU-T, Recommendation F.851, ITU-T, 1995.

[122] H. Matsue and H. Yamamoto, “Mobility management concept based on distributed processing environment for personal communications,” in *Proc. IEEE ICUPC'94*, (San Diego), 1994.

[123] Y. B. Lin, “Modeling techniques for large-scale PCS networks,” *IEEE Communications Magazine*, vol. 35, no. 2, pp. 102–107, 1997.

[124] I. Iida, T. Nishigaya, and K. Murakami, “DUET: An agent-based personal communications network,” *IEEE Communications Magazine*, vol. 33, no. 11, pp. 44–49, 1995.

[125] N. Funabiki and Y. Takefuji, “A neural network parallel algorithm for channel assignment problems in cellular radio networks,” *IEEE Trans. Vehic. Technol.*, vol. VT-41, no. 4, pp. 430–437, 1992.

[126] M. Duque-Anton, D. Kunz, and B. Ruber, “Channel assignment for cellular radio using simulated annealing,” *IEEE Trans. Vehic. Technol.*, vol. VT-42, no. 1, pp. 14–21, 1993.

[127] H. Furukawa and Y. Akaiwa, “Self-organized reuse partitioning (SORP), a distributed dynamic channel assignment method,” *Technical report of IEICE*, vol. RSC92-126, pp. 61–66, 1993. (in Japanese).

[128] J. P. Buzen, “Computational algorithms for closed queueing networks with exponential servers,” *Comm. ACM*, vol. 16, no. 9, pp. 527–531, 1973.

[129] Y. Furuya and Y. Akaiwa, “Channel segregation, a distributed adaptive channel allocation scheme for mobile communication systems,” *Trans. IEICE*, vol. E74, no. 6, pp. 1531–1537, 1991.