Call Admission Control, Bandwidth Adaptation, and Scheduling in Cellular Wireless Internet: Analytical Models and Performance Evaluation

by

Dusit Niyato

B. Engg., King Mongkut's Institute of Technology Ladkrabang, Thailand, 1999

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

in the Department of Electrical and Computer Engineering

We accept this thesis as conforming
to the required standard

Prof. E. Hossain, Supervisor, Dept. of Electrical & Computer Engineering

---

Dr. J. Diamond, TR*Labs* and Dept. of Electrical & Computer Engineering

---

Prof. S. Noghanian, Dept. of Electrical & Computer Engineering

---

Prof. J. Misic, External Examiner, Dept. of Computer Science

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

Cell Admission Control, Bandwidth Adaptation, and Scheduling in Cellular Wireless Internet: Analytical Models and Performance Evaluation

BY

Dusit Niyato

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Supervisor: Prof. E. Hossain

ABSTRACT

In this thesis, we address admission control, bandwidth adaptation and scheduling problems in a cellular wireless Internet environment. While the admission control is responsible for deciding whether an incoming call/connection can be accepted or not, bandwidth adaptation and scheduling are used to allocate the available resources among the ongoing calls/connections adaptively. We provide an extensive survey of the existing admission control algorithms. The issues related to and the approaches for designing admission control and bandwidth adaptation in fourth-generation (4G) cellular wireless networks are discussed. An admission control method considering the quality of service (QoS) requirements in both the wireless and the wired part of the networks is presented.

At the mobile end, we propose a service differentiation model for QoS-sensitive and best-effort traffic. Traffic scheduling is used to grant access to wireless channel. Admission control is used for QoS-sensitive traffic to limit the number of ongoing connections so that the QoS requirements can be met.

A performance model for cellular networks with adaptive bandwidth allocation by using Markov arrival process (MAP) for call arrival and phase-type (PH) distribution for channel holding time is presented. In the presence of time-varying traffic, the transient behavior of a cellular wireless network is analyzed and an adaptive admission control method based on the transient analysis is presented. The analytical framework to investigate the call-level and the packet-level performances are presented. We consider hard capacity systems (i.e., TDMA and FDMA). By using this framework, the interdependencies among the call-level and the packet-level QoS metrics can be examined, and the optimal values for the system parameters and the admission criterion can be obtained.

We assume that the cellular wireless network internetworks with the Differentiated Services (DiffServ)-based wired Internet. For this cellular wireless Internet access scenario, we present an optimization formulation to obtain the traffic shaping parameters at the mobile end so that the packet delay is minimized.

Examiners:

Prof. E. Hossain, Supervisor, Dept. of Electrical & Computer Engineering

---

Dr. J. Diamond, TR*Labs* and Dept. of Electrical & Computer Engineering

---

Prof. S. Noghanian, Dept. of Electrical & Computer Engineering

---

Prof. J. Misic, External Examiner, Dept. of Computer Science

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to thank

- Professor Ekram Hossain for his guidance and support,
- Jeff Diamond for his helpful comments,
- TR*Labs* for research funding and facilities,
- Teerawat Issariyakul for his suggestion and discussion,
- my family, Usa (my mom), Tawat (my dad), Wittaya (my brother) for their efforts to give me to have joyful life,
- my love, Amporn Udomsakakul,

and all of my friends in Thailand and Canada.

Chapter 1

Introduction

Cellular wireless technology today has become the prevalent technology for wireless networking. Not only mobile phones but also other types of devices such as laptops and *Personal Digital Assistant (PDA)* can connect to Internet via cellular infrastructure. These mobile devices are often capable of running multimedia applications (e.g., video, images). Therefore, cellular networks need to provide quality of service (QoS) guarantee to different types of data traffic in a mobile environment. A call admission control (CAC) scheme aims at maintaining the delivered QoS to the different calls (or users) at the target level by limiting the number of ongoing calls in the system. One major challenge in designing a CAC arises due to the fact that the cellular network has to service two major types of calls: new calls and handoff calls. The QoS performances related to these two types of calls are generally measured by new call blocking probability and handoff call dropping probability. In general, users are more sensitive to dropping of an ongoing and handed over call than blocking a new call. Therefore, a CAC scheme needs to prioritize handoff calls over new calls by minimizing handoff dropping probability.

Again, bandwidth adaptation and scheduling are necessary mechanisms for achieving high utilization of the wireless resources (e.g., channel bandwidth) while satisfying the QoS requirements for the users. These two techniques are closely related to call admission control, and in fact these three mechanisms jointly determine the call-level and the packet-level QoS for the different types of traffic in the cellular wireless air interface. For example, upon arrival of a new call or handoff call, bandwidth adaptation can be performed to degrade the channel allocations for some calls (still maintaining the QoS requirements) so that the new call can be admitted. Scheduling mechanisms impact the packet-level system dynamics (e.g, queueing behavior), and

therefore, packet-level QoS. The packet-level dynamics can be exploited for designing efficient call admission control methods.

The call admission control (CAC) and the adaptive channel adaptation (ACA) mechanisms are generally treated as the network layer (above layer-2) functionalities in the wireless transmission protocol stack (Figure 1.1). The scheduling and the adaptive modulation and coding (AMC) are layer-2 and layer-1 (i.e., physical) functionalities, respectively.

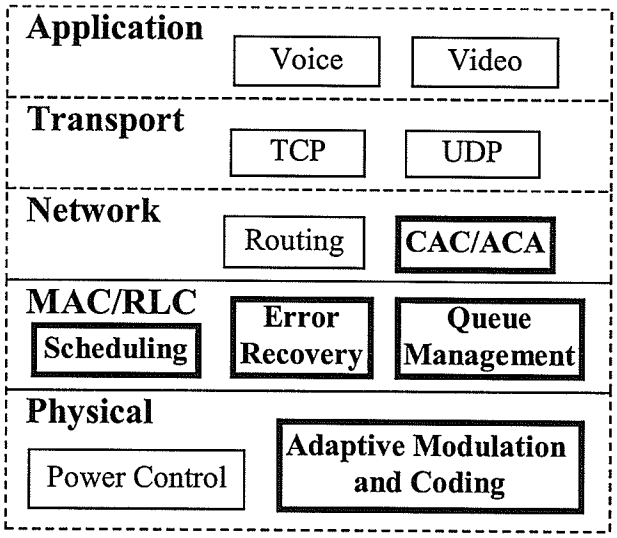


Figure 1.1. *Standard protocol stacks and their components.*

Cellular networks are envisioned to interwork with the *Differentiated Services (DiffServ)* [1]-based wired Internet in the next-generation wireless networks. The DiffServ architecture for such an integrated network would be attractive from the scalability point of view, since DiffServ networks provide group-based QoS rather than flow-based QoS as in the *Integrated Services (IntServ)*-based networks. In DiffServ networks, there are two main components, edge routers and core routers. While the core routers are used within a DiffServ domain to form the infrastructure, edge routers connect a DiffServ domain to the external networks. In an integrated cellular and DiffServ architecture, edge routers would decide whether an incoming connection can

be admitted or not.

1.1 Objectives, Motivations, and Scopes of the Thesis

The main objective of this thesis is to study the call admission control, bandwidth adaptation and scheduling problems in a conventional cellular wireless network and in the next-generation cellular wireless Internet. We intend to develop analytical models to study the performance of the different CAC mechanisms under different system parameter settings and their impact on both the call-level and the packet-level QoS performances.

Comprehensive analytical models are required to study the interdependencies among the different system parameters and the performance measures and also to explore the inter-layer protocol interactions. Such models can be used to obtain the performance measures efficiently compared to computer simulations. Again, based on the analytical models optimization formulation can be developed and solved so that optimal system parameter setting can be obtained. In summary, for design and engineering of call admission control, bandwidth adaptation and scheduling, comprehensive analytical frameworks would be required.

While traditionally the call admission control and the bandwidth adaptation problems have been addressed for voice-oriented cellular networks, there is a need to revisit these problems and the related issues for data-oriented packet-switched cellular networks. For example, while the major performance measures in circuit-switched networks are call blocking and call dropping probabilities, in packet-switched networks the principle performance measures are packet delay and packet loss ratio. The CAC and the bandwidth adaptation schemes for the next-generation wireless networks must take both the call-level and the packet-level performance measures into account.

In this thesis, we provide an overview of issues and approaches in designing CAC schemes for the next-generation (e.g., fourth-generation (4G)) wireless networks. A CAC algorithm that considers both QoS performance in wireless air interface and DiffServ domain for connecting to the Internet is presented.

Call admission control would be also required at the end mobile to prioritize among different types of connections. We propose a system architecture for service differentiation among wireless connections in which two types of traffic (i.e., QoS-sensitive and best-effort) are considered. The CAC is used for QoS-sensitive queue to limit the number of connections so that the performances of the ongoing connections do not deteriorate. An analytical model based on discrete time Markov chain is presented. We also formulate and solve an optimization problem to obtain near-optimal parameter setting.

At the base station, CAC is an important component to guarantee call-level performance to the users. However, most of the existing CAC schemes are for single class of users (i.e., voice call) and they ignore the traffic burstiness which is common in operational environment. We propose an analytical model for CAC with multiple classes of services in which the burstiness of the traffic is captured.

To enhance the network resource utilization, ACA is necessary. ACA allocates available channels among the ongoing calls according to the traffic load in the cell. Most of the analytical works on ACA in the literature, considered Poisson call arrival rate and exponential channel holding time. However, studies have shown that channel holding time in micro and pico-cellular environments is not exponentially distributed. Therefore, we present a new analytical model by considering Markov arrival process (MAP) for call arrival and phase-type (PH) distribution for channel holding time. Both MAP and PH models are general in which correlation in the inter-arrival and service time can be captured.

Most of the analytical models for CAC in the literature considered the performance only at steady state. However, since in an actual environment call arrival and channel holding time strongly depend on time of the day and day of the week, the steady state might be never reached. Therefore, we present transient analysis for CAC under time-varying traffic. By using our model, transient behavior of the system can be investigated. We also propose adaptive call admission control in which resource reservation is dynamically adjusted according to traffic load and transient behavior of the system.

We present novel queueing models for analyzing both call and packet-level in hard capacity wireless systems (i.e., Time-Division Multiple Access (TDMA) and

Frequency-Division Multiple Access (FDMA)). ACA is used to adaptively allocate available channels to multimedia calls (i.e., real-time, non-real-time and best-effort). In the physical layer, we consider adaptive modulation and coding (AMC) and we capture the effects of correlated channel fading by using a finite state Markov chain (FSMC) model. Automatic repeat request (ARQ) is used for error control such that the reliability of the transmission can be ensured. An application of the proposed model for optimal channel pool partitioning among three multimedia service classes is presented.

For an integrated cellular and DiffServ network, we apply optimization technique to obtain parameters of traffic shaper at DiffServ edge router employed with generalized processor sharing traffic scheduling. Based on optimal token bucket parameters, traffic delay can be minimized while increasing the resource utilization.

1.2 System Architecture

We consider a cellular wireless Internet architecture (Figure 1.2) with three major components, namely, mobile node, base station and DiffServ edge router. There are two types of traffic at the mobile node - QoS-sensitive and best-effort, and two separate queues are used for these traffic. There is no performance guarantee for the traffic in the best-effort queue, and therefore, no admission control is performed for this type of traffic. However, the number of ongoing connections for the QoS-sensitive traffic needs to be limited so that the QoS performances do not degrade below the desired level. The service between the QoS-sensitive queue and the best-effort queue is differentiated by fair scheduling with work-conserving property. Adaptive modulation and coding is used for transmission between the mobile and the base station. Automatic repeat request is used to retransmit erroneous packet to ensure reliability of transmission.

We consider hard capacity systems (i.e., FDMA, and TDMA). Call admission control is employed to limit the number of ongoing calls, prioritize handoff calls over new calls, and reserve resources for the different classes of calls. Also, adaptive channel (bandwidth) allocation is used to allocate available channels to the ongoing calls according to the traffic load in a cell. This base station is connected to the

DiffServ edge router. *Generalized processor sharing (GPS)* [2] is used to differentiate services from different connections.

The major wireless protocol components considered in this thesis (i.e., CAC, ACA, error control, queue management and adaptive modulation and coding) are shown in the transmission protocol stack in Figure 1.1.

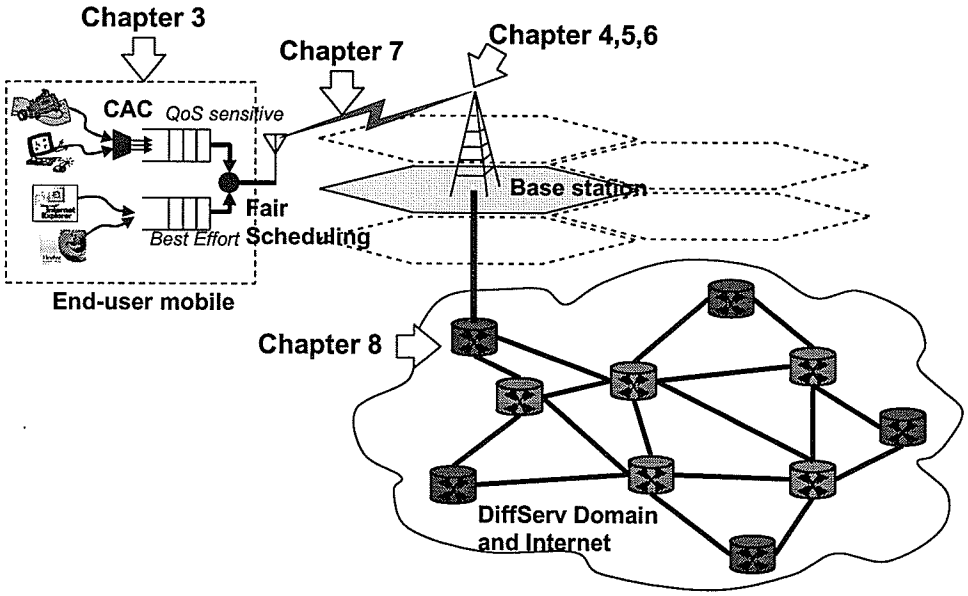


Figure 1.2. *System architecture.*

1.3 Organization of This Thesis

The organization of this thesis is as follows:

- **Chapter 2** provides a background and survey on the existing CAC algorithms in the literature. The issues and design approaches for CAC schemes in 4G wireless networks are discussed. Then, a two-tier CAC scheme which considers QoS at both the wireless and the wired part of the network is proposed.
- **Chapter 3** presents a service differentiation model at the mobile considering two types of traffic, namely, QoS-sensitive and best-effort. CAC is used at

the QoS-sensitive queue to limit the number of ongoing connections, and fair scheduling is used to serve both of these two queues. Adaptive modulation and coding is used at the physical layer to increase the transmission rate by exploiting the dynamic channel variations. An optimization problem is formulated and the near-optimal parameter settings are obtained.

- **Chapter 4** presents an analytical model for CAC considering burstiness in the call arrival pattern. This burstiness in the call-level traffic is represented by a *Markov modulated Poisson process*. The model considers multiple classes of service in a cellular network. An optimization problem is formulated and the optimal number of reserved channels for handoff calls is obtained.
- **Chapter 5** presents an analytical model for CAC in cellular networks with Markov call arrival process and phase-type channel holding time. This model is general and applicable to any system in which call interarrival and channel holding times are not exponentially distributed.
- **Chapter 6** presents a novel analytical model for CAC with static and adaptive bandwidth allocations in a cellular network with time-varying traffic. Both steady and transient state performances are analyzed. Based on the transient behavior of the system, an adaptive CAC algorithm is introduced.
- **Chapter 7** presents an analytical framework for cross-layer study of call-level and packet-level QoS in wireless mobile multimedia networks. CAC and ACA are used to limit number of ongoing calls and allocate available channels for three service classes (i.e., real-time, non-real-time and best-effort). The model considers adaptive modulation and coding in the physical layer, and ARQ as well as queue management at the link layer for non-real-time traffic. Examples on the applications of the proposed model for resource allocation for multimedia services are presented.
- **Chapter 8** deals with the problem of choosing optimal parameter settings for traffic shaping under the generalized processor sharing scheduling policy used at a DiffServ edge router. By using a searching technique, the token bucket size and the token generation rate for a token bucket traffic shaper are obtained.
- **Chapter 9** summarizes the contributions of this thesis and outlines a few directions for future research.

Chapter 2

Call Admission Control for QoS Provisioning in 4G Wireless Networks: Issues and Approaches

2.1 Introduction

Supporting multimedia applications with different quality of service (QoS) requirements in the presence of diversified wireless access technologies (e.g., 3G cellular, IEEE 802.11 WLAN, Bluetooth) is one of the most challenging issues for the forth-generation (4G) wireless networks. In such a network, depending on the bandwidth, mobility, and application requirements, users will be able to switch among the different access technologies in a seamless manner. Efficient radio resource management and call admission control (CAC) strategies will be key components in such a heterogeneous wireless system supporting multiple types of applications with different QoS requirements.

A call admission control scheme aims at maintaining the delivered QoS to the different calls (or users) at the target level by limiting the number of ongoing calls in the system. One major challenge in designing a CAC arises due to the fact that the network has to service two major types of calls: new calls and handoff calls. The QoS performances related to these two types of calls are generally measured by new call blocking probability and handoff call dropping probability. In general, users are more sensitive to dropping of an ongoing and handed over call than blocking a new call. Therefore, a CAC scheme needs to prioritize handoff calls over new calls to keep

the handoff dropping probability below some threshold (e.g., 5%). Also, the new call blocking probability should be maintained below or at the target level. After all, the resource utilization should be maximized while achieving the QoS requirements.

In contrast to the traditional voice-oriented circuit-switched cellular wireless networks, the 4G networks will be based on packet-switching at the wireless interface and will be internetworked with IP-based Internet. Therefore, while designing a CAC scheme for such a network, packet-level performance measures (e.g., packet dropping probability and packet transmission delay) at both the wireless interface and the wired interface (e.g., at the IP-aware wireless router/base station) will need to be considered in addition to the call-level performance measures.

For the evolving 4G networks, the traditional and existing CAC algorithms (e.g., those for 3G systems) needs to be modified with a view to

- Handling the handoff calls between the networks (i.e., *vertical handoff*). Resource reservation and call admission control become more complicated due to the presence of heterogeneous wireless access environment in which the mobile terminals have the ability to connect to different types of networks.
- Prioritizing the different types of calls. Some calls might use real-time applications which have more strict QoS requirements than those using non-real-time applications. CAC must be performed based on different QoS requirements.
- Taking into account packet-level performances. 4G wireless networks will operate purely on packet-based data transfer. The CAC algorithm must consider not only call-level QoS but also the packet-level QoS. In other words, the CAC algorithm needs to evaluate the availability of the network resources by taking into account the packet-level performance statistics.
- Internetworking with the IP-based Internet. The CAC algorithm should be aware of the availability of the network resources at the wireless-Internet gateway and in the wired network so that wireless resources are not wasted due to dropping of packets at the wired-part of the network.

The rest of the chapter is organized as follows. The general model and the components of call admission control and its classifications are presented in Section 2. A survey of the traditional CAC schemes is presented in Section 3. The challenges and the issues in designing the CAC schemes for 4G networks are outlined in Section 4.

Section 5 presents the architecture of a novel CAC method, which considers resource allocation and management at both the air interface and the wireless-Internet gateway. Summary are stated in Section 6.

2.2 CAC: General Model and Classification

2.2.1 Components of CAC

In general, a CAC scheme has three main components: information management, resource reservation, and admission control. These three components collaborate with each other by exchanging information (Figure 2.1) with a view to achieving specific CAC objectives such as minimizing QoS degradation or maximizing the revenue.

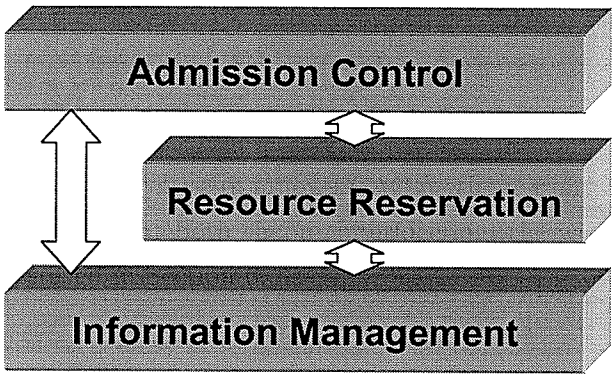


Figure 2.1. *Components of a call admission control mechanism.*

2.2.1.1 Information management

This is required for storing, exchanging, and maintaining the state of the network. The type and amount of information (e.g., number of channels used by ongoing calls, position and velocity of a mobile) depends on the types of the CAC algorithm. These information can be either exchanged among the cells or used locally.

2.2.1.2 Resource Reservation

This is required to reserve the resources according to the users' needs. This component can consult the information management component and use optimization and prediction techniques to calculate the amount of resources to be reserved for handoff calls and maximize the resource utilization.

2.2.1.3 Admission Control

The admission control component is responsible for making decision on whether an incoming call (either a new call or a handoff call) can be accepted or not. This component consults the information management and resource reservation components. Most of the admission control algorithms are rule-based, i.e., based on examining the pre-defined conditions. The outcome of the algorithm could be either to accept the call, reject the call, or queue the call until resources become available.

As an example, for a guard channel-based CAC scheme [3], the information management component maintains the current number of busy channels, and the resource reservation component reserves a pool of channels for handoff calls. The admission control module uses information from both the components to decide whether a call can be accepted or not.

2.2.2 Threshold-Based Mechanism: The General CAC Principle

The concept of threshold-based CAC is applicable for both hard- and soft-capacity wireless systems. A threshold-based CAC is based on the availability of resource stored in vector $\mathbf{I}$ [4]. The objective of a threshold-based CAC is to maintain every element in $\mathbf{I}$ less than that in the threshold stored in vector $\mathbf{I}_{th}$. These thresholds are defined based on the congestion condition of the system. When a call arrives, the algorithm estimates the increase $\Delta\mathbf{I}$ that the incoming call would affect to the current value of $\mathbf{I}$. Generally, the policy of CAC is based on the condition

$$\mathbf{I} + \Delta\mathbf{I} < \mathbf{I}_{th}. \tag{2.1}$$

If this condition is satisfied, the incoming call is accepted, otherwise it is rejected or queued. With this policy, the CAC scheme needs to be developed by considering

the elements of vector $\mathbf{I}$, which are the performance measures of the system, the way to estimate the increase $\Delta\mathbf{I}$, and the optimal value of the threshold $\mathbf{I}_{th}$. In this case, the threshold can be set statically without considering the current status of the network (static scheme). However, adaptive CAC algorithms (e.g., [5]) can adjust the thresholds dynamically resulting in better performance over static schemes.

In case of systems with hard capacity (i.e., TDMA and FDMA systems), the elements of matrix $\mathbf{I}$ can be simply the number of occupied channels, and the increase in the resource usage $\Delta\mathbf{I}$ can be the number of channels required by a incoming call. In this case, the thresholds can be chosen so that the target QoS levels are achieved.

2.2.3 Classification

2.2.3.1 Centralized and Distributed Approaches

A call admission control algorithm can operate in a centralized or in a distributed fashion. In the former case, the CAC algorithm is executed in a central site (e.g., mobile switching center (MSC)). In this case, information from every cell needs to be transferred to the central site and the CAC needs to be performed remotely from the local cell. In case of distributed CAC, the CAC algorithm is executed locally at the base station of each cell.

A distributed CAC algorithm can follow either a collaborative or a local approach. In the former case, information is exchanged among the neighboring cells for resource reservation and admission control while the decision is made locally. In the latter case, information collection and decision making are done locally. Although the collaborative approach can provide more accurate information for CAC decision, it incurs more communication overhead.

2.2.3.2 Traffic Descriptor-Based and Measurement-Based Approaches

A call admission control policy can use either a traffic descriptor-based or a measurement-based approach. In the traffic descriptor-based approach, it is assumed that the knowledge about the traffic pattern of the incoming calls is available. Therefore, a CAC algorithm can simply determine the expected amount of resource usage by summing all the resources used by all ongoing calls and the incoming calls together. If

this is less than some predefined threshold, the incoming call is accepted, otherwise the call is rejected. Although traffic descriptor-based CAC is simple, it is relatively conservative, since the ongoing calls will not use maximum amount of resource as specified in the descriptor all the time.

Instead of using explicit traffic descriptors, the information on the traffic pattern can be obtained by measuring the characteristics of the call. In that case, call admission control decision can be made dynamically based on the actual state of the network. Measurement-based CAC schemes are based on this principle.

Most of the CAC algorithms in the hard-capacity cellular networks are traffic descriptor-based. Most of CAC algorithms in CDMA systems are measurement-based in which SIR information is measured and used to ensure the QoS of the ongoing calls. The CAC algorithms used in WLANs mostly adopt a measurement-based approach.

2.2.3.3 Classification Based on the Granularity of Resource Control

A taxonomy of CAC algorithms by considering the granularity of the resource control was presented in [6]. Three different criteria were used to categorize a CAC algorithm. The first criterion is the type of information used by decision making process for call admission control. Generally, CAC algorithms consider resource usage of the mobiles, but some of the algorithms consider the mobility patterns of the users. In the latter case, the accuracy of the resource reservation can be improved by taking the direction and the speed of the users into account.

The second design criterion is the spatial distribution (position and movement) of the mobiles which can be either uniform or non-uniform. The third criterion is based on how the information is organized and manipulated by the CAC algorithms. The information can be on the aggregate of flows or on the per-user basis and the CAC algorithms use the information on the group of mobiles or on individual mobile, respectively. Based on these criteria, the granularity of the CAC algorithms are different. For example, the algorithm which considers the resource usage of all ongoing calls assuming uniform spatial distribution (e.g., the guard channel scheme) has the largest, and the algorithm which considers the mobility of individual user assuming non-uniform spatial distribution has the smallest granularity of resource control.

2.3 Traditional CAC Approaches in Cellular Networks

2.3.1 Guard Channel Approach

To prioritize handoff calls over new calls, some channels (referred to as guard channels) are reserved for handoff calls [3]. Specifically, if the total number of available channels is C and the number of guard channels is $C - K$, a new call is accepted if the total number of channels used by ongoing calls (i.e., busy channels) is less than the threshold K, while a handoff call is always accepted if there is an available channel. According to this channel reservation, the threshold must be chosen such that the handoff call dropping probability is minimized while the system can admit as many incoming calls as possible.

Although the guard channel scheme with a static threshold is easy to implement, it may not be efficient. Higher channel utilization could be achieved through adaptation of the threshold according to the state of the network.

A more general scheme, namely, the *fractional guard channel scheme* was introduced in [7]. In this case, an incoming call is accepted with certain probability which depends on the number of busy channels. In other words, when the number of busy channels becomes larger, the probability for accepting a new call becomes smaller and vice versa. This helps to keep the handoff call dropping probability lower than the desired value and also avoid congestion in the cell.

2.3.2 Partitioning and Sharing Approach

This approach for reserving channels and admitting new calls is based on the concepts of *complete sharing* and *complete partitioning*. In case of complete sharing, the handoff calls and new calls can use all the available channels. In contrast, in case of complete partitioning the channels reserved for handoff and new calls are not shared between these types of calls.

Instead of using pure complete sharing, which is unable to prioritize the calls, and complete partitioning, which is relatively conservative, a hybrid model for resource reservation and CAC was proposed in [8]. In this hybrid scheme, the channels are

divided into three categories: channels dedicated for new calls, channels shared among the new calls and handoff calls, and channels reserved for handoff calls. By combining complete partitioning and complete sharing, resource reservation becomes flexible to control the performance of the system. This type of hybrid resource reservation can handle calls with different priority levels as well.

2.3.3 Collaborative Approach Based on Estimation

This is a distributed approach for call admission control. In this case, information is exchanged among the neighboring cells for resource reservation and admission control, while the admission control decision is made locally. CAC algorithms of this type were proposed in [9] which use estimates of call dropping and call blocking probabilities. The maximum number of ongoing calls N is estimated from the following:

$$P_{hd} = \frac{1}{2}\text{erfc}\left(\frac{N - \bar{m}}{\sigma}\right) \tag{2.2}$$

where P_{hd} is the target call dropping probability, and $\bar{m}$ and σ denote the mean and variance of the number of calls in the home cell, respectively. The mean and variances are approximated from the number of users in the home cell and the neighboring cells. The call blocking probability $P_{nb}(t)$ at time $t-1$ to t is estimated locally as follows:

$$P_{nb}(t) = (1-\omega)P_{nb}(t-1) + \omega\frac{s(t)}{r(t)} \tag{2.3}$$

where $s(t)$ and $r(t)$ are the number of blocked calls and the number of calls that arrived during time interval $t-1$ to t, respectively, and ω is the weight used for calculating the exponential weighted moving average. The decision on whether an incoming call is accepted or rejected is made based on (2.2) and (2.3)).

2.3.4 Non-Collaborative Approach Based on Prediction

In a pico-cellular wireless network with high user mobility, exchanging information among the cells to make resource reservation and admission control might incur significant control overhead. Therefore, CAC algorithms designed based on local information (e.g., history of bandwidth usage) would be desirable. In such a case, resource

reservation is based only on local information in the home cell which is used to predict the resource needed in the future [5].

In [5], two prediction techniques were used: *Wiener filtering* and *time series analysis* (e.g., ARMA (autoregressive moving average) model). In the former case, the prediction can be done directly from the historic data, whereas in the latter case the time series model needs to be constructed and the corresponding parameters need to be estimated so that the prediction can be performed based on this model afterwards. Such a local predictive approach to call admission control was shown to perform as good as a collaborative approach when the traffic fluctuation is moderate.

2.3.5 Mobility-Based Approach

Mobility-based approaches exploit the user mobility information for efficient call admission control. For example, in [10], based on user mobility information *shadow clustering* concept was introduced to estimate future resource requirements in a wireless network with microcellular architecture. The idea here is that every mobile terminal with an active wireless connection exerts an influence upon the cells (and their base stations) in the vicinity of its current location and along its direction of travel.

Figure 2.2 shows the shadow cluster for a mobile in cell C which is moving towards north. The cells in a shadow cluster usually have different levels of predicted traffic intensity. For example, in Figure 2.2, the predicted traffic intensity in cells denoted by B in the vicinity of cell C will be more than that in the cells denoted by A. To calculate the shadow cluster and the corresponding levels of intensity, the information on call holding time, current direction, velocity, and position of the active mobile terminal need to be considered.

The base station in the home cell must inform its neighboring base stations of the location and mobility of the active mobile terminal (e.g., this information can be obtained from global positioning system). In a cell, the amount of resources reserved for handoff calls is based on the number of calls moving to that cell and the corresponding probabilities. The estimated resource usage $C_{u_j}(t)$ in cell j at time t is expressed as follows:

$$C_{u_j}(t) = C_{u_j}(t-1) - C_{u_j}^{out}(t) + C_{u_j}^{in}(t) \tag{2.4}$$

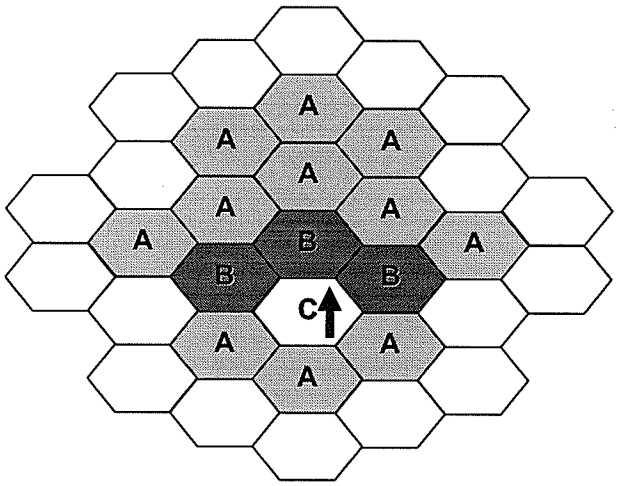


Figure 2.2. *Shadow clustering: C is the home cell of active mobile terminal, B is the bordering neighbor, and A denote nonbordering neighbor.*

where $C_{u_j}^{out}(t)$ is the estimated amount of resources that will be freed by active users whose calls will be either terminated or handed over to other cells and $C_{u_j}^{in}(t)$ is the estimated amount of resources that will be occupied by active mobile terminals moving from neighbors cells within the shadow cluster. Both of them are functions of the probability of active mobile terminal x moving from cell k to cell j at time t ($P_{x,k,j}(t)$) as follows:

$$C_{u_j}^{out}(t) = \sum_{\forall x, \forall k, \forall j} (1 - P_{x,k,j}(t))\, c(x) \tag{2.5}$$

$$C_{u_j}^{in}(t) = \sum_{\forall x, \forall k, \forall j} P_{x,k,j}(t) c(x) \tag{2.6}$$

where $c(x)$ is the resource used by active mobile terminal x.

Although the mobility information-based CAC schemes can improve the efficiency of the resource reservation and admission control, calculating the probabilities $P_{x,k,j}(t)$ would be non-trivial and also real-time exchange of control messages among the cells would incur large communication overhead.

2.3.6 Pricing-Based Approach

A pricing-based approach to call admission control was proposed in [11], where the objective is to maximize the utility of the wireless resources. The utility is generally defined as the users' level of satisfaction with perceived QoS. For example, the utility is a decreasing function of new call blocking and handoff call dropping probabilities. However, maximizing the utility of the network might not maximize the revenue of the service provider. The tradeoff between the user satisfaction and revenue is illustrated in Figure 2.3. Specifically, for higher user satisfaction, more resources should be allocated to each user. In contrast, to maximize revenue under flat rate pricing, the allocations need to be degraded to accommodate more number of users. Therefore, a CAC scheme can be designed such that the optimal point can be obtained.

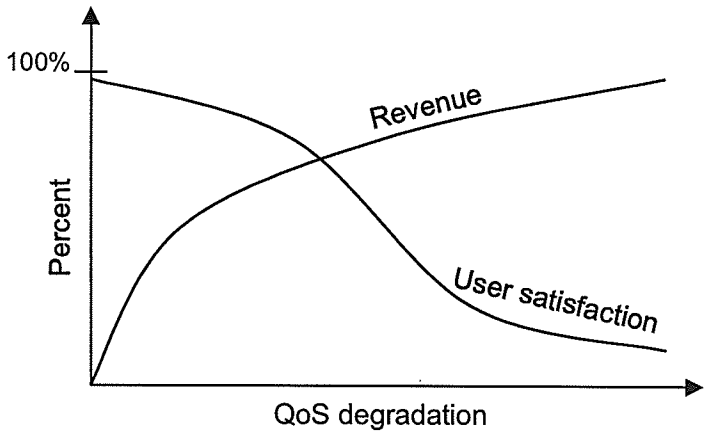


Figure 2.3. *Tradeoff between QoS degradation and network revenue.*

In [11], the optimal point between utility and revenue was determined in terms of the new call arrival rate, and a pricing scheme was developed to achieve this optimal effective arrival rate in the network. In this case, the QoS metric P_b referred to as the grade of service (GoS) is defined as follows:

$$P_b = \alpha P_{nb} + \beta P_{hd} \tag{2.7}$$

where α and β are the weights corresponding to the new call blocking and handoff call dropping probabilities, respectively, and $\alpha + \beta = 1$.

The metric P_b can be defined as a function of new call arrival rate λ_n (i.e., $P_b = g(\lambda_n)$). Then, the utility function becomes $U = h(P_b)$. Assuming a flat rate pricing, the revenue depends on the number of admissible users which again depends on the new call arrival rate $f(\lambda_n)$. The optimal value of the new call arrival rate λ_n^* that maximizes the total utility $U(\lambda_n) = f(\lambda_n) \times h[g(\lambda_n)]$ can be calculated by differentiating $U(\lambda_n)$, and finding the point at which the slope equals to zero.

Based on this optimal new call arrival rate, the pricing scheme is developed by changing the cost of a call. The pricing scheme adjusts the fee dynamically by taking the state of the network into account. If the network is congested, it will charge *peak hour price* $p(t)$ which is higher than the normal hour price p_0. According to this pricing scheme, the demand function which describes the reaction of users to the change of price is expressed by

$$D[p(t)] = \exp\left(-\left(\frac{p(t)}{p_0} - 1\right)^2\right), \quad p(t) \geq p_0 \tag{2.8}$$

and the peak hour price is calculated by considering the state of the network. With this pricing scheme, a user has an incentive not to initiate a call during peak hours so that the congestion of the network can be avoided.

2.3.7 Call Admission Control in CDMA Systems

In a CDMA network, due to the soft-capacity feature, the admission control decision should be based on the state of the ongoing calls (e.g., interference level). The CAC approaches used in hard-capacity systems based on the assumption of time-invariant cell capacity may degrade the system utilization in a CDMA system. Also, due to the soft handoff feature, the length of a handoff process becomes longer than that of hard handoff, and the CAC algorithm must consider this duration into account.

CAC schemes based on the estimation/measurement of current state of interference and SIR were proposed in the literature [12]. While in the interference-based approaches the objective is to keep interference from all sources below the acceptable level, the SIR-based approaches emphasize the SIR requirement for each call considering the statistical factors such as voice activity, fading and shadowing in the channel.

A utility and pricing-based CAC scheme for a CDMA system was proposed in [13], in which the admission control decision is based not only on the availability of the resources, but also on the price and the corresponding level of service received. If U_k is the utility function (which depends on the received SIR) for user k and P_k is the transmission power, the total utility of the system is

$$U = \sum_k (U_k - \sigma P_k) \tag{2.9}$$

considering the fact that the transmission of the user incurs negative utility of σP_k to the system because of the interference. The negative utility here is a linear function of the transmission power P_k (σ is a constant). If the price for the code channel is ρ_c and the unit price for transmission power is ρ_p, the revenue of the system is

$$R = \sum_k (\rho_c + \rho_p P_k - \sigma P_k). \tag{2.10}$$

To maximize the system revenue, optimization techniques can be applied to obtain the optimal prices.

2.3.8 Admission Control in Wireless LANs and Wireless PANs

The general approach for call admission control in WLAN is that an estimation of the network throughput is made based on the channel access mechanism at the wireless nodes. Then a CAC decision is made heuristiscally based on the estimated throughput and the QoS requirements of the incoming call [14].

In a wireless personal area network (WPAN) such as *Bluetooth*, each piconet consists of a master device and several slave devices, and all channel access in controlled by the master (i.e., by using E-limited polling). Specifically, the master device governs the channel access by polling each slave device for data transmission. In E-limited polling, the master allows a slave device to transmit up to a maximum number of packets or until there is no packet left in the queue. With channel access based on E-limited polling and given the traffic description (e.g., mean rate), a performance model based on vacation queuing was proposed in [15]. The mean access delay at the slave device and mean cycle time of the piconet derived from the analytical model can be used to make decision on the admission control.

The main ideas of all the above CAC approaches are summarized in Table 2.1.

Table 2.1. *Different Approaches for call admission control in cellular wireless networks.*

Approach	Main Idea
Guard channel	Some portion of the wireless resources is reserved for handoff calls so that handoff call dropping probability can be maintained below the target level.
Fractional guard channel	New calls are gradually blocked according to the current status (i.e., the number of ongoing calls) of the network.
Collaborative	The neighboring cells exchange information about the network status so that the resource reservation can be made in advance accurately.
Non-collaborative	By using prediction techniques (e.g., ARMA model, Wiener filtering) to project the amount of the resources required locally so that the resources can be reserved in advance without the need for information exchange among neighboring cells
Mobility-based	Mobility information (i.e., position and direction of movement) of the mobiles can be used to enhance the accuracy of the resource reservation.
Pricing-based	Dynamic pricing is used to limit the call arrival rate so that the maximum utility and revenue of the system is achieved.

Table 2.2. *Challenges in call admission control for 4G wireless network.*

Requirements	Description
Heterogeneous environment	4G system will consist of several types of wireless access technologies, and therefore, CAC schemes must be able to handle vertical handoff and special mode of connection such as ad hoc on cellular.
Multiple types of services	4G systems need to accommodate different types of users and applications with different QoS requirements
Adaptive bandwidth allocation	With multimedia applications, system utilization and QoS performances can be improved by adjusting the bandwidth allocation depending on the state of the network and users' QoS requirements.
Cross-layer design	Both call- and packet-level QoSs need to be considered for designing CAC algorithms so that not only the call dropping and call blocking probabilities, but also the packet delay and packet dropping probabilities can be maintained at the target level.

2.4 Call Admission Control in 4G Wireless Networks

The diverse QoS requirements for multimedia applications and the presence of different wireless access technologies pose significant challenges in designing efficient CAC algorithms for 4G wireless networks. Table 2.2 summarizes some of these major challenges.

In a heterogeneous environment such as the one shown in Figure 2.4, each network, connected to each other via an Internet, will be responsible for making the decision on whether an incoming call can be accepted or not. That is, the call admission control will be performed at the edge of the network and each type of network will have its own CAC mechanism. If the call can be admitted, the required authentication, authorization, and accounting will need to be performed. Routing layer protocols

such as such as Mobile IP can be used to transfer the transport level connections from one network to another in a seamless manner.

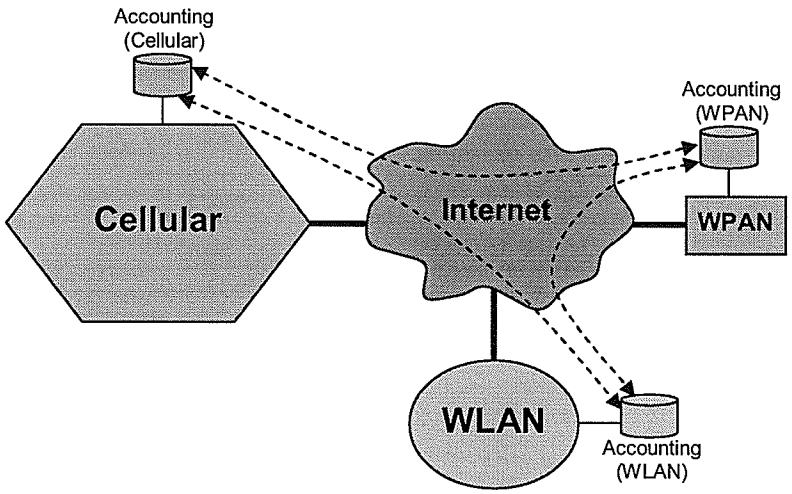


Figure 2.4. *Heterogeneous structure of a 4G wireless system.*

2.4.1 Heterogeneous Networking

Due to the seamless connection and global mobility requirements in a 4G system, a call in one particular network must be able to roam and be handed over to another network transparently. This is called vertical handoff and for this several issues need to be addressed. The usual signal-strength-based handoff initiation may not be enough, and other system parameters such as the congestion level at the networks need to be considered as well. For instance, mobile users with non-real-time applications can be handed over to WLANs in order to mitigate congestion in the cellular networks. All these factors will impact the call holding time distribution in each network.

Interoperability is also an issue in vertical handoff. The system must ensure that different types of mobile nodes and networks are able to operate with each other in a heterogeneous environment. Also, there can be ad hoc multi-hop connections both in the cellular networks and the WLANs. The ad hoc multi-hop topology can be used to relay traffic from a mobile node to the central base station or access point [16].

From the CAC point of view, vertical handoff results in new sub-type of handoff calls. A CAC algorithm must determine the priority of this type of calls over new calls. A new performance metric, namely, the vertical handoff call dropping probability, should be determined and should be maintained below the acceptable threshold. Also, the issues of call dropping and/or queuing need to be addressed. For example, if a cellular network cannot accept a call vertically handed over from WLAN, the call can be dropped or may stay connected with the WLAN and wait until the cellular network is able to accommodate the call.

2.4.2 Multiple Classes of Services and Interoperability with DiffServ-Based IP Networks

The CAC algorithms should be designed to support multiple classes of services each with specific QoS requirements. The service classes can be similar to those used in the Differentiated Services (*DiffServ*) [1] IP networks. This would enable seamless integration of wireless networks with the IP-based Internet.

DiffServ is one of the key technologies for providing QoS in the Internet. Instead of providing QoS on per-flow basis as in the Integrated Services (*IntServ*) model, *DiffServ* operates based on group of flows by aggregating several IP-level flows which have the same QoS requirements into the same group. For a packet, *DiffServ* routers can identify the group by using type of service (TOS) field in the IP packet header and also manage the traffic to satisfy the QoS requirements at the aggregate level.

In the *DiffServ* domain, there are two main types of components: edge routers and core routers. While an edge router is connected to the outside network equipment (e.g., wireless base station), a core router is connected the equipment in the same domain. Resource allocations in both these types of routers are based on service level agreement (SLA) which is issued and negotiated by the service users.

In the *DiffServ* domain three groups of traffic services are provided. When internetworking 4G wireless systems with *DiffServ*-based IP networks, the conversational and the streaming calls can be mapped into premium and assured forwarding service classes, respectively, while the non-real-time services can be mapped into the best effort service class. With this architecture, the CAC module at *DiffServ* edge router should be able to negotiate appropriate SLA with *DiffServ* domain, and the decision

on accepting or rejecting a call should be made based on both the availability of wireless resources and the negotiated SLA.

2.4.3 Adaptive Bandwidth Allocation

Due to the diversity of applications and QoS requirements for the mobile users and the dynamic nature of the wireless channel quality, adaptive bandwidth allocation (ABA) would be necessary to improve the utilization of the wireless network resources. Therefore, call admission control strategies should be designed taking this adaptive bandwidth allocation into account. With ABA, when the network conditions are favorable, the quality of a call can be upgraded by assigning more resources. However, when the network becomes congested, the amount of bandwidth allocated to some ongoing calls will be revoked to accommodate more incoming calls so that the call dropping and blocking probabilities can be maintained at the target level. In [17], such an ABA algorithm was proposed which tries to allocate a target level of bandwidth to a connection as much as possible.

Again, adaptive bandwidth allocation is needed during vertical handoff. The acceptable bandwidth should be negotiated and the CAC strategy should be based on the result of negotiation. For example, when a call handed over to a cellular networks from a WLAN, bandwidth adaptation will be required for that call.

2.4.4 Cross-Layer Design

For a wireless network, cross-layer optimization can lead to significant improvement in the transmission protocol stack performance [18]. In the context of CAC, cross-layer design principle should be applied to capture both call and the packet-level (at the radio link level) QoS performances.

Traditionally, the CAC schemes have been based on the call-level QoS measures only, although some of the CAC schemes consider the physical layer parameters such as SIR into account. However, the radio link level performance (e.g., resulting from the different scheduling and error control schemes) have not been considered while designing a CAC scheme. In contrast to a traditional voice-oriented circuit-switched network, in a purely packet-switched wireless network, the QoS will need to be de-

scribed in terms of both call-level (e.g., call blocking and call dropping probabilities) and packet-level performance metrics (e.g., packet transmission delay and packet dropping probability). Therefore, a new call should be admitted only if the "quality" of all the calls (including the incoming call) in terms of the packet-level performances can be maintained at the desired level.

2.5 CAC for 4G Networks: Architecture and Example

2.5.1 A Novel CAC Architecture

We introduce a novel CAC architecture for 4G wireless networks. The CAC module is divided into two sub-modules (i.e., two-tier CAC): one for the wireless part and the other for the wired part (Figure 2.5).

In the wireless part, the CAC needs to handle multiple classes of calls as well as calls due to vertical handoff from other types of networks (e.g., WLAN). If the call is used for data transfer, adaptive bandwidth allocation can be applied to increase resource utilization. Moreover, CAC in the wireless part must consider the nature of capacity of the systems (i.e., soft or hard) so that the resource reservation and admission control can be performed optimally.

The CAC submodule for the wired part, which internetworks with the *DiffServ* domain, is important in the sense that the dropping probability for the packets already transmitted over the air interface should be made as small as possible (to minimize wastage of wireless resources) and the packet delay should not violate the agreed SLA bound. Since the wireless resources are the scarcest resources in the system, the CAC submodule in the wired part must ensure that the wired-network can maintain the QoS of traffic from wireless users (already transmitted across the wireless links) at the desired level.

Both the call-level and the packet-level performance requirements need to be satisfied in the wireless part. Packet-level QoS performances in the wireless part can be maintained through adaptive bandwidth allocation and proper scheduling mechanisms. The call-level performances depend on the resource reservation and the admis-

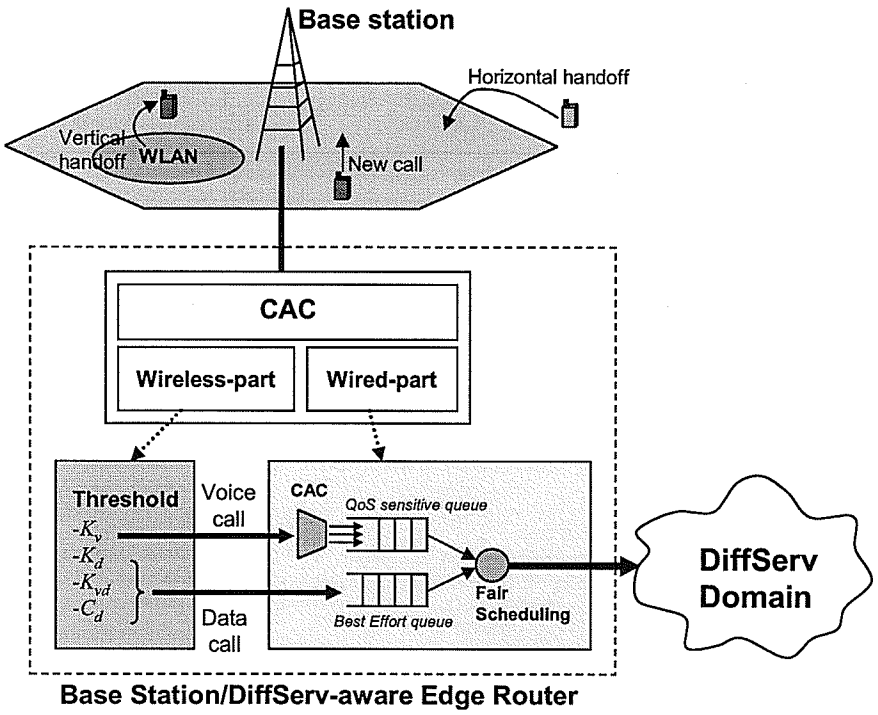


Figure 2.5. *System model for the proposed CAC scheme.*

sion control strategy in the wireless part. However, in the wired part, only packet-level QoS requirements need to be satisfied.

2.5.2 Example: A Two-Tier CAC Algorithm

Using the above architecture, we show an example of a two-tiered CAC algorithm which considers admission control at both the wireless and the wired part of the system. Threshold-based call admissions are employed to maintain call and packet-level QoS. *A call will be admitted only if it can be accepted by both of the CAC components for the air interface and the wired part of the system.* For performance analysis, design, and engineering of the system, the corresponding analytical models are developed. Typical numerical results based on the analytical modeling are also presented.

2.5.2.1 Tier-I: CAC Scheme in the Wireless Part

We assume that there are multiple types of users and adaptive bandwidth allocation is used to increase the utilization of the wireless network resources. Vertical handoff from/to other types of networks is also taken into account. A threshold-based resource reservation and admission control is used.

The base station serves two types of calls: voice and data calls, and both of these types of calls share a pool of channels. The number of channels in the cell is fixed at C (i.e., hard-capacity), and one voice call requires only one channel. For a data call, ABA is used to adjust channel allocation according to the state of the network. Under light load conditions, a data call is allocated as many channels as the user requests. Under heavy load conditions, each data call will receive at least one channel to maintain the connection.

To minimize handoff call dropping probability, the thresholds for new calls (i.e., both voice and data calls) are set at K_v and K_d, respectively. However, since data calls can be vertically handed over from other networks (e.g., WLANs), the CAC mechanism prioritizes these vertical handoff calls by using the threshold K_{vd} in which $K_d \leq K_{vd}$. Therefore, $K_{vd} - K_d$ channels are reserved particularly for horizontal and vertical handoff calls, and $C - K_v$ channels are reserved for horizontal handoff calls. With these thresholds, the priority of a horizontal handoff call is the highest, and the priority of a new call is the lowest. Again, voice calls are prioritized over data calls by limiting the number of accepted data calls when the total number of ongoing calls is equal or greater than threshold C_d ($K_d \leq K_{vd} \leq C_d$).

The ABA algorithm includes mechanisms to increase and decrease the amount of bandwidth allocated to the data calls. These mechanisms work as follows: when either a voice call or a data call arrives, if sufficient amount of resources is not available, some of the ongoing data calls (randomly chosen) are downgraded to give the required amount of resources to the incoming call. Similarly, when the call departs, the bandwidth freed will be randomly assigned to upgrade the ongoing data calls. However, if there is no ongoing data call which can be downgraded, the incoming call is rejected.

Under the above assumptions on multiple types of calls, ABA, and vertical handoff, the system can be modeled by using a continuous Markov chain. The arrivals of the

new voice calls, handoff voice calls, new data calls, horizontal and vertical handoff data calls are assumed to follow a Poisson process and the corresponding average rates (calls per minute) are denoted by $\lambda_n^{(v)}$, $\lambda_h^{(v)}$, $\lambda_n^{(d)}$, $\lambda_{hh}^{(d)}$, $\lambda_{vh}^{(d)}$, respectively. We assume that the call holding time for both voice and data calls are exponentially distributed and the mean values are $1/\mu_v$ and $1/\mu_d$, respectively. The state space of the system is

$$\Phi = \{(\mathcal{V}, \mathcal{D}),\ 0 \leq \mathcal{V} \leq C, 0 \leq \mathcal{D} \leq C_d\}. \tag{2.11}$$

where $\mathcal{V}$ and $\mathcal{D}$ are the number of ongoing voice and data calls, respectively.

The call-level performances of the system can be determined from the steady state probabilities. From this model, we are able to obtain new call blocking and handoff call dropping probabilities for both voice and data calls as well as the vertical handoff call dropping probabilities for data calls.

To evaluate the performance, we assume that there are 40 channels in a cell and the average call holding times for voice and data calls are 5 and 10 minutes, respectively. We set the values for the different thresholds as follows: $K_v = 36$, $C_d = 36$, $K_{vd} = 34$, and $K_d = 32$. Typical variations in new voice call blocking probability under different new and handoff voice call arrival rates are presented in Figure 2.6. As expected, this blocking probability increases as the arrival rates increase.

Figure 2.7 shows typical variations in the vertical handoff call dropping probability (for data calls) under different data call arrival rates. This dropping probability also increases with increasing arrival rate, however at a slower rate compared to that for a voice call because data call arrival rate is smaller for voice calls.

Figure 2.8 shows typical variations in the average bandwidth allocation for data calls under the different voice call arrival rates. This result shows the effect of prioritizing voice calls over data calls. Specifically, when the rate of arrival of voice calls increases, the ongoing data calls need to be degraded to accommodate incoming voice calls.

2.5.2.2 Tier-II: CAC Scheme in the Wired Part

As shown in Figure 2.5, the *DiffServ*-aware edge router has two transmission queues: QoS queue and best-effort (BE) queue, with size U and V packets, respectively. The QoS queue is used for voice packets while the best-effort queue is used for data packets.

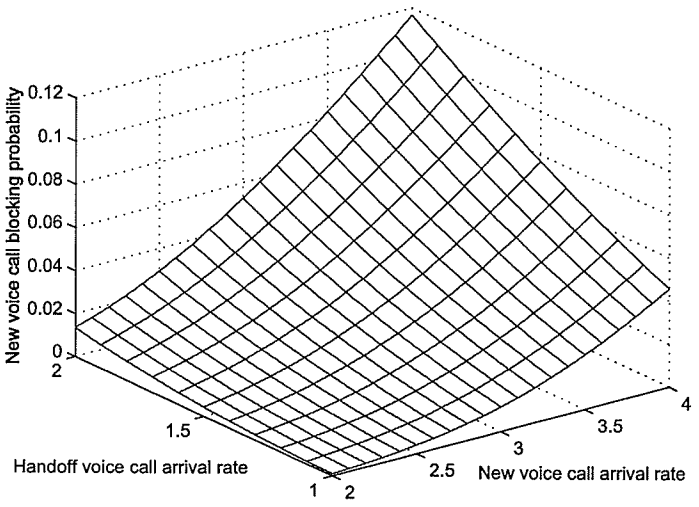


Figure 2.6. *New voice call blocking probaiblity under different new and handoff voice call arrival rates.*

A CAC mechanism is applied at the QoS queue to guarantee packet-level QoS.

We assume that the router serves the queues in a time-division multiplexing fashion using fixed-size time slot and only one packet is transmitted during one time slot. The router uses a fair scheduling mechanism which is based on the packetized version of the generalized processor sharing (GPS) [2]. With this type of traffic scheduling, the fairness is maintained in the sense that the packets in one queue will not affect the performance of those in the other queue beyond minimum performance guarantee. The amount of service for queue i, $S_i(t_1, t_2)$ ($i \in \{QoS, BE\}$) in time $[t_1, t_2)$ is governed by the weight ϕ_i, and for the two queues in Figure 2.5, if both the queues are backlogged during period $[t_1, t_2)$, the following property is maintained:

$$\frac{S_{QoS}(t_1, t_2)}{S_{BE}(t_1, t_2)} = \frac{\phi_{QoS}}{\phi_{BE}} \tag{2.12}$$

where ϕ_{QoS} and ϕ_{BE} are the weights for the QoS and best effort queues in the model. With the work conserving property of the above fair scheduling, if one queue is not backlogged, the available service will be allotted to the other queue. The state space

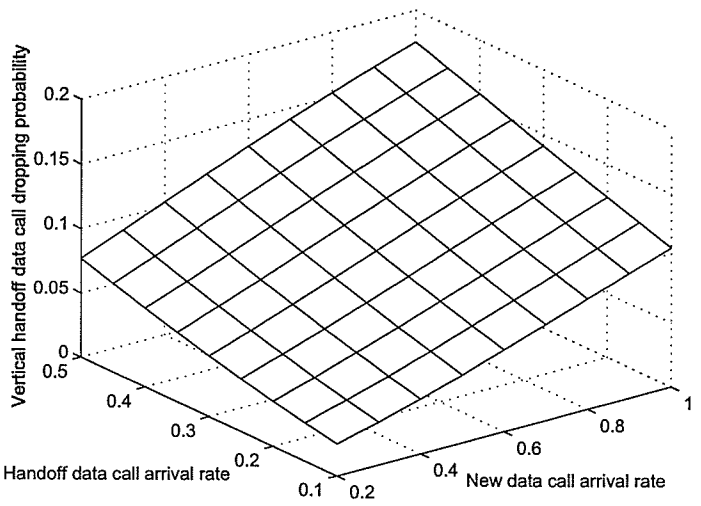


Figure 2.7. *Vertical handoff data call dropping probability under various new and handoff data call arrival rates.*

of this system can be expressed as follows:

$$[\Xi = \{(\mathcal{X}, \mathcal{Y}, \mathcal{Z}),\ 0 \leq \mathcal{X} \leq U, 0 \leq \mathcal{Y} \leq V, 0 \leq \mathcal{Z} \leq T\} \tag{2.13}$$

where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ represent the number of packets in the QoS queue, the number of packets in the best effort queue, and the number of calls admitted to the QoS queue, respectively.

A threshold-based CAC mechanism is used to ensure that for the QoS traffic the packet-level performance measures (e.g., packet dropping probability and average delay) are maintained below the acceptable level. The threshold T limits the number of calls to the QoS queue. Specifically, when a call arrives, the CAC algorithm checks whether the number of ongoing calls is less than the threshold. If so, the new call is admitted, otherwise the call is rejected. *This threshold can be set according to the congestion condition in* DiffServ *domain.*

We assume that call arrival follows a Poisson process with mean λ and the call holding time is exponentially distributed with mean $1/\mu$. For each flow, the packet arrival follows a Bernoulli process with the probability of arrival of one packet in a time slot being a_{QoS}. We assume that the probability a_{QoS} is the same for all QoS

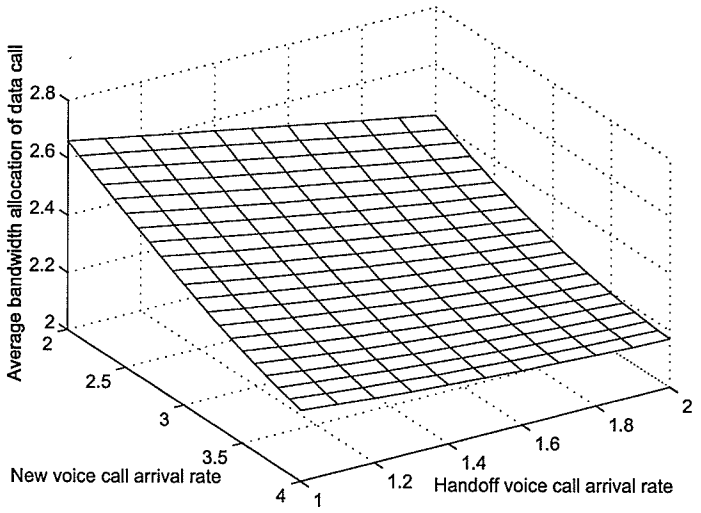


Figure 2.8. *Average bandwidth of data call under different new and handoff call arrival rates.*

sensitive flows. For the best-effort queue, the probability that i packets arrive in one time slot is denoted by a_{BE}^{i}.

We obtain the packet-level QoS measures from the model. The length of time slot is assumed to be 10^{-3} second and the weights for the queues are $\phi_{QoS} = 0.7$ and $\phi_{BE} = 0.3$. We vary the probability of arrival of one packet in the best-effort queue a_{BE}^{1} and obtain the performance results. Figure 2.9 shows typical variations in the mean waiting time W_{QoS} when the buffer size for each of the QoS and best effort queues is 20 packets. For the QoS queue, the packet arrival probability of voice call is fixed at 0.2 (i.e., $a_{QoS} = 0.2$) in this case.

We observe that W_{QoS} increases with increasing a_{BE}^{1} up to a certain point after which W_{QoS} becomes constant. When the traffic in the best effort queue grows, the QoS queue receives less amount of service. Until the best effort queue uses all its allocated service, the waiting time in the QoS queue does not increase because packets in this queue still receive at least the guaranteed amount of service determined by ϕ_{QoS}.

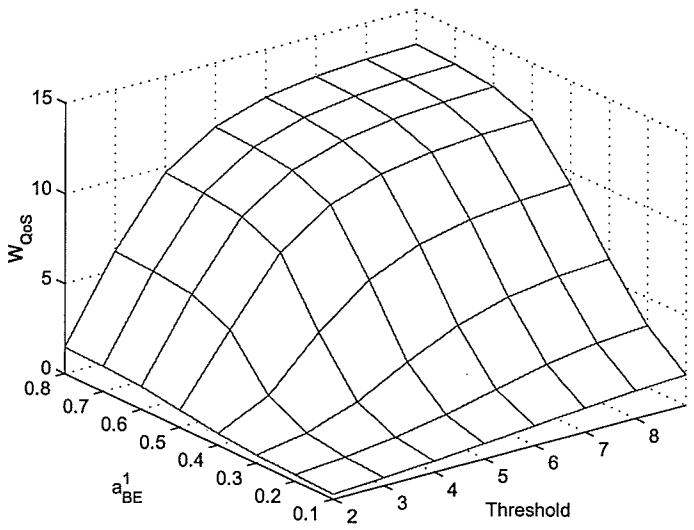


Figure 2.9. *Average waiting time for the packets in the QoS queue.*

2.6 Chapter Summary

We have presented a comprehensive survey on the issues and approaches to call admission control in the next-generation (e.g., 4G) wireless networks. Starting with the general model and classifications of the CAC strategies, different CAC schemes proposed in the literature (for cellular wireless, WLAN, WPAN) have been reviewed and the challenges in designing efficient CAC schemes for 4G systems have been outlined.

To this end, we have introduced a two-tiered CAC architecture for 4G networks to ensure QoS in both the wireless and the wired parts. In the general architecture, the CAC decision should be based on both the call-level and the packet-level performance metrics into account. We have given an example of a two-tier CAC scheme considering two types of services (voice and data) based on this architecture. Data calls due to vertical handoff from other types of networks have been also considered.

In this CAC scheme, for the wireless part a threshold-based mechanism is used to provide higher priority to voice calls over data calls and adaptive bandwidth allocation is used to increase utilization of channel resources by minimizing call blocking and call dropping probabilities. The CAC component in the wired part uses a threshold-based

policy for admission control to ensure packet-level QoS for the voice traffic through a fair scheduling mechanism. The threshold can be adjusted to accommodate different packet-level QoS requirements. Analytical models for performance evaluation of the proposed CAC scheme have been outlined and typical numerical results have been presented. Although in the example CAC scheme we have not explicitly considered the packet-level QoS in the wireless part of the system, a model similar to the one used in the wired part can be used for this purpose.

The following issues on call admission control for QoS provisioning in the future-generation IP-based wireless mobile networks need to be addressed:

- Consideration of traffic shaping policies (e.g., using token-bucket shaper) at the mobile nodes and optimizing the CAC policies accordingly. Specifically, considering the bursty nature of the data traffic, for given traffic shaping parameters, dynamic bandwidth allocation and CAC policies can be devised (e.g., based on intelligent traffic prediction techniques) so that the wireless resource utilization can be maximized while satisfying the required QoS assurance.
- Pricing models for CAC in heterogeneous wireless access networks with a view to maximizing the total revenue in the network while providing acceptable level of QoS to the users.
- The end-to-end QoS should be considered for CAC. In our model, some feedback- or measurement-based scheme can be used to adjust the threshold in the CAC submodule in wired part to take the network condition of the *DiffServ* domain into account.
- Analytical models for performance evaluation of CAC schemes under realistic traffic patterns.

Chapter 3

Service Differentiation in Wireless Networks: A Unified Analysis

3.1 Introduction

In addition to supporting best-effort data services, next-generation broadband wireless networks will need to provide quality-of-service (QoS) guarantee for multimedia users. In essence, traffic scheduling is the major component to provide service differentiation between QoS-sensitive and best-effort traffic in the wireless access networks. Several fair traffic scheduling disciplines based on generalized processor sharing [2] were proposed in the literature [19]-[20]. Along with traffic scheduling, a CAC mechanism is also necessary to control the number of accepted connections to guarantee that the performance requirements for all the admitted connections in the network can be met. Again, adaptive multi-rate transmission (e.g., through adaptive modulation and coding (AMC)) according to the wireless channel variations at the physical layer would be a key feature of broadband wireless systems such as IEEE 802.16. Analytical model for performance evaluation (at both the packet-level and the connection-level) and engineering of broadband wireless networks is necessary which can incorporate all these features into account in a unified way.

Packet-level delay performance in a polling-based wireless access network was analyzed in [21]. In [22], several types of CAC policies for mobile networks were presented and analyzed. We observe that most of the works in the literature considered either packet-level QoS or call-level QoS. Although the work in [23] proposed a model for both call and packet-level performance analysis, but the authors did not take traffic scheduling into account. In particular, packet-level performances depend not only

on the radio link level transmission and error control mechanisms, but also on the resource sharing mechanism among multiple users.

A resource allocation strategy, namely, enhanced staggered resource allocation (ESRA) method, was proposed in [24] the objective of which was to maximize the number of concurrent transmissions, so that the throughput can be maximized. However, the buffer dynamics at the radio link level queue (and hence the queueing performance) was not analyzed. Effects of multi-rate transmission (achieved through adaptive modulation and coding) on radio link level queueing performance were analyzed in [25] for a single-user scenario (i.e., without scheduling).

In this chapter, we present a unified queueing analytical framework for a service differentiation model in a wireless access network with scheduling and CAC. The model considers two types of traffic, namely, QoS-sensitive traffic for real-time multimedia applications and best-effort (BE) traffic for applications such as web and e-mail. Two separate queues are used to accommodate the aggregated traffic from the QoS-sensitive and best-effort flows. This configuration is compatible with the *DiffServ* [1] model for service differentiation in which one queue is used for QoS-sensitive flows (expedited forwarding, EF) and another one is used for best-effort flows (assured forwarding, AF). A work-conserving traffic scheduling scheme based on packetized generalized processor sharing (PGPS) is used to allocate the channel resources between the QoS-sensitive and the best-effort flows in a time-slotted wireless transmission scenario. To satisfy the performance requirements for the QoS-sensitive flows, a threshold-based CAC scheme is used to limit the number of connection of the QoS-sensitive queue while no admission control is used for the best-effort queue.

Moreover, the multi-rate transmission feature in the physical layer, which can be achieved through adaptive modulation and coding (e.g., in IEEE 802.16 systems), is also taken into account. Note that, the time-slotted transmission in this model is consistent with the TDMA (Time Division Multiple Access)/TDD (Time Division Duplex) mode in IEEE 802.16. In addition, the service differentiation model to support both QoS-sensitive and best-effort traffic can be applied for provisioning of both polling and best-effort services in IEEE 802.16.

From the analytical framework, the various QoS measures such as connection blocking probability for the QoS-sensitive users, and packet dropping probability,

average queue length, and packet delay distribution can be obtained for both the QoS-sensitive and best-effort queues. In particular, the inter-dependencies among the connection-level and the packet-level performance measures can be analyzed and the parameters for the scheduling and CAC can be determined according to the QoS requirements.

3.2 System Model and Assumptions

3.2.1 System Components

We consider a system with two transmission queues: QoS-sensitive queue and best-effort (BE) queue. Fair scheduling is used to differentiate services between these two queues. A threshold-based CAC mechanism is employed for the QoS-sensitive queue to limit the number of admitted connections so that the performances for this traffic type can be maintained at the desired level (Figure 8.1). Since performance guarantee is not necessary for best-effort traffic, no CAC is applied to the best-effort queue. The length of the QoS-sensitive queue is assumed to be finite, while the size of the best-effort queue can be either finite or infinite, depending on the deployment scenario.

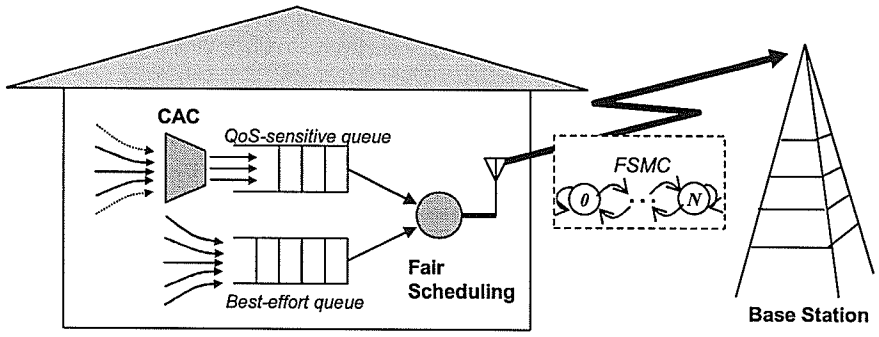


Figure 3.1. *System model.*

In the physical layer, adaptive modulation and coding (AMC) is used to enhance the transmission rate according to channel state information (CSI), and we utilize

a finite state Markov chain (FSMC) channel model for the different transmission modes of AMC. The transmission time is slotted and a time-division multiplexing (TDM)-based channel access is assumed where a batch of packets can be transmitted during one time slot depending on the channel condition (and hence the transmission mode). Infinite persistent automatic repeat request (ARQ) is used to ensure reliable data transmission.

3.2.2 Wireless Channel Model and Multi-rate Transmission

A finite state Markov chain (FSMC) model is used to represent multiple states of a slow Nakagami-m fading channel where each state corresponds to one transmission mode of AMC. With an N state FSMC, the signal to noise ratio (SNR) at the receiver γ can be partitioned into $N+1$ non-overlapping intervals by thresholds Γ_n ($n \in \{0, 1, \ldots, N\}$) where $\Gamma_0 = 0 < \Gamma_1 < \ldots < \Gamma_{N+1} = \infty$. The channel is said to be in state n if $\Gamma_n \leq \gamma < \Gamma_{n+1}$, and in this state n bits are transmitted per symbol using 2^n-QAM which corresponds to transmission rate n. To avoid possible transmission error, no packet is transmitted when $n = 0$.

Assuming that the channel is slowly fading (i.e., transitions occur only between adjacent states), the state transition matrix for the FSMC can be expressed as follows:

$$\mathbf{P}_c = \begin{bmatrix} \zeta_{0,0} & \zeta_{0,1} & \cdots & 0 \\ \zeta_{1,0} & \zeta_{1,1} & \zeta_{1,2} & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & \zeta_{N-1,N-2} & \zeta_{N-1,N-1} & \zeta_{N-1,N} \\ 0 & \cdots & \zeta_{N,N-1} & \zeta_{N,N} \end{bmatrix}. \tag{3.1}$$

The transition probability from state n to n' ($n' \in \{n-1, n, n+1\}$) $\zeta_{n,n'}$, average SNR and average packet error rate ($\overline{PER}_n$) can be obtained as in [25]. Based on the packet error rate, assuming an independent packet error process, the probability that m out of n packets are successfully transmitted in one time slot can be obtained as follows:

$$\theta_{m,n} = \binom{n}{m} \theta^m (1-\theta)^{n-m}, \quad \theta = 1 - \overline{PER}_n. \tag{3.2}$$

3.2.3 Fair Scheduling

For the QoS-sensitive and the BE queues, we consider a fair scheduling mechanism which is based on the packetized version of the generalized processor sharing (GPS) [2]. With this type of traffic scheduling, the service for each queue is governed by the corresponding weight, and if both the queues are backlogged during period $[t_1, t_2)$, then $\frac{S_q(t_1,t_2)}{S_b(t_1,t_2)} = \frac{\phi_q}{\phi_b}$, where ϕ_q and ϕ_b denote scheduler weights for the QoS and the best-effort queue, respectively.

Due to the work conserving property of fair scheduling, if the number of packets in one queue is zero, the another queue will receive the entire service. On the other hand, if both queues are backlogged, each of them will receive service according to its weight. The scheduling in our model is based on temporal fairness. Specifically, at steady state if every flow is backlogged, the probability that the time slot is allocated to a particular flow is determined by its weight.

3.3 Queueing Analytical Model

3.3.1 CAC for the QoS-Sensitive Queue

When a new connection arrives, the CAC algorithm checks whether the number of ongoing connections is less than the predefined threshold T. If it is true, then the arriving connection is admitted, otherwise it is rejected. We assume that connection arrival follows a Poisson process [22] with mean ρ and the connection holding time is exponentially distributed with mean $1/\mu$.

With mean rate ρ, the probability that a Poisson events occur in time interval t is given by

$$f_a(\rho) = \frac{e^{-\rho t}(\rho t)^a}{a!}. \tag{3.3}$$

If the length of the interval t (i.e., a time slot) is very small compared to the connection arrival rate and holding time, we can assume that there is at most one connection arrival and departure in one time slot. Then, the probability transition matrix for the CAC process is given by (3.4) where the elements inside matrix $\mathbf{C}$ expressed can

be expressed by (3.5)-(3.6).

$$\mathbf{C} = \begin{bmatrix} c_{0,0} & c_{0,1} & & & & \\ c_{0,1} & c_{1,1} & c_{1,2} & & & \\ \ddots & \ddots & \ddots & & & \\ & c_{i-1,i} & c_{i,i} & c_{i,i+1} & & \\ & \ddots & \ddots & \ddots & & \\ & & c_{T-2,T-1} & c_{T-1,T-1} & c_{T-1,T} & \\ & & & c_{T-1,T} & c_{T,T} \end{bmatrix}. \tag{3.4}$$

$$c_{i,i+1} = \begin{cases} F_1(\rho), & i = 0 \\ F_1(\rho) f_0(i\mu) & 0 \le i \le T-1. \end{cases} \quad c_{i,i-1} = \begin{cases} f_0(\rho) F_1(i\mu), & 1 \le i \le T-1 \\ F_1(i\mu), & i = T. \end{cases} \tag{3.5}$$

$$c_{i,i} = \begin{cases} f_0(\rho), & i = 0 \\ F_1(\rho) F_1(i\mu) + f_0(\rho) f_0(i\mu), & 0 \le i \le T-1 \\ f_0(i\mu), & i = T. \end{cases} \tag{3.6}$$

Here, $F_a(\rho) = \sum_{j=a}^{\infty} f_j(\rho)$ denotes the complementary distribution for the number of connection arrivals. Note that, this model is a special type of $G/G/c$ queue in which the customer departure process ($c_{i,i-1}$) is determined by the number of ongoing connections ($i\mu$).

The steady state probability for the number of ongoing connections can be obtained by solving $\boldsymbol{\pi}_q \mathbf{C} = \boldsymbol{\pi}_q$ and $\boldsymbol{\pi}_q \mathbf{1} = 1$, where $\mathbf{1}$ is a column matrix of ones, $\boldsymbol{\pi}_q$ is row matrix of steady state probability in which each column (i.e., $[\boldsymbol{\pi}_q]_{c+1}$) corresponds to the number of connections c in the system ($c = 0, 1, \ldots, T$). Note that, $[\boldsymbol{\pi}_q]_i$ indicates the element at column i of row matrix $\boldsymbol{\pi}_q$.

The average number of connections in the system ($\overline{c}$) and average packet arrival rate in the QoS queue[1] ($\overline{\lambda}_q$) per time slot are calculated as follows:

$$\overline{c} = \sum_{c=1}^{T} c \times [\boldsymbol{\pi}_q]_{c+1}, \quad \overline{\lambda}_q = \sum_{c=1}^{T} \left(\lambda_q \times c \times [\boldsymbol{\pi}_q]_{c+1} \right) \tag{3.7}$$

where λ_q is the average number of packet arrivals per time slot per connection which we assume to be identical for all connections. This assumption is valid if the same traffic shaper (e.g., leaky bucket) is used for all the connections. The connection blocking probability (i.e., the probability that the system has T ongoing connections) can be obtained as $P_{bl} = [\boldsymbol{\pi}_q]_{T+1}$.

[1]We use the terms 'QoS-sensitive queue' and 'QoS queue' interchangeably in this chapter.

3.3.2 Fair Scheduling and CAC

We model the fair scheduling process with FSMC channel model by using a discrete-time Markov chain. The state space of the system is

$$\Delta = \{(Y, X, C),\ 0 \leq Y \leq B_b, 0 \leq X \leq B_q, 0 \leq C \leq N\} \tag{3.8}$$

where X, Y and C represent the number of packets in the QoS queue, the number of packets in the best-effort queue, and the channel state, respectively. Since for real-time traffic the delay has to be bounded, we assume that the size of the QoS-sensitive queue is finite with B_q packets. However, the size of the best-effort queue B_b can be either infinite or finite depending on the system and service requirements and we model both the cases in this chapter.

3.3.2.1 Model for the QoS Queue

While the packet arrival probability is obtained from truncated Poisson process with maximum A packets can arrive in one time slot (i.e., A is obtained such that $F_A(\lambda) < 10^{-4}$), the departure process depends on the modulation level and the packet error rate. With queue size B_q, the probability transition matrix $\mathbf{Q}_y$ for the QoS queue, when the number of packets in the best-effort queue is y, is defined as follows:

$$\mathbf{Q}_y = \begin{bmatrix}
\mathbf{q}_{0,0}^{(y)} & \mathbf{q}_{0,1}^{(y)} & \cdots & \mathbf{q}_{0,A}^{(y)} & & & & & \\
\mathbf{q}_{1,0}^{(y)} & \mathbf{q}_{1,1}^{(y)} & \mathbf{q}_{1,2}^{(y)} & \cdots & \mathbf{q}_{1,A}^{(y)} & & & & \\
\vdots & \vdots & \ddots & \ddots & \ddots & & & & \\
\hline
\mathbf{q}_{N,0}^{(y)} & \cdots & \mathbf{q}_{N,N-1}^{(y)} & \mathbf{q}_{N,N}^{(y)} & \mathbf{q}_{N,N+1}^{(y)} & \cdots & \mathbf{q}_{B,A}^{(y)} & & \\
 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & \mathbf{q}_{x,x-N}^{(y)} & \cdots & \mathbf{q}_{x,x-1}^{(y)} & \mathbf{q}_{x,x}^{(y)} & \mathbf{q}_{x,x+1}^{(y)} & \cdots & \mathbf{q}_{x,A}^{(y)} \\
\hline
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & & \mathbf{q}_{B_q,B_q-N}^{(y)} & \cdots & \mathbf{q}_{B_q,B_q-1}^{(y)} & \mathbf{q}_{B_q,B_q}^{(y)}
\end{bmatrix} \tag{3.9}$$

where the element $\mathbf{q}_{x,x'}^{(y)}$ of matrix $\mathbf{Q}_y$ is the probability that the number of packets in the QoS queue is x in the current time slot and it becomes x' in the next time slot.

Let $\mathbf{D}_m^{(y)}$ be the probability matrix corresponding to successful transmission of m packets from the QoS queue when there are y packets in best-effort queue ($m, n \in \{0, 1, \ldots, N\}$), and this matrix can be obtained as follows:

$$\left[\mathbf{D}_m^{(y)}\right]_{n+1,n+1} = \theta_{m,n}, \qquad y = 0 \tag{3.10}$$

$$\left[\mathbf{D}_m^{(y)}\right]_{n+1,n+1} = \theta_{m,M_{(q)}(n)}, \quad y > 0 \text{ for } m = 0, \ldots, M_{(q)}(n) \tag{3.11}$$

where $\theta_{m,n}$ is obtained from (3.2). Note that, the matrix $\mathbf{D}_m^{(y)}$ is of the same size as the channel state transition $\mathbf{P}_c$. $M_{(q)}(n)$ is the function that return the number of packets that QoS sensitive queue can transmit in channel state n, and $M_{(b)}(n)$ denotes the function that returns the number of packets that best-effort can transmit if both queues are backlogged. These functions are chosen such that on average both queues receive transmission rate according to their weights and therefore we can formulate optimization problem as follows:

$$\text{Minimize:} \quad \left\| \frac{\sum_{n=0}^{N} \left(M_{(q)}(n) \left[\boldsymbol{\pi}_c\right]_{n+1} \right)}{\sum_{n=1}^{N} n \left[\boldsymbol{\pi}_c\right]_{n+1}} - \phi_q \right\| \tag{3.12}$$

$$\text{Subject to:} \quad M_{(q)}(n) \in \{0, 1, \ldots, N\} \tag{3.13}$$

$$M_{(b)}(n) = n - M_{(q)}(n) \tag{3.14}$$

where $\boldsymbol{\pi}_c$ is obtained from solving $\mathbf{P}_c\boldsymbol{\pi}_c = \boldsymbol{\pi}_c$ and $\boldsymbol{\pi}_c\mathbf{e} = 1$. This optimization can be simply solved by enumeration method.

Now, the elements in the first and second parts of matrix $\mathbf{Q}_y$ can be obtained as follows:

$$\mathbf{q}_{x,x-r}^{(y)} = \mathbf{P}_c \sum_{m-a=r} \left(f_a(\overline{\lambda}_q)\mathbf{I} \right) \mathbf{D}_m^{(y)} \quad \text{for } r = 1, 2, \ldots, R \tag{3.15}$$

$$\mathbf{q}_{x,x+s}^{(y)} = \mathbf{P}_c \sum_{a-m=s} \left(f_a(\overline{\lambda}_q)\mathbf{I} \right) \mathbf{D}_m^{(y)} \quad \text{for } s = 1, 2, \ldots, A \tag{3.16}$$

$$\mathbf{q}_{x,x}^{(y)} = \mathbf{P}_c \sum_{a=m} \left(f_a(\overline{\lambda}_q)\mathbf{I} \right) \mathbf{D}_m^{(y)} \tag{3.17}$$

where $m \in \{0, 1, 2, \ldots, R\}$ and $a \in \{0, 1, 2, \ldots, A\}$ represent the number of packets transmitted and the number of packet arrivals, respectively, $\mathbf{I}$ is the identity matrix which has the same size as $\mathbf{P}_c$.

Considering both the packet arrival and the packet departure events, (3.15), (3.16), and (3.17) above represent the transition probability matrices for the cases when the number of packets in queue decreases by r packets, increases by s packets, and do not change, respectively. Note that, the maximum total packet transmission rate can be greater than the number of packets in queue, and the decrease in the number of packets cannot be less than the number of packets in the queue. Therefore, the maximum number by which the number of packets in the queue can decrease is $R = \min(N, x)$.

The third part of matrix $\mathbf{Q}_y$ ($\{x = B_q - A + 1, B_q - A + 2, \ldots, B_q\}$) has to capture the packet dropping effect due to buffer unavailability. Let $\hat{\mathbf{q}}^{(y)}_{x,x+i}$ denote the probability transition in the matrix $\mathbf{Q}_y$ in the case that there is no dropped packets, (3.15) becomes

$$\mathbf{q}^{(y)}_{x,x+s} = \begin{cases} \sum_{i=s}^{A} \hat{\mathbf{q}}^{(y)}_{x,x+i}, & x+s = B_q \\ \mathbf{0}, & x+s > B_q \end{cases} \tag{3.18}$$

for $x = B_q$, for $x + s \geq B_q$ and (3.17) becomes

$$\mathbf{q}^{(y)}_{x,x} = \hat{\mathbf{q}}^{(y)}_{x-1,x-1} + \sum_{i=1}^{A} \hat{\mathbf{q}}^{(y)}_{x,x+i} \quad \text{for } x = B_q. \tag{3.19}$$

Eqs. (3.18) and (3.19) indicate the case that the queue will be full if the number of incoming packets is greater than the available space in the queue. In other words, the transition probability to the state that the queue is full can be calculated as the sum of all the probabilities that make the number of packets in the queue equal to or larger than the queue size B_q.

3.3.2.2 Model for the Best-Effort Queue

We assume that packet arrival probability of best-effort traffic is obtained from truncated Poisson process in which the maximum number of packets that can arrive at the best-effort queue in one time slot is A (i.e., A is obtained such that $F_A(\lambda) < 10^{-4}$). In particular, the probability of arrival of a packets ($0 \leq a \leq A$) can be obtained from (3.3). Note that, unlike the QoS queue, there is no CAC for the BE queue.

The transition matrix for the entire system $\mathbf{P}$ with infinite queue size can be expressed as

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_{0,0} & \mathbf{p}_{0,1} & \cdots & \mathbf{p}_{0,A} & & & & & \\ \mathbf{p}_{1,0} & \mathbf{p}_{1,1} & \mathbf{p}_{1,2} & \cdots & \mathbf{p}_{1,A} & & & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & & & & \\ \mathbf{p}_{N,0} & \cdots & \mathbf{p}_{N,N-1} & \mathbf{p}_{N,N} & \mathbf{p}_{N,N+1} & \cdots & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & \mathbf{p}_{y,y-N} & \cdots & \mathbf{p}_{y,y-1} & \mathbf{p}_{y,y} & \mathbf{p}_{y,y+1} & \cdots & \mathbf{p}_{y,A} \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \tag{3.20}$$

where the element $\mathbf{p}_{y,y'}$ indicates that there are y packets in best-effort queue in this time slot and it becomes y' in the next time slot. These elements are calculated based on the model for the QoS queue.

Similar to the model for the QoS queue, let $\mathbf{E}_m$ be the probability matrix corresponding to successful transmission of m packets from the best-effort queue ($m = 0, 1, \ldots, N$ and $x = 0, 1, \ldots, B_q$), and it can be obtained as follows:

$$[\mathbf{E}_m]_{x(N+1)+n+1,x(N+1)+n+1} = \theta_{m,n}, \qquad x = 0 \tag{3.21}$$

$$[\mathbf{E}_m]_{x(N+1)+n+1,x(N+1)+n+1} = \theta_{m,M_{(b)}(n)}, \quad x > 0 \text{ for } m = 0, \ldots, M_{(b)}(n) \tag{3.22}$$

This matrix $\mathbf{E}_m$ is of the same size as transition matrix $\mathbf{Q}_y$.

The elements in matrix $\mathbf{P}$ can be obtained as follows:

$$\mathbf{p}_{y,y-r} = \mathbf{Q}_y \sum_{m-a=r} (f_a(\lambda_b)\mathbf{I})\, \mathbf{E}_m \quad \text{for } r = 1, 2, \ldots, R \tag{3.23}$$

$$\mathbf{p}_{y,y+s} = \mathbf{Q}_y \sum_{a-m=s} (f_a(\lambda_b)\mathbf{I})\, \mathbf{E}_m \quad \text{for } s = 1, 2, \ldots, A \tag{3.24}$$

$$\mathbf{p}_{y,y} = \mathbf{Q}_y \sum_{a=m} (f_a(\lambda_b)\mathbf{I})\, \mathbf{E}_m \tag{3.25}$$

where $m \in \{0, 1, 2, \ldots, R\}$ and $a \in \{0, 1, 2, \ldots, A\}$ represent the number of packets transmitted and the number of packet arrivals, respectively. The maximum number of successfully transmitted packets from the best-effort queue is $R = \min(N, y)$.

Eqs. (3.23), (3.24), and (3.25) represent the transition probability matrices for the cases when the number of packets in best-effort queue decreases by r packets, increases by s packets, and do not change, respectively. Note that, all of these matrices incorporate all possible combinations of transitions in channel states and the number of packets in both the QoS and the best-effort queues. However, if the best-effort queue is finite (i.e., $B_b < \infty$), the bottom part of matrix $\mathbf{P}$ (i.e., row $B_b - A + 1$ to B_b) needs to capture the packet dropping effect. In this case, it can be obtained in the same way as that for $\mathbf{Q}_y$ which resulted in eqs. (3.18) and (3.19).

3.3.3 Steady State Probability

To evaluate the QoS performance measures, the steady state probabilities for the system states are required. If the size of the best-effort queue is finite, the steady state probability of the system $\boldsymbol{\pi}$ can be simply obtained by solving $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{1} = 1$. The steady state probability $\pi(n, x, y)$ that the channel is in state n and

that there are x and y packets in the QoS and the best-effort queues, respectively, can be extracted directly from matrix $\boldsymbol{\pi}$ as follows:

$$\begin{aligned} \pi(n,x,y) &= [\boldsymbol{\pi}]_{col(n,x,y)}, \quad \text{where} && (3.26) \\ col(n,x,y) &= (y \times (B_q+1) \times (N+1)) + (x \times (N+1)) + n + 1. && (3.27) \end{aligned}$$

In the case that the size of the best-effort queue is infinite, we apply *matrix-geometric* method [28] to obtain the steady state probabilities. For this, we re-block the matrix $\mathbf{P}$ to obtain the transition probability matrix in form as shown in Figure 3.2.

When the stability condition, namely, $\boldsymbol{\delta}\mathbf{Z}_2\mathbf{1} > \boldsymbol{\delta}\mathbf{Z}_0\mathbf{1}$, where $\boldsymbol{\delta} = \boldsymbol{\delta}\mathbf{Z}$, $\boldsymbol{\delta}\mathbf{1} = 1$, and $\mathbf{Z} = \mathbf{Z}_0 + \mathbf{Z}_1 + \mathbf{Z}_2$ is satisfied, then the matrix $\mathbf{R}$, which is the minimal non-negative solution of $\mathbf{R} = \mathbf{Z}_0 + \mathbf{R}\mathbf{Z}_1 + \mathbf{R}^2\mathbf{Z}_2$ can be determined such that $\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_i\mathbf{R}$. This matrix $\mathbf{R}$ can be obtained iteratively from $\mathbf{R}(k+1) = \mathbf{Z}_0 + \mathbf{R}(k)\mathbf{Z}_1 + \mathbf{R}^2(k)\mathbf{Z}_2$ until $|\mathbf{R}(k+1) - \mathbf{R}(k)|_{i,j} < \epsilon$, $\forall i,j$ (e.g., $\epsilon = 10^{-9}$). Next, we calculate $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ by solving following equations:

$$\mathbf{B}[\mathbf{R}] = \begin{bmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{W} & \mathbf{Z}_1 + \mathbf{R}\mathbf{Z}_2 \end{bmatrix}, \quad [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1] = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1]\,\mathbf{B}[\mathbf{R}], \quad \boldsymbol{\pi}_0\mathbf{1} + \boldsymbol{\pi}_1(\mathbf{I} - \mathbf{R})^{-1}\mathbf{1} = 1. \tag{3.28}$$

Since $\boldsymbol{\pi}_i$ consists of $A-1$ states of different number of packets in the best-effort queue, the steady state probability $\pi(n,x,y)$ can be extracted as follows:

$$\begin{aligned} \pi(n,x,y) &= \left[\boldsymbol{\pi}_{i(x)}\right]_{col(n,x,y)}, \quad \text{where} \\ col(n,x,y) &= (i(x) \times (A-1) \times (B_q+1) \times (N+1)) + \\ &\quad (y \times (B_q+1) \times (N+1)) + (x \times (N+1)) + n + 1 \\ i(x) &= \left\lfloor \frac{x}{A} \right\rfloor && (3.29) \end{aligned}$$

where $i(x)$ determines the block for calculating steady state probability for x packets in QoS queue.

3.3.4 QoS Measures

Packet-level QoS measures including the average number of packets in queue (i.e., average queue length), packet dropping probability, average delay, and delay distribution can be obtained based on the steady state probability $\pi(n,x,y)$. However,

$$
\mathbf{P} = \begin{bmatrix}
\mathbf{U} & \mathbf{V} & & & \\
\mathbf{W} & \mathbf{Z}_1 & \mathbf{Z}_0 & & \\
 & \mathbf{Z}_2 & \mathbf{Z}_1 & \mathbf{Z}_0 & \\
 & & \ddots & \ddots & \ddots
\end{bmatrix}, \quad \text{where}
$$

$$
\mathbf{U} = \begin{bmatrix}
\mathbf{p}_{0,0} & \mathbf{p}_{0,1} & \cdots & \mathbf{p}_{0,A-2} & \mathbf{p}_{0,A-1} \\
\vdots & \vdots & \cdots & \vdots & \vdots \\
\mathbf{p}_{N,0} & \mathbf{p}_{N,1} & \cdots & \mathbf{p}_{N,A-2} & \mathbf{p}_{N,A-1} \\
 & \ddots & \vdots & \vdots & \vdots \\
 & & \mathbf{p}_{A-1,A-N-1} & \cdots & \mathbf{p}_{A-1,A-1}
\end{bmatrix}
$$

$$
\mathbf{V} = \begin{bmatrix}
\mathbf{p}_{0,A} & & & & \\
\vdots & \ddots & & & \\
\mathbf{p}_{N,A} & \cdots & \mathbf{p}_{N,N+A} & & \\
\vdots & \vdots & \vdots & \ddots & \\
\mathbf{p}_{A-1,A} & \mathbf{p}_{A-1,A+1} & \cdots & \mathbf{p}_{A-1,2A-2} & \mathbf{p}_{A-1,2A-1}
\end{bmatrix}
$$

$$
\mathbf{W} = \mathbf{Z}_2 = \begin{bmatrix}
\mathbf{0} & \mathbf{p}_{A,A-N} & \cdots & \mathbf{p}_{A,A-2} & \mathbf{p}_{A,A-1} \\
 & \ddots & \ddots & & \vdots \\
 & & & & \mathbf{p}_{A+N-1,A-1} \\
 & & & & \mathbf{0}
\end{bmatrix}
$$

$$
\mathbf{Z}_1 = \begin{bmatrix}
\mathbf{p}_{A,A} & \mathbf{p}_{A,A+1} & \cdots & \mathbf{p}_{A,2A-2} & \mathbf{p}_{A,2A-1} \\
\vdots & \vdots & \cdots & \vdots & \vdots \\
\mathbf{p}_{A+N,A} & \mathbf{p}_{A+N,A+1} & \cdots & \mathbf{p}_{A+N,2A-2} & \mathbf{p}_{A+N,2A-1} \\
 & \ddots & \ddots & \ddots & \vdots \\
 & & \mathbf{p}_{2A-1,2A-N-1} & \cdots & \mathbf{p}_{2A-1,2A-1}
\end{bmatrix}
$$

$$
\mathbf{Z}_0 = \begin{bmatrix}
\mathbf{p}_{A,2A} & & & & \\
\vdots & \ddots & & & \\
\mathbf{p}_{A+N,2A} & \cdots & \mathbf{p}_{A+N,2A+N} & & \\
\vdots & \vdots & \ddots & \ddots & \\
\mathbf{p}_{2A-1,2A} & \mathbf{p}_{2A-1,2A+1} & \cdots & \mathbf{p}_{2A-1,3A-2} & \mathbf{p}_{2A-1,3A-1}
\end{bmatrix}.
$$

Figure 3.2. *Transition probability matrices.*

in the case that the size of best-effort queue is infinite, the calculation needs to be truncated at $B_b = \tau$ such that $1 - \sum_{n=0}^{N} \sum_{x=0}^{B_q} \sum_{y=0}^{\tau} \pi(n,x,y) < \epsilon$.

3.3.4.1 Average Queue Length

The average queue length for the QoS-sensitive queue ($\overline{x}$) and the best-effort queue ($\overline{y}$) are calculated as follows:

$$\overline{x} = \sum_{x=0}^{B_q} x \left(\sum_{y=0}^{B_b} \sum_{n=0}^{N} \pi(n,x,y) \right), \quad \overline{y} = \sum_{y=0}^{B_b} y \left(\sum_{x=0}^{B_q} \sum_{n=0}^{N} \pi(n,x,y) \right). \tag{3.30}$$

3.3.4.2 Packet Dropping Probability

The packet dropping probability can be obtained based on the average number of dropped packets per time slot [29]. For the QoS queue, given that there are x packets in the queue and the number of packet arrivals is s, the number of dropped packets is $s - (B_q - x)$ if $s > B_q - x$, and zero, otherwise. The average number of dropped packets per time slot can then be obtained as follows:

$$\overline{x}_{drop} = \sum_{y=0}^{B_b} \sum_{n=0}^{N} \sum_{x=0}^{B_q} \left(\sum_{s=B_q-x+1}^{A} \pi(n,x,y) \left(\sum_{j=0}^{N} \left[\mathbf{q}_{x,x+s}^{(y)} \right]_{n,j+1} \right) (s - (B_q - x)) \right) \tag{3.31}$$

where $\left[\mathbf{q}_{x,x+s}^{(y)} \right]_{i,j}$ is the element of matrix $\mathbf{q}_{x,x+s}^{(y)}$ at row i and column j. The factor $\left(\sum_{j=0}^{N} \left[\mathbf{q}_{x,x+s}^{(y)} \right]_{n,j+1} \right)$ in (3.31) indicates the total probability that the number of packets in the QoS queue increases by s when the number of packets in the best-effort queue is y. This probability differs from the probability of packet arrival, because we have to consider the successfully transmitted packet(s) in the same time slot as well.

After determining the average number of dropped packets per time slot, the probability that an incoming packet is dropped is given by $P_{dr}^{(q)} = \frac{\overline{x}_{drop}}{\overline{\lambda}_q}$, where $\overline{\lambda}_q$ is the average packet arrival rate at the QoS queue which can be obtained from (3.7).

For the best-effort queue, when the queue size is finite, this performance measures can be obtained in the similar way. First, we obtain the average number of dropped packets per time slot as follows:

$$\overline{y}_{drop} = \sum_{y=0}^{B_b} \sum_{n=0}^{N} \sum_{x=0}^{B_q} \left(\sum_{s=B_b-y+1}^{A} \pi(n,x,y) \left(\sum_{i=0}^{B_q} \sum_{j=0}^{N} [\mathbf{p}_{y,y+s}]_{k,l} \right) (s - (B_b - y)) \right)$$

$$k = x \times (N+1) + n + 1$$
$$l = i \times (N+1) + j + 1$$

and then packet dropping probability is calculated as follows $P_{dr}^{(b)} = \frac{\overline{y}_{drop}}{\lambda_b}$.

3.3.4.3 Queue Throughput

The throughput (in packets/time slot) for the QoS queue and the best-effort queue are obtained as $\eta_q = \overline{\lambda}_q(1 - P_{dr}^{(q)}), \quad \eta_b = \lambda_b(1 - P_{dr}^{(b)})$.

3.3.4.4 Average Delay for a Packet

Using the effective arrival rate, the average delay for a packet in the QoS queue ($\overline{d}_q$) and in the best-effort queue ($\overline{d}_b$) can be obtained by applying Little's law as follows: $\overline{d}_q = \frac{\overline{x}}{\eta_q}, \quad \overline{d}_b = \frac{\overline{y}}{\eta_b}$.

3.3.4.5 Delay Distribution

We utilize the concept of *absorbing Markov chain* to determine the delay distributions for packets in both the QoS queue and the best-effort queue (assuming that the size of the best-effort queue is finite). The general form of the transition probability matrix of an absorbing Markov chain is

$$\mathbf{P}_{abs} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Phi} & \mathbf{\Omega} \end{bmatrix} \tag{3.32}$$

where $\mathbf{\Omega}$ is the transient state transition matrix, and $\mathbf{\Phi}$ is the transition matrix to absorbing state. Note that, if there is only one absorbing state, matrix $\mathbf{\Phi}$ can be simply obtained from $\mathbf{\Phi} = \mathbf{1} - \mathbf{\Omega 1}$. Let $\boldsymbol{\alpha}$ denote the initial transient state probability matrix. The probability mass function and the distribution for the time required to reach the absorbing state can be expressed as follows:

$$f_{(d)}(w) = \boldsymbol{\alpha}\mathbf{\Omega}^{w-1}\mathbf{\Phi 1}, \quad F_{(d)}(W) = \sum_{w=1}^{W-1} f_{(d)}(w). \tag{3.33}$$

Based on matrix $\mathbf{P}$ in (3.20), we can establish the transient state transition matrix for both the QoS queue and the best-effort queue assuming th at the absorbing system

state is the state in which the number of packets in the queue is zero (i.e., the tagged packet departs queue). The delay for a packet can be measured as the required number of time slots (since the arrival of the packet) for the system to reach the absorbing state. Specifically, delay is measured from the first time slot that packet is in the queue until that packet leaves the queue. Note that, in this case, there is no arrival in that particular queue while the process moves towards the absorbing state.

For the QoS queue, we delete the first $N+1$ rows and columns in matrix $\mathbf{Q}_y$ (from (3.9)) for $y = 0, 1, \ldots, B_q$ and then we have

$$\mathbf{P}_{abs} = \left[\begin{array}{c|cccc} 1 & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \hline 1-\tilde{\mathbf{p}}_{0,0}\mathbf{1} & \tilde{\mathbf{p}}_{0,0} & & & \\ \vdots & \vdots & \ddots & & \\ 1-\sum_{i=0}^{N}\tilde{\mathbf{p}}_{N,N-i}\mathbf{1} & \tilde{\mathbf{p}}_{N,0} & \cdots & \tilde{\mathbf{p}}_{N,N} & \\ \vdots & & \ddots & \ddots & \ddots \\ 1-\sum_{i=0}^{N}\tilde{\mathbf{p}}_{B_b,B_b-i}\mathbf{1} & & & \tilde{\mathbf{p}}_{B_b,B-N} & \tilde{\mathbf{p}}_{B_b,B_b} \end{array}\right] \tag{3.34}$$

where $\tilde{\mathbf{p}}_{y,y'}$ is obtained from the modified $\mathbf{Q}_y$ without considering any packet arrival (i.e., with $\overline{\lambda}_q = 0$). We can establish initial transient state probability matrix from steady state probability $\boldsymbol{\pi}$ by deleting elements corresponding to $\pi(n, 0, y)$ and then normalizing it by using

$$\boldsymbol{\alpha} = \frac{\tilde{\boldsymbol{\pi}}}{\tilde{\boldsymbol{\pi}}\mathbf{1}} \quad \text{where} \quad \tilde{\boldsymbol{\pi}} = \left[\begin{array}{cccc} \tilde{\pi}(0) & \tilde{\pi}(1) & \cdots & \tilde{\pi}(B_b) \end{array}\right] \tag{3.35}$$

where $\tilde{\pi}(y)$ denotes modified steady state probability matrix when there are y packets in the best-effort queue which is obtained from

$$\tilde{\pi}(y) = \left[\begin{array}{ccc} [\boldsymbol{\pi}]_{y(N+1)(B_q+1)+N+2} & \cdots & [\boldsymbol{\pi}]_{(y+1)(N+1)(B_q+1)} \end{array}\right]. \tag{3.36}$$

For the best-effort queue, the transient state transition matrix is obtained by deleting the first $(N+1)(B_q+1)$ rows and columns of $\mathbf{P}$ so that we have

$$\mathbf{P}_{abs} = \left[\begin{array}{c|cccc} 1 & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \hline \mathbf{p}'_{1,0}\mathbf{1} & \mathbf{p}'_{1,1} & & & \\ \vdots & \vdots & \ddots & & \\ \mathbf{p}'_{N,0}\mathbf{1} & \mathbf{p}'_{N,1} & \cdots & \mathbf{p}'_{N,N} & \\ & & \ddots & \ddots & \ddots \\ & & & \mathbf{p}'_{B_b,B_b-N} & \mathbf{p}'_{B_b,B_b} \end{array}\right] \tag{3.37}$$

where $\mathbf{p}'_{y,y'}$ is modified using $\lambda_b = 0$. The initial transient state probability matrix

is obtained from

$$\alpha = \frac{\begin{bmatrix} \pi(1) & \pi(2) & \cdots & \pi(B_b) \end{bmatrix}}{\begin{bmatrix} \pi(1) & \pi(2) & \cdots & \pi(B_b) \end{bmatrix} \mathbf{1}} \tag{3.38}$$

where $\boldsymbol{\pi}(y)$ denotes the original steady state probability matrix for y packets in the best-effort queue and can be obtained from

$$\boldsymbol{\pi}(y) = \begin{bmatrix} [\pi]_{y(N+1)(B_q+1)+1} & \cdots & [\pi]_{(y+1)(N+1)(B_q+1)} \end{bmatrix}. \tag{3.39}$$

3.4 Performance Evaluation

3.4.1 Parameter Setting

We consider adaptive modulation with five transmission rates (i.e., $N = 5$) where the maximum transmission rate is achieved for 64-QAM. The values of a_n and g_n for fitting the packet error rate curve are the same as in [25]. The length of a time slot is 2 ms and the packet size is 1,080 bits. For fading channel, we assume a Nakagami-m channel with parameter $m = 1.1$. The size of the QoS-sensitive queue is 30 packets while that of the best-effort queue is infinite.

The assumed values for the other parameters are as follows: average SNR, $\overline{\gamma} = 15$ dB, $\phi_q = 0.7$, $\phi_b = 0.3$, $\rho = 0.4$, $\mu = 1/15$, $\lambda_q = 0.2$ packet per time slot per connection, $\lambda_b = 0.7$ packet per time slot, $\theta = 0.98$ and CAC threshold $T = 5$. Note that, we vary some of these parameters according to the evaluation scenarios, while the rest remain fixed according to the aforementioned setting.

3.4.2 Numerical Results and Discussions

3.4.2.1 Impact of CAC Threshold on Connection-Level and Packet-Level Performances

The connection blocking probability varies with connection arrival rate and CAC threshold. It increases/decreases with increasing arrival rate/CAC threshold, which is quite expected. Although a larger CAC threshold allows more number of connections to be admitted, since the radio resource (i.e., transmission rate) remains the same, packet-level performance degrades, for example, the packet dropping probability in

the QoS-sensitive queue increases with increasing threshold. Also, for a particular threshold, since the CAC mechanism limits the number of admitted connections to that value, the packet dropping probability becomes constant after certain arrival rate. Therefore, when the connection arrival rate is known, the CAC threshold can be selected such that the desired packet-level performance can be guaranteed. We do not plot these intuitive results for brevity.

3.4.2.2 Impact of Traffic Load on the Performance of QoS and BE Queue

As expected, the average packet delay increases as the connection arrival rate and/or the packet arrival rate in the QoS queue (ρ) increases (Figure 3.3(a)). Also, for relatively small ρ, due to the work-conserving property of fair scheduling, the BE queue would be allotted more than its assigned service rate. Therefore, smaller values of ρ result in smaller average delay for packets in the BE queue (Figure 3.3(b)).

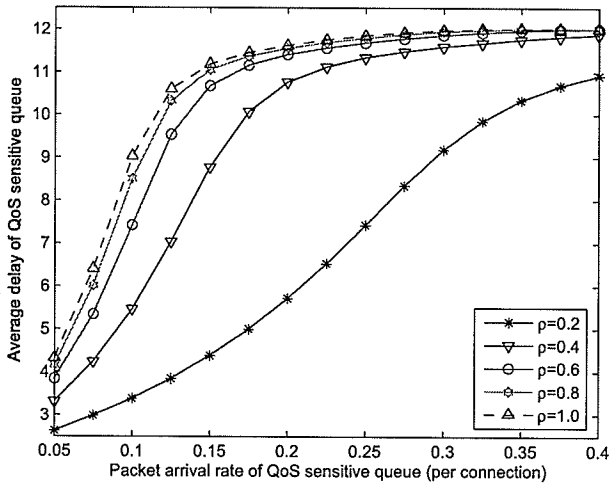
3.4.2.3 Effects of Physical Layer on the Queueing Performance

Figure 3.4 shows typical variations in packet dropping probability in the QoS queue under different SNR and packet error rate. In this case, we can define matrix $\mathbf{P}_c$ as a function of average SNR, packet error rate and the maximum number of modulation and coding level as follows $\mathbf{P}_c = \mathbf{P}_c\left(\overline{\gamma}, \overline{PER}, N\right)$. The packet dropping probability decreases as the average SNR increases, however, the rate of decrease becomes smaller with increasing SNR. Also, the impact of channel quality on the queueing performance at higher average SNR is less significant than that at lower SNR. Moreover, we observe that smaller packet error rate results in smaller packet dropping probability. Even though smaller packet error rate requires larger SNR threshold (i.e., lower modulation level at the same average SNR), the impact of re-transmission mechanism for the erroneous packets (i.e., due to the infinite persistent ARQ) is more significant on the queueing performance.

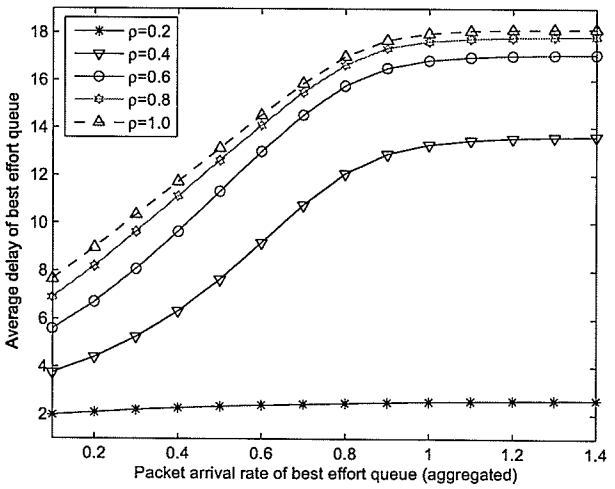
3.5 Chapter Summary

We have presented a model for service differentiation between the QoS-sensitive traffic and the best-effort traffic in wireless networks. In this model, fair scheduling is used

to prioritize QoS traffic over best-effort traffic and a threshold-based CAC is used to limit the number of connections for the QoS-sensitive service. A queueing analytical framework has been developed for this service differentiation model which also takes the multi-rate transmission feature in the physical layer also into account. From the analytical model, various QoS measures (at both the connection-level and the packet-level) can be obtained. The numerical results have shown that CAC combined with fair scheduling can provide the required level of QoS to both the QoS-sensitive and the best-effort traffic.



(a)



(b)

Figure 3.3. *Impact of connection arrival rate on average delay in (a) QoS queue, and (b) BE queue.*

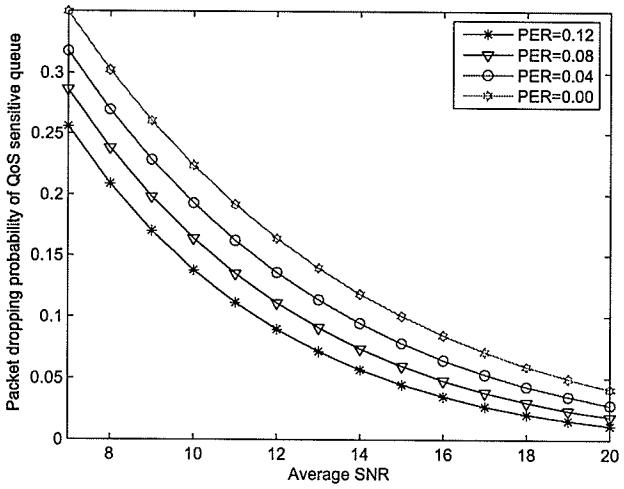


Figure 3.4. *Packet dropping probability of the QoS-sensitive queue under different packet error rates.*

Chapter 4

Multi-Service Cellular Mobile Networks with MMPP Call Arrival Patterns: Modeling and Analysis

4.1 Introduction

In a wireless mobile network, two important performance parameters are handoff call dropping and new call blocking probabilities. These refer to the probabilities that a handoff call is dropped and a newly initiated call is blocked by the call admission controller, respectively, due to the unavailability of radio channels. Since users are more sensitive to the dropping of ongoing calls than the blocking of new calls, CAC policies are generally designed such that the the handoff call dropping probability is minimized. To analyze the network performance in terms of these parameters, accurate and computationally efficient system model is required. Such a model can be used to find optimal system configuration in an adaptive radio resource management framework [31].

The evolving fourth-generation (4G) cellular networks will interwork with wireless LANs and wireless PANs to provide ubiquitous network connectivity to the mobile users with different classes of quality of service (QoS) requirements [32]. In such a heterogeneous environment the call arrival patterns in a cellular network are expected to be quite non-uniform and bursty. For such a network, a general performance model is required which is able to capture the fluctuations in the call arrival rates and take into account the existence of multiple classes of users.

Performance analysis of a cellular wireless network for a single class of users was presented in [3]. A model for performance evaluation of a multi-service cellular wireless network was proposed in [33] where the authors used a recursive formula approximation followed by a Markov model to obtain new call blocking and handoff call dropping probabilities as well as system utilization. The model can support arbitrary channel holding time distribution (e.g., gamma and hyperexponential) by using *phase type distribution* [28], even though the cost of using phase type distribution for arbitrary approximation is the larger matrix size. An approximation technique was used in [34] to analytically model micro- and pico-cellular wireless networks with arbitrary cell topology in a high mobility environment. The approximation is based on moment matching of handoff events by using single cell decomposition analysis. The model can approximate non-Poisson arrival rates and is suitable for heterogeneous traffic environment.

A two-dimensional Markov model for performance analysis under threshold CAC was proposed in [35]. In [36], a Markov model was used to analyze a *dynamic channel allocation (DCA)* scheme in which channels in a particular cell can be borrowed from another group of cells. Performance measures were derived considering both single class and two classes of users.

A performance analysis model for CAC in a mobile cellular network was presented in [22] in which the channel holding time for new calls and handoff calls were assumed to be non-identical and a two-dimensional Markov model was used to keep track of the number of new calls and handoff calls in the system. A performance analysis model based on *Stochastic Petri Net (SPN)* was proposed in [37] for multi-class mobile networks with different QoS requirements in terms of number of channels needed, channel holding time, and the number of guard channels. In [7], the concept of *fractional guard channel* was proposed, where to make optimal resource usage, calls are accepted with certain probability.

Our work in this chapter complements the above works in that we propose a Markov model for performance analysis of a multi-service cellular wireless network which captures the different arrival rates of new calls and handoff calls. Since the amount of traffic in a cellular wireless network may vary depending on the time of the day [38], using a pure Poisson arrival process is not enough to obtain accurate

performance parameters of the network in terms of the new call and handoff call blocking probabilities. Specifically, using the maximum/average call arrival rate for performance analysis may result in over estimation/under estimation of call dropping probabilities.

In the proposed model, this variation of arrival rates is taken into account by using a Markov modulated Poisson process (MMPP) and multiple classes of users in the network are considered through a multi-dimensional Markov model. Based on this model, we present an optimization formulation to maximize channel utilization while maintaining the ratio of the handoff call dropping probability and the new call blocking probability below some desired level. We also adopt a computationally efficient approach to solve the multi-dimensional Markov model so that on-line use of the proposed model for system engineering would be feasible.

4.2 System Model and Analysis

4.2.1 Channel Allocation/Reservation and Occupancy Model

We consider a cellular mobile wireless network which uses a fixed channel allocation scheme. Each cell has N channels and we assume that the arrival of new and handoff calls during time phase r follows Poisson distribution with rates $\lambda_{N,r}$ and $\lambda_{H,r}$, respectively. We also assume that the channel holding time for both new calls and handoff calls follows exponential distribution with mean $1/\mu_r$. For channel reservation, we assume a guard channel scheme in which a fixed number of channels $(N - K_r)$ are reserved for handoff calls during phase r. For example, if $N = 10$ and $K_1 = 8$, then 2 channels are reserved for handoff calls during time phase 1.

4.2.2 Model for Call Arrival Rate with MMPP Arrival

Here, variations in the call arrival rate is modeled by an MMPP. The MMPP is a doubly stochastic process where the intensity of the corresponding Poisson process is defined by the state of Markov chain. The parameters for the MMPP model for call arrival can be defined as

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,R} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,R} \\ \vdots & \vdots & & \vdots \\ p_{R,1} & p_{R,2} & \cdots & p_{R,R} \end{bmatrix} \tag{4.1}$$

$$\mathbf{\Lambda}_N = \begin{bmatrix} \lambda_{N,1} & & & \\ & \lambda_{N,2} & & \\ & & \ddots & \\ & & & \lambda_{N,R} \end{bmatrix} \quad \mathbf{\Lambda}_H = \begin{bmatrix} \lambda_{H,1} & & & \\ & \lambda_{H,2} & & \\ & & \ddots & \\ & & & \lambda_{H,R} \end{bmatrix} \tag{4.2}$$

where $\mathbf{\Lambda}_N$ and $\mathbf{\Lambda}_H$ define the arrival rate patterns for new calls and handoff calls, respectively, R is the maximum number of phases in the MMPP, $p_{r,s}$ is a transition rate from phase r to s, and λ_r is the arrival rate at phase r. For this model, there exists a matrix $\mathbf{x}$ such that $\mathbf{xP} = \mathbf{0}$ and $\mathbf{xe} = 1$, where $\mathbf{e}$ is column matrix of 1. The mean arrival rate for this MMPP can be calculated as

$$\bar{\lambda}_N = \mathbf{x}\mathbf{\Lambda}_N\mathbf{e} \tag{4.3}$$

$$\bar{\lambda}_H = \mathbf{x}\mathbf{\Lambda}_H\mathbf{e}. \tag{4.4}$$

For example, if the transitions among the different phases are sequential and circular (specifically, the system stays in each phase r with mean duration of $1/p_{r,r+1}$ except in phase R for which the mean duration is $1/p_{R,1}$), the state transition diagram is as shown in Figure 4.1. Such a model is able to represent the call arrival pattern with a high degree of autocorrelation [38].

4.2.3 Single Class of Users and MMPP Call Arrival Pattern

According to the above MMPP call arrival model and channel allocation, the state space for this Markov model (for a single class of users) can be defined as

$$\Delta = \{(P, U); 1 \leq P \leq R, 0 \leq U \leq N\} \tag{4.5}$$

where P and U represent the phase and the number of busy channels (i.e., the number of channels used by ongoing calls), respectively. Then the transition matrix is given

by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{B}_0 & & & & \\ \mathbf{C}_1 & \mathbf{A}_1 & \mathbf{B}_1 & & & \\ & \mathbf{C}_k & \mathbf{A}_k & \mathbf{B}_k & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{C}_{N-1} & \mathbf{A}_{N-1} & \mathbf{B}_{N-1} \\ & & & & \mathbf{C}_N & \mathbf{A}_N \end{bmatrix} \tag{4.6}$$

Each row k of matrix $\mathbf{Q}$ corresponds to the number of busy channels. For the elements in matrix $\mathbf{Q}$ (i.e., $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_N$; $\mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_{N-1}$;$\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N$), the rows correspond to the different phases of arrival. The size of each of these matrices is $R \times R$, where R is the order of MMPP model. In general, matrix $\mathbf{A}_k$ can be defined as

$$\mathbf{A}_k = \begin{bmatrix} a_{1,1}^k & p_{1,2} & \cdots & & & \\ p_{2,1} & a_{2,2}^k & p_{2,3} & \cdots & & \\ p_{3,1} & \cdots & a_{i,i}^k & p_{i,i+1} & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & \cdots & p_{R-1,R-2} & a_{R-1,R-1}^k & p_{R-1,R} \\ & & & \cdots & p_{R,R-1} & a_{R,R}^k \end{bmatrix} \tag{4.7}$$

where $a_{i,i}^k$s denote the diagonal elements of $\mathbf{A}_k$ and are given as follows:

$$a_{i,i}^k = -1 \times (\sum_{\forall j} b_{i,j}^k + \sum_{\forall j} c_{i,j}^k + \sum_{\forall j \neq i} a_{i,j}^k) \tag{4.8}$$

where $a_{i,j}^k$ is the element of matrix $\mathbf{A}_k$. Note that, (4.8) is used to normalize the transition matrix $\mathbf{A}_k$ in (4.7).

In (4.6), $\mathbf{C}_k$ and $\mathbf{B}_k$ are diagonal matrices. The matrix $\mathbf{C}_k$ corresponds to call departure during a phase and its elements are $k\mu_i$. Matrix $\mathbf{B}_k$ represents arrival of new calls and handoff calls and its elements $b_{i,i}^k$, $i \in \{1, \ldots, R\}$ are defined as follows:

$$b_{i,i}^k = \begin{cases} \lambda_{N,i} + \lambda_{H,i}, & 0 \leq k < K_i \\ \lambda_{H,i}, & K_i \leq k \leq R \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

Let $\boldsymbol{\pi}$ denote the matrix of steady state probabilities of the system. For example, $\pi(r, u)$ denotes the steady state probability of state (r, u), in which u channels are in

use and the arrival phase is r, and this can be obtained as follows:

$$\pi(r,u) = [\boldsymbol{\pi}]_{(u\times R)+r} \tag{4.10}$$

where $[\boldsymbol{\pi}]_i$ denotes the element at column i of matrix $\boldsymbol{\pi}$. From this steady state probability, we can obtain new call blocking probability (P_b) and handoff call dropping probability (P_d) as follows:

$$P_b = \sum_{r=1}^{R} \sum_{u=K_r}^{N} \pi(r,u) \tag{4.11}$$

$$P_h = \sum_{r=1}^{R} \pi(r,N). \tag{4.12}$$

Also, the channel utilization can be obtained as follows:

$$P_u = \frac{\sum_{r=1}^{R} \left(\sum_{u=1}^{N} u\pi(r,u) \right)}{N}. \tag{4.13}$$

4.2.4 Multiple Classes of Users

For multiple classes of users, a multi-dimensional Markov model is required where each of the dimensions in the model represents one class of users. For example, if there are three classes in the system, the Markov model will have three dimensions. To illustrate how to obtain the model, suppose we have two classes of users which correspond to voice and data call, respectively. Each of the voice call requires c_v channels and each of the data call requires c_d channels. In this case, the state space becomes

$$\Delta = \{(V,D); 0 \le V \le M_v, 0 \le D \le M_d\}, \tag{4.14}$$

where V and D represent the number of voice and data calls in the system, respectively, and M_v and M_d are the maximum number of ongoing voice and data calls, respectively, which can be calculated as $M_v = \lfloor N_v/c_v \rfloor$ and $M_d = \lfloor N_d/c_d \rfloor$. Note that, N_v and N_d are the maximum number of channels that can be used by voice and data calls, respectively.

The transition diagram of the model is shown in Figure 4.2. In this model, the number of guard channels for voice calls is $N_v - K^{(v)}$ and that for data calls is $N_d - K^{(d)}$.

The mean arrival rates of new calls and handoff calls and the mean channel holding time for voice and data calls are $\lambda_N^{(v)}$, $\lambda_H^{(v)}$, $1/\mu^{(v)}$, $\lambda_N^{(d)}$, $\lambda_H^{(d)}$, and $1/\mu^{(d)}$, respectively.

The transition matrix in this case is as follows:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{B}_0 & & & & \\ \mathbf{C}_1 & \mathbf{A}_1 & \mathbf{B}_1 & & & \\ & \mathbf{C}_k & \mathbf{A}_k & \mathbf{B}_k & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{C}_{M_v-1} & \mathbf{A}_{M_v-1} & \mathbf{B}_{M_v-1} \\ & & & & \mathbf{C}_{M_v} & \mathbf{A}_{M_v} \end{bmatrix}. \tag{4.15}$$

The rows of matrix $\mathbf{Q}$ correspond to the number of voice calls. Since the maximum number of voice calls is limited by M_v, the maximum size of matrix $\mathbf{Q}$ is M_v+1. Each row of the matrices inside $\mathbf{Q}$ (i.e., $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_{M_v}$;$\mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_{M_v-1}$;$\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_{M_v}$) corresponds to the number of data calls in the system. Since the maximum number of data calls is M_d, the maximum size of each of these matrices is $M_d + 1$.

Let $a_{i,j}^k$ denote items inside matrix $\mathbf{A}_k$ in (4.16). Matrices $\mathbf{B}_k$ and $\mathbf{C}_k$ are diagonal with elements $b_{i,i}^k$ and $c_{i,i}^k$, $i \in \{0, \ldots, M_d\}$, respectively. Then the values for $a_{i,j}^k$, $b_{i,i}^k$ and $c_{i,i}^k$ can be expressed by (4.8), (4.17), (4.18), (4.19) and (4.20), where $\Phi = kc_v + ic_d$, $\Phi' = kc_v + (i+1)c_d$, $\theta = ic_d$, and $\theta' = (i+1)c_d$.

$$\mathbf{A}_k = \begin{bmatrix} a_{0,0}^k & a_{0,1}^k & & & \\ a_{i,i-1}^k & a_{i,i}^k & a_{i,i+1}^k & & \\ & \ddots & \ddots & \ddots & \\ & a_{M_d-1,M_d-2}^k & a_{M_d-1,M_d-1}^k & a_{M_d-1,M_d}^k \\ & & a_{M_d,M_d-1}^k & a_{M_d,M_d}^k \end{bmatrix}. \tag{4.16}$$

$$a_{i,i+1}^k = \begin{cases} \lambda_N^{(d)} + \lambda_H^{(d)}, & 0 \le \Phi' < K^{(d)} \\ \lambda_H^{(d)}, & K^{(d)} \le \Phi' \le N_v; K^{(d)} \le \theta' \le N_d \\ 0, & \text{otherwise.} \end{cases} \tag{4.17}$$

$$a_{i,i-1}^k = \begin{cases} i \times \mu^{(d)}, & 0 \le \Phi < N_v; 0 \le \theta < N_d \\ 0, & \text{otherwise.} \end{cases} \tag{4.18}$$

$$b_{i,i}^k = \begin{cases} \lambda_N^{(v)} + \lambda_H^{(v)}, & 0 \le \Phi < K^{(v)} \\ \lambda_H^{(v)}, & K^{(v)} \le \Phi \le N_v; K^{(v)} \le \theta \le N_d \\ 0, & \text{otherwise.} \end{cases} \tag{4.19}$$

$$c_{i,i}^{k} = \begin{cases} k \times \mu^{(v)}, & 0 \leq \Phi < N_v; 0 \leq \theta < N_d \\ 0, & \text{otherwise.} \end{cases} \tag{4.20}$$

Let $\pi(u_v, u_d)$ denote the probability that there are u_v voice calls and u_d data calls in the system. Then, the new call blocking probability for voice calls ($P_b^{(v)}$) and data calls ($P_b^{(d)}$) can be obtained as follows:

$$P_b^{(v)} = \sum_{K^{(v)} \leq (u_v+1)c_v + u_d c_d \leq N} \pi(u_v, u_d) \tag{4.21}$$

$$P_b^{(d)} = \sum_{K^{(d)} \leq u_v c_v + (u_d+1)c_d \leq N} \pi(u_v, u_d). \tag{4.22}$$

The handoff call dropping probability for voice calls ($P_d^{(v)}$) and data calls ($P_d^{(d)}$) can be obtained as follows:

$$P_d^{(v)} = \sum_{N_v \leq (u_v+1)c_v + u_d c_d \leq N} \pi(u_v, u_d), \tag{4.23}$$

$$P_d^{(d)} = \sum_{N_d \leq (u_d+1)c_d + u_v c_v \leq N} \pi(u_v, u_d), \tag{4.24}$$

Also, the channel utilization is given by

$$P_u = \frac{\sum_{u_v=1}^{M_v} \left(\sum_{u_d=1}^{M_d} (u_v c_v + u_d c_d) \pi(u_v, u_d) \right)}{N}. \tag{4.25}$$

4.2.5 System with Two Classes of Users and MMPP Arrival Pattern

In this section, we combine the two models described earlier to obtain the system model for two classes of calls and MMPP call arrival patterns. Thus, the state space of the system is

$$\Delta = \{(P, V, D), 1 \leq P \leq R, 0 \leq V \leq \lfloor N_v/c_v \rfloor, 0 \leq D \leq \lfloor N_d/c_d \rfloor\} \tag{4.26}$$

and the state transition diagram for the model with sequential and circular MMPP is shown in Figure 4.3.

Let $\lambda_{N,r}^{(v)}$, $\lambda_{H,r}^{(v)}$, and $1/\mu_r^{(v)}$ ($\lambda_{N,r}^{(d)}$, $\lambda_{H,r}^{(d)}$, and $1/\mu_r^{(d)}$) denote the average values of new call arrival rate, handoff call arrival rate and channel holding time, respectively, for

voice and data calls in phase r of the MMPP, and $K_r^{(v)}$ and $K_r^{(d)}$ denote the number of guard channels for voice and data calls, respectively. To obtain the transition matrix, we have to replace the elements inside matrix $\mathbf{Q}$ in (4.15) for the two-class model by those of the MMPP call arrival model.

In this model, the matrices inside $\mathbf{Q}$ (i.e., $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_{M_v}; \mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_{M_v-1}; \mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_{M_v}$) correspond to the number of data calls and the maximum size of these matrices is $M_d + 1$, where M_d denotes the maximum possible number of data calls in the system. The rows of these matrices $\mathbf{a}_{i,j}^k$, $\mathbf{b}_{i,j}^k$, and $\mathbf{c}_{i,j}^k$ correspond to the phases of the MMPP. Note that, these matrices are the elements of the matrices $\mathbf{A}_k$, $\mathbf{B}_k$, and $\mathbf{C}_k$, respectively.

The element at row r column s (i.e., $\left[\mathbf{a}_{i,j}^k\right]_{r,s}$) is defined as follows:

$$\left[\mathbf{a}_{i,i}^k\right]_{r,s} = \begin{cases} p_{r,s}, & r \neq s \\ -1 \times \left(\sum_{\forall r} \left[\mathbf{b}_{i,i}^k\right]_{r,s} + \sum_{\forall r} \left[\mathbf{c}_{i,i}^k\right]_{r,s} + \sum_{r \neq s} \left[\mathbf{a}_{i,i}^k\right]_{r,s}\right), & r = s \end{cases} \tag{4.27}$$

$$\left[\mathbf{a}_{i,i+1}^k\right]_{r,r} = \begin{cases} \lambda_{N,r}^{(d)} + \lambda_{H,r}^{(d)}, & (k \times c_v) + (i \times c_d) < K_r^{(d)}, \\ \lambda_{H,r}^{(d)}, & K_r^{(d)} \leq (k \times c_v) + (i \times c_d) < N_d, \\ 0, & \text{otherwise} \end{cases} \tag{4.28}$$

$$\left[\mathbf{a}_{i,i-1}^k\right]_{r,r} = \begin{cases} i \times \mu_r^{(d)}, & (k \times c_v) + (i \times c_d) \leq N_d c_d \\ 0, & \text{otherwise} \end{cases} \tag{4.29}$$

where $p_{r,s}$ is the element at row r column s of matrix $\mathbf{P}$ of the MMPP model and all other elements are zero. The matrix $\mathbf{a}_{i,i}^k$ corresponds to the changes in the phases of the MMPP, $\mathbf{a}_{i,i+1}^k$ corresponds to data call arrival, and the matrix $\mathbf{a}_{i,i-1}^k$ corresponds to data call departure. Similarly, the elements of the matrices $\mathbf{b}_{i,i}^k$ and $\mathbf{c}_{i,i}^k$ which are used to represent the voice call arrival and departure, respectively, are obtained as follows:

$$\left[\mathbf{b}_{i,i}^k\right]_{r,r} = \begin{cases} \lambda_{N,r}^{(v)} + \lambda_{H,r}^{(v)}, & (k \times c_v) + (i \times c_d) < K_r^{(v)} \\ \lambda_{H,r}^{(v)}, & K_r^{(v)} \leq (k \times c_v) + (i \times c_d) < N_v \\ 0, & \text{otherwise} \end{cases} \tag{4.30}$$

$$\left[\mathbf{c}_{i,i}^k\right]_{r,r} = \begin{cases} k \times \mu_r^{(v)}, & (k \times c_v) + (i \times c_d) \leq N_v c_v \\ 0, & \text{otherwise.} \end{cases} \tag{4.31}$$

Let $\pi(r, u_v, u_d)$ denote the steady state probability that there are u_v voice calls and u_d data calls in the system when the arrival phase is r. Then, the new call blocking probability for voice call ($P_b^{(v)}$) and data call ($P_b^{(d)}$) can be obtained as follows:

$$P_b^{(v)} = \sum_{r=1}^{R} \sum_{K_r^{(v)} \le (u_v+1)c_v + u_d c_d \le N} \pi(r, u_v, u_d) \tag{4.32}$$

$$P_b^{(d)} = \sum_{r=1}^{R} \sum_{K_r^{(d)} \le u_v c_v + (u_d+1)c_d \le N} \pi(r, u_v, u_d). \tag{4.33}$$

The handoff call dropping probability for voice call ($P_d^{(v)}$) and data call ($P_d^{(d)}$) can be obtained as follows:

$$P_d^{(v)} = \sum_{r=1}^{R} \sum_{N_v \le (u_v+1)\times c_v \le N} \pi(r, u_v, u_d) \tag{4.34}$$

$$P_d^{(d)} = \sum_{r=1}^{R} \sum_{N_d \le (u_d+1)\times c_d \le N} \pi(r, u_v, u_d). \tag{4.35}$$

The channel utilization is given by

$$P_u = \frac{\sum_{r=1}^{R} \sum_{u_v=1}^{M_v} \left(\sum_{u_d=1}^{M_d} (u_v c_v + u_d c_d) \pi(r, u_v, u_d) \right)}{N}. \tag{4.36}$$

4.2.6 Calculation of Steady State Probabilities

In order to obtain the new call blocking and the handoff call dropping probabilities, we have to calculate the steady state probability for each state. One way to calculate these probabilities is by solving matrix $\boldsymbol{\pi}$ from the equations $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$, where $\mathbf{Q}$ is the transition matrix obtained from our model, $\mathbf{e}$ is a column vector of 1, and $\boldsymbol{\pi}$ contains the elements of $\pi(r, u)$, $\pi(u_v, u_d)$, and $\pi(r, u_v, u_d)$ for the system with MMPP arrival model and multiclass service. However, when the number of channels in system becomes large, the size of matrix $\mathbf{Q}$ will grow rapidly.

The steady state probabilities can be obtained more efficiently by using recursion (instead of solving the entire matrix) [39]. Following this approach, first we calculate $\mathbf{E}_k$ from

$$\mathbf{E}_k = \mathbf{A}_k + \mathbf{C}_k(-\mathbf{E}_{k-1}^{-1})\mathbf{B}_{k-1} \quad 1 \le k \le M \tag{4.37}$$

where $\mathbf{E}_0 = \mathbf{A}_0$ and $\mathbf{A}_k$, $\mathbf{B}_k$, and $\mathbf{C}_k$ are matrices inside $\mathbf{Q}$, and $M = N$ for the MMPP arrival model and $M = M_v$ for the two-class model. The vector $\boldsymbol{\pi}_k$ which is the element at column k of row matrix $\boldsymbol{\pi}$, can be iteratively calculated by using the following equations:

$$\boldsymbol{\pi}_M \mathbf{E}_M = \mathbf{0} \tag{4.38}$$

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_{k+1} \mathbf{C}_{k+1} (-\mathbf{E}_k^{-1}), \quad 0 \leq k \leq M-1 \tag{4.39}$$

$$\sum_{k=0}^{M} \boldsymbol{\pi}_k \mathbf{e} = 1. \tag{4.40}$$

Starting with the matrix $\boldsymbol{\pi}_M$, the other steady state probability matrices can be obtained by using the above iterative algorithm.

4.2.7 Estimation of MMPP Parameters

To obtain the performance measures from our model, information related to call arrival and channel holding time would be required. Formal techniques for trace fitting and parameter estimation can be used to obatin these information from system traces.

In [40], a trace fitting procedure for MMPP was proposed which takes both the autocovariance and the marginal distribution into account. The MMPP model is constructed based on combining two basic MMPP models together, so that one MMPP model (with 2^L phases) is used to capture the autocovariance and the other (with X phases) is used for marginal distribution.

The first step in this procedure is to estimate the autocovariance by using a weighted sum of exponential functions. Next, the parameters of the MMPP model with X phases are approximated within the constraint of autocovariance determined before. Then, the complete model for the MMPP is constructed by superposing these two basic MMPP models.

4.2.8 Optimal Number of Guard Channels

To enhance the performance of the system, the number of guard channels for handoff calls need to be dynamically adjusted according to the call arrival rate. For this, we formulate an optimization problem with the objective of maximizing the utilization of

the channel usage (P_u) while maintaining the ratio between the handoff call dropping probability and the new call blocking probability below some desired threshold. This threshold (ω), which essentially prioritizes handoff calls over new calls, is defined through the grade-of-service (GoS) as follows:

$$GoS = P_b + \omega P_d \tag{4.41}$$

where ω is the weight of handoff call dropping probability over new call blocking probability and a typical value for ω is 10 [41].

The optimization formulation for two-class of users is as follows:

$$\text{Maximize} \quad P_u \tag{4.42}$$

subject to:

$$\frac{P_b^{(v)}}{P_d^{(v)}} \geq \omega \quad \text{and} \quad \frac{P_b^{(d)}}{P_d^{(d)}} \geq \omega. \tag{4.43}$$

From the analytical model for two-classes of users as presented in Subsection 4.2.4, performance metrics such as channel utilization, new call blocking and handoff call dropping probabilities for both voice and data calls can be determined as functions of $K^{(v)}$ and $K^{(d)}$, which denote the number of guard channels for voice calls and data calls, respectively. Given the new call and handoff call arrival rate for both voice and data calls, the optimal number of guard channels for both types of calls can be obtained by enumeration method.

4.3 Numerical and Simulation Results

4.3.1 Parameter Settings

We use a 3-phase MMPP model for the incoming calls and it is arbitrarily chosen as follows:

$$\mathbf{P} = \begin{bmatrix} -0.01 & 0.01 & 0 \\ 0 & -0.01 & 0.01 \\ 0.02 & 0 & -0.02 \end{bmatrix} \tag{4.44}$$

$$\Lambda_1 = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 1.5 \end{bmatrix} \quad \Lambda_H^{(v)} = \begin{bmatrix} 0.5 & & \\ & 1 & \\ & & 0.5 \end{bmatrix} \tag{4.45}$$

where $\Lambda_H^{(v)}$ represents the arrival pattern of handoff voice calls (i.e., arrival rate in phase r is $\lambda_{H,r}^{(v)} = \left[\Lambda_H^{(v)}\right]_{r,r}$). The arrival rate for new voice calls is defined as $\Lambda_N^{(v)} = \rho\Lambda_1$, where ρ is a measure of the intensity of new call arrivals. The arrival rates for new data calls and handoff data calls are assumed to be 50% of those for voice calls, respectively.

Figure 4.4 shows typical traffic traces for new call arrival which are generated by using the MMPP model and the Poisson model for $\rho = 1$. For the Poisson arrival model, the same mean rates (as calculated from (4.3) and (4.4)) as those for the MMPP model are used. As is evident, the MMPP arrivals are more bursty than those due to the Poisson model.

An event-driven simulator is used to obtain the performance results in a single-cell environment. The channel holding time for both handoff calls and new calls is assumed to be exponentially distributed. A threshold-based CAC method is used, where the threshold is determined based on the number of guard channels. We assume that the the number of channels available for voice and data calls is 30. A voice call requires 1 channel and a data call requires 2 channels and the average channel holding time for a voice and a data call is assumed to be 5 and 10 minutes, respectively.

4.3.2 Model Validation

To validate the correctness of the model, we compare the results obtained from this model with those obtained by simulations. We set the threshold for voice and data calls to 28 and 26 (i.e., $K^{(v)} = 28$ and $K^{(d)} = 26$), respectively.

Typical variations in the new call blocking and handoff call dropping probabilities with the intensity parameter ρ are shown in Figure 4.5. As is evident from Figure 4.5, the simulation results follow the analytical results very closely. As the new call arrival rate increases, the new call blocking and handoff call dropping probabilities increase. However, since some guard channels are reserved for handoff calls, the handoff call dropping probability for both voice and data calls are smaller than new call blocking probabilities. Additionally, the performance results obtained for the

traditional Poisson arrival process, with the same mean arrival rate as that for the MMPP, are also shown in the same figure for comparison. Also, we observe that with the traditional Poisson-based model, the performance measures are quite optimistic and relatively different from the actual performance results when the call arrival pattern is bursty in nature.

The performance results on new call blocking and handoff call dropping probabilities are shown in Figure 4.6 for another scenario where the total number of data calls is limited to 14. We observe that the handoff data call dropping probability in this case increases significantly while both the voice call blocking and voice call dropping probabilities decrease. This is due to the fact that lesser number of channels are now available for data calls while more channels are available for the voice calls. This shows the effect of prioritization of voice calls over data calls by limiting the number of channels for data calls.

4.3.3 Performance Results with Optimization

Figure 4.7 shows variations in channel utilization along with the new call and handoff call dropping probabilities with new call arrival rate for voice calls (i.e., $\lambda_N^{(v)}$) when the CAC threshold K (and hence the number of guard channels) for voice and data calls is obtained by using the optimization model, The value of ω is set to be 10 in this case so that the ratio between the new call blocking and the handoff call dropping probability for both voice and data calls can be maintained larger than 10 (Figures. 4.7(a) and (b)).

In this scenario, since the new voice call arrival rate increases, the number of guard channels for voice calls will need to be increased to maintain voice call dropping probability at desired level. Also, the number of guard channels for data calls will also change accordingly so that the channel utilization is maximized while at the same time maintaining the desired ratio between the handoff data call dropping probability and new data call blocking probability.

We also present the results from a static (i.e., non-optimal) allocation, in which we set $K^{(v)} = 24$ and $K^{(d)} = 24$ to minimize the handoff call dropping probabilities of both voice and data calls. However, with these number of guard channels, the system becomes conservative so that more number of new calls are blocked. With

optimal setting, the ratio between the handoff call dropping probability and the new call blocking probability can be maintained at the desired level, while achieving the as much channel utilization as possible (Figure 4.7(c)).

The proposed framework can be used to determine the optimal number of guard channels dynamically under bursty call arrival patterns (i.e., at the different phases of MMPP). Figure 4.8 shows typical variations in new call blocking and handoff call dropping probabilities under varying call arrival intensity (ρ), where the threshold K for both voice calls and data calls in all the arrival phases is chosen dynamically based on ρ. We observe that the ratio between new call blocking probabilities and handoff dropping probabilities for both voice and data calls is maintained at the desired level (i.e., $\omega = 10$ from (4.43)), while the channel utilization increases compared with that for static guard channel threshold (e.g., for $K_r^{(v)} = 24$ and $K_r^{(d)} = 24$ $\forall r$). The improvement due to optimal guard channel thresholds is more significant at high traffic load conditions. Therefore, based on the traffic load condition, the guard channel thresholds can be dynamically adjusted to control the admission of new calls so that the channel utilization can be improved.

4.4 Chapter Summary

We have presented a Markov model to analyze the new call and the handoff call blocking probabilities under MMPP call arrival patterns in a multi-service cellular mobile network. Analytical results have been compared with the simulation results to validate the model and also have been compared with the results for the traditional Poisson model for call arrival. Although the results have been shown only for voice calls and data calls, the analytical methodology can be used for more than two classes of users. We have observed that when the call arrival pattern is bursty in nature, the analytical results for the MMPP-based call arrival model are much closer than those for the traditional Poisson-based model. This observation suggests that the traditional performance models of cellular wireless networks are not suitable for the environment with the burst call arrival rate. Finally, our proposed analytical model would be useful for radio resource allocation and CAC in future-generation wireless mobile networks.

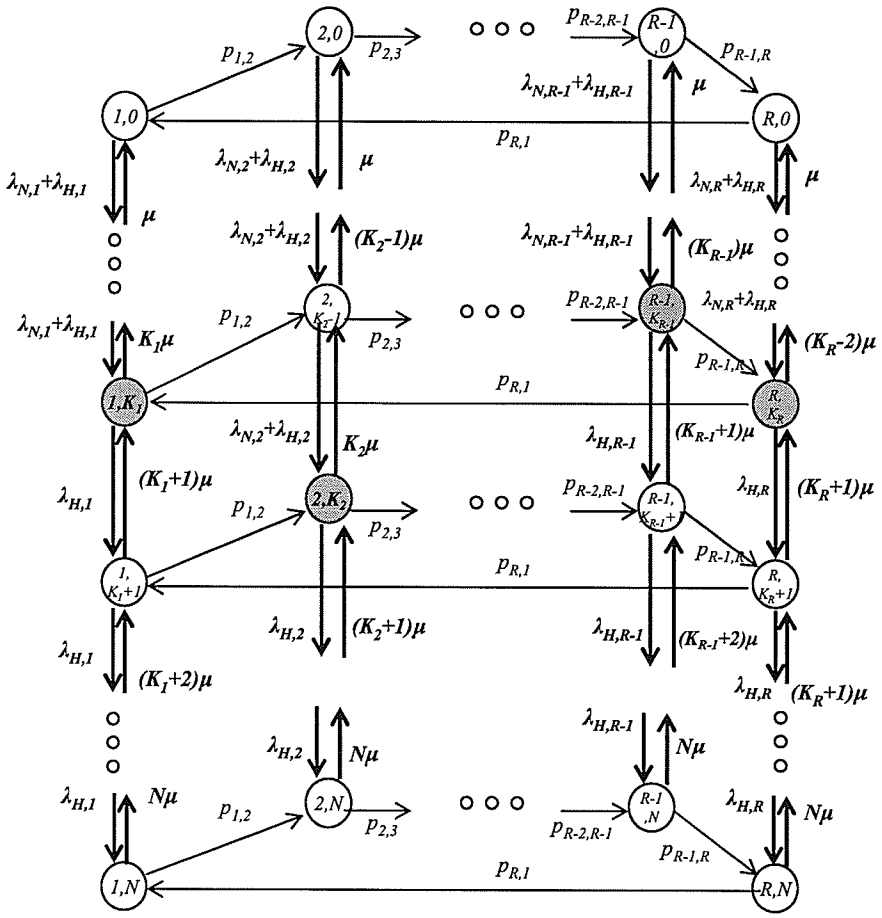


Figure 4.1. *Model for sequential and circular MMPP-based call arrival pattern with R different phases where K_r is the guard channel threshold in each phase.*

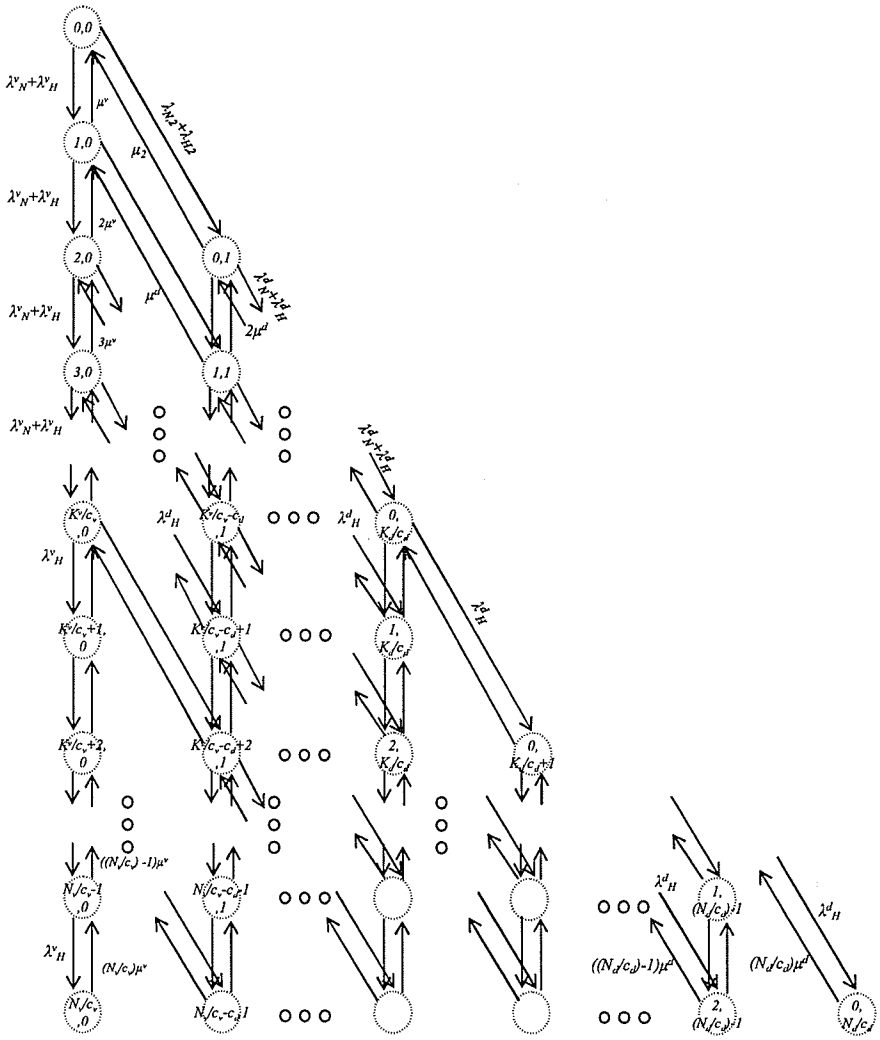


Figure 4.2. *General model for two classes of users.*

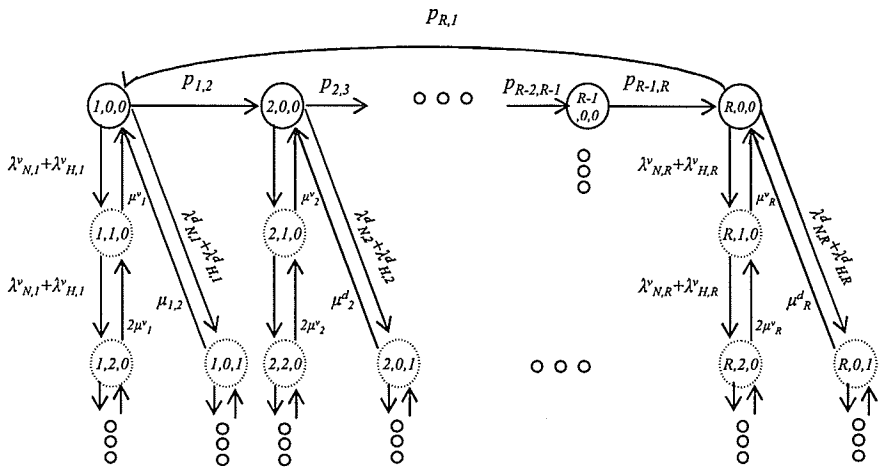


Figure 4.3. *Markov model for a system with two classes of users and MMPP call arrival pattern.*

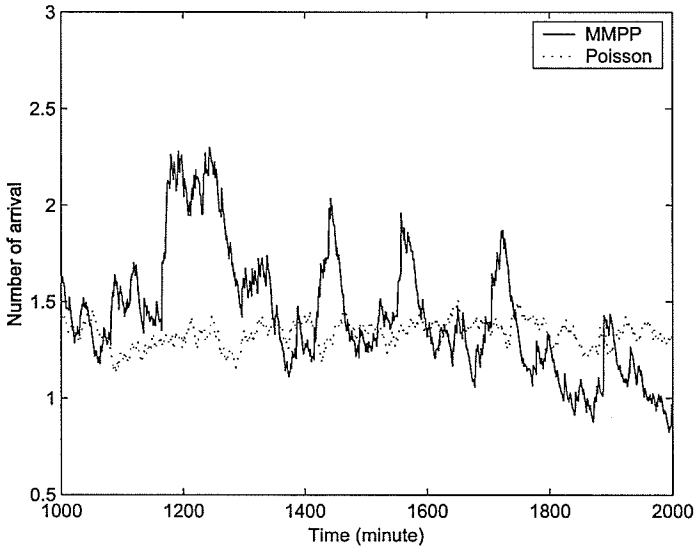


Figure 4.4. *Typical trace of MMPP traffic arrival.*

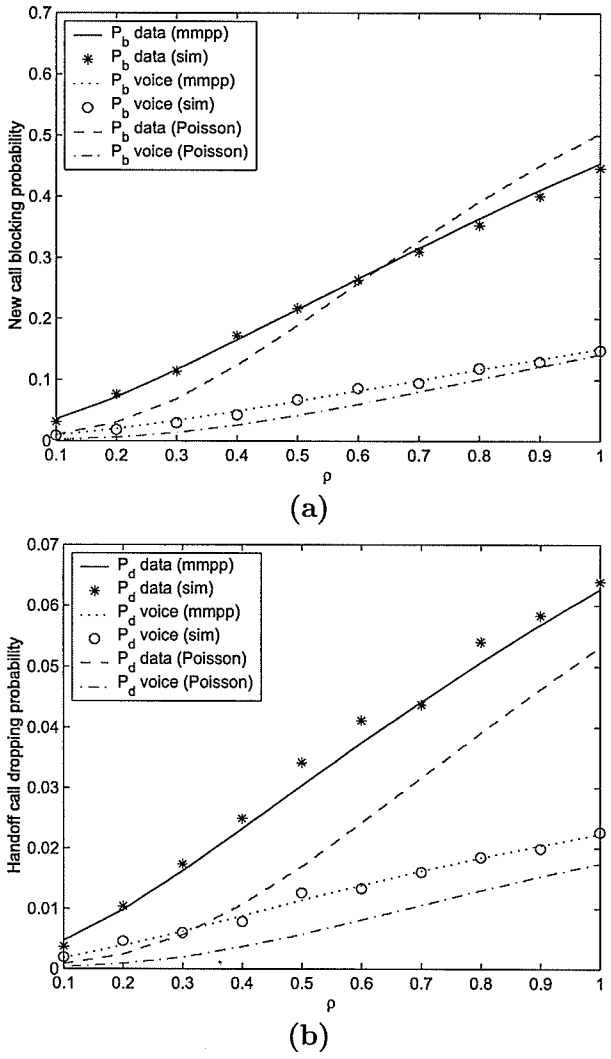


Figure 4.5. *Variations in (a) new call blocking probability (b) handoff call dropping probability with different traffic intensity ρ ($N = 30$).*

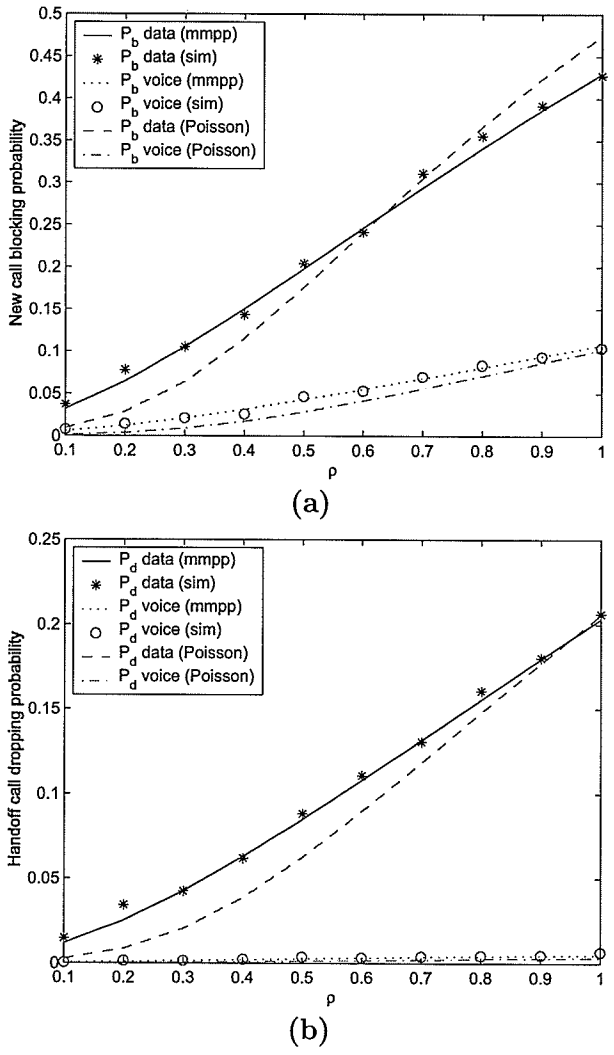
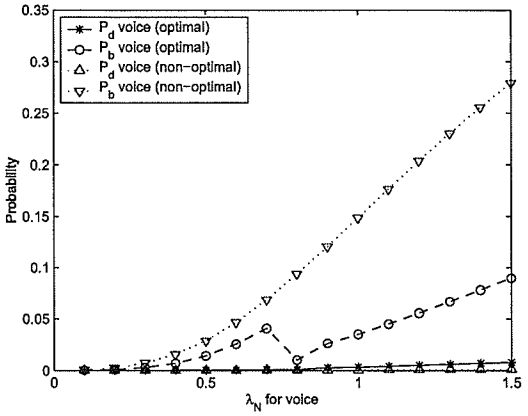
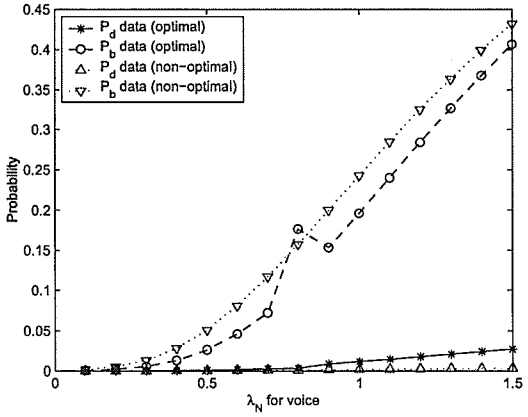


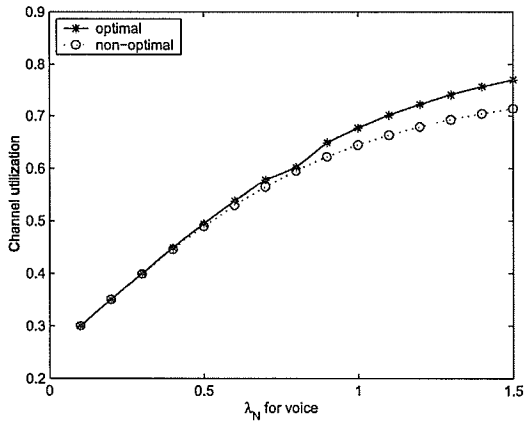
Figure 4.6. *Variations in (a) new call blocking probability and (b) handoff call dropping probability when the the number of data calls is limited to 14 (for $N = 30$).*



(a)

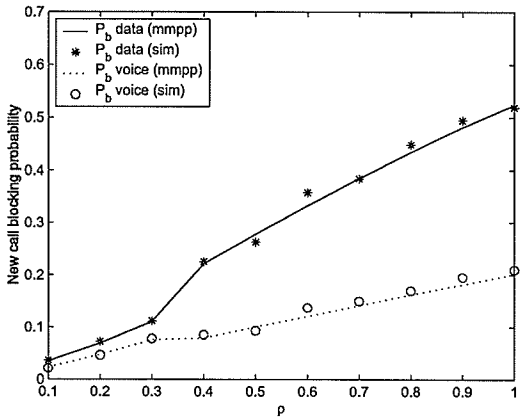


(b)

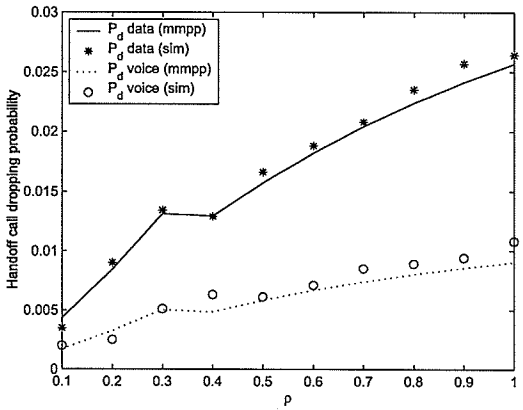


(c)

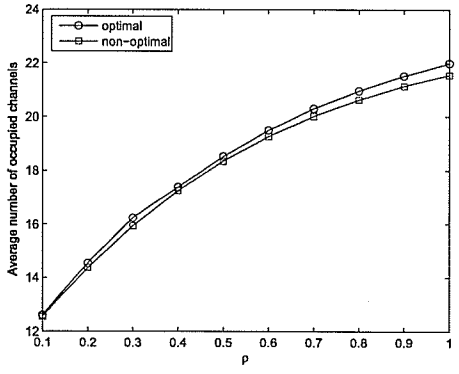
Figure 4.7. *(a) Performance measures for voice calls, (b) performance measures for data calls, and (c) channel utilization.*



(a)



(b)



(c)

Figure 4.8. *Variations in (a) new call blocking probability, (b) handoff call dropping probability and (c) channel utilization with different number of guard channels in each phase of MMPP (for $N = 30$) .*

Chapter 5

Adaptive Bandwidth Allocation in Cellular Mobile Networks Under Markov Call Arrival Process and Phase-Type Channel Holding Time Distribution

5.1 Introduction

In wireless mobile multimedia networks adaptive bandwidth allocation (ABA) [42] is necessary to maximize the utilization of radio channels while keeping the quality-of-service (QoS) of a multimedia call at the acceptable level. ABA can minimize the number of blocked new calls and the number of dropped handoff calls by adjusting the allocated bandwidth of ongoing calls and allowing the incoming calls to be serviced without degrading the QoS of the ongoing calls below the acceptable level.

The performance analysis models for ABA were proposed in [58]-[59] based on the assumptions that arrivals of new calls and handoff calls follow Poisson distribution and the distribution of channel holding time is exponential. However, the distributions for channel holding time of new calls and handoff calls are different from exponential [22] and models such as sum of hyper-exponential (SOHYP) [43] and hyper-Erlang [44] are more realistic than exponential-based models.

In this chaper, we propose an analytical framework for ABA in which call arrivals

and channel holding time are modeled by Markov Arrival Process (MAP) [45] and phase-type distribution, respectively. Due to the use of MAP, the framework is flexible to capture autocorrelation in the incoming call process as well as the burstiness (e.g., through MMPP which is special type of MAP). The phase-type distribution is general enough to represent many of the well-known probability distributions. We examine the performance of ABA for two different CAC schemes, namely, the guard-channel scheme [3] and the call thinning scheme [46].

Phase-type distribution was used in [47] to model service time of the calls in cellular networks. The model with MAP call arrival was proposed in [48]. However, these models were developed for mobile networks where the bandwidth allocation to the calls is non-adaptive (i.e., static). In [49], phase-type models for different channel holding time distributions such as SOHYP and hyper-Erlang were introduced so that the parameters for the phase-type distribution can be obtained. However, fitting techniques [50] can also be applied to the trace data with arbitrary distribution to obtain the phase-type parameters.

5.2 System Model and Assumptions

5.2.1 ACA and CAC

We consider the wireless cellular networks in which bandwidth of ongoing calls can be adjusted adaptively according to the states of the network. The total bandwidth in a cell is constant and is denoted by c. We assume that the bandwidth for a call is chosen from a set of discrete values $B = \{b_1, b_2, \ldots, b_n\}$ where $b_i < b_{i+1}$ for $i = 1, \ldots, n-1$. The minimum and the maximum amount of bandwidth allocated to a call is b_1 and b_n, respectively. The bandwidth requirement for a mobile is denoted by b_{req} ($b_n \geq b_{req} \geq b_1$). If the amount of bandwidth allocated to a mobile is less than b_{req}, then a degradation in call quality occurs, however, the call is not dropped.

We consider two schemes, namely, the guard channel scheme and the call thinning scheme for CAC of new calls and handoff calls. In the guard channel scheme, a portion of the available bandwidth is reserved for handoff calls. In other words, incoming new calls are accepted if the number of ongoing calls in the cell is less than predefined threshold τ. Hence, the bandwidth reserved for handoff calls is $c - \tau$.

Call thinning scheme is the generalized version of *fractional guard channel scheme* [7]. We consider a thinning scheme in which a new call is admitted with *acceptance probability* α. This probability is set according to the state of the network (e.g., the current number of ongoing calls). Let $\alpha(x)$ denote the acceptance probability when the number of ongoing call is x. When $\alpha(x) > \alpha(x+1), \quad x \in \{0, 1, \ldots, c-1\}$, the number of admitted new calls become thinner if the number of ongoing calls increases. The advantage of call thinning scheme is the ability to smooth the traffic admission rate rather than cutting the call acceptance at a certain level of load [46].

5.2.2 Bandwidth Adaptation Algorithm

We consider a fairness-based bandwidth adaptation algorithm [17] which works in such a way that the allocated bandwidth to the ongoing calls will not differ from each other by more than one step. The bandwidth of an ongoing call is also allowed to be degraded below bandwidth requirement b_{req} to minimize new call blocking and handoff call dropping probabilities.

The complete description of our bandwidth adaptation algorithm for guard channel scheme is shown in Algorithm 5.2.1. Let w_{allc} and $\mathbf{b}_{allc}$ denote the expected bandwidth for an incoming call and the bandwidth vector of ongoing calls, respectively. When a call arrives, the cell performs admission control by checking whether the total number of ongoing calls is less than the threshold t. If this condition is satisfied or if the incoming call is a handoff call, the cell tries to allocate maximum bandwidth to the incoming call; otherwise, the incoming new call is blocked. However, if the available bandwidth is not enough to allocate maximum bandwidth to an incoming call, the adaptation algorithm is invoked. The adaptation algorithm will randomly select an ongoing call with the current maximum bandwidth (i.e., $\max(\mathbf{b}_{allc})$) and degrade allocated bandwidth of that call one step. At this point, expected bandwidth for incoming call increases one step. This operation is iteratively performed until the expected bandwidth for an incoming call is equal to the current minimum bandwidth of all ongoing calls (i.e., $\min(\mathbf{b}_{allc})$). In contrast, if every call has the minimum bandwidth b_1, none of the ongoing call can be degraded. Therefore, an incoming call is dropped.

For call thinning scheme, line 1 of this algorithm would be changed to admit the

Algorithm 5.2.1: ACA(*inputs* : type of incoming call, $K, C, \mathbf{c}_{allc}, c_1, c_{max}$)

if (((incoming call is a new call) **and** (number of ongoing calls $< K$))
or (incoming call is a handoff call))

then
- **if** available bandwidth $\geq c_{max}$
  - **then** assign c_{max} to incoming call
  - **else**
    - $w_{allc} \leftarrow 0$
    - **while** $\max(\mathbf{c}_{allc}) > c_1$ **and** $w_{allc} < \min(\mathbf{c}_{allc})$
      - **do**
        - randomly select one call with number of channels $\max(\mathbf{c}_{allc})$
        - decrease number of channels for the selected call by one
      - $w_{allc} \leftarrow w_{allc} + 1$
    - **if** $w_{allc} > 0$
      - **then** accept incoming call with number of channels w_{allc}
      - **else** reject incoming call

else reject incoming new call

new call based on a Bernoulli trial with probability $\alpha(x)$.

In the event that a call is terminated or is handed over to a neighboring cell, the bandwidth of that call is released. In this case, some of the ongoing calls will be upgraded to higher bandwidth level. Our upgrade scheme randomly selects a call with the current minimum bandwidth or $\min(\mathbf{b}_{allc})$. The bandwidth of that call is upgraded by one step. This routine is performed until all released bandwidth is allocated or all of the ongoing calls have the maximum bandwidth (b_n).

From our bandwidth adaptation algorithm, the number of calls ($m_i(B, c, x)$) with allocated bandwidth b_i can be calculated from the number of ongoing calls x as follows:

$$m_i(B,c,x) = \begin{cases} x - \left\lfloor \frac{c - b_i x}{b_{i+1} - b_i} \right\rfloor, & i = \hat{i} \\ \left\lfloor \frac{c - b_i x}{b_{i+1} - b_i} \right\rfloor, & i = \hat{i} + 1 \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

where $\hat{i} = \hat{i}(B, c, x) = \max\{j | b_j \leq c/x\}$. The number of degraded calls $d(x)$, when the number of ongoing calls is x, can be obtained from $d(x) = m_i(B, c, x), \ b_i < b_{req}$.

The average bandwidth at the certain number of calls is calculated as follows:

$$\bar{b}(x) = \begin{cases} \frac{c}{x} & \lfloor \frac{c}{b_{i+1}} \rfloor \leq x \leq \lfloor \frac{c}{b_i} \rfloor \\ b_n & \lfloor \frac{c}{b_n} \rfloor \geq x. \end{cases} \tag{5.2}$$

For example, with total bandwidth/cell of 30 units, $B \in \{1, 2, 3\}$, and $b_{req} = 2$, when the number of ongoing calls is 25, $m_2(B, 30, 25) = 5$, $m_1(B, 30, 25) = 20$, $d(25) = 20$, and $\bar{b}(25) = 1.2$ units.

5.3 Formulation of the Queueing Model and Analysis

5.3.1 Call Arrival and Channel Holding Time Distribution

For MAP arrival process, we use the matrices $\mathbf{C}_N$, $\mathbf{C}_H$, and $\mathbf{C}_0$ to describe the process of a new call arrival, handoff arrival, and no arrival, respectively. The size of these matrices is $K \times K$ where K is the number of states in the arrival process. Let $\lambda_{i,k}^{(N)}$ ($\lambda_{i,k}^{(H)}$), $i, k \in \{1, \ldots, K\}$ denote the elements in $\mathbf{C}_N$ ($\mathbf{C}_H$) corresponding to transition from state i to k as a new call (handoff call) arrives. The elements $\lambda_{i,k}$, $i \neq k$ in matrix $\mathbf{C}_0$ correspond to the transitions without any arrival. Since $\mathbf{C} = \mathbf{C}_0 + \mathbf{C}_N + \mathbf{C}_H$, the diagonal elements of matrix $\mathbf{C}_0$ should be negative to satisfy the condition $\mathbf{Ce} = 0$, where $\mathbf{e}$ is column vector of 1.

We can obtain π_{MAP} which is the steady state probability of MAP by solving $\pi_{MAP}\mathbf{C} = 0$ and $\pi_{MAP}\mathbf{e} = 1$. According to this steady state probability, the mean rate of the new call and handoff call can be calculated from $\lambda_N = \pi\mathbf{C}_N\mathbf{e}$ and $\lambda_H = \pi\mathbf{C}_H\mathbf{e}$, respectively.

We use phase-type distribution to model the channel holding time, because it is general enough to fit well-known channel holding time distributions (e.g., sum of hyperexponential (SOHYP) and Hyper-Erlang). Also, with phase-type distribution standard techniques (e.g., Quasi-Birth Death Process (QBD)) can be used to analyze the system performances.

The phase-type distribution is defined by an absorbing Markov chain with k transient states and one absorption state. The parameters of phase-type distribution are $(\beta, \mathbf{S})$, in which $\mathbf{S}$ is the transition matrix of the transient states for M phases, and

β is the probability vector of each phase at time zero. The size of β and $\mathbf{S}$ are $1 \times M$ and $M \times M$, respectively. Continuous phase-type distribution can be shown in matrix form as

$$\begin{bmatrix} \mathbf{S} & S^0 \\ \mathbf{0} & 0 \end{bmatrix} \tag{5.3}$$

where the bottom row is the absorbing state and vector S^0 contains the rate to absorption state from transient states. The probability density function (pdf) and the mean of phase-type distribution can be obtained as follows:

$$f(x) = \beta e^{\mathbf{S}x} S^0 \tag{5.4}$$

$$\overline{x} = \beta(\mathbf{I} - \mathbf{S}^{-1})\mathrm{e}. \tag{5.5}$$

There are two approaches to obtain the parameters of phase-type distribution for a channel holding time: analytical and empirical approaches. An analytical approach can provide a systematic method to obtain parameters [49]. In contrast, an empirical approach uses the fitting technique [50] to obtain the parameters from the data set and can be more flexible to be applied with arbitrary distributions (e.g., Gamma distribution). In our model we use a general phase-type distribution, and therefore, both approaches can be used.

5.3.2 Markov Model

As was shown in [22], the channel holding time for the new calls and the handoff calls can be different. In our model we consider the number of new calls and handoff calls separately while describing the the system states.

For guard channel scheme, the state space of the model is

$$\begin{aligned} \Delta = \{(N, H, A, P);\ & 0 \le N \le t, 0 \le H \le c, \\ & 1 \le A \le K, 1 \le P \le M\} \end{aligned} \tag{5.6}$$

where N and H represent the number of new calls and handoff calls, respectively, P and A represent the phases of departure and the states of arrival process.

For call thinning scheme, the state space of model is similar to that for the guard channel scheme, but the maximum number of new calls is c ($0 \le x_n \le c$). The system state transition diagrams with the guard channel and the call thinning schemes are shown in Figure 5.1 and Figure 5.2, respectively.

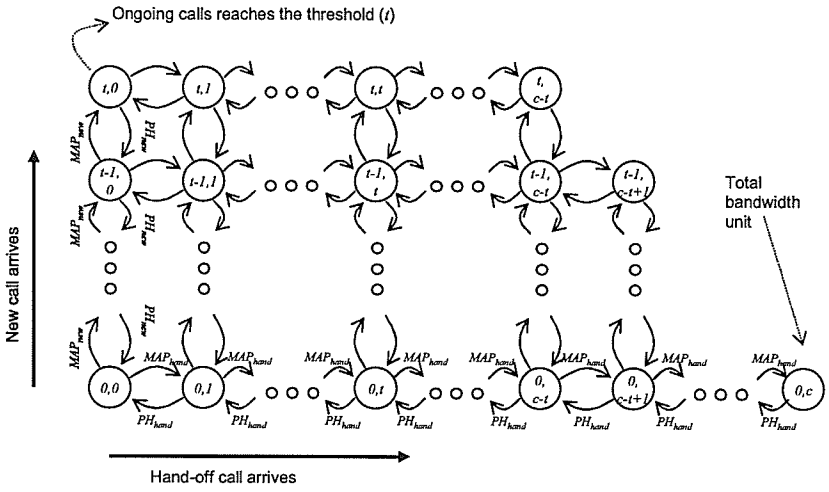


Figure 5.1. *State transition diagram for guard channel scheme where each state represents the number of new calls x_n and the number of handoff calls x_h.*

The rate transition matrices for both models have a block tri-diagonal form with MAP arrival and phase-type service time, as shown in (5.7) below

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{B}_0 & & & & \\ \mathbf{D}_1 & \mathbf{A}_1 & \mathbf{B}_1 & & & \\ & \mathbf{D}_k & \mathbf{A}_k & \mathbf{B}_k & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{D}_{N-1} & \mathbf{A}_{N-1} & \mathbf{B}_{N-1} \\ & & & & \mathbf{D}_N & \mathbf{A}_N \end{bmatrix} \tag{5.7}$$

where each row of elements in matrix $\mathbf{Q}$ represents the number of new calls, thus, $N = t + 1$ for guard channel scheme and $N = c + 1$ for call thinning scheme. The rows of inside matrices ($\mathbf{A}_k,\ k \in \{0, \ldots, N\}$, $\mathbf{B}_k,\ k \in \{0, \ldots, N-1\}$, and $\mathbf{D}_k,\ k \in \{1, \ldots, N\}$) represent the number of handoff calls. Since both the guard channel and the call thinning schemes involve the arrival process only, the matrix $\mathbf{D}_k$, representing departure process of new call, are identical and diagonal in both models.

Let $\mathbf{d}_{l,l}^k$ denote the diagonal elements of matrix $\mathbf{D}_k$ when the number of new calls

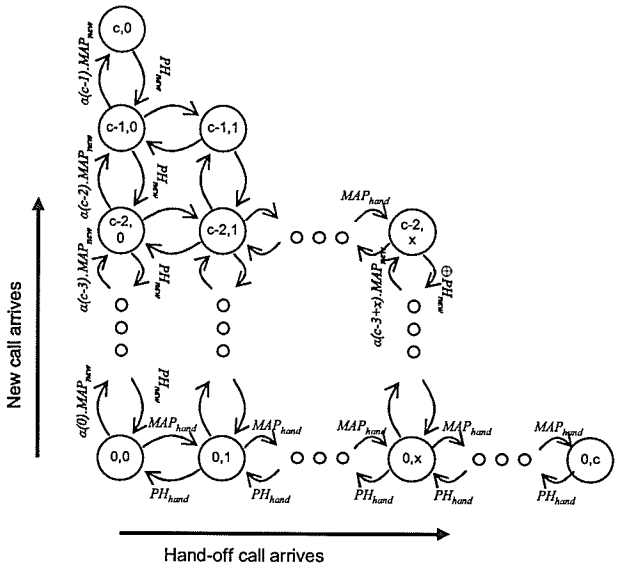


Figure 5.2. *State transition diagram for call thinning scheme where each state represents the number of new calls x_n and the number handoff calls x_h.*

is k and the number of handoff calls is l. We obtain the following:

$$\mathbf{d}^k_{(l,l)} = \begin{cases} \mathbf{I}_C \otimes (l \times \mathbf{S}^0_n \beta_n), & l > 1 \\ \mathbf{I}_C \otimes (\mathbf{S}^0_n), & l = 1 \end{cases} \tag{5.8}$$

where $\mathbf{S}^0_h$, β_h and $\mathbf{S}^0_n$, β_n are the parameters of phase-type distributions for the channel holding times of handoff and new call, respectively, $\otimes$ denotes kronecker product operator, and $\mathbf{I}_C$ is an identity matrix with size the same size of matrix $\mathbf{C}$.

For guard channel scheme, $\mathbf{B}_i$ are the diagonal matrix with elements $\mathbf{b}^k_{(l,l)} = \mathbf{I}_{S_n} \otimes \mathbf{C}_N \otimes \mathbf{I}_{S_h}$, $k + l < t$ where $\mathbf{C}_N$ is an arrival process of the new call. Matrix $\mathbf{A}_k$ is the Markov process for handoff calls and is defined in (5.9), where $\mathbf{a}^k_{l,l+1} = \mathbf{I}_{S_n} \otimes \mathbf{C}_H \otimes \mathbf{I}_{S_h}$ represents the arrival process and $\mathbf{a}^k_{(l,l-1)}$ represents the departure process of handoff calls, and are given by (5.10) and (5.11), where $\oplus$ is the kronecker sum defined as $\mathbf{X} \oplus \mathbf{Y} = (\mathbf{X} \otimes \mathbf{I}_Y) + (\mathbf{I}_X \otimes \mathbf{Y})$.

$$\mathbf{a}^k_{(l,l-1)} = \begin{cases} \mathbf{I}_{S_n} \otimes \mathbf{I}_C \otimes (l \times \mathbf{S}^0_h \beta_h), & l > 1 \\ \mathbf{I}_{S_h} \otimes \mathbf{I}_C \otimes (\mathbf{S}^0_h), & l = 1. \end{cases} \tag{5.10}$$

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{a}^k_{0,0} & \mathbf{a}^k_{0,1} & & & & & \\ \mathbf{a}^k_{1,0} & \mathbf{a}^k_{1,1} & \mathbf{a}^k_{1,2} & & & & \\ & \mathbf{a}^k_{l,l-1} & \mathbf{a}^k_{l,l} & \mathbf{a}^k_{l,l+1} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \mathbf{a}^k_{c-k-1,c-k-2} & \mathbf{a}^k_{c-k-1,c-k-1} & \mathbf{a}^k_{c-k-1,c-k} & \\ & & & & \mathbf{a}^k_{c-k-1,c-k} & \mathbf{a}^k_{c-k,c-k} & \end{bmatrix}. \tag{5.9}$$

$$\mathbf{a}^k_{(l,l)} = \begin{cases} \bigoplus_{i=1}^k \mathbf{S}_n \oplus \mathbf{C}_0 & l = 0 \\ \bigoplus_{i=1}^k \mathbf{S}_n \oplus \mathbf{C}_0 \oplus \bigoplus_{i=1}^l \mathbf{S}_h & 0 < l < t-k, l+k < t \\ \bigoplus_{i=1}^k \mathbf{S}_n \oplus (\mathbf{C}_0 + \mathbf{C}_N) \oplus \bigoplus_{i=1}^l \mathbf{S}_h & 0 < l < t-k, l+k \geq t \\ \bigoplus_{i=1}^k \mathbf{S}_n \oplus (\mathbf{C}_0 + \mathbf{C}_N + \mathbf{C}_H) \oplus \bigoplus_{i=1}^l \mathbf{S}_h & l = t-k. \end{cases} \tag{5.11}$$

In call thinning scheme, the differences are at $\mathbf{B}_k$ and $\mathbf{a}^k_{l,l}$ which are the matrices inside $\mathbf{A}_k$. The matrices $\mathbf{B}_k$ are diagonal with elements $\mathbf{b}^k_{(l,l)} = (\alpha(k+l)\mathbf{C}_N) \otimes \mathbf{I}_{S_n}$. The matrices $\mathbf{a}^k_{l,l}$ are defined in (5.12). Note that, factors $\bigoplus_{i=1}^k \mathbf{S}_n$ and $\bigoplus_{i=1}^l \mathbf{S}_h$ can be approximated by using method in [51], so that the steady state probability can be calculated more efficiently.

In order to obtain QoS measures of the adaptive bandwidth allocation, we have to calculate the steady state probability for each state. One way to calculate these probabilities is by solving matrix $\boldsymbol{\pi}$ from $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$, where $\mathbf{Q}$ is the transition matrix. However, when the number of channels becomes large, the size of matrix $\mathbf{Q}$ grows rapidly. Another way to solve the steady state probability is by recursion as was proposed in [39]. Following this approach, first we calculate $\mathbf{E}_k$ from

$$\mathbf{E}_k = \mathbf{A}_k + \mathbf{D}_k(-\mathbf{E}^{-1}_{k-1})\mathbf{B}_{k-1}, \qquad 1 \leq k \leq N \tag{5.13}$$

$$\mathbf{a}^k_{(l,l)} = \begin{cases} \bigoplus_{i=1}^k \mathbf{S}_n \oplus (\mathbf{C}_0 + (1-\alpha(k+l))\mathbf{C}_N) & l = 0 \\ \bigoplus_{i=1}^k \mathbf{S}_n \oplus (\mathbf{C}_0 + (1-\alpha(k+l))\mathbf{C}_N) \oplus \bigoplus_{i=1}^l \mathbf{S}_h & 0 < l < c \\ \bigoplus_{i=1}^k \mathbf{S}_n \oplus (\mathbf{C}_0 + (1-\alpha(k+l))\mathbf{C}_N + \mathbf{C}_H) \oplus \bigoplus_{i=1}^l \mathbf{S}_h & l = c. \end{cases} \tag{5.12}$$

where $\mathbf{A}_k$, $\mathbf{B}_k$, and $\mathbf{D}_k$ are matrices inside $\mathbf{Q}$, $\mathbf{E}_0 = \mathbf{A}_0$, and $N = t$ for guard channel scheme or $N = c$ for call thinning scheme. The vector $\boldsymbol{\pi}_k$ contains the probability of each state, which represents the number of new calls x_n, number of handoff calls x_h, state of arrival process a and phase of channel holding time p. This vector can be calculated by using the following equations:

$$\boldsymbol{\pi}_N \mathbf{E}_N = \mathbf{0} \tag{5.14}$$

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_{k+1} \mathbf{D}_{k+1} (-\mathbf{E}_k^{-1}), \quad 0 \leq k \leq N-1 \tag{5.15}$$

$$\sum_{k=0}^{N} \boldsymbol{\pi}_k \mathbf{e} = 1. \tag{5.16}$$

We can obtain the steady state probability from the last matrix at item N first and then use the recursive algorithm to obtain all other probabilities.

5.3.3 QoS Measures

We consider five QoS measures, namely, the new call blocking and handoff call dropping probabilities (indicating the proportion of failed incoming calls), average service bandwidth of the cell, user outage probability [53], and call degradation probability [58]. These QoS measures can be calculated from the steady state probability $\pi(x_n, x_h)$, which is the sub-vector of $\boldsymbol{\pi}$, by summing probabilities in all states of arrival and all phases of channel holding time at the same number of new calls and handoff calls.

- *New Call Blocking Probability*

  In guard channel scheme, an incoming new call is blocked if the number of ongoing calls is equal or greater than the threshold which is used to reserve bandwidth for the handoff calls. Thus, we have

  $$p_{nb}^{g} = \sum \pi(x_n, x_h)\mathbf{e}, \quad \lfloor \tfrac{t}{b_1} \rfloor \leq x_n + x_h. \tag{5.17}$$

  However, in call thinning scheme the new calls can be blocked randomly based on an acceptance probability (α) and is given as follows:

  $$p_{nb}^{t} = \sum (1 - \alpha(x_n + x_h))\pi(x_n, x_h)\mathbf{e} \tag{5.18}$$

  where $\alpha(x)$ is the acceptance probability when the number of ongoing calls is x.

- *Handoff Call Dropping Probability*

  Handoff call dropping probability is defined as the probability that handed over calls from the neighboring cells is dropped. In our model, the handoff calls are dropped only when all of the ongoing calls allocated with minimum bandwidth. Therefore, the handoff call dropping probability is given by

$$p_{hd} = \pi(x_n, x_h)\mathbf{e}, \quad \lfloor \tfrac{c}{b_1} \rfloor = x_n + x_h. \tag{5.19}$$

- *Average Bandwidth of the Cell*

  Average bandwidth of the cell indicates the expected bandwidth received by all ongoing calls and depends on the number of calls in the cell. If there are a few ongoing calls, the cell can allocate maximum bandwidth to every call. In contrast, if the number of calls increases, some of ongoing calls will be degraded, so that average bandwidth will be decreased. The average bandwidth of the cell can be obtained from

$$\bar{b}_{cell} = \bar{b}(x_n + x_h)\pi(x_n, x_h)\mathbf{e} \tag{5.20}$$

  where $\bar{b}(x)$ is an average bandwidth when the number of ongoing calls in the cell is x. Note that, in this model we assume that channel holding times are independent of bandwidth allocated to the call.

- *User Outage Probability*

  User outage probability is the probability that the cell has at least one call with bandwidth below the required level (b_{req}). The higher the user outage probability, the more is the possibility that a call will receive bandwidth below an acceptable level. The user outage probability can be obtained from

$$p_{out} = \sum_{\{x_n, x_h | \lfloor \frac{c}{b_{req}} \rfloor \leq x_n + x_h\}} \pi(x_n, x_h)\mathbf{e}. \tag{5.21}$$

- *Call Degradation Probability*

  Call degradation probability is the probability that the calls are allocated with bandwidth less than the desired value b_{req}. It indicates the level of call degradation and is given by

$$p_{deg} = \sum_{\forall x_n, \forall x_h} \pi(x_n, x_h) \frac{d(x_n + x_h)}{x_n + x_h} \tag{5.22}$$

where $d(x)$ is the number of degraded calls when the number of ongoing calls in the cell is x.

5.4 Numerical Results and Discussions

5.4.1 Parameter Setting

We assume that the available bandwidth per cell is 20 units, $B \in \{1, 2\}$, and $b_{req} = 2$. For the guard channel scheme 4 units of bandwidth are reserved for handoff calls. For the call thinning scheme, the acceptance probabilities are set as follows:

$$\alpha(x) = \begin{cases} 1, & x \in \{0, .., 13\} \\ -0.667x + 3.167, & x \in \{14, .., 18\} \\ 0, & x \in \{19, .., 20\}. \end{cases} \tag{5.23}$$

According to (5.23), the cell throttles the incoming new calls linearly if it becomes congested, (i.e., the number of ongoing calls lies between 14 to 18).

For arrival process, we use Markov Modulated Poisson Process (MMPP) which is a special type of MAP. In MMPP, each state of an arrival has its own mean arrival rate and the state transition changes the current arrival rate. We use the following parameters

$$\begin{aligned} \mathbf{C}_0 &= \begin{bmatrix} -1.02 & 0.02 \\ 0.1 & -4.6 \end{bmatrix}, \quad \mathbf{C}_N = \begin{bmatrix} 0.66 & 0 \\ 0 & 3.0 \end{bmatrix} \\ \mathbf{C}_H &= \begin{bmatrix} 0.34 & 0 \\ 0 & 1.5 \end{bmatrix} \end{aligned} \tag{5.24}$$

for which the mean rates for new calls, handoff calls, and total calls are 1.06, 0.54, and 1.6 calls per minute, respectively. Traces of handoff call arrival generated from this parameter setting are shown in Figure 5.3. The trace shows the burstiness of arrival rate which cannot be captured by normal Poisson process. Since the calculation of factors $\bigoplus_{i=1}^{k} \mathbf{S}_n$ and $\bigoplus_{i=1}^{l} \mathbf{S}_h$ is time consuming in case of phase type distribution, we assume exponential distribution with mean 5 minutes for channel holding time.

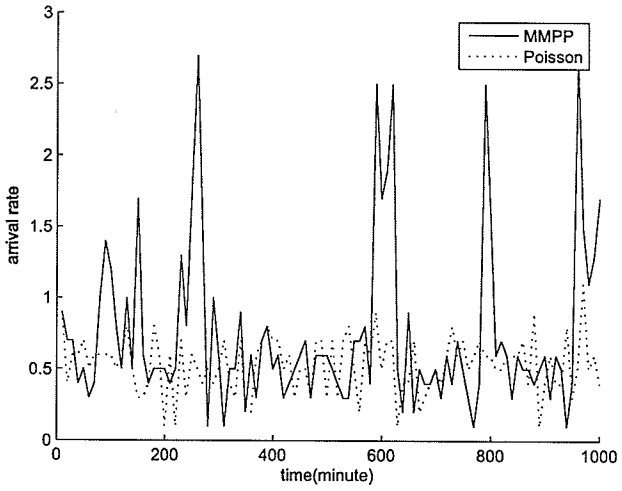


Figure 5.3. *Traces of handoff call arrival based on the example MAP parameters.*

5.4.2 Numerical Results

We obtain numerical results on the QoS measures from our model with MAP and phase-type distributions (MAP/PH). Based on the matrices in (5.24), we obtain results when the the mean arrival rate of new calls is varied (Figures 5.4-5.7).

We observe that with adaptive bandwidth allocation, handoff call dropping probability can be minimized by adjusting the bandwidth allocation of ongoing calls. Secondly, the arrival process and distribution of a channel holding time affect the QoS performances in ABA environment.

Thirdly, as expected, as the mean arrival rate increases, the new call blocking, handoff call dropping, user outage and call degradation probabilities increase while the average bandwidth of the cell decreases.

Moreover, with the same mean and distribution for call arrival and channel holding time, the observed QoS measures become different for the different CAC schemes. Although the average bandwidth of the cell and the outage probability are similar for both the guard channel and the call thinning schemes (Figures 5.5-5.6), the call degradation probabilities in case of call thinning scheme are much smaller than those

for the guard channel scheme(Figure 5.6). This difference comes from the fact that the call thinning scheme uses an acceptance probability to smooth the admitted new calls before the cell becomes congested. With the proper acceptance probability setting, the call thinning scheme can be used to minimize call degradation probability which is an important QoS measure of an ABA system.

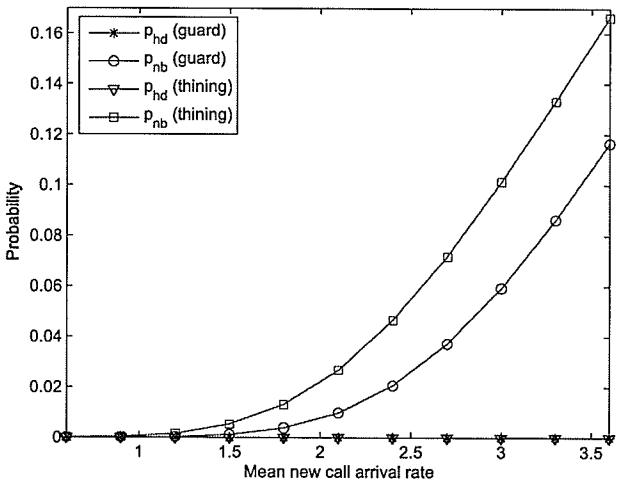


Figure 5.4. *New call blocking and handoff call dropping probabilities for varying mean of arrival rate of new calls.*

5.5 Chapter Summary

We have presented an analytical framework for adaptive bandwidth allocation in cellular mobile networks. We have used MAP for modeling call arrival and phase-type distribution for modeling channel holding time. We consider the fact that the distributions for channel holding time for the new calls and handoff calls can be different. Two types of CAC (i.e., guard channel and call thinning scheme) have been examined. Using the framework, various performance metrics (i.e., new call blocking probability, handoff call dropping probability, average bandwidth of the cell, user outage probability, and call degradation probability) have been derived.

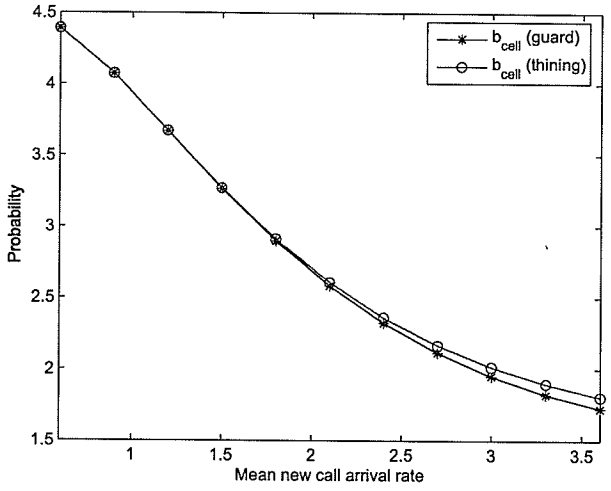


Figure 5.5. *Average bandwidth of the cell for MAP/PH model.*

The numerical results obtained from the model have shown that the ABA can minimize handoff call dropping probability, while some calls might experience service degradation below an acceptable level. We have observed that the call thinning scheme is able to smooth the rate of admitted new calls to avoid congestion in the cell.

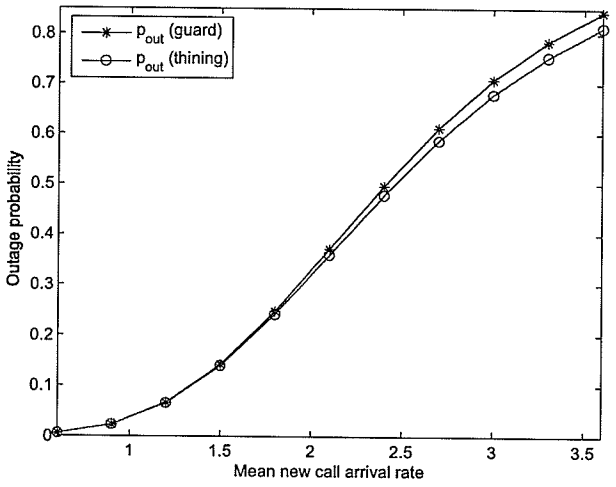


Figure 5.6. *User outage probability of the cell for MAP/PH model.*

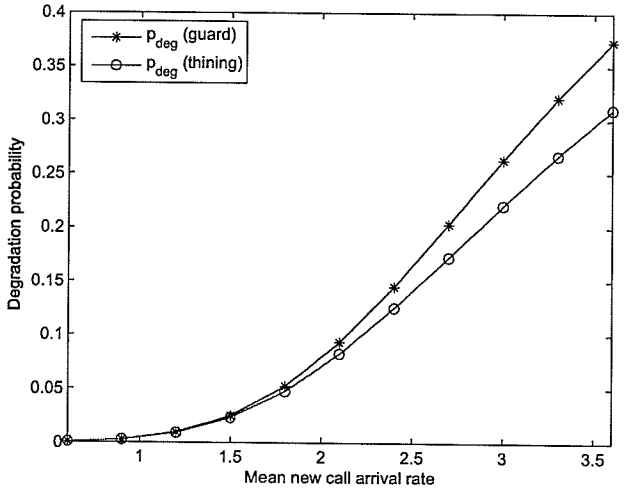


Figure 5.7. *Call degradation probability of the cell for MAP/PH model.*

Chapter 6

Performance Analysis and Adaptive Call Admission Control in Cellular Mobile Networks with Time-Varying Traffic

6.1 Introduction

A system exhibits transient behavior when it is not in the steady state, i.e., during the transition period until the system reaches an equilibrium state [54]. The transient behavior is important especially for a time-varying system since the system may never reach the steady state. Transient analysis based on Markov models is widely used to compute performance measures in reliability analysis ([55]-[56]). *Uniformization* is a well-known Markov-based transient analysis method with appealing properties such as numerical stability and controllable computation error. Transient analysis for cellular systems was used in [57] to investigate the time dependent packet-level performance measures (e.g., packet blocking probability).

However, in cellular mobile networks it is also crucial to analyse call-level quality of service (QoS) under different bandwidth allocation schemes (e.g., static and adaptive) and CAC strategies. In case of static bandwidth allocation, the bandwidth allocated to a call remains fixed over the entire connection period. Static bandwidth allocation is used mainly for voice and constant bit rate services. Alternatively, adaptive bandwidth allocation (ABA) allows the bandwidth of ongoing calls to be degraded to

accommodate handoff calls or more new calls. ABA techniques can be used for multimedia services with flexible QoS requirements in which the transmission rate can be varied by adjusting the encoding scheme. By using ABA, the number of blocked calls can be minimized and the resource utilization can be increased. By using proper CAC policies, call-level QoS for both static and adaptive bandwidth allocation systems can be maintained at the acceptable level.

Performance analysis of a CAC scheme based on the *guard channel concept* was done in [3]. Instead of using one value of threshold for admission control of the new calls, the concept of *fractional guard scheme* was introduced in [7] (where a new call is admitted with a certain probability) and the various types of CAC policies were analysed in [22]. These models considered *static* bandwidth allocation and derived the QoS performance measures including new call blocking and handoff call dropping probabilities. Analytical models for performance evaluation of ABA were proposed in [58]-[59] and QoS measures related to degradation of the service quality were obtained. However, most of these models in the literature dealt with the steady state behavior only and the models were developed for using off-line to obtain the QoS performances under pre-determined system parameters.

We propose an analytical model for transient performance analysis of both static and adaptive bandwidth allocation in cellular mobile networks under time-varying traffic pattern. Based on the analytical model, using an optimization approach, we also propose an on-line adaptive CAC mechanism which dynamically adjusts the CAC threshold for new calls. Typical numerical results are presented and validated by simulation.

The rest of this chapter is organized as follows. Section 6.2 describes the system models for both static and adaptive bandwidth allocation under time-varying traffic. The transient analysis method (i.e., uniformization) is presented in Section 6.3. Section 6.4 presents an on-line adaptive CAC method based on numerical optimization. The results from the analytical models and simulations are presented in Section 6.5. Conclusions are stated in Section 6.6.

6.2 System Model Under Time-Varying Traffic

6.2.1 Call Arrival and Bandwidth Allocation

We assume that both new and handoff call arrivals follow Poisson process and the channel holding time is exponentially distributed. However, the arrival rates for both new calls and handoff calls as well as the channel holding time are time-varying. We consider both static (i.e., the bandwidth allocated to each call is static and is determined at the call initiation time) and adaptive bandwidth allocation (i.e., the bandwidth allocated to the ongoing calls can be adjusted according to the state of the network).

When a call arrives, a CAC policy is responsible to make a decision on accepting or rejecting an incoming call. Since the users are more sensitive to dropping an on-going call than blocking a newly initiated call, some portion of bandwidth is reserved for handoff calls. We assume a *guard channel scheme* [3] in which out of the C bandwidth units in a cell, $C - K$ units are reserved for handoff calls in order to minimize the handoff call dropping probability.

6.2.2 Analytical Model for Static Bandwidth Allocation Under Time-Varying Traffic

The continuous finite state Markov chain model for static bandwidth allocation with guard channel-based CAC (i.e., new calls are accepted only if the amount of bandwidth used by ongoing calls is less than the threshold $K(t)$) is shown in Figure 6.1. Note that, the arrival rate of new calls $\lambda_N(t)$ and handoff calls $\lambda_H(t)$ as well as the channel holding time for both types of calls $1/\mu(t)$ are time-varying. Moreover, the amount of bandwidth reserved for handoff calls is adjustable in each time interval.

The state space of this model is $\Delta = \{(x),\ 0 \leq x \leq C\}\}$, where x represents the number of ongoing calls. The infinitesimal generator matrix for this Markov process can be expressed by (6.1) as follows:

$$\mathbf{Q}(t) = \begin{bmatrix} q_{0,0}(t) & q_{0,1}(t) & & & & \\ q_{1,0}(t) & q_{1,1}(t) & q_{1,2}(t) & & & \\ & q_{i,i-1}(t) & q_{i,i}(t) & q_{i,i+1}(t) & & \\ & & \ddots & \ddots & \ddots & \\ & & & q_{C,C-1}(t) & q_{C,C}(t) \end{bmatrix} \tag{6.1}$$

where

$$q_{i,i+1}(t) = \begin{cases} \lambda_N(t) + \lambda_H(t), & i \in \{0, 1, \ldots, K-1\} \\ \lambda_H(t), & i \in \{K, \ldots, C-1\} \end{cases} \tag{6.2}$$

$$q_{i,i}(t) = \begin{cases} -\lambda_N(t) - \lambda_H(t) - i\mu(t), & i \in \{0, 1, \ldots, K-1\} \\ -\lambda_H(t) - i\mu(t), & i \in \{K, \ldots, C-1\} \\ -C\mu(t), & i = C \end{cases} \tag{6.3}$$

$$q_{i,i-1}(t) = \; i\mu(t), \quad i \in \{1, \ldots, C\}. \tag{6.4}$$

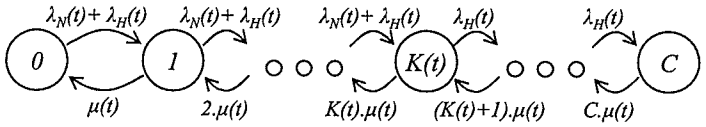


Figure 6.1. *Markov chain model for static bandwidth allocation under time-varying traffic.*

Let $\pi_r(t) = \Pr\{x(t) = r\}$ be the transient probabilities that the system stays in state r at time t conditioned on the initial state. We can obtain the new call blocking probability $p_{nb}^{static}(t)$ and the handoff call dropping probability $p_{hd}^{static}(t)$, at time t as follows:

$$p_{nb}^{static}(t) = \sum_{r=K}^{C} \pi_r(t) \tag{6.5}$$

$$p_{hd}^{static}(t) = \pi_C(t). \tag{6.6}$$

6.2.3 Analytical Model for Adaptive Bandwidth Allocation Under Time-Varying Traffic

Let us assume that a call can be allocated bandwidth from the set $B = \{b_1, b_i, \ldots, b_n\}$ in which $b_i < b_i+1$, $i \in \{1, \ldots, n-1\}$. Then, the minimum and maximum bandwidth for a call are b_1 and b_n, respectively. However, because of the adaptation algorithm, bandwidth of the call might be degraded below the target level which is defined as b_{tar} [60].

We consider the fairness-based bandwidth adaptation algorithm (BAA) proposed in [17]. With threshold K, if a new call arrives and the bandwidth used by ongoing calls is larger than K, the incoming call is rejected. Otherwise, the cell tries to allocate the maximum possible bandwidth to the incoming call. However, if the available bandwidth is not enough, the BAA is executed. The BAA degrades the bandwidth of an ongoing call (randomly chosen) with the current maximum bandwidth by one step. This degradation process is iteratively performed until the free bandwidth is equal to the current minimum bandwidth of all ongoing calls. The incoming call is accepted and allocated with the free bandwidth. However, if every call has the minimum bandwidth b_1 (and hence no bandwidth adaptation is performed), the incoming call is rejected.

In the event that a call is terminated or a handoff occurs, the bandwidth of that call is released. Then, the bandwidth of some ongoing calls are upgraded. The upgrade scheme randomly upgrades a call with the current minimum bandwidth until all released bandwidth is allocated or all of the ongoing calls have the maximum bandwidth (b_n).

Considering the fact that the distributions for the channel holding time of new calls and handoff calls can have different means [22], especially in the environment with a variety of mobile platforms [61], the system state should describe both the number of new calls and handoff calls, so that the state space can be expressed as

$$\Delta = \{(N, H) | 0 \leq N \leq C, 0, \leq H \leq C\} \tag{6.7}$$

where N and H represent the number of ongoing new calls and handoff calls, respectively. The corresponding Markov chain is shown in Figure 6.2, where the new call and handoff call arrival rates at time t are denoted by $\lambda_N(t)$ and $\lambda_H(t)$, respectively.

The distributions of channel holding time for new calls and handoff calls have means $1/\mu_N(t)$ and $1/\mu_H(t)$, respectively. Note that, in Figure 6.2, the states shown in dotted-lines are unreachable states and have transition rate equal to zero. Normally, these states can be eliminated from the model. However, to simplify the transient analysis and to keep the size of the infinitesimal generator matrix $\mathbf{Q}(t)$ constant at all time these states are retained.

The generator matrix for this model is in block tri-diagonal form and can be expressed as

$$\mathbf{Q}(t) = \begin{bmatrix} \mathbf{A}_0(t) & \mathbf{B}_0(t) & & & & \\ \mathbf{D}_1(t) & \mathbf{A}_1(t) & \mathbf{B}_1(t) & & & \\ \ddots & \ddots & \ddots & & & \\ & \mathbf{D}_j(t) & \mathbf{A}_j(t) & \mathbf{B}_j(t) & & \\ & \ddots & \ddots & \ddots & & \\ & & \mathbf{D}_{C-1}(t) & \mathbf{A}_{C-1}(t) & \mathbf{B}_{C-1}(t) & \\ & & & \mathbf{D}_C(t) & \mathbf{A}_C(t) \end{bmatrix} \tag{6.8}$$

where each row in matrix $\mathbf{Q}(t)$ represents the number of new calls. The matrix $\mathbf{D}_j(t)$, representing departure of new call, is diagonal. Let $\mathbf{d}^j_{l,l}(t)$ denote the diagonal elements of matrix $\mathbf{D}_j(t)$ when the number of new calls is j and the number of handoff calls is l at time t. We have $\mathbf{d}^j_{(l,l)}(t) = j\mu_N(t),\ j \leq K$. The diagonal matrix $\mathbf{B}_j(t)$ represents the arrival of new calls and the diagonal elements of this matrix can be expressed as $\mathbf{b}^j_{(l,l)}(t) = \lambda_N(t),\ j + l < K(t)$. The rows of the matrices $\mathbf{A}_j(t),\ j \in \{0, \ldots, C\}$, $\mathbf{B}_j(t),\ j \in \{0, \ldots, C-1\}$, and $\mathbf{D}_j(t),\ j \in \{1, \ldots, C\}$ represent the number of handoff calls. Matrix $\mathbf{A}_j(t)$ is the Markov process for handoff calls at time t and is defined in (6.9),

$$\mathbf{A}_j(t) = \begin{bmatrix} \mathbf{a}^j_{0,0}(t) & \mathbf{a}^j_{0,1}(t) & & & & & \\ \mathbf{a}^j_{1,0}(t) & \mathbf{a}^j_{1,1}(t) & \mathbf{a}^j_{1,2}(t) & & & & \\ & \mathbf{a}^j_{l,l-1}(t) & \mathbf{a}^j_{l,l}(t) & \mathbf{a}^j_{l,l+1}(t) & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \mathbf{a}^j_{C-j-1,C-j-2}(t) & \mathbf{a}^j_{C-j-1,C-j-1} & \mathbf{a}^j_{C-j-1,C-j} \\ & & & & \mathbf{a}^j_{C-j-1,C-j} & \mathbf{a}^j_{C-j,C-j} \end{bmatrix} \tag{6.9}$$

$$\mathbf{a}^{j}_{(l,l)}(t) = \begin{cases} -j\mu_N(t) - \lambda_N(t) - \lambda_H(t) & l = 0 \\ -j\mu_N(t) - \lambda_N(t) - \lambda_H(t) - l\mu_H(t) & 0 < l < K - j, l + j < K \\ -j\mu_N(t) - \lambda_H(t) - l\mu_H(t) & 0 < l < K - j, l + j \geq K \\ -j\mu_N(t) - l\mu_H(t) & l = K - j \end{cases} \tag{6.10}$$

where $\mathbf{a}^{j}_{l,l+1}(t) = \lambda_H(t)$ represents the arrival and $\mathbf{a}^{j}_{(l,l-1)}(t) = l\mu_H(t)$ represents the departure of handoff calls and $a^{j}_{(l,l)}(t)$ is given by (6.10).

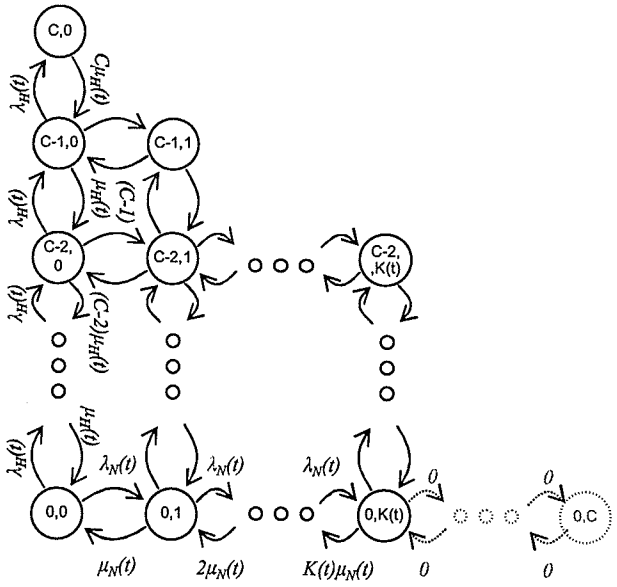


Figure 6.2. *Markov chain model for adaptive bandwidth allocation under time varying traffic in which the name of the state denote the number of handoff and new calls, respectively.*

Let $\pi_{(u,v)}(t) = \Pr\{n(t) = u, h(t) = v\}$ be the transient probabilities that the system has u new calls and v handoff calls at time t conditioned on the initial state. We can obtain the QoS measures including new call blocking probability $p^{aba}_{nb}(t)$, handoff call dropping probability $p^{aba}_{hd}(t)$, and call degradation probability $p_{deg}(t)$ (defined as the probability that any calls are allocated with bandwidth less than target level b_{tar})

[58] at time t as follows:

$$p_{nb}^{aba}(t) = \sum_{\{(u,v)|\lfloor \frac{K}{b_1} \rfloor \leq u+v\}} \pi_{(u,v)}(t) \tag{6.11}$$

$$p_{hd}^{aba}(t) = \pi_{(u,v)}(t), \qquad \left\lfloor \frac{C}{b_1} \right\rfloor = u + v \tag{6.12}$$

$$p_{deg}(t) = \sum_{u,v} \pi_{(u,v)}(t) \frac{d(u+v)}{u+v} \tag{6.13}$$

where $d(x)$ is the number of degraded calls when the number of ongoing calls in the cell is x.

6.3 Transient Analysis

In the model for static bandwidth allocation under time-varying traffic, let $\pi_{r,r0}(0,t) = \Pr\{x(t) = r | x(0) = r0\}$ be the probability that the system is in state r at time t given the initial state $r0$ at time 0. If we consider the system at time t, we have $\pi_r(t) = \Pr\{X(t) = r\}$ which is conditioned on the initial state. These transient probabilities are the elements of $\boldsymbol{\pi}(t) = [\pi_0(t), \pi_1(t), \ldots, \pi_C(t)]$ which is a row vector of the state probability distribution at time t. To calculate $\boldsymbol{\pi}(t)$, we have to solve the Kolmogorov-forward equations in (6.14) which can be expressed in the matrix form as in (6.15).

$$\begin{aligned}
\frac{\mathrm{d}\pi_0(t)}{\mathrm{d}t} &= -q_{0,0}(t)\pi_0(t) + q_{0,1}(t)\pi_0(t), \\
\frac{\mathrm{d}\pi_i(t)}{\mathrm{d}t} &= -\pi_i(t)q_{i,i}(t) + \pi_i(t)(q_{i,i+1}(t) + q_{i,i-1}(t)), \quad i \in \{1, \ldots, C-1\} \\
\frac{\mathrm{d}\pi_C(t)}{\mathrm{d}t} &= -q_{C,C}(t)\pi_C(t) + q_{C,C-1}(t)\pi_C(t),
\end{aligned} \tag{6.14}$$

$$\frac{\mathrm{d}\boldsymbol{\pi}(t)}{\mathrm{d}t} = \boldsymbol{\pi}(t)\mathbf{Q}(t). \tag{6.15}$$

By solving (6.15) based on the assumption that the elements in the generator matrix are constant in time interval $[0,t)$ or $\mathbf{Q}(t) = \mathbf{Q}$, we have

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0).e^{\mathbf{Q}.t}. \tag{6.16}$$

The *Uniformization* technique [81] (also called *Jensen's Method* or *Randomization Method*) is a general and efficient way to obtain the solution from this equation. In this technique, the new transition probability matrix $\mathbf{P}$ is defined as

$$\mathbf{P} = \frac{\mathbf{Q}}{\Lambda} + I \tag{6.17}$$

where $\Lambda \geq \min_i (|q_{i,i}|)$. In other words, Λ is greater than or equal to the absolute value of the minimum diagonal element in $\mathbf{Q}$. Therefore, we have

$$\begin{aligned} \pi(t) &= \pi(0)e^{(\mathbf{P}-I)\Lambda t} = \pi(0)e^{\mathbf{P}\Lambda t}e^{-\Lambda t} \\ &= \pi(0)\sum_{n=0}^{\infty}\mathbf{P}^n\frac{(\Lambda t)^n}{n!}e^{-\Lambda t}. \end{aligned} \tag{6.18}$$

However, to reduce the computation time, the limit of summation δ is to be set such that the truncation error remains below ϵ. That is,

$$\pi(t) = \pi(0)\sum_{n=0}^{\delta}\mathbf{P}^n\frac{(\Lambda t)^n}{n!}e^{-\Lambda t} \tag{6.19}$$

where

$$1 - e^{-\Lambda t}\sum_{n=0}^{\delta}\frac{(\Lambda t)^n}{n!} \leq \epsilon. \tag{6.20}$$

In the case that the call arrival rate and the mean channel holding time are time-varying, let $\mathbf{P}(t_i)$ be the transition probability matrix of the process $\mathbf{Q}(t_i)$, whose elements are constant in $[t_i, t_{i-1})$. After uniformization we obtain the transient probability at time t_i conditioned on the previous time t_{i-1} as follows:

$$\pi(t_i) = \pi(t_{i-1})\sum_{n=0}^{\delta}\mathbf{P}^n(t_i)\frac{(\Lambda t)^n}{n!}e^{-\Lambda t}. \tag{6.21}$$

The same technique can be used for transient analysis in the case of adaptive bandwidth allocation.

The challenge of using uniformization is how to choose the length of iteration to determine the transient behavior. If the length is short, the model may provide accurate results but the total computation overhead would be high. Techniques such as *finite-difference methods* [54] and *adaptive uniformization* [63] were proposed to minimize the computation time. However, to make the analytical model simple, we use fixed length of iteration which is short enough to obtain accurate results.

6.4 Adaptive CAC

In this section, we propose an adaptive CAC policy based on the transient analysis. In our model an incoming handoff call is accepted if there is enough bandwidth available (in case of static bandwidth allocation) or there are some calls which can be degraded (in case of adaptive bandwidth allocation). In contrast, admission of new calls is controlled by the threshold K which is calculated dynamically by using numerical optimization.

The objective of the proposed adaptive CAC policy is to minimize the increase in call-level QoS measures such as new call blocking probability (under both static and adaptive bandwidth allocation) during successive adaptation intervals (e.g., t_{s-1} and t_s) by adjusting the threshold K. The *near-optimal* solution for the threshold $K(t_s)$ is obtained by using enumeration method. To obtain the solution, we assume that the arrival rates of new calls and handoff calls during the next adaptation interval are known in advance. Estimation techniques such as Kalman filtering [31] can be applied to project these arrival rates. The solution of $K(t_s)$ is obtained by solving the optimization problem based on matrix $\mathbf{Q}(t_s)$.

In case of static bandwidth allocation there are two main QoS measures: new call blocking and handoff dropping probabilities. Since the handoff calls have to be prioritized over new calls, the handoff call dropping probability is to be kept below the acceptable level τ_{hd}^{static}, while the increase in new call blocking probability should be minimized. Hence, the optimization problem is formulated as follows:

$$\text{minimize} \quad p_{nb}^{static}(t_s, K(t_s)) \tag{6.22}$$

subject to:

$$p_{hd}^{static}(t_s, K(t_s)) \leq \tau_{hd}^{static} \tag{6.23}$$

where $p_{nb}^{static}(t_s, K(t_s))$ and $p_{hd}^{static}(t_s, K(t_s))$ are defined as functions of time t_s and guard channel $K(t_s)$.

In case of adaptive bandwidth allocation, the handoff call dropping probability is still the most important QoS measure. However, the new call blocking probability can be improved by allowing the bandwidth of ongoing calls to be degraded. Therefore, the handoff call dropping probability and new call dropping probabilities are maintained below the acceptable levels τ_{hd}^{aba} and τ_{nb}^{aba}, respectively, and the increase in call

degradation probability should be minimized. Therefore, the optimization problem in this case is formulated as follows:

$$\text{minimize} \quad p_{deg}(t_s) \tag{6.24}$$

subject to

$$p_{hd}^{aba}(s_i) \leq \tau_{hd}^{aba} \quad \text{and} \quad p_{nb}^{aba}(s_i) \leq \tau_{nd}^{aba}. \tag{6.25}$$

Alternatively, the constraint in the above optimization problem can be modified if we want to maintain the call degradation probability below a certain level τ_{deg}, in which case the objective would be to minimize the increase in new call blocking probability.

6.5 Numerical and Simulation Results

We assume that there are 40 bandwidth units per cell and each call uses 2 units. Also, there is no ongoing call at time zero i.e., $\pi(0) = [1, 0, \ldots, 0]$. The length of iteration for the transient analysis is set to 1 minute. The arrival rate and channel holding time of new calls and handoff calls as well as the threshold K are set according to the evaluation scenarios.

For the case with static bandwidth allocation, we set the threshold K to 36 bandwidth units. That is, if the bandwidth used by ongoing calls in the cell is equal or greater than 36 units , an incoming new call will be blocked. The new call and the handoff call arrival rates are 3 and 2 calls per minute, respectively. The average channel holding time for both types of calls is 5 minutes. Figure 6.3 shows the new call blocking and the handoff call dropping probabilities from both steady and transient state analyses.

We observe that the system spends around 12 minutes in the transient state before reaching the steady state where p_{nb}^{static} and p_{hd}^{static} are 0.4829 and 0.071, respectively. The simulation results confirm the convergence of the new call blocking and handoff call dropping probabilities. It is evident that the steady state analysis might not give the accurate results if the system is not in steady state.

In case of time-varying arrival rates of new calls and handoff calls (for the trace shown in Figure 6.4), the new call blocking and the handoff call dropping probabilities are shown in Figures 6.5-6.6 when the threshold is fixed at $K = 36$. The same performance measures are shown in Figures 6.7-6.8 when the threshold $K(t)$ is adaptively

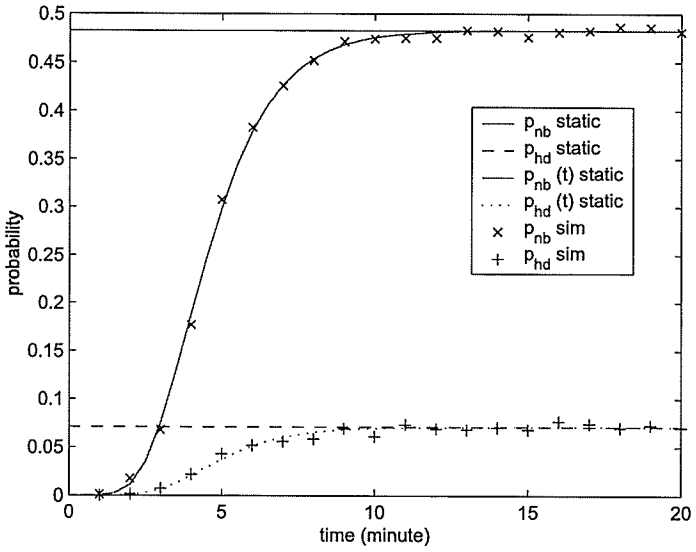


Figure 6.3. *New call blocking and handoff call dropping probabilities from transient analysis ($p_{nb}^{static}(t)$ and $p_{hd}^{static}(t)$), steady state analysis (p_{nb}^{static} and p_{hd}^{static}) and simulation (p_{nb} sim and p_{hd} sim).*

adjusted (as shown in Figure 6.9) according to the proposed adaptive CAC. In this case, we set the length of adaptation interval to 1 minute and $\tau_{hd}^{static} = 0.05$.

We observe that with dynamic adjustment of the threshold the handoff calls dropping probability remains below the acceptable level (0.05) and also the new call blocking probability decreases during some periods (e.g., during periods 15-25, 35-25, and 55-60 in Figure 6.5 and Figure 6.7). The reason is that when the arrival rate of handoff calls decreases, some of the reserved bandwidth $C - K(t)$ can be yielded to the new calls, so that the blocking probability becomes smaller. We also observe that most of the time the results from transient analysis agrees with simulation results rather than with those from steady state analysis.

For adaptive bandwidth allocation, we assume that there are 30 bandwidth units per cell, $B \in \{1, 2\}$ and $b_{tar} = 2$. The arrival rates are as shown in Figure 6.4 and the average channel holding times of new call and handoff call are 4 and 6 minutes, respectively. Figure 6.10 shows the results for the proposed adaptive CAC

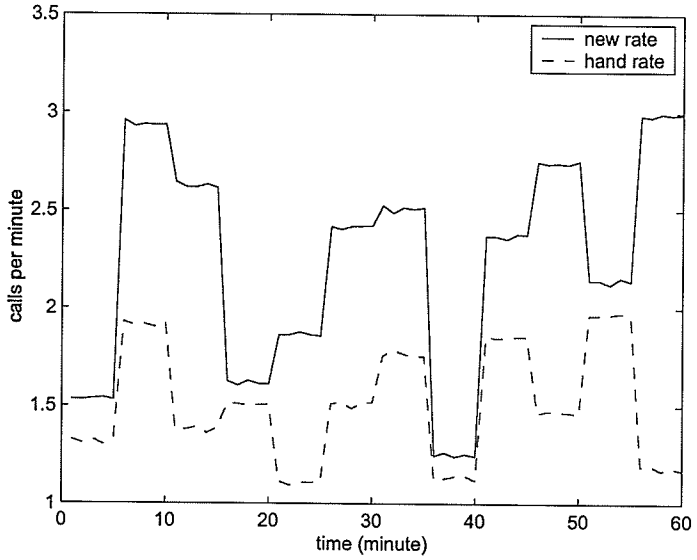


Figure 6.4. *Traces of new call and hand off call arrival rates.*

with acceptable handoff call dropping probability of 0.02 ($\tau_{hd}^{aba} = 0.02$) and new call blocking probability of 0.1 ($\tau_{nb}^{aba} = 0.1$). The objective here is to minimize the increase in call degradation probability.

Figure 6.11 shows typical results for the adaptive CAC with constrained call degradation probability of 0.35 ($\tau_{deg} = 0.35$). The objective function in this case is to minimize the increase in the new call blocking probability. We observe that with ABA the new call blocking and handoff call dropping probabilities are smaller than those in the static bandwidth allocation case because of the use of the bandwidth adaptation algorithm.

With adaptive CAC, the handoff call dropping, new call blocking, and call degradation probabilities can be controlled by applying numerical optimization technique with the transient analysis model. However, there are some discrepancies between the analytical and the simulation results. Our hypothesis is that the length of iteration of uniformization might not be suitably chosen. To alleviate this problem, some advanced technique (e.g., adaptive uniformization) will be investigated in our future work.

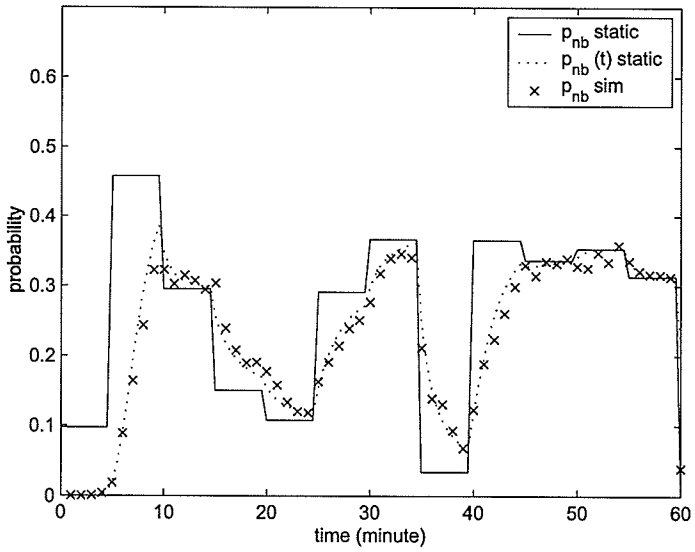


Figure 6.5. *New call blocking probability from steady state analysis (p_{nb}^{static}), transient analysis ($p_{nb}^{static}(t)$), and simulations (p_{nb} sim) when the threshold for new calls is fixed.*

6.6 Chapter Summary

We have presented a framework for analyzing the call-level transient performances of cellular mobile networks under time-varying traffic pattern. The *uniformization* technique has been used to obtain the call-level QoS parameters under both static and adaptive bandwidth allocation. Based on the analytical model, we have also developed a threshold-based on-line adaptive CAC scheme. Numerical results from both the steady state and the transient analyzes have been compared. The results have shown that the system spends some period of time in transient state before reaching the steady state, and therefore, transient analysis is needed to obtain the accurate QoS performances during certain period of time. In addition, our proposed adaptive CAC can successfully control the QoS performances at the desired level under time-varying traffic, even thought the performances of our proposed guard channel adaptation is slightly better than that of static scheme.

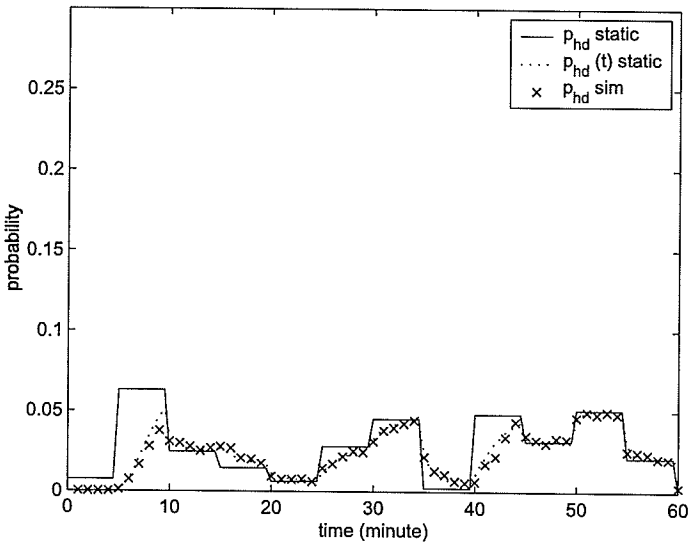


Figure 6.6. *Handoff call dropping probability from steady state analysis (p_{hd}^{static}), transient analysis ($p_{hd}^{static}(t)$), and simulation (p_{hd} sim) when the threshold for new calls is fixed.*

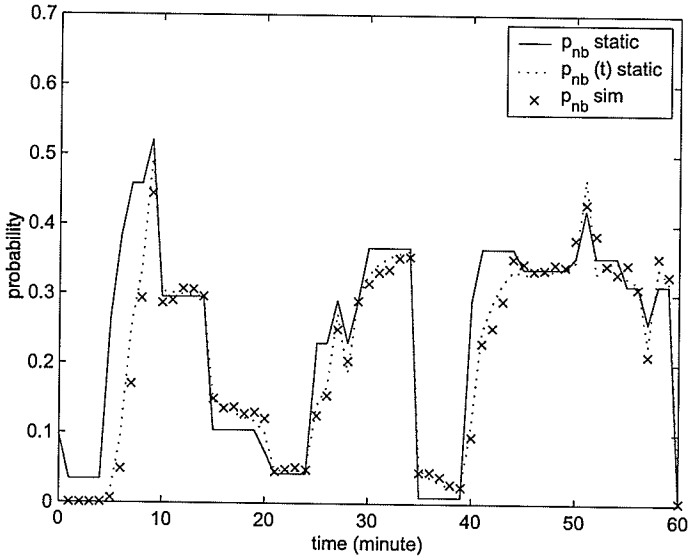


Figure 6.7. *New call blocking probability from steady state analysis (p_{nb}^{static}), transient analysis ($p_{nb}^{static}(t)$), and simulations (p_{nb} sim) for the proposed adaptive CAC.*

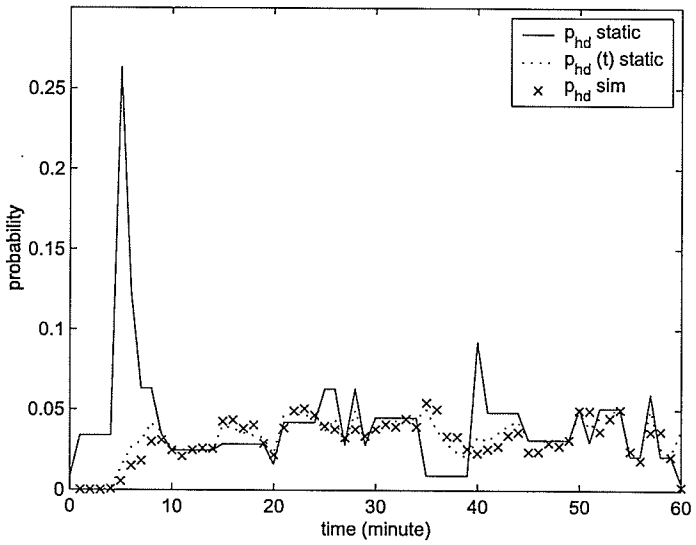


Figure 6.8. *Handoff call dropping probability from steady state analysis (p_{hd}^{static}), transient analysis ($p_{hd}^{static}(t)$), and simulations (p_{hd} sim) for the proposed adaptive CAC policy.*

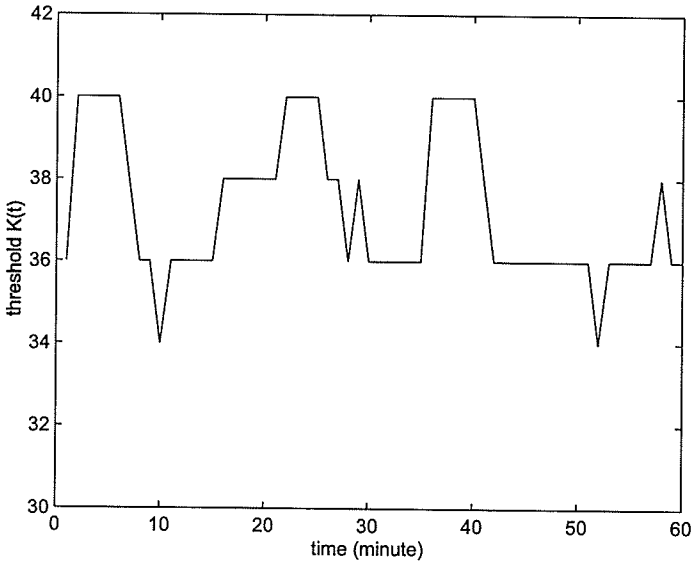


Figure 6.9. *Adjustment of threshold $K(t)$ under adaptive CAC (for static bandwidth allocation).*

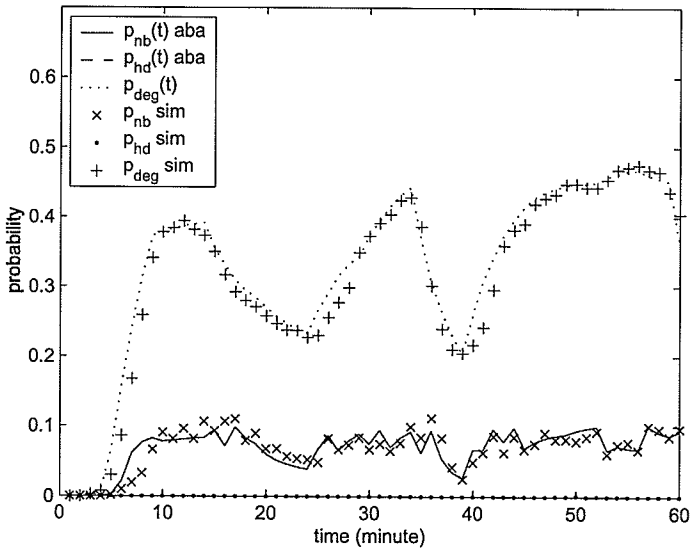


Figure 6.10. *QoS measures for ABA from transient analysis and simulation.*

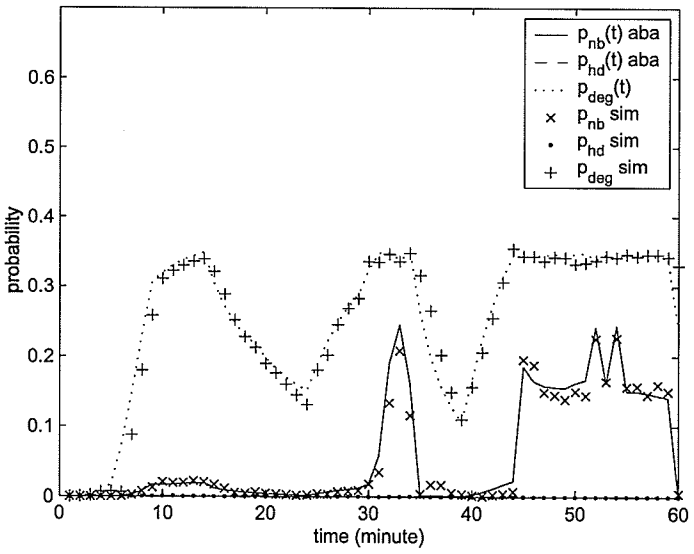


Figure 6.11. *QoS measures for ABA with constrained call degradation probability of 0.35.*

Chapter 7

A Novel Analytical Framework for Integrated Cross-Layer Study of Call-Level and Packet-Level QoS in Mobile Wireless Multimedia Networks

7.1 Introduction

To guarantee quality-of-service (QoS) in a wireless mobile multimedia network, CAC is crucial to decide whether an incoming connection can be accepted or not. This decision is made based on the QoS requirements of the users and the state of the network. In such a network, the call level QoS metrics (e.g., new call blocking and handoff call dropping probabilities) as well as the packet-level QoS metrics (e.g., packet dropping probability and packet delay) need to be maintained at the desired level. Again, in a wireless multimedia network adaptive channel (or bandwidth) allocation (ACA) can be used to maximize the utilization of radio channels while maintaining the QoS requirements of the calls at the acceptable levels.

In this chapter, we present an analytical framework to investigate the impacts of CAC and ACA on the call-level and the packet-level performances in a multimedia mobile wireless network using adaptive modulation in the physical layer. We consider three types of traffic, namely, real-time (e.g., voice), non-real-time and best-

effort traffic (e.g., file transfer), which are modeled by Markov modulated Poisson process (MMPP), Poisson process, and batch transmission process, respectively. In our system model, a guard channel scheme [3] is used for CAC to prioritize handoff calls over new calls, and a fairness-based ACA algorithm is applied to allocate available channels to the ongoing calls. In the data link layer, finite drop tail queueing is used for non-real-time traffic. In the physical layer, we use finite state Markov channel (FSMC) model to capture adaptive modulation and coding (AMC) in a Nakagami fading channel. AMC is considered here to enhance the transmission rate by changing modulation level according to the channel quality, i.e., the signal-to-noise ratio (SNR) at the receiver.

Using the analytical model, various packet-level QoS performance measures of non-real-time traffic (e.g., packet dropping probability due to the lack of buffer space, average queue length, average delay, throughput, and delay distribution) are obtained. For real-time traffic, packet loss due to delay transmission is obtained. For best-effort traffic, we derive the delay distribution corresponding to transmission of fixed-sized files. The impact of user mobility on the packet-level performances is also demonstrated. The presented analytical model is validated by extensive simulations. Also, we demonstrate the application of the proposed model for obtaining the system parameter settings so that a target level of QoS can be achieved.

7.2 Related Work

Analytical models for performance evaluation of different CAC algorithms in cellular networks were proposed in [22]. In [59] and [64], analytical models for call-level performance evaluation under adaptive channel allocation were proposed. In [17], in-call performance measures (i.e., *degradation ratio* and *upgrade/degrade frequency*) under adaptive bandwidth allocation were derived. However, an integrated evaluation of call-level and packet-level QoS measures in presence of CAC and adaptive channel allocation was not performed.

For a wireless system, packet-level performance at the radio link level under different radio resource (e.g., transmission power and rate) management strategies were analyzed in the literature. Although the general problem of radio resource management

was studied in [65], the radio link level queueing aspects were ignored. Radio resource management techniques for multiservice code division multiple access (CDMA) networks were proposed in [78]. To ensure packet-level QoS in a data-oriented CDMA network, traffic scheduling schemes based on generalized processor sharing were used in [79].

An analytical model for both call-level and packet-level performance evaluation in DS (Direct Sequence)-CDMA networks was proposed in [80]. However, this model considered only a single-packet buffer and ignored the radio link level queueing dynamics. Again, the performance analysis was not exact.

In [29], a Markov-based model was presented to analyze the radio link level packet dropping process under automatic repeat request (ARQ)-based error control. A model for analyzing radio link level delay (i.e., queueing delay, transmission delay, and resequencing delay) for selective repeat ARQ (SR-ARQ) was presented in [66]. However, all these works considered only the packet-level performances under single user system with single transmission channel.

To enhance the spectrum efficiency, adaptive modulation is commonly used in 2.5G/3G wireless systems. In such a system, the transmission rate can be increased by adaptively adjusting the modulation level according to the channel quality. In [25], an analytical model to derive packet loss rate, average throughput and average spectral efficiency under adaptive modulation was presented. Although queueing analysis for packet-level performance evaluation under adaptive modulation in a multiple-user system was presented in [67], call-level parameters and their impacts on packet-level and call-level performances were not investigated.

Wireless multimedia networks need to accommodate different types of traffic with different QoS requirements. In [69], an analytical framework for rate adaptive encoding of MPEG video was presented, however, no queueing analysis for packet-level and call-level performance evaluation was performed. In [70], a resource management strategy for multimedia wireless networks was proposed and both call-level and user-level QoS (in terms of allocated bandwidth) were investigated. However, this work did not consider the packet-level QoS as well as the impact of fading channel.

Rate control is widely used in wired networks to control the packet generation at the traffic source to avoid network congestion. For multimedia transmission, the

impacts of rate control on the QoS performance was analyzed in [71]. In [72], a traffic shaping scheme based on token bucket was proposed to maintain QoS in the UMTS networks. However, the impact of multi-rate transmission at the physical wireless channel (achieved through adaptive modulation) as well as the queueing performances were not investigated.

7.3 System Model and Assumptions

We consider uplink transmissions from mobiles to a base station in a cellular architecture with multiple channels available in a cell for ongoing calls (Figure 7.1). The available channels at the base station are partitioned into three groups (i.e., C_{rt}, C_{nrt} and C_{be}) to be used by real-time (strictly delay-sensitive but loss-insensitive), non-real-time (loss-sensitive but moderately delay-sensitive), and best-effort services (loss-sensitive)[1], respectively. That is, new calls and handoff calls of certain service class are allocated channels from the corresponding pool of channels. Since each service class uses different pool of channels, states of of one service class will no affect other service classes.

For new calls and handoff calls, CAC and ACA methods are used at the base station for each service class separately. At the network layer[2], while the CAC algorithm is used to determine whether an incoming call can be accepted or not, the ACA algorithm is responsible for allocating the available channels among the calls. Since the CAC and ACA of all service classes are performed in the same manner, the analytical model of call-level are applicable for all service classes while the packet arrival processes are different.

As we will see later in this chapter, these two methods in fact operate in a complementary fashion. At the physical layer, we consider an FSMC model for channel fading and multi-rate transmission in which the transmission rate can be adjusted according to the instantaneous SNR and the target bit error rate (BER).

[1]We will show later in this chapter how an optimal partitioning of the channels at the base station can be obtained to accommodate different service classes.

[2]In this chapter, this refers to any sub-layer/layer above layer-2 in the wireless protocol stack.

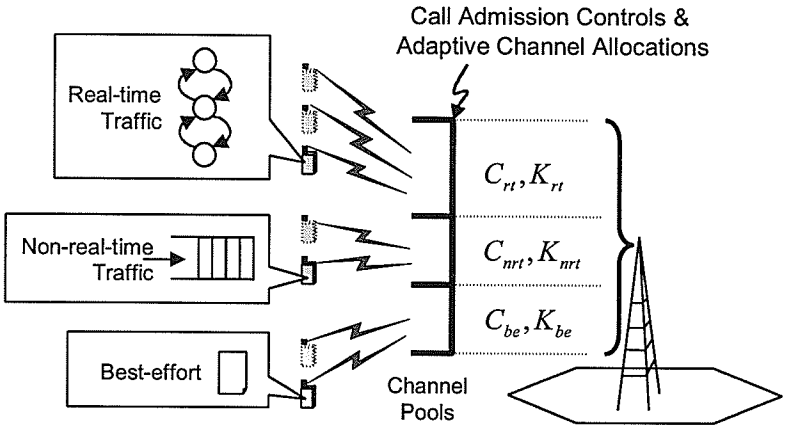


Figure 7.1. *System Model.*

7.3.1 Wireless Channel Model and Multi-rate Transmission

An FSMC model is a useful model for analyzing radio channel with non-independent fading (and hence bursty channel errors). A slowly varying Nakagami-m fading channel is represented by the FSMC model and each state of the FSMC corresponds to one transmission mode for AMC. With an N state FSMC, the SNR at the receiver γ can be partitioned into $N+1$ non-overlapping intervals by thresholds Γ_n ($n \in \{0, 1, \ldots, N\}$), where $\Gamma_0 = 0 < \Gamma_1 < \ldots < \Gamma_{N+1} = \infty$. The channel is said to be in state n if $\Gamma_n \leq \gamma < \Gamma_{n+1}$. In this state n bits can be transmitted per symbol using 2^n-QAM (Quadrature Amplitude Modulation) which corresponds to transmission rate n. To avoid possible transmission error, no packet is transmitted when $n = 0$.

Assuming that the channel is slowly fading (i.e., transitions occur only between adjacent states), the state transition matrix for the FSMC can be expressed as follows [25]:

$$\zeta = \begin{bmatrix} \zeta_{0,0} & \zeta_{0,1} & \cdots & 0 \\ \zeta_{1,0} & \zeta_{1,1} & \zeta_{1,2} & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & \zeta_{N-1,N-2} & \zeta_{N-1,N-1} & \zeta_{N-1,N} \\ 0 & \cdots & \zeta_{N,N-1} & \zeta_{N,N} \end{bmatrix}. \tag{7.1}$$

The transition probability from state n to n' ($n' \in \{n-1, n, n+1\}$), $\zeta_{n,n'}$ can be obtained as follows:

$$\zeta_{n,n+1} = \frac{N_{n+1} \times t}{\Pr(n)}, \quad n = 0, \ldots, N-1 \tag{7.2}$$

$$\zeta_{n,n-1} = \frac{N_n \times t}{\Pr(n)}, \quad n = 1, \ldots, N \tag{7.3}$$

$$\zeta_{n,n} = \begin{cases} 1 - \zeta_{n,n+1} - \zeta_{n,n-1}, & 0 < n < N \\ 1 - \zeta_{0,1}, & n = 0 \\ 1 - \zeta_{N,N-1}, & n = N \end{cases} \tag{7.4}$$

where t is the length of a transmission time slot and N_n is the level crossing-rate at Γ_n of state n and it can be estimated from

$$N_n = \sqrt{2\pi \frac{m\Gamma_n}{\overline{\gamma}}} \frac{f_d}{\Gamma(m)} \left(\frac{m\Gamma_n}{\overline{\gamma}}\right)^{m-1} \exp\left(-\frac{m\Gamma_n}{\overline{\gamma}}\right). \tag{7.5}$$

To calculate the packet error rate (PER) when the channel state is n, we use the following approximation [25]:

$$PER_n(\gamma) \approx \begin{cases} 1, & 0 < \gamma < \Gamma_{pn} \\ a_n \exp(-g_n \gamma), & \gamma \geq \Gamma_{pn} \end{cases} \tag{7.6}$$

where a_n, g_n and Γ_{pn} are obtained by fitting the exact PER curve. Then the average packet error rate $\overline{PER}_n$ corresponding to transmission rate n can can be obtained as follows:

$$\begin{aligned} \overline{PER}_n &= \frac{1}{\Pr(n)} \int_{\Gamma_n}^{\Gamma_{n+1}} a_n \exp(-g_n\gamma) p_\gamma(\gamma) \mathrm{d}\gamma \\ &= \frac{1}{\Pr(n)} \frac{a_n}{\Gamma(m)} \left(\frac{m}{\overline{\gamma}}\right)^m \frac{\Gamma(m, e_n\Gamma_n) - \Gamma(m, e_n\Gamma_{n+1})}{(e_n)^m}, \quad n = 1, \cdots, N \end{aligned} \tag{7.7}$$

Here, $\Pr(n)$ is the probability that the channel state is n, which is given by

$$\Pr(n) = \frac{\Gamma(m, m\Gamma_n/\overline{\gamma}) - \Gamma(m, m\Gamma_{n+1}/\overline{\gamma})}{\Gamma(m)} \tag{7.8}$$

where $\overline{\gamma}$ is the average SNR, m is the Nakagami fading parameter ($m \geq 0.5$), $\Gamma(m)$ is the Gamma function, $\Gamma(m, \gamma)$ is the complementary incomplete Gamma function, and $e_n = m/\overline{\gamma} + g_n$.

7.3.2 Packet Transmission and Error Control

We consider a time-division multiplexing (TDM)-based packet transmission scenario where the size of each packet is L *bits*. The length of a time slot is denoted by t which is assumed to be equal to the time interval required to transmit one packet using the basic modulation level (i.e., the number of bits per symbol is one). The number of packets that can be transmitted in a channel during one time slot (n_p) depends on the corresponding modulation level and $n_p \in \{0, 1, 2, \ldots, N\}$.

Since real-time traffic is delay-intolerant we do not assume any error control for this type of traffic in this chapter. For non-real-time and best-effort traffic, an infinite persistent automatic repeat request (ARQ) protocol is used. That is, the erroneous packets will be re-transmitted until they are successfully received at the base station. Assuming an independent packet error process, the probability that l out of n packets are successfully transmitted in one time slot can be obtained as follows:

$$\theta_{l,n} = \binom{n}{l} \theta^l (1-\theta)^{n-l} \tag{7.9}$$

where θ is the probability of successful transmission for a packet (i.e., $\theta = 1 - \overline{PER}_n$) when the transmitter uses modulation level n. We also assume that the transmission status for the packet transmitted in the previous time slot is made available to the mobile before transmissions in the current time slot start.

7.3.3 Traffic Sources

7.3.3.1 Real-Time Traffic

This type of traffic is strictly delay-sensitive but loss-insensitive (e.g., voice or real-time video). Therefore, upon availability of the channel resources, the generated packets from a source are transmitted immediately. We use Markov modulated Poisson process (MMPP), which is able to capture burstiness in the traffic arrival process, to model a real-time traffic source. With MMPP, the packet arrival rate λ_s is determined by the state s of the Markov chain, and the total number of states is S (i.e., $s = 1, 2, \ldots, S$). The MMPP process can be represented by $\mathbf{U}$ and $\mathbf{\Lambda}$, in which the former is the transition probability matrix of the modulating Markov chain, and the

latter is the matrix corresponding to the Poisson arrival rates. These matrices are defined as follows:

$$\mathbf{U} = \begin{bmatrix} u_{1,1} & \cdots & u_{1,S} \\ \cdots & \cdots & \cdots \\ u_{S,1} & \cdots & u_{S,S} \end{bmatrix}, \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_S \end{bmatrix}. \tag{7.10}$$

Discrete-time MMPP (dMMPP) [40] provides similar result to MMPP in the continuous time when the interval of time slot becomes very small relative to the smallest mean time in the continuous matrix. In this case, the rate matrix $\mathbf{\Lambda}$ is represented by diagonal probability matrix $\mathbf{\Lambda}_a$ when the number of packets arriving in one frame is a. Each element of $\mathbf{\Lambda}_a$ can be obtained from $f_a(\lambda_s)$ as follows:

$$\begin{aligned} \mathbf{\Lambda}_a &= \begin{bmatrix} f_a(\lambda_1) & \cdots & f_a(\lambda_S) \end{bmatrix}, \quad \cdots \\ \mathbf{\Lambda}_A &= \begin{bmatrix} F_A(\lambda_1) & \cdots & F_A(\lambda_S) \end{bmatrix} \end{aligned}$$

where the probability that a packets arrive during time interval t with mean rate λ is given by

$$f_a(\lambda) = \frac{e^{-\lambda t}(\lambda t)^a}{a!} \tag{7.11}$$

where $a \in \{0, 1, \ldots, A\}$ and A is the maximum batch size for packet arrival which is the queue size (i.e., $A = X$) in this case. The complementary cumulative probability mass function for this arrival process is given by

$$F_a(\lambda) = \sum_{j=a}^{\infty} f_j(\lambda, t). \tag{7.12}$$

In this chapter, we consider the packet loss probability (due to the unavailability of channels) as the QoS metric for this type of traffic.

7.3.3.2 Non-Real-Time Traffic

This type of traffic is moderately delay-sensitive, however, loss-insensitive. We assume Poisson process for this traffic source. Packets are queued in a transmission queue at the mobile node and upon transmission failure ARQ-based error recovery is performed. In this chapter, we consider the packet dropping probability (due to finite size of the buffer) and packet delay as the QoS metrics for this type of traffic.

7.3.3.3 Best-Effort Traffic

For this traffic class, we consider a file transfer scenario where fixed size (F packets) files (e.g., images) are transmitted from the mobile to the base station as a batch of packets. In case of transmission failure, ARQ-based error recovery is used. We consider the delay distribution as the QoS metric for this type of traffic.

7.3.4 ACA and CAC

The number of channels allocated to ongoing calls can be adjusted adaptively according to the number of ongoing calls in the corresponding service class. The total number of channels available for one particular service class is assumed to be C (i.e., $C \in \{C_{rt}, C_{nrt}, C_{be}\}$). The number of channels allocated to a call is chosen from a set of discrete values $B = \{c_1, c_2, \ldots, c_{max}\}$ where $c_i < c_{i+1}$ and $c_i \in \mathbb{N}$. The minimum and the maximum number of channels that can be allocated to a call is given by c_1 and c_{max}, respectively. In this chapter, we assume with loss of generality that $c_1 = 1$ which denotes the minimum number of channels that an ongoing call requires to maintain the connection. This minimum number of channels corresponds to minimum allocated time slot or frequency band in TDMA and FDMA systems, respectively.

There are two types of incoming calls: handoff calls and new calls. To prioritize handoff calls over new calls, guard channel scheme is used to reserve a certain number of channels for handoff calls. In other words, incoming new calls are accepted if the number of ongoing calls in the service class is less than a predefined threshold K (i.e., $K \in \{K_{rt}, K_{nrt}, K_{be}\}$). Therefore, the number of channels reserved for handoff calls is $C - K$.

7.3.5 Adaptative Channel Allocation Algorithm

For a particular service class, let w_{allc} and $\mathbf{c}_{allc}$ denote the number of channels which can be allocated to an incoming call and the vector corresponding to the number of channels that are currently allocated to the ongoing calls, respectively. When a new call arrives, admission control is performed by checking whether the total number of ongoing calls is less than the guard channel threshold K (Algorithm 7.3.1). If this

Algorithm 7.3.1: ACA(*inputs* : type of incoming call, $K, C, \mathbf{c}_{allc}, c_1, c_{max}$)

if (((incoming call is a new call) **and** (number of ongoing calls $< K$))
or (incoming call is a handoff call))

then
- **if** available bandwidth $\geq c_{max}$
  - **then** assign c_{max} to incoming call
  - **else**
    - $w_{allc} \leftarrow 0$
    - **while** $\max(\mathbf{c}_{allc}) > c_1$ **and** $w_{allc} < \min(\mathbf{c}_{allc})$
      - **do**
        - randomly select one call with number of channels $\max(\mathbf{c}_{allc})$
        - decrease number of channels for the selected call by one
      - $w_{allc} \leftarrow w_{allc} + 1$
    - **if** $w_{allc} > 0$
      - **then** accept incoming call with number of channels w_{allc}
      - **else** reject incoming call

else reject incoming new call

condition is satisfied or if the incoming call is a handoff call, the base station tries to allocate maximum number of channels (c_{max}) to the incoming call; otherwise, the incoming new call is blocked. However, if the available number of channels is not enough to allocate c_{max} channels, the adaptation algorithm is invoked.

A fairness-based channel adaptation algorithm is used so that the maximum difference in the number of allocated channels for every ongoing call is one, and all of the ongoing calls have the same probability to be degraded and upgraded when a call arrives or departs, respectively. The adaptation algorithm randomly selects an ongoing call with the current maximum number of channels (i.e., $\max(\mathbf{c}_{allc})$) and then decreases the number of channels for that call by one (i.e., $w_{allc} \leftarrow w_{allc} + 1$). At the same time, the number of channels available for the incoming call increases by one. This operation is iteratively performed until the number of channels that can be allocated to an incoming call is equal to the minimum of the number of channels currently allocated to all ongoing calls (i.e., $\min(\mathbf{c}_{allc})$). In contrast, if every call is allocated with the minimum number of channels c_1, none of the ongoing calls can be degraded. In this case, an incoming new call/handoff call is blocked/dropped.

If an ongoing call is terminated or handed over to a neighboring cell, the corresponding channels are released, and some of the ongoing calls will be upgraded to have more number of channels. The upgrade procedure selects a call with the minimum number of channels (i.e., $\min(\mathbf{c}_{allc})$) and then the number of channels allocated to that call is increased by one. This routine is performed until all released channels are allocated or all of the ongoing calls have the maximum number of channels (c_{max}).

For the above fairness-based ACA, we can calculate the number of calls $m_i(B, C, u)$ which are allocated with c_i channels at the certain point of time based on the number of ongoing calls u as follows:

$$m_i(B,C,u) = \begin{cases} u - \left\lfloor \frac{C-c_i u}{c_{i+1}-c_i} \right\rfloor, & i = \hat{i} \\ \left\lfloor \frac{C-c_i u}{c_{i+1}-c_i} \right\rfloor, & i = \hat{i}+1 \\ 0, & \text{otherwise} \end{cases} \tag{7.13}$$

where $\hat{i} = \hat{i}(B,C,u) = \max\{j | c_j \leq C/u\}$.

The average number of allocated channels for a certain number of ongoing calls u is calculated as follows:

$$\overline{c}(u) = \begin{cases} \frac{C}{u} & \lfloor \frac{C}{c_{i+1}} \rfloor \leq u \leq \lfloor \frac{C}{c_i} \rfloor \\ c_{max} & \lfloor \frac{C}{c_{max}} \rfloor \geq u. \end{cases} \tag{7.14}$$

For example, with $C = 20$ and $B = \{1, 2, 3\}$, when the number of ongoing calls is 15, we have $m_2(B, 20, 15) = 5$ calls, $m_1(B, 20, 15) = 10$ calls and $\overline{c}(15) = 1.33$ channels.

7.3.6 Adaptive Traffic Shaping

To maintain the QoS performance, adaptive traffic shaping is used to control the packet generation rate for a real-time/non-real-time traffic source. Since the transmission rate corresponding to a call should vary depending on the number of allocated channels c and channel quality (i.e., average SNR, $\overline{\gamma}$), the traffic shaper adjusts the arrival rate $\tilde{\lambda}(c, \lambda)$ when the number of allocated channels changes (i.e., after a call arrives and/or departs the cell). Therefore, if the packet generation rate from a traffic source is λ, the traffic shaper will limit the arrival rate based on the average transmission rate $\overline{l}_c$ corresponding to c allocated channels (which will be derived in the next section) as follows:

$$\tilde{\lambda}(c, \lambda) = \min(\lambda, \phi \times \overline{l}_c) \tag{7.15}$$

where ϕ is an adjustable traffic shaping parameter. Note that, if we set ϕ to a very large number (e.g., $\phi = \infty$), the traffic shaper will be disabled since the packet arrival rate would be the same as the packet generation rate.

Traffic shaping can be implemented either in the data link layer or in the application layer. In the data link layer, incoming packets can be dropped randomly such that the arrival rate conforms to $\tilde{\lambda}(c, \lambda)$. In the application layer, for example, multimedia adaptation techniques can be used to reduce the refresh rate or the resolution of the media such that packet generation rate is controlled to $\tilde{\lambda}(c, \lambda)$.

7.4 Formulation of the Markov Models

We develop a discrete-time Markov chain (DTMC) model to analyze the call-level and the packet-level performances for the test calls in a particular service class under the system model described earlier. We consider a test call that stays forever in the system. When c channels are allocated, we assume only first c channels are always allocated to the test call. The number of packets that can be transmitted in one time slot of the test call depends on the number of allocated channels and the modulation index (i.e., the achievable number of bits per symbol) used in each of these channels. For a test call, the transmission probability matrix corresponding to the modulation index used in the different channels can be derived analytically from the average SNR and the target BER.

Since the number of assigned channels to a test call depends on the number of ongoing calls we first formulate the call-level model. Note that, this model is a general one and can be applied to any service class. Then the queueing models for the real-time and the non-real-time traffic sources are presented. The performance model for the best-effort traffic (i.e., file transfer scenario) is presented afterwards.

7.4.1 Transmission Probability Matrices

We can establish the transmission probability matrices $\mathbf{D}_l^{(k)}$ ($l \in \{0, 1, \ldots, N\}$ and $\mathbf{D}_l^{(k)} \in \mathbb{R}_{1\times(N+1)^k}$) for one channel (i.e., $k = 1$) as follows:

$$\left[\mathbf{D}_l^{(1)}\right]_{n+1} = \begin{cases} 1, & l = n \\ 0, & \text{otherwise} \end{cases} \tag{7.16}$$

$$\left[\mathbf{D}_l^{(1)}\right]_{n+1} = \theta_{l,n} \tag{7.17}$$

for transmission without error control (7.16), and with infinite persistent ARQ (7.17), respectively. Note that, $\left[\mathbf{D}_l^{(k)}\right]_j$, the element at column j of matrix $\mathbf{D}_l^{(k)}$, corresponds to the probability of a decrease of l in the number of packets in the queue when the channel state is l and k channels are allocated to the queue.

In each time slot, there are c channels allocated to a random chosen call, and we assume that the average SNR is the same in each of these c channels while the fading processes in these channels are independently varying. Based on (7.1), the channel state transition matrix for any allocated channels c can be expressed as follows:

$$\boldsymbol{\zeta}_c = \bigotimes_{i=1}^{c} \boldsymbol{\zeta} \tag{7.18}$$

where $\otimes$ denotes Kronecker product. The transmission probability matrix for c FSMC channels can be obtained as follows:

$$\mathbf{D}_l^{(c)} = \sum_{\{i,j|i+j=l\}} \mathbf{D}_i^{(c-1)} \otimes \mathbf{D}_j^{(1)}, \quad i,j \in \{0,1,\ldots,N\} \tag{7.19}$$

for $l = 0,1,\ldots,N \times 2$. For more than two channels, the transmission matrices can be obtained in a similar way. We can calculate the average transmission rate for c FSMC channels as follows:

$$\bar{l}_c = \sum_{l=1}^{N\times c} l \times \left(\boldsymbol{\pi}_f \left(\mathbf{D}_l^{(c)}\right)^T\right) \tag{7.20}$$

where $\mathbf{1}$ is the column vector of ones, and $\boldsymbol{\pi}_f$ (i.e., steady state probability of channel state of c channels) is obtained by solving

$$\boldsymbol{\pi}_f \boldsymbol{\zeta}_c = \boldsymbol{\pi}_f \text{ and } \boldsymbol{\pi}_f \mathbf{1} = 1. \tag{7.21}$$

7.4.2 Call-Level Markov Model

We formulate a two-dimensional DTMC for which the states of a test call are observed at the end of each transmission time slot. The state space for this DTMC is

$$\Delta_{call} = \{(\mathcal{U}, \mathcal{B}); 0 \leq \mathcal{U} \leq C, 0 \leq \mathcal{B} \leq 1\} \tag{7.22}$$

where $\mathcal{U}$ and $\mathcal{B}$ denote the state of number of allocated channels to a call chosen at random at each call arrival or departure, respectively. In particular, c_{i+1} and c_i channels are allocated to a call chosen at random if $\mathcal{B} = 1$ and $\mathcal{B} = 0$, respectively. According to this state space, a two dimensional DTMC is established. The first dimension of the Markov chain represents the number of ongoing calls, and the second dimension represents the number of allocated channels to a test call. Note that, the more the number of ongoing calls, the fewer the number of allocated channels to each call.

The variation in the number of ongoing calls is a stochastic process which can be modeled as a special type of $M/M/C/K$ queue. In particular, there are two types of calls in a cell, namely, new calls and handoff calls. The arrival process for these calls is assumed to be Poisson with rates $\rho_{(n)}$ and $\rho_{(h)}$, respectively. The channel holding time for both of these types of calls is assumed to be exponentially distributed with mean $1/\mu$.

For the second dimension of the Markov chain, the number of allocated channels to a random chosen call depends on the number of ongoing calls in the corresponding service class. The transition probability matrix can be expressed as follows:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_{0,0} & \mathbf{q}_{0,1} & & & & \\ \mathbf{q}_{1,0} & \mathbf{q}_{1,1} & \mathbf{q}_{1,2} & & & \\ \ddots & \ddots & \ddots & & & \\ & \mathbf{q}_{u-1,u} & \mathbf{q}_{u,u} & \mathbf{q}_{u,u+1} & & \\ & \ddots & \ddots & \ddots & & \\ & & \mathbf{q}_{C-2,C-1} & \mathbf{q}_{C-1,C-1} & \mathbf{q}_{C-1,C} & \\ & & & \mathbf{q}_{C,C-1} & \mathbf{q}_{C,C} \end{bmatrix} \tag{7.23}$$

where element $\mathbf{q}_{u,u'}$ indicates that there are u ongoing calls in time slot T and there are u' calls in slot $T+1$. The elements $\mathbf{q}_{u,u'}$ can be obtained as follows:

$$\begin{aligned} [\mathbf{q}_{u,u+1}]_{k,1} &= \frac{F_1(\alpha,t) \times f_0(u\mu,t) \times m_{i+1}(B,C,u+1)}{m_{i+1}(B,C,u+1) + m_i(B,C,u+1)} \\ [\mathbf{q}_{u,u+1}]_{k,2} &= \frac{F_1(\alpha,t) \times f_0(u\mu,t) \times m_i(B,C,u+1)}{m_{i+1}(B,C,u+1) + m_i(B,C,u+1)} \\ [\mathbf{q}_{u,u-1}]_{k,1} &= \frac{f_0(\alpha,t) \times F_1(u\mu,t) \times m_i(B,C,u-1)}{m_{i-1}(B,C,u-1) + m_i(B,C,u-1)} \end{aligned} \tag{7.24}$$

$$
\begin{aligned}
[\mathbf{q}_{u,u-1}]_{k,2} &= \frac{f_0(\alpha,t) \times F_1(u\mu,t) \times m_{i+1}(B,C,u-1)}{m_{i-1}(B,C,u-1) + m_i(B,C,u-1)} \\
[\mathbf{q}_{u,u}]_{k,k} &= f_0(\alpha,t) \times f_0(u\mu,t) + F_1(\alpha,t) \times F_1(u\mu,t)
\end{aligned}
$$

where $k = 1, 2$ and α represents the total call arrival rate which is obtained based on the CAC algorithm as follows:

$$
\alpha(u) = \begin{cases} \rho_{(n)} + \rho_{(h)} & u < K \\ \rho_{(h)} & K \leq u < C. \end{cases} \tag{7.25}
$$

Due to the fairness-based channel adaptation, the size of the matrix $\mathbf{q}_{u,u'}$ can be reduced to 2×2 (since the maximum difference in the number of allocated channels among the ongoing calls is one), and the first and the second row correspond to c_{i+1} and c_i, respectively, for a certain number of ongoing calls u. The transition probabilities in these matrices $\mathbf{q}_{u,u'}$ depend on the number of calls in the next time slot while the number of allocated channel to a test call depend on factor $\frac{m_{i+1}(B,C,u)}{m_{i+1}(B,C,u)+m_i(B,C,u)}$ and $\frac{m_i(B,C,u+1)}{m_{i+1}(B,C,u)+m_i(B,C,u)}$.

The call-level performance measures (i.e., new call blocking probability, handoff call dropping probability, average number of ongoing call, and average number of allocated channels to an ongoing call) can be obtained from steady state probability $\boldsymbol{\pi}_{call}$ which can be obtained by solving $\boldsymbol{\pi}_{call}\mathbf{Q} = \boldsymbol{\pi}_{call}$ and $\boldsymbol{\pi}_{call}\mathbf{1} = 1$. The steady state probability for the state $\pi_{call}(u,c)$, which corresponds to the case that the number of ongoing calls is u and the number of allocated channels is c, can be decomposed from matrix $\boldsymbol{\pi}_{call}$ as follows:

$$
\begin{aligned}
\pi_{call}(0,0) &= [\boldsymbol{\pi}_{call}]_1 \\
\pi_{call}(u,c_{i+1}) &= [\boldsymbol{\pi}_{call}]_{u\times 2}, \quad \text{if } m_{i+1}(B,C,u) > 0 \\
\pi_{call}(u,c_i) &= [\boldsymbol{\pi}_{call}]_{u\times 2+1}, \quad \text{if } m_i(B,C,u) > 0
\end{aligned} \tag{7.26}
$$

where $[\boldsymbol{\pi}_{call}]_j$ indicates the element at column j of row matrix $\boldsymbol{\pi}_{call}$. The average number of ongoing calls is obtained from

$$
\overline{u} = \sum_{u=1}^{C} u \times \sum_{\forall c} \pi_{call}(u,c). \tag{7.27}
$$

New call blocking probability P_{nb} and handoff call dropping probability P_{hd} are obtained from

$$
P_{nb} = \sum_{u=K}^{C} \sum_{\forall c} \pi_{call}(u,c), \quad P_{hd} = \sum_{\forall c} \pi_{call}(C,c). \tag{7.28}
$$

The average number of allocated channels for each call is

$$\overline{c} = \sum_{c=1}^{c_{max}} c \times \sum_{\forall u} \pi_{call}(u, c). \tag{7.29}$$

7.4.3 Modeling for Real-Time Traffic

The objective is to derive the packet loss rate for a real-time traffic source which is modeled as an MMPP. The state space in this case is

$$\Delta_r = \{(\mathcal{M}, \mathcal{U}, \mathcal{B}, \mathcal{A}); 1 \leq \mathcal{M} \leq S, 1 \leq \mathcal{U} \leq C, 0 \leq \mathcal{B} \leq 1\} \tag{7.30}$$

where $\mathcal{U}$ is the number of ongoing calls, $\mathcal{B}$ is the level of number of allocated channels, and $\mathcal{M}$ represents the arrival state of MMPP source, and $\mathcal{A}$ is the channel state.

Now, since we observe queue state at the test mobile which stays forever in the system, the transition probability matrix for the call-level model needs to be modified such that the minimum number of calls is one as follows:

$$\mathbf{Q}' = \begin{bmatrix} \mathbf{q}_{1,1} & \mathbf{q}_{1,2} & & & & \\ \mathbf{q}_{2,1} & \mathbf{q}_{2,2} & \mathbf{q}_{2,3} & & & \\ \ddots & \ddots & \ddots & & & \\ & \mathbf{q}_{u-1,u} & \mathbf{q}_{u,u} & \mathbf{q}_{u,u+1} & & \\ & \ddots & \ddots & \ddots & & \\ & & \mathbf{q}_{C-2,C-1} & \mathbf{q}_{C-1,C-1} & \mathbf{q}_{C-1,C} & \\ & & & \mathbf{q}_{C,C-1} & \mathbf{q}_{C,C} \end{bmatrix} \tag{7.31}$$

and all element $\mathbf{q}_{u,u'}$ can be obtained from (7.24)). Note that, this matrix $\mathbf{Q}'$ is similar to that in (7.23) in which the first row and column, corresponding to state that $\mathcal{U} = 0$, is removed since we consider embedded Markov chain when the system has at least one ongoing call.

When the system state transition matrix above is combined with the channel state transition matrix, it becomes

$$\mathbf{R} = \mathbf{Q}' \otimes \boldsymbol{\zeta}_{c_{max}} \tag{7.32}$$

where $\mathbf{R} \in \mathbb{R}_{2C(N+1)^{c_{max}} \times 2C(N+1)^{c_{max}}}$.

The vector $\mathbf{E}_l$ (i.e., $\mathbf{E}_l \in \mathbb{R}_{1 \times 2C(N+1)^{c_{max}}}$) indicates packet departure process corresponding with evolution of number of ongoing calls, number of allocated channels

to a test call and channel state and can obtained from

$$\mathbf{E}_l = \left[\begin{array}{cccc} \mathbf{d}_l^{(1)} & \mathbf{d}_l^{(2)} & \cdots & \mathbf{d}_l^{(C)} \end{array} \right] \tag{7.33}$$

where

$$\mathbf{d}_l^{(u)} = \left[\begin{array}{cc} \mathbf{D}_l^{(c_{\hat{i}(B,C,u)+1})} \otimes \mathbf{1}_{j(C,1)} & \mathbf{D}_l^{(c_{\hat{i}(B,C,u)})} \otimes \mathbf{1}_{j(C,0)} \end{array} \right] \tag{7.34}$$

where $\mathbf{1}_{j(u,d)}$ is an vector of ones with size $j(u,d)$ and $j(u,d) = (N+1) \times (c_{max} - c_{\hat{i}+d})$. The purpose of this vector is to enlarge the size of $\mathbf{E}_l$ to match with $\mathbf{R}$. This block $\mathbf{D}_l^{(c_{\hat{i}(B,C,u)})} \otimes \mathbf{I}_{j(u,d)}$ corresponds to the case that $c_{\hat{i}(B,C,u)}$ channels are allocated to the interested call and $c_{max} - c_{\hat{i}+d}$ channels are allocated to the other calls. We can calculate the average offered transmission rate (i.e., transmission rate without packet arrival) for each mobile from

$$\overline{l} = \sum_{l=1}^{N \times c_{max}} l \times \left(\boldsymbol{\pi}_r \left(\mathbf{E}_l \right)^T \right) \tag{7.35}$$

where $\boldsymbol{\pi}_r$ (i.e., steady state probability corresponding to number of ongoing calls, number of allocated channels, and channel state) is obtained by solving

$$\boldsymbol{\pi}_r \mathbf{R} = \boldsymbol{\pi}_r \quad \text{and} \quad \boldsymbol{\pi}_r \mathbf{1} = 1. \tag{7.36}$$

With adaptive traffic shaping, the Poisson arrival matrix for the dMMPP becomes

$$\begin{aligned} \boldsymbol{\Lambda}_a &= \left[\begin{array}{ccc} f_a(\tilde{\lambda}(c,\lambda_1)) & \cdots & f_a(\tilde{\lambda}(c,\lambda_S)) \end{array} \right], \quad \cdots \\ \boldsymbol{\Lambda}_A &= \left[\begin{array}{ccc} F_A(\tilde{\lambda}(c,\lambda_1)) & \cdots & F_A(\tilde{\lambda}(c,\lambda_S)) \end{array} \right]. \end{aligned} \tag{7.37}$$

The matrix corresponding to the packets (among the generated packets) which cannot be transmitted in one time slot (and therefore lost) is obtained from

$$\mathbf{K}_z = \sum_{a-l=z} \boldsymbol{\Lambda}_a \otimes \mathbf{E}_l \quad \text{for } z = 1, 2, \ldots, A. \tag{7.38}$$

Therefore, the average number of packets lost per time slot is given by

$$\overline{x}_{loss} = \sum_{z=1}^{A} z \times \left(\boldsymbol{\pi}_{ur} \left(\mathbf{K}_z \right)^T \right) \tag{7.39}$$

where $\boldsymbol{\pi}_{ur}$ is obtained from solving $\boldsymbol{\pi}_{ur} \left(\mathbf{U} \otimes \mathbf{R} \right) = \boldsymbol{\pi}_{ur}$ and $\boldsymbol{\pi}_{ur} \mathbf{1} = 1$. Note that, $(\mathbf{U} \otimes \mathbf{R})$ represents the transition probability matrix of the entire system (i.e., MMPP

state, number of ongoing calls, level of number of channels allocated to a test call, and channel state). Then, the packet loss rate is obtained from

$$P_{loss} = \frac{\overline{x}_{loss}}{\overline{\lambda}_{mmpp}}. \tag{7.40}$$

Here, $\overline{\lambda}_{mmpp}$ is the average packet generation rate of the MMPP source which is obtained from

$$\overline{\lambda}_{mmpp} = \sum_{a=1}^{A} a \times \left(\boldsymbol{\pi}_m \left(\boldsymbol{\Lambda}_a\right)^T\right) \quad \text{where} \quad \boldsymbol{\pi}_m \mathbf{U} = \boldsymbol{\pi}_m \text{ and } \boldsymbol{\pi}_m \mathbf{1} = 1. \tag{7.41}$$

Note that, $\boldsymbol{\pi}_m$ denotes the steady state probability vector corresponding to the different states of an MMPP source. Throughput of real-time call can be obtained from

$$\eta_{rt} = \overline{\lambda}_{mmpp} - \overline{x}_{loss}. \tag{7.42}$$

7.4.4 Queueing Model for Non-Real-Time Traffic

For Non-real-time traffic, we relax the arrival process of MMPP by using batch Bernoulli in which the arrival process is statistical identical independent with truncated Poisson. However, call-level model are the same with that of real-time traffic. We consider a discrete-time queueing model for a test call with non-real-time traffic where the packets arriving in time slot T cannot be served until the next time slot $T+1$ at the earliest. With the call under consideration, we assume that the queue size is finite (i.e., X packets). Any incoming packet will be dropped if the queue is full. The number of packets transmitted in each time slot depends on the allocated number of channels and the corresponding transmission rates.

7.4.4.1 System State Space

The state space is as follows:

$$\Delta = \{(\mathcal{X}, \mathcal{U}, \mathcal{B}, \mathcal{A}); 0 \leq \mathcal{X} \leq X, 1 \leq \mathcal{U} \leq C, 0 \leq \mathcal{B} \leq 1\} \tag{7.43}$$

where $\mathcal{U}$ is the number of ongoing calls, $\mathcal{B}$ is the level of number of allocated channels, $\mathcal{X}$ is the number of packets in the queue and $\mathcal{A}$ is the channel state.

For this case, we can establish the diagonal matrix corresponding to the packet arrival process with adaptive traffic shaping as follows:

$$\mathbf{G}_a = \begin{bmatrix} f_a(\tilde{\lambda}(c_{max},\lambda)) \times \mathbf{I}_{(N+1)\times c_{max}} & & & \\ & \ddots & & \\ & & f_a(\tilde{\lambda}(c(s),\lambda)) \times \mathbf{I}_{(N+1)\times c_{max}} & \\ & & & \ddots \\ & & & f_a(\tilde{\lambda}(c_1,\lambda)) \times \mathbf{I}_{(N+1)\times c_{max}} \end{bmatrix} \tag{7.44}$$

and for $\mathbf{G}_A$, function $f_a(.)$ becomes $F_a(.)$. The average packet arrival rate due to adaptive traffic shaping can be obtained from

$$\overline{a} = \sum_{a=1}^{A} a \times (\boldsymbol{\pi}_r \mathbf{G}_a \mathbf{1}) . \tag{7.45}$$

7.4.4.2 Transition Matrix

The Markov chain in this model can be developed by modifying the model for the real-time traffic by incorporating the stochastic process for the variations of the number packets in the queue. While packet arrivals are independent of the number of packets in the queue, the packet departure process depends on the number of allocated channels and the corresponding channel states.

The state transition matrix $\mathbf{P}$ in this case can be defined as in (7.46). The rows of matrix $\mathbf{P}$ represent the number of packets in the queue and the rows in matrix $\mathbf{p}_{x,x'}$ represent the number of ongoing calls, the number of allocated channels to a test call, and the channel state.

$$\mathbf{P} = \left[\begin{array}{cccccc} \mathbf{p}_{0,0} & \cdots & \mathbf{p}_{0,A} & & & \\ \vdots & \ddots & \ddots & \ddots & & \\ \hline \mathbf{p}_{N_m,0} & \cdots & \mathbf{p}_{N_m,N_m} & \cdots & \mathbf{p}_{N_m,N_m+A} & \\ & \ddots & \ddots & \ddots & \ddots & \ddots \\ & \mathbf{p}_{x,x-N_m} & \cdots & \mathbf{p}_{x,x} & \cdots & \mathbf{p}_{x,x+A} \\ \hline & & \ddots & \ddots & \ddots & \ddots \\ & & & \mathbf{p}_{X,X-N_m} & \cdots & \mathbf{p}_{X,X} \end{array}\right] . \tag{7.46}$$

Since in one time slot several packets can arrive and be transmitted, matrix $\mathbf{P}$ is divided into three parts. The first part, from row 0 to $N_m - 1$, indicates the case

that the maximum total transmission rate is greater than the number of packets in the queue and none of the incoming packets is dropped. The second part, from row N_m to $X - A$, represents the case in which the maximum packet transmission rate is equal to or less than the number of packets in queue and none of the incoming packets is dropped. The third part, from row $X - A + 1$ to X, indicates the case that some of the incoming packets are dropped due to the lack of queue space. Since the maximum total offered packet transmission rate can be greater than the number of packets in the queue, the maximum amount by which the number of packets in queue can decrease is obtained from

$$N'_m = \min(N_m, x) \tag{7.47}$$

where x is the number of packets in queue. Therefore, for the packet departure process, the diagonal matrix corresponding to the maximum number of packets that can be transmitted (N'_m) when there are x packets in the queue ($\mathbf{E}^{(x)}_{N'_m}$) can be expressed as follows:

$$\mathbf{E}^{(x)}_{N'_m} = \sum_{l=x}^{N_m} \mathrm{diag}(\mathbf{E}_l) \quad \text{if } N'_m = x \tag{7.48}$$

and

$$\mathbf{E}^{(x)}_{l} = \mathrm{diag}(\mathbf{E}_l) \quad \text{for } l = 0, 1, \ldots, N'_m \tag{7.49}$$

where function $\mathrm{diag}(\mathbf{E}_l)$ changes vector $\mathbf{E}_l$ into a corresponding diagonal matrix. That is $[\mathrm{diag}(\mathbf{x})]_{i,j} = [\mathbf{x}]_i \, \delta_{i,j}$ where $\delta_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$.

The elements without dropped packet in the first and second parts of the matrix $\mathbf{P}$ can be obtained as follows:

$$\hat{\mathbf{p}}_{x,x-y} = \mathbf{R} \sum_{\{l,a|l-a=y\}} \mathbf{G}_a \times \mathbf{E}^{(x)}_l \quad \text{for } y = 1, 2, \ldots, N'_m \tag{7.50}$$

$$\hat{\mathbf{p}}_{x,x+z} = \mathbf{R} \sum_{\{a,l|a-l=z\}} \mathbf{G}_a \times \mathbf{E}^{(x)}_l \quad \text{for } z = 1, 2, \ldots, A \tag{7.51}$$

$$\hat{\mathbf{p}}_{x,x} = \mathbf{R} \sum_{\{a,l|a=l\}} \mathbf{G}_a \times \mathbf{E}^{(x)}_l \tag{7.52}$$

where $l \in \{0, 1, 2, \ldots, N'_m\}$ and $a \in \{0, 1, 2, \ldots, A\}$ represent the number of transmitted packets and the number of packet arrivals, respectively. Considering both the

packet arrival and the departure events, (7.50), (7.51), and (7.52) above represent the transition probability matrices for the cases when the number of packets in the queue decreases by y packets, increases by z packets, and does not change, respectively.

The third part of the matrix $\mathbf{P}$ $(x = X - A + 1, X - A + 2, \ldots, X)$ has to capture the packet dropping effect. Let $\hat{\mathbf{p}}_{x,x+i}$ denote the element of matrix $\mathbf{P}$ in the case that there is no dropped packet, (7.51) becomes

$$\mathbf{p}_{x,x+z} = \begin{cases} \sum_{i=z}^{A} \hat{\mathbf{p}}_{x,x+i} & \text{for } x+z \geq X \\ \hat{\mathbf{p}}_{x,x+z} & \text{otherwise} \end{cases} \tag{7.53}$$

for $x + z \geq X$ $(z = X - x, \ldots, A)$, and (7.52) becomes

$$\mathbf{p}_{x,x} = \begin{cases} \hat{\mathbf{p}}_{x,x} + \sum_{i=1}^{A} \hat{\mathbf{p}}_{x,x+i} & \text{for } x = X \\ \hat{\mathbf{p}}_{x,x} & \text{otherwise.} \end{cases} \tag{7.54}$$

Eqs. (7.53) and (7.54) indicate that the queue will be full if the number of incoming packets is greater than the space available in the queue. In other words, the transition probability corresponding to the state in which the queue is full can be calculated as the sum of all the probabilities that make the number of packets in the queue larger than the queue size X.

7.4.4.3 QoS Measures

To obtain the queueing performance measures, the steady state probabilities for the system states would be required. Since the size of the queue is finite (i.e., $X < \infty$), the probability matrix $\boldsymbol{\pi}$ is obtained by solving the equations $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{1} = 1$, where $\mathbf{1}$ is column matrix of one. The matrix $\boldsymbol{\pi}$ contains the steady state probabilities for the states with feasible combinations of the state variables $(\mathcal{X}, \mathcal{U}, \mathcal{B}, \mathcal{A})$. Hence, this matrix can be decomposed as $\pi(x, u, c, s)$ which is the steady state probability that the test call is allocated with c channels when the number of ongoing calls is u, there are x packets in the queue corresponding to that call and the channel state is s. Using the steady state probabilities, the various queueing performance measures can be obtained.

Average Queue Length: The mean number of packets in the queue is obtained as follows:

$$\overline{x} = \sum_{x=1}^{X} x \left(\sum_{u=1}^{C} \sum_{c=0}^{1} \sum_{\forall s} \pi(x, u, c, s) \right). \tag{7.55}$$

Packet Dropping Probability: The packet dropping probability can be obtained based on the average number of dropped packets per time slot [29]. Given that there are x packets in the queue and the queue size increases by z, the number of dropped packets is $z-(X-x)$ for $z > X-x$, and zero otherwise. The average number of dropped packets per time slot is obtained as in (7.56)

$$\overline{x}_{drop} = \sum_{x,u,b,s} \sum_{z=\max(0,X-x+1)}^{A} [\pi]_{\text{colx}(x,u,b,s)} \left(\sum_{u',b',s'} [\mathbf{p}_{x,x+z}]_{\text{colu}(u,b,s),\text{colu}(u',b's')} \right) \times \max(0, z-(x-X)) \tag{7.56}$$

where function $\text{colx}(x,u,b,s)$ indicates row of matrix $\mathbf{P}$ corresponding to x packets in queue of test call, number of ongoing calls is u, level b of number of channels allocated to test call and channel state is s. Similarly, function $\text{colu}(u,b,s)$ indicates the column of matrix $\mathbf{p}_{x,x'}$ corresponding to number of ongoing calls is u and level b of number of channels allocated to test call and channel state is s. After calculating the average number of dropped packets per time slot, we obtain the probability that an incoming packet is dropped as follows:

$$P_{drop} = \frac{\overline{x}_{drop}}{\overline{a}} \tag{7.57}$$

where $\overline{a}$ is the average arrival rate per time slot due to adaptive traffic shaping (as given by (7.45)).

Queue Throughput: It measures the average number of packets transmitted in one time slot. We calculate this measure based on the fact that if a packet from the source is not dropped due to adaptive traffic shaping, it will be transmitted eventually. Hence, the queue throughput (packets/time slot) can be obtained from

$$\eta_{nrt} = \overline{a}(1 - P_{drop}). \tag{7.58}$$

Average Packet Delay: The average packet transmission delay is defined as the number of time slots required for a packet to be transmitted since its arrival at the queue. We apply Little's law to obtain this performance measure as follows:

$$\overline{d} = \frac{\overline{x}}{\eta} \tag{7.59}$$

where η is the throughput (same as the effective arrival rate at the queue) and $\overline{x}$ is the average queue length.

Delay Distribution: To analyze the delay distribution for a packet in the queue, we utilize the concept of *absorbing Markov chain.* The general form of the transition probability matrix of an absorbing Markov chain is

$$\mathbf{P}_{abs} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{\Phi} & \mathbf{\Omega} \end{array}\right] \tag{7.60}$$

where the top most row denotes the absorbing state, $\mathbf{\Omega}$ is the transient state transition matrix, and $\mathbf{\Phi}$ is the transition matrix to the absorbing state. Note that, if there is only one absorbing state, matrix $\mathbf{\Phi}$ can be simply obtained from $\mathbf{\Phi} = \mathbf{1} - \mathbf{\Omega 1}$. If $\boldsymbol{\beta}$ denotes the initial transient state probability matrix, the probability mass function $f_{(d)}(w)$ and the distribution $F_{(d)}(w)$ for the time (i.e., the number of time slots w) required to reach the absorbing state can be expressed as follows:

$$f_{(d)}(w) = \boldsymbol{\beta}\mathbf{\Omega}^{w-1}\mathbf{\Phi 1}, \quad F_{(d)}(w) = \sum_{j=1}^{w-1} f_{(d)}(j). \tag{7.61}$$

We can establish an absorbing Markov chain from matrix $\mathbf{P}$ in (7.46) by assuming that the absorbing system state is the state at which the number of packets in the queue becomes zero. The delay for a packet can be measured as the required number of time slots (since the arrival of the packet) for the system to reach the absorbing state. Note that, in this case, there is no arrival (i.e., $\lambda = 0$) while the process moves towards the absorbing state. Therefore, we have

$$\mathbf{\Omega} = \left[\begin{array}{c|cccc} \mathbf{I} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \hline \mathbf{p}'_{1,0} & \mathbf{p}'_{1,1} & & & \\ \vdots & \vdots & \ddots & & \\ \mathbf{p}'_{N_m,0} & \mathbf{p}'_{N_m,1} & \cdots & \mathbf{p}'_{N_m,N_m} & \\ & & \ddots & \ddots & \ddots \\ & & & \mathbf{p}'_{X,X-N_m} & \mathbf{p}'_{X,X} \end{array}\right] \tag{7.62}$$

where $\mathbf{p}'_{x,x-y}$ denotes the probability matrix corresponding to successful transmission of y packets ($y \in \{0, 1, \ldots, N_m\}$) when the packet arrival rate is zero.

The initial transient state probability matrix $\boldsymbol{\beta}$ can be obtained based on the steady state probability $\boldsymbol{\pi}$ by ignoring the state with zero packet in the queue as follows:

$$\boldsymbol{\beta} = \frac{\boldsymbol{\pi}'}{\boldsymbol{\pi}'\mathbf{1}}, \quad \text{where} \quad \boldsymbol{\pi}' = \left[\begin{array}{ccc} [\boldsymbol{\pi}]_{r+1} & \cdots & [\boldsymbol{\pi}]_{r\times X} \end{array}\right]. \tag{7.63}$$

7.4.5 Model for File Transfer

We derive the distribution for file transfer delay from a mobile node to the base station. The size of a file is assumed to be F packets. To obtain the delay distribution, we use matrices $\mathbf{\Psi}(x,w)$ (i.e., $\mathbf{\Psi}(x,w) \in \mathbb{R}_{2C(N+1)^{c_{max}} \times 2C(N+1)^{c_{max}}}$) whose elements represent the probability that x packets are successfully transmitted in w time slots. If there are x packets which are to be transmitted in w slots, and y packets are successfully transmitted in the current slot, there remains $x-y$ packets to be transmitted in $w-1$ slots. Therefore, we can write the following recursive relations:

$$\mathbf{\Psi}(x,w) = \sum_{y=0}^{N_m} \mathbf{p}'_{x,x-y} \mathbf{\Psi}(x-y, w-1) \tag{7.64}$$

$$\mathbf{\Psi}(0,0) = \mathbf{I}. \tag{7.65}$$

The probability that the delay is w slots can be calculated as

$$F_{(df)}(w) = \boldsymbol{\pi}_r \mathbf{\Psi}(F+1, w)\mathbf{1} \tag{7.66}$$

where $\boldsymbol{\pi}_r$, which denotes the steady state probability corresponding to the number of calls, the number of allocated channels and the channel states, is obtained from (7.36). We can obtain probability mass function (*pmf*) of delay w as follows:

$$f_{(df)}(w) = F_{(df)}(w) - F_{(df)}(w-1), \quad \text{for } w \in \{1,2,\ldots\} \tag{7.67}$$

where $F_{(df)}(0) = 0$ since we assume that a packet requires at least one time slot to be transmitted.

Note that, we can obtain the delay distribution corresponding to the transmission of multiple files by using discrete convolution [75]. Let $f^{(i)}_{(df)}(w)$ denote *pmf* of delay w for file i. Then, the *pmf* of delay for transferring files i and j is obtained as follows:

$$f^{(i,j)}_{(df)}(k+1) = \sum_{l=0}^{k} f^{(i)}_{(df)}(l) \times f^{(j)}_{(df)}(k-l). \tag{7.68}$$

7.5 Results and Discussions

7.5.1 Parameter Setting

We consider a single-cell environment with 10 channels (i.e., $C = 10$) allocated to each service class. The set of channel allocation is $B = \{1,2,3\}$ (i.e., $c_1 = 1$ and

$c_{max} = 3$). We assume that the new call arrival rate for any specific type of call is $\rho_{(n)} = 0.25$ calls per minute and the handoff call arrival rate is half of the new call arrival rate. The mean channel holding time for both new calls and handoff calls is 20 minutes (i.e., $1/\mu = 20$).

The length of a time slot is 2 ms and the packet size is 1,080 $bits$. The packet arrival rate for a Poisson source is 2.0 packets per time slot (i.e., $\lambda = 2.0$), and the value of the traffic shaping parameter ϕ is set to 1.0. For real-time traffic, we consider a three-state MMPP source (i.e., $S = 3$) with the following parameters:

$$\mathbf{U} = \begin{bmatrix} 0.6 & 0.4 & \\ 0.2 & 0.4 & 0.4 \\ & 0.3 & 0.7 \end{bmatrix}, \mathbf{\Lambda} = \kappa \begin{bmatrix} 1 & & \\ & 2 & \\ & & 1.5 \end{bmatrix} \tag{7.69}$$

where κ indicates the traffic intensity of an MMPP source.

We consider adaptive modulation with two transmission rates (i.e., $N = 2$) where the maximum transmission rate is achieved for 4-QAM. The values of a_n and g_n for fitting the packet error rate curve are the same as in [25]. For fading channel, we assume a Nakagami-m channel with $m = 1.1$. The queue size at a mobile for the non-real-time traffic is 30 packets (i.e., $X = 30$). The average SNR for each channel is 15 dB (i.e., $\overline{\gamma} = 15$), the packet error rate is 0.03 (i.e., $\theta = 0.97$), and $f_d = 15$ Hz. Note that, we vary some of the above parameters according to the evaluation scenarios, while the rest remain fixed according to the aforementioned setting.

In order to validate the correctness of our proposed model, an event-driven simulator is used to obtain the performance results in a single-cell environment. The call arrival follows a Poisson process, and channel holding time for both handoff calls and new calls is assumed to be exponentially distributed. The queueing dynamics (due to packet arrival and packet transmission) at the mobile is simulated on a time slot basis. The number of packets that can be transmitted in each time slot depends on the current number of allocated channels and the instantaneous SNR of each channel. We keep track of the queue for the test call since the call arrives until it leaves the cell or the call terminates.

7.5.2 Numerical and Simulation Results

The analytical framework presented above can be used to investigate the QoS performances for the different types of traffic under different system parameter settings. We show how a cross-layer evaluation of the network performance can be performed using the proposed analytical framework. Specifically, we demonstrate the impacts of call-level parameters, channel quality and user mobility on the packet-level performances. Typical results obtained from the analytical model are validated by simulations.

7.5.2.1 QoS Performance for Real-Time Traffic

Typical variations in new call blocking and handoff call dropping probabilities with new call arrival rate are shown in Figure 7.2. As expected, when the new call arrival rate increases, new call blocking and handoff call dropping increase. Similar performance is observed for other service classes as well.

The impact of new call arrival rate on the average transmission rate for a mobile is shown in Figure 7.3. We observe that the lower the new call arrival rate, the higher is the transmission rate. Due to the adaptive channel allocation, when the new call arrival rate is low, more channels can be allocated to the ongoing calls. Consequently, the transmission rate increases.

Figure 7.4 shows variations in the packet loss rate for the real-time traffic under different channel holding time. The packet loss rate increases with increasing traffic intensity. Again, it decreases as the channel holding time decreases, because ACA can allocate more number of channels to a mobile which results in higher transmission rate.

7.5.2.2 QoS Performance for Non-Real-Time Traffic

The queue-length distributions under different packet arrival rates and without any traffic shaping (i.e., $\phi = \infty$) are shown in Figure 7.5(a). As expected, the queue length increases with increasing packet arrival rate. Typical variations in the average queueing delay for a packet under different packet arrival rate and channel quality (i.e., average SNR) are shown in Figure 7.5(b). When the channel quality becomes better, the transmitter can utilize higher modulation level, and therefore, more packets can be transmitted in one time slot. Consequenty, the average delay decreases with increasing

average SNR. The simulation results for all the above cases follow the numerical results very closely.

For a tagged packet in the queue, the delay distribution is shown in Figure 7.6 under different packet arrival rate. This delay statistics obtained from the analytical model would be useful for choosing parameters such as the packet arrival rate for moderately delay-sensitive traffic so that the performance requirements can be satisfied. For example, to maintain the delay smaller than 10 time slots, the packet arrival rate should be less than or equal 2.0 packets per time slot (Figure 7.6).

Impacts of new call arrival rate and channel holding time on packet dropping probability are shown in Figure 7.7. Since both of these call-level parameters affect the number of ongoing calls, they impact the queueing performances. As the call arrival rate and/or channel holding time increases, fewer number of channels are allocated to a call. We observe that, when the channel holding time is not too high (e.g., $1/\mu = 10$), the ACA algorithm can efficiently allocate the available channels among the ongoing calls, and therefore, the packet dropping probability decreases significantly (Figure 7.7(a)). Also, higher packet error rate results in higher packet dropping probability due to buffer overflow (Figure 7.7(b)). Note that, with an infinite persistent ARQ protocol, erroneous packets need to be retransmitted until they are successfully received at the base station.

For non-real-time traffic, the effects of traffic shaping on the achievable packet arrival rate and packet dropping performances under varying channel quality are shown in Figure 7.8. Since the traffic shaper adaptively adjusts the packet arrival rate according to the number of allocated channels and channel quality, achievable packet arrival rate at the queue is higher while the packet dropping probability is lower compared to the case with no traffic shaping. Specifically, if ACA allocates more number of channels, the traffic shaper will increase the packet arrival rate as much as the achievable transmission rate allows. However, if the number of allocated channels is reduced, the traffic shaper will control the packet arrival rate accordingly. In this way, the adaptive traffic shaper at the mobile exploits the information on both the number of allocated channels and the channel quality to efficiently control the packet arrival rate so that the QoS performances can be maintained at the desired level.

7.5.2.3 QoS Performance for Best-Effort Traffic

Both the call-level parameters (i.e., channel holding time) and channel quality (i.e., average SNR) affect the distribution of file transfer delay (Figure 7.9). Specifically, longer channel holding time and/or lower average SNR result in higher delay. From Figure 7.9(a), we observe that with 4-QAM modulation (i.e., 2 symbols per Hz), the highest value for the cumulative probability of delay corresponds to $\lceil 40/(3 \times 2) \rceil = 7$, $\lceil 40/(2 \times 2) \rceil = 10$, and $\lceil 40/(1 \times 2) \rceil = 20$ time slots when the number of allocated channels is 3, 2, and 1, respectively. Similar results are obtained under different file sizes (Figure 7.9(b)).

7.5.2.4 Impact of User Mobility

We demonstrate how the presented analytical framework can be used to quantitatively analyze the impacts of user mobility and cell-geometry on packet-level QoS. For this, we calculate handoff call arrival rate and channel holding time (i.e., the time for which a call occupies channel(s) in a particular cell) from call holding time (i.e., the time length of a call) and cell dwell time (i.e., the time a call spends in a given cell before it is handed off to another cell). If we assume that the call holding time (T_c) and the cell dwell time (T_d) are exponentially distribution with average $1/\mu_c$ and $1/\mu_d$, respectively, channel holding time is also exponentially distributed with average $1/\mu = \frac{1}{\mu_c+\mu_d}$ [76].

The cell dwell time for a call can be calculated based on the cell size and the speed of the mobile. Specifically, if μ_d is the average rate at which a mobile with speed V crosses a series of cells each having an area S_{cell} and boundary length L_{cell}, μ_d can be obtained as follows [76]: $\mu_d = \frac{VL_{cell}}{\pi S_{cell}}$. As a special case, if we consider circular cell of radius r, we have $\mu_d = \frac{2V}{\pi r}$. Similarly, the probability of handoff can be obtained from $P_{hoff} = \frac{\mu_d}{\mu_d+\mu_c}$, and the handoff call arrival rate can be approximated from new call arrival rate as follows:

$$\rho_{(h)} \approx \frac{P_{hoff}}{1 - P_{hoff}} \rho_{(n)} \approx \frac{\mu_c}{\mu_d} \rho_{(n)}. \tag{7.70}$$

Typical variations in packet dropping probability under varying mobile speed and call holding time are shown in Figure 7.10(a) (considering a circular cell-geometry with cell radius of 400 m). We observe that the call holding time has significant

impact on the queueing performance. Specifically, longer call holding time results in higher packet dropping probability. The number of available channels also impacts the packet dropping performance (Figure 7.10(b)). The queueing performance may not be impacted significantly under varying mobile speed when the number of available channels is sufficiently high compared to the system load.

7.6 Application of the Analytical Model

7.6.1 Optimal Parameter Setting

Given the desired QoS performances, using the analytical framework, we are able to obtain the optimal setting for the system parameters. For example, to maintain a target packet dropping probability (e.g., $P_{drop}^{tar} \leq 0.1$), given the other system parameters, we can formulate the following optimization problem:

$$\text{Minimize:} \quad |P_{drop}(\lambda) - P_{drop}^{tar}| \tag{7.71}$$

where the packet dropping probability is a function of the packet arrival rate $P_{drop}(\lambda)$. With this formulation, the decision variable is λ. We can use the *Golden Section Search* [77] algorithm to find the minima of (7.71) and the corresponding optimal value of $\tilde{\lambda}$. Starting at a given initial interval a_1 and b_1 ($a_1 < b_1$), this algorithm (Algorithm 7.6.1) proceeds by evaluating the function $f(x) = |P_{drop}(x) - P_{drop}^{tar}|$.

Algorithm 7.6.1: GOLDEN SECTION SEARCH($inputs : a_1, b_1, tol$)

$b_1 \leftarrow a_1 + (1-\tau)(b_1 - a_1), F_b \leftarrow f(b_1)$
$d_1 \leftarrow b_1 - (1-\tau)(b_1 - a_1), F_d \leftarrow f(d_1)$
$\tilde{\lambda} \leftarrow b_1$
repeat
if $F_b < F_d$

$$\textbf{then} \begin{cases} \tilde{\lambda} \leftarrow b_k \\ a_{k+1} \leftarrow a_k, b_{k+1} \leftarrow d_k, d_{k+1} \leftarrow b_k \\ b_{k+1} \leftarrow a_{k+1} + (1-\tau)(b_{k+1} - a_{k+1}) \\ F_d \leftarrow F_b, F_b \leftarrow f(b_{k+1}) \end{cases}$$

$$\textbf{else} \begin{cases} \tilde{\lambda} \leftarrow d_k \\ a_{k+1} \leftarrow b_k, b_{k+1} \leftarrow b_k, b_{k+1} \leftarrow d_k \\ d_{k+1} \leftarrow b_{k+1}(1-\tau)(b_{k+1} - a_{k+1}) \\ F_b \leftarrow F_d, F_d \leftarrow f(d_{k+1}) \end{cases}$$

until $b_{k+1} - a_{k+1} < tol$
return $(\tilde{\lambda})$

To use the above algorithm, we set $\tau = \frac{3-\sqrt{5}}{2}$, $tol = 10^{-3}$, $a_1 = 0.01$, and $b_1 = 5$. Figure 7.11 shows variations in the maximum arrival rate under different channel holding time when the packet dropping probability is less than 0.1. As expected, the maximum allowable arrival rate increases as the average SNR increases, and smaller channel holding time results in higher maximum packet arrival rate. This arrival rate can be used to set the traffic shaping parameter ϕ accordingly.

7.6.2 Optimal Allocation of Channel Resources at the Base Station for Multimedia Services

We can formulate an optimization problem to obtain an optimal partitioning of the total number of available channels in the cell C_{cell} for each of the service classes (i.e., C_{rt} channels for real-time service, C_{nrt} channels for non-real-time service and C_{be} channels for best-effort service). For call-level performance measures, the objective of this optimal channel partitioning problem is to maximize the average number of channels allocated to the ongoing calls (i.e., $\overline{c}_{rt}$, $\overline{c}_{nrt}$ and $\overline{c}_{be}$) while maintaining the new

call blocking and the handoff call dropping probabilities below the target thresholds (i.e., P_{nb}^{tar} and P_{hd}^{tar}, respectively). Mathematically, the formulation can be expressed as follows:

$$\text{Maximize:} \quad w_{rt}\overline{c}_{rt} + w_{nrt}\overline{c}_{nrt} + w_{be}\overline{c}_{be} \tag{7.72}$$

$$\text{Subject to:} \quad P_{nb}^{(rt)}, P_{nb}^{(nrt)}, P_{nb}^{(be)} \leq P_{nb}^{tar} \tag{7.73}$$

$$P_{hd}^{(rt)}, P_{hd}^{(nrt)}, P_{hd}^{(be)} \leq P_{hd}^{tar} \tag{7.74}$$

$$C_{rt} + C_{nrt} + C_{be} = C_{cell} \tag{7.75}$$

where w_{rt}, w_{nrt}, and w_{be} denote the weights corresponding to real-time, non-real-time, and best-effort service classes, respectively. For numerical results, we use $\rho_{(n)}^{(rt)} = \rho_{(n)}^{(nrt)} = 0.1$ for new call arrival rates of real-time and non-real-time traffic and we assume that the handoff arrival rate is half of the new call arrival rate. Also, we set $1/\mu_{(rt)} = 10$, $1/\mu_{(nrt)} = 15$, $1/\mu_{(be)} = 10$, and $w_{rt} = 3$, $w_{nrt} = 2$, $w_{be} = 1$. The total cell capacity is 30 channels (i.e., $C_{cell} = 30$) and target thresholds for new call blocking and handoff call dropping probability are 0.1 and 0.05 (i.e., $P_{nb}^{tar} = 0.1$ and $P_{hd}^{tar} = 0.05$), respectively. We vary the new call arrival rate $\rho_{(n)}^{(be)}$ for the best-effort traffic and search for the optimal partitioning. The number of channels for each service class and the average number of allocated channels per call are shown in Figures. 7.12(a) and (b), respectively.

As expected, when the traffic load for the best-effort service class increases, more number of channels need to be allocated to this service class to maintain the new call blocking and the handoff call dropping probabilities below the target thresholds. In this example, with the chosen values of the weights, as the traffic load increases, best-effort service takes channels from the non-real-time service first as long as the call-level QoS for the non-real-time traffic can be maintained, and then it takes channels from the real-time service. Consequently, the number of allocated channels per call for each service class decreases as the traffic load in the cell increases. In the above formulation, the priority for each service class can be chosen by properly choosing the weighting values. Specifically, in this case, real-time and best-effort services receive the highest and the lowest number of channels (Figure 7.12(b)), respectively.

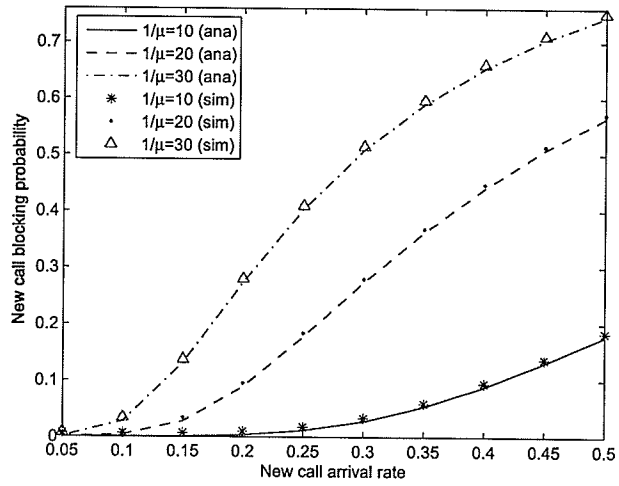
Note that, similar optimization formulation can developed considering both the call-level and the packet-level QoS together in the objective function and modifying

the constraints accordingly.

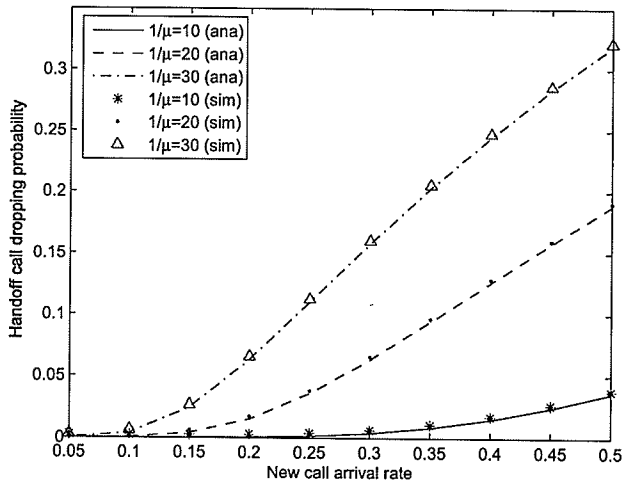
7.7 Chapter Summary

We have presented a novel analytical framework for cross-layer performance evaluation in a multiuser cellular wireless network under Markov fading channel model. With guard channel-based CAC and fairness-based ACA at the network layer, ARQ-based error control at the link layer, and adaptive modulation at the physical layer, we have analyzed the QoS performances for real-time, non-real-time, and best-effort traffic. Specifically, for non-real-time traffic various queueing performance measures (e.g., queueing delay and packet dropping probability) have been obtained while for real-time and best-effort traffic packet loss rate and packet delay distribution have been obtained. For real-time and non-real-time traffic, the framework also includes an adaptive traffic shaping mechanism which takes the number of allocated channels and channel quality into account to control the packet arrival rate so that the QoS performances can be maintained at a desired level.

We have presented extensive performance evaluation results obtained from both analysis and simulation. The results have shown how the physical layer parameters (e.g., channel quality) and the call-level parameters (e.g., call arrival rate and channel holding time) impact the packet-level QoS. Impact of user mobility on packet-level performance has been demonstrated through a basic mobility model. The analytical framework is, however, general enough to be used with any mobility model. We have also demonstrated applications of the proposed framework in obtaining the optimal system parameter setting and the optimal channel partitioning at the base station for different multimedia services under QoS constraint. After all, the proposed analytical framework will be useful for multimedia wireless network system design and engineering.



(a)



(b)

Figure 7.2. *(a) New call blocking probability and (b) handoff call dropping probability from analytical model and simulation.*

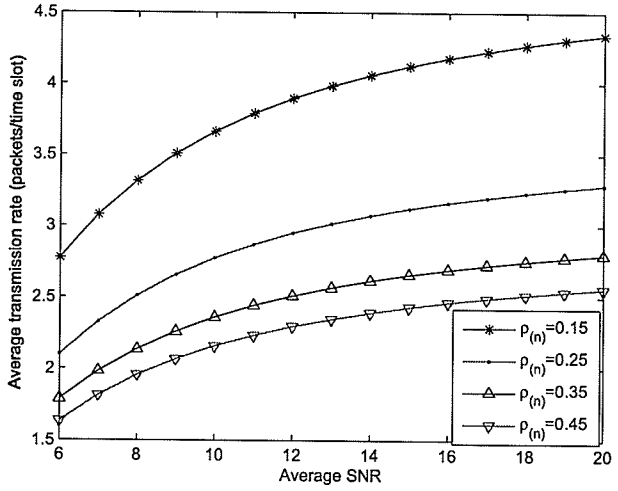


Figure 7.3. *Transmission rate of mobile under different new call arrival rates.*

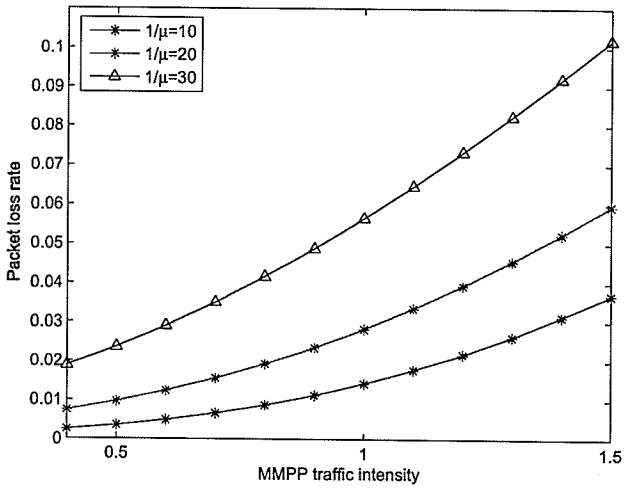
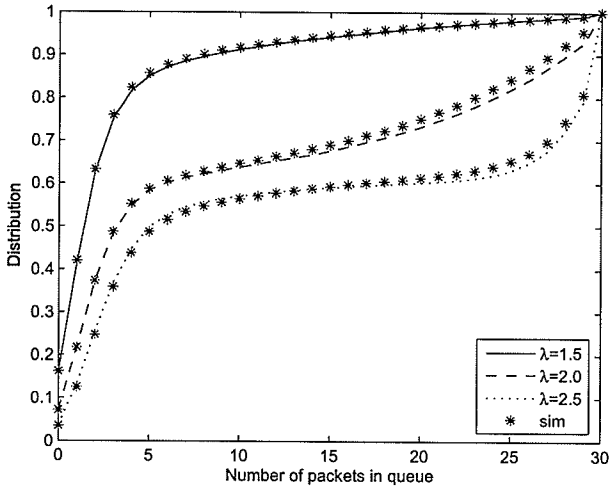
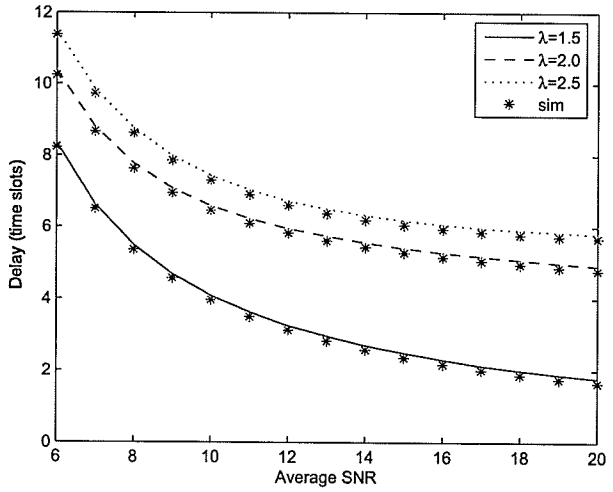


Figure 7.4. *Packet loss rate of real-time traffic.*



(a)



(b)

Figure 7.5. *(a) Queue distribution and (b) average queueing delay obtained from analytical model and simulations.*

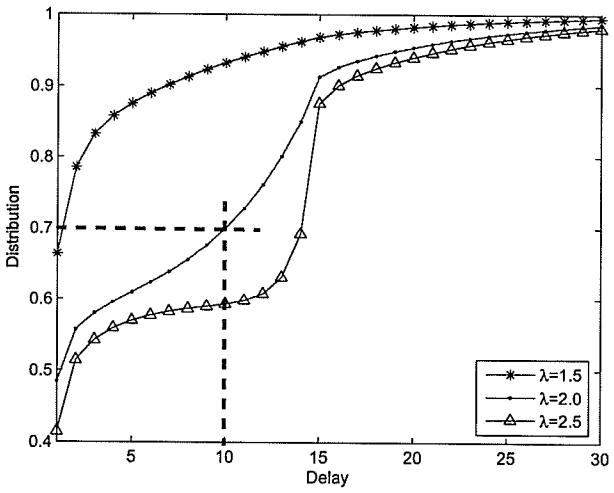
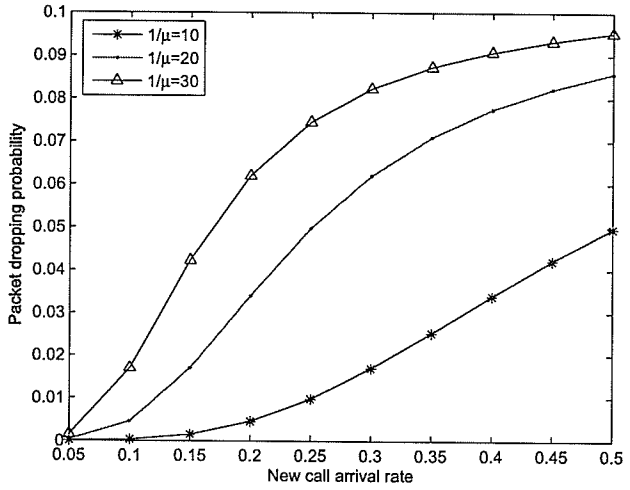
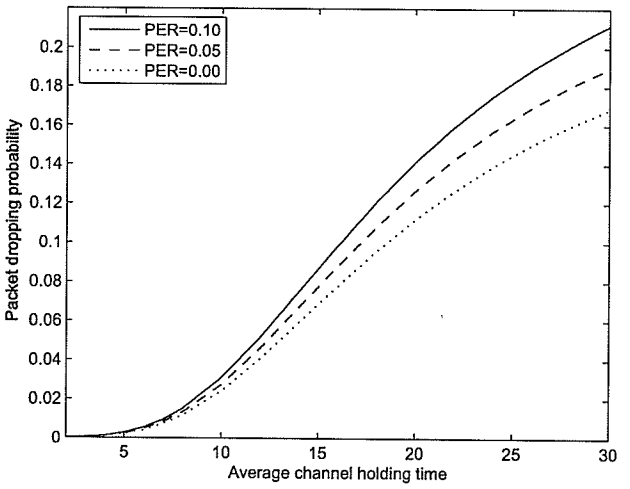


Figure 7.6. *Delay distribution.*

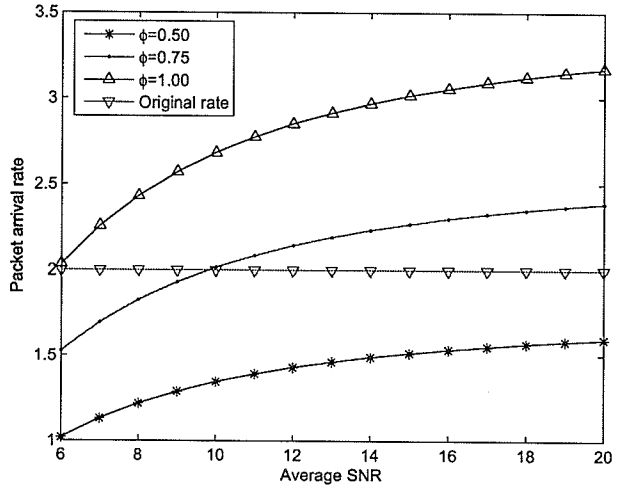


(a)

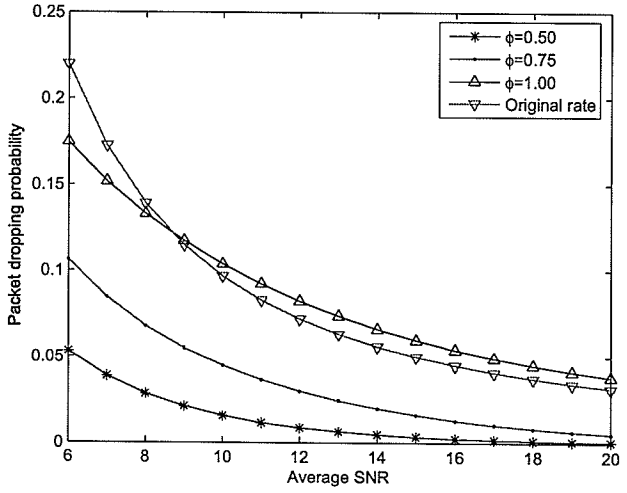


(b)

Figure 7.7. *Packet dropping probability under varying (a) call arrival rate and (b) channel holding time.*

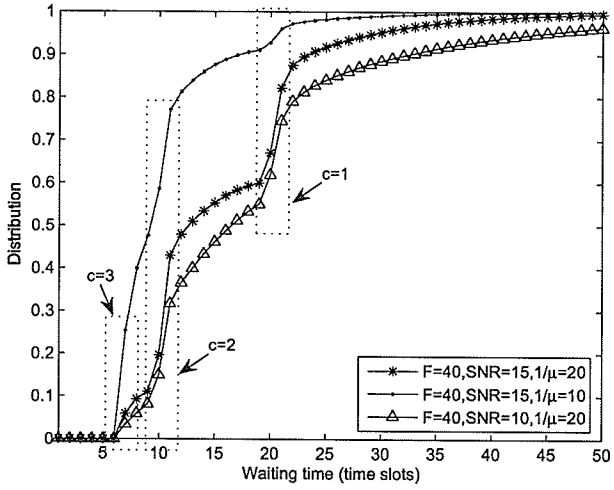


(a)

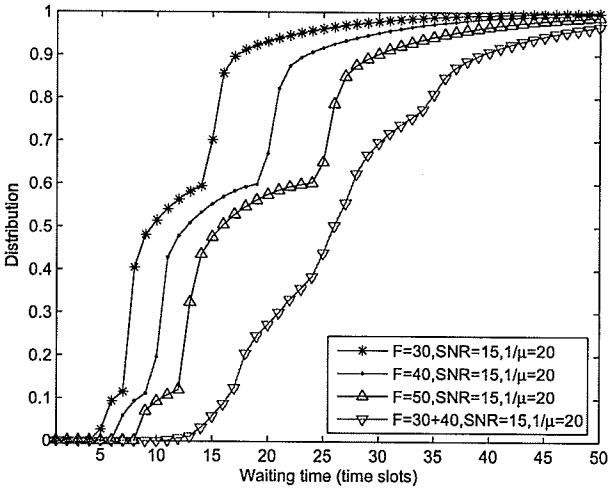


(b)

Figure 7.8. *(a) Packet arrival rate and (b) packet dropping probability of Poisson traffic source with adaptive traffic shaper.*

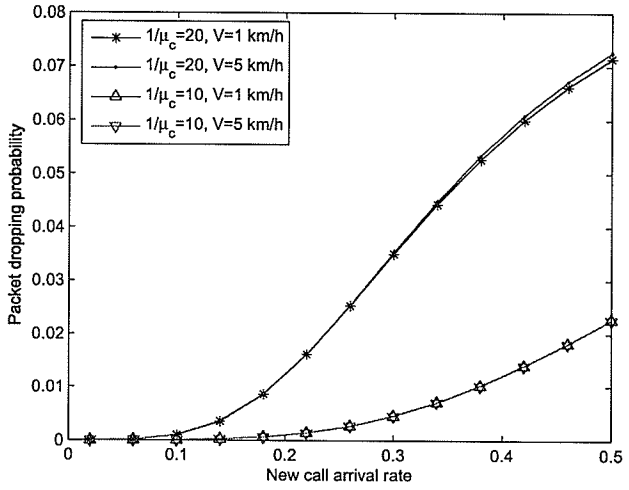


(a)

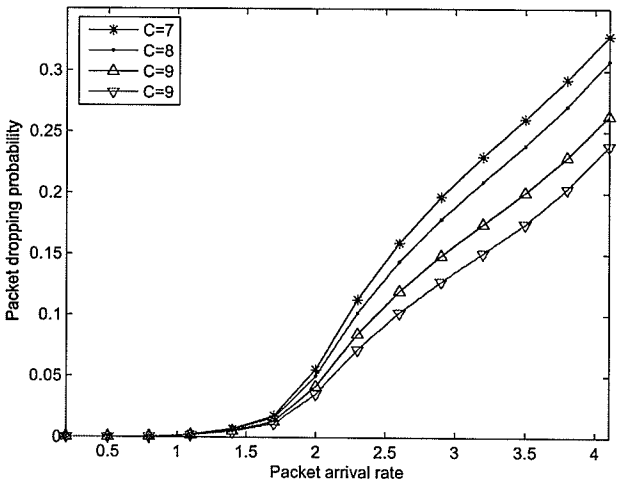


(b)

Figure 7.9. *Waiting time distribution of file transfer traffic under (a) different call and channel parameter settings and (b) different file sizes.*



(a)



(b)

Figure 7.10. *Variations in packet dropping probability with (a) channel holding time and (b) number of channels.*

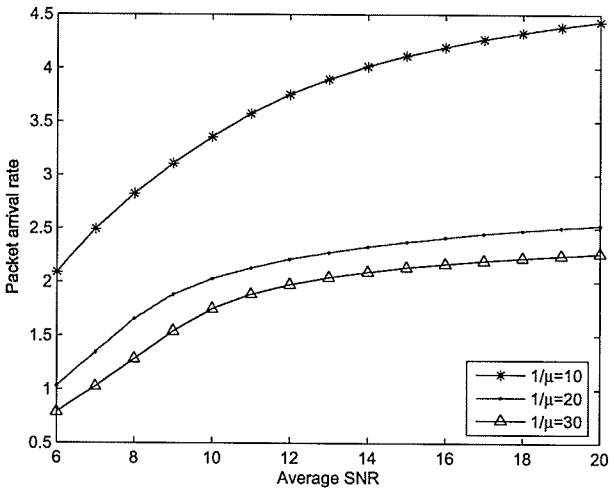
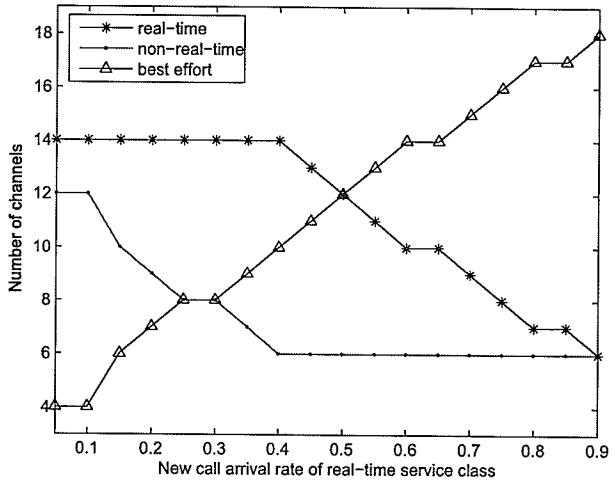
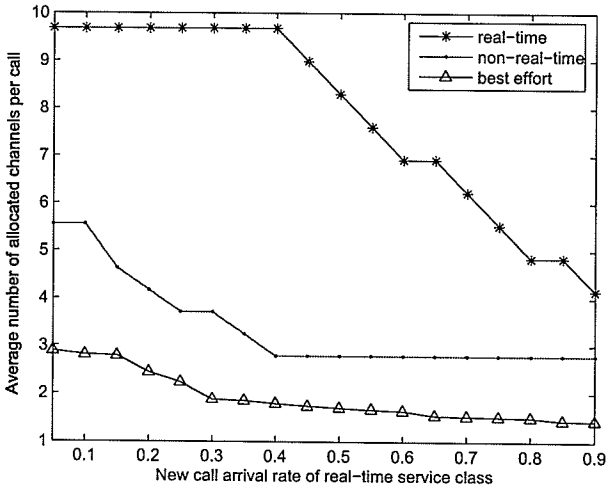


Figure 7.11. *Maximum packet arrival rate to maintain packet dropping probability less than 0.1.*



(a)



(b)

Figure 7.12. *Variations in (a) channel pool partitioning and (b) average number of allocated channels per call for each service class.*

Chapter 8

On Optimizing Token Bucket Parameters at the Network Edge Under Generalized Processor Sharing (GPS) Scheduling

8.1 DiffServ Wireless Edge Router

Quality-of-service (QoS) is one of the most important issues for providing multimedia services over the Internet. Normally, the QoS guarantees between service providers and users are determined by the service level agreement. Specifically, the data streams from users are required to conform to some 'shape' and the service providers must guarantee a certain level of throughput, maximum delay, and/or maximum packet loss rate for the agreed upon traffic pattern.

Different applications require different QoS guarantees and service differentiation is achieved through scheduling algorithms. Group-based service differentiation approaches such as *DiffServ* [1] aggregate traffic flows with similar QoS guarantees into a small number of groups and traffic scheduling is operated based on these groups.

A number of scheduling disciplines were proposed ([2] and [84]) in the literature. Most of them were designed to provide delay and throughput guarantees as well as fairness. Generalized processor sharing (GPS) is a work-conserving scheduling discipline which can provide fairness between connections. When the traffic sources are policed and shaped through token-bucket shapers, with GPS scheduling, different

levels of QoS are provided to different classes of traffic by assigning different weights to the classes. However, in order to guarantee QoS, an admission control policy is also required. A number of CAC schemes for GPS schedulers were developed [85]-[87]. The role of the CAC is to control the number of admitted sources and their weights in order to provide guarantees on delay and throughput. In addition to controlling the weights, the traffic shaper parameters can be adjusted (while the weights remain fixed) to achieve the desired QoS performances

In this chapter, we investigate the optimal setting of the token bucket parameters for traffic sources with known profiles, such that the minimum delay bound is achieved. We consider traffic sources for which trace data are provided. This scenario could be typical for some stored video or some other on-demand applications. While the seminal results in [2] require a linear form for the traffic bounding function, if we have access to the traffic trace in advance, we can obtain a more detailed , non-linear description of a function which bounds the amount of traffic from the source in an interval of any duration. In such a case, if we want to apply the results of [2], we must find a linear function which bounds this traffic profile. There are many choices for such a linear function, and we address the problem of the choice of that function.

We proceed by choosing one source at a time and performing a local search in the parameter space which minimizes the delay bound for that source. We also introduce a second method, which we refer to as the *composite delay envelope* method which is based on a random selection of token bucket parameters and a process of combining the different envelope functions which are used in [2]. In addition, because the trace of the traffic is provided,we can use the non-linear traffic bound to obtain slightly better results than we could obtain by assuming a greedy, linear bounded source.

8.2 System Model

We consider a system which consists of several traffic sources, a generalized process sharing (GPS) scheduler, and an output link. The structure of the system is shown in Figure 8.1. Each traffic source is controlled by a token bucket shaper. The service rate for source i is determined by the weight ϕ_i. We assume, as in [2], that the incoming link bandwidths and buffer sizes for each source are infinite.

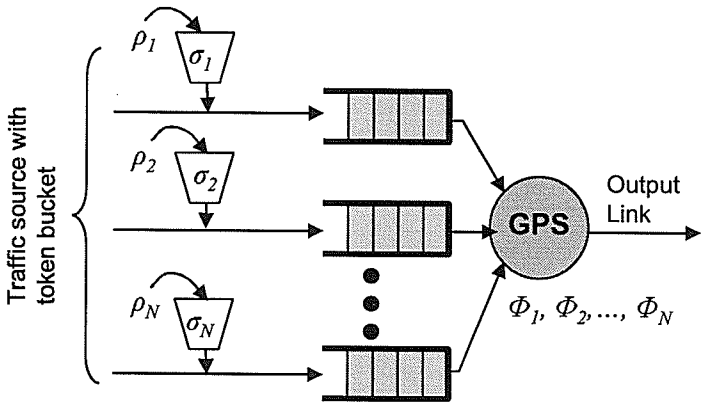


Figure 8.1. *System model*

8.2.1 Generalized Processor Sharing (GPS) Scheduler and the Delay Bounds

8.2.1.1 Basic Delay Bound for GPS Scheduler

With N sessions, the GPS scheduling discipline uses the weights ϕ_i ($i \in \{1, \ldots, N\}$) to determine $S_i(\tau, t]$, the amount of data from session i served by the GPS server in the time interval $(\tau, t]$. The operation of the GPS server satisfies:

$$\frac{S_i(\tau, t)}{S_j(\tau, t)} \geq \frac{\phi_i}{\phi_j}, \quad j = 1, 2, \ldots, N. \tag{8.1}$$

If C denotes the total service rate, session i will have a guaranteed minimum offered service rate (when it is busy) of

$$g_i = \frac{\phi_i}{\sum_j \phi_j} C. \tag{8.2}$$

Token bucket traffic shapers are used to shape and police traffic entering into the network. A token bucket shaper is parameterized by the bucket size σ_i which limits the maximum burstiness of the incoming traffic, and the token rate ρ_i, which is the bucket filling rate. We denote the total traffic arriving from source i in the interval $[t_1, t_2]$ by the function $A_i(t_1, t_2)$. Traffic coming from a token bucket policed source

with bucket size and filling rate σ_i and ρ_i will satisfy the condition

$$A_i(t_1, t_2) \leq \sigma_i + \rho_i(t_2 - t_1).$$

If we have equality in the above relation, so that $A_i(t_1, t_2) = \sigma_i + \rho_i(t_2 - t_1)$, for all $t_2 > t_1$, then we say that source i is greedy from time t_1. The main result in [2] provides that, for a number of token bucket constrained sources with parameters σ_i and ρ_i, if the condition $\sum_i \rho_i < 1$ is satisfied, then the maximum delay for any source is bounded above by the maximum delay for that source in a scenario where all sources are greedy from time zero, the beginning of a system busy period. From this, we have the following bound on packet delay:

$$D_i = \frac{\sigma_i}{g_i}. \tag{8.3}$$

8.2.1.2 A Tighter Delay Bound

This delay bound, the maximum queueing delay a session can experience in a GPS scheduler, obtained from (8.3) is loose, because it assumes that the service rate for session i is kept constant at a rate g_i till the backlog is cleared. However, the service rates for backlogged sessions may increase when other sessions have cleared their backlog. This is because the excess resources from the sessions which have cleared their backlog are distributed among the currently backlogged sessions [88].

Figure 8.2 illustrates a case where the delay bound (D_i) given by the guaranteed rate is greater than the actual worst case delay, denoted in the figure by t_i. When other sessions empty their backlog, the slope of the attained service curve increases. At time t_1, one session has cleared its backlog, and the service rate of source i increases. Again, at time t_2 another session has cleared its backlog and the service rate for session i increases. At t_3, this session has cleared backlog, so t_i provides a tighter delay bound than D_i.

In [88], a method was outlined for calculating this tighter delay bound based on the main theorem in [2]. This provides the computational details for computing the maximum delay for each source when each source is greedy from time zero.

We omit the computational details here, but observe that the calculation of the delay bound for a source i depends on the values of σ_k, ρ_k, and ϕ_k, for all sources k

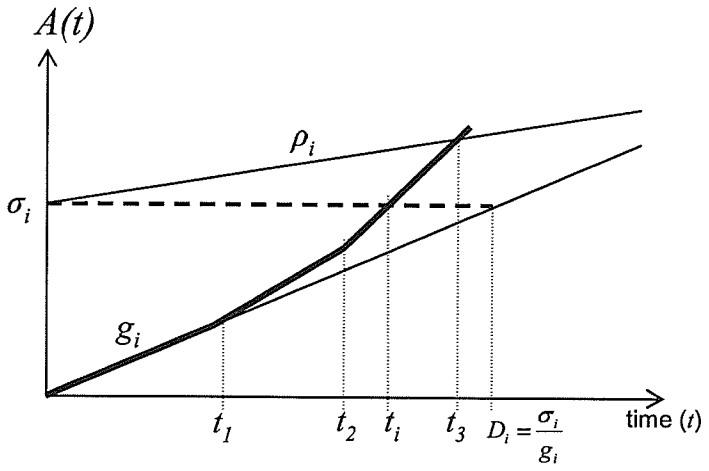


Figure 8.2. *Service and arrival of traffic for four greedy sessions served by a GPS server.*

and therefore, we can express the delay bounds for all the sources by $D_i^*(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\phi})$ in which $\boldsymbol{\sigma}$, $\boldsymbol{\rho}$, and $\boldsymbol{\phi}$ denote vectors containing the values of σ, ρ, and ϕ, respectively.

8.2.2 Queue Length and Delay Envelope

Stronger results still can be obtained from [2] in the form of a delay or queue length envelope which bounds the delay or queue length experienced by an arriving packet by a function of the time since the beginning of the last busy period for that source. This result is contained in the following Lemma.

Lemma 8.2.1 *(Lemma 10 in [2]) Suppose that time t is contained in a session p busy period that begins at time τ. Then*

$$\hat{S}_p(0, t-\tau) \leq S_p(\tau, t)$$

where $S_p(t_1, t_2)$ and $\hat{S}_p(t_1, t_2)$ represent (respectively) the service attained by session p and the service obtained by session p in a scenario where all sessions are greedy from time zero.

We can express the effect of this result on the queue length via the following corollary.

Corollary 8.2.2 *Suppose that time t is contained in a session p busy period that begins at time τ. Also assume that the function A_p which gives the session arrivals ($A_p(t_1, t_2)$ represents the amount of arrivals in the time interval $(t_1, t_2]$) is bounded by some function $B_i(t)$ so that $A_p(t_1, t_2) \leq B_i(t_2 - t_1)$ for all (t_1, t_2) such that $t_1 \leq t_2$. Then*

$$Q(t) \leq B_i(t - \tau) - \hat{S}_p(0, t - \tau).$$

In [2] and [88], the case of $B_i(t) = \sigma_p + \rho_p t$ in Corollary 8.2.2, was considered, and the calculations outlined in [88] effectively calculated the envelope function $B_i(t - \tau) - \hat{S}_p(0, t - \tau)$ which is piecewise linear when B_i is linear. A similar result can be obtained for the delay of a packet. If $D_p(s)$ represents the delay experienced by a packet arriving at time s then

$$D_p(s) = inf\{t \geq s : S_p(0, t) = A_p(0, s)\} - s.$$

But if we combine Lemma 8.2.1 with the bound $A_p(t_1, t_2) \leq B_i(t_2 - t_1)$, we have

$$D_p(s) \leq \hat{D}_p(s) = inf\{t \geq s : \hat{S}_p(0, t - \tau(s)) = B_i(s - \tau(s))\} - s$$

where $\tau(s)$ represents the last busy period start before time s. If the bounding function B_i is linear, the delay envelope $\hat{D}_p(*)$ will be piecewise linear as is the queue length envelope function.

We are interested in the case where the bounding function $B_i(*)$ is not linear. In this case, we can apply the results of [2] and [88] to calculate a bound on $\hat{S}_p(0, t - \tau)$ for any choice of the parameter set $\cup_i\{\sigma_i, \rho_i\}$ as long as $B_i(t) \leq \sigma_p + \rho_p t$ for all sessions p. We attempt to find the best choice of these parameters, in order to obtain the best bound possible in the case where $B_i(*)$ is not linear. We can also use the bound in Corollary 8.2.2 directly, and need not replace B_i in Corollary 8.2.2 by a linear bound except for in the calculation of $\hat{S}_p(0, t - \tau)$.

8.2.3 Formulation of the Optimization Problem

We formulate an optimization problem to obtain σ_i and ρ_i for source i ($i \in \{1, \ldots, N\}$) which minimizes the delay bound for a particular source j.

As an example, we use some traces of video traffic sources to illustrate the process. To determine possible values for the σ_i and ρ_i, we first have to calculate the maximum

amount of traffic $B_i(\tau)$,which can arrive from source i during a time interval or length τ as follows:

$$B_i(\tau) = \max_s \int_s^{s+\tau} A_i(s, s+\tau)\mathrm{d}s, \quad 0 \leq s \leq c \tag{8.4}$$

where c is the end time of the sampled traffic in the trace file. Figure 8.3 shows an example of traffic arrival function $A_i(0,t)$ and $B_i(\tau)$ as a function of time t and duration τ respectively. Given $B_i(\tau)$ for any value of ρ_i, we can obtain the minimal value of σ_i such that $B_i(\tau)$ is bounded by the corresponding linear function:

$$\sigma_i = \max_\tau \left[B_i(\tau) - \rho_i \tau\right]. \tag{8.5}$$

Since, here, σ_i is a function of ρ_i, we use the notation $\sigma *_i (\rho_i)$ and hence $\boldsymbol{\sigma} = [\sigma *_1 (\rho_1), \sigma *_i (\rho_2), \ldots, \sigma *_N (\rho_N)]$. An example plot of σ_i against ρ_i is shown in Figure 8.4.

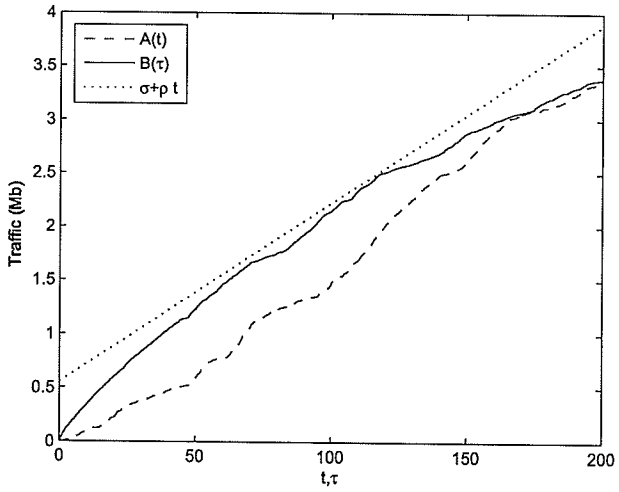


Figure 8.3. *An example of maximum amount of traffic arrival $B_i(\tau)$.*

With N sources, we can formulate an optimization problem to obtain the minimum delay bound for a particular source.The optimization problem is defined as follows:

$$\text{minimize} \quad D_i^*(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\phi}) \tag{8.6}$$

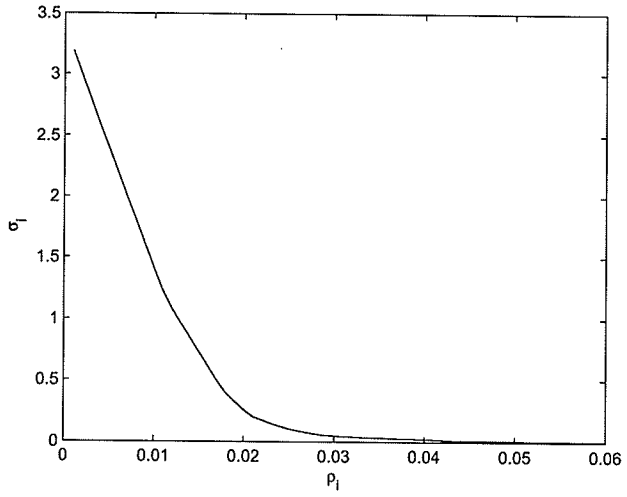


Figure 8.4. *Variation in σ_i with ρ_i according to $B_i(\tau)$.*

subject to

$$\sum_{i=1}^{N} \rho_i \leq 1 \tag{8.7}$$

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}*(\boldsymbol{\rho}) \tag{8.8}$$

given

$$\boldsymbol{\sigma}*(\boldsymbol{\rho}) \quad \text{and} \quad \boldsymbol{\phi}. \tag{8.9}$$

In other words, we want to find the optimal point on the curve illustrated in Figure 8.4 such that the minimum delay bound is achieved. To solve this optimization problem, we use the Neldel-Mead simplex (direct search) method [62]. We perform the optimization separately for each source, where in each case, the objective function represents the maximum delay bound for the current source of interest.

8.2.4 Composite Delay Envelope

We also introduced an alternate method for finding delay bounds for sources which have non-linear traffic bounding functions. We introduce the concept of a *composite*

delay envelope. To obtain this composite delay envelope, we generate a number of instances of the delay envelope by randomly choosing ρ_i, and setting σ_i accordingly, as in the previous section. Rather than performing a local search in the space of $\boldsymbol{\sigma}*(\boldsymbol{\rho})$ with the maximum delay bound as objective function, we choose a number of points in the space at random. Each point provides a delay envelope. The intersection of all of these delay envelopes yields a composite delay envelope which is also a delay envelope.

In order to cover the space of possible values of $\boldsymbol{\rho}$ in some reasonable fashion, we chose to parameterize the search in terms of the sum $y = \sum_i \rho_i$ which must be smaller than 1.0 in order for the delay envelope obtained to be valid (i.e., this is the condition that all sources are able to clear their backlog). At each iteration, we chose N values X_i uniformly at random in the interval $[0, 1]$ and set

$$x_i = \frac{X_i(y - \sum_i L_i)}{\sum (U_i - L_i)X_i} \tag{8.10}$$

$$\rho_i = L_i + (U_i - L_i)x_i \tag{8.11}$$

for $i \in \{1, \ldots, N\}$, where U_i and L_i are the maximum and minimum useful values of ρ_i. The maximum and minimum values of ρ_i can be determined from

$$U_i = B_i(1), \tag{8.12}$$

$$L_i = A_i(c)/c \tag{8.13}$$

where c is ending time.

This guarantees that $\sum_i \rho_i = y$ and that the values of ρ are within the correct range of interest, for each iteration. Each random instance of $\boldsymbol{\rho}$ gives a different delay envelope for the particular source. Let the total number of random instances be K and let $\hat{D}_p^k(t)$ denote the delay envelope of source i from random instance k at time t. To obtain the composite delay envelope for a particular source j, we take the minimum value of $\hat{D}_p^k(t)$ calculated from the set of random instances $\boldsymbol{\rho}_k$ as follows:

$$\hat{D}_j^{com}(t) = \min_k \hat{D}_p^k(t), \quad 0 \leq t \leq c, k \in \{1, \ldots, K\}. \tag{8.14}$$

Note that, from a random $\boldsymbol{\rho}_k$, the corresponding $\boldsymbol{\sigma}_k$ can be obtained from the function $\boldsymbol{\sigma}*(\boldsymbol{\rho})$ which is, in turn obtained from the bounding functions $B_i(*)$. An illustration of this composite delay envelope is shown in Figure 8.5.

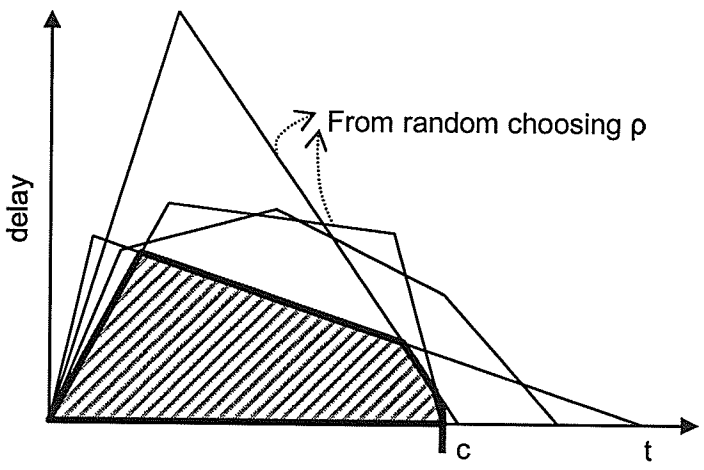


Figure 8.5. *An illustration of* composite delay envelope.

8.3 Numerical Results

To illustrate the results from our proposed optimization, we use ten sources with $B_i(\tau)$ as shown in Figure 8.6. The weights of the sources is set as ϕ = [0.13 0.18 0.21 0.16 0.82 0.24 0.26 0.34 0.42 0.41], proportional to the mean rates, and the link capacity is 3.321 *Mbps* (yielding a 95 % utilization overall). The solution from solving the optimization problem for source $j = 1$ is compared with those obtained from a random choice of ρ_i in Figure 8.7. It is clear from this figure that the optimization is doing something useful.We also observe the convergence process to the solution of the optimization algorithm in Figure 8.8 where we show the value of the objective function, the maximum delay bound for source 1 versus the value of ρ_1. The optimal solution, in this view, appears to depend very strongly on the value of ρ_1 and favors a particular value rather strongly. The solution in this case is observed to be 0.0428 (this is the normalized rate which is obtained from dividing the actual rate by the link speed). Interestingly, this value for the optimal token generation rate is close to the mean rate of the traffic sources. This is a somewhat unexpected result and indicates that one of the simplest allocation of the values of ρ_i may be the best.

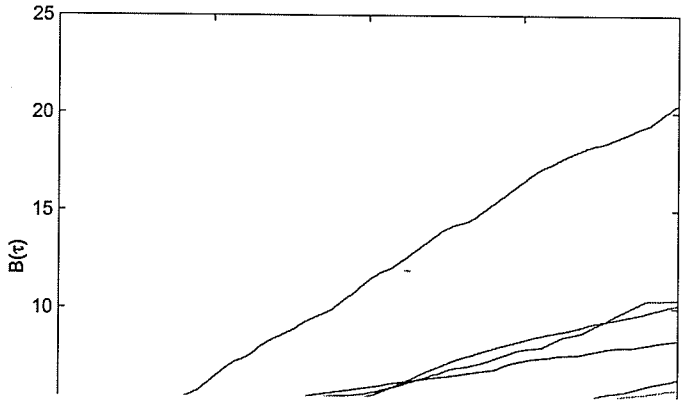
Figure 8.9 and Figure 8.10 show the delay envelopes obtained from three different

Table 8.1. *The delay from $\boldsymbol{\rho}$ obtained from optimization formulation and the mean data rate*

Source	Delay (optimization)	Delay (mean rate)
1	29.6451	30.8250
2	47.4002	47.8502
3	29.1386	30.4085
4	85.5327	85.9427
5	14.6674	14.9174
6	17.8599	18.2399
7	32.9989	33.8489
8	53.8916	54.5816
9	27.7079	28.1479
10	22.5568	23.0224

methods for source 1 and source 2 respectively. The first two use the token bucket parameters obtained from the local search method, but the method 1 curve uses the linear bound on $B_i(*)$ whereas the method 2 curve uses the non-linear $B_i(*)$ directly. The method 3 curve results from the randomized composite delay envelope method. For the composite delay envelope method, we used 500 random instances (i.e., $K = 500$) of $\boldsymbol{\rho}$. We observe that the maximum delay from the composite delay envelope method is very close to the maximum delay obtained from the other methods. However, the method 2 envelope is much closer to the composite delay envelope compared to the method 1 curve. If we are interested in more than the maximum delay alone, and wish to consider moments of the waiting time or some other measures, this may be an advantage.

As shown in Table. 8.1, the solution of our optimization formulation is observed to be close to the mean data rate of the traffic sources. Therefore, choosing the mean rate for each source for the corresponding token bucket filling rate appears to be a good choice although not necessarily optimal.


25
20
15
10
B(τ)

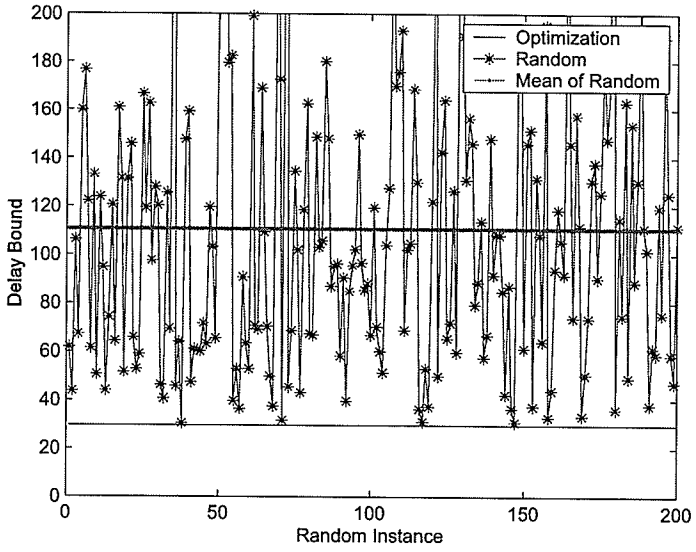


Figure 8.7. *Delay from the solution of the optimization model compared with that with random selection of* ρ_i.

local search method and the composite envelope method is close to the mean rate of the traffic source. Our conclusion is that using the mean rate for token generation rate and the corresponding bucket size can give good delay bounds for a particular traffic source. For the future work, estimation technique will be applied to obtain traffic parameters, so that optimization can be performed in an online fashion.

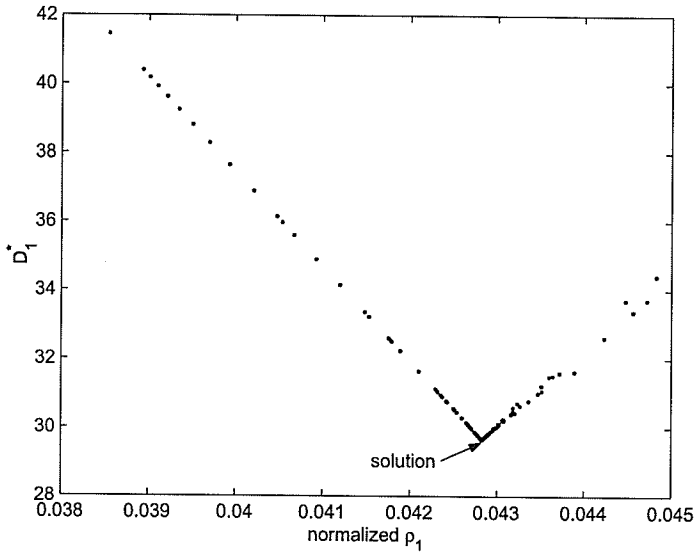


Figure 8.8. *The convergence of solution* ρ_1 *from* Nelder-Mead simplex method.

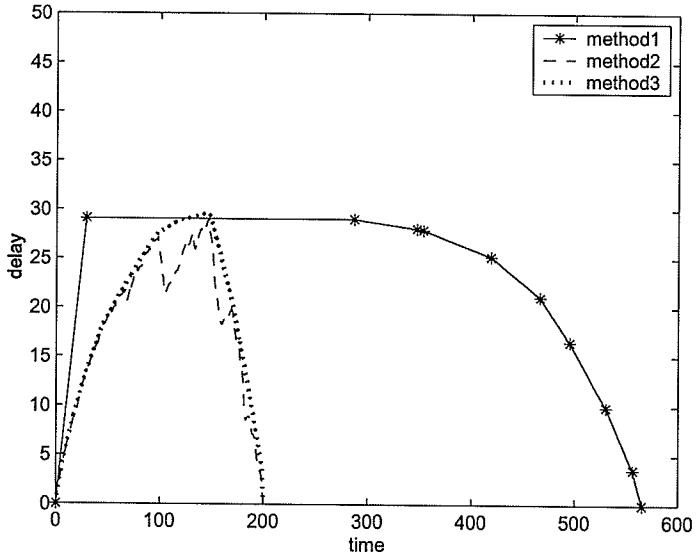


Figure 8.9. *Composite delay envelope for source* $j = 1$.

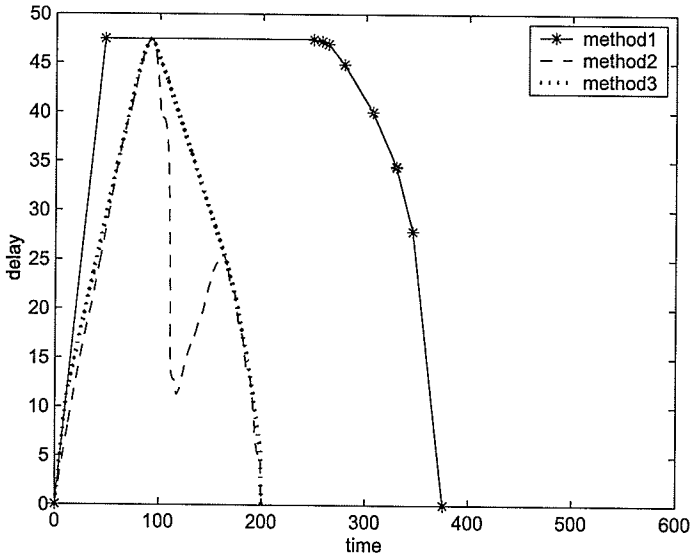


Figure 8.10. *Composite delay envelope method for source* $j = 2$.

Chapter 9

Conclusion

This chapter provides a summary of the works presented in this thesis and outlines a few directons for future research.

9.1 Summary

The following provides a summary of the works:

- *Issues and approach in 4G wireless networks:* We have presented a survey on the issues and approaches to design call admission control in the next-generation (e.g., 4G) wireless technologies. These issues include heterogeneous wireless networks, multiple types of services, adaptive bandwidth allocation and cross-layer design for both call and packet level performances. We have also introduced a two-tiered CAC architecture for 4G networks to ensure QoS in both the wireless and the wired parts. In the general architecture, the CAC decision should be based on both the call-level and the packet-level performance metrics. Our two-tier CAC scheme considered two types of services (voice and data), and data calls due to vertical handoff from other types of networks have been also taken into account.
- *Service differentiation in wireless networks:* We have presented a model for service differentiation between the QoS-sensitive and the best-effort traffic in wireless networks. Fair scheduling is used to prioritize different traffic types and a threshold-based CAC is used to limit the number of connections for the QoS-sensitive service. From the analytical model, various QoS measures (at both the connection-level and the packet-level) can be obtained. The numerical

results have shown that CAC combined with fair scheduling can provide the target level of QoS to both the QoS-sensitive and the best-effort traffic. Using the model, the CAC threshold and the parameters for fair scheduling can be optimally chosen so that the system utility is maximized.

- *Multi-service cellular mobile networks with MMPP call arrival patterns:* We have presented a Markov model to analyze the new call and the handoff call blocking probabilities under MMPP call arrival patterns in a multi-service cellular mobile network. We have observed that when the call arrival pattern has burstiness, the analytical results for the MMPP-based call arrival model can provide more accurate results than those for the traditional Poisson-based model. This observation suggests that the traditional performance models of cellular wireless networks are not suitable for the environment with bursty call arrival patterns.
- *Adaptive bandwidth allocation in cellular mobile networks under Markov call arrival process and phase-type channel holding time distribution:* We have presented an analytical framework for adaptive bandwidth allocation in cellular mobile networks. Using the framework, various performance metrics (i.e., new call blocking probability, handoff call dropping probability, average bandwidth of the cell, user outage probability, and call degradation probability) have been derived. The numerical results obtained from the model have shown that the ABA can minimize handoff call dropping probability, while some calls might experience service degradation below an acceptable level.
- *Performance analysis in cellular mobile networks with time-varying traffic:* We have presented a framework for analyzing the call-level transient performances of cellular mobile networks under time-varying traffic pattern. Based on the analytical model, we have also developed a threshold-based on-line adaptive CAC scheme. Numerical results from both the steady state and the transient analyzes have been compared. The results have shown that the system spends some period of time in transient state before reaching the steady state, and therefore, transient analysis is needed to obtain the accurate QoS performances during certain period of time. In addition, our proposed adaptive CAC can successfully control the QoS performances at the desired level under time-varying

traffic.

- *Analytical framework for integrated cross-layer study in mobile wireless multimedia networks:* We have presented a novel analytical framework for cross-layer performance evaluation in a multiuser cellular wireless network under Markov fading channel model. We have analyzed the QoS performances for real-time, non-real-time, and best-effort traffic. For real-time and non-real-time traffic, the framework also considers an adaptive traffic shaping mechanism which takes the number of allocated channels and channel quality into account to control the packet arrival rate. The results have shown how the physical layer parameters (e.g., channel quality) and the call-level parameters (e.g., call arrival rate and channel holding time) impact the packet-level QoS. Effects of user mobility on packet-level performance has been demonstrated through a basic mobility model. We have also demonstrated applications of the proposed framework in obtaining the optimal system parameter setting and channel allocation at the base station for different multimedia services under QoS constraints.
- *Optimizing token bucket parameters at DiffServ edge router under generalized processor sharing (GPS) Scheduling:* We have presented an optimization model to obtain the parameters of a token-bucket traffic shaper for a source with a non-linear traffic bound. If the trace of the traffic (e.g., the non-linear traffic bound) is provided, a relationship between the bucket size and the token generation rate can be obtained. Based on this result, and given the weights for the different traffic connections, a local search technique is used to obtain parameters which will approximately minimize the maximum average delay for traffic sources. We have observed that the optimal token generation rate obtained from both the local search technique and the composite envelope method is close to the mean rate of the traffic source. We have concluded that using the mean rate for token generation rate and the corresponding bucket size can give good delay bounds for a particular traffic source.

9.2 Future Work

We outline a few directions for further research in the area of call admission control, bandwidth adaptation and scheduling as follows:

- *Distributed admission control*: With the proliferation of data services in the cellular networks, it is expected that there will be a large amount of traffic due to short-lived data flows (e.g., for applications such as retrieval of weather information, neighborhood information). For this type of flows, centralized admission control may not be very efficient in terms of control overhead and the call processing time, and therefore, distributed admission control would be preferable. Design and analysis of distributed CAC for cellular system would be an interesting topic to research on.
- *Adaptive resource reservation and bandwidth adaptation for call admission control considering user mobility*: For a threshod-based CAC, it is necessary to selection the threshold value in an intelligent way. For example, user mobility information can be exploited to set this threshold value and reserve the resources accordingly for handoff calls/flows. Again, both the call-level and the packet-level performances need to be considered.
- *Admission control policy based on delay statistics*: Since one of the most important QoS metrics for real-time traffic is the delay, admission control for this type of traffic should consider the impact of admitting an incoming call/connection on the delay statistics of the ongoing calls. This delay statistics would depend on the scheduling policy, bandwidth adaptation mechanism and the link level error control policy. Queueing performance models can be used to calculate delay statistics so that an admission control method can use this information to decide whether an incoming connection can be accepted or not.

Bibliography

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF, RFC 2475, 1998.

[2] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 1, pp. 344-357, June 1993.

[3] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, VT-35(3):77-92, Aug. 1986.

[4] L. Badia, M. Zorzi, and A. Gazzini, "A model for threshold comparison call admission control in third generation cellular systems," in *Proceedings of IEEE International Conference on Communications 2003 (ICC'03)*, vol. 3, pp. 1664-1668, May 2003.

[5] Tao Zhang et al., "Local Predictive Resource Reservation for Handoff in Multimedia Wireless IP Networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp.1931-1941, Oct. 2001.

[6] R. Jain and E.W. Knightly, "A framework for design and evaluation of admission control algorithms in multi-service mobile networks," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 1999 (INFOCOM'99)*, pp. 1027-1035, March 1999.

[7] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 1996 (INFOCOM'96)*, vol. 1, pp. 43-50, Mar. 1996.

[8] B. Epstein and M. Schwartz, "Reservation strategies for multi-media traffic in a wireless environment," in *Proceedings of IEEE Vehicular Technology Conference 1995 (VTC'95)*, vol. 1, pp. 16-169, July 1995.

[9] B. Epstein and M. Schwartz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 523-534, Mar. 2000.

[10] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A resource estimation and call

admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 1-12, Feb. 1997.

[11] J. Hou, J. Yang, and S. Papavassiliou, "Integration of pricing with call admission control to meet QoS requirements in cellular networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 9, pp. 898-910, Sept. 2002

[12] J. Zhang, J. Huai, R. Xiao, and B. Li, "Resource Management in the next-generation DS-CDMA cellular networks," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 52-58, Aug. 2004.

[13] P. Liu, P. Zhang, S. Jordan, and M.L. Honig, "Single-cell forward link power allocation using pricing in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 533-543, March 2004.

[14] D. Pong and T. Moors, "Call admission control for IEEE 802.11E contention access mechanism," in *Proceedings of IEEE Global Telecommunications Conference 2003 (GLOBECOM'03)*, vol. 1, pp. 174-178, Dec. 2003.

[15] J. Misic, K. L. Chan, and V. B. Misic, "Admission control in Bluetooth piconets," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 890-911, May 2004.

[16] A. Safwat, "A-cell: a novel multi-hop architecture for 4G and 4G+ wireless networks," in *Proceedings of IEEE Vehicular Technology Conference 2003 (VTC'03)*, vol. 5, pp. 2931-2935, Oct. 2003.

[17] C. T. Chou and K.G. Shin, "Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service," *IEEE Transactions on Mobile Computing*, vol. 3, no. 1, pp. 5-17, Jan.-Mar. 2004.

[18] G. Carneiro, J. Ruela, and M. Ricordo, "Cross-layer design in 4G wireless terminals," *IEEE Wireless Communications*, vol. 11, no. 2, pp. 7-13, Apr. 2004.

[19] S. Lu, V. Bharghavan, and R. Srikant, "Fair Scheduling in Wireless Packet Networks," *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 473-489, Aug. 1999.

[20] T. S. E. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 1998 (INFOCOM'98)*, pp. 1103-111, March 1998.

[21] R. Fantacci and L. Zoppi, "Performance evaluation of polling systems for wireless local communication networks," *IEEE Transactions on Vehicular Technology*, vol. 49, pp. 2148-2157, Nov. 2000.

[22] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 371-382, March 2002.

[23] M. Cheung and J. W. Mark, "Resource allocation in wireless networks based on joint packet/call levels QoS constraints," in *Proceedings of IEEE Global Telecommunications Conference 2000 (GLOBECOM'00)*, vol. 1, pp. 271-275, Nov. 2000.

[24] K. K. Leung and A. Srivastava, "Dynamic allocation of downlink and uplink resource for broadband services in fixed wireless networks," in *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 5, pp. 990 - 1006, May 1999.

[25] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1746 - 1755, Sept. 2004.

[26] J. Wang and B. Ravindran, "Time-utility function-driven switched Ethernet: Packet scheduling algorithm, implementation, and feasibility analysis," *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 2, pp. 119 - 133, Feb. 2004.

[27] E. D. Jensen, "Asynchronous decentralized real-time computer systems," *Real-Time Computing*, W.A. Halang and A.D. Stoyenko, eds., NATO Advanced Study Inst., Oct. 1992.

[28] M. F. Neuts, "Matrix Geometric Solutions in Stochastic Models - An Algorithmic Approach," *John Hopkins Univ. Press, Baltimore, MD*, 1981.

[29] M. Zorzi, "Packet dropping statistics of a data-link protocol for wireless local communications," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 1, pp. 71-79, Jan. 2003.

[30] R. Hooke and T.A. Jeeves, "Direct search solution of numerical and statistical problems," *J. Ass. Comput. Mach.*, pp. 212-29, 1961.

[31] M. Achir, Y. M. Ghamri-Doudane, and G. Pujolle, "Predictive resource allocation in cellular networks using Kalman filters," in *Proceedings of IEEE International Conference on Communications 2003 (ICC'03)*, vol. 2, pp. 974-981.

[32] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *IEEE Wireless Communications*, vol. 11, no. 3, pp. 8-15, June 2004.

[33] P. Fazekas, S. Imre, and M. Telek, "Modeling and analysis of broadband cellular networks with multimedia connections," *Telecommunication Systems*, vol. 19(3-4), pp. 263-288, 2002.

[34] K. Mitchell, K. Sohraby, A. van de Liefvoort, and J. Place, "Approximation models of wireless cellular networks using moment matching," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 11, Nov. 2001, pp. 2177 - 2190.

[35] D. Z. Deniz and N. O. Mohamed, "Performance analysis of the threshold CAC strategy for multimedia traffic in wireless networks," in *Proceedings of IEEE International Conference on Communications 2003 (ICC'03)*, pp. 1312 - 1317.

[36] J. Misic and Y. B. Tam, "Non-uniform traffic issues in DCA wireless multimedia networks," *Wireless Networks*, 2003, pp. 605-622.

[37] B. Li, S. T. Chanson, and C. Lin, "Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks," *Wireless Networks*, vol. 4, no. 4, July 1998, pp. 279-290.

[38] Y. Shu, M. Yu, J. Liu, and O. W. W. Yang, "Wireless traffic modeling and prediction using seasonal ARIMA models," in *Proceedings of IEEE International Conference on Communications 2003 (ICC'03)*, vol. 3, pp. 1675-1679.

[39] D. P. Gaver, P.A. Jacobs, and G. Latouche, "Finite birth-and-death models in randomly changing environments," *Adv. Applied Prob.*, vol. 16, 1984, pp. 715-731.

[40] P. Salvador, R. Valadas, and A. Pacheco. "Multiscale fitting procedure using Markov modulated Poisson processes," *Telecommunication Systems*, vol. 23, pp. 123-148, 2003.

[41] S. Boumerdassi and A.-L. Beylot, "Adaptive channel allocation for wireless PCN," *Mobile Networks and Applications*, vol. 4, pp. 111 - 116, 1999.

[42] S. Singh, "Quality of service guarantees in mobile computing," *Computer Communications*, no. 19, pp. 359-371, 1996.

[43] P. Orlik and S.S. Rappaport, "A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, June 1998, pp.788-803.

[44] Y. Fang, "Hyper-Erlang distribution model and its application in wireless mobile networks," *Wireless Networks*, vol. 7, pp. 211-219, 2001.

[45] M.F. Neuts, "Structured Stochastic Matrices of M/G/1-type and their Applications," Marcel Dekker, New York, NY, 1989.

[46] Y. Fang, "Thinning schemes for call admission control in wireless networks," *IEEE Transactions of Computer*, vol. 52, no. 5, pp. 685-687, May 2003.

[47] A. Jayasuriya, D. Green, and J. Asenstorfer, "Modelling service time distribution in cellular networks using phase-type service distribution," in *Proceedings of IEEE International Conference on Communications 2001 (ICC'01)*, vol. 2, pp. 440-444, June 2001.

[48] A.S. Alfa and Wei Li, "A homogeneous PCS network with Markov call arrival process and phase-type cell residence time," *Wireless Networks*, vol 8, pp. 597-605, 2002.

[49] T. K. Christensen, B.F. Nielsen, and V.B. Iversen, "Phase-type models of channel-holding times in cellular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 725-733, May 2004.

[50] A. Horvath and M. Telek, "PhFit: A general phase-type fitting tool," *Computer Performance Evaluation Tools*, pp. 82-91, 2002.

[51] J. E. Diamond and A. S. Alfa, "Approximation method for M/PH/1 retrial queues with phase type inter-retrial time," *European Journal of Operational Research*, vol. 113, pp. 620-631, 1999.

[52] L. Breuer, "Parameter estimation for a class of BMAPs," in *Advances in Algorithmic Methods for Stochastic Models*, 87-97, Notable Publications, 2000.

[53] Y. Cao and V.O.K. Li, "Utility-oriented adaptive QoS and bandwidth allocation in wireless networks," in *Proceedings of IEEE International Conference on Communications 2002 (ICC'02)*, vol. 5, pp. 3071-3075, May 2002.

[54] D. W. Dormuth and A. S. Alfa, "Two finite-difference methods for solving MAP(t)/PH(t)/1/K queueing models," *Kluwer Queueing System*, vol. 27, pp. 55-78, 1997.

[55] J. A. Carrasco, "Solving dependability/performability irreducible Markov models using regenerative randomization," *IEEE Transactions on Reliability*, vol. 52, no. 3, pp. 319-329, Sept. 2003.

[56] J. A. Carrasco, "Transient analysis of some Rewarded Markov models using Randomization with Quasistationarity Detection," *IEEE Transactions on Computers*, vol. 53, no. 9, pp. 1106-1120, Sept. 2004.

[57] J. Zhang and E. J. Coyle, "The transient performance analysis of voice/data integrated networks," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 1990 (INFOCOM'90)*, vol. 3, pp. 963-968, June 1990.

[58] S. Kim, T. Kwon, and Y. Choi, "Call admission control for prioritized adaptive multimedia services in wireless/mobile networks," in *Proceedings of IEEE Vehicular Technology Conference 2000 (VTC'00)*, pp.1536-1540, May 2000.

[59] N. Nasser and H. Hassanein, "Connection-level performance analysis for adaptive bandwidth allocation in multimedia wireless cellular networks," in *Proceedings of IEEE Phoenix Conference on Computers and Communications (IPCCC'04)*, pp. 61-68, Apr. 2004.

[60] T. Kwon, Y. Choi, C. Bisdikian, and M. Naghshineh, "Call admission control for adaptive multimedia in wireless/mobile networks," in *Proceedings of ACM Workshop on Wireless Mobile Multimedia (WOWMOM'98)*, Oct. 1998.

[61] S. S. Rappaport, "Blocking, hand-off and traffic performance for cellular communications with mixed platforms," in *Proceedings of Instrument Electronics Engineering*, vol.140, no.5, pp.389-401, 1993.

[62] J. A. Neldel and R. Mead, "A simplex method for function minimisation," *The Computer Journal 7*, pp. 308-313, 1965.

[63] A. P. A. van Moorsel and B.R. Haverkort, "Adaptive uniformization," *Communications in Statistics - Stochastic Models*, vol. 10, no. 3, pp. 619-647, 1994.

[64] S. Kim, T. Kwon, and Y. Choi, "Call admission control for prioritized adaptive multimedia services in wireless/mobile networks," in *Proceedings of IEEE Vehicular Technology Conference 2000 (VTC'00)*, pp.1536-1540, May 2000.

[65] J. Ye, J. Hou, and S. Papavassiliou, "A comprehensive resource management framework for next generation wireless networks," *IEEE Transactions on Mobile Computing*, vol. 1, no. 4, Oct.-Dec. 2002, pp. 249-264.

[66] J. G. Kim and M. M. Krunz, "Delay analysis of selective repeat ARQ for a Markovian source over a wireless channel," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 5, pp. 1968-1981, Sept. 2000.

[67] L. B. Le, E. Hossain and A. S. Alfa, "Queuing analysis for radio link level scheduling in a multi-rate TDMA wireless network." in *Proceedings of IEEE Global Telecommunications Conference 2004 (GLOBECOM'04)*, vol. 6, pp. 4061-4065, Dec. 2004.

[68] T. V. J. Ganesh Babu, T. Le-Ngoc and J. F. Hayes, "Performance of a priority-based dynamic capacity allocation scheme for wireless ATM systems," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 355-369, Feb 2001.

[69] L. Galluccio, F. Licandro, G. Morabito, and G. Schembra, "An analytical framework for the design of intelligent algorithms for adaptive-rate MPEG video encoding in next-generation time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 369-384, Feb. 2005.

[70] G. Schembra, "A resource management strategy for multimedia adaptive-rate traffic in a wireless network with TDMA access," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, pp. 65-78, Jan. 2005.

[71] R. T. Sheu, and J. L.C. Wu, "Performance analysis of rate control with scaling QoS parameters for multimedia transmissions," *IEE Proceedings Communications*, vol. 150, no. 5, pp. 361-366, Oct. 2003.

[72] F. Y. Li and N. Stol, "QoS provisioning using traffic shaping and policing in 3rd-generation wireless networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC'02)*, vol. 1, pp. 139-143, March 2002.

[73] O. B. Akan and I. F. Akyildiz, "ARC: the analytical rate control scheme for real-time traffic in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 4, pp. 634-644, Aug. 2004.

[74] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Transactions on Communications*, vol. 47, no. 6, pp. 884-895, June 1999.

[75] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, Siam Philadelphia, 2000.

[76] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Personal Communications*, vol. 3, no. 6, pp. 4-9, Dec. 1996.

[77] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scienti c Computing.* Cambridge University Press, 1988.

[78] D. Zhao, X. Shen, and J. W. Mark, "Radio resource management for cellular CDMA systems supporting heterogeneous services," *IEEE Transactions on Mobile Computing*, vol. 2, no. 2, pp. 147-160, Apr. 2003.

[79] L. Xu, X. Shen, and J.W. Mark, "Dynamic fair scheduling with QoS constraints in multimedia wideband CDMA cellular networks," *IEEE Transactions on Wireless Communications*, vol. 3, no. 1, pp. 60-73, Jan. 2004.

[80] T. C. Wong, J. W. Mark, and K. C. Chua, "Joint connection level, packet level, and link layer resource allocation for variable bit rate multiclass services in cellular DS-CDMA networks with QoS constraints," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1536-1545, Dec. 2003.

[81] S. M. Ross, "Stochastic Processes,", *John Wiley and Sons*, 1983.

[82] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks", in *Proceedings of ACM Special Interest Group on Data Communications (SIGCOMM'97)*, pp. 63-74, Sept. 1997.

[83] C.R. Jon, Bennett, and Hui Zhang, "WF^2Q: Worst-case fair weighted fair queuing," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 1996 (INFOCOM'96)*, 1996.

[84] D. Hang, H.R. Shao, W. Zhu, and Y.Q. Zhang, "TD^2FQ: an integrated traffic scheduling and shaping scheme for DiffServ networks," in *Proceedings of IEEE Workshop on High Performance Switching and Routing (HPSR'01)*, pp. 78-82, May 2001.

[85] D. Nandita, J. Kuri, and H.S. Jamadagni, "Optimal call admission control in generalized processor sharing (GPS) schedulers," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 2001 (INFOCOM'01)*, vol. 1, pp. 468-477, Apr. 2001.

[86] R. Szabo, P. Barta, F. Nemeth, J. Biro, and C.-G. Perntz, "Call admission control in generalized processor sharing (GPS) schedulers using non-rate proportional weighting of sessions," in *Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00)*, vol. 3, pp. 1243-1252, Mar. 2000.

[87] F.N. Nemeth, P. Barta, and J. Biro, "Network level call admission control algorithms for generalized processor sharing scheduling discipline," in *Proceedings of IEEE International Symposium on Computers and Communication (ISCC'03)*, vol. 2, pp. 1299-1305, July 2003.

[88] R. Szabo, P. Barta, F. Nemeth, and J. Biro, "Non-rate proportional weighting of generalized processor sharing schedulers," in *Proceedings of IEEE Global Telecommunications Conference 1999 (GLOBECOM'99)*, vol. 2, pp. 1334-1339, Dec. 1999.

VITA

Surname:	Niyato	***Given Names:***	Dusit
Place of Birth:	Thailand	***Date of Birth:***	Oct. 24, 1978

Educational Institutions Attended

King Mongkut's Institute of Technology Ladkrabang (KMITL)	1994 to 1999
King Mongkut's Institute of Technology Ladkrabang (KMITL)	1999 to 2001

Degrees Awarded

B.E.	KMITL	1999
M.E.	KMITL	2001

Honors and Awards

Thai Graduated Institute Science and Technology (TGIST), Scholarship	1999-2001
Telecommunication Research Labs (TR*Labs*) Research Scholarship	2003-2005

Journal Publications

1. **D. Niyato** and E. Hossain, "A Novel Analytical Framework for Integrated Cross-Layer Study of Call-Level and Packet-Level QoS in Wireless Mobile Multimedia Networks," submitted to *IEEE Transactions on Mobile Computing.*
2. **D. Niyato** and E. Hossain, "Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks," submitted to *IEEE Transactions on Mobile Computing.*
3. **D. Niyato** and E. Hossain, "Call-Level and Packet-Level Quality of Service and User Utility in Rate-Adaptive Cellular CDMA Networks: A Queueing Analysis," submitted to *IEEE Transactions on Mobile Computing.*
4. **D. Niyato** and E. Hossain, "Service Differentiation in Broadband Wireless Access Networks: A Unified Analysis," submitted to *IEEE Transactions on Wireless Communications.*
5. **D. Niyato** and E. Hossain, "Multi-Service Cellular Mobile Networks with MMPP Call Arrival Patterns: Modeling and Analysis," submitted to *Elsevier Computer Communications.*

6. **D. Niyato** and E. Hossain, "Call Admission Control for QoS Provisioning in 4G Wireless Networks: Issues and Approaches," accepted for publication in the Special Issue of *IEEE Network* on "4G Network Technologies for Mobile Telecommunications".
7. **D. Niyato** and E. Hossain, "Adaptive fair subcarrier/rate allocation in multi-rate OFDMA networks: Radio link level queuing performance analysis," submitted to *IEEE Transactions on Vehicular Technology.*

Conference Publications

1. **D. Niyato** and E. Hossain, "A Game Theoretic Approach to Bandwidth Allocation and Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks," submitted to *IEEE International Conference on Communications 2006 (ICC'06)*, Istanbul, Turkey.
2. **D. Niyato** and E. Hossain, "Delay-Based Admission Control Using Fuzzy Logic for OFDMA Broadband Wireless Networks," submitted to *IEEE International Conference on Communications 2006 (ICC'06)*, Istanbul, Turkey.
3. **D. Niyato** and E. Hossain, "Joint Bandwidth Allocation and Connection Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks," submitted to *IEEE International Conference on Communications 2006 (ICC'06)*, Istanbul, Turkey.
4. **D. Niyato** and E. Hossain, "Call-Level and Packet-Level Performance Analysis of Call Admission Control and Adaptive Channel Allocation in Cellular Wireless Networks," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.
5. **D. Niyato** and E. Hossain, "Queue-Aware Uplink Bandwidth Allocation for Polling Services in 802.16 Broadband Wireless Networks," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.
6. **D. Niyato** and E. Hossain, "Call-Level and Packet-Level Performance Modeling in Cellular CDMA Networks," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.
7. **D. Niyato** and E. Hossain, "Connection Admission Control Algorithms for OFDMA Wireless Networks," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.
8. **D. Niyato** and E. Hossain, "Queueing Analysis of OFDM/TDMA Systems," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.

9. T. Issariyakul, **D. Niyato**, E. Hossain, and A. S. Alfa, "Exact Distribution of Access Delay in IEEE 802.11 DCF MAC," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.

10. **D. Niyato**, J. Diamond and E. Hossain, "On Optimizing Token Bucket Parameters at the Network Edge Under Generalized Processor Sharing (GPS) Scheduling," to be presented in *IEEE Global Telecommunications Conference 2005 (GLOBECOM'05)*, St. Louis, MO.

11. **D. Niyato**, E. Hossain, and A. S. Alfa, "Performance analysis and adaptive call admission control in cellular mobile networks with time-varying traffic," in Proc. of *IEEE International Conference on Communications 2005 (ICC'05)*, Seoul, Korea, May 2005.

12. **D. Niyato** and E. Hossain, "Analysis of fair scheduling and connection admission control in differentiated services wireless networks," in Proc. of *IEEE International Conference on Communications 2005 (ICC'05)*, Seoul, Korea, May 2005.

13. **D. Niyato**, E. Hossain, and A. S. Alfa, "Performance analysis of multi-service cellular wireless networks for time-dependent call arrival patterns," in Proc. of *IEEE Global Telecommunications Conference 2004 (GLOBECOM'04)*, Dallas, TX, USA, Nov.-Dec. 2004.

14. **D. Niyato**, C. Srinilta, "Load balancing algorithms for Internet video and audio server," in Proc. of *IEEE International Conference on Networks 2001 (ICON'01)*, Bangkok, Thailand, Oct. 2001.

VITA

Surname: Niyato *Given Names:* Dusit
Place of Birth: Thailand *Date of Birth:*

Educational Institutions Attended

King Mongkut's Institute of Technology Ladkrabang (KMITL)	1994 to 1999
King Mongkut's Institute of Technology Ladkrabang (KMITL)	1999 to 2001

Degrees Awarded

B.E.	KMITL	1999
M.E.	KMITL	2001

Honors and Awards

Thai Graduated Institute Science and Technology (TGIST), Scholarship	1999-2001
Telecommunication Research Labs (TR*Labs*) Research Scholarship	2003-2005

Journal Publications

1. **D. Niyato** and E. Hossain, "A Novel Analytical Framework for Integrated Cross-Layer Study of Call-Level and Packet-Level QoS in Wireless Mobile Multimedia Networks," submitted to *IEEE Transactions on Mobile Computing.*
2. **D. Niyato** and E. Hossain, "Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks," submitted to *IEEE Transactions on Mobile Computing.*
3. **D. Niyato** and E. Hossain, "Call-Level and Packet-Level Quality of Service and User Utility in Rate-Adaptive Cellular CDMA Networks: A Queueing Analysis," submitted to *IEEE Transactions on Mobile Computing.*
4. **D. Niyato** and E. Hossain, "Service Differentiation in Broadband Wireless Access Networks: A Unified Analysis," submitted to *IEEE Transactions on Wireless Communications.*
5. **D. Niyato** and E. Hossain, "Multi-Service Cellular Mobile Networks with MMPP Call Arrival Patterns: Modeling and Analysis," submitted to *Elsevier Computer Communications.*