ABSTRACT

EFFICIENT RESOURCE ALLOCATION AND CALL ADMISSION CONTROL IN HIGH CAPACITY WIRELESS NETWORKS

by
Jiongkuan Hou

Resource Allocation (RA) and Call Admission Control (CAC) in wireless networks are processes that control the allocation of the limited radio resources to mobile stations (MS) in order to maximize the utilization efficiency of radio resources and guarantee the Quality of Service (QoS) requirements of mobile users. In this dissertation, several distributed, adaptive and efficient RA/CAC schemes are proposed and analyzed, in order to improve the system utilization while maintaining the required QoS.

Since the most salient feature of the mobile wireless network is that users are moving, a Mobility Based Channel Reservation (MBCR) scheme is proposed which takes the user mobility into consideration. The MBCR scheme is further developed into PMBBR scheme by using the user location information in the reservation making process. Through traffic composition analysis, the commonly used assumption is challenged in this dissertation, and a New Call Bounding (NCB) scheme, which uses the number of channels that are currently occupied by new calls as a decision variable for the CAC, is proposed.

This dissertation also investigates the pricing as another dimension for RA/CAC. It is proven that for a given wireless network there exists a new call arrival rate which can maximize the total utility of users, while maintaining the required QoS. Based on this conclusion, an integrated pricing and CAC scheme is proposed to alleviate the system congestion.

EFFICIENT RESOURCE ALLOCATION AND CALL ADMISSION CONTROL IN HIGH CAPACITY WIRELESS NETWORKS

by
Jiongkuan Hou

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering

May 2003

APPROVAL PAGE

EFFICIENT RESOURCE ALLOCATION AND CALL ADMISSION CONTROL IN HIGH CAPACITY WIRELESS NETWORKS

Jiongkuan Hou

Dr. Symeon Papavassiliou, Dissertation Advisor — Date
Assistant Professor of Electrical and Computer Engineering, NJIT

Dr. Nirwan Ansari, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Sirin Tekinay, Committee Member — Date
Assistant Professor of Electrical and Computer Engineering, NJIT

Dr. Yu-Dong Yao, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, Stevens Institute of Technology

Dr. Sotirios Ziavras, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author: Jiongkuan Hou

Degree: Doctor of Philosophy

Date: May 2003

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2003
- Master of Science in Electrical Engineering,
  Northern Jiaotong University, Beijing, P. R. China, 1999
- Bachelor of Science in Electrical Engineering,
  Northern Jiaotong University, Beijing, P. R. China, 1993

Major: Electrical Engineering

Presentations and Publications:

Jiongkuan Hou and Symeon Papavassiliou,
"A Dynamic Reservation Based Call Admission Control Algorithm for Wireless Networks Using the Concept of Influence Curve,"
Journal of Telecommunications Systems, 22:1-4, 299-319, 2003.

Jian Ye, Jiongkuan Hou and Symeon Papavassiliou,
"A Comprehensive Resource Management Framework for Next Generation Wireless Networks,"
IEEE Transactions on Mobile Computing, Vol. 1 No. 4, pp. 249-264, Oct-Dec 2002.

Jiongkuan Hou, Jie Yang and Symeon Papavassiliou,
"Integration of Pricing with Call Admission Control to Meet QoS Requirements in Cellular Networks,"
IEEE Transactions on Parallel and Distributed Systems, Vol. 13, Number. 9, pp. 898-910, September 2002.

Jiongkuan Hou and Yuguang Fang,
"Mobility-based Call Admission Control Schemes for Wireless Mobile Networks,"
Wireless Communications and Mobile Computing, Vol. 1, Issue. 3, pp. 269-282, 2001.

Symeon Papavassiliou, Jiongkuan Hou and Sebnem Ozer,
"Admission Control in Wireless Networks,"
The Wiley Encyclopedia of Telecommunications, John Wiley & Sons Inc., January 2003.

Jian Ye, Jiongkuan Hou and Symeon Papavassiliou,
"Integration of Advanced Reservation and Bandwidth Reconfiguration Based Admission Control in Wireless Networks with Multimedia Services,"
Proceedings of IEEE MWN 2003, May 2003.

Jian Ye, Jiongkuan Hou and Symeon Papavassiliou,
"Mobile Agent Based Framework for Mobility Assisted Channel Reservation in Wireless Networks,"
Proceedings of 36th Annual Conference on Information Sciences and Systems, pp. 833-838, Princeton, March 2002.

Jiongkuan Hou, Jie Yang and Symeon Papavassiliou,
"Integration of Pricing with Call Admission Control for Wireless Networks,"
Proceedings of IEEE VTC 2001/Fall, pp. 1344-1348, October, 2001.

Jiongkuan Hou and Symeon Papavassiliou,
"Influence-Based Channel Reservation Scheme for Mobile Cellular Networks,"
Proceedings of IEEE ISCC 2001, pp. 218-223, July 2001.

Jiongkuan Hou, Yuguang Fang and Ali Akansu,
"Mobility-based Channel Reservation Scheme for Wireless Mobile Networks,"
Proceedings of IEEE WCNC 2000, pp. 527-531, Chicago, September 2000.

To my wife Jimei and my parents

ACKNOWLEDGMENT

I am indebted to many individuals for their care and support given to me during my doctoral studies. First of all, I would like to express my deep gratitude to Dr.. Symeon Papavassiliou. As my research supervisor, he has provided me with constant encouragement, insightful comments and invaluable suggestions, which benefited not only the completion of this dissertation, but also my career in a long time to come. I really enjoy working with him, and I recall that I was so lucky to be part of his group since he was so concerned about his students.

I am also grateful to Dr. Yuguang Fang for leading me into this exciting research area and helping me to develop the necessary skills to be a successful researcher.

I would like to thank Dr. Nirwan Ansari, Dr. Sirin Tekinay, Dr. Yu-Dong Yao and Dr. Sotirios Ziavras for their precious time spent in reviewing this work and for their valuable suggestions during its development.

I would also like to thank the former and present members of the Broadband, Mobile and Wireless Networking Laboratory (BMWN) and the New Jersey Center for Multimedia Research (NJCMR), especially, Jie Yang, Jian Ye, Surong Zeng, Feihong Chen, Sebnem Ozer, Dequan Liu and Zhicheng Ni, for their friendship, discussion and advice. I am really pleased to be associated with these talented colleagues.

Thanks are also due to the Department of ECE and the New Jersey Center for Wireless Telecommunications for their support and effort in providing a friendly research environment.

Special thanks should go to my parents and my brother, since they always loved me, believed in me and encouraged me in my study. My final acknowledgments go to my wife, Jimei, for her dedicated sacrifice, support, understanding and encouragement. This dissertation could not be completed without her presence beside me.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Motivation

A wireless network is typically organized into geographical regions called cells. Radio resources are assigned (statically or dynamically) to each cell to serve the mobile stations (MS) inside the cell. Resource Allocation (RA) and Call Admission Control (CAC) are processes that control the allocation of these limited radio resources to MSs to maximize the utilization efficiency of radio resources and to guarantee the Quality of Service (QoS) requirements of mobile users.

Due to the user mobility, the limitation of wireless resources and the harsh radio frequency environment, RA and CAC processes are more complicated for wireless networks than those for wired networks. Since a user may change its radio cell (handoff) a number of times during the lifetime of its connection, if the availability of the radio resources at the handoff target cell cannot be guaranteed (*i.e.,* when congestion occurs), the connection will be terminated prematurely (dropped), which presents a serious QoS degradation.

With the advancement in the field of wireless communications the recent years, the problem of congestion control has gained high research and practical importance. On one hand, the population of wireless/mobile users is growing at a rapid rate and wireless systems are using micro/pico cellular architectures to provide a higher capacity. Due to the smaller coverage area of micro/pico cells and the characteristics of the multi-path and shadow fading environment, handoff events occur at a much higher rate. On the other hand, wireless networks are expected to support multiple service types, such as voice, data and video, each of which has different QoS requirements. The limited radio resources must be allocated fairly and efficiently among the different users, in order to maximize the system utilization while meeting the users' QoS requirements.

These challenging requirements motivate the design of efficient, distributed, adaptive and scalable RA and CAC schemes that can support multimedia services.

1.2 Background

In wireless networks, the radio resources are organized as logical channels in the forms of time slots, frequency bandwidth or codes. Two types of calls share these channels: the new calls and the handoff calls. New calls are those initiated by mobile users in the current cell, while the handoff calls are those initiated in other cells and handed off into the current cell. When a call arrives at a cell in which a channel is not available, the call is said to be blocked. The *new call blocking probability* (P_{nb}) and *handoff call blocking probability* (P_{hb}) are two of the most significant connection level QoS metrics in wireless mobile networks. In cellular systems, each arriving call, whether it is a new call or a handoff call, will be allocated one or several channels according to its traffic characteristics and QoS requirements, if it is accepted for service. Thus, new calls and handoff calls compete for the usage of the same resources in a cell. Since the number of channels in a cell is limited, the new call blocking probability and handoff call blocking probability cannot be decreased simultaneously.

From the users' point of view, a call being forced to terminate during the service is more annoying than a call being blocked at its start. Hence, usually the handoff call blocking probability is much more stringent than the new call blocking probability. Therefore, handoff calls are commonly given a higher priority in accessing the wireless channels.

In recent years, considerable efforts have focused on the handoff priority-based Resource Allocation and Call Admission Control problems and many schemes that range from static to dynamic strategies have been proposed in literature. They can be roughly classified into three categories:

(1) *Guard Channel Schemes*: A number of channels in each cell are reserved for the exclusive use by handoff calls; the rest of the channels are shared by both new and handoff calls. Guard channel schemes can be further divided into static schemes and dynamic schemes.

The static guard channel schemes are also called fixed reservation or cut-off priority schemes. In such schemes, the number of channels reserved for handoff purposes is fixed for each cell, and it is calculated according to the knowledge of the traffic pattern of the area and the estimation of the channel occupancy time distribution. Based on the assumptions that call arrival follows Poisson process and channel holding time is exponentially distributed, Hong and Rappaport modeled the system by birth-death processes and calculated and compared the new call blocking probability and handoff call blocking probabilities for both prioritized and non-prioritized handoff procedures [34]. Ramjee *et al.* proved that guard channel policy is optimal for the MINOBJ problem, *i.e.,* minimizing the penalties associated with blocking of new and handoff calls [70].

The fixed reservation schemes are very simple in their implementation since no communication and computation overheads are involved. However, such schemes are not flexible in the sense that they do not use the traffic information of the current cell and the neighboring cells, and therefore, they cannot adapt to the real-time network conditions. Dynamic guard channel schemes are proposed to overcome the disadvantages of the static schemes to some extent.

Oliveira *et al.* set the number of channels to be reserved as a function of the requested bandwidth of connections or as a function of the number of connections [64]. The base station keeps monitoring the handoff dropping probability and the utilization of the channels, and based on these parameters the reservation is adjusted.

Choi and Shin predicted the probability that a call will be handed off to a certain neighboring cell based on the aggregate history of handoffs observed in each cell in order to determine the number of channels needed to be reserved [15]. The base station records the number of handoff call drops and adjusts the reservation values by changing the size of the estimation window. Levine *et al.* introduced the shadow cluster concept to estimate the future resource requirements based on the current movement pattern of the mobile users [53].

(2) *Queuing Priority Schemes:* In Queuing Priority Schemes, when all channels are occupied, either new calls are queued while handoff calls are blocked [52], or new calls are blocked while handoff calls are queued [81], or both calls are queued [10]. In standard PSTN, the queueing of new calls is impractical since the signaling needed for the dialing is done on the communication channel itself. Queueing of new call would therefore result in multiple redials that would unnecessarily occupy some communication channels. In cellular systems, the setup of a call is done on a separate control channel, which can provide the system with a way of queueing new calls without affecting the transmission channels [31]. Queueing of handoff calls is possible due to the fact that there is a finite time interval between the time that the received signal level drops below the handoff threshold and the time that the call is terminated due to insufficient signal level [72, 81].

The queueing of new call is proposed for a cutoff priority system to improve the total carried traffic [31]. A two dimensional Markov chain model is constructed to analyze the system and calculate the performance metrics such as handoff call blocking probability, average delay of new calls and average number of busy channels. Tekinay and Jabbari proposed a measurement based handoff queueing scheme [81]. The base station keeps measuring the receiving signal power of each queued call and the queue is dynamically reordered as new measurement results are available: the mobile station that has the lowest signal power is put at the head of the queue. Chang *et al.* proposed to integrate the guard channel scheme and the queueing scheme to handle handoff calls [10]. If no channels are available, both the new calls and handoff calls are put into their respective queues. To be more practical, the authors also considers the situation that the new calls will renege from the queue if the waiting time is longer than the patience time and handoff call will drop if the waiting time is longer than the handoff time. Signal flow graphs and Mason's formula [51] are used to obtain the performance metrics such as blocking probabilities and average waiting time.

(3) *Channel Borrowing Schemes:* When all the channels in a cell are occupied, the cell borrows channels from other cells to accommodate the incoming handoff calls. One problem associated with the channel borrowing scheme is *Channel Locking.* That is, cells within the required minimum channel reuse distance from a cell that borrows a channel cannot use the same channel [67].

Chang *et al.* suggested a two-phase channel borrowing scheme [11]: first if channel borrowing is necessary, the channel is borrowed from neighboring cells by an impact-based borrowing strategy; then channel reallocation procedures are used to further improve the efficiency. Jiang and Rappaport proposed a Prioritized Channel Borrowing Without Locking Scheme which makes borrowed channels to be used with reduced power to guarantee no co-channel interference and hence solve the problem of channel locking [44]. Since the borrowed channel can only be used in part of the cell, channel rearrangement and cutoff priority are used to discourage excessive borrowing and promote a more uniform grade of service throughout the service area.

An alternative to channel borrowing scheme is the sub-rating scheme [56]: when there is no available channel at the arrival of a handoff call, one of the currently occupied channels is divided into two half rate channels to serve the handoff call. In other words, instead of borrowing channels from neighbor cells, a sub rate channel is borrowed from users in the current cell.

Some other papers in the literature provided additional models to analyze the wireless systems and calculate the performance metrics of RA/CAC schemes. Lin *et al.* analyzed the handoff traffic of the cellular network with Poisson new call arrival, and derived a general formula for handoff call arrival rate [57], as follows:

$$\lambda_h = \frac{(1-p_0)\eta[1-f_m^*(\mu)]}{\mu[1-(1-p_f)f_m^*(\mu)]}\lambda_0 \tag{1.1}$$

where λ_0 is the new call arrival rate, $1/\eta$ is the average cell dwell time (the time that a mobile station spends in a cell whether it is active or not, cell dwell time is a measure of user mobility), $1/\mu$ is the average call holding time (how long the call will last if it is not terminated prematurely), p_0 is the new call blocking probability, p_f is the handoff call blocking probability and $f_m^*(s)$ is the Laplace-Stieltjes transform for the cell dwell time distribution. The above equation shows that the handoff call arrival process is not an independent process. It is generated during the operation of the system and is a function of other system parameters.

1.3 Outline of the Dissertation

In this dissertation, several distributed RA/CAC schemes that are flexible, adaptive and efficient in order to solve the congestion control problem are introduced, analyzed and evaluated.

Chapter 2 proposes and evaluates the Mobility Based Channel Reservation (MBCR) scheme. The basic idea behind the MBCR scheme is that a moving user, in addition to its resource requirements in the current cell, exerts some influence on the channel allocation in neighboring cells. Such an influence is related to the moving pattern of this user (speed and direction), and it can be calculated statistically. In Chapter 2, the concept of influence curve is introduced, which provides an estimate of the resource requirements that the ongoing calls in current cell impose on a neighboring cell. Based on this concept, a channel reservation scheme is proposed, which is capable of dynamically and adaptively adjusting the number of channels that should be reserved for handoff purposes in each cell. The proposed algorithm can be carried out in a distributed way: each cell collects its current traffic condition, calculates the influence and sends the results to all its neighbors periodically.

The MBCR scheme is further developed into Predictive Mobility-Based Bandwidth Reservation (PMBBR) scheme. Based on the history location information, the future moving speed and direction of each user can be predicted, which can be used to further refine the reservation making process and reduce the likelihood of false reservations. Furthermore, since the next generation wireless networks are going to be based on the packet-switching technology, the radio resources that are allocated to the users in PMBBR are in the form of bandwidth instead of logic channels. By integration PMBBR with flexible QoS management and call admission control processes, a comprehensive resource management framework for next generation wireless networks is obtained.

Chapter 2 also investigates the User Channel Holding Time, which is an important system parameter that influences heavily the accuracy and performance of the analytical models. Through traffic composition analysis, in this dissertation, the commonly used assumption in the literature is challenged, and it is demonstrated that the average channel holding time for handoff calls is always less than that for new calls for a wireless network having multiple platforms. Based on this conclusion, Chapter 3 presents the New Call Bounding (NCB) scheme. If too many new calls are accepted in a cell (for instance the new calls arrive in bursts), there will be fewer channels available in a relatively longer time, and the cell will be then congested. NCB scheme places a direct limitation on the number of new calls admitted to a cell in order to prevent potential congestion.

In Chapter 4, the role of pricing is investigated as an additional dimension of the call admission control process in order to provide users with some incentives to use wireless resources efficiently. Traditional CAC schemes that mainly focus on the tradeoff between new call blocking probability and handoff call blocking probability cannot solve the problem of congestion in wireless networks. It is first proven that for a given wireless network there exists a new call arrival rate which can maximize the total utility of users, while maintaining the required QoS. Based on this argument, the integration of pricing and call admission control is proposed, where the price is adjusted dynamically based on the current network conditions. Through extensive simulation studies, it is indicated that the proposed integrated approach achieves to efficiently alleviate the network congestion by re-shaping the incoming traffic load.

Finally, Chapter 5 concludes this dissertation and summarizes its main contributions.

CHAPTER 2

MOBILITY BASED CHANNEL RESERVATION AND CALL ADMISSION CONTROL

From the discussion presented in Chapter 1, it can be observed that most of the guard channel schemes do not explicitly take the mobility of users into consideration. Cheung and Mark has demonstrated that user mobility has a profound effect on QoS provisioning [13]. The most salient feature of the mobile wireless network is mobility. Hence, in order to make a reservation scheme effectively adapt to the changing network traffic conditions, the user mobility information must be incorporated in the channel reservation process. Levine *et al.* introduced the shadow cluster concept to estimate the future resource requirements based on the current movement pattern of mobile users [53]. Donis *et al.* proposed Virtual Cell Area Determination schemes to improve the resource allocation based on the subscriber mobility patterns [19]. However, the strength of these schemes depend on the accuracy of the knowledge of users' movement patterns, such as the trajectory of a mobile user, which is difficult to predict in a real system. Moreover, since they are centralized schemes, the signaling during the call as well as the equipment requirements are greatly increased.

One critical issue associated with all the reservation based CAC schemes is how the reservation is determined and implemented. In static guard channel schemes or the cutoff priority schemes, the number of guard channels is determined based on the prior knowledge of the cell traffic and the call blocking requirements. Obviously, the performance will degrade if the cell traffic does not conform to the prior knowledge. Thus, it is better to use dynamic channel reservation schemes which adjust the number of guard channels with the network traffic. In order to determine an optimal or near optimal reservation value, one must first answer the following question: *When to reserve channels for the incoming handoff calls?* If the reservation is made at the time when it is needed, the resulting scheme will definitely achieve better performance. Such timing is closely related with user mobility, and will be discussed in Section 2.4.1.

Handoffs occur when mobile users are moving during the call connection. Thus, a good reservation scheme should be designed based on the users' mobility pattern. Mobility patterns are determined by many factors, such as mobile users' destinations, the geographical layout of the wireless network, the traffic condition in the network, *etc.* Therefore, it is not easy to characterize in great details the mobility pattern of each specific user. However, it should be noted that the network performance is the collective outcome of all users in the network, and therefore, the statistical users' mobility patterns are more useful.

Based on these observations, the Mobility-Based Channel Reservation (MBCR) scheme and the Predictive Mobility-Based Bandwidth Reservation (PMBBR) are proposed in this chapter. The remaining of this chapter is organized as follows. Section 2.1 provides the description of the proposed mobility-based channel reservation scheme. Specifically, in Subsection 2.1.1, the concept of influence curve is introduced, and based on this in Subsections 2.1.2 through 2.1.4 the detailed description of the proposed channel reservation scheme and the associated call admission control procedures are provided. The performance evaluation of the proposed scheme is provided in Section 2.2. The performance of MBCR is compared, in terms of achieved blocking probabilities and traffic capacity under certain QoS requirements, with the corresponding performance of a cellular system with fixed and dynamic guard channel schemes. Section 2.3 provides a general analytical framework and model which takes into account the different new call and handoff call channel holding times. Section 2.4 discusses the PMBBR scheme which uses the user position information to further improve the performance. In Section 2.4, a flexible QoS management based call admission control scheme is also introduced which can be used in combination with the PMBBR to address the resource allocation problem in wireless network that supports multiple service types.

2.1 Mobility-Based Channel Reservation (MBCR)

Consider a wireless mobile network in which each cell is equipped with C channels. In order to assign higher priority to handoff calls, C_h channels out of the total C channels can be reserved for the incoming handoff calls. In this chapter, mobile users are classified into two classes according to their velocities: high speed users (vehicular users) and low speed users (pedestrians), for illustrative purposes. The average cell dwell time of a high speed user is shorter than that of a low speed user. Based on such a classification, the handoff probability of each class is predicted and reservations are made accordingly. Although the classification used here is coarse, the technique can be easily generalized to handle more general situations.

2.1.1 Influence Curves

In order to understand the reasoning behind the proposed scheme, the following observations are made:

(1) A user is more likely to request a handoff in the far future than in the near future after it enters a cell ("enter" means either the initiation of a new call or a successful handoff of an ongoing call into this cell), which implies that the handoff probability (the probability that a call needs at least one handoff during the remaining call life) is a function of the time elapsed after a call enters a cell;

(2) After dwelling in a cell for the same length of time, a high-speed user is more likely to request a handoff than a low-speed user, which implies that the handoff probability is also related to the class of a user.

Due to the existence of call handoffs, resource requirements among cells are no longer independent: when a call enters a cell, it does not only consume a channel in the current cell, but also generates certain requirements on the channels in the neighboring cells (with certain probability). In other words, an ongoing call in the current cell exerts some influence on the channel assignment in the neighboring cells. From the aforementioned observations, it can be concluded that the extent of such influence can be characterized by both the elapsed time and the class that the call belongs to. The number of channels to be reserved, C_h, has a close relationship to the extent of the influence. The more influence a call exerts on its neighboring cells, the more likely channels should be reserved in the neighboring cells to maintain the QoS requirement of this call. In order to characterize such influence, the concept of influence curve is introduced as follows.

Let $f_h(t)$ and $f_l(t)$ denote the cell-dwell-time probability density functions (*pdf*) of the high-speed users and the low-speed users, respectively. If a high-speed user enters the cell at time instant t, the probability that it will request a handoff after time instant T is:

$$\begin{aligned} & \Pr(\text{this call will request a handoff sometime after } T) \\ = & \Pr(\text{this call will stay in current cell before } T) \\ = & \int_0^{T-t} f_h(\tau)d\tau \\ = & L_h(t,T) \qquad (2.1) \end{aligned}$$

Similarly, $L_l(t,T)$ can be obtained by substituting $f_h(t)$ with $f_l(t)$.

Let $\alpha_{i,j}$ ($j \in N_i$) be the directional factor, *i.e.*, the probability that the handoff target cell is cell j when the call is being served in the cell i, where $\sum_{j \in N_i} \alpha_{i,j} = 1$, and N_i denotes the set of the neighboring cells to the cell i. For a totally random movement pattern in a homogeneous cellular network, the users move to all possible directions with equal probabilities, $\alpha_{i,j} = \frac{1}{|N_i|}$ for all the $j \in N_i$, where $|N_i|$ denotes the cardinality of the set N_i. For a cellular network with hexagonal layout, each cell has six neighbors and the directional factors for this case will be 1/6. In an environment (such as highway) that users' movements follow a highly directional pattern, some factors can be much greater than others (heterogeneous network).

With $L_l(t,T)$, $L_h(t,T)$ and $\alpha_{i,j}$, the influence curve for an ongoing high-speed call or low-speed call is defined as follows:

$$I(i,j,t,T) = \begin{cases} \alpha_{i,j} L_h(t,T) & \text{for a high speed call} \\ \alpha_{i,j} L_l(t,T) & \text{for a low speed call} \end{cases} \tag{2.2}$$

The influence curve characterizes the influence exerted on cell j at time instant T by an ongoing call which enters the cell i at time instant t.

2.1.2 Mobility-Based Channel Reservation

With the influence curve for every ongoing call, the number of channels needed to be reserved in each cell can be further determined. The total influence that all the ongoing calls in cell i exert on cell j is

$$I_{i,j} = \sum_{k \in S_i} \alpha_{i,j} L(t_k, T) \tag{2.3}$$

where S_i is the set of all the currently ongoing calls in cell i, $L(t_k,T)$ can be either $L_l(t_k,T)$ or $L_h(t_k,T)$ depending on the class of the call. The influence between neighboring cells is shown in Figure 2.1. As mentioned before, the number of channels needed to be reserved has a close relationship to the extent of the influence. In this dissertation, this number is chosen to be proportional to the extent of the influence. Thus the number of the reserved channels in cell j for calls in cell i is defined as

$$R_{i,j} = BI_{i,j} \tag{2.4}$$

where B is a tunable constant. The effect of B is discussed in Section 2.2.2. Based on the previous definitions, it can be concluded that at time T, cell j needs to reserve totally

$$R_j = \sum_{i \in N_j} R_{i,j} \tag{2.5}$$

channels for possible handoff calls from its neighboring cells. The corresponding values R_j for the various cells can be calculated in a distributed way without the need for any central coordination. The only communication requirement is that neighboring cells exchange information with each other: cell i should report $R_{i,j}$ to its neighbor cell j. Since the users are mobile, the information exchange must be done regularly (periodically) to guarantee that a cell can always have the latest information about the reservation requirements of its neighbors. It should be noted here that there is a tradeoff between the information exchange period, the accuracy of the information and the estimation of values R_j, and the involved communication overhead.

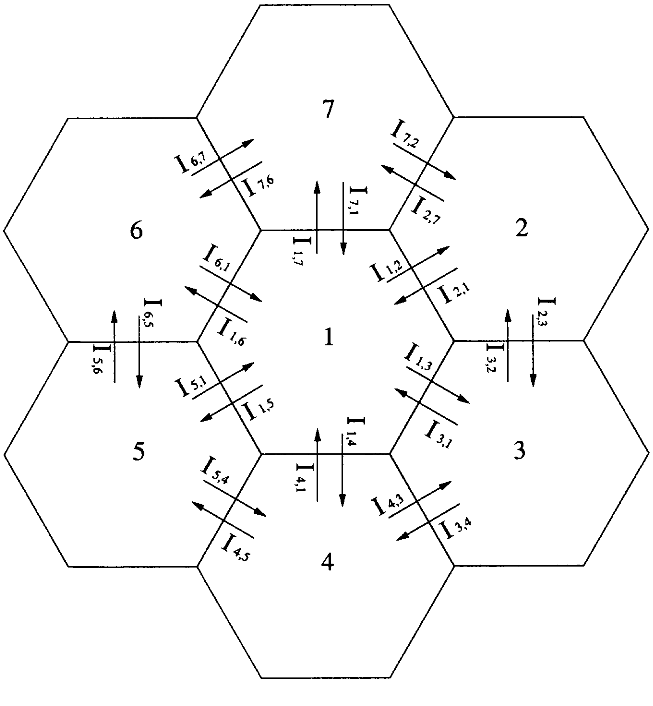


Figure 2.1 The influence between neighboring cells.

2.1.3 A Special Case

If the cell dwell times for both classes of users have exponential distributions, then $f_h(\tau) = \mu_h e^{-\mu_h \tau}$ and $f_l(\tau) = \mu_l e^{-\mu_l \tau}$, where $1/\mu_h$ and $1/\mu_l$ are the average cell dwell times for high-speed users and low-speed users, respectively. If it is further assumed that users are moving in a random movement pattern in a cellular network with hexagonal layout, following the above procedures (from Equation (2.1) to (2.5)), the influence and reservation value are obtained.

$$\begin{aligned} L_h(t,T) &= 1 - e^{-\mu_h(T-t)} \\ L_l(t,T) &= 1 - e^{-\mu_l(T-t)} \\ I(i,j,t,T) &= \begin{cases} \frac{1}{6}(1 - e^{-\mu_h(T-t)}) & \text{for high-speed user} \\ \frac{1}{6}(1 - e^{-\mu_l(T-t)}) & \text{for low-speed user} \end{cases} \end{aligned}$$

$$R_{i,j} = \frac{B}{6}\left[\sum_{k \in S_i^h}(1 - e^{-\mu_h(T-t_k)}) + \sum_{k \in S_i^l}(1 - e^{-\mu_l(T-t_k)})\right] \tag{2.6}$$

where t_k is the enter time of ongoing call k to the cell of interest, S_i^h is the set of all the ongoing high-speed calls in cell i and S_i^l includes all the ongoing low-speed calls in cell i.

2.1.4 Call Admission Control

Call Admission Control (CAC) is used to determine whether an incoming call (mostly a new call) is admitted for service or not. In order to meet the desired QoS requirements of admitted calls, some calls have to be blocked although current network resources (channels) are still available. Ramjee *et al.* proposed a general setting for CAC: guard channel schemes can be formulated by call admission probability [70], *i.e.,* if i is a decision variable (for example, the number of busy channels), a new arriving call is admitted into the network with probability $P(i)$. Depending on the choice of $P(i)$, different CAC procedures can be obtained. In this section, based on the above defined MBCR scheme, two different CAC procedures are proposed.

Integral MBCR. At time instant T, cell j calculates R_j according to Equation (2.5). Note that R_j may not be an integer. In this scheme, the final target number of reserved channels is calculated as:

$$\widetilde{R_j} = f(R_j) \tag{2.7}$$

where

$$f(x) = \begin{cases} \lceil x \rceil & \text{if } x - \lfloor x \rfloor \geq 0.5 \\ \lfloor x \rfloor & \text{otherwise} \end{cases} \tag{2.8}$$

The following policy is set up:

$$P_{new} = \begin{cases} 1 & B_{used} \leq C - \widetilde{R_j} - B_{new} \\ 0 & B_{used} > C - \widetilde{R_j} - B_{new} \end{cases} \tag{2.9}$$

where P_{new} is the admission probability for new calls, B_{used} and B_{new} are the number of used channels and the number of channels required by the incoming new call, respectively.

Fractional MBCR. In the above scheme, the reservation request is rounded to an integer number of channels. However, some information carried by the fractional part is lost during the rounding. For example, if the R_j is 2.6, $\widetilde{R_j}$ becomes 3, and therefore, the reservation is 15% more than the requirement. In order to fully use the information, fractional reservation is introduced. If R_j has integral part R_j^I and fractional part R_j^F, the scheme is defined as:

$$P_{new} = \begin{cases} 1 & B_{used} \leq C - R_j^I - B_{new} - 1 \\ 1 - R_j^F & B_{used} = C - R_j^I - B_{new} \\ 0 & B_{used} > C - R_j^I - B_{new} \end{cases} \tag{2.10}$$

It should be noted that the R_j value calculated according to Equation (2.5) is the *target reservation requirement* in cell j. However, sometimes, especially when traffic load is heavy, the number of available channels may be less than the target value. In this case, the above suggested CAC schemes still try to reserve the target values ($\widetilde{R_j}$ or R_j) during the valid period of R_j (*i.e.*, before R_j is updated). This means that if the available channels (*i.e.*, total number of channels minus the used ones) in a cell are less than the target reservation value ($\widetilde{R_j}$ or R_j), then at the time of reservation cell j reserves the available channels. However, as channels are released due to call completion or call handoff, these channels do not become available to new calls, and are kept for the handoff purposes only, until the reservation value R_j is reached, or until target reservation value is updated.

2.2 Performance Evaluation of MBCR Scheme

As mentioned before, the proposed MBCR scheme mainly aims to introduce the concepts of mobility and influence curve into the channel assignment and call admission control processes, and quantify the gains that can be achieved through this approach. In this section, the performance of MBCR is evaluated in a heterogeneous networking environment. The performance of MBCR is compared, in terms of achieved blocking probabilities and traffic capacity under certain QoS requirements, with the corresponding performance of a cellular system with fixed and dynamic guard channel schemes.

2.2.1 Model and Assumptions

In this simulation study, a more realistic system is considered as shown in Figure 2.2. The wireless network under consideration consists of 37 cells, each of which has six neighboring cells. The cells are wrapped around to eliminate the border effect. Solid thin lines in the figure represent the existence of narrow streets in the cell, while solid thick lines represent the main streets in a cell. Depending on the classes of users that they mainly serve, as well as their corresponding geographic layout, there are three different types of cells:

- Type 1 cells: cells along the main streets (*e.g.,* cell 8, 10, 30, ...). There are more high-speed users in these cells and the users are moving in a highly directional pattern;

- Type 2 cells: cells that represent the residential areas or shopping malls (*e.g.,* cells 1, 4, 18, ...). Users in these cells are mainly low-speed users that move randomly;

- Type 3 cells: cells with narrow streets (*e.g.,* cells 5, 25, 35, ...). The traffic pattern presented in these cells falls in between of type 1 and type 2.

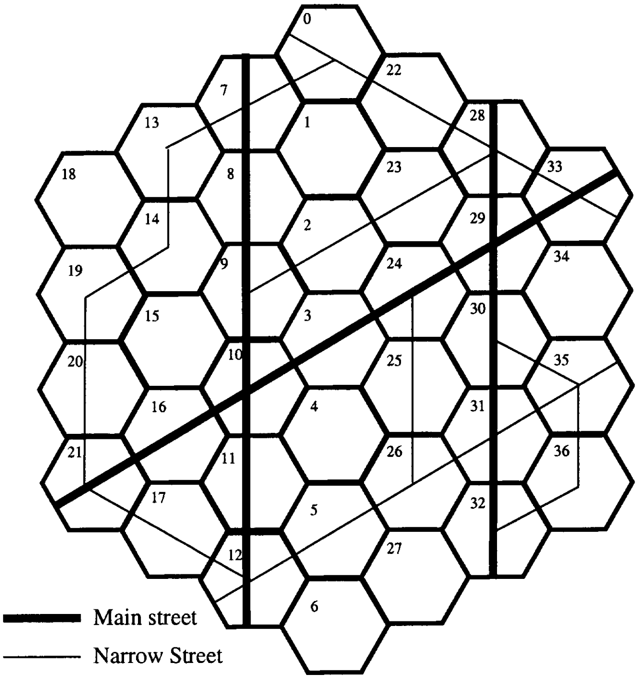


Figure 2.2 The layout of cellular system under consideration.

Other system parameters and assumptions are summarized as follows:

(1) Each cell has $C = 40$ channels;

(2) The arrival of new calls initiating in each cell forms a Poisson process with rate λ_n;

(3) Each call requires only one channel for service, $B_{new} = 1$;

(4) The life time of each call is exponentially distributed with mean 240 seconds;

(5) The cell dwell time probability density functions $f_h(t)$ and $f_l(t)$ follow exponential distributions with mean value 120 seconds and 600 seconds, respectively;

(6) The new call requests are generated by either high-speed mobiles or by low-speed mobiles with probability P_{high} or $1 - P_{high}$, respectively, where

$$P_{high} = \begin{cases} 70\% & \text{for type 1 cells} \\ 0\% & \text{for type 2 cells} \\ 30\% & \text{for type 3 cells} \end{cases}$$

(7) A cell reports the target reservation to all its neighboring cells every 30 seconds and the Integral MBCR is used as CAC;

It should be noted here that the directional factors are actually different for each user class. For example, in a cell that has a main street that goes through it, if a high-speed user request a handoff, it has only two possible target cells (along the street), while for a low-speed user, the corresponding handoff target cell can be any one of the neighboring cells with higher probabilities to cells along the street direction. Since this study concerns the statistical movement behavior of all the users in a cell, and not the corresponding moving behavior of the individual users, in the following, the directional factors ($\alpha_{i,j}$) are defined to be the weighted average of that of low-speed users and high-speed users, where the weight depends on the composition of traffic in each cell.

2.2.2 Comparison with Fixed Reservation Schemes

The QoS metrics considered in this study are the new call blocking probability (P_{nb}) and handoff call blocking probability (P_{hb}). As mentioned before, cells have different traffic characteristics, which result in different P_{high} and directional factors. Three typical cells (cell 1, cell 8 and cell 10) are chosen to investigate their performance. Cell 1 belongs to type 2, cell 8 and cell 10 belong to type 1: cell 10 has two crossing main streets within its coverage area, while cell 8 has only one.

In order to study and observe how the proposed scheme can control and adjust the channel reservation as the network conditions change, two sets of experiments are performed and the corresponding results are presented. The first set of experiments aim to compare the new call/handoff call blocking probability of the proposed scheme and the fixed reservation scheme, and demonstrate that proposed scheme can dynamically reserve channels for each cell and for different offered traffic loads, therefore, letting the system adaptively adjust to the changing traffic conditions. In this set of experiments, each cell is offered with the same new originating traffic load λ_n. However, as the system operates and evolves, the overall offered load to each cell changes, as a result of the generated handoff calls due to the continuous moving of the mobile users.

Figures 2.3, 2.4, 2.5 and 2.6 compare the new call/handoff call blocking probabilities of the MBCR scheme ($B = 0.3$, here the choice of B is based on experimentation) and the fixed reservation schemes for each cell. It can be observed form these figures that for the fixed reservation schemes, as the number of reserved channel (C_h) increases, the new call blocking probability increases and the handoff call blocking probability decreases. Under MBCR scheme, the reservation changes dynamically as the offered new call traffic load changes. For example, in cell 10, when the new call arrival rate is 0.065, the performance of MBCR scheme lies between two fixed reservation schemes with $C_h = 3$ and $C_h = 4$ respectively; while on the other end of the simulation range, $\lambda_n = 0.13$, MBCR scheme works like the fixed reservation scheme with $C_h = 6$. Comparing the results of cell 10 with the ones for cell 8 and cell 1, it can be seen that the within the given new call arrival rate, the dynamic range of cell 8 goes from $C_h = 2$ to somewhere between $C_h = 3$ and

$C_h = 4$; for cell 1 the range is between $C_h = 0$ and $C_h = 1$. It can be concluded from these figures that MBCR scheme can dynamically reserve channels for each cell and for different offered traffic loads. Its dynamic range covers several fixed reservation schemes therefore, overcoming the disadvantages of fixed reservation schemes and simplifying the channel allocation process.

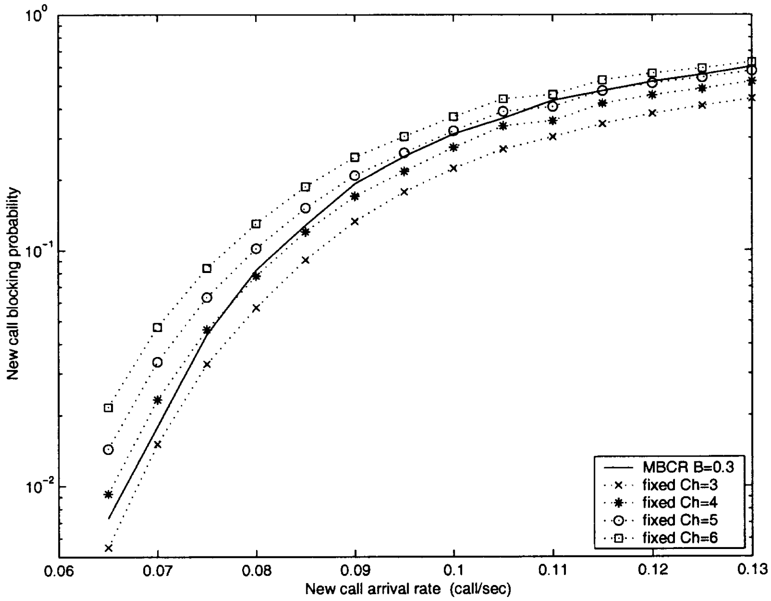


Figure 2.3 New call blocking probabilities for cell 10: $B = 0.3$.

Figure 2.7 compares the performance of these three cells using the MBCR scheme with $B = 0.3$. It can be observed that the operation points of cells are different from one another. For example, if the QoS requirement of a mobile user is $P_{nb} < 1\%$ and $P_{hb} < 0.1\%$, cell 10 works well when new call arrival rate is lower than 0.065, while the corresponding thresholds for cell 8 and cell 1 are 0.087 and 0.115, respectively.

Two factors contribute to this difference. First, the handoff call arrival rates (λ_h) are different for different cells. As mentioned before, cells have different traffic characteristics, and in proposed model, the handoff processes are tightly coupled with cell dwell time which is determined by the traffic of the cell. In cell 10, which has two main streets in its coverage area, most of the users in its neighboring cells are

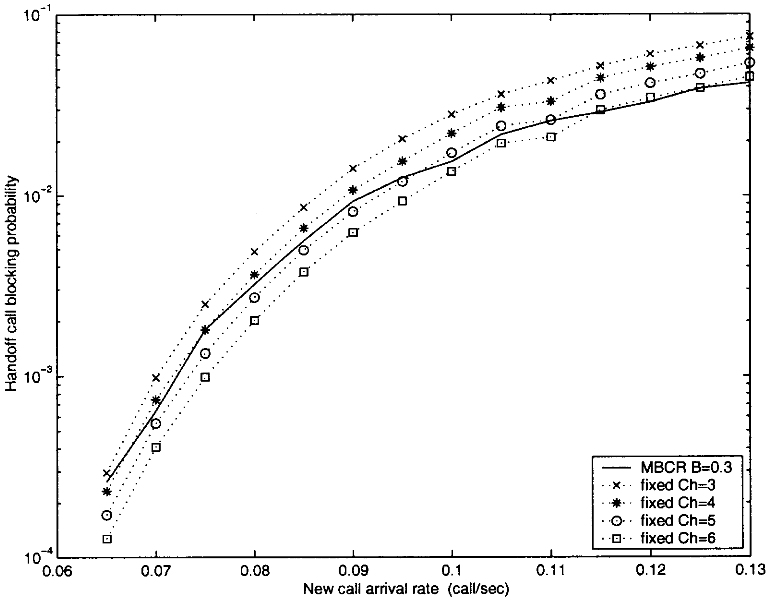


Figure 2.4 Handoff call blocking probabilities for cell 10: $B = 0.3$.

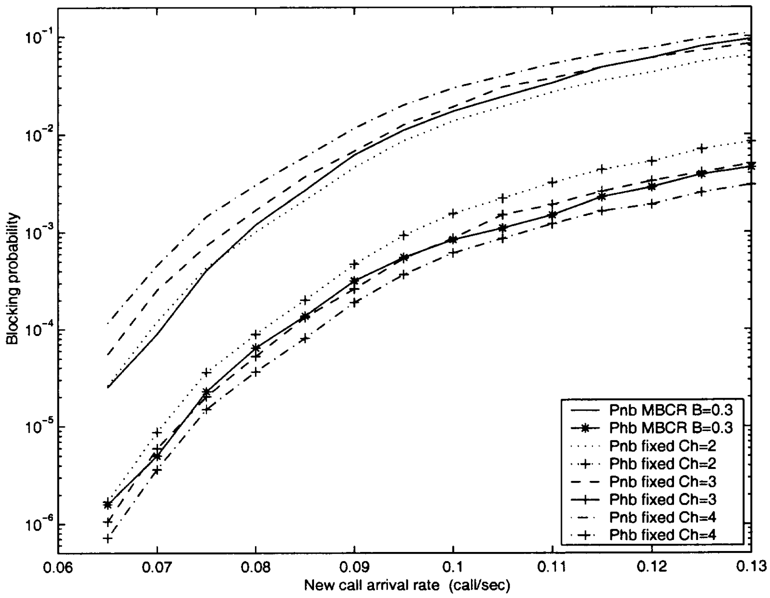


Figure 2.5 Call blocking probabilities for cell 8: $B = 0.3$.

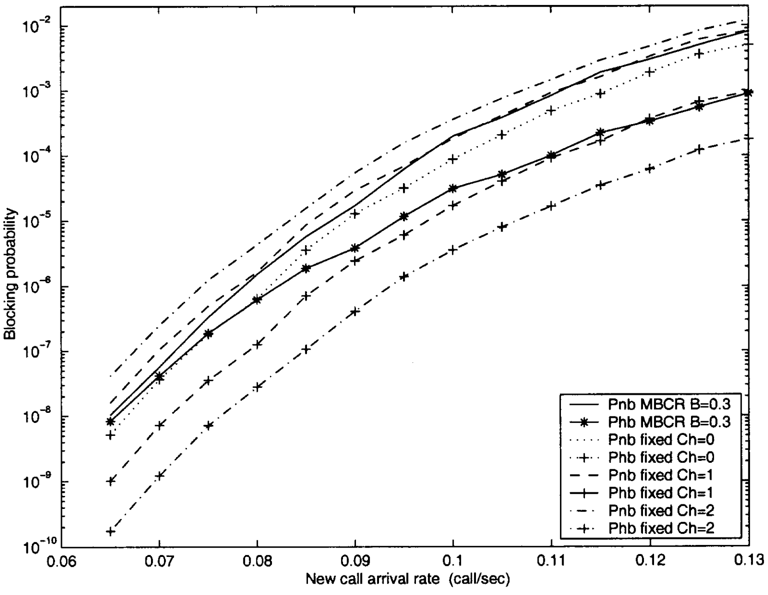


Figure 2.6 Call blocking probabilities for cell 1: $B = 0.3$.

high-speed users; while for cell 1, on the other extreme, the users in its neighboring cells are mainly low-speed users. Therefore, the handoff arrival rate for cell 10 is much higher than that of cell 1 when all the cells are loaded with the same new call arrival rate. Figure 2.8 shows the generated handoff arrival rates against the given new call arrival rates for the three cells under consideration. In real systems, the new call arrival rates of busy cells, such as cell 10, are much higher, and it will accordingly result in even higher handoff arrival rates. In order to handle such traffic, more channels should be allocated to the busy cells dynamically during the busy periods.

The second reason of the results in Figure 2.7 is that channel holding times are different for different cells. Channel holding time is the time that a call occupies a channel in a cell ([22, 23]). For new calls, the channel holding time is determined by call holding time, cell dwell time distribution and new call traffic composition (P_{high}); for handoff calls the channel holding time is determined by the residual call holding time, cell dwell time distribution and the traffic composition of handoff traffic. Table 2.1 compares the channel holding time and traffic composition for new call and handoff calls, where Tc_{new} and $Tc_{handoff}$ are the average channel holding times for

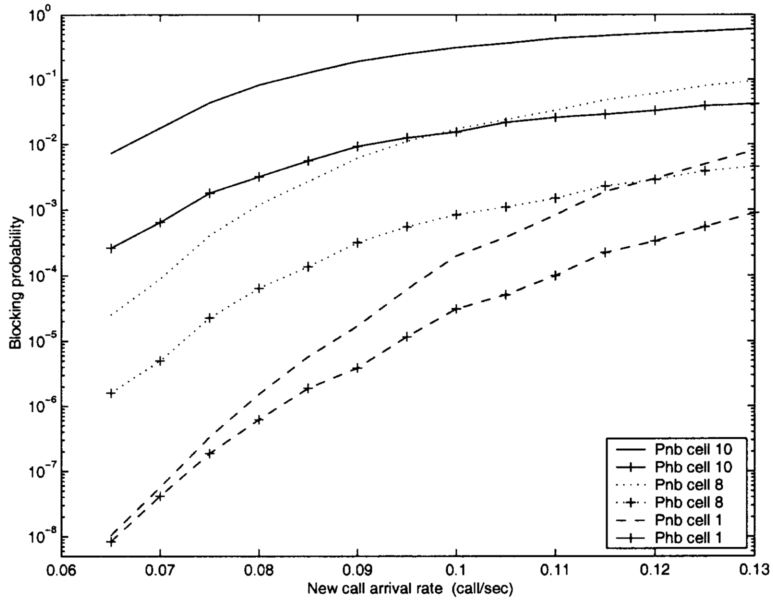


Figure 2.7 Call blocking probabilities of MBCR scheme for different cells: $B = 0.3$.

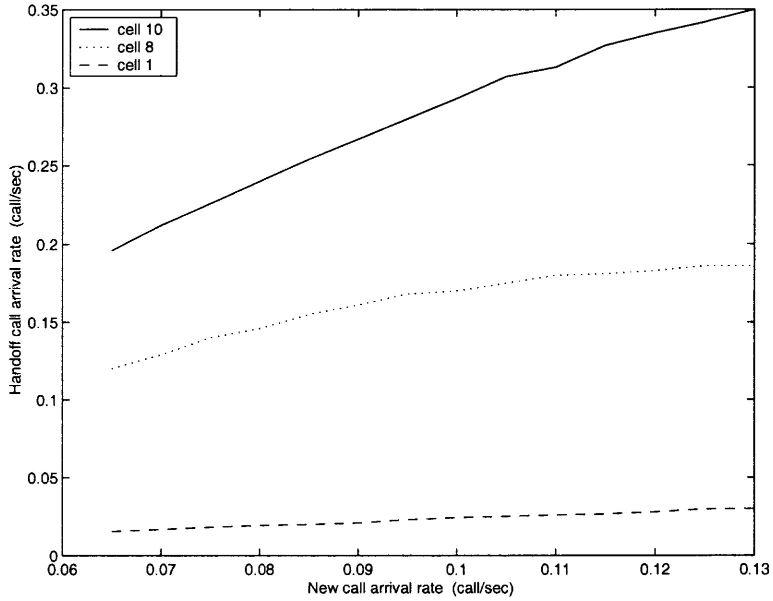


Figure 2.8 Handoff call arrival rates for different cells.

new calls and handoff calls respectively, and P_{hfhigh} is the percentage of high-speed calls in the handoff traffic.

Table 2.1 Channel Holding Time for Different Cells

cell	P_{high}	Tc_{new}	P_{hfhigh}	$Tc_{handoff}$
10	70%	107.3sec	92%	87sec
8	70%	107.4sec	85%	93.8sec
1	0%	171.1sec	0%	171.1sec

It is observed that for cells with combined traffic (cell 10 and cell 8), channel holding time for handoff calls is always shorter than that of the new calls. That is due to the fact that low-speed calls tend to terminate in the originating cell, while there are more high-speed users in the handoff traffic than in the new call traffic. Furthermore, it has been demonstrated by Xie *et al.* that the pdf of the speeds of handoff terminals follows the "Biased Sampling formula" which favors the high speed terminals [84]. The analysis of guard channel systems with different channel holding times for new calls and handoff calls is discussed in detail in Section 2.3.

Figure 2.9 shows the effect of different values of B on the corresponding call blocking probabilities. As mentioned before, B is a constant parameter representing the relationship between the influence and reservation. With the increase of B, the number of reserved channels also increases, which will cause the new call blocking probability to increase and handoff call blocking probability to decrease. The choice of B depends on the user's QoS requirement and the system capacity.

The second set of experiments demonstrate that the proposed algorithm is capable of controlling and adjusting dynamically the channel reservation values to the appropriate levels, in order to offer the same quality of service to the ongoing calls throughout the system (similar handoff call blocking probability). During this experiment, the same level of total offered load (*i.e.*, composite new call offered load

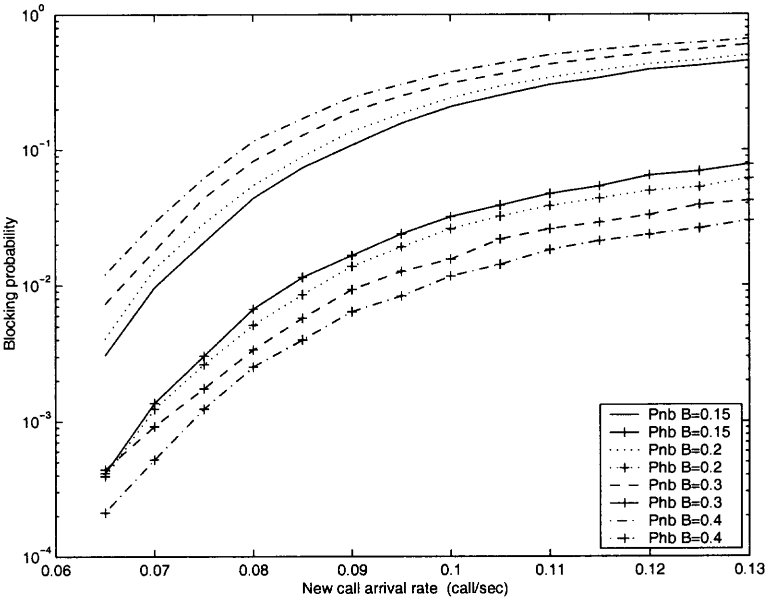


Figure 2.9 Call blocking probabilities for cell 10: different values of B.

and handoff call offered load) is maintained to each cell and the performance of the new call blocking probability and the handoff call blocking probability is presented as a function of the total overall offered load per cell. However, the combination of the traffic (handoff calls vs. new calls) in each cell can be different, although the total composite traffic per cell is the same. The corresponding results for the fixed reservation scheme and MBCR scheme are compared in Figure 2.10 and 2.11. It can be observed from these figures that the MBCR scheme provides similar handoff call blocking probability to all the users, independent of the area they are moving towards (target cell) and independent of the traffic distribution in each cell (*i.e.*, handoff calls vs. new calls), under the assumption that the overall composite offered load (*i.e.*, handoff call load plus new call load) presented to each cell is the same. On the other hand, the handoff call blocking probabilities for fixed reservation schemes are different from cell to cell. These results demonstrate that MBCR scheme achieves to balance the handoff call blocking probabilities throughout the whole system, which actually corresponds to an extra increase of 10% of the maximum traffic load that could be

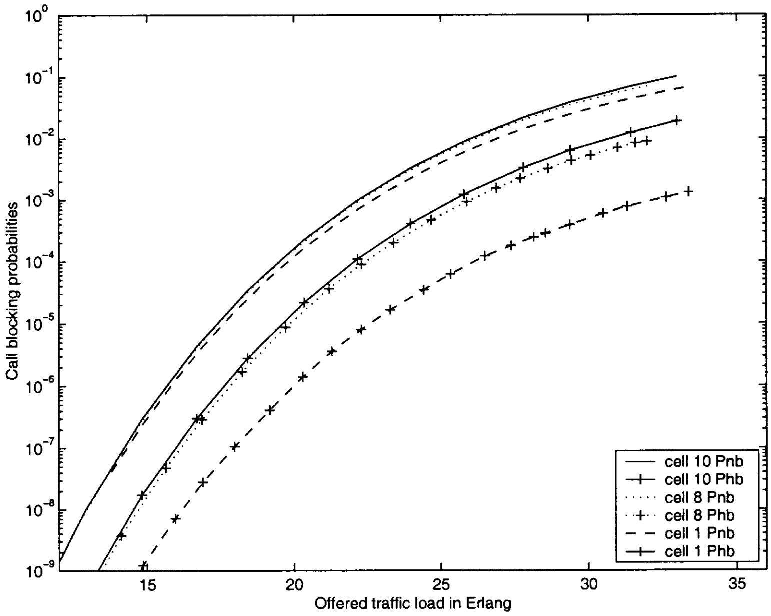


Figure 2.10 Call blocking probabilities vs. total offered load for fixed reservation scheme: $C_h = 2$.

handled by a system with specific objectives and requirements on the handoff blocking probability of the ongoing calls.

2.2.3 Comparison with PRP Dynamic Scheme

In this study, the performance of MBCR scheme is compared with that of a representative dynamic channel reservation scheme, namely, the Predictive Reservation Policy (PRP) [6].

The main principle of PRP is that it dynamically reserves channels when the number of communications in progress grows in a given cell. When the number of occupied channels $N(t)$ reaches the defined threshold k or a multiple of k in a cell, the cell requests reservation of resources in the neighboring cells for which the probability of transition is high. If the neighbors have free channels, the reservations take place immediately. Otherwise, the PRP algorithm waits for a free channel. Two thresholds are considered: if the transition probability is lower than a given value p, the reservation threshold is k_0, else it is k_1.

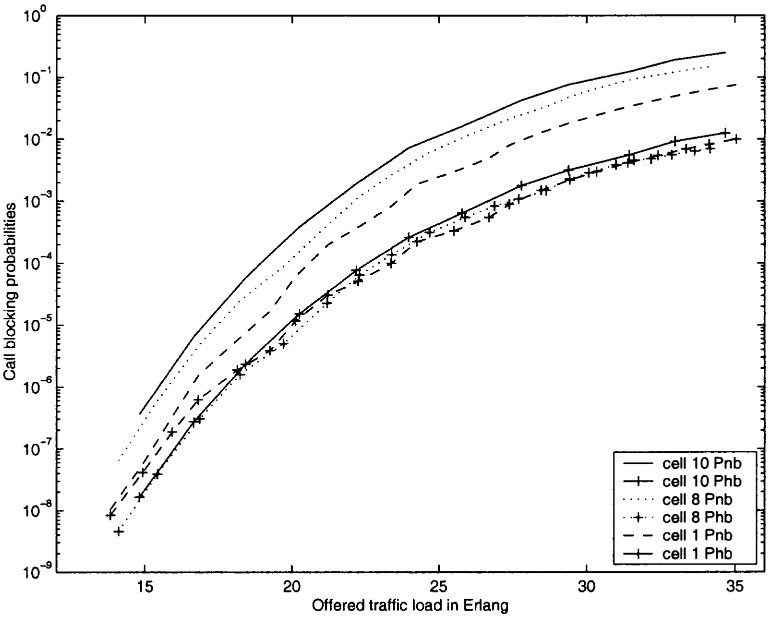


Figure 2.11 Call blocking probabilities vs. total offered load for MBCR scheme: $B = 0.3$.

In the following, the PRP scheme is compared with the proposed MBCR scheme for the system described in the previous section. Regarding the PRP, the following specific parameters are used:

- $p = \frac{1}{6}$;
- $k_0 = 15$ and $k_1 = 10$;

Figures 2.12, 2.13 and 2.14 compare the average number of reserved channels, the new call blocking probability and the handoff call probability of PRP scheme and MBCR scheme for cell 10 and cell 8. It is observed that the PRP scheme reserves much more channels than MBCR scheme in both cells. As a result, the PRP scheme makes the handoff call blocking probabilities unnecessary low. And the penalty for PRP is that too many new calls are blocked and the resource utilization is low. Moreover, Figure 2.15 shows the achieved performance of PRP scheme if the same level of total offered load (*i.e.*, composite new call offered load and handoff call offered load) is maintained to each cell as the second set of experiments of previous section. It can

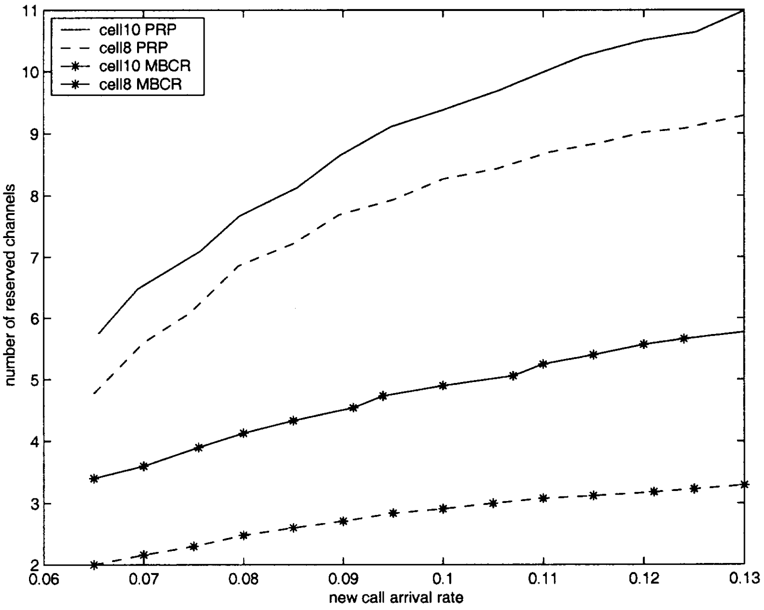


Figure 2.12 Comparison of average number of reserved channels.

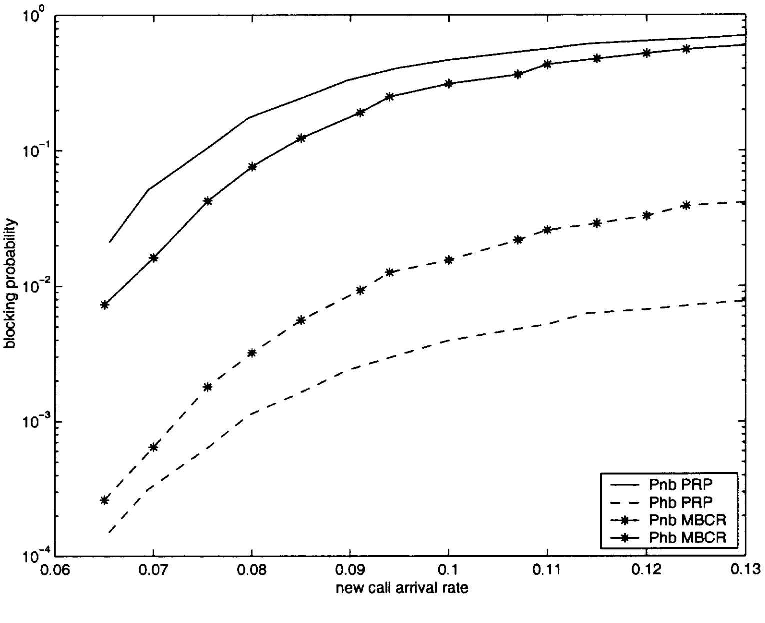


Figure 2.13 Comparison of call blocking probabilities for cell 10.

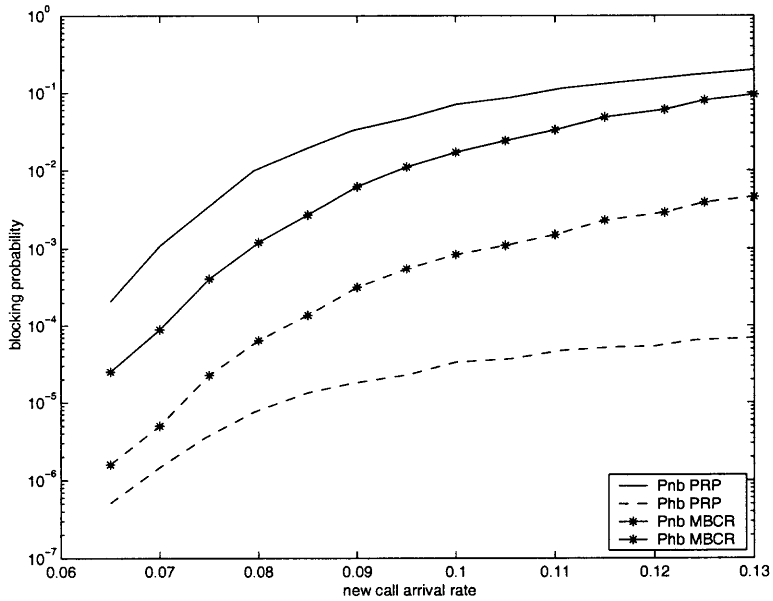


Figure 2.14 Comparison of call blocking probabilities for cell 8.

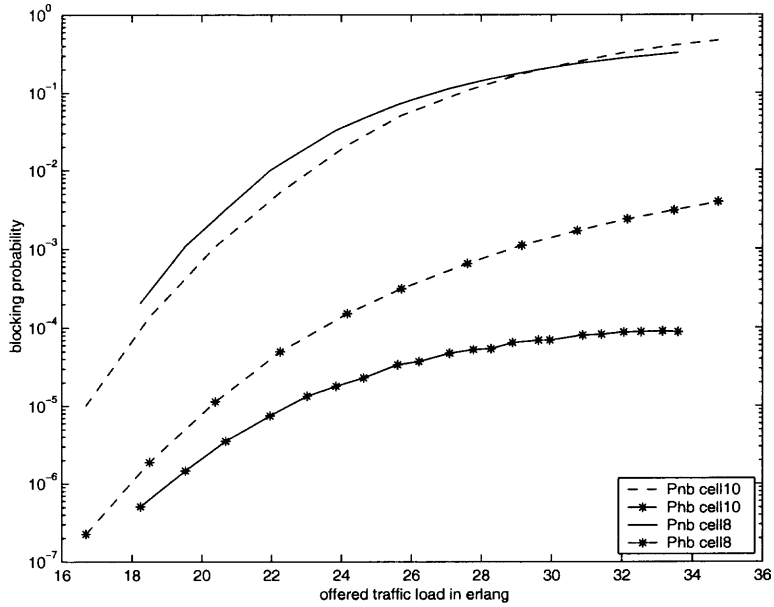


Figure 2.15 Call blocking probabilities vs. offered traffic load.

be observed that PRP dynamic scheme, just like the fixed guard channel schemes, cannot balance the handoff call blocking probabilities throughout the whole system. The reason is that since PRP does not use the user mobility information, although the reservation changes dynamically, it could not guarantee the same level of QoS throughout the system.

2.3 Analysis of Guard Channel Scheme with Different New/Handoff Call Channel Holding Time

In some traditional research efforts, when trying to analyze the call blocking probabilities of guard channel schemes, researchers always assume that new calls and handoff calls have the same channel holding time $1/\mu$. Under this assumption the problem is simplified, in the sense that system can be modeled by a one dimensional Markov chain and the new call blocking probability (P_{nb}) and handoff call blocking probability (P_{hb}) can be calculated accordingly.

However, some recent analysis [22, 23] showed that the above assumption (same channel holding time) does not hold for a cellular network, and in Section 2.2.2, it is observed that the handoff calls have shorter average channel holding time than new calls in cells with multiple mobility platforms. In this case, in order to analyze and calculate the corresponding blocking probabilities, it is necessary to use a two dimensional Markov chain model instead of the one dimensional model.

The transition diagram of the system with different channel holding time is given in Figure 2.16, where λ_n, λ_h, $1/\mu_n$ and $1/\mu_h$ are new call arrival rate, handoff call arrival rate, average new call channel holding time and average handoff call holding time, respectively; C is the capacity of the cell and C_h is the number of guard channels. The state space is defined as:

$$S = \{(n_1, n_2) | 0 \leq n_1 \leq C - C_h, 0 \leq n_1 + n_2 \leq C\} \qquad (2.11)$$

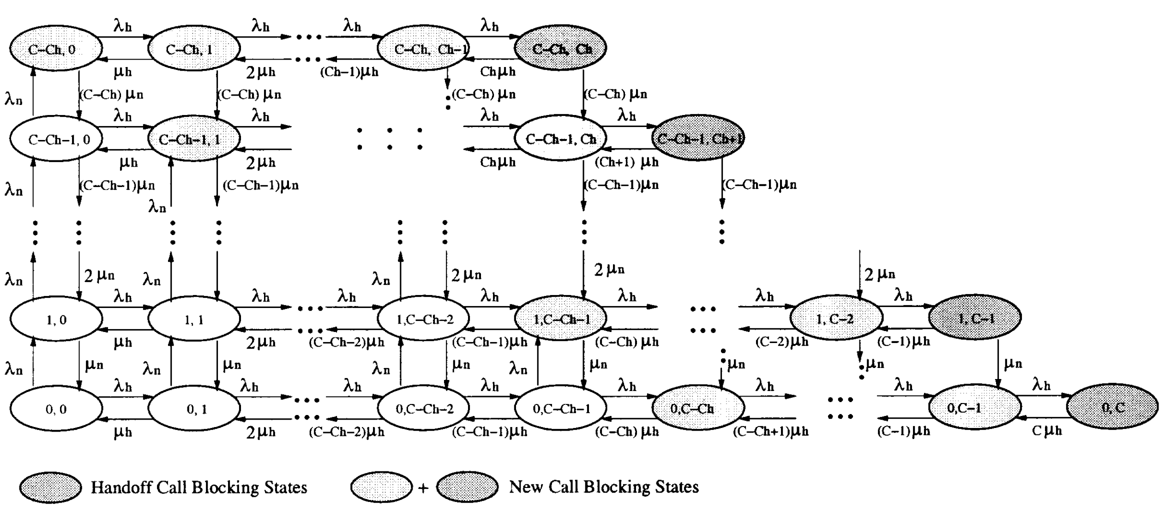


Figure 2.16 The state transition diagram.

where n_1 is the number of admitted new calls and n_2 is the number of admitted handoff calls. From the diagram, the state-transition equations can be obtained as shown below.

(i) If $n_1 = 0$, then

$$(\lambda_n + \lambda_h)P_{0,0} = \mu_h P_{0,1} + \mu_n P_{1,0} \qquad \text{for } n_2 = 0;$$

$$(\lambda_n + \lambda_h + n_2\mu_h)P_{0,n_2} = \lambda_h P_{0,n_2-1} + (n_2+1)\mu_h P_{0,n_2+1} + \mu_n P_{1,n_2}$$

$$\text{for } 1 \le n_2 \le C - C_h - 1;$$

$$(\lambda_h + n_2\mu_h)P_{0,n_2} = \lambda_h P_{0,n_2-1} + (n_2+1)\mu_h P_{0,n_2+1} + \mu_n P_{1,n_2}$$

$$\text{for } C - C_h - 1 < n_2 < C;$$

$$(n_2\mu_h)P_{0,n_2} = \lambda_h P_{0,n_2-1} \qquad \text{for } n_2 = C;$$

(ii) If $1 \le n_1 \le C - C_h - 1$, then

$$(\lambda_n + \lambda_h + n_1\mu_n)P_{n_1,0} = \lambda_n P_{n_1-1,0} + (n_1+1)\mu_n P_{n_1+1,0} + \mu_h P_{n_1,1} \qquad \text{for } n_2 = 0;$$

$$(\lambda_n + \lambda_h + n_1\mu_n + n_2\mu_h)P_{n_1,n_2} = \lambda_n P_{n_1-1,n_2} + \lambda_h P_{n_1,n_2-1} + (n_1+1)\mu_n P_{n_1+1,n_2} +$$

$$(n_2+1)\mu_h P_{n_1,n_2+1} \qquad \text{for } 1 \le n_2 \le C - C_h - n_1 - 1;$$

$$(\lambda_h + n_1\mu_n + n_2\mu_h)P_{n_1,n_2} = \lambda_n P_{n_1-1,n_2} + \lambda_h P_{n_1,n_2-1} + (n_1+1)\mu_n P_{n_1+1,n_2} +$$

$$(n_2+1)\mu_h P_{n_1,n_2+1} \qquad \text{for } n_2 = C - C_h - n_1;$$

$$(\lambda_h + n_1\mu_n + n_2\mu_h)P_{n_1,n_2} = \lambda_h P_{n_1,n_2-1} + (n_1+1)\mu_n P_{n_1+1,n_2} + (n_2+1)\mu_h P_{n_1,n_2+1}$$

$$\text{for } C - C_h - n_1 + 1 \le n_2 \le C - n_1 - 1;$$

$$(n_1\mu_n + n_2\mu_h)P_{n_1,n_2} = \lambda_h P_{n_1,n_2-1} \qquad \text{for } n_2 = C - n_1;$$

(iii) If $n_1 = C - C_h$, then

$$(\lambda_h + n_1\mu_n)P_{n_1,0} = \lambda_n P_{n_1-1,0} + \mu_h P_{n_1,1} \qquad \text{for } n_2 = 0;$$

$$(\lambda_h + n_1\mu_n + n_2\mu_h)P_{n_1,n_2} = \lambda_h P_{n_1,n_2-1} + (n_2+1)\mu_h P_{n_1,n_2+1} \quad \text{for } 1 \le n_2 \le C - n_1 - 1;$$

$$(n_1\mu_n + n_2\mu_h)P_{n_1,n_2} = \lambda_h P_{n_1,n_2-1} \qquad \text{for } n_2 = C - n_1;$$

From the diagram, it can be observed that the given Markov chain is not of product form, and thus there is no closed form solution to it. Therefore, state probabilities must be calculated directly from the above equations and the normalization condition:

$$\sum_{(n_1,n_2)\in S} P_{n_1,n_2} = 1; \tag{2.12}$$

With the state probabilities of all the possible states, the new call blocking probability (P_{nb}) and the handoff call blocking probability (P_{hb}) can be obtained as:

$$\begin{aligned} P_{nb} &= \sum_{n_1=0}^{C-C_h} \sum_{n_2=C-C_h-n_1}^{C-n_1} P_{n_1,n_2} \\ P_{hb} &= \sum_{n_1=0}^{C-C_h} P_{n_1,C-n_1} \end{aligned}$$

The analytical results calculated according to above method are compared with the simulation results in Figure 2.17. From this figure, it can be observed that the proposed model matches the simulation results perfectly.

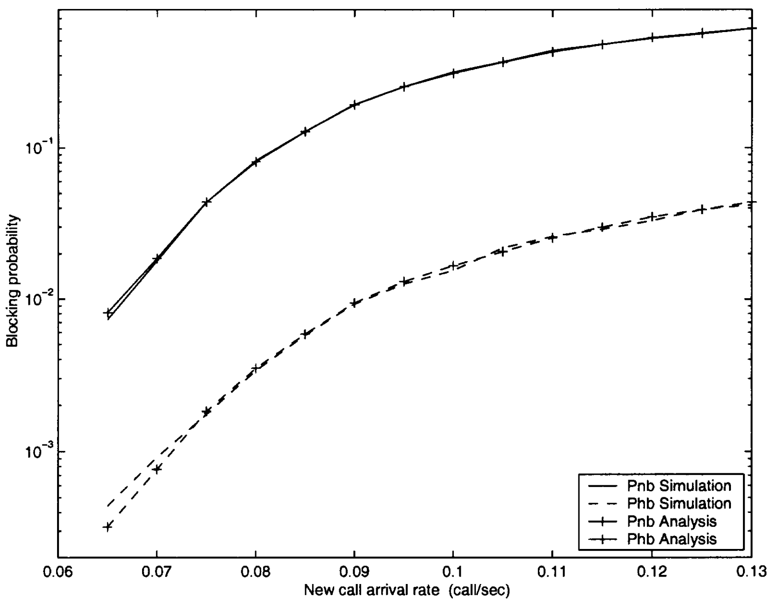


Figure 2.17 Comparison of simulation results and analysis results.

2.4 Predictive Mobility-Based Bandwidth Reservation Scheme

Wireless geolocation technology has received considerable attention over the past few years [8]. Among the basic functions of wireless geolocation is to figure out the position of mobile stations in the service area. Although originally used and developed to support the FCC E-911 requirements [8], this new function facilitates several applications which will benefit businesses as well as consumers. At the same time, the real time position information can also be used by resource management mechanism to carry out functions such as resource allocation and call admission control.

This section presents the Predictive Mobility-Based Bandwidth Reservation Scheme (PMBBR) as an improved version of Mobility-based Channel Reservation Scheme (MBCR) discussed above. The PMBBR scheme takes advantage of the geolocation technology: based on the history user location information, the future moving pattern of users can be predicted. By integrating these predictions into the reservation making process, the PMBBR scheme achieves to optimize the efficiency of handoff mechanisms and minimize, if not eliminate, the unnecessary reservation of resources, and therefore, improve the system capacity and throughput. Furthermore, bandwidth reconfiguration based call admission control processes are developed that may allow the efficient resource re-distribution in a cell to balance the QoS among all the mobile users in the cell, especially when users with flexible QoS requirements are supported in the system. By combining these two approaches, an integrated resource management strategy is proposed that can be implemented in next generation wireless networks that support multimedia services (data, voice, video, *etc.*).

2.4.1 Description of PMBBR Scheme

The original MBCR scheme implies that each user requires one logical channel for service. However, for next generation wireless networks that support multiple types of services, each user may have different bandwidth requirements. Therefore, it is important to enhance the MBCR scheme to support different bandwidth requirements for different users and/or classes of service. In the following, this enhanced scheme is referred to as Mobility-based Bandwidth Reservation (MBBR) scheme.

In the MBBR scheme, the influence curve for each ongoing call k is defined as:

$$I_k(i, j, t_k, T) = BW_k \alpha_{i,j} L(t_k, T) \tag{2.13}$$

where BW_k is the bandwidth requirement of the call under consideration, $\alpha_{i,j}$ and $L(t_k, T)$ follow the definitions in Section 2.1

The amount of bandwidth cell j needs to reserve at time T can be calculated based on Equations (2.3), (2.4) and (2.5):

$$R_j = \sum_{i \in N_j} R_{i,j} = B \sum_{i \in N_j} \sum_{k \in S_i} I_k(i, j, t_k, T) \tag{2.14}$$

where $R_{i,j}$ is the amount of bandwidth needed to be reserved in cell j for ongoing calls in cell i, N_j denotes the set of all neighboring cells to the cell j, S_i denotes the set of all the currently ongoing calls in cell i, and B is a tunable constant which models the relationship between influence and reservation.

With the moving speed and direction predictions, the MBBR scheme can be improved in the following aspects:

(1) *The directional factors.* In MBBR scheme, the directional factors are statistical values for all the users in current cell, which means that the influence value of each single user is distributed to all the neighbor cells according to directional factors. With the predicted moving direction and current position, the handoff target cell can be accurately calculated, so that the reservation is made only in one cell, and therefore, the resource waste in other neighbor cells can be eliminated.

(2) *The handoff probability.* Let t_k^{hf} denote the time interval in the future that the user k under consideration needs to request a handoff, which can be calculated based on the current position and the predicted moving speed and direction of the user. The probability that this ongoing call will request a handoff in the future can be calculated as:

$$\begin{aligned} P_k &= \Pr(\text{this call will request a handoff }) \\ &= \Pr(\text{the residual call lifetime is longer than } t_k^{hf}) \\ &= \int_{t_k^{hf}}^{\infty} f(\tau) d\tau \end{aligned} \tag{2.15}$$

where $f(\tau)$ is the residual call life time probability density function. If the call life time is assumed to be exponentially distributed with mean value T_l, the above probability equals: $e^{-t_k^{hf}/T_l}$

(3) *Reservation Timing.* A critical issue that influences the performance of a bandwidth reservation mechanism is the actual time that the reservation is made for the incoming handoff calls. If the reservation is made at the time that it can be used at the near future, such a scheme can achieve better performance. Otherwise, the reservation could result in waste of resources: if the reserved resources are not

used by a handoff call, the system will incur unnecessary new call blocking. This problem is also addressed in [14], in which the concept of threshold distance (TD) is introduced to reduce the likelihood of false reservations: users inside the TD circle will not submit reservation requests. However, since users may have different moving speed, a high speed user that is currently located inside the TD circle may move out of the coverage region of the cell earlier than a low speed user that is outside the TD circle. Therefore, distance alone is not a good solution to the problem of reservation timing. Based on these considerations, a *reservation advance time* t_{thre} is set. Reservations are made only for those calls which will request handoffs in the near future, *i.e.*, $t_k^{hf} \leq t_{thre}$.

Taking all the above factors into consideration, the PMBBR scheme determines the total amount of bandwidth to be reserved in cell j as:

$$R_j = B \sum_{i \in N_j} \sum_{k \in S_i'} BW_k P_k \tag{2.16}$$

where S_i' is the set of those ongoing calls which are currently in cell i and, according to the prediction, are going to handoff to cell j within time interval t_{thre}. Notice that, since the number of mobile users in the network and the location history of each user keep changing as time evolves, the bandwidth reservation should be re-calculated periodically.

2.4.2 Bandwidth Reconfiguration Based Call Admission Control

This section introduces the detailed call admission control and proposes resource reconfiguration mechanism to be used for new and handoff calls. In the following, for simplicity and without loss of generality, it is assumed that there are two classes of traffic in the network:

- Class 1 (realtime traffic): The desired bandwidth for each class 1 user is BW_1^u. If this requirement cannot be met, the user may have the option to continue at a lower bandwidth requirement BW_1^l, for instance by adjusting the coding rate so that the video/audio quality is still acceptable.

- Class 2 (non-realtime traffic): The desired bandwidth for each class 2 user is BW_2^u and there are no strict QoS requirements. However, some flexible QoS requirements are defined for this service type. The user could specify a set of acceptable QoS levels that correspond to bandwidth requirements that range from a lower bound bandwidth requirement BW_2^l to a maximum bandwidth requirement BW_2^u, and expect a QoS varying in the specified range.

The basic underlying principles of the proposed CAC scheme for the case of multiple classes of service with different priorities are: handoff calls always have higher priority than new calls and class 1 traffic has higher priority than class 2 traffic to access bandwidth resources.

The following notations are used throughout the rest of the chapter:

- BW_{total}: total bandwidth capacity of a cell;
- BW_1^{used}: total bandwidth used by all the class 1 users in a cell;
- BW_2^{used}: total bandwidth used by all the class 2 users in a cell;
- $BW_2^{minused}$: minimum bandwidth that should be kept for current class 2 users being served, to meet their minimum bandwidth requirements. $BW_2^{minused}$ can be calculated as:

$$BW_2^{minused} = N_{class2} \times BW_2^l \tag{2.17}$$

  where, N_{class2} is number of class 2 calls in service;

- BW_1^{res}: bandwidth reservation request in a cell for class 1 users (for handoff purposes) from all neighbor cells. BW_1^{res} can be calculated using Equation

(2.16) for only class 1 traffic as follows:

$$BW_1^{res} = B \sum_{i \in N_j} \sum_{k \in S'_{i,1}} BW_1^u P_{k,1} \tag{2.18}$$

where $P_{k,1}$ is the handoff probability of class 1 user k, and $S'_{i,1}$ is the set of those ongoing class 1 calls which are currently in cell i and are going to handoff to cell j in less than t_{thre}.

- BW_2^{res}: bandwidth reservation request in a cell for class 2 users (for handoff purposes) from all neighbor cells. BW_2^{res} can be calculated using Equation (2.16) for only class 2 traffic, as follows:

$$BW_2^{res} = B \sum_{i \in N_j} \sum_{k \in S'_{i,2}} BW_2^u P_{k,2} \tag{2.19}$$

where $P_{k,2}$ is the handoff probability of class 2 user k and $S'_{i,2}$ is the set of those ongoing class 2 calls which are currently in cell i and are going to handoff to cell j in less than t_{thre}.

The following paragraphs describe conceptually how the proposed CAC scheme operates. For a new connection, the proposed scheme works as follows. For a class 1 new call, the scheme first attempts to allocate the desired amount of bandwidth BW_1^u, if it is available. Otherwise, the scheme tries to accept the new class 1 call at the degraded quality (*i.e.,* at bandwidth BW_1^l), however, still acceptable to the user (according to the pre-specified characteristics and requirements of a class 1 user). If there is still not sufficient bandwidth available to do so, the call is rejected. It should be noted here that the available bandwidth is calculated based on the total bandwidth capacity, the bandwidth currently used by active calls and the bandwidth that has been reserved in the current cell for handoff purposes. The bandwidth reservation at each step is determined by the PMBBR scheme described in Equations (2.18) and (2.19), for class 1 and class 2 traffic, respectively. It should be also pointed out here that the bandwidth to be reserved in a cell represents the collective effect of the influence that the class 1 (class 2) traffic from neighboring cells exerts on the cell under consideration, and does not refer specifically to individual users. Therefore, this bandwidth is available to be used for all the calls that belong to the corresponding class as they may move into the target cell.

For a class 2 new call, the scheme first attempts to allocate the desired amount of bandwidth BW_2^u, the call is accepted if there is enough bandwidth available. Otherwise, the call is rejected. In this case, the new class 2 call is not accepted at degraded quality where less bandwidth is required although such bandwidth could be available at the time of the call generation. The reason lies in that a large number of class 2 calls supported at the lowest acceptable bandwidth (*i.e.*, BW_2^l) will saturate the system. It is desirable to keep some class 2 users operate at higher bandwidth, and therefore, make possible the resource borrowing from ongoing class 2 calls to accommodate future class 1 and class 2 handoff calls.

For handoff connections, the proposed scheme works as follows. For a class 1 handoff call, it first attempts to allocate the desired amount of bandwidth BW_1^u. If this is not available, the scheme tries to continue supporting the handoff class 1 call at slightly degraded quality (*i.e.*, allocate bandwidth BW_1^l). If the available bandwidth is still not sufficient enough to do so, the bandwidth reconfiguration is initiated which borrows bandwidth from current class 2 calls in this cell that are supported with bandwidth higher than their minimum requirements. The detailed procedure of the bandwidth reconfiguration as well as the selection of the class 2 users to participate in this process is described later in this section. The handoff call is dropped only when reconfiguration process could not get enough bandwidth to support it.

For a class 2 handoff call, the proposed scheme first attempts to allocate the desired amount of bandwidth BW_2^u. If this is not available, the scheme tries to continue supporting the handoff class 2 call at some lower bandwidth within the user pre-specified range (*i.e.*, between BW_2^l and BW_2^u). At this step, the system attempts to allocate to the handoff class 2 call the maximum available bandwidth (within its pre-specified range) without involving the bandwidth reconfiguration process. As a last attempt to accept this handoff call, the scheme tries to allocate the minimum required bandwidth BW_2^l to this call by initiating the bandwidth reconfiguration process. Again, as can be observed by the order of the above steps the bandwidth reconfiguration process is invoked as a last resort in an attempt to minimize the overhead associated with this process and to minimize the number of other class 2 calls impacted. Finally the handoff call is dropped only when reconfiguration process could not get enough bandwidth to support it.

As can be observed in the above procedure, under certain situations, bandwidth reconfiguration is required to serve the handoff call at the cost of degrading class 2 call(s) currently in service. In the following, a reconfiguration strategy is discussed which minimizes the impact of bandwidth reconfiguration on current class 2 calls in the cell. A virtual queue is established and maintained for all the class 2 calls in each cell. The queueing policy is the handoff time t_{hf}^k (the time interval in the future that the user k will request a handoff, as defined in Section 2.4.1), the user with lowest handoff time is placed at the head of the queue. Each time a class 2 (new or handoff) request is accepted in the cell with bandwidth more than BW_2^l, it is put at the appropriate position of this virtual queue (based on the predicted handoff time). The queue is reordered each time new t_{hf}^k is available. Whenever bandwidth is needed from current class 2 call(s), the call at the head of the queue decreases its bandwidth occupation until enough bandwidth is spared or the lower bound BW_2^l is met. The above procedure is repeated if necessary. This reconfiguration strategy guarantees that the number of impacted class 2 users is minimized and the degradation time is also reduced to the minimum possible values.

The above resource reconfiguration based CAC process always attempts to provide the required resources to meet the service quality of class 1 calls that present more strict QoS requirements, whether these are new call attempts or handoff calls. At the same time, non-realtime traffic (class 2 traffic) that has been accepted into the wireless networks can maintain high successful handoff rate.

2.4.3 Performance Analysis

This section presents the performance analysis of the proposed scheme. Specifically, the performance results of the proposed integrated strategy, where both advanced bandwidth reservation (via the PMBBR scheme) and QoS management (via the call admission control and resource reconfiguration scheme) are implemented, are compared with the corresponding results of a conventional system where the fixed bandwidth reservation is implemented, in terms of achievable new call and handoff call blocking probabilities. Furthermore, in order to gain some insights into the individual impact of the different components of the proposed integrated solution, the performance of a system, where only the direction-based advanced bandwidth reservation (via the PMBBR scheme) is implemented while the bandwidth reconfiguration based call admission control component is not implemented, is investigated. The following sections first describe the model and assumptions used throughout the performance study, and then present the corresponding results of the comparative study.

Model and Assumptions. The wireless network used throughout this study is composed of 37 cells, each of which has six neighboring cells. The cell radius is set to be 1000 meters. In order to approximate the performance of a large cellular system the cells are wrapped around to eliminate the border effect. The arrival of new calls initiated in each cell forms a Poisson process with rate λ. The life time of each call is exponentially distributed with mean 240 seconds [34, 53]. Additional system and traffic parameters are summarized in Table 2.2. It should be noted here that the proposed framework aims to suggest a general approach that supports the seamless operation and provides flexibility for the resource management in the next generation wireless networks that support multimedia services. The bandwidth values for the different classes of service were chosen for illustration purposes. New voice coding technologies and emerging data applications may bring different bandwidth requirements to the wireless networks, which however can be easily fit into the framework. Throughout the simulation study, it is assumed that the desired bandwidth requirement for class 1 (*e.g.,* voice) users is 30 Kbps and for class 2 (data) users is 50 Kbps. These parameters are selected based on some current realistic systems and some other research efforts that have been reported in the literature on this topic [64]. For instance, a GPRS terminal is able to download data at the speed up to 40-50Kbps and the voice data is sent at 22.8Kbps.

Table 2.2 Simulation Parameters

Parameter	Value	Description
BW_{total}	$1000Kbps$	Total Bandwidth Capacity of a cell
BW_1^u	$30Kbps$	Desired bandwidth requirement for class 1 users
BW_1^l	$25Kbps$	Lower bound bandwidth requirement for class 1 users
BW_2^u	$50Kbps$	Desired bandwidth requirement for class 2 users
BW_2^l	$5Kbps$	Lower bound bandwidth requirement for class 2 users

The mobility model used throughout this study is as follows. When a new call is initiated, the corresponding MS is assigned with a random initial position inside the cell, a random moving direction and an initial moving speed which is chosen according to a uniform distribution in the interval $[0, V_{max}]$mile/hr. The speed (v) and direction

(ϕ) are updated every time interval Δt, according to the following model:

$$v_{new} = \begin{cases} \min\{\max[v_{old} + \Delta v, 0], V_{max}\} & \text{when } p \leq 0.9 \\ 0 & \text{otherwise} \end{cases} \tag{2.20}$$

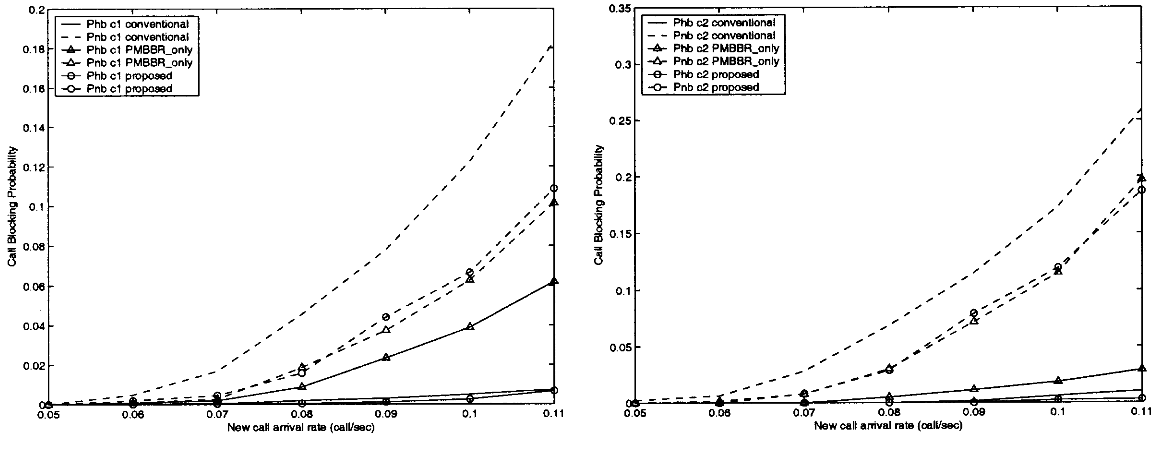
$$\phi_{new} = \phi_{old} + \Delta\phi \tag{2.21}$$

where Δv models the acceleration/deceleration of the mobile user and is a uniformly distributed random variable over the interval $[-5mile/hr, 5mile/hr]$; $\Delta\phi$ characterizes the user's change in moving direction and is a uniformly distributed variable over the interval $[-\Delta\phi_{max}, \Delta\phi_{max}]$; p is a uniformly distributed random variable over the interval $[0, 1]$. The use of variable p allows us to simulate the situation where a mobile user may stop occasionally during the course of moving. Based on this mobility model, two mobility patterns are considered in this study: high speed pattern and low speed pattern. In the high speed pattern, the mobility parameters are set to be $V_{max} = 60mile/hr, \Delta\phi_{max} = \pi/4$ which correspond to highly directional, fast moving traffic (*e.g.,* highway traffic); for the low speed pattern, $V_{max} = 30mile/hr, \Delta\phi_{max} = \pi/2$ which corresponds to a less directional, slow moving traffic (*e.g.,* downtown traffic). The mobility update interval (Δt) is chosen to be $10sec$ throughout this study.

Numerical Results. The corresponding numerical results for two different test traffic scenarios regarding the composition of class 1 and class 2 traffic are presented in the following. Specifically, in scenario 1, 10% of the new call attempts are class 2 calls, while in scenario 2, 50% of the new calls are class 2 calls. Note that in the proposed CAC and resource reconfiguration scheme only class 2 calls can lend bandwidth to handoff calls, and therefore the number of ongoing class 2 calls in a cell will influence the performance of the system.

Figure 2.18 compares the handoff call and new call blocking probabilities of the proposed system with the corresponding results of the conventional system for different new call arrival rates (λ) under test traffic scenario 1, with users moving in the high speed pattern. In the legends of the figures, "c1" and "c2" stand for class 1 and class 2 users respectively, "proposed" indicates that the results are obtained under the proposed integrated system where both position-assisted advanced bandwidth reservation (PMBBR) and resource reconfiguration are implemented, "PMBBR_only" represents a system where only the advanced bandwidth reservation (via the PMBBR scheme) is implemented, while the bandwidth reconfiguration based call admission control component is not implemented, and finally "conventional" corresponds to the case of a conventional system. The conventional system uses the fixed bandwidth reservation, where the reservation value represents a fixed percentage of the total capacity of the cell. In the following study, the corresponding reservation value of the conventional system for class 1 users is $BW_1^{res} = 30Kpbs$ and for class 2 users is $BW_2^{res} = 50Kpbs$. The reservation values are selected based on experimentation with the objective of keeping similar handoff blocking probability for both the proposed system and the conventional system. The call admission control procedure for conventional system is the same as the one proposed in section 2.4.2 except that the bandwidth reconfiguration is not used. It can be observed from the figure that for the given parameters, the proposed system and the conventional system have similar handoff call blocking probabilities. However, the proposed system can significantly decrease the new call blocking probabilities for both user classes, which demonstrates that the proposed system can admit more users than the conventional system while still guarantee the same level of QoS for handoff calls.

The reason is two-fold: first, by using PMBBR algorithm, the bandwidth reservation is only made for those users that will request handoff in the near future and the reservation value can be dynamically adjusted according to the predicted user



(a) Class 1 users

(b) Class 2 users

Figure 2.18 Call blocking probabilities for high mobility pattern under test traffic scenario 1

speed and direction. As can be seen by Figure 2.18, compared with the "conventional" scheme, for both user classes, the "PMBBR_only" scheme achieves to decrease significantly the new call blocking probability at the cost of a slight increase in the handoff call blocking. Then by utilizing the bandwidth reconfiguration based CAC component the "proposed" scheme further improves the performance, by decreasing significantly the handoff call blocking probabilities with no impact on the new call blocking probabilities. This means that the bandwidth reconfiguration component works complementary to the PMBBR scheme and eliminates any potential impact on handoff call blocking probability that may be introduced by the PMBBR scheme. This is because the bandwidth reconfiguration procedure makes some ongoing class 2 users to reduce their current bandwidth usage to spare some bandwidth for the incoming handoff calls, which allows the system to achieve better handoff performance at relatively less reservation values. Less reservation value gives new calls better chance to be granted admission to the system. Therefore, the bandwidth reconfiguration component and the PMBBR scheme work together to improve the system resource utilization and guarantee seamless operation.

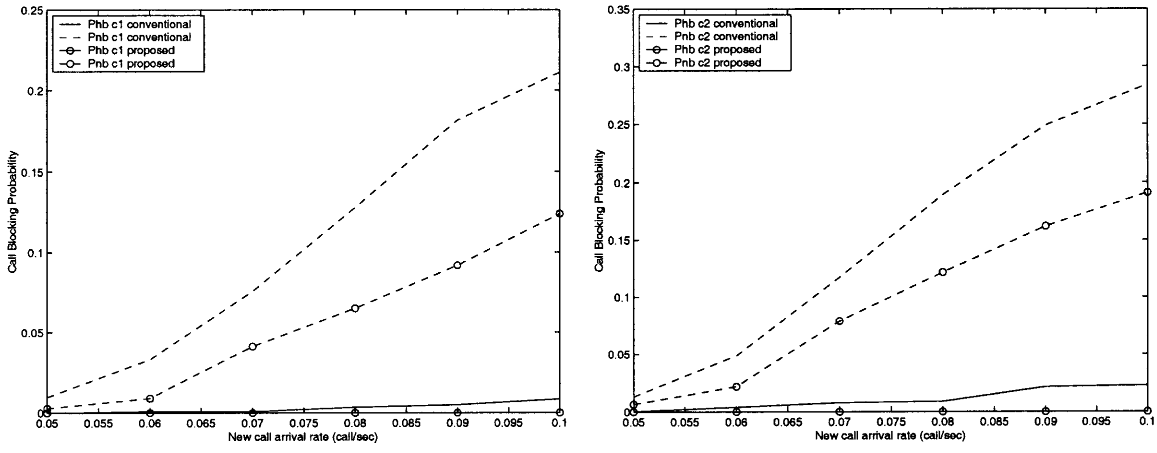
It should be noted here that the connection level QoS improvement achieved by the proposed system is obtained at the cost of slightly decreasing the bandwidth actually used by class 2 users. Table 2.3 lists the average used bandwidth by each class 2 user. For the worst case where new call arrival rate $\lambda = 0.11$, resource reconfiguration makes the average class 2 users' used bandwidth to decrease by only 14%. In other words, the proposed scheme can significantly improve the connection level QoS by slightly degrading the performance of class 2 service calls. This is due to the fact that a class 2 user may lend bandwidth to incoming handoff calls only when a cell is congested and the degraded class 2 call can have a chance to obtain larger bandwidth when it handoff to a less busy cell. This degradation policy (as described in section 2.4.2) can also guarantee that both the number of the degraded class 2 users and the degradation time are minimized.

Table 2.3 Average Bandwidth Used by Each Class 2 User

λ (call/sec)	0.05	0.06	0.07	0.08	0.09	0.1	0.11
proposed (Kbps)	50	49.6	49	48.1	46.85	45.46	42.5
conventional (Kbps)	50	50	49.9	49.77	49.75	49.67	49.42

Figure 2.19 presents the corresponding numerical results for test traffic scenario 2 where 50% of the new calls belong to class 2 traffic. It can be observed that under this traffic configuration the proposed system can achieve to provide an even better performance improvement: the new call blocking probabilities are significantly decreased for both user classes, and the handoff call blocking probabilities are zero in the offered new call arrival rate range from 0.05 to 0.1. As a result, calls for both classes are never forced to termination and the users obtain seamless connections once they are admitted into the system. Since each admitted class 2 call can spare some

bandwidth (up to $45 Kbps$ in the numerical study) to accommodate the handoff calls, if there are more class 2 users in service, the handoff call can always have sufficient bandwidth to borrow from when the reserved bandwidth is not enough to support it.



(a) Class 1 users

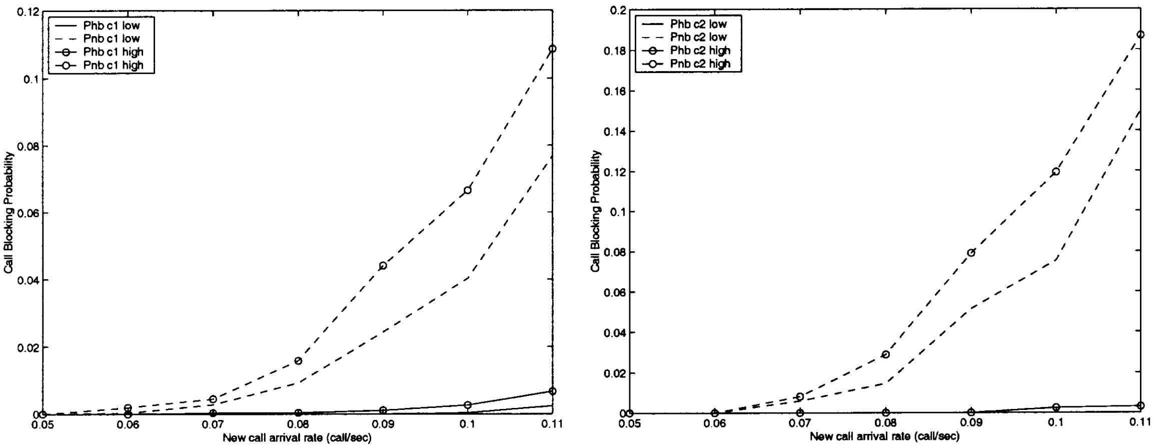
(b) Class 2 users

Figure 2.19 Call blocking probabilities with high mobility pattern under test traffic scenario 2

Figure 2.20 shows the corresponding blocking probabilities that can be achieved by the proposed system for different mobility patterns (high vs. low) under test traffic scenario 1. As can be observed from the figure that the proposed scheme is capable of achieving good performance even when the users move in the low speed pattern: the proposed scheme achieves to provide very low handoff failure rates (in the magnitude of 10^{-4}) for both user classes, which are similar to the results obtained under the high speed mobility pattern. These results indicate that regarding the achieved handoff call blocking probabilities, the proposed system is not very sensitive to the user mobility pattern. Both the PMBBR algorithm and the bandwidth reconfiguration based CAC contribute to such insensitivity. The PMBBR makes bandwidth reservation based on the influence so that the proposed algorithm is capable of adjusting the reserved bandwidth to the appropriate level according to the current traffic conditions.

Moreover, the reservation advance time of the PMBBR controls the reservation timing so that the likelihood of false reservation can be decreased even when users move in the less directional low speed pattern. The bandwidth reconfiguration procedure is concerned with the balance of the bandwidth usage between users and user classes, and its functionality is independent of the user mobility in nature.

It can also be observed from Figure 2.20 that the new call blocking probability for the high speed pattern is higher than that for the low speed pattern. This is due to the fact that for the same new call arrival rate, higher user mobility results in higher handoff arrival rate [37]. Higher handoff traffic load, on one hand requires more bandwidth to be reserved, while on the other hand it requires more class 2 calls to reduce their current bandwidth usage to lend some bandwidth to the incoming handoff calls, which in turn results in less bandwidth released by call completions and handoff-outs. Hence, the bandwidth available to new calls is less and the new call blocking probability is higher for the high speed pattern.



(a) Class 1 users

(b) Class 2 users

Figure 2.20 Call blocking probabilities for different mobility patterns under test traffic scenario 1

CHAPTER 3

NEW CALL BOUNDING SCHEME (NCB)

3.1 New Call Bounding Scheme

As can be observed from literature studying that all guard channel schemes use the number of occupied channels as a decision variable: when this number exceeds some certain threshold, arriving new calls are blocked and only handoff calls are accepted. However, it may well happen that a large number of new calls are accepted into the system, which may result in congestion in neighboring cells due to the handoffs of these new calls in the future. This is the case when calls arrive in bursts.

When congestion occurs, the QoS of both new and handoff calls is impacted. In order to avoid network congestion, a *New Call Bounding* (NCB) scheme is proposed in this chapter. The NCB scheme directly controls the number of admitted new calls and uses the number of channels that are currently occupied by new calls as a decision variable for the CAC. More specifically, the scheme works as follows: when a new call which requires B_{new} channels for service arrives at a cell, the probability that this call will be admitted (P_{new}) is:

$$P_{new} = \begin{cases} 1 & B_{usednew} \leq N_{bnd} \ \& \ B_{used} \leq C - B_{new} \\ 0 & otherwise \end{cases} \tag{3.1}$$

where $B_{usednew}$ and B_{used} are the number of channels that is currently used by new calls and number of busy channels in this cell respectively and N_{bnd} is a given bound for new calls. The idea behind this scheme is that the system would rather accept fewer calls than drop ongoing calls in the future, so it controls the number of accepted new calls directly. The performance of the NCB scheme is investigated through both simulation study and Markov chain analysis in the following sections.

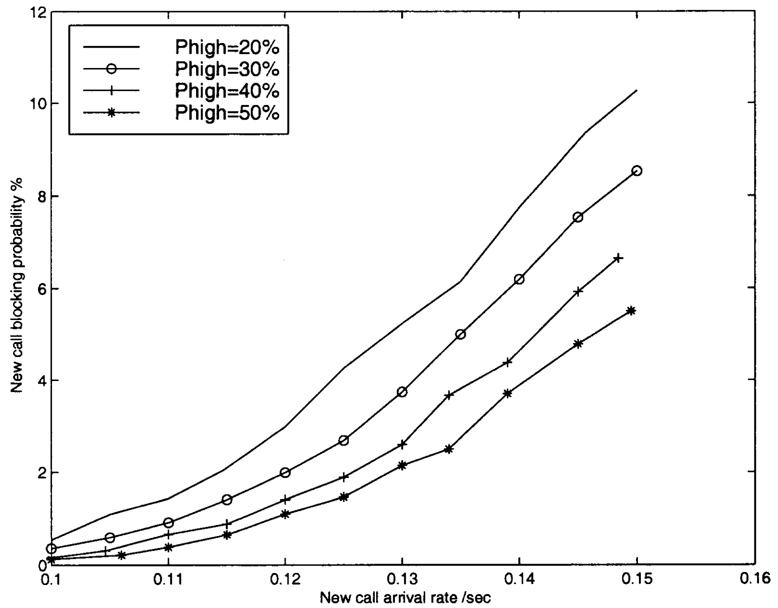


Figure 3.1 New call blocking probability for NCB scheme with different traffic composition: $N_{bnd} = 25$.

3.2 Numerical Study

The performance of the NCB scheme is evaluated in a homogeneous network model. The only difference between the homogeneous model and the heterogeneous model proposed in Section 2.2.1 is that in a homogeneous model the geographic layout of the cells is not assumed. Therefore, users follow a random movement pattern and all the direction factors are 1/6.

Figures 3.1 and 3.2 present the new call blocking probability and handoff call blocking probability of NCB scheme with different new call traffic composition and $N_{bnd} = 25$. As can be seen by these figures, with the decrease of percentage of high speed user in new call traffic, the handoff call blocking probability is decreased and the new call blocking probability is increased. In order to explain this observation, the user *channel holding time* is investigated. Simulation results are collected and listed in Table 3.1, where Tc_{new} and $Tc_{handoff}$ are the average channel holding times of new calls and handoff calls, respectively, and P_{hfhigh} is the percentage of high speed calls in the handoff traffic. The table shows that with the decrease of the percentage

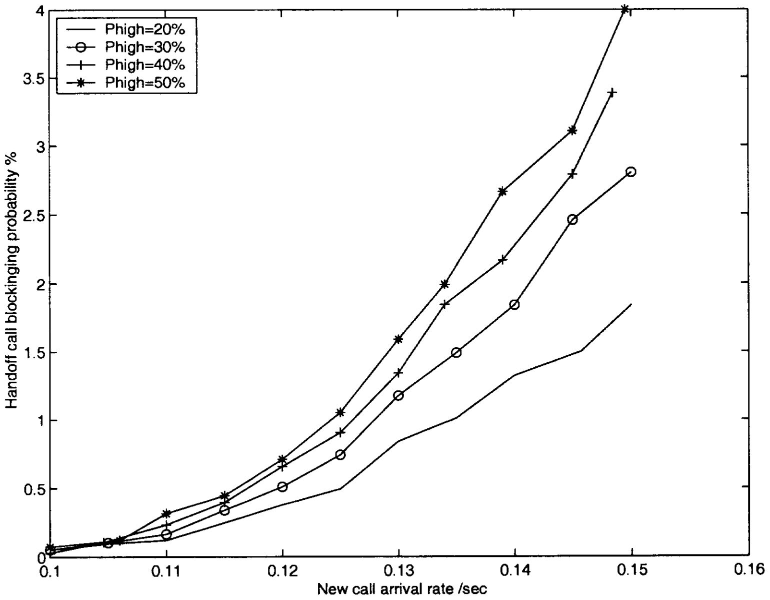


Figure 3.2 Handoff call blocking probability for NCB scheme with different traffic composition: $N_{bnd} = 25$.

Table 3.1 Channel Holding Time for Different Traffic Composition

P_{high}	Tc_{new}	P_{hfhigh}	$Tc_{handoff}$
20%	152.9sec	55%	120.5sec
30%	144.1sec	70%	109.2sec
40%	135.1sec	79%	101.2sec
50%	125.6sec	85%	95.5sec

of high speed users, the average channel holding time of new calls become longer. When a new call is admitted, it tends to occupy the channel for a longer time. Recall that the NCB scheme uses the number of accepted new users as a decision variable, thus when the channel holding time for new calls is longer, the traffic intensity for new calls increases, which leads to the increase of new call blocking probability.

Table 3.1 also shows that the channel holding time of handoff calls is always shorter than that of new calls (same fact is also observed in MBCR study, Section 2.2.2). The reason is that the low speed calls will more likely finish their session in the originating cell, and therefore, the handoff traffic is dominated by high speed calls whose channel holding time is shorter. Since new calls have longer average channel holding time, it is implied that new calls are more reluctant to release channels than handoff calls once they are accepted. As mentioned above, if call arrives in bursts, the *neighboring cells* will be congested due to future handoffs. To make the situation even worse, if a large number of new calls are accepted in a cell, there will be fewer channel available in a relatively longer time and thus *the current cell* is also congested. In order to avoid this situation, it is desirable to limit the number of the admitted new calls. The NCB scheme is suitable for the so called "hot-spot" cells such as subway station and stadium, where calls tend to arrive in bursts.

3.3 Analysis of NCB Scheme

In order to model the system that carries out NCB call admission control scheme, the following assumptions, typical of teletraffic analysis, are made:

- New call attempts and handoff call attempts follow Poisson processes with arrival rates λ_n and λ_h, respectively.
- Channel holding times of new calls and handoff calls are exponentially distributed with mean $1/\mu_n$ and $1/\mu_h$, respectively.
- Let C be the total number of channels in a cell and K be the bound for new calls ($K = N_{bnd}$).

The NCB scheme can be analyzed by using a two-dimensional Markov chain whose transition diagram is shown in Figure 3.3.

The state space of two-dimensional Markov chain is:

$$S = \{(n_1, n_2)|0 \leq n_1 \leq K, n_1 + n_2 \leq C\}$$

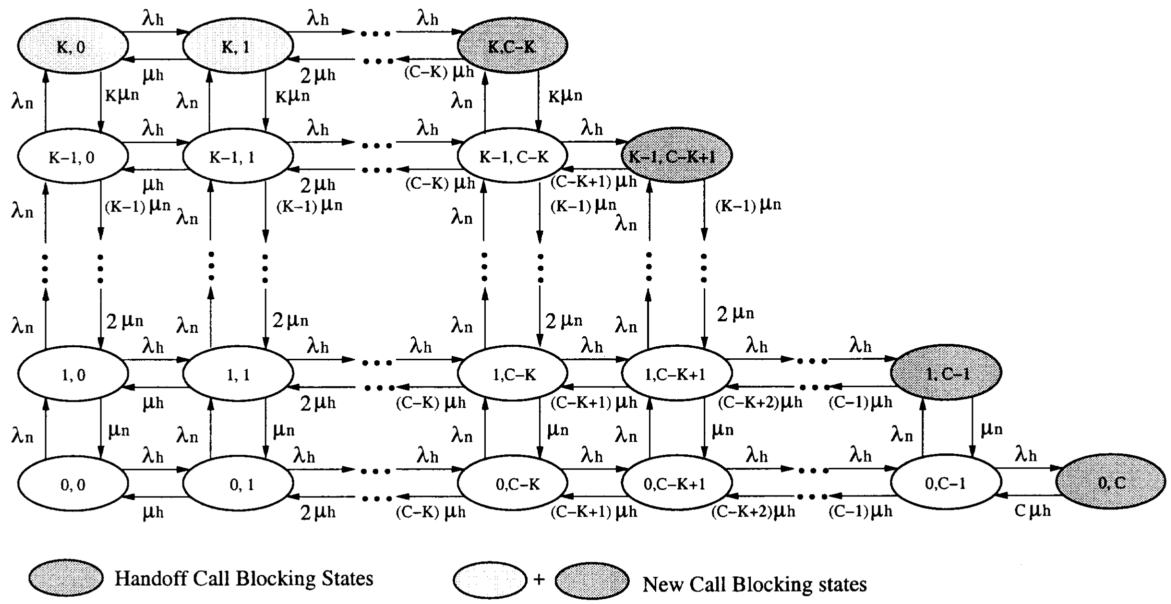


Figure 3.3 State transition diagram for the new call bounding scheme.

where n_1 denotes the number of new calls initiated in the cell and n_2 is the number of admitted handoff calls in the cell. Let $q(n_1, n_2; \bar{n}_1, \bar{n}_2)$ denote the probability transition rate from state (n_1, n_2) to the state $(\bar{n}_1, \bar{n}_2)$, then state-transition equations can be obtained as:

$$\begin{aligned}
q(n_1, n_2; n_1 - 1, n_2) &= n_1\mu_n & (0 < n_1 \le K, 0 \le n_2 \le C) \\
q(n_1, n_2; n_1 + 1, n_2) &= \lambda_n & (0 \le n_1 < K, 0 \le n_2 \le C) \\
q(n_1, n_2; n_1, n_2 - 1) &= n_2\mu_h & (0 \le n_1 \le K, 0 < n_2 \le C) \\
q(n_1, n_2; n_1, n_2 + 1) &= \lambda_h & (0 \le n_1 \le K, 0 \le n_2 < C)
\end{aligned}$$

where (n_1, n_2) is a feasible state in S. Let $p(n_1, n_2)$ denote the steady-state probability that there are n_1 new calls and n_2 handoff calls in the cell. Let $\rho_n = \lambda_n/\mu_n$ and $\rho_h = \lambda_h/\mu_h$. It is easy to verify that:

$$p(n_1, n_2) = \frac{\rho_n^{n_1}}{n_1!} \cdot \frac{\rho_h^{n_2}}{n_2!} \cdot p(0,0), \quad 0 \le n_1 \le K, n_1 + n_2 \le C, n_2 \ge 0 \qquad (3.2)$$

From the normalization equation,

$$p(0,0) = \left[\sum_{0 \le n_1 \le K, n_1 + n_2 \le C} \frac{\rho_n^{n_1}}{n_1!} \cdot \frac{\rho_h^{n_2}}{n_2!} \right]^{-1} = \left[\sum_{n_1=0}^{K} \frac{\rho_n^{n_1}}{n_1!} \sum_{n_2=0}^{C-n_1} \frac{\rho_h^{n_2}}{n_2!} \right]^{-1} \tag{3.3}$$

Based on this, the formulae for new call blocking probability p_{nb} and handoff call blocking probability p_{hb} can be obtained as follows:

$$\begin{aligned} p_{nb} &= \frac{\sum_{n_2=0}^{C-K} \frac{\rho_n^K}{K!} \cdot \frac{\rho_h^{n_2}}{n_2!} + \sum_{n_1=0}^{K-1} \frac{\rho_n^{n_1}}{n_1!} \cdot \frac{\rho_h^{C-n_1}}{(C-n_1)!}}{\sum_{n_1=0}^{K} \frac{\rho_n^{n_1}}{n_1!} \sum_{n_2=0}^{C-n_1} \frac{\rho_h^{n_2}}{n_2!}} \\ p_{hb} &= \frac{\sum_{n_1=0}^{K} \frac{\rho_n^{n_1}}{n_1!} \cdot \frac{\rho_h^{C-n_1}}{(C-n_1)!}}{\sum_{n_1=0}^{K} \frac{\rho_n^{n_1}}{n_1!} \sum_{n_2=0}^{C-n_1} \frac{\rho_h^{n_2}}{n_2!}} \end{aligned} \tag{3.4}$$

Figure 3.4 compares numerical results calculated according to Equation (3.4) with the simulation results. It can be observed that two results match perfectly. This analytical model can be used to determine the appropriate new call bound value (N_{nbd}) for the given system parameters and QoS requirements at the system planning phase.

3.4 Integration of MBCR and NCB

If the NCB scheme is used in combination with Integral MBCR scheme, constraints are imposed on the number of channels occupied by the new calls, as well as on the number of channels occupied and reserved. The new call admission probability for this scheme is as follows:

$$P_{new} = \begin{cases} 1 & B_{used} \le C - \widetilde{R_j} - B_{new} \ \& \ B_{usednew} \le N_{bnd} \\ 0 & otherwise \end{cases} \tag{3.5}$$

Figures 3.5 and 3.6 show the effect of new call bounding. It can be observed that compared with MBCR only, the MBCR with new bound scheme achieves to further decrease the handoff call blocking probability, at the cost of a slight increase in the new call blocking probability. By integrating new call bound with MBCR, the admission

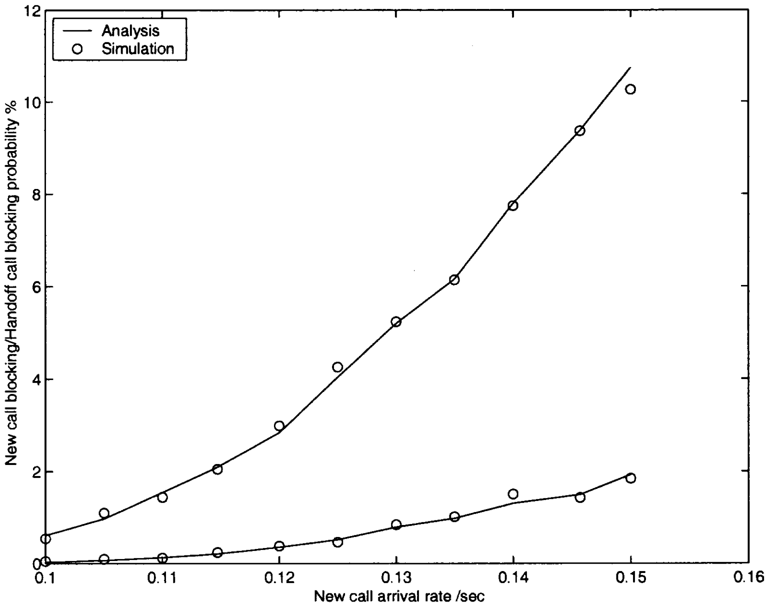


Figure 3.4 Comparison of simulation results and analysis results of NCB scheme: $P_{high} = 0.2$.

policy for new calls becomes stricter: a new call can be admitted only when both of the requirements are met. In this case, more new calls are blocked in order to prevent the congestion, and as a result handoff calls are given even higher priority, which means that once a call is admitted, it can obtain a better service. As observed by Figures 3.5 and 3.6, the lower the new bound, the stricter the limit, hence, the higher the P_{nb} and the lower the P_{hb}.

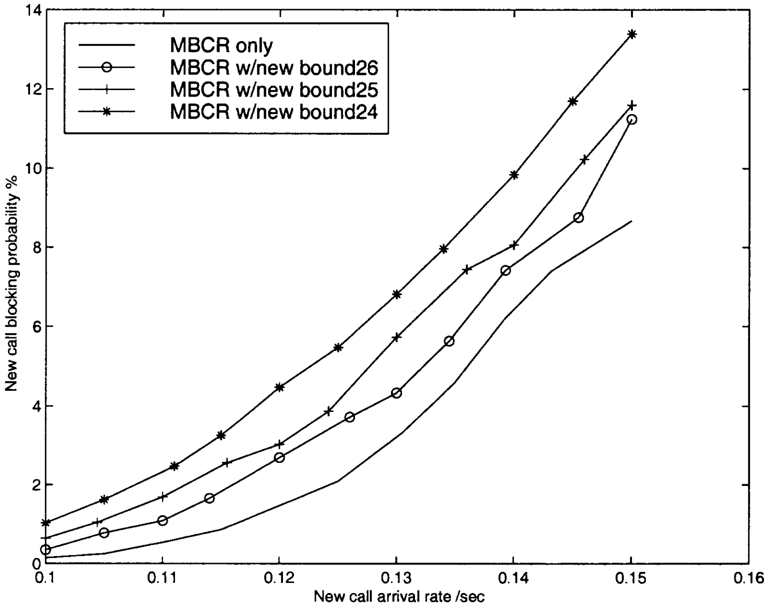


Figure 3.5 New call blocking probability for different new call bounds: $P_{high} = 0.2$.

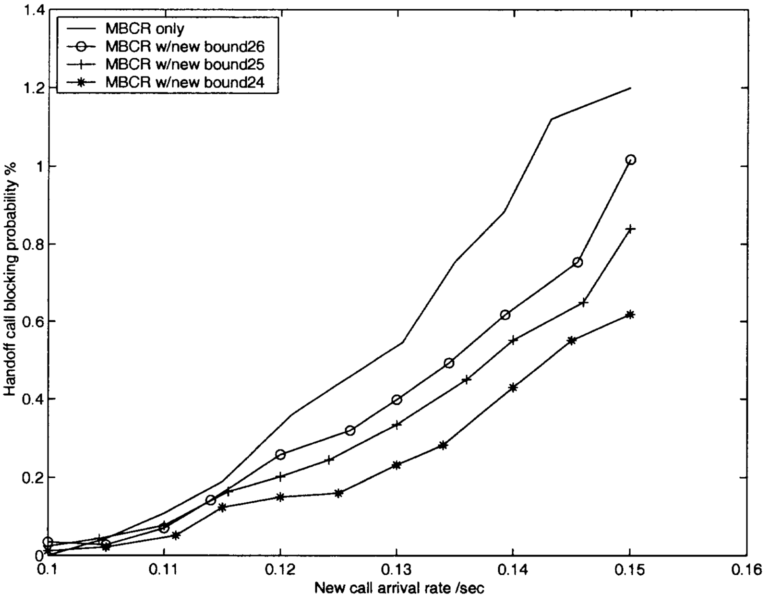


Figure 3.6 Handoff call blocking probability for different new call bounds: $P_{high} = 0.2$.

CHAPTER 4

INTEGRATING PRICING WITH CALL ADMISSION CONTROL

4.1 Introduction

As discussed in Chapter 1, various handoff priority-based RA/CAC schemes have been proposed in literature and can be roughly classified into three categories: Guard Channel Schemes, Queuing Priority Schemes and Channel Borrowing Schemes. These research efforts are mainly focused on how to adjust the tradeoff between new call blocking probability and handoff call blocking probability: decreasing the handoff call blocking probability at the cost of increasing the new call blocking probability of this cell (Guard Channel schemes) or other cells (Channel Borrowing schemes). Within a certain dynamic range of call arrival rate, these schemes can improve the system performance. However, it can also be observed from the results presented by these research efforts that with the increase of call arrival rate, both the new call blocking probability and the handoff call blocking probability increase. When the call arrival rate is temporarily very high (for example in busy hours), no matter how the parameters are adjusted, these schemes cannot guarantee the QoS to users. For example, in order to keep the handoff call blocking probability under a given threshold, a dynamic guard channel CAC scheme must increase the number of guard channels. This will result in a large number of new calls being blocked, which is also a great penalty to the QoS. In this case, it is said that the offered traffic is beyond the capacity of a cell or the cell is overloaded and congested.

In the Queuing Priority Scheme, if there are no available channels at the time of new/handoff call arrival, arriving calls are placed in queue(s) and wait for the currently occupied channels to be released. However, in a practical system the assumption of infinite queue is not realistic: a handoff call may not wait in the queue for long time since the user may move out of the handoff area; due to caller's impatience a new call will not stay in the queue for a long time either [10]. If the overload lasts for a long period, the Queuing Priority Scheme cannot achieve a better performance either.

The main reason of QoS degradation described above stems from the fact that resources in a wireless network, such as time slots, code and power, are shared by all the users. When one user is admitted into the network, it will cause QoS degradation to other users. In terms of Economy, this phenomenon is recognized as *Externality* [82]. In general, it can be observed that the most serious QoS violation (Externality) occurs when the system is congested. However, the current CAC schemes cannot avoid congestion, since they do not provide incentives for users to use the channel resources effectively.

Network users act independently and sometimes "selfishly", without considering the current network traffic conditions. Hence, system overload situations are unavoidable. With the emerging 3G and 4G services, conditions will become even worse since users are allowed to use more bandwidth resources to transmit large volume data or even real-time video [63]. If each user requests the resources that maximize his/her individual level of satisfaction, the total utility of the community will decrease, so that there must be some mechanism to provide incentives for users to behave in ways that improve overall utilization and performance. In commercial networks, this can be most effectively achieved through pricing.

Network pricing has recently been embraced by researchers in the multi-service broadband networks [16, 17, 43] not only as an economic issue and element, covering the infrastructure expenses and operational expenses through charging the end users, but also as a resource management issue. The aggregate traffic load on a wireless network is the result of many users' individual decisions about whether and how to use the network resources, and these decisions are affected by the incentives these users encounter when using the resources of a wireless network. These incentives can take many forms. One of the most important incentives is the monetary incentive [16]–raising the unit price could make some of the users request less resources. This provides another dimension for the design of CAC schemes that can be used in wireless networks as well. In this chapter, pricing is integrated with CAC to address the problem of congestion. The proposed approach combines concepts from network design and engineering with concepts from economics and user behavior to provide an overall call admission strategy that simultaneously: (a) alleviates the network congestion; (b) meets the QoS requirements of users and (c) uses the network resources efficiently.

The above goals are accomplished by integrating in the call admission control process the following concepts and elements: (a) view the wireless network as a public good, which should be efficiently used so that the social welfare can be maximized, (b) implicitly implement a distributed user based mechanism to direct the user in prioritizing his/her call by providing negative incentives according to the current network condition, and therefore, shaping the traffic, and (c) take into account the user behavior and the user demand function.

The remaining of this chapter is organized as follows. In Section 4.2, it is proven that for a given wireless network there exists an optimal new call arrival rate where the total utility of the users can be maximized. In Section 4.3, a detailed description of the model and operation of the proposed system is provided, while Section 4.4 describes how the price is set dynamically according to traffic load. The performance evaluation of the proposed approach and scheme is presented in Section 4.5, while Section 4.6 extends the study and results to the case of multiple service types.

4.2 Optimal Call Arrival Rate to Maximize Total Utility and Meet QoS Requirements

Quality of Service is described in CCITT Recommendation E. 800 as: "The collective effect of service performance which determines the degree of satisfaction of a user of the service." This ties the quality of service to the user's perception of the service. Meeting users' QoS requirements can be more appropriately expressed as maximizing users' level of satisfaction towards the service, which is the ultimate goal of network provisioning. While the traditional network performance metrics, such as call blocking probability and bandwidth usage, fall short in capturing user's perception of application performance, the *Utility Function* provides means to quantify the relations between users experience and network performance metrics. In terms of economics, *utility functions* describe users' level of satisfaction with the perceived Quality of Service [16, 43]; the higher the utility, the more satisfied the users. In general, utility function characterizes how sensitive users are to the changes in QoS. It is sometimes useful to view the utility functions as of money a user is willing to pay for certain QoS. Some utility functions have been suggested in literature in order to model the customer behavior and evaluate the corresponding pricing policies. For example, in [16], Cocchi *et al.* proposed utility functions for four types of applications: electronic mail (Email), file transfer service (FTP), remote login service (Telnet) and real-time packetized voice. For Email applications, it is assumed that utility is a decreasing function of both average delay and the percentage of messages not delivered within a delay bound of five minutes; for remote login, utility decreases as the average packet round-trip time increases. In this chapter, the utility function is defined as function of the new and handoff call blocking probabilities, which represent the main QoS metrics in cellular networks.

This section proposes and proves a theorem which states that there exists a new call arrival rate where the total user utility is maximized, and therefore, the network resources are optimally utilized. A wireless network that carries out Guard Channel CAC scheme is considered; the arrival process of new calls is assumed to be Poisson and the channel holding time is assumed to follow exponential distribution. The parameters of the system are given, including the total number of channels, the number of guard channels, the average new call channel holding time and average handoff call channel holding time, so that the performance of the system depends on the new call arrival rate (λ_n) and handoff call arrival rate (λ_h) [34, 35, 50]. Lin *et al.* also proved in [57] that handoff call arrival rate is a function of new call arrival rate and other system parameters. Therefore, the following study focuses on how the total utility changes with the new call arrival rate. The analysis is based on the following definitions, observations and assumptions.

Definition 1. The average number of admitted users (N) is defined as a function of new call arrival rate, *i.e.*, $N = f(\lambda_n)$. $f(\lambda_n)$ is a differentiable and monotonically increasing continuous function of λ_n with the following properties:

$$0 \leq f(\lambda_n) < C; \qquad f'(\lambda_n) > 0; \qquad f(\lambda_n = 0) = 0; \qquad \lim_{\lambda_n \to \infty} f(\lambda_n) = C \tag{4.1}$$

where C is the total number of channels assigned to this cell.

Definition 2. The Quality of Service metric P_b is defined as a weighted sum of new call blocking probability (P_{nb}) and handoff call blocking probability (P_{hb}):

$$P_b = \alpha P_{nb} + \beta P_{hb} \tag{4.2}$$

where α and β are constants that denote the penalty associated with rejecting new calls and handoff calls respectively, with $\beta > \alpha$ to reflect the higher cost of blocking a handoff call. P_b is also referred to as cost function since it characterizes the cost on QoS when blocking a new call or a handoff call. Since both P_{nb} and P_{hb} are monotonically increasing functions of λ_n, $P_b = g(\lambda_n)$ is also a monotonically increasing function of λ_n. Function $g(\lambda_n)$ has the following properties:

$$0 \leq g(\lambda_n) < 1; \qquad g'(\lambda_n) > 0; \qquad g(\lambda_n = 0) = 0; \qquad \lim_{\lambda_n \to \infty} g(\lambda_n) = 1 \tag{4.3}$$

The general properties and characteristics of the user utility function are described below. As mentioned before, utility function models network users' preference. It is argued here that when the cost function (P_b) increases, users will observe higher call blockings and the level of user satisfaction decreases. Please also note that when P_b is small, the satisfaction degradation caused by the increase of P_b is not significant; as P_b becomes large, the satisfaction degradation will be substantial, which is referred to as the diminishing margin property [73]. Therefore, the following assumption is made:

Assumption 1. The utility function of a single user (U_s) is a differentiable and monotonically decreasing concave function of the QoS parameter P_b. Therefore, $U_s = h(P_b)$ where function $h(P_b)$ has the following properties::

$$h(P_b) \geq 0; \qquad h'(P_b) < 0; \qquad h''(P_b) < 0 \tag{4.4}$$

Note that U_s achieves maximum value at $P_b = 0$, which means that if the blocking probability is 0% the user has the highest level of satisfaction, therefore, $U_s(P_b = 0) = U_s^{max}$. Moreover, although different applications may have different QoS requirements, and therefore, different utility functions, without loss of generality, it can be assumed that there exists a P_b^{max} such that $U_s(P_b) = 0$ for all $P_b \geq P_b^{max}$. This means that when call blocking probability is very high, the user satisfaction is zero. In a realistic wireless system, P_b^{max} represents the threshold value (maximum) of P_b that can be tolerated so that the Quality of Service is considered acceptable. Based on the above definitions and assumptions, the following theorem is proven:

Theorem 1. For a given wireless network, there exists an optimal new call arrival rate λ_n^* that maximizes the total utility U, where U is defined as:

$$\begin{aligned} U &= N \times U_s \\ &= f(\lambda_n) \times h(P_b) \\ &= f(\lambda_n) \times h[g(\lambda_n)] \end{aligned}$$

Proof. **(i)** From the properties of function $g(\lambda_n)$ and $h(P_b)$,

$$\frac{dU_s}{d\lambda_n} = h'[g(\lambda_n)]g'(\lambda_n) < 0 \tag{4.5}$$

which means that U_s is a continuous and monotonically decreasing function of λ_n, with $U_s(\lambda_n) \geq 0$, $U_s(\lambda_n = 0) = U_s^{max}$ and $U_s(\lambda_n = \lambda_n^{max}) = 0$, where λ_n^{max} is the new call arrival rate that satisfies $g(\lambda_n^{max}) = P_b^{max}$.

(ii) Since both N and U_s are continuous functions over the closed interval $\lambda_n \in [0, \lambda_n^{max}]$, U, as a product of N and U_s, is also a continuous function over the same interval. According to Extreme Value Theorem[1][29], it can be concluded that function U has a minimum and a maximum value on that interval.

- From (i) and Equation (4.1), the minimum value is obtained at the endpoints of the closed interval. Specifically, total utility U is zero either when $\lambda_n = 0$ (*i.e.*, $U(\lambda_n = 0) = 0$) which corresponds to the case that no user is in the system or when $\lambda_n = \lambda_n^{max}$ (*i.e.*, $U(\lambda_n = \lambda_n^{max}) = 0$) which corresponds to the case that single user's utility becomes zero.

- Assuming the maximum value of total user utility to be U_{ub}, and by Extreme Value Theorem it can be concluded that there exist **at least one** λ_n value, denoted by λ_n^0, over interval $(0, \lambda_n^{max})$ that makes

$$U(\lambda_n = \lambda_n^0) = U_{ub} \tag{4.6}$$

[1] **Extreme Value Theorem:** Every function that is continuous over a closed interval has a maximum and a minimum on that interval.

(iii) The number of different values of λ_n^0 depends on the second order derivatives of functions f, g and h.

- If there exists only one λ_n^0 that satisfies Equation (4.6), the optimal new call arrival rate for the system is $\lambda_n^* = \lambda_n^0$;
- If there exist more that one different values of λ_n that satisfy Equation (4.6), the optimal new call arrival rate is set to be: $\lambda_n^* = \sup\{\lambda_n^0 \mid U(\lambda_n^0) = U_{ub}\}$ which is the highest new call arrival rate value that can maximize the total user utility.

□

Maximum total user utility also means that channel resources are most efficiently utilized. When $\lambda_n < \lambda_n^*$, users can get a better quality than their QoS requirements, but some channel resources are wasted and from the perspective of the service provider, this means less revenue. On the other hand, when $\lambda_n > \lambda_n^*$, a large number of users are blocked when trying to initiate their calls or when trying to handoff to another cell in the middle of a call, which means that the QoS degrades and may become unacceptable. In this case, although on average, more channels are used, due to the increasing handoff failures, it is hard for a user to finish his/her call successfully and as a result the "effective" utilization of channel resources is low. Therefore, $\lambda_n = \lambda_n^*$ is a point where the number of satisfied users is maximized and channel resources are most efficiently used. When $\lambda_n > \lambda_n^*$, both the total user utility and the QoS decrease with any further increase of λ_n and the cell enters the congestion state. From the view point of QoS guarantee, it is ideal for a system to operate at the optimal traffic load (λ_n^*) or below.

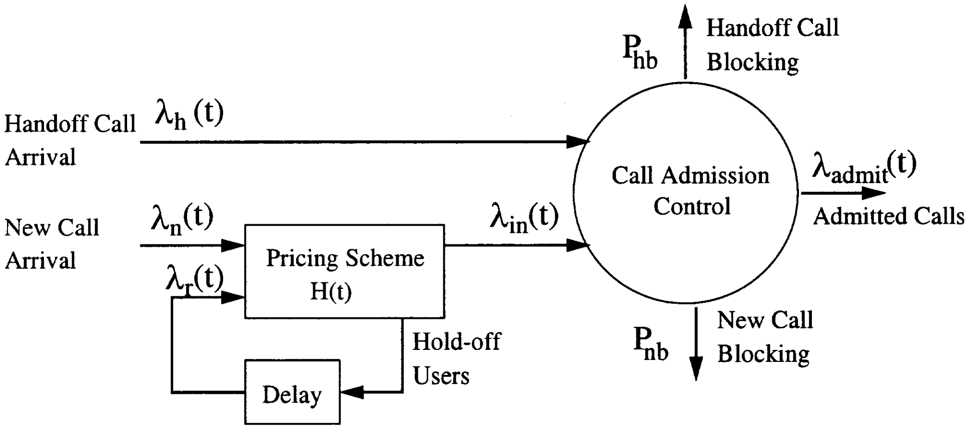


Figure 4.1 Integration of pricing scheme with call admission control.

4.3 System Model

Current wireless networks use flat pricing schemes: users are charged with fixed rate or based on the time of the day. The major advantage of these schemes is that the billing and accounting processes are simple. However, the price is independent of the current state of the network, or any dependence is fixed and is based on decisions that have been made during the planning phase of the system and may not correspond to the actual system conditions. Hence, such systems cannot provide enough incentives for users to avoid congestion, and furthermore, cannot react effectively to the dynamic and sometimes unpredictable variation of the network usage and conditions. This chapter proposes a new scheme which integrates the congestion pricing with call admission control to address this problem. Figure 4.1 provides a schematic representation of the proposed approach and model.

The system is composed of two functional blocks: CAC block and Pricing block. Here, a guard channel scheme is used at the CAC block. The pricing block works as follows: when the traffic load is less than the optimal value, $\lambda_n < \lambda_n^*$, a *normal price* is charged to each user. The normal price is the price that is acceptable to every user. When the traffic load increases beyond the optimal value, dynamic *peak hour price* will be charged to users who want to place their calls at this time. Base stations broadcast the current unit price to users when they try to place calls. It should be noted here that according to the proposed scheme the decision about the peak hour price is based on the network conditions and not only on the time of the day. This means that the price is continuously and dynamically adjusted according to the changes in the system condition as the system evolves. The following points should also be noted in Figure 4.1:

- The handoff call arrival rate λ_h is determined by the new call arrival rate λ_n and other system parameters. The handoff calls do not go through the pricing block, since they are continuation of previously admitted calls and their operation is governed by the price agreed at the time of the call acceptance (*i.e.*, price does not change during the operation of the call);

- During the period that dynamic peak hour price is charged to users, if some users are not willing to accept the extra charge, they will choose not to place their calls at this time. These users can make their calls later when the network conditions change and the price decreases. This generates another traffic stream to the pricing block–the retry traffic, whose arrival rate is denoted by λ_r in Figure 4.1. Different mechanisms that model the retry traffic can be implemented depending on the behavior of users that retry to place their calls. In Figure 4.1, this behavior is represented by the block labeled "Delay".

- Since the traffic load varies according to the time of the day, the above used traffic rates (λ_n, λ_h, λ_r, λ_{in} and λ_{admit}) are all functions of time t.

The system function of pricing block ($H(t)$) is defined as the percentage of the incoming users that will accept the price at time t, *i.e.*,

$$(\lambda_n(t) + \lambda_r(t))H(t) = \lambda_{in}(t) \tag{4.7}$$

where $\lambda_{in}(t)$ is the rate of input traffic to CAC block. The congestion pricing block should be designed in such a way that by adjusting $H(t)$ according to current traffic condition, $\lambda_{in}(t)$ always meets the following requirement:

$$\lambda_{in}(t) \leq \lambda_n^* \tag{4.8}$$

where λ_n^* is the optimal new call arrival rate obtained in Section 4.2. This constraint guarantees that the cell will not be congested, and therefore, the quality of service requirements of the callers in service can be guaranteed.

4.4 Resource Pricing According to Traffic Load

As mentioned before, monetary incentive can influence the way that users use resources and is usually characterized by demand functions. Demand function describes the reaction of users to the change of price. Different demand functions have been proposed in literature [17, 27]. In this chapter, the following demand function [27] is used.

$$D[p(t)] = e^{-(\frac{p(t)}{p_0}-1)^2} \qquad p(t) \geq p_0 \tag{4.9}$$

where p_0 is the normal price, $p(t)$ is the price charged to users at time t which is the sum of normal price and extra peak hour price (if applicable). $D[p(t)]$ denotes the percentage of users that will accept this price. Note that $D(p_0) = 1$, which means that the normal price is acceptable to all users. In the proposed scheme, $p(t)$ is determined by current traffic load $\lambda_n(t) + \lambda_r(t)$. The control function of $H(t)$ is realized by users' reaction to the current price, therefore:

$$H(t) = D[p(t)] \tag{4.10}$$

Combining Equations (4.7) and (4.8),

$$H(t) = \frac{\lambda_{in}(t)}{\lambda_n(t) + \lambda_r(t)} \leq \frac{\lambda_n^*}{\lambda_n(t) + \lambda_r(t)} \tag{4.11}$$

Taking into account Equation (4.9) and (4.10), in order to obtain the desired QoS, at time t the price should be set as :

$$p(t) = D^{-1}\left(\min\left(\frac{\lambda_n^*}{\lambda_n(t)+\lambda_r(t)}, 1\right)\right) \tag{4.12}$$

4.5 Performance Analysis

In this section, the performance of the proposed integrated pricing and call admission control is evaluated in terms of congestion prevention, achievable total user utility and obtained revenue. The achieved performance of different variations of the proposed integrated approach (the variations model the potential behavior of users with regard to pricing and call blocking) is compared with the corresponding results of conventional systems where pricing is not taken into consideration in the call admission control process. It is observed that the proposed integrated scheme achieves to alleviate the system congestion occurrences and meet the QoS requirements of the users in service, while other conventional CAC schemes fail to do so. Moreover, considerable improvements are also observed on both the achieved total user utility and the obtained revenue.

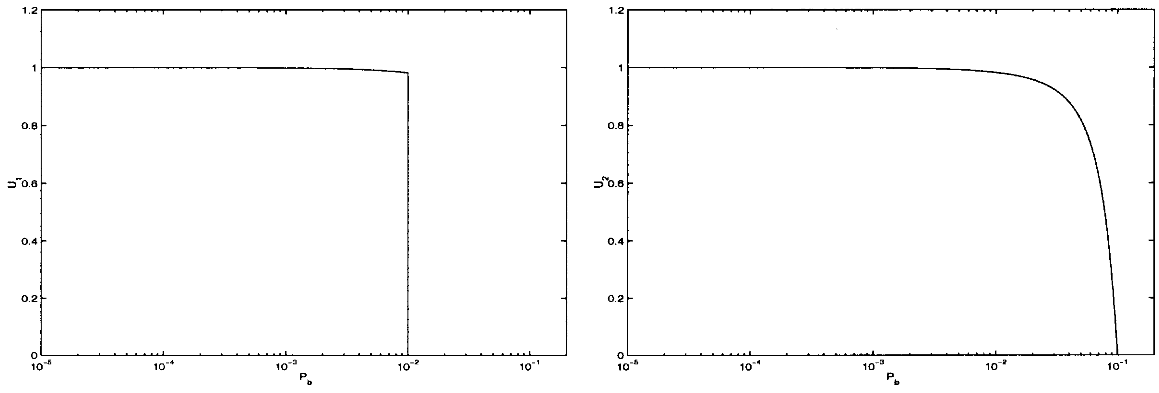
Section 4.5.1 introduces specific utility functions that used throughout the study. In Section 4.5.2, the basic assumptions about the system under consideration as well as the specific experiments performed are described in detail. Section 4.5.3 provides the corresponding results of the different settings compared.

4.5.1 Utility Functions vs. "Elastic"/ "Inelastic" QoS Requirements

In Section 4.2, instead of suggesting an explicit utility function, only a qualitative description of the properties of user utility function was presented. Exact user utility functions can be obtained through field tests and user surveys. Many wireless operators and researchers have conducted sophisticated surveys to find the utility function for various QoS metrics [45]. In the remaining, without loss of generality, two user utility functions U_1 and U_2 are assumed, as defined in Equations (4.13) and (4.14) respectively.

$$U_1 = \begin{cases} 1 - e^{30(P_b - 0.1)} & \text{when } 0 \leq P_b \leq 0.01 \\ 0 & \text{when } P_b > 0.01 \end{cases} \tag{4.13}$$

$$U_2 = \max(1 - e^{30(P_b - 0.1)}, 0) \tag{4.14}$$



(a) U_1.

(b) U_2.

Figure 4.2 User utility functions: U_1 and U_2.

A schematic representation of these two utility functions is shown in Figure 4.2. These two utility functions differ in their corresponding QoS requirements. The first utility function (U_1) describes the case that users have "hard" QoS requirement, which means that users will not accept the service when QoS is worse than a pre-specified threshold. From Equation (4.13) and Figure 4.2(a), it can be observed that when P_b is more than 1%, the utility becomes zero. This is also referred to as "inelastic" QoS requirement [74]. The second utility function (U_2) describes the situation that users' satisfaction gradually decreases with the increase of blocking probability, which means that users can tolerate some degree of service outage. U_2 corresponds to a soft or "elastic" QoS requirement [74].

4.5.2 Model and Assumptions

The parameters used throughout the performance evaluation are as follows:

(1) Each cell is assigned $C = 40$ channels, and 2 of them are used as guard channels.

(2) Each call requires only one channel for service.

(3) For both new calls and handoff calls, the call duration times are exponentially distributed with mean $(\frac{1}{\mu})$ 240 seconds, and the cell dwell times are also exponentially distributed with mean $(\frac{1}{\eta})$ 120 seconds. Although the price changes dynamically, it is assumed here that the call duration is independent of the price [76]. Through congestion pricing experiments, Shih *et al.* found that users do not terminate their calls earlier with the increase of price, "because users do not know how long time the price increase will last". Moreover, as mentioned before, once a call has been accepted in the system its operation is governed by the price agreed at the time of call acceptance (*i.e.,* the price of a specific call does not change during the operation of the call).

(4) The arrival of new calls initiating in each cell forms a Poisson process with rate $\lambda_n(t)$. The variation of new call arrival rate during a 24-hour period used throughout this study is shown in Figure 4.3.

(5) Parameters α and β in Equation (4.2) are set to be $\frac{1}{3}$ and $\frac{2}{3}$ respectively, which means that handoff calls are treated twice more important than new calls.

(6) In the following numerical study, $U_s = U_1$ (defined in Section 4.5.1). The optimal new call arrival rate for this system is calculated to be $\lambda_n^* = 0.12$ call/sec. Based on the analysis in Section 4.2, this is the optimal operation

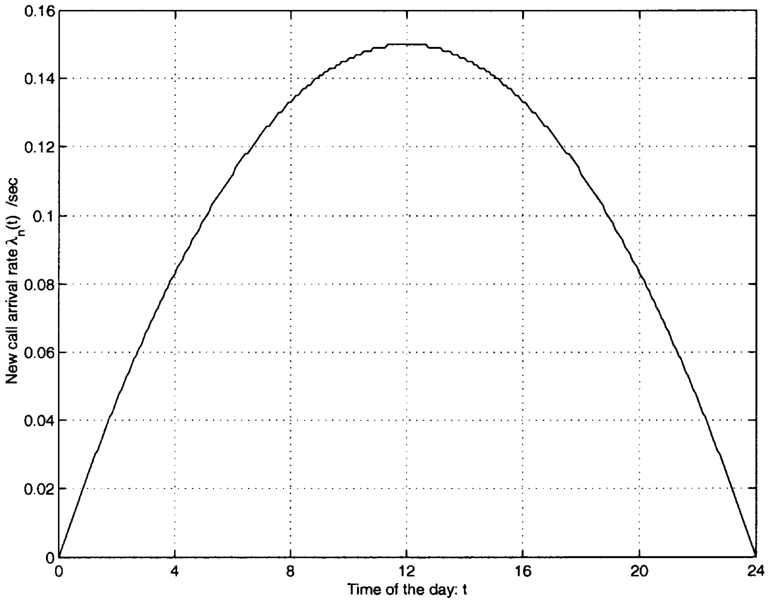


Figure 4.3 Input new call arrival rate as function of time.

point of the system in the sense that at this point, the total user utility can be maximized given that the hard QoS requirement ($P_b \leq 0.01$) is met.

In Equation (4.12), the traffic load input into the pricing block ($\lambda_n(t) + \lambda_r(t)$) is used to calculate the price. However, in a real system it is very difficult, if not impossible, to obtain traffic load in a realtime fashion. In this study, the estimated traffic load is used to calculate the price. The 24-hour period is divided into 5-minute sections. At the end of each section, the average offered traffic load ($\lambda_n(t) + \lambda_r(t)$) during this section is calculated and the price is determined using Equation (4.12). This price will be used in the next 5-minute section.

In order to study and observe how the proposed scheme can solve the problem of congestion in wireless networks, five experiments are performed. The first two experiment setups correspond to conventional systems that do not implement dynamic pricing mechanism while the last three experiments are variations of the proposed scheme which integrates pricing and call admission control. The specific setting for each experiment is as follows:

No. 1: No pricing block is implemented. Users blocked by CAC (blocked users) retry after waiting some time. This is referred to as Conventional System with Retry (CSwR).

No. 2: No pricing block is implemented. One third of the blocked users leave the system and the rest wait and retry. This is referred to as Conventional System with Retry and Loss (CSwRL).

No. 3: A user that does not accept the current price (hold-off users) waits for some time and retries, while blocked users do not retry and they are cleared from the system. This is the scenario implied by Figure 2.1. This is referred to as Pricing System with Hold-off Retry (PSwHR).

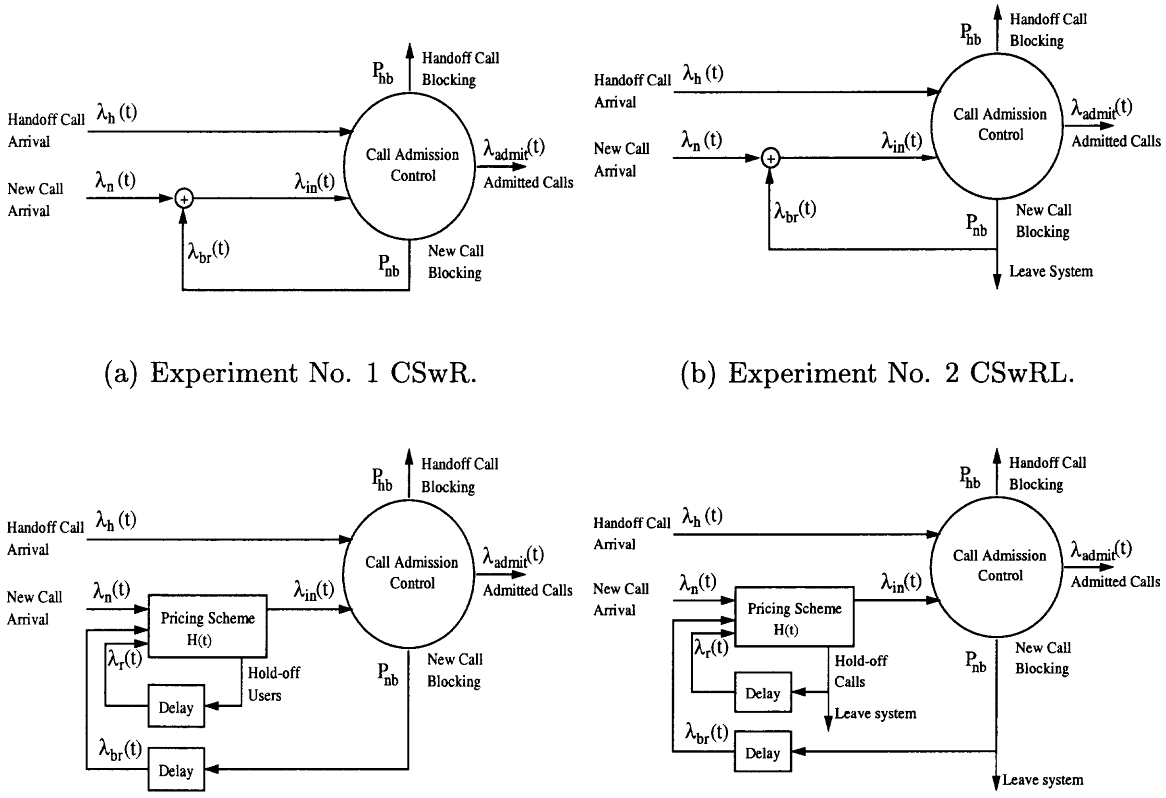
No. 4: Both hold-off users and blocked users retry after waiting some time. This is referred to as Pricing System with Retry (PSwR).

No. 5: One third of the hold-off users and one third of blocked users leave the system, and the rest hold-off and blocked users will wait some time and retry. This is referred to as Pricing System with Retry and Loss (PSwRL).

The corresponding diagrams of CSwR, CSwRL, PSwR and PSwRL (experiments 1, 2, 4 and 5) are shown in Figure 4.4. They all take the blocked call retry into consideration, hence, there is a new traffic stream input to the pricing block: the blocked retry traffic whose rate is denoted by $\lambda_{br}(t)$. The corresponding diagram of PSwHR (experiment 3) is shown in Figure 4.1. For the comparison purposes, all the five experiments use the same fixed guard channel scheme (2 guard channels) in the "Call Admission Control" block.

4.5.3 Numerical Results and Discussion

Congestion Prevention. Figures 4.5 and 4.6 show the results of conventional systems (CSwR and CSwRL) that do not use pricing in the call admission control



(a) Experiment No. 1 CSwR.

(b) Experiment No. 2 CSwRL.

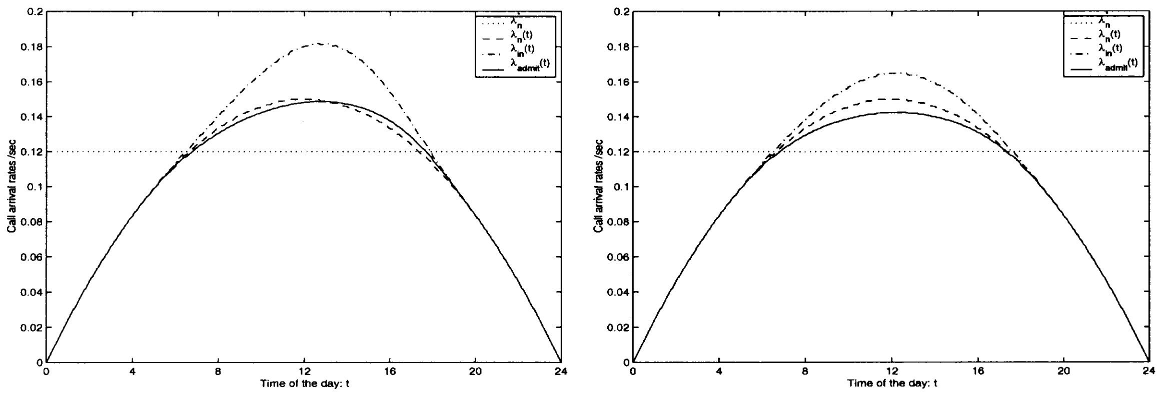
(c) Experiment No. 4 PSwR.

(d) Experiment No. 5 PSwRL.

Figure 4.4 The diagrams of experiments No.1, No.2, No.4 and No.5.

process to control the traffic. Figure 4.5(a) and (b) show the traffic rates (vertical axis) at different points of the system for experiments CSwR and CSwRL, respectively (horizontal axis corresponds to the 24-hour period). Specifically, four curves are presented in each of these figures. The optimal new call arrival rate is denoted by λ_n^*; the new call arrival rate is denoted by $\lambda_n(t)$; the rate of traffic input to CAC block is denoted by $\lambda_{in}(t)$ and the traffic rate of calls that are admitted into the system is denoted by $\lambda_{admit}(t)$. It can be first observed that for the given system parameters and input traffic pattern, the system works within a wide range of channel loads, from a light load at midnight hours to heavy load at noon hours (the highest λ_{in}

value for experiment CSwR is 0.182 *call/sec* which corresponds to utilization factor[2] $\gamma = 84.3\%$). It is also observed from this figure that for both experiments, $\lambda_{in}(t)$ exceeds the optimal operation value in noon hours. From Figure 4.6, it can be seen that when traffic load is heavy (*e.g.,* in noon hours), P_b can be as high as 8% (for experiment CSwR) and 6% (for experiment CSwRL). These values are far beyond users' minimum QoS requirement 1%, and therefore, it can be concluded that cells are seriously congested during this period. As a result the QoS of on-going calls is also affected significantly.



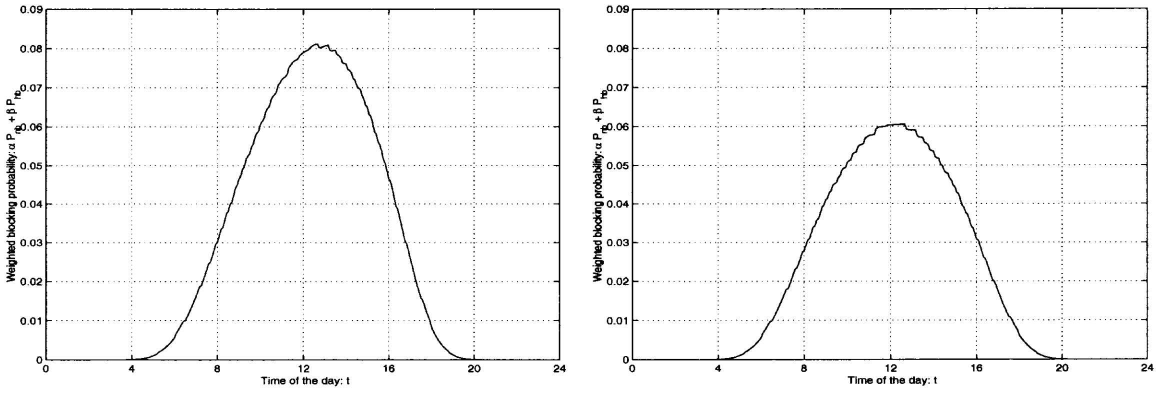
(a) CSwR.

(b) CSwRL.

Figure 4.5 Experiments CSwR and CSwRL: Traffic rates at different points.

The corresponding results of experiments PSwHR, PSwR and PSwRL are shown in Figures 4.7 through 4.9. All these three experiments use pricing to prevent congestion occurrence. The difference among these three schemes is that users have different behavior modes. When encountered with unacceptable price or denial of channel access request, in experiment PSwHR, all hold-off users (users that did not accept the current price) may choose to retry while the blocked users are cleared from

[2] The utilization factor is calculated as follows: $\gamma = \frac{\lambda_{in}(1-P_{nb})+\lambda_h(1-P_{hb})}{C(\mu+\eta)}$. Note that this is different from the overall offered load (traffic intensity) given by $\rho = \frac{\lambda_{in}+\lambda_h}{C(\mu+\eta)}$, which is even higher.



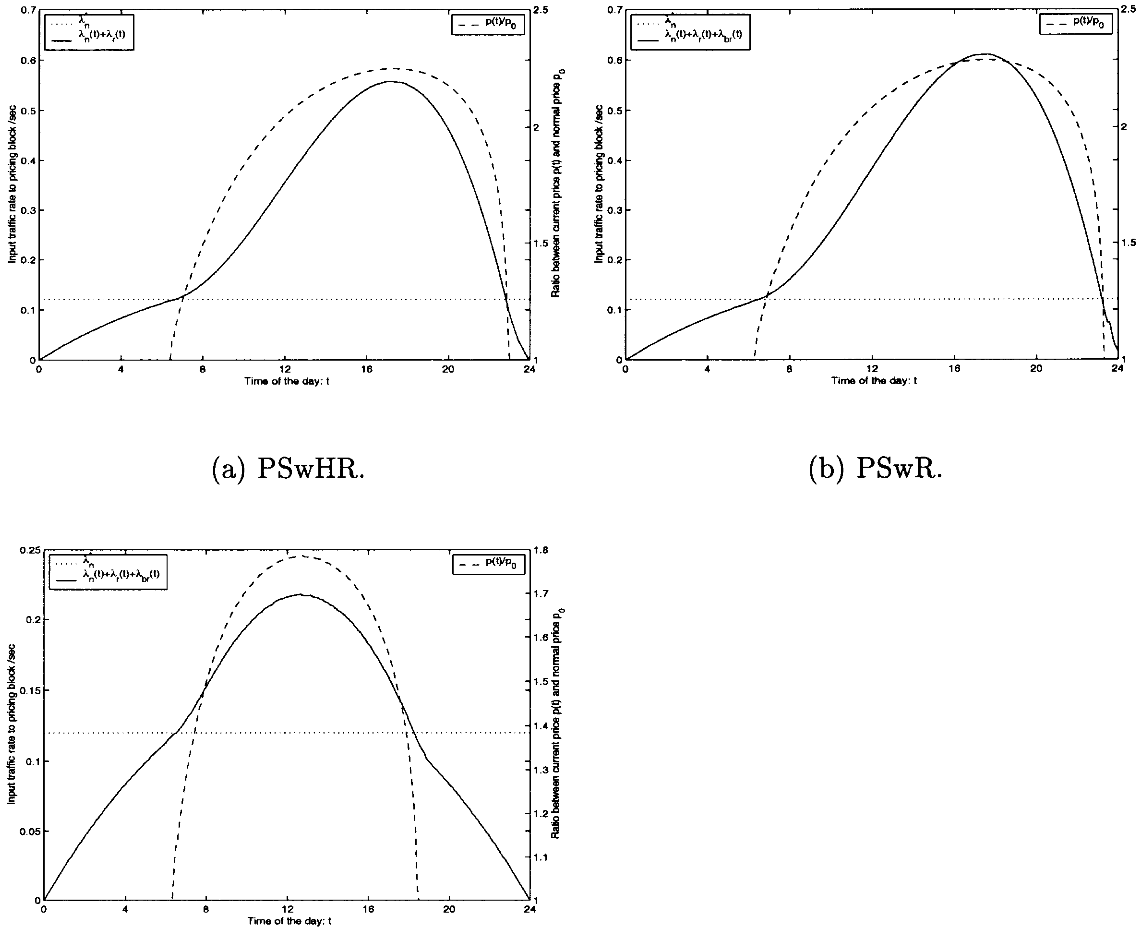
(a) CSwR.

(b) CSwRL.

Figure 4.6 Experiments CSwR and CSwRL: Weighted call blocking probabilities.

the system; in experiment PSwR, all of the hold-off and blocked users may choose to retry; in experiment PSwRL, part of the hold-off and blocked users may leave the system due to caller's impatience and the rest choose to retry. In general, these three experiments model and represent the various user reactions to pricing and CAC.

Figure 4.7 also shows how the price is adjusted according to the change of offered traffic load. For the given new call arrival rate variation, when the offered traffic load into the pricing block is more than the optimal new call arrival rate (6:30AM to 11:00PM in Figure 4.7(a)), the ratio $p(t)/p_0$ becomes more than 1, which means that the pricing mechanism comes into effect and the peak hour prices are charged to users. The heavier the traffic load, the higher the price, so that the less the percentage of users that would like to access the network, as suggested by Equation (4.12). In the proposed scheme, there is no central control mechanism to determine which user can access the channel resources. Each base station just sets the price according to current traffic load of the cell, and it is the individual user's decision on whether to accept this price or not that controls the input traffic load to the system at this time. This implicitly implements a distributed user-based prioritization scheme where the priority is set by user's reaction to current price.



(a) PSwHR.

(b) PSwR.

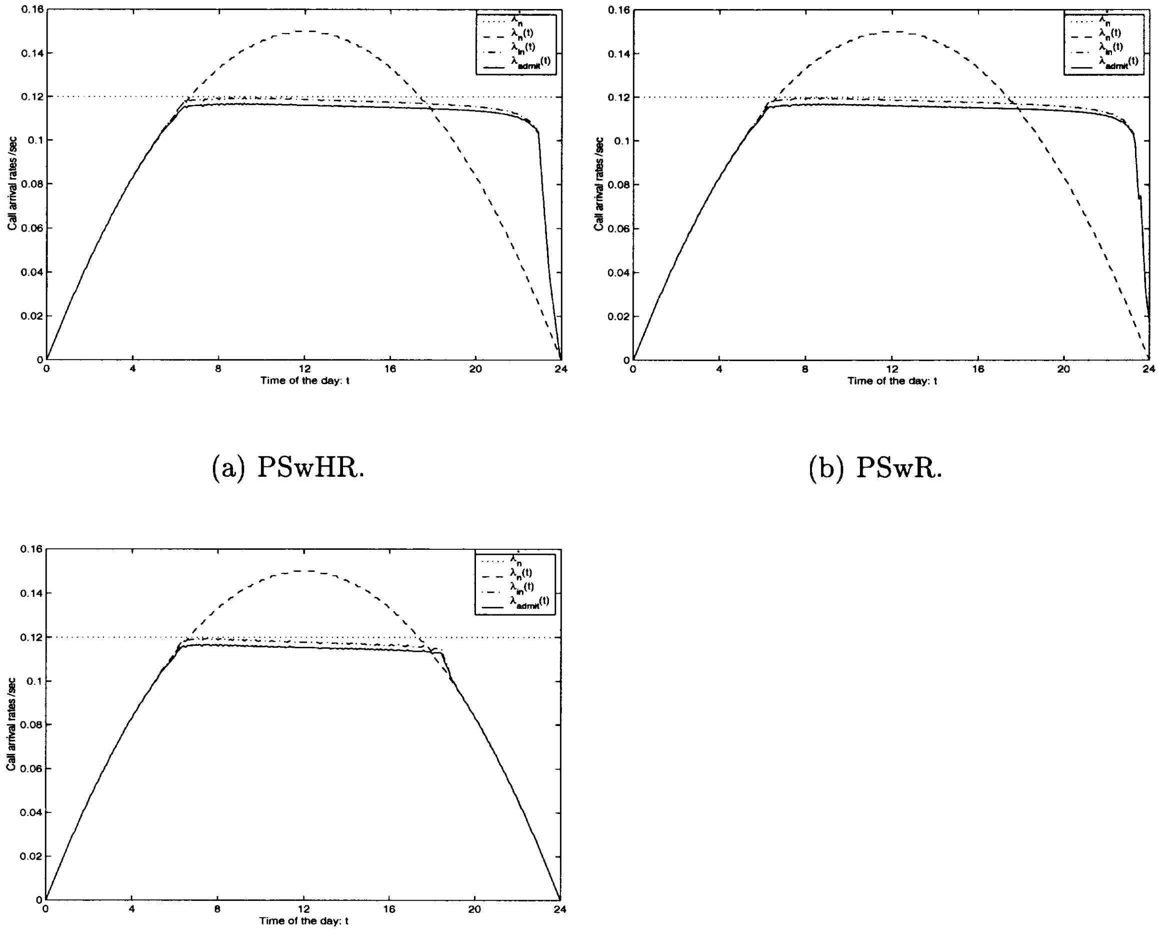
(c) PSwRL.

Figure 4.7 Experiments PSwHR, PSwR and PSwRL: Setting price according to traffic conditions.

Figure 4.8 shows the traffic rates at different points of the system for experiments PSwHR, PSwR and PSwRL. From this figure, it can be first observed that no matter how the hold-off and blocked users behave, the inputs to the CAC block are always lower than the optimal rate, *i.e.*, $\lambda_{in}(t) < \lambda_n^*$, which means that the cell is not congested. The reason is that the price is adjusted based on the user demand function and current traffic load (Equation (4.12)) so that the price is always set to the appropriate value to guarantee that the traffic rate going through the pricing block is less than the optimal value. This result is justified by Figure 4.9. From this figure, it is observed that the weighted call blocking probabilities are always lower than 1%, which means that the QoS of the users who accept the current price can be guaranteed. Comparing Figure 4.8 with Figure 4.5, it can be observed that the pricing block works like a traffic shaper, which can "move" part of the peak hour (6:30AM to 6:00PM in Figure 4.8(a)) traffic to relatively idle hours that follow the peak (6:00PM to 11:00PM in Figure 4.8(a)). The traffic being "moved" is composed of users that do not accept the peak hour price, however, they are willing to retry to place their calls at a later time. Therefore, the traffic load input to the CAC block presents a "flat" shape during the peak hours and following few hours, and during that "flat" period the system achieves maximum total utility.

From Figure 4.8, it is also observed that the shaped traffics ($\lambda_{in}(t)$) for PSwHR, PSwR and PSwRL are different. PSwR has the longest "flat" period, PSwHR has a slightly shorter one, and PSwRL's "flat" period is much shorter than the previous two. This difference is due to different user behavior modes in these three experiments. Compared with PSwR, PSwHR clears blocked users from the system. Since call blocking probability is limited to be less than 1%, whether blocked calls will retry or be cleared from the system only slightly affect the output of the experiments. Compared with PSwR, in PSwRL one third of the hold-off and blocked users will leave the system. In peak hours, in order to guarantee the QoS of users in service, high congestion prices may be charged to users ($p(t)$ can be as high as $1.8p_0$ in Figure 4.7(c)), therefore, the percentage of users that do not accept the price can be high (up to 46%). Since a large number of users leave the system during peak hours, the traffic volume that can be "moved" is relatively less which further results in a much shorter "flat" period for PSwRL than that for PSwR.

As mentioned before, the results in Figures 4.7 through 4.9 have been obtained for the utility function U_1 (defined in Section 4.5.1) that corresponds to the case of hard QoS requirements. If users have soft QoS requirements (*e.g.,* use of utility



(a) PSwHR.

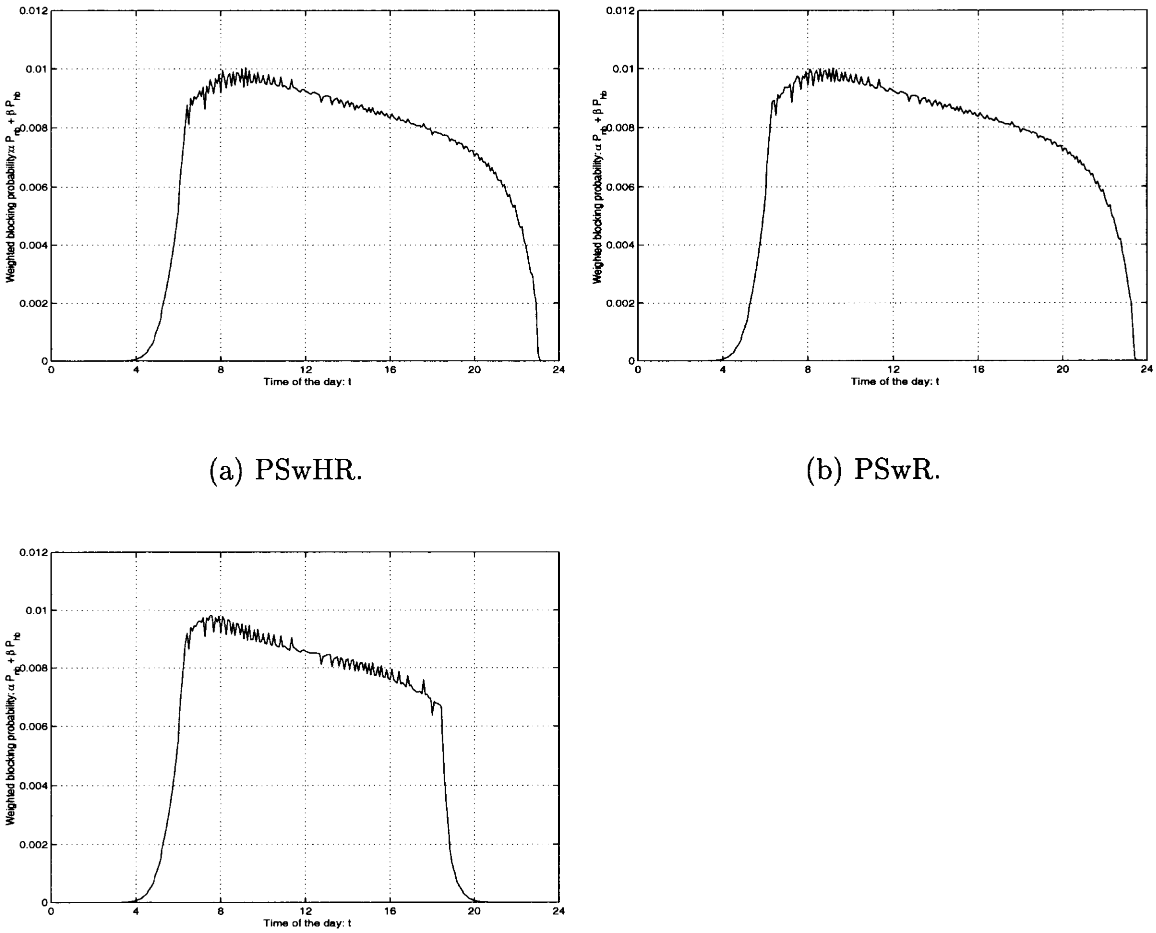
(b) PSwR.

(c) PSwRL.

Figure 4.8 Experiments PSwHR, PSwR and PSwRL: Traffic rates at different points.

function U_2), it can be calculated that $\lambda_n^* = 0.136$ call/sec and at this point $P_b = 2.5\%$ which means that more users are blocked. In this case, it is expected that PSwHR and PSwRL have even shorter "flat" periods.

From the above results and discussion, it can be concluded that no matter how users behave, the proposed integrated approach can always prevent the occurrence of congestion. Furthermore, the proposed scheme provides a dynamic method to operate the system at the optimal point, where the number of new calls that can be accepted in the system is maximized under the constraints of guaranteeing the required QoS for a given user utility function.



(a) PSwHR.

(b) PSwR.

(c) PSwRL.

Figure 4.9 Experiments PSwHR, PSwR and PSwRL: Weighted sum of call blocking probabilities.

Increasing Total User Utility. As discussed in Section 4.2, total user utility is a measure of how efficiently the channel resources are used. Table 4.1 compares the achieved total user utility during a one-day period for the five different experiments under evaluation. From this table, it is observed that no matter how the hold-off and blocked users behave, compared with the conventional CAC schemes, integrating

Table 4.1 Total User Utility for Each Experiment

Experiment	CSwR	CSwRL	PSwHR	PSwR	PSwRL
Total Utility ($\times 10^6$)	0.70	0.74	1.99	2.04	1.82

pricing with CAC always increases the total user utility significantly. Comparing the results of experiments PSwHR, PSwR and PSwRL, it can be found that PSwR has the largest achievable total user utility while PSwRL has the lowest one. The reason lies in that in PSwR all the hold-off users and blocked users choose to retry while in PSwRL some of users give up retrying, so that PSwR can serve more users than PSwRL, and PSwR has longer "flat" period than PSwRL, as can be seen by Figure 4.8.

Comparison of Obtained Revenue. Previous results demonstrate that integrating pricing with CAC can effectively prevent congestion occurrence and increase the total user utility. In this subsection, the impact of the integration of pricing in the call admission control process on the network operators/providers is studied from the generated revenue point of view. The average revenue can be calculated as follows:

$$R = \int_0^{T_0} \lambda_{admit}(t) p(t) \tau dt \qquad (4.15)$$

where $T_0 = 24hrs = 86400secs$ and τ is the average duration time of calls. The corresponding revenues for the five different experiments are shown in Table 4.2. Comparing the results of experiments PSwR and PSwRL with their counterpart

Table 4.2 Revenue for Each Experiment

Experiment	CSwR	CSwRL	PSwHR	PSwR	PSwRL
Revenue ($\times 10^6$)	2.07	2.03	3.63	3.73	2.58

experiments CSwR and CSwRL, respectively, it is observed that when pricing is introduced into the call admission control process, the revenue for operators is increased significantly as well (82% for experiment PSwR over CSwR and 28% for experiment PSwRL over CSwRL). The increase in revenue during busy hours (when $\lambda_n > \lambda_n^*$) is due to the fact that with the suggested demand function, when congestion occurs the effect of decrease in the number of accepted users can be compensated by the higher congestion price charged to users. For example, at peak hours, the λ_{admit} for CSwR and PSwR are 0.1487 and 0.119, respectively, while the price charged to each user are p_0 and $2.3p_0$, respectively. However, it should be noted here that PSwR can guarantee the required QoS while CSwR fails to do so. The overall increase in revenue during the hours following the busy hours is due to the combined effect of both the increased arrival traffic load (due to the traffic shaping feature of the proposed scheme *i.e.*, $\lambda_{in} \geq \lambda_n$ as can be observed from Figure 4.8) and increased price (*i.e.*, $p(t)/p_0 \geq 1$ as can be observed from Figure 4.7).

4.6 Maximizing the Total User Utility in Wireless Networks with Multiple Service Types

The development of wireless networks and of new emerging services, along with the evolution of communication infrastructures into multiplexed and multiple service-class networks, provides the capability and need to support a wide variety of applications. In order to allocate the resources efficiently among users that have different resource and QoS requirements, this section investigates how the total users utility varies with different traffic loads and studies the problem of maximizing the total user utility in wireless networks with multiple service types. Based on this, the input traffic space is divided into congestion space and operation space to avoid system congestion and provide users with satisfactory service.

4.6.1 Maximizing Total User Utility

In the following, a wireless networks that can support $T + 1$ different types of calls/applications is considered, where each one of them may present different QoS and bandwidth requirements. For simplicity in the presentation and without loss of generality it is assumed that type 0 corresponds to real-time services such as conversational voice, while types 1 to T correspond to T different types of data applications, that may range from web-browsing to transaction services, e-mail *etc.* The total arrival process of new calls is assumed to follow Poisson process with rate λ, while the type i $(i = 0, 1, \cdots, T)$ user channel holding time is assumed to follow exponential distribution with mean τ_{ch}^{i}. Upon acceptance, a type 0 call always requests bandwidth BW_0, while each type i $(i = 1, \cdots, T)$ data call can be associated with multiple levels of bandwidth assignment. Note that for each type i data user, different bandwidth assignment will result in different quality of service and therefore different levels of user satisfaction, which will be reflected by the corresponding utility function. It is also assumed that a type i data call has L_i different service choices (levels) where each level corresponds to certain bandwidth requirement $BW_i^j, (j = 1, 2, \cdots, L_i)$. Therefore, there are totally $K+1 = 1+\sum_{i=1}^{T} L_i$ classes of calls to be served in the system, where each class k corresponds to a certain application type i working on a certain service level j. Each call arrival is assumed to be of class k with some probability γ_k with $\sum_{k=0}^{K} \gamma_k = 1$. Please note that if each application type has only one service level, *i.e.*, $L_i = 1$, in this special case $K = T$.

If the parameters of the system (total capacity, service rate, *etc.*,) are given, the performance of the system depends on the offered traffic load (rate and composition). This section investigates how the total utility changes with the new call arrival rate for any given traffic composition, *i.e.*, γ_k values. In section 4.2, the problem of maximizing the total user utility for a single class of service was studied. This section extends this approach to the case of multiple service types.

For any fixed input traffic composition, the average number (N_k) of admitted class k users ($k = 0, 1, \cdots, K$) is defined as function of the new call arrival rate, *i.e.*, $N_k = f_k(\lambda)$, where $f_k(\lambda)$ are differentiable and monotonically increasing continuous functions of λ with the following properties:

$$0 \leq f_k(\lambda) < C/BW_k; f_k'(\lambda) > 0; f_k(\lambda = 0) = 0; \tag{4.16}$$

where C denotes the total capacity of the cell under consideration and BW_k is the bandwidth allocated to class k users based on the application type and service level.

For each class, a cost function is defined as the composite call blocking probability P_k represented by the weighted sum of new call blocking probability (P_k^{nb}) and handoff call blocking probability (P_k^{hb}) experienced by users of type i:

$$P_k = \alpha P_k^{nb} + \beta P_k^{hb} \qquad k = 0, 1, \cdots, K \tag{4.17}$$

where α and β are constants that denote the penalty associated with rejecting new calls and handoff calls respectively, with $\beta > \alpha$ to reflect the higher cost of blocking a handoff call. Because both P_k^{nb} and P_k^{hb} are monotonically increasing functions of λ, $P_k = g_k(\lambda)$ is also a monotonically increasing function of λ:

$$0 \leq g_k(\lambda) < 1; g_k'(\lambda) > 0; g_k(\lambda = 0) = 0; \lim_{\lambda \to \infty} g_k(\lambda) = 1 \tag{4.18}$$

Following the **Assumption 1** made in section 4.2, the general properties and characteristics of the user utility function are described below. The utility function models the network users' preference. It is argued that when the cost function (P_k) increases the users will observe higher blocking and their level of satisfaction decreases. Please also note that when P_k is small, the satisfaction degradation caused by the increase of P_k is not significant; as P_k becomes large, the satisfaction degradation will be substantial. For data users, *i. e.*, type $1, \cdots, T$, the user's preference to service is also related to the amount of bandwidth actually allocated to them. Taking these factors into consideration, the following assumption is made throughout this section:

Assumption 2. The utility function of a single type 0 user is a differentiable and monotonically decreasing concave function of the QoS parameter P_k $(k = 0)$. Therefore, $U_0 = h_v(P_0)$ where function $h_v(P_0)$ has the following properties:

$$h_v(P_0) \geq 0; \qquad h'_v(P_0) < 0; \qquad h''_v(P_0) < 0 \tag{4.19}$$

The utility function (U_k) of a single class k data user (whose application type is i and service level is j) is assumed to be:

$$U_k = h_d^k(P_k)BW_k/BW_i^{max} \qquad \text{for } k = 1, 2, \cdots, K \tag{4.20}$$

where BW_i^{max} denotes the maximum bandwidth that can be allocated to a single type i application, which the class k user belongs to, *i.e.*, $BW_i^{max} = \max_{j=1,2,\cdots,L_i}(BW_i^j)$ and function h_d^k models the class k users' perception to the call blocking probabilities, and it has the same properties with h_v. Note that all classes belonging to type i share the same function format of h_d^k.

For all users, U_k achieves maximum value at $P_k = 0$ $(k = 0, 1, \cdots, K)$, which means that if the blocking probability is 0% the user has the highest level of satisfaction. Therefore, the following is assumed here: $U_k(P_k = 0) = U_k^{max}$. Although different applications may have different QoS requirements and therefore different utility functions, for all practical purposes it can be assumed that there exists P_k^{max} such that $U_k(P_k) = 0$ when $P_k \geq P_k^{max}$. This means that when call blocking probability is very high, the user satisfaction is zero. In a realistic wireless system P_k^{max} represents the threshold value (maximum) of P_k that can be tolerated by class k users so that the Quality of Service is considered acceptable. The total user utility is defined as:

$$U = \sum_{k=0}^{K} N_k U_k = \mathscr{F}(\lambda, \gamma_0, \gamma_1, \gamma_2, \cdots, \gamma_K) \tag{4.21}$$

Based on the above definitions and assumptions the following theorem the following theorem can be proven:

Theorem 2. For a given wireless network and given traffic composition $\Gamma_0 = (\gamma_{0,0}, \gamma_{1,0}, \cdots, \gamma_{K,0})$, there exists an optimal new call arrival rate $\lambda^*(\Gamma_0)$ that maximizes the total utility U, *i.e.*, $\mathscr{F}(\lambda^*(\Gamma_0), \Gamma_0) \geq \mathscr{F}(\lambda, \Gamma_0)$ for $0 \leq \lambda < \infty$.

Proof. **(i)** The total user utility can be rewritten as:

$$U = f_0(\lambda) h_v(g_0(\lambda)) + \sum_{k=1}^{K} f_k(\lambda) h_d^k(g_k(\lambda)) \tag{4.22}$$

From the properties of function h_v, h_d^k and g_k,

$$\begin{aligned} \frac{dU_0}{d\lambda} &= h_v'[g_0(\lambda)]g_0'(\lambda) < 0 \\ \frac{dU_k}{d\lambda} &= h_d^{k'}[g_k(\lambda)]g_k'(\lambda) < 0 \qquad \text{for } k = 1, 2, \cdots, K \end{aligned}$$

which means that for all users the single user utility is a continuous and monotonically decreasing function of λ, with $U_k(\lambda) \geq 0$, $U_k(\lambda = 0) = U_k^{max}$ and $U_k(\lambda = \lambda_k^m) = 0$, where λ_k^m is the new call arrival rate that satisfies $g_k(\lambda_k^m) = P_k^{max}$. It is also set that $\lambda^{max} = \max_{k=0,1,\cdots,K}(\lambda_k^m)$.

(ii) Since both N_k and U_k are continuous functions over the closed interval $\lambda \in [0, \lambda^{max}]$, U as defined in equation (4.22), is also a continuous function over the same interval. According to Extreme Value Theorem [29], it can be concluded that function U has a minimum and a maximum value on that interval.

- From (i) and equation (4.16), the minimum value is obtained at the endpoints of the closed interval. Specifically, total utility U is zero either when $\lambda = 0$

($i.e.$, $U(\lambda = 0) = 0$) which corresponds to the case that there is no user in the system, or when $\lambda = \lambda^{max}$ ($i.e.$, $U(\lambda = \lambda^{max}) = 0$) which corresponds to the case that each single user's utility becomes zero.

- Let U_{ub} denote the maximum value of total user utility. Then by Extreme Value Theorem [29] it can be concluded that there exist at least one λ value(s), denoted by λ_n, $(n = 0, 1, \cdots)$ over interval $(0, \lambda^{max})$ such that:

$$U(\lambda = \lambda_n) = U_{ub} \tag{4.23}$$

(iii) The number of different values of λ_n depends on the second order derivatives of functions f_k, g_k h_v and h_d^k.

- If there exists only one λ_n $(n = 0)$ that satisfies equation (4.23), the optimal new call arrival rate for the system is $\lambda^*(\Gamma_0) = \lambda_0$;
- If there exist more that one different values of λ that satisfy equation (4.23), the optimal new call arrival rate is set to be: $\lambda^*(\Gamma_0) = \sup_{n \in \{0,1,\cdots\}}\{\lambda_n \mid U(\lambda_n) = U_{ub}\}$, which is the highest new call arrival rate value that can maximize the total user utility.

□

For given utility functions, these optimal values can be easily calculated for each traffic composition Γ using $K + 1$ dimensional Markov Chain model.

4.6.2 Operation Space and Congestion Space

As discussed in Section 4.2, the maximum total user utility also means that bandwidth resources are most efficiently utilized. When $\lambda > \lambda^*(\Gamma_0)$, a large number of users are blocked when trying to initiate their calls or when trying to handoff to another cell in the middle of a call, which means that the QoS degrades and may become unacceptable. Although on average the total resource utilization is higher, due to the increasing number of handoff failures it is hard for a user to finish the call successfully, and as a result the "effective" utilization of bandwidth resources is low. In this case, both the total user utility and the single user utility decrease with any further increase of λ and the cell is congested. From the viewpoint of congestion prevention, it is ideal for a system to operate at the optimal traffic load ($\lambda^*(\Gamma_0)$) or below when traffic composition is Γ_0.

Let S denote a $K+2$ dimensional space, $S = \{(\lambda, \gamma_0, \gamma_1, \gamma_2, \cdots, \gamma_K) \mid 0 \leq \lambda < \infty,\ 0 \leq \gamma_k \leq 1, \sum_{k=0}^{K} \gamma_k = 1\}$, where each point in S is a possible input traffic pattern into the system. Based on the theorem proposed in section 4.6.1 and the above discussion, the space S can be divided into two sub-spaces: Operation Space (OS) and Congestion Space(CS):

$$OS = \{(\lambda, \gamma_0, \gamma_1, \gamma_2, \cdots, \gamma_K) \mid 0 \leq \gamma_k \leq 1, \sum_{k=0}^{K} \gamma_k = 1, \quad 0 \leq \lambda \leq \lambda^*(\gamma_0, \gamma_1, \gamma_2, \cdots, \gamma_K)\} \tag{4.24}$$

and $CS = S - OS$.

Numerical results are presented in the following to demonstrate the existence of optimal new call arrival rates and the division of the traffic space into Operation Space (OS) and Congestion Space (CS). A wireless network that carries out the Guard Channel CAC scheme is considered where the capacity of each cell is C units of bandwidth, with C_h units out of the C reserved for handoff purposes only. In the experiment presented here, for simplicity in the presentation and without loss of generality, it is assumed that there are two types of calls (*i.e.,* $T = 1$): type 0 with $BW_0 = 1$, and type 1 calls that operate on a single service sub-class (*e.g.,* $L_1 = 1$ and $K = 1$) with $BW_1^1 = 2$. The single user utilities for type 0 and type 1 users are assumed to follow the following equations respectively (for $0 \leq P_0$):

$$U_0 = h_v(P_0) = \max(1 - e^{(P_0 - 0.01)}/1 - e^{(-0.01)}, 0)$$

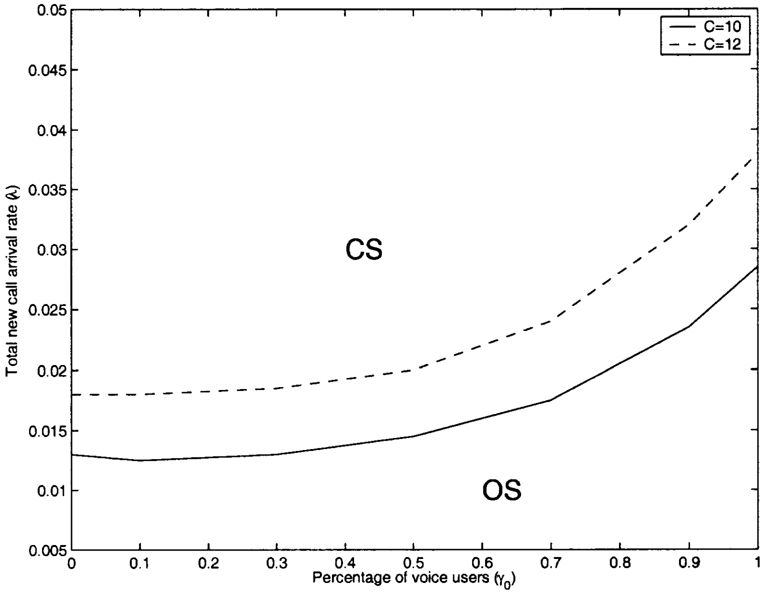


Figure 4.10 Operation Space (OS) and Congestion Space(CS)

$$U_1 = \frac{BW_1^1}{BW_1^{max}} h_d^1(P_1) = \max(1 - e^{(P_0 - 0.1)}/1 - e^{(-0.1)}, 0)$$

where $BW_1^{max} = BW_1^1$ since $L_1 = 1$. It should be noted that data users and voice users have different perception and tolerance towards call blocking. The corresponding numerical results for two different cell capacities $C = 10$ and $C = 12$ are presented in Figure 4.10 assuming that $C_h = 1$ and $\tau_{ch}^0 = \tau_{ch}^1 = 72$ seconds. For every point in each curve, the corresponding Y-axis value indicates the optimal new call arrival rate for the given traffic composition (represented by the X-axis value). The area below the curve is the OS and the area above is the CS.

CHAPTER 5

CONCLUSIONS

Resource allocation (RA) and Call Admission Control (CAC) in wireless networks are processes that control the allocation of the limited radio resources to mobile stations (MS) to maximize the utilization efficiency of radio resources and guarantee the (QoS) requirements of mobile users. Wireless networking has enjoyed dramatic increase in popularity over the last few years. The advances in hardware design and the increased user requirements for mobility and geographic dispersion have generated a tremendous need for the support of multiple classes of services with certain QoS requirements. Large-scale deployment of multimedia services over wireless networks depends heavily on the offered QoS, network reliability and cost effectiveness of the services. Therefore, more efficient RA/CAC schemes need to be designed to handle these emerging challenges.

Through the review of current research efforts in this area, it is observed that most existing schemes do not take the user mobility into consideration, so that they cannot effectively adapt to the real time network conditions. Moreover, these schemes do not provide incentives for users to use the limited radio resources efficiently, therefore, they cannot prevent the occurrence of network congestion. In this dissertation, several RA/CAC schemes are proposed and analyzed to solve the above mentioned problems.

Due to user mobility, an ongoing call, in addition to its channel requirements in the current cell, exerts some influence on the channel allocation in neighboring cells. The proposed MBCR scheme quantifies this influence and based on this makes resource reservations in neighboring cells. The MBCR algorithm can be carried out in a distributed way: each cell collects its current traffic condition, calculates the influence and sends the results to all its neighbors periodically. Extensive simulation study demonstrated that MBCR scheme outperforms the fixed guard channel schemes and other dynamic schemes in that it enables the system to automatically adjust to the changing traffic conditions, provides same quality of service to the ongoing calls throughout the system and simplifies the channel pre-allocation process.

In order to improve the applicability and efficiency of the MBCR schemes, the MBCR scheme is further developed into PMBBR scheme, which uses geolocation information in the reservation making process. Based on the history location information, the future moving speed and direction of each user can be predicted. These predictions are further used to refine the calculation of handoff probabilities and direction factors and to solve the problem of reservation timing. By taking into account the multiple service types with different QoS requirements, a bandwidth reconfiguration based call admission control strategy is proposed, which takes advantage of the service priorities and the flexible QoS requirements of certain service types. The proposed CAC scheme can be integrated with the PMBBR scheme to maximize the efficient use of available bandwidth in next generation wireless networks, that may support multiple classes of service with various levels of quality ranging from strict to flexible and soft QoS requirements. The performance analysis indicated that the integrated approach is capable of alleviating the problem of over-reservation, supporting seamless operation throughout the wireless network and increasing significantly the system capacity.

Most of the existing guard channel schemes use the total number of occupied channels as the decision variable for CAC. However, they cannot prevent the potential cell congestion, especially when users arrive in bursts as expected in current and future wireless networks. The traffic composition analysis and the MBCR simulation study revealed the fact that the channel holding time for handoff calls is shorter than that for new calls. Based on this conclusion, the New Call Bounding (NCB) scheme, which uses the number of admitted new calls as the decision variable for CAC, is proposed and analyzed.

Furthermore, this dissertation investigates the role of pricing as an additional dimension of the call admission control process in order to efficiently and effectively control the use of wireless network resources, and proposes to integrate dynamic pricing with CAC. It is proven that for a given wireless network there exists a new call arrival rate which can maximize the total utility of users, while maintaining the required QoS. Based on this result, a pricing strategy is developed, which can dynamically adjust the price according on the current network conditions. The proposed scheme works as a traffic shaper which implicitly implements a distributed user-based traffic prioritization. The problem of network congestion, as well as the achievable total user utility and obtained revenue, has been extensively studied via simulation. The corresponding numerical results indicate that the proposed scheme can alleviate the problem of congestion in wireless networks, while at the same time achieves to meet the required user QoS and maximize the efficient use of channel resources.

This integrated approach is also extended into multiple service-class network environment. It is further proven that in a wireless network, for any traffic composition there always exits a new call arrival rate that maximizes the total users utility. Based on these values, the input traffic space can be divided into Operation Space and Congestion Space. These results can provide guidelines and insights for the development of efficient call admission control schemes, resource allocation algorithms and pricing schemes in order to increase the total benefit of the user community and avoid the system congestion.

REFERENCES

1. A. S. Acampora and M. Naghshineh, "An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 8, October 1994, pp. 1365-1374.

2. S. Aidarous and T. Plevyak (Editors), "Telecommunications Network Management into the 21st century: Techniques, Standards, Technologies, and Applications," *IEEE Press*, 1994.

3. S. Aidarous and T. Plevyak (Editors), "Telecommunications Network Management: Technologies and Implementations," *IEEE Press*, 1998.

4. N. Bartolini and I. Chlamtac, "Call Admission Control in Wireless Multimedia Networks," *Proceedings of IEEE PIMRC'02*, pp. 285-289, 2002.

5. D. Bertsekas and R. Gallager, *Data Networks 2nd Ed.* Prentice Hall, Upper Saddle River, NJ 1992.

6. S. Boumerdassi and A. Beylot, "Adaptive Channel Allocation for Wireless PCN," *Mobile Networks and Applications*, No. 4 1999.

7. G. Brasche and Bernhard Walke, "Concepts, Services, and Protocols of the New GSM Phase 2+ General Packet Radio Service," *IEEE Communications Magazine*, August 1997.

8. J. J. Caffery, Jr. and G. L. Stuber, "Overview of Radiolocation in CDMA Cellular Systems," *IEEE Communications Magazine*, pp. 38-45, April 1998.

9. C. Chang, C. J. Chang and K. Lo, "Analysis of a Hierarchical Cellular System with Reneging and Dropping for Waiting New and Handoff Calls," *IEEE Transactions on Vehicular Technology*, Vol. 48, No. 4, July 1999.

10. C. J. Chang T. T. Su and Y. Y. Chiang, "Analysis of a Cutoff Priority Cellular Radio System with Finite Queueing and Reneging/Dropping," *IEEE/ACM Transactions on Networking*, Vol. 2, No. 2, April 1994.

11. K. N. Chang, J. T. Kim, C. S. Yim and S. Kim, "An Efficient Borrowing Channel Assignment Scheme for Cellular Mobile Systems," *IEEE Transactions on Vehicular Technology*, Vol. 47, No. 2, May 1988.

12. C. Chao and W. Chen, "Connection Admission Control for Mobile Multiple-Class Personal Communications Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, October 1997.

13. M. Cheung and J. W. Mark, "Effect of Mobility on QoS Provisioning in Wireless Communication Networks," *Proceedings of IEEE WCNC'99*, New Orleans, September 1999.

14. M. Chiu and M. A. Bassiouni, "Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 3, pp. 510-522, March 2000.

15. S. Choi and K. G. Shin, "Predictive and Adaptive Bandwidth Reservation for Hand-Offs in QoS-Sensitive Cellular Networks," *Proceedings of the ACM SIGCOMM'98*, September 1998.

16. R. Cocchi, S. Shenker, D. Estrin and L. Zhang, "Pricing in Computer Networks: Motivation, Formulation and Example," *IEEE/ACM Transactions on Networking*, Vol. 1, No. 6, December 1993.

17. C. Courcoubetis, V. A. Siris and G. D. Stamoulis, "Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks," *Proceedings of IEEE GLOBECOM 1996*, London.

18. L. A. Dasilva, "QoS-Enabled Networks: A Survey," *IEEE Communication Surveys*, Second Quarter 2000.

19. G. Donis-Hernández, D. Muñoz-Rodriguez and S. Tekinay, "ATM Virtual Cell Area Determination in PCS Networks," *Proceedings of 5th IEEE International Conference on Universal Personal Communications*, Vol. 2, pp. 697-701, September 1996.

20. B. M. 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, No. 3, March 2000.

21. M. Falkner, M. Devetsikiotis and I. Lambadaris, "An Overview of Pricing Concepts for Broadband IP Networks," *IEEE Communications Surveys*, 2nd Quarter, 2000.

22. Y. Fang and I. Chlamtac, "Teletraffic Analysis and Mobility Modeling for PCS Networks," *IEEE Transactions on Communications*, Vol. 47, No. 7, pp. 1062-0172, July 1999.

23. Y. Fang I. Chlamtac and Y. B. Lin, "Modeling PCS Networks Under General Call Holding Times and Cell Residence Time Distribution," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, pp. 893-906, December 1997.

24. Y. Fang, "Hyper-Erlang Distributions and Traffic Modeling in Wireless and Mobile Networks," *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), 1999*, pp. 398-402.

25. Y. Fang and I. Chlamtac, "A New Mobility Model and Its Application in the Channel Holding Time Characterization in PCS Networks," *Proceedings of IEEE INFOCOM '99*, pp. 20-27.

26. Y. Fang, I. Chlamtac and Y. Lin, "A General Formula for Handoff Rate in PCS Networks," *Proceedings of IEEE PIMRC '98*, pp. 330-334.

27. P. C. Fishburn and A. M. Odlyzko, "Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet," *ICE'98*, pp. 128-139

28. B. Gavish and S. Sridhar, "Threshold Priority Policy for Channel Assignment in Cellular Networks," *IEEE Transactions on Computers*, Vol. 46, No. 3, March 1997.

29. C. Goffman, *The Calculus: an Introduction,* Harper & Row, Publisher, 1971

30. G. Goldszmidt and Y. Yemini, "Delegated Agents for Network Management," *IEEE Communications Magazine*, pp. 66-70, March 1998.

31. R. Guérin, "Queueing-Blocking System with Two Arrival Streams and Guard Channels," *IEEE Transactions on Communications*, Vol. 36, No. 2, pp. 153-163, February 1988.

32. C. Ho and C. Lea, "Improving Call Admission Policies in Wireless Networks," *Wireless Networks*, 5(1999)257-265.

33. L. L. Ho, D. J. Cavuto, S. Papavassiliou and A. G. Zawadzki, "Adaptive and Automated Detection of Service Anomalies in Transaction-Oriented WAN's': Network Analysis, Algorithms, Implementation and Deployment," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 5, May 2000.

34. 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,* Vol. 35, pp. 77-92, August 1986.

35. J. Hou, Y. Fang and A. Akansu, "Mobility-based Channel Reservation Scheme for Wireless Mobile Networks," *Proceedings of IEEE WCNC 2000*, pp. 527-531, Chicago, September 2000.

36. J. Hou and Y. Fang, "Mobility-based Call Admission Control Schemes for Wireless Mobile Networks," *Wiley International Journal on Wireless Communications and Mobile Computing*, Vol. 1, Issue 3, pp. 269-282, 2001.

37. J. Hou and S. Papavassiliou, "Influence-Based Channel Reservation Scheme for Mobile Cellular Networks," *Proceedings of IEEE ISCC 2001*, pp. 218-223, July 2001.

38. J. Hou and S. Papavassiliou, "A Dynamic Reservation Based Call Admission Control Algorithm for Wireless Networks Using the Concept of Influence Curve," *Journal of Telecommunications Systems*, 22:1-4, pp. 299-319, 2003.

39. J. Hou, J. Yang and S. Papavassiliou, "Integration of Pricing with Call Admission Control for Wireless Networks," *Proceedings of IEEE VTC 2001/Fall*, pp. 1344-1348, October, 2001.

40. 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. 13, No. 9, pp. 898-910, September 2002.

41. L. Hu and S. S. Rappaport, "Personal Communication System Using Multiple Hierarchical Cellular Overlays," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 2, pp. 406-415, February 1995.

42. D. L. Jagerman, "Some Properties of the Erlang Loss Function," *The Bell System Technical Journal*, Vol. 53, No. 3, pp. 525-551, March 1974.

43. H. Ji, J. Y. Hui and E. Karasan, "GoS-Based Pricing and Resource Allocation for Multimedia Broadband Networks," *Proceedings of IEEE INFOCOM 1996*, pp. 1020-1027, 1996.

44. H. Jiang and S. S. Rappaport, "Prioritized Channel Borrowing Without Locking: A Channel Sharing Strategy for Cellular Communications," *IEEE/ACM Transactions on Networking*, Vol. 4, No. 2, April 1996.

45. Z. Jiang, H. Mason, B. J. Kim, N. K. Shankaranarayanan, and P. Henry, "A Subjective Survey of User Experience for Data Application for Future Cellular Wireless Network," *Proceedings of Symposium on Applications and the Internet 2001*, pp. 167-175, January 2001.

46. I. Katzela and M. Naghshineh, "Channel Assignment Schemes for Cellular Mobile Telecommunication System: A Comprehensive Survey," *IEEE Personal Communications*, June 1996.

47. J. S. Kaufman, "Blocking in a Shared Resource Environment," *IEEE Transactions on Communications*, Vol. 29, No. 10, October 1981.

48. J. Keilson and O. C. Ibe, "Cutoff Priority Scheduling in Mobile Cellular Communication System," *Transactions on Communications*, Vol. 43, No. 2/3/4, February/March/April 1995.

49. K. R. Krishnan, "The Convexity of Loss Rate in an Erlang Loss System and Sojourn in an Erlang Delay System with Respect to Arrival rate and Service Rate," *IEEE Transactions on Communications*, Vol. 38, No. 9, September 1990.

50. M. D. Kulavaratharasah and A. H. Aghvami, "Teletraffic Performance Evaluation of Microcellular Personal Communication Networks (PCN's) with Prioritized Handoff Procedures," *IEEE Transactions on Vehicular Technology*, Vol. 48, pp. 137-152, January 1999.

51. B. C. Kuo, *Automatic Control Systems, 5th Ed.*, Prentice Hall, Englewood Cliffs, NJ, 1987.

52. V. K. N. Lau and S. V. Maric, "Mobility of Queued Call Requests of a New Call Queuing Technique for Cellular Systems," *IEEE Transactions on Vehicular Technology*, Vol. 47, No. 2, May 1998.

53. D. A. Levine, I. F. Akylidiz 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, No. 1, February 1997.

54. B. Li, C. Lin and S. T. Chanson, "Analysis of a Hybrid Cutoff Priority Scheme for Multiple Classes of Traffic in Multimedia Wireless Networks," *Wireless Network*, Vol. 4, No. 4, pp. 279-290, July 1998.

55. J. Li, N. B. Shroff and E. K. P. Chong, "Channel Carrying: A Novel Handoff Scheme for Mobile Cellular Networks," *IEEE/ACM Transactions on Networking*, Vol. 7, No. 1, February 1999.

56. Y. B. Lin, A. R. Noerpel and D. J. Harasty, "The Sub-rating Channel Assignment Strategy for PCS Hand-offs," *IEEE Transactions on Vehicular Technology*, Vol. 45, No. 1, pp. 122-130, February 1996.

57. Y. B. Lin, S. Mohan and A. Noerepel, "Queueing Priority Channel Assignment Strategies for PCS Hand-Off and Initial Access," *IEEE Transactions on Vehicular Technology*, Vol. 43, No. 3, August 1994.

58. Y. Ma, J. J. Han and K. S. Trivedi, "Call Admission Control for Reducing Dropped Calls in Code Division Multiple Access (CDMA) Cellular Systems," *Proceedings of IEEE INFOCOM 2000*, pp. 100-110, March 2000.

59. J. K. MacKie-Mason and H. R. Varian, "Pricing Congestible Network Resources," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, September 1995.

60. H. De Meer, A. La Corte, A. Puliafito, and O. Tomarchio, "Programmable Agents for Flexible QoS Management in IP Networks," *IEEE Journal on Selected Areas in Communication*, 18(2), February 2000.

61. M. Naghshineh and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 4, May 1996.

62. M. Naghshineh and A. S. Acampora, "QoS Provisioning in Micro-cellular Networks Supporting Multiple Classes of Traffic," *Wireless Networks*, 2(1996) 195-203.

63. S. Ohmori, Y. Yamao and N. Nakajima, "The Future Generations of Mobile Communications Based on Broadband Access Technologies," *IEEE Communications Magazine*, Vol. 38, Issue 12, pp. 134-142, December 2000.

64. C. Oliveira, J. B. Kim and T. Suda, "An Adaptive Bandwidth Reservation Scheme for High-Speed Multimedia Wireless Network," *IEEE Journal on Selected Area in Communications*, Vol. 16, No. 6, pp. 858-874, August 1998.

65. L. Ortigoza-Guerrero and A. H. Aghvami, "A Prioritized Handoff Dynamic Channel Allocation Strategy for PCS," *IEEE Transactions on Vehicular Technology*, Vol. 48, No. 4, August 1999.

66. K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. Makela, R. Pichna and J. Vallstrom, "Handoff in Hybrid Mobile Data Networks," *IEEE Personal Communications*, April 2000.

67. S. Papavassiliou, S. Ozer and J. Hou, "Admission Control in Wireless Networks," *The Wiley Encyclopedia of Telecommunications*, John Wiley & Sons, Inc, January 2003.

68. H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," *IEEE Communications Magazine*, pp. 82-91, November 1996.

69. P. Ramanathan, K. M. Sivalinga, P. Agrawal and S. Kishore, "Dynamic resource Allocation Schemes During Handoff for Mobile Multimedia Wireless Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 7, pp. 1270-1283, July 1999.

70. R. Ramjee, D. Towsley and R. Nagarajan, "On optimal call admission control in cellular networks," *Wireless Networks*, Vol. 3, No. 1, pp. 29-41, March 1997.

71. S. S. Rappaport and C. Purzynski, "Prioritized Resource Assignment for Mobile Cellular Communication Systems with Mixed Services and Platform Types," *IEEE Transactions on Vehicular Technology*, Vol. 45, No. 3, August 1996.

72. T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, Upper Saddle River, NJ 1995.

73. V. Shah, N. B. Mandayam and D. J. Goodman, "Power Control for Wireless Data based on Utility and Pricing," *Proceedings of 9th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1427-1432, 1998.

74. S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, pp. 1176-1188, September 1995.

75. S. J. Shenker, "Making Greed Work in Networks: A Game-theoretic Analysis of Switch Service Disciplines," *IEEE/ACM Transactions on Networking*, Vol. 3, No. 6, December 1995.

76. J. S. Shih, R. H. Katz and A. D. Joseph, "Pricing Experiments for a Computer-Telephony-Service Usage Allocation," *Proceedings of IEEE GLOBECOM 2001*, November 2001.

77. M. Sidi and D. Starobinski, "New Call Blocking versus Handoff blocking in cellular networks," *Wireless Network*, Vol. 3, No. 1, pp. 15-27, March 1997.

78. B. Sklar, *Digital Communications: Fundamentals and Applications*, Prentice Hall, Englewood Cliffs, NJ 1988.

79. A. S. Tanenbaum, *Computer Networks 3rd Ed.*, Prentice Hall, Upper Saddle River, NJ 1996.

80. S. Tekinay and B. Jabbari, "Handover and Channel Assignment in Mobile Cellular Networks," *IEEE Communication Magazine*, pp. 42-46, November 1991.

81. S. Tekinay and B. Jabbari, "A Measurement-Based Prioritization Scheme for Handovers in Mobile Cellular Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 10, No. 8, May 1992.

82. H. R. Varian, *Microeconomic Analysis (2nd Ed.)*, W. W. Norton and Company, 1987.

83. Q. Wang and J. M. Peha, "State-Dependent Pricing and Its Economic Implications," *Telecommunication Systems*, 18:4, 315-329, 2001.

84. H. Xie and D. J. Goodman, "Mobility Models and Biased Sampling Problem," *Proceedings of 2nd IEEE International Conference on Universal Personal Communications*, Vol. 2, pp. 803-807, October 1993.

85. W. Yang and E. Geraniotis, "Admission Policies for Integrated Voice and Data Traffic in CDMA Packet Radio Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 4, pp. 654-664, May 1994.

86. J. Ye, J. Hou and S. Papavassiliou, "Mobile Agent Based Framework for Mobility Assisted Channel Reservation in Wireless Networks," *Proceedings of 36th Annual Conference on Information Sciences and Systems*, pp. 833-838, Princeton, March 2002.

87. 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, pp. 249-264, 2002.

88. J. Ye, J. Hou and S. Papavassiliou, "Integration of Advanced Reservation and Bandwidth Reconfiguration Based Admission Control in Wireless Networks with Multimedia Services," *Proceedings of IEEE MWN 2003*, May 2003.