Optimal resource management in multi-service mobile cellular networks

Yang, Xun

2008

Yang, X. (2008). Optimal resource management in multi-service mobile cellular networks. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/13224

https://doi.org/10.32657/10356/13224

Optimal Resource Management in Multi-service Mobile Cellular Networks

Yang Xun

School of Electrical and Electronic Engineering

A thesis submitted to Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2008

To my husband, my parents and my dear sister

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Feng Gang, for helping me so much from a bachelor degree holder to a Ph.D. He has been an ideal supervisor from every aspect, both in terms of technical and instructive advice on my research and in terms of general advice on my career. I believe I am not able to complete this thesis without his advice and encouragement. I also would like to present my appreciation to my co-supervisor, Prof. Huang Guangbin, for his generous help.

I benefited from the collaborations and discussions with the members of Special Interest Group on Networking (SIGNet) of ICIS, NTU. I am also grateful to Xie Feng, Long Fei, Fang Can and Xue Daojun. I had a great time working with them and learned a lot from the group interactions. A special thank goes to my husband, He Xuzhou, my parents and sister whom I love deeply. They are always supportive to me.

I would also like to express my gratitude to the anonymous reviewers for their valuable comments on my submitted papers.

Table of Contents

Summary

In the design of mobile cellular networks, resource management (RM) plays a critical role in Quality of Service (QoS) provisioning. At call level, three main RM optimization problems: maximizing system utilization, minimizing system cost and minimizing call blocking probability are investigated extensively in traditional mono-service mobile networks. With the increase of Internet access and many data applications, traffic load in future mobile cellular networks presents significant asymmetry between uplink and downlink. Traditional RM schemes, which may be optimal for either one of three RM optimization problems in mono-service mobile networks, becomes inappropriate in such a multi-service environment. In this thesis, we focus on the research of two key RM issues: call admission control (CAC) and bandwidth allocation (BA), for the RM optimization problems in multi-service mobile networks.

We first study the *MAXU* problem, which is defined as maximizing system utilization subject to constraints on call blocking probabilities. In multi-service mobile networks, bandwidth allocated on uplink and downlink is different in order to satisfy asymmetric traffic load brought by some data applications. Since it is difficult to promptly adjust bandwidth allocation on uplink and downlink according to the change of traffic load in a system, the mismatch of bandwidth allocation and traffic load results in low bandwidth utilization. In such an environment, traditional CAC schemes may admit superfluous Real-Time (RT) calls or Non-Real-Time (NRT) calls and thus lead to bandwidth waste. We propose and evaluate two new CAC schemes to address the low bandwidth utilization problems in such bandwidth asymmetry networks. Our design objective is to improve bandwidth utilization while retaining handoff call dropping probabilities of both RT can NRT calls at a reasonably low level. By determining the admissible regions for RT calls and NRT calls, the proposed schemes prevent a specific call class from overusing bandwidth resources. Mathematical analysis and simulation experiments are employed to evaluate the performance of the proposed

schemes and some existing schemes. Numerical results show that the proposed schemes can achieve better performance in terms of call dropping and blocking probability and bandwidth utilization compared with some existing schemes, even those performing well in bandwidth asymmetry mobile networks.

Next, we focus on minimizing average system cost (*MINCost*) problem in multi-service mobile cellular networks. By modeling admission decision as a Markov decision process (MDP) and analyzing the corresponding value function, we obtain some monotonicity properties of the optimal policy. These properties suggest that the optimal admission control policy for the bandwidth asymmetry mobile networks have a threshold structure and the threshold specified for a call class may change with system states. Because of the prohibitively high complexity for computing the thresholds in a system with a large state-space, we propose a heuristic CAC policy called Call-Rate-based Dynamic Threshold (CRDT) policy to approximate the theoretical optimal policy based on the insights obtained from the modeling and the analytical study on the properties of the optimal policy. The CRDT policy is efficient and can be easily implemented. Numerical results show that the proposed CRDT policy provides a sub-optimal solution to the optimal policy for the MINCost problem in the bandwidth asymmetry mobile networks.

Subsequently, we turn to study the problem of minimizing new call blocking probabilities with hard constraints on handoff call dropping probabilities (*MINBlock*) in multi-service mobile cellular networks. Different from traditional mono-service networks, different call classes may have different constraints on handoff call dropping probabilities in a multi-service mobile network, which makes the derivation of thresholds for various call classes more complicated. In this work, we investigate how to find appropriate thresholds based on the system information from not only local cell but also neighboring cells. Based on that, we propose a new distributed multi-service admission control scheme (DMS-AC) to handle the MINBlock problem in multi-service mobile wireless networks. By computing and setting different thresholds for different call classes, the proposed CAC scheme controls the admission of new calls and thus avoids handoff call dropping probabilities of different call classes from exceeding the predefined constraints and at the same time new call blocking probabilities are also minimized.

In a dynamic traffic network, the traffic load in the system changes over time. When the offered

traffic load exceeds the control range of a employed CAC scheme, the QoS of some call classes cannot be guaranteed. In order to satisfy the QoS requirements of different call classes in a dynamic traffic load environment, we propose bandwidth re-allocation as a complementary mechanism for CAC in bandwidth asymmetry mobile networks. Based on the proposed DMS-AC scheme, we investigate *when* and *how* to adjust bandwidth allocation on uplink and downlink in a multi-service mobile network with bandwidth asymmetry under dynamic traffic load conditions. Our design objective is to improve system bandwidth utilization while satisfying the call-level QoS requirements of different call classes. When the traffic load brought by some call classes under the dynamic traffic conditions in a system exceeds the control range of DMS-AC, bandwidth re-allocation process is activated and the admission control policy will try to meet the QoS requirements under the adjusted bandwidth allocation. We explore the relationship between admission thresholds and bandwidth allocation by identifying certain constraints to verify the feasibility of the adjusted bandwidth allocation. We conduct extensive simulation experiments to validate the effectiveness of the proposed bandwidth re-allocation scheme. Numerical results show that when traffic pattern with certain bandwidth asymmetry between uplink and downlink changes, the system is able to re-allocate the bandwidth on uplink and downlink adaptively. With the designed bandwidth re-allocation scheme in conjunction with the proposed DMS-AC, the QoS requirements of different call classes can be guaranteed under dynamic traffic conditions and in the mean time the system bandwidth utilization is improved significantly.

Our work in this thesis is an essential extension for resource management in the design of multi-service mobile cellular networks, especially for bandwidth asymmetry mobile networks. By studying and analyzing the special features of the multi-service mobile networks, we investigate main call-level RM optimization problems in a bandwidth asymmetry environment, and propose some efficient and effective RM schemes based on comprehensive analysis and mathematical models. We believe that our work can bring some insights to the research work in the area of RM design in multi-service mobile cellular networks.

List of Figures

List of Tables

List of Abbreviations

BA	: Bandwidth Allocation
BS	: Base Station
CAC	: Call Admission Control
CDMA	: Code Division Multiple Access
CRDT	: Call-Rate-based Dynamic Threshold
DA	: Different Time-slot Allocation
DCA	: Distributed Call Admission Control
DL	: Downlink
DMS-AC	: Distributed Multi-service Admission Control
DMTBR	: Dynamic Multiple-Threshold Bandwidth Reservation
DP	: Dynamic Partition
DTBR	: Dual Threshold Bandwidth Reservation
FDD	: Frequency Division Duplex
GC	: Guard Channel
GSM	: Global System for Mobile Communication System
HTTP	: Hypertext Transfer Protocol
IMT-2000	: International Mobile Telecommunications-2000
LFGC	: Limited Fractional Guard Channel
MAXU	: Maximizing System Resource Utilization
MDP	: Markov Decision Process
MINBlock	: Minimizing New Call Blocking Probabilities with Hard Constraints on Handoff Call Blocking Probabilities
MINCost	: Minimizing Average System Cost

MINOBJ	: Minimizes an Objective Function
MLC	: Most Likely Cluster
MMS	: Multimedia Message Service
MT	: Mobile Terminal
NRT	: Non Real Time
QoS	: Quality of Service
RM	: Resource Management
RCS	: Restricted Complete Sharing
RT	: Real Time
SA	: Same Time-slot Allocation
TDD	: Time Division Duplex
UL	: Uplink
UMTS	: Universal Mobile Telecommunications System

List of Mathematical Symbols

P_h	: dropping probability of handoff call
P_n	: blocking probability of new call
P_h^i	: dropping probability of class i handoff call
P_n^i	: blocking probability of class i new call
Γ_s	: asymmetry factor
Γ_{NRT}	: asymmetry factor of NRT calls
B_u	: total uplink bandwidth
B_d	: total downlink bandwidth
B_{RT}^u	: uplink bandwidth used by RT calls
B_{RT}^d	: downlink bandwidth used by RT calls
B_{NRT}^u	: uplink bandwidth used by NRT calls
B_{NRT}^d	: downlink bandwidth used by NRT calls
B_{GC}	: capacity of guard channels
B_{CC}	: capacity of common channels
B_{NRT}	: capacity of NRT channels
B_r	: total bandwidth of C_r
B_l	: total bandwidth of C_l
C_r	: right cell
C_l	: left cell
b_i^u	: required uplink bandwidth of a class i call
b_i^d	: required downlink bandwidth of a class i call
b_{RT}^u	: required uplink bandwidth of an RT call

b^d_{RT} : required downlink bandwidth of an RT call

b^u_{NRT} : required uplink bandwidth of an NRT call

b^d_{NRT} : required downlink bandwidth of an NRT call

λ_i : mean call arrival rate of call class i

λ_{RT} : mean call arrival rate of RT call

λ_{NRT} : mean call arrival rate of NRT call

$1/\mu_i$: mean service time of call class i

$1/\mu_{RT}$: mean service time of RT call

$1/\mu_{NRT}$: mean service time of NRT call

$R^u_{h,i}(\pi)$: remaining uplink bandwidth that could be used by class i handoff calls when the system state is π

$R^u_{n,i}(\pi)$: remaining uplink bandwidth that could be used by class i new calls when the system state is π

$R^d_i(\pi)$: remaining downlink bandwidth that could be used by the calls of class i when the system state is π

U_{up} : uplink bandwidth utilization

U_{down} : downlink bandwidth utilization

$g(\cdot)$: cost function

G_k : cost of the k_{th} stage

Λ_x : overall system transition rate

$V(\cdot)$: value function

e_i : i_{th} unity vector

ρ_{RT} : traffic load brought by RT calls

ρ_{NRT} : traffic load brought by NRT calls

φ^l_i : call dropping probability of call class i in C_l

φ^r_i : call dropping probability of call class i in C_r

$\hat{\varphi}^l_i$: computed call dropping probability of call class i in C_l

$\hat{\varphi}^r_i$: computed call dropping probability of call class i in C_r

$\varphi^l_{i,j}$: probability that the cell C_l is at one of the overload states of call class j, which results due to the admission of class i calls.

$\varphi_{i,j}^r$: probability that the cell C_r is at one of the overload states of call class j, which results due to the admission of class i calls.

ϕ_i^r : new call blocking probability of call class i in C_r.

ϕ_i^l : new call blocking probability of call class i in C_l.

$\hat{\phi}_i^r$: computed new call blocking probability of call class i in C_r.

$\hat{\phi}_i^l$: computed new call blocking probability of call class i in C_l.

$P_r(n_i)$: probability that there are n_i class i calls in C_r during a control period

$P_l(n_i)$: probability that there are n_i class i calls in C_l during a control period

$P_{i,r}^s$: probability that a class i call in C_r remains in the same cell during a control period

$P_{i,l}^s$: probability that a class i call in C_l remains in the same cell during a control period

$P_{i,rl}^m$: probability that a class i call moves from C_r to C_l during a control period

$P_{i,lr}^m$: probability that a class i call moves from C_l to C_r during a control period

$\mathbf{S}$: set of the feasible states of a cell

s_k : the k_{th} state of $\mathbf{S}$

$\mathbf{S_{i,j}}$: set of states for call class j, such that when a system is at a state $s_k(s_k \in \mathbf{S_{i,j}})$, it can reach the overload states of class j with the increase of the number of class i calls in the system

$N_{i,j}(s_k)$: minimum number of class i calls that let the system enter the overload states of call class j when system is at state s_k

$Q(\cdot)$: integral over the tail of a Gaussian distribution

Th_i : threshold of call class i

γ_u : traffic load on uplink

γ_d : traffic load on downlink

Chapter 1

Introduction

1.1 Background and Motivations

Accompanying the booming of Internet and popularization of cell phone, laptop and other mobile devices, new applications such as rich voice, video, Internet access, web browsing, data transmission and multimedia will be supported in future mobile cellular networks [1]. According to the estimation of Universal Mobile Telecommunications System (UMTS) Forum, there will be over 50% daily traffic brought by mobile data in western Europe, which is more than two times of that brought by voice service [2]. It has been widely accepted that the coming trend is the combination of mobile wireless networks and Internet and multiple services will be supported in a mobile environment [1]. With the increase of mobile data users, multi-service mobile cellular networks present some distinctive features compared with traditional mono-service mobile networks.

1. Handoff: One of the notable features of mobile cellular network is that a mobile user is able to change the point of attachment to a mobile network during an ongoing communication session. This phenomenon is called handoff. If there are no sufficient resources in the target cell for a handoff call, this call will be dropped. Since it is more annoying to be disrupted during a communication session, how to decrease handoff call dropping is a critical problem in the design of mobile cellular networks. With the increase of mobile users, the micro-cells may not have sufficient resources to support so many users. One of the applicable

methods is to reduce the size of original cells [3,4]. Thus micro-cells may split into pico-cells, where handoff happens more often than that in the micro-cell networks. On the other hand, data or multimedia calls may also have handoff attempts in multi-service mobile networks and different call classes may have various QoS requirements. These factors make handoff more complicated than that in mono-service networks. We need to analyze the new possible problems in the multi-service mobile networks and design appropriate schemes to guarantee the QoS of different multi-service users.

2. Limited bandwidth resources: Compared with wired networks, the radio bandwidth of certain frequency of mobile networks is scares. Although the mobile systems are evolving from the second-generation, such as Global System for Mobile communication system (GSM), to the third-generation (3G), such as UMTS, and the maximum data transmission rate increases from $9.6kbps$ (GSM) to $2Mbps$ (UMTS, low speed mode), it still cannot satisfy some data transmission requirements of many multi-servcie users. Table 1.1 and Table 1.2 show the service classes and the provided system bandwidth of different modes defined in International Mobile Telecommunications-2000 (IMT-2000) systems [5], which is the umbrella specification of all 3G systems. We can find that the bandwidth is only $384kbps$ when a user is on move with low rate and it is even lower when the users' moving speed increases. The limited bandwidth resource makes how to use system resources efficiently in such multi-service mobile networks a critical issue. Therefore, resource management (RM) scheme should be designed delicately to maximize system utilization and at the same time satisfy the QoS requirements of users.

Table 1.1: Call classes defined in IMT-2000 systems

	Class-A	Class-B	Class-C	Class-D
Bit rate	$4-25kbps$	$32-384kbps$	$64/144/384/2048kbps$	NA
Service example	voice	video	Internet access	E-mail

Table 1.2: IMT-2000 system bandwidth requirements

Indoor and office	Pedestrian	Vehicular up to 120kmph
$2Mbps$	$384kbps$	$144kbps$

3. Asymmetric traffic load: In multi-service mobile networks, many applications which are popular in Internet will be supported in a mobile wireless environment. Different from traditional

voice service, many data services present significant asymmetric bandwidth requirements between uplink and downlink. A representative example is web browsing service. When we check a web page, a short request message through the hypertext transfer protocol (HTTP) is sent to a server initially, and then a large amount of information including web page text, background image, pictures included in the web page, video clips and even multimedia files will be downloaded from the server to the end user. In such client-server application, the traffic on uplink is usually much lighter than that on downlink. Fig. 1.1 shows the offered traffic asymmetry for each service category of 3G multimedia services [6]. From this figure we can find that all services bring much more traffic load on downlink than on uplink except simple voice and multimedia message service (MMS). This type of asymmetry becomes usual as more and more data services become available. In order to improve system utilization in such an asymmetric traffic load environment, it is necessary to allocate different bandwidth between uplink and downlink. Unfortunately, only little existing research work focuses on RM in such bandwidth asymmetry networks since the bandwidth allocation is always symmetric in traditional mono-service mobile networks. How to find efficient RM schemes to guarantee the QoS of different call classes and improve system utilization in the bandwidth asymmetry environment challenges the traditional RM schemes and motivates us to investigate RM problems in multi-service mobile networks.

Considering the above features of multi-service mobile networks, it is necessary to study how to maximize system utilization (*MAXU*), how to minimize average system cost (*MINCost*) and how to decrease call blocking probability of different call classes (*MINBlock*) in the design of RM schemes in multi-service mobile wireless networks. These problems are three main optimization problems of call-level resource management. In this Ph.D research, I concentrate on two key issues of call-level RM, call admission control (CAC) and bandwidth allocation (BA), for solving these optimization problems in multi-service mobile networks, especially in the networks with bandwidth asymmetry. The objective is to design more effective and efficient RM schemes in multi-service mobile wireless networks.

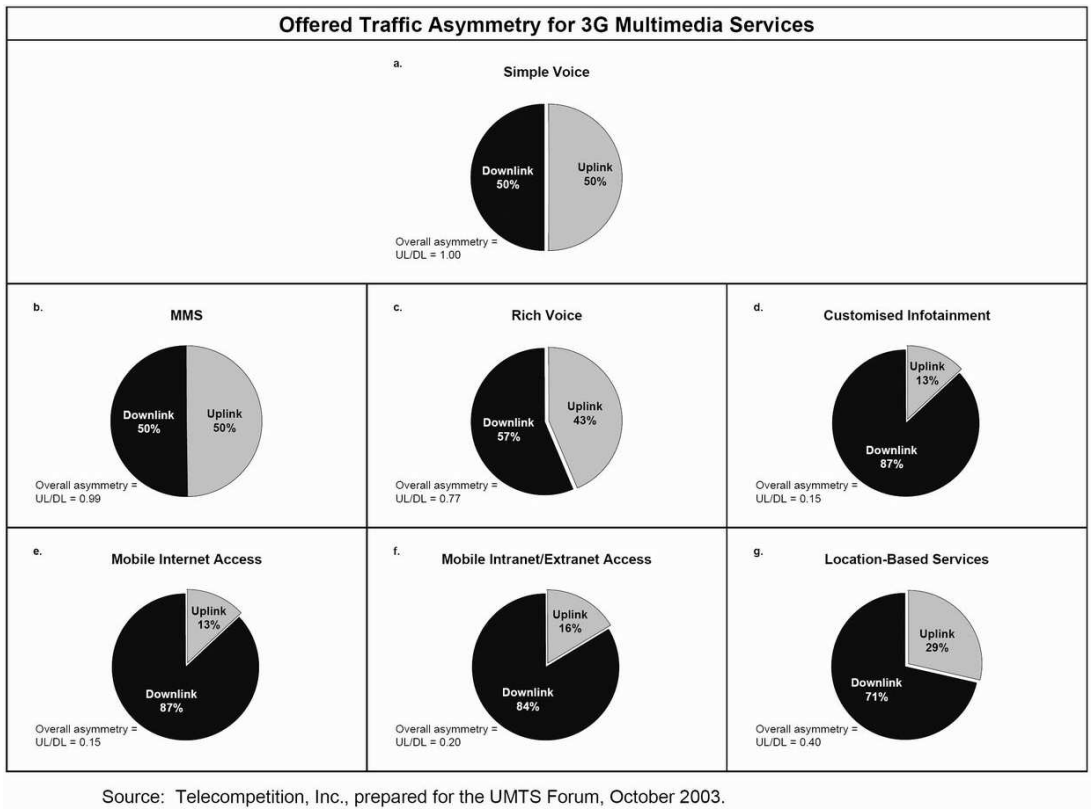

Figure 1.1: A comparison of offered traffic asymmetry by service category.

1.2 Resource Management Problems and Design Objectives

1.2.1 Maximizing System Resource Utilization ($MAXU$)

The $MAXU$ problem is defined as maximizing system bandwidth utilization subject to the constraints on call dropping and blocking probabilities. In multi-service mobile cellular networks, both symmetric traffic service, such as voice, and asymmetric traffic service, such as Internet access, will be supported. Considering asymmetric traffic load between uplink and downlink in such networks, the bandwidth allocated on two links should also be asymmetric in order to improve system bandwidth utilization [7]. It is proved that the system with asymmetric bandwidth allocation always outperforms the symmetric bandwidth allocation in a multi-service environment [8]. In such bandwidth asymmetry networks with dynamic traffic load, the bandwidth allocation on uplink and downlink cannot be adjusted too often since it needs to rearrange all the ongoing calls in a cell [9]. As a result, a bandwidth allocation may be kept in a relatively long time period (maybe one or several hours) after it is determined based on the average traffic load on uplink and downlink.

However, the traffic pattern in the system keeps changing in a relatively small time scale. The mismatch of dynamic traffic load and system bandwidth allocation on uplink and downlink could result in two new problems:

(i) If too many bandwidth-symmetric calls are accepted, more downlink bandwidth resources might be wasted;

(ii) If too many bandwidth-asymmetric calls are accepted, more uplink bandwidth might be wasted.

Both problems may result in a low bandwidth utilization. In order to improve bandwidth utilization in multi-service mobile networks, it is necessary to control the portion of different class calls admitted in the system. As an essential tool for traffic control, CAC has been extensively studied in wired networks and mono-service mobile networks [10–19] in two decades. However, a little work focuses on the research of CAC in bandwidth asymmetry multi-service mobile networks. In order to improve system bandwidth utilization and at the same time guarantee the QoS of different call classes, we need to investigate and design effective CAC scheme for solving the MAXU problem in a bandwidth asymmetry environment. In our work, we address the MAXU problem by identifying and analyzing the main problems that may cause the low bandwidth utilization in multi-service mobile wireless networks and then propose two new CAC schemes to address these problems. Our design objective is to control the admission of Real-Time (RT) calls, which bring symmetric traffic load, and Non-Real-Time (NRT) calls, which bring asymmetric traffic load, to match the bandwidth asymmetry and thus to maximize system resource utilization.

1.2.2 Minimizing System Average Cost (*MINCost*)

Different from the MAXU problem, the MINCost problem concerns about minimizing a linear objective cost function to obtain the minimum average cost. When a call request is accepted or denied, it generates certain revenue or cost accordingly. The revenue/cost may relate with call dropping or blocking probability, system income or other measurements. In traditional mono-service networks, the MINOBJ problem, which is similar to the MINCost problem, was studied in [19] and the guard-channel scheme [16] has been proven be optimal. [20] studied maximizing reward problem and demonstrate the submodularity for the 2-classes problem. The authors investigated the optimal

admission control for the large-capacity system and showed that the trunk reservation policy is optimal when the calls in a system have identical service time. However, it is unrealistic to require that the calls of different classes have identical service time in a multi-service environment. Thus, it is necessary to study the MINCost problem under some general conditions such as considering more call classes and different call durations in a bandwidth asymmetry environment. On the other hand, finding heuristic dynamic CAC scheme is also indispensable since it is time-consuming to find a optimal solution when system state-space is large.

In our work, we regard CAC as a decision process which decides whether or not to accept an arrival call subject to the MINCost problem. In order to minimize the average cost brought by call dropping/blocking of different call classes, we need to find optimal solution for the MINCost problem in the multi-service mobile wireless networks. To the best of our knowledge, few existing work focuses on modeling and analysis of the MINCost CAC problem especially in bandwidth asymmetry networks. We formulate admission control into a Markov decision process (MDP) model and analyze the corresponding value function. Some monotonicity properties of the value function for bandwidth asymmetry mobile networks are identified. These properties suggest that the optimal policy in such environment have a threshold structure and the thresholds may vary with system states. Because of the prohibitively high complexity of computing the values of the thresholds in a system with large state-space, we also propose a heuristic scheme based on our insights obtained in modeling and analysis. Our objective is to find a heuristic scheme which can be readily implemented and the performance of the scheme is expected to close to that obtained by applying the policy from the MDP model in a dynamic traffic load system.

1.2.3 Minimizing New Call Blocking Probabilities with Hard Constraints on Handoff Call Dropping Probabilities (*MINBlock*)

Different from traditional mono-service mobile networks, various RT and NRT call classes are supported in multi-service mobile networks and most of them have handoff attempts. For different call classes, the highest tolerable handoff call dropping probabilities may be different. In order to guarantee handoff call dropping probabilities of different call classes under certain constraints, there is always a tradeoff between the admission of handoff calls and new calls. Since the new

calls cannot be sacrificed too much, how to minimize new call blocking probabilities of various call classes with hard constraints on handoff call dropping probabilities (MINBlock) in a dynamic multi-service mobile networks challenges the existing RM schemes.

In mono-service mobile wireless networks, Limited Fractional Guard Channel (LFGC) scheme has been proposed to address the MINBlock problem [19]. Similar as the Guard Channel scheme [16], the LFGC scheme reserves $C - T$ channels out of total C channels for handoff calls while T channels can be used by both new and handoff calls. When the number of used channels is equal to T, a new call is accepted with probability β. When the number of used channels is greater than T, only handoff calls can be accepted. Although the LFGC scheme has been proved be optimal for the MINBlock problem in mono-service networks [19], it is hard to be extended in multi-service mobile networks. For multi-service mobile networks, the Dual Threshold Bandwidth Reservation scheme (DTBR) as been proposed in [21]. In the DTBR scheme, the total bandwidth of a cell are divided into three parts by two thresholds K_1 and K_2 ($K_1 > K_2$). When the number of channels occupied is less than the threshold K_2, no data calls can be accepted; when the number of channels occupied is more than the threshold K_1, only handoff voice calls can be accepted. The handoff voice call will be dropped if there are no enough free channels. No matter the LFGC scheme or the DTBR scheme, the most critical problem is how to compute the critical parameters, such as T and β used in the LFGC scheme and K_1 and K_2 suggested in the DTBR scheme, in a dynamic traffic load environment. Because of the relatively small state-space of mono-service networks, the LFGC scheme finds the appropriate values of T and β by employing bisection search. The situation becomes more complicated in multi-service environment. In [21], the author did not describe how to compute K_1 and K_2. In our research, we study how to find appropriate thresholds for different call classes in multi-service mobile networks. Our design objective is to find a effective way to compute the thresholds for different call classes and thus guarantee the handoff call dropping probability under predefined constraints. At the same time, the new call blocking probability of different call classes should also be minimized.

In a dynamic traffic load environment, there is no such an RM scheme that is able to guarantee the QoS of different call classes all the time. We need to find a complimentary mechanism to improve the system performance further when the traffic load exceeds the control range of the employed CAC

scheme in a multi-service mobile wireless network. Because of the asymmetric traffic load brought by some data applications, future mobile networks are expected to present significant bandwidth asymmetry between uplink and downlink in order to improve system utilization. In such bandwidth adjustable networks, it is natural to consider bandwidth re-allocation as a complimentary strategy for a CAC scheme.

Compared with CAC, BA in multi-service mobile networks is a relatively new research topic in recent years. In traditional mono-service mobile networks, bandwidth allocated on uplink and downlink in a cell is always same due to the symmetric traffic load on two links. In multi-service mobile networks, the traffic load brought by different call classes exhibits significant asymmetry between uplink and downlink. Some systems have been designed to improve system resource utilization of such asymmetric traffic load networks. In IMT-2000 proposals, two transmission modes–frequency division duplex (FDD) mode and time division duplex (TDD) mode–are suggested [22]. Among these, the CDMA system with TDD mode (CDMA/TDD) is attractive as it can support variable bandwidth asymmetry [22]. That is, bandwidth allocation can be readily adjusted between uplink and downlink. A bandwidth adjustable CDMA/TDD system has been proposed for traffic unbalance networks in [23]. Jeong et al. suggested that the number of time slots in a TDD frame on uplink and downlink of a cell be reset according to the traffic pattern of a cell. For deterministic traffic parameters and mobility characteristics, fixed bandwidth allocation is able to provide an optimal solution for resource allocation problem in mobile networks with bandwidth asymmetry [23, 24]. However, many emerging applications and services with bursty and variable bandwidth requirements call for new treatments of network resource management, in order to satisfy application needs and improve network resource utilization. Furthermore, in multi-service mobile networks, the traffic generated by some applications is time-dependent. For example, the bandwidth asymmetry caused by some data applications could be significantly higher than usual during peak hours in some particular cells. Due to mobility, some users with certain applications may handoff from one cell to another causing the change of traffic load asymmetry in that cell. Therefore, it is necessary to adjust bandwidth allocation on uplink and downlink dynamically. In [22], the authors proved that the system with different time-slot allocations for different cells always outperforms that with the same time-slot allocation, if the time slots on uplink

and downlink are properly allocated. However, there is little known work which addresses how to "properly" allocate bandwidth on uplink and downlink. On the other hand, since bandwidth re-allocation on uplink and downlink may affect all the ongoing calls in a system [23], we should limit bandwidth re-allocation frequency and perform bandwidth re-allocation when it is "necessary". Although it is suggested that a system allocates bandwidth on uplink and downlink according to the traffic load [23, 25], we still do not know when the system needs to adjust the bandwidth on uplink and downlink. To the best of our knowledge, there is no similar work in the literatures that addresses dynamic bandwidth allocations on uplink and downlink in bandwidth asymmetry mobile wireless networks with changing traffic load and pattern. All these reasons motivate us to study *when* and *how* to adjust bandwidth allocation on uplink and downlink in multi-service mobile cellular networks. Our objective is to develop an effective dynamic bandwidth allocation scheme that can adapt to the changing traffic conditions in multi-service mobile networks and collaborates with CAC scheme to provide the desired QoS of different call classes and in the mean time utilize system resources in the best way.

The MAXU problem, the MINCost problem and the MINBlock problem are main optimization problems which are focused on different aspects of call-level resource management. Since the system parameters, such as bandwidth utilization, call blocking probability and average cost, are interrelated, these optimization problems are not totally independent from each other. For the MAXU problem, system bandwidth utilization is the major parameter that needs to be optimized. However, the call dropping/blocking probability, especially that of some high priority call classes, should also be managed at a reasonable low level. When we consider the MINBlock problem, not only handoff call dropping probability should be guaranteed below hard constraint but also new call blocking probability should not be violated too much. Otherwise, system resources will be wasted. It is necessary to consider BA to collaborate with CAC scheme to improve system bandwidth utilization and at the same time solve the MINBlock problem in a traffic asymmetry multi-service mobile network. System cost is a general concept, which may be related with call dropping/blocking probability as that in [26] or totally determined by call prices defined by a service provider. When we design CAC scheme for handling the MINCost problem, we concentrate on the average system cost over a long period of time, which is determined by profit and cost of call

admission and rejection, respectively. Indeed, such profit/cost may be related with call priority, call dropping/blocking probability, call bandwidth requirements, etc. The MINCost problem can be regarded as a more general optimization problem which may combine several system parameters. In this thesis, the proposed CAC and BA schemes designed for each optimization problem could be used independently or cooperatively according to the design goal of system and thus obtaining the optimized system performance.

1.3 Thesis Contributions

In this Ph.D thesis, we address two prominent RM issues, CAC and BA, by considering three main call-level optimization problems in multi-service mobile cellular networks, especially in bandwidth asymmetry environment. Our work can be regarded as an indispensable extension of traditional RM in multi-service mobile wireless networks. The thesis contributions for addressing these RM optimization problems can be summarized as follows:

1. We identify two new problems which lead to low bandwidth utilization in bandwidth asymmetry networks due to the mismatch between traffic load and bandwidth allocation. We propose two dynamic CAC schemes to handle the MAXU problem in multi-service mobile celllular networks. By determining the appropriate admissible regions for RT calls and NRT calls, the proposed CAC schemes are able to prevent RT/NRT calls from overusing uplink/dowlink bandwidth and thus improve system bandwidth utilization. We also employ mathematical analysis to evaluate the proposed schemes and numerical results demonstrate that the results match that obtained from the analytical model well. We also conduct comprehensive experiments in a realistic scenario to study and evaluate the performance of the proposed two CAC schemes. The experiment results demonstrate that the proposed schemes can avoid the low bandwidth utilization problems in a bandwidth asymmetry mobile network while the proposed Scheme 2 can guarantee the dropping probability of handoff NRT calls at a low level without deteriorating the dropping/blocking probability of RT calls when the arrival rate of handoff NRT calls is not high. Compared with some existing multi-service CAC schemes such as Jeon's scheme and the DTBR scheme, Scheme 2 can achieve much higher bandwidth

utilization when traffic load changes in bandwidth asymmetry networks. At the same time, it guarantees the dropping probabilities of handoff RT calls and handoff NRT calls below some reasonably low levels.

2. We consider admission control as a decision process and formulate it as a Markov decision process (MDP) model for addressing the MINCost problem in multi-service mobile cellular networks. We analyze the corresponding value function of the formulated MDP model and extend the properties in [20] in a multi-service mobile networks with bandwidth asymmetry. We prove some monotonicity properties of the optimal admission policy in bandwidth asymmetry mobile networks. These properties indicate that the optimal policy in such environment has a threshold structure and the thresholds of different call classes may vary with system states. Based on our insights obtained in modeling and analysis, we propose a heuristic policy called Call-Rate-based Dynamic Threshold (CRDT) policy. A notable feature of the proposed CRDT policy is that the thresholds of different call classes can be computed readily. Numerical results show that the performance of the proposed CRDT policy is very close to that of the optimal policy obtained from the MDP model and better than that of other two known policies, which are also proposed for bandwidth asymmetry multi-service mobile wireless networks.

3. In order to guarantee handoff call dropping probability of different call classes under some predefined hard constraints, it is critical to determine appropriate thresholds in multi-service mobile networks. We propose a CAC scheme named Distributed Multi-Service Admission Control (DMS-AC) to determine proper threshold for each call class according to the dynamic traffic pattern in a system. In multi-service mobile networks, the admission of a call affects not only handoff call dropping probability of this call class but also that of other call classes. Thus, the situation becomes more complicated to compute the threshold of each call class in multi-service networks than that in mono-service networks. By analyzing the relationship between the admission of different call classes, we decompose all system overload states into the overload states of different call classes and study how the admission of calls from a specific class results in the overload states of other call classes. Based on the system states of local cell and the information from neighboring cells, DMS-AC computes and sets different

thresholds for each call class to prevent the new calls from overusing system resources and control the number of potential handoff calls from the local cell to the neighboring cells. We conduct experiments by considering five different traffic conditions. Numerical results show that DMS-AC is able to guarantee handoff call dropping probabilities of different call classes under certain constraints with the expense of sacrificing more new NRT calls, which has the lowest priority. Since more NRT calls are blocked, downlink bandwidth utilization is also decreased comparing with some existing schemes, which could not guarantee the QoS of handoff calls in the experiment scenarios due to accepting too many NRT calls.

In order to further improve the system performance under dynamic traffic load conditions, we investigate bandwidth allocation for the MINBlock problem. We address two basic BA problems: *when* and *how* to adjust bandwidth allocation to guarantee the QoS of different call classes in a multi-service mobile wireless network, which have not been intensively studied in the literatures so far. The proposed DMS-AC scheme is used as a trigger for activating the BA scheme. When DMS-AC cannot find the feasible thresholds of some call classes or the blocking probabilities of the new calls exceeds some predefined upper bounds, it indicates the QoS requirements of those call classes cannot be guaranteed. In such a situation, the system may adjust the bandwidth allocation on uplink and downlink and re-compute the call admission thresholds until the proper thresholds are determined for each call class in the cell. By investigating the bandwidth re-allocation problem based on DMS-AC scheme, we find that the bandwidth allocated on uplink and downlink should not only be proportional to the traffic load as suggested in [23], but also satisfy certain constraints, which are obtained from the derivations of the thresholds of DMS-AC. We use these constraints to verify the feasibility of a bandwidth allocation and thus let the bandwidth allocation closely collaborate with the admission control to provide a good solution to "when" and "how" to adjust bandwidth allocation in multi-service mobile networks. The MINBlock problem in multi-service mobile networks is solved gracefully by the collaboration of the proposed DMS-AC and BA schemes.

1.4 Thesis Organization

The structure of this thesis is as follows. In Chapter 2, we review some major existing admission control and bandwidth allocation schemes. We analyze their pros and cons in different mobile cellular network conditions. Chapter 4 studies the new problems that result in the low bandwidth utilization in bandwidth asymmetry networks. Two new admission control schemes are proposed to handle the MAXU problem in multi-service mobile networks. In Chapter 5, we investigate the MINCost problem in asymmetric bandwidth allocation mobile networks. By formulating admission control as a Markov discission process, we find the optimal admission control policy should have a threshold structure. We also propose a heuristic policy, the CRDT policy, based on our insights obtained in modeling and analysis. In Chapter 6 and 7, we focus on MINBlock problem in multi-service mobile networks with dynamic traffic load. Chapter 6 proposes a distributed admission control scheme and details the process for computing the thresholds of different call classes. In Chapter 7, we explore the relationship between admission control and bandwidth allocation. Certain constraints for verifying the feasibility of the adjusted bandwidth allocation are established. Based on the proposed CAC scheme, we answer when and how to adjust bandwidth allocation on uplink and downlink and propose an efficient bandwidth allocation scheme. Finally, we present our conclusions and discuss future research directions in Chapter 8.

Chapter 2

Review of Admission Control and Bandwidth Allocation in Mobile Networks

In two decades, resource management is studied, designed and refined continually to satisfy the changing QoS requirements in different network environments. As indispensable components of resources management, call admission control and bandwidth allocation in mobile wireless networks attract many researchers. In this chapter, we review some of the existing call admission control and bandwidth allocation schemes related to our research work and evaluate the performance of some representative ones.

2.1 Call Admission Control in Mobile Networks

As a critical component of resource management, call admission control has been studied extensively and hundreds of CAC-related literatures have been published in recent years [27]. CAC schemes can be classified based on different QoS parameters such as signal quality, call dropping probability, system revenue/cost, packet-level parameters, etc. In this section, we present a review of CAC schemes designed for different optimization problems.

2.1.1 CAC for Minimizing Call Blocking Probability

In mobile wireless networks, handoff call blocking probability (P_h) is one of the critical QoS measurements and should be controlled at a reasonably low level since it is more undesirable to block an ongoing call than a new call. In order to decrease P_h, there is always a tradeoff between the

admission of handoff calls and new calls. That is, in a resource sharing system, the admission of low priority calls, such as new calls, are always limited in order to reserve enough system resources to satisfy the QoS requirements of high priority calls such as handoff calls. How to decrease handoff call blocking probability and at the same time minimize new call blocking probability (P_n) and improve system utilization is a critical problem for the design of call-level admission control scheme.

In order to decrease P_h, Guard Channel (GC) scheme [16] was proposed in traditional mono-service mobile networks. As shown in Fig. 2.1, GC scheme reserves certain amount of channels (guard channels) for handoff calls exclusively. When the number of free channels in the system is less than the amount of guard channels, only handoff calls can be accepted and thus the handoff call dropping probability decreases significantly compared with the Complete Sharing scheme. Since only handoff calls are allowed to use the reserved guard channels, the reduction of P_h comes at the expense of higher P_n. Therefore, the number of guard channels reserved for handoff calls have to be properly computed in order to avoid sacrificing more new calls unnecessarily. An enhanced guard channel scheme, Limited Fractional Guard Channel (LFGC) scheme has been proposed in [19] and proven to be optimal for minimizing P_n with a hard constraint on P_h (MINBLOCK problem) and minimizing the number of needed channels with a hard constraint on both P_h and P_n (MINC problem). Similar to GC scheme, as shown in Fig. 2.2, LFGC reserves $C - T$ channels out of total C channels as guard channels for handoff calls. The new calls can be accepted without any limitations when the number of used channels is less than T. If the number of used channels is equal to T, the new calls are accepted with probability β. The new calls cannot be accepted if only guard channels are available. Since the values of T and β decide the performance of LFGC, the authors illustrated how to compute T and β comprehensively. However, even if all the critical parameters are computed appropriately based on the current traffic load in the system, it does not mean P_h can be guaranteed since the traffic load in the system is variable.

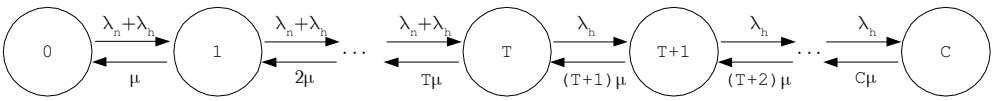

Figure 2.1: State-transition diagram of GC scheme.

GC scheme was also extended to multi-service mobile networks as in [21]. The authors proposed a Double Threshold Bandwidth Reservation (DTBR) scheme. In DTBR scheme, the channels of

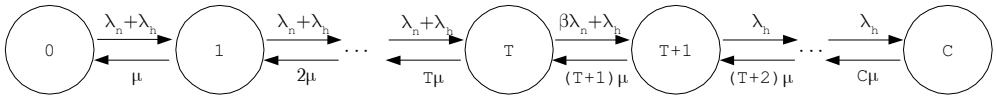

Figure 2.2: State-transition diagram of LFGC scheme.

each cell are divided into three parts by two thresholds K_1 and K_2. When the number of channels occupied is less than the threshold K_2, both data and voice traffic can be admitted into the system. When the number of channels occupied is over K_2, no data traffic is admitted. When the number of channels occupied is more than the threshold K_1, only handoff voice calls can be accepted. The handoff voice call will be dropped only if there is no available channel. The DTBR is still a static scheme and it does not consider dynamic traffic load condition. Moreover, the authors did not illustrated how to find appropriate values of two critical parameters K_1 and K_2.

In [28], the authors compared the performance of six different CAC schemes which include bandwidth reservation scheme, i.e., GC scheme. They found that none of the algorithms can significantly outperform the reservation scheme when the system traffic load is known. However, it is unrealistic to expect to know dynamic system traffic load accurately beforehand. It is necessary to design dynamic CAC schemes to improve the system performance in varying traffic load networks. In [26], the authors suggested re-computing T and β in LFGC based on the estimation of call arrival rate when the current setting cannot guarantee the QoS requirements of handoff and new calls. However, the authors did not present the estimating process in detail. In [18, 29], the authors suggested reserving bandwidth in all neighboring cells for each accepted new/handoff call. If bandwidth reservation succeeds in all neighboring cells, the arrival call is accepted. Otherwise, it is rejected. It is obviously inefficient since the handoff mobile host will move to only one of the neighboring cells ultimately and all the pre-reserved resources in other neighboring cells will be wasted. Therefore, *how much* bandwidth should be dynamically reserved in *which* cell is the critical issue for dynamic bandwidth-reservation-based CAC.

So far, a plenty of dynamic bandwidth-reservation-based CAC schemes have been proposed [8, 30–54]. Normally, such CAC schemes include two major components: 1) Estimating or forecasting users' mobility, such as handoff target cell, the dwell time of a user staying at the local cell, etc. 2) Dynamically reserving bandwidth or adjusting the reserved bandwidth according to the estimation. In [30], Levine *et al.* proposed the shadow cluster mechanism to dynamically reserve

bandwidth in potential target cells based on the observation that every mobile terminal with an active wireless connection exerts an influence upon the cells (and their base stations) in the vicinity of its current location and along its direction of travel. A shadow cluster system can be viewed as a message system where mobile terminals inform the base stations in their neighborhood about their requirements, position, and movement parameters. With these information, base stations predict future demands, reserve resources accordingly. The coverage of a shadow cluster for a given active mobile mainly consists of the cell where the mobile is currently present (i.e., the center of the shadow cluster) and all its adjacent cells along the direction of travel. This area changes when the mobile call hands-off to other cells, thus a tentative shadow cluster needs to be implemented for every new call as well as every handoff call. Simulations show that the shadow cluster mechanism is able to reduce the percentage of dropped calls in a controlled fashion. The efficiency of this scheme depends on the accuracy of prediction of the future mobile movement, which makes it most suitable for a strong directional environment such as the highway. Most Likely Cluster (MLC) model was proposed in [55, 56]. The MLC model considers that the cells that are situated along a mobile user's direction have higher directional probabilities and are more likely to be visited than those that are situated outside of this direction. Bandwidth resources required by each handoff request are reserved in each MLC cell during a certain estimated time interval. In [38,40,43,53], the authors suggested predicting the mobile position based on some positioning technologies such as Global Positioning System (GPS) or digital road maps to improve the accuracy of the estimation. All these schemes are per-call based and system needs to trace the mobility of individual mobile user. The bandwidth is pre-reserved for the handoff calls in the predicted target cell. However, keeping track of each mobile's mobility over time is too costly and it is inapplicable when the number of mobile users is large.

One of the straightforward methods to adjust the reserved bandwidth for handoff calls is according to the estimate of the handoff call arrival rate. In [31], the reserved guard bandwidth for handoff calls is adjusted according to the current estimate of the instantaneous handoff call arrival rate so as to keep the handoff call blocking probability close to the objective while not deteriorating the new call blocking significantly. In [8], the authors proposed dynamic guard channel scheme for multi-class services in bandwidth asymmetry networks. By estimating the handoff call arrival rates

of different call classes in a certain period of time, the proposed scheme dynamically reserve different guard channels on uplink and downlink separately for different call classes. As pointed in [31], the instantaneous handoff call arrival rate at a test cell for the next estimation interval depends on the handoff initiation process, the number of active mobile terminals (MT) with ongoing calls in the neighboring cells, the mobility patterns of the active MTs in terms of speed and direction during the estimation interval, the sizes of the cells currently resided by the active MTs, and the remaining call durations of the ongoing calls. All these information needs to be measured or exchanged between neighboring cells. Since the existing dynamic guard channel admission schemes are developed under the assumption of perfect estimation, it may not be possible in a highly non-stationary environment and thus resulting in failures to maintain targeted blocking/dropping probabilities. [54] presents the fairly adjusted multi-mode dynamic guard bandwidth scheme, which is a dynamic guard-based scheme over Code Division Multiple Access (CDMA) systems with predictive adaptation control to adapt interference-based guard-loading-limits under non-stationary call arrival condition; and reactive adaptation control to counteract estimation errors of arrival rate. When the predictive adaptation control policy mode is not able to maintain long-term call blocking or dropping targets due to estimation errors, this will trigger reactive adaptation control policy modes that include temporary blocking (preemption) of one or more lower priority classes subject to fairness constraints to ensure that low priority classes are not preempted at all costs during estimation error recovery.

In order to simplify the estimation, it has been proposed to estimate user's mobility based on aggregate handoff history in [33, 41, 57]. In [33, 57], the proposed schemes are based on the assumption that handoff behavior of a mobile user will be probabilistically similar to the mobile users which came from the same previous cell and are now residing in the current cell. The proposed schemes decide how much bandwidth should be reserved in each neighboring cell by utilizing an aggregate history of observations based on caching the mobile information of handoff calls, such as the identifier of the target neighboring cell and the sojourn time of the mobile user in the current cell. The information that needs to be measured or exchanged between adjacent cells is reduced. In [41], the authors presented two methods that use local information alone to predict the resource demands of handoff calls and determine resource reservation levels for future handoff calls in mobile wireless networks. Their basic idea is to use the current and the past values of required bandwidth

resources for handoff calls in the cell to predict the future values directly. Since the current and the past bandwidth required by handoff calls can be measured by a base station locally, prediction can be performed by utilizing local information only.

As a critical QoS measurement, call blocking probability is also employed as a criterion for dynamic bandwidth reservation [36, 42, 44–48, 51]. In [36], handoff call blocking probabilities of different call classes are used as the admission criteria in multiple-class mobile wireless networks. The pre-reserved bandwidth for handoff calls of each call class are computed based on traffic load measurements in order to keep the relative call blocking probability below a specific value. In [42, 44–48, 51], the reserved bandwidth for handoff calls is adjusted according to the measured or estimated handoff call blocking probability. When handoff calls are blocked or P_h is greater than a certain threshold, the reserved guard bandwidth is increased. On the other hand, the guard bandwidth is decreased when P_h is less than a specific threshold. Such dynamic bandwidth reservation schemes is working in a reactive manner since the adjustment of reserved bandwidth is based on the "past" information, i.e., the measurement of P_h. If the past information cannot reflect the future trend, the adjustment may be unnecessary or inaccurate.

Queueing handoff or new call requests, with or without reserving bandwidth for handoff calls, is another method to reduce the call dropping/blocking probability. Fig. 2.3 shows the change of the power of a handoff mobile user between two cells. When the signal power is below the handoff

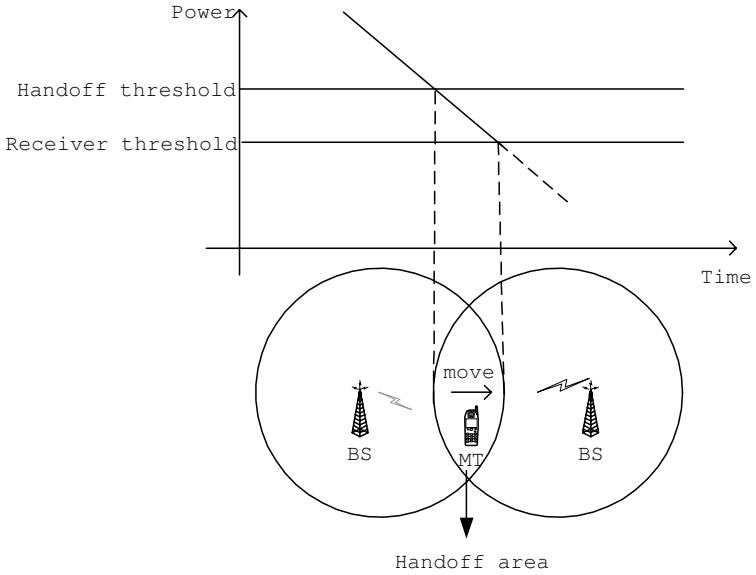

Figure 2.3: Signal power of handoff MT and handoff area.

threshold, the mobile user enters the handoff area and starts handoff process. The handoff call will be dropped if there are no enough free channels in the target cell for the handoff call before the signal power less than the receiver threshold. Since the mobile user spends certain period of time in the handoff area, it is reasonable to hold the unacceptable handoff requests in the waiting queue. In [16,58], based on the bandwidth reservation scheme, the authors used a waiting queue to hold the unacceptable handoff requests and thus reduce P_h further. Similarly, in [59], the handoff calls are queued until enough channels are available. Moreover, the queue is dynamically reordered according to the measurements of the users' power level. The queued user who has lower signal power will be served first. In [60, 61], new calls are queued until more free channels are available except the guard channels. By employing queue to hold the call requests, such Q-based bandwidth reservation schemes not only minimize the handoff call blocking probability but also increase the total carried traffic of the system. Since reserving guard channels may decrease system bandwidth utilization and block more new calls sometimes, a pure buffer-based CAC scheme was proposed in [62]. The authors suggested buffering both handoff and new calls when no free channels are available. There is no guard channel for handoff calls but handoff calls has higher priority than new calls. The handoff calls are always buffered at the priori position of new calls. Although there are no guard channels for handoff calls, the proposed scheme does not violate P_h to much and at the same time it decreases the new call blocking probability. In [63–65], handoff calls arriving at the base station (BS) are queued in two separate and finite queues based on their priority if all channels are busy. By using receiver signal strength, base station can estimate the remaining time of the users and the user who has least remaining time has highest priority and will be served first. With the increase of mobile users, the size of cells will be reduced in order to support more users [3, 4] and thus the time of handoff users staying in the handoff area is also decreased. The unacceptable handoff requests cannot be hold long time enough before it is dropped. Therefore, the Q-based admission control may be ineffective in the picocell mobile wireless networks.

Different from bandwidth reservation CAC, threshold-type CAC scheme limits the number of the admission of calls by using threshold without reserving bandwidth. Distributed CAC scheme (DCA) is a threshold style CAC scheme proposed in [17]. By using threshold to limit the admission of new calls, DCA guarantees the overload probability of the local cell and all the neighboring

cells under the upper bound of P_h and thus satisfys the QoS requirements of handoff calls. Since the accepted new calls will handoff to other cells in "future" with certain probability, limiting the admission of new calls reduces the number of handoff calls from the local cell to the neighboring cells and thus reduce the handoff call blocking probability. However, it is unacceptable to sacrifice the new calls too much and it is necessary to balance the admission of handoff calls and new calls. In [66, 67], the authors extended DCA by considering P_h and P_n together. They revised the call admission conditions employed in [17] and required admission of every new call cannot let the overload probability of the local cell and neighboring cells exceed a predefined P_{QoS}, which is the linear combination of P_h and P_n. The simulation results show that the system capacity gain was improved significantly. More modified distributed admission control schemes were proposed in [68–71]. In [71], the authors estimated the time-dependent dropping probability in a cell and the derivation is based on the solution to the evolution equation of the occupancy distribution, which greatly improves over the Gaussian approximation used in [17]. The multiple handoffs scenario is also considered and the estimation of the call dropping probability is based on the call transition probabilities between nearest as well as second and third nearest neighboring cells. The call dropping probability yields an expression for the acceptance ratio, which is the maximum fraction of new calls to be admitted into cell in the coming control period. By stochastically accepting each new call with certain probability, the proposed scheme avoids a sudden overload of the network at the beginning of the control period during congestion, leading to more effective and stable control.

2.1.2 CAC for Maximizing System Revenue

In a mobile wireless network, call's admission or rejection will bring certain revenue or cost to the system. The problem is to find a control policy in order to accept/reject the arriving calls as a function of the current system state in order to maximize the average revenue. This problem has been studied in [72] by using stochastic knapsack model. The classical knapsack problem is how to pack a knapsack of integer volume F with objects from K different classes in order to maximize profit. In a telecommunications system, a variety of traffic types (e.g., voice, video, data, etc.) are supported and they share the limited bandwidth resource of the system. Different traffic

types may have different bandwidth requirements and holding-time distributions. By modeling the total system bandwidth as the knapsack, the traffic types as the object and the bandwidth requirements as the object volumes, the problem of optimally accepting calls in order to maximize average revenue is equivalent to the stochastic knapsack problem [72]. Since the optimal policy is in general complicated especially when system state-space is large, the authors searched high-performing policies with a simple structure in the coordinate convex policies. Although it may exclude the best optimal one, the coordinate convex policies can provide product-form solutions for the associated equilibrium state probabilities, from which all of the performance measures of interest can be determined. The authors proved that the optimal policy is of the threshold type for a wide range of parameters when the number of traffic type is two. This conclusion has been validated in [19] by considering minimizing a objective function of two blocking probabilities. In [19], the authors considered the problem of finding an admission control policy that minimizes a linear objective function of the new and handoff call blocking probabilities (MINOBJ problem). Indeed, the penalty or cost of the system can be associated with the call blocking probabilities directly and MINOBJ problem is equivalent to the cost minimization problem. The authors proved that GC scheme [16] is optimal for the MINOBJ problem. More research work of the CAC for optimizing system revenue can be found in [20, 73–82]. In [73], the authors considered the networks with a variety of traffic classes (e.g., data, voice, video, etc.) sharing certain bandwidth resources, each of which has its own traffic requirement and reward function. The problem of dynamically allocating the capacity of each circuit among the traffic classes is addressed in the literature. As an optimal allocation policy is extremely hard to find, the authors applied a different methodology by which they found the bound of the optimal expected reward, and proposed a specific threshold policy-the Restricted Complete Sharing (RCS) scheme-that yields a reward sufficiently close to this bound. In [74], the authors discussed the special case of [73] where the service rates and the reward per acceptance do not depend on the customers' class (under certain constraints), and the optimality of a trunk reservation policy is established. [77, 81] studied CAC for revenue maximization subject to several predetermined call-level and/or packet level QoS constraints by using reinforcement learning algorithm [83]. In [78–80], CAC for maximizing network revenue integrated with call pricing is investigated. In [20], Altman et al. studied reward maximizing problem of multi-class CAC in a

resources sharing system. The authors demonstrated the sub-modularity for the 2-classes problem and established some properties of optimal policies for such resources-sharing system. Moreover, the authors formulated CAC problem in a resource-sharing system into a fluid model and studied the optimal admission control for the large-capacity system. They showed that the trunk reservation policy is optimal when the calls in the system have identical service time. When the call duration does not depend on the call class, the system model is reduced to a one-dimensional state-space model. Indeed, such a one-dimensional state-space model has been studied in [19] when the number of call classes is two and a similar conclusion has been drawn, which indicated the guard channel policy is optimal. However, it is unrealistic to require that the calls of different classes have identical service time in a multi-service environment. This limitation is loosed in [82] by employing a new assumption that the decision maker of the call admission control knows the duration of the call when the call arrives.

2.2 Bandwidth Re-allocation for Bandwidth Asymmetry Mobile Networks

Compared with the "long" history of the research of CAC, it is a new research topic to study the bandwidth re-allocation between uplink and downlink in a bandwidth asymmetry mobile wireless network. Since only simple voice service is supported in the traditional mobile wireless networks, it is not necessary to consider asymmetric traffic in the traditional mono-service networks. With more and more data applications supported in mobile wireless networks, it is widely accepted that future mobile wireless networks will present distinct traffic asymmetry between uplink and downlink. In order to improve system utilization of such mobile wireless networks, CDMA/TDD system has been proposed [9, 84]. In the proposed system of [9], the number of uplink time slots in a TDD frame differs from that of downlink. Moreover, the difference can be reset by the network operator according to the traffic pattern. In [22], the authors compared two different time slot allocation strategies, same time-slot allocation (SA) and different time-slot allocation (DA), for CDMA/TDD systems. SA strategy requires all cells within a service area have same time slot allocation. In DA strategy, the time slot allocation may be different from cell to cell according to the level of traffic

asymmetry of each cell. In multi-service mobile wireless networks, the level of traffic asymmetry may be significantly different from cell to cell. In this case, the slot allocation should be varied cell by cell to maximize the frequency utilization. However, DA strategy will result in crossed-slot interference between two adjacent cells. For example, let us consider two adjacent cells, cell A and cell B, and DA strategy is used. The time slot allocation in cell A and cell B is shown in Fig. 2.4 (a). From this figure we can find that several time slots are allocated to uplink (UL) in cell A while

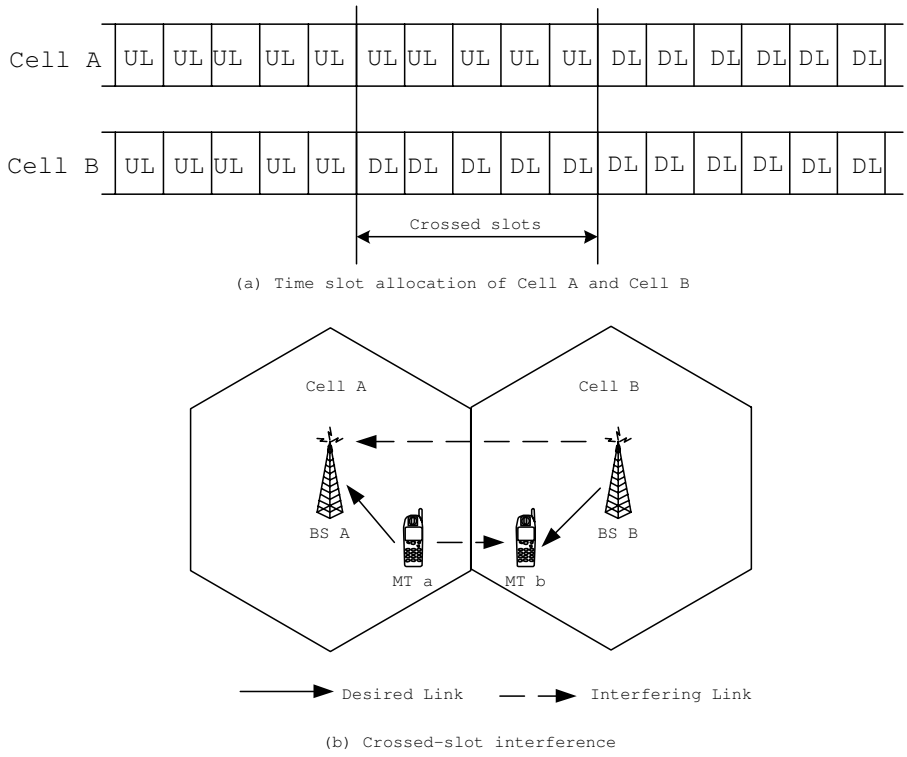

Figure 2.4: Crossed-slot interference.

some slots are allocated to downlink (DL) in cell B during a same time period. If MT a (b) and base station (BS) B (A) transmit signals during the same time slots as shown in Fig. 2.4 (b), the uplink (downlink) channel in a cell will be interfered by the downlink (uplink) of the adjacent cell, which results in capacity degradation. This phenomenon is called crossed-slot interference. In [22], the author computed the capacity of CDMA/TDD systems with DA strategy and compared it with that of SA strategy. They found that the DA strategy always outperforms SA strategy if the TDD slots are properly allocated. To the best of our knowledge, current research work of bandwidth re-allocation (time slot re-allocation) commonly focuses on crossed-slot interference problem of DA

strategy [85–87]. Since the signal interference problem exceeds the research area of this thesis, we do not illustrate it further. Our research work about BA in bandwidth asymmetry mobile networks focus on two main problems: 1) When to adjust the bandwidth allocations on uplink and downlink; 2) How to find appropriate value of bandwidth allocated on uplink and downlink. Although we cannot find sufficient related work about these two problems, it makes our work more meaningful for multi-service mobile wireless networks.

2.3 Summary

In this chapter, we provide an overview of call admission control and bandwidth allocation mechanisms for mobile wireless networks. We described a lot of schemes with their main features for each of the class. This work may provide some valuable hints for the further design and development of call admission control and bandwidth allocation mechanisms.

Chapter 3

System and Traffic Model

3.1 System Model

We consider a multi-service mobile cellular network, where traffic load and bandwidth allocation on uplink and downlink could be asymmetric. The system resource is bandwidth or channels, which can be regarded as the bandwidth units. A representative system that support asymmetric bandwidth allocation is CDMA/TDD system or TD-CDMA system and the industry standard of such system is WCDMA-TDD mode. One of the most significant benefits of TDD (Time Division Duplex) is that TDD supports variable asymmetry, which means an operator can dictate how much capacity is allocated to downlink versus uplink. In such CDMA/TDD system, the system resources can be understood from two aspects. The first is the time slots in a frame at TDD level and the second is the tolerable total received signal power in each slot at CDMA level. Since CDMA is an interference-limited system, total received power should be restricted at a proper level in order to maintain adequate transmission quality. Given a system, the "tolerable total received power" from all mobile users can be interpreted as a "tolerable aggregate data rate" in the system, which means when the load equals tolerable aggregate data rate, the bit error rate is maintained under a certain value which is defined as system design specification [9]. The "tolerable aggregate data rate" in the system can also be interpreted as the system maximum bandwidth. We assume that the total bandwidth of each cell in the system is time-invariant. In practice, the capacity of a cell

(the total resource in the cell) may vary with the traffic load of home and neighboring cells because of interferences. Since we investigate how the system asymmetry and traffic asymmetry affect the system performances, we assume that the total resources in a cell are time invariant for the ease of illustration. In such multi-service mobile networks, both traffic load and bandwidth allocation on uplink and downlink of each cell in the system could be asymmetric. We use B_u and B_d to denote the uplink and downlink bandwidth of a cell, respectively. We define system *asymmetry factor*, denoted by Γ_s, as

$$\Gamma_s = \frac{total\ downlink\ bandwidth}{total\ uplink\ bandwidth} = \frac{B_d}{B_d}\,, \tag{3.1}$$

which is used to represent the degree of system bandwidth asymmetry. Due to different traffic pattern, we assume that the bandwidth allocation could be different from cell to cell, which means Γ_s of different cells may be different.

In our research, we first consider the design of resource management scheme in a single-cell system in Chapter 4 and 5, where only one cell is considered as shown in Figure 3.1 (a). Calls from different call classes may be generated in the cell or handoff from other cells. The call leaves the cell by terminating the call or handing-off to neighboring cells. The calls that are originally generated in a cell are called new calls, while the calls handoff between cells are called handoff calls. Then we extend the research to multiple cells in Chapter 6 and 7 as shown in Figure 3.1 (b) and (c). In Chapter 6, we consider a simple two-cell system as shown in Figure 3.1 (b), where two cells are named as C_l and C_r. Calls may be generated in C_l or C_r and also can handoff between these two cells. Call termination or handing off to the cells out of the two-cell system are regarded as leaving the system. We also extend the proposed CAC and BA schemes to multi-cell system as shown in Figure 3.1 (c), which is composed by seven cells.

The proposed admission control scheme and bandwidth allocation are implemented at base station of each cell. The base station acts as admission controller. When it receives a call connection request from a mobile user, it decides whether or not to accept the call according to certain admission control scheme. If the call cannot be accepted, the call request will be rejected immediately. In our work, we do not consider request buffer for holding the unacceptable call requests. In the design of resource management schemes in multi-cell system in Chapter 6 and 7, we assume that some system state information, such as the number of calls of a specific call class and call handoff

probability etc., can be exchanged between two neighboring cells. The exchange frequency could be determined by the proposed resource management or other requirements. We may increase the exchange frequency in a dynamic traffic load environment in order to accurately estimate some critical parameters, such as mean call arrival rate, in the design of resource management schemes.

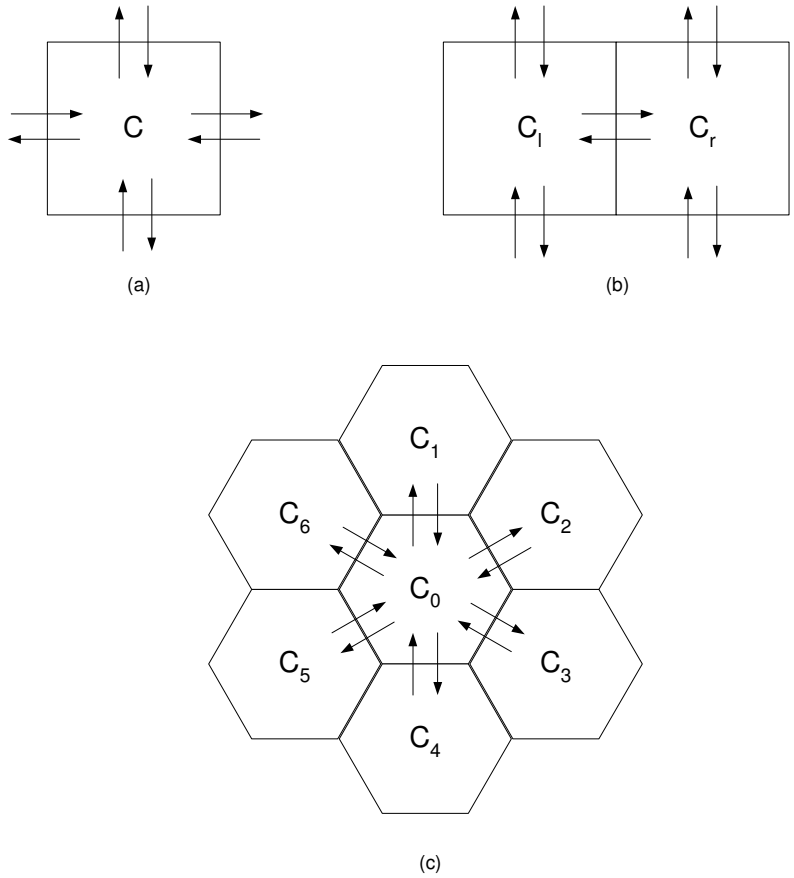

Figure 3.1: Single cell, two-cell and multi-cell network model

3.2 Traffic Model

We consider the system that supports multiple call classes. The calls from different call classes share certain bandwidth resources of a cell. Calls from the same class have the same bandwidth requirements on uplink and downlink. We assume that a call of class i demands b_i^u and b_i^d units of bandwidth on uplink and downlink, respectively. In order to simplify our analysis, we assume that the arrival of class i calls is according to Poisson distribution with mean λ_i. Indeed, this assumption is not compulsory, especially in the design of RM schemes in our research. We may

periodically estimate the mean call arrival rate by scaling. We also assume that the call connection holding time of the call class i is exponentially distributed with mean $1/\mu_i$. For different call classes, mean call arrival rate and mean call service time could be different. The traffic load brought by a class i call can be expressed as λ_i/μ_i. The QoS requirements in terms of the highest tolerable call dropping probability and the upper bound of call blocking probability of different call classes could be different. In our research, the call priority is determined by the highest tolerable call dropping probability (for handoff calls) and the upper bound of call blocking probability (for new calls). Since it is more undesirable to block an ongoing call, handoff calls always have higher priority than new calls. The handoff calls in a system may belong to different call classes and the call class which has the most strict requirement on the highest tolerable call dropping probability is assigned the highest priority. For the new call, the call class which has the lowest upper bound of call blocking probability has the highest priority. In Chapter 4 and 5, we classified all arrival calls to two call classes: Real-time call and Non-real-time call. The RT call such as voice call or video supported voice call requires the some bandwidth on uplink and downlink. The NRT call such as Internet access may bring asymmetric traffic load between two links. We use NRT asymmetry factor to represent the asymmetry degree of NRT calls, which is denoted by Γ_{NRT}. Γ_{NRT} can be expressed as

$$\Gamma_{NRT} = \frac{B^d_{NRT}}{B^u_{NRT}} , \tag{3.2}$$

where B^d_{NRT} and B^u_{NRT} denote the total uplink bandwidth and downlink bandwidth used by NRT calls in the system, respectively. We also further classify NRT calls to N subclasses, which have different bandwidth requirement and QoS requirements. Let n_i denote the number of the NRT calls in the system. Γ_{NRT} can be expressed as

$$\Gamma_{NRT} = \frac{\sum_{i=0}^{N-1} n_i b^d_i}{\sum_{i=0}^{N-1} n_i b^u_i} \tag{3.3}$$

or

$$\Gamma_{NRT} = \frac{\sum_{i=0}^{N-1} a_i b^d_i}{\sum_{i=0}^{N-1} a_i b^u_i} , \tag{3.4}$$

where a_i is defined as

$$a_i = \frac{n_i}{\sum_{i=0}^{N-1} n_i}, \tag{3.5}$$

which is the ratio of class i NRT calls over all NRT calls. In a large time scale, a_i can be regarded as statistically fixed and thus Γ_{NRT} too. In Chapter 6 and 7, we consider a more general traffic model. We assume that the system can support M call classes. Besides mean call arrival rate and mean service time, the calls of the same class have the same handoff probability, which is defined as the probability that a call hands-off from one cell to another. The handoff probability of different directions may be different. For example, in two-cell system shown in Figure 3.1 (b), the handoff probability from C_r to C_l could be different from that from C_l to C_r. The handoff probability of a specific call class can be estimated according to user movement patterns such as moving speed, moving direction etc. Some researches have been conducted to estimate these probabilities [17, 30, 53] and we do not discuss it further in this chapter.

For call level resource management design, there are four major QoS measurements: handoff call dropping probability, new call blocking probability, average system cost and system bandwidth utilization. In the following four chapters, we will focus on the optimization of different QoS measures. In Chapter 4, we concentrate on maximization of system bandwidth utilization. Chapter 4 is focused on minimization of system average cost. Chapter 6 and 7 are focused on minimizing call blocking probability.

Chapter 4

Maximizing Resource Utilization in Bandwidth Asymmetry Mobile Networks

In multi-service mobile cellular networks, asymmetric bandwidth allocation has been proposed to satisfy the requirements of asymmetric traffic load introduced by some data applications. However, it is difficult to promptly adjust bandwidth allocation on uplink and downlink according to the dynamics of traffic load. Inappropriate CAC schemes may admit superfluous RT calls or NRT calls and thus lead to low bandwidth utilization in such bandwidth asymmetry networks. In this chapter, we propose and evaluate two new CAC schemes to address the problems caused by the mismatch of bandwidth allocation and traffic changing in multi-service mobile networks with bandwidth asymmetry. By determining admissible regions for RT calls and NRT calls, the proposed schemes prevent the calls of a specific class from overusing bandwidth resources. The design objective is to improve bandwidth utilization while retaining the call dropping probabilities of handoff RT and NRT calls at a reasonable low level. Mathematical analysis and simulation experiments are employed to study and compare the performance of the proposed schemes with that of the existing schemes. Numerical results show that the proposed schemes can achieve better performance in terms of call dropping probability and bandwidth utilization compared with some existing schemes, even those performing well in bandwidth asymmetry mobile cellular networks.

4.1 Introduction

One of the distinctive features of future mobile cellular networks is that multi-services such as voice, data, video, and multimedia will be supported over wireless infrastructures [5, 88]. Unlike the traditional voice communication, the demands for bandwidth resources on uplink and downlink could be asymmetric for many multi-service applications. For example, Internet access, which is a representative service supported by the next generation mobile networks, exhibits evident asymmetric bandwidth demands on uplink and downlink. For some client-server applications, the traffic on uplink is usually much lighter than that on downlink where data, voice or even video traffic can be carried. With the rapid growth of data traffic, future mobile cellular networks are expected to present distinctive traffic asymmetry between uplink and downlink [5].

In multi-service mobile cellular networks with asymmetric traffic load, if we allocate equal bandwidth on both uplink and downlink, the system capacity could be limited by downlink [9]. This results in bandwidth waste and resource utilization degradation. The resource utilization can be improved by allocating different bandwidth on uplink and downlink [7]. It is proved that the system with asymmetric bandwidth allocation will outperform that with symmetric bandwidth allocation in traffic asymmetry environment. One example of the systems that support asymmetric bandwidth allocation is CDMA/TDD (Time Division Duplex) system or the TD-CDMA system and the industry standard of such system is WCDMA-TDD mode. One of the most significant benefits of TDD is that TDD supports variable asymmetry, which means an operator can dictate how much capacity is allocated to downlink versus uplink. Some resource allocation strategies have been proposed [9, 89]. However, such strategies cannot be implemented readily since they need to rearrange all the ongoing calls in a cell [9]. Since the traffic pattern in a system may keep changing in a relatively small time scale, it is difficult to promptly adjust the bandwidth allocation on uplink and downlink accordingly. Two new problems may arise under such circumstance: (1) if too many bandwidth-symmetric calls are accepted, more downlink bandwidth resources might be wasted; (2) if too many bandwidth-asymmetric calls are accepted, some uplink bandwidth might be wasted. Both problems may result in a low bandwidth utilization. Therefore, an appropriate CAC policy is essential for such mobile wireless networks to maximize the bandwidth utilization.

Although many CAC schemes are proposed for the multi-service mobile networks, few existing

CAC schemes consider the asymmetric traffic load brought by multi-class services, which is one of the most notable features in future mobile networks [5]. In order to achieve good performance in such a system with asymmetric traffic load, Jeon et al. proposed a multi-guard-channel scheme [8]. In Jeon's scheme, the size of guard bandwidth for each traffic class on uplink and downlink is determined separately. The reserved bandwidth is proportional to the call arrival rate, the mean call duration and the required bandwidth of each call class. This scheme tries to reserve optimal guard bandwidth for each call class by estimating call arrival rate of each class. Jeon's scheme achieves good performance in terms of handoff call dropping probability and new call blocking probability. The authors also proved that the proposed scheme can achieve better bandwidth utilization in asymmetric bandwidth allocation environment than that in the symmetric environment. Since the scheme does not consider the limitation introduced by bandwidth asymmetry, it cannot avoid the low bandwidth utilization problem in bandwidth asymmetry networks.

Because of user's handoff in addition to the bandwidth asymmetry between uplink and downlink, CAC becomes more complicated in multi-service mobile cellular networks. In this chapter, we address how to maximize system utilization and at the same time guarantee the QoS requirements of different call classes in bandwidth asymmetry mobile networks. We first identify and analyze the main problems that may cause low bandwidth utilization in such multi-service mobile networks and then propose two new CAC schemes to address the problems. Our design objective is to control the admission of RT calls and NRT calls to match the bandwidth asymmetry and thus to maximize network resource utilization. In the proposed Scheme 1, the bandwidth that can be used by RT calls and NRT calls is determined by setting the admissible region for NRT calls. This admissible region is also used as a threshold for both handoff and new NRT calls. When total bandwidth used by NRT calls reaches the threshold, both handoff NRT calls and new NRT calls will be blocked. Since handoff NRT calls have higher priority than the new NRT calls, in the proposed Scheme 2, we modify Scheme 1 and set threshold for new NRT calls only. When the bandwidth used by NRT calls reaches the threshold, only new NRT calls will be blocked. From the numerical results, we find that both Scheme 1 and Scheme 2 can achieve good bandwidth utilization in such environment. However, the proposed Scheme 2 can achieve much lower call dropping probability of handoff NRT calls than that of Scheme 1 by making a tradeoff between handoff NRT calls and new NRT calls. Compared

with some existing CAC schemes, Scheme 2 exhibits its better performance in terms of bandwidth utilization and call dropping probability in multi-service mobile wireless networks. Moreover, the proposed schemes have a lower implementation complexity compared with some existing schemes which also implement asymmetric bandwidth allocation, such as Jeon's scheme [8].

The rest of this chapter is organized as follows. In Section 4.2, we identify and analyze the problems caused by bandwidth asymmetry in multi-service mobile cellular networks and elaborate on the proposed CAC schemes. In Section 4.3, we present the performance analysis of the proposed schemes by using Markov model. In Section 4.4, we present the numerical results with discussions and compare the performance of the proposed schemes with that of some existing schemes. Finally, we conclude this chapter in Section 4.5.

4.2 Bandwidth Reservation Based CAC Schemes

4.2.1 Problem Formulation

Let us consider a multi-service mobile cellular network, where two types of calls, RT call and NRT call, are supported. RT call such as voice call requires the same bandwidth on uplink and downlink while NRT call such as web browsing requires more downlink bandwidth. Both RT call and NRT call may have handoff attempts. Since it is more undesirable to block a handoff call than a new call, handoff calls have higher priority than new calls. Given that NRT calls can tolerate much longer delay than RT calls, RT calls should have higher priority than NRT calls. We arrange the priorities of different call class in descending order as follows: handoff RT call, handoff NRT call, new RT call and new NRT call.

We consider the system at the steady state with heavy traffic load. If no bandwidth is wasted, the uplink bandwidth and the downlink bandwidth used by RT calls and NRT calls should satisfy:

$$B^u_{RT} + B^u_{NRT} = B_u \tag{4.1}$$

$$B^d_{RT} + B^d_{NRT} = B_d \ , \tag{4.2}$$

where B^u_{RT} (B^u_{NRT}) and B^d_{RT} (B^d_{NRT}) denote the bandwidth used by RT (NRT) calls on uplink

and downlink, respectively. The total uplink bandwidth and downlink bandwidth are denoted by B_u and B_d, respectively. Combining (4.2) and (4.1) yields

$$B^d_{NRT} - B^u_{NRT} = B_d - B_u \tag{4.3}$$

or

$$B^u_{NRT} = \frac{\frac{B_d}{B_u} - 1}{\frac{B^d_{NRT}}{B^u_{NRT}} - 1} B_u \ . \tag{4.4}$$

Note that $\frac{B_d}{B_u}$ is just the system asymmetry factor denoted by Γ_s which has been defined in (3.1). The asymmetry factor of the NRT calls is defined as

$$\Gamma_{NRT} = \frac{B^d_{NRT}}{B^u_{NRT}} \ . \tag{4.5}$$

Then (4.4) becomes

$$B^u_{NRT} = \frac{\Gamma_s - 1}{\Gamma_{NRT} - 1} B_u \ . \tag{4.6}$$

Since B_u can be determined by the total system bandwidth and the system asymmetry factor Γ_s, we can find from above equation that the bandwidth which can be used by the NRT calls is totally determined by Γ_s and Γ_{NRT} when the system utilization is maximized. Since the reassignment of bandwidth on uplink and downlink cannot be executed frequently [9], Γ_s can be regarded as statistically fixed.

We divide the NRT calls into different classes based on bandwidth requirements of applications. We assume that NRT calls are classified to N sub-classes. Let n_i denote the number of class i NRT calls in the system. (4.5) can be rewritten as

$$\Gamma_{NRT} = \frac{\sum_{i=0}^{N-1} a_i b^d_i}{\sum_{i=0}^{N-1} a_i b^u_i} \ , \tag{4.7}$$

where a_i is defined as

$$a_i = \frac{n_i}{\sum_{i=0}^{N-1} n_i} \ , \tag{4.8}$$

which is the ratio of class i NRT calls over all NRT calls. In a large time scale, a_i can be regarded as statistically fixed and thus Γ_{NRT} too. As Γ_s and Γ_{NRT} are statistically fixed, there exists unique B^u_{NRT} to achieve the maximum system utilization. We name this value B_{NRT}. If we can guarantee that the uplink bandwidth used by NRT calls statistically equals to B_{NRT}, the system utilization can be maximized. However, it is difficult to ensure that since the traffic in the system may keep changing in a relatively small time scale. Thus two problems which may result in low bandwidth utilization arise. We use a simple example to illustrate these problems (Figure 4.1). Let downlink bandwidth be 1.5 times of uplink bandwidth ($\Gamma_s = 1.5$). Assume there are two different call classes, RT calls and NRT calls, in the system. An RT call needs the same amount of bandwidth on both uplink and downlink while an NRT calls requires more downlink bandwidth ($\Gamma_{NRT} = 5$). Assume that the system is at a saturated situation and from (4.6) we know that the ratio of the uplink bandwidth used by the NRT calls over the total uplink bandwidth should be 12.5%. If the ratio is greater or smaller than 12.5%, a certain amount of bandwidth (uplink or downlink) could be wasted. In an extreme case, when the ratio of the uplink bandwidth used by the NRT calls over total uplink bandwidth is smaller than 5% (Figure 4.1 Case 1), too many RT calls are accepted and the RT calls will overuse the uplink bandwidth. As a result, only a small amount of uplink bandwidth can be used by the NRT calls. Since downlink has higher capacity than uplink and the RT calls require the same amount of bandwidth on both uplink and downlink, the remaining uplink bandwidth is too little to support sufficient NRT calls to use all the remaining downlink bandwidth. As a result, more than 20% downlink bandwidth will be wasted. On the other hand, when the ratio is greater than 20% (Figure 4.1 Case 2), too many NRT calls are accepted. They will use up the downlink bandwidth since the NRT calls require more downlink bandwidth than uplink bandwidth. In this case, the rejection of arriving calls is due to insufficient downlink bandwidth although there is unused uplink bandwidth. As a result, more than 30% uplink bandwidth will be wasted. In these two cases, both the RT calls and the NRT calls cannot be accepted any more, although there is unused bandwidth on downlink or uplink in the system.

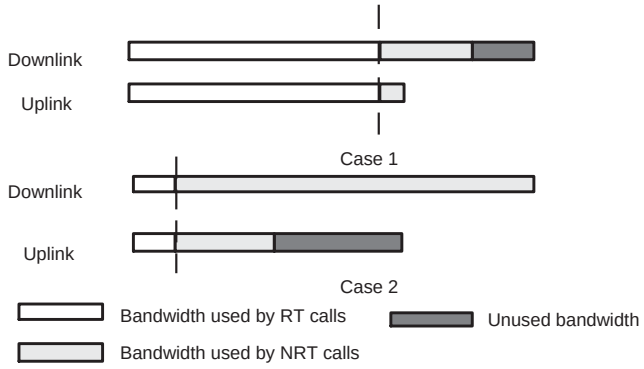

Figure 4.1: Illustration of the problems in multi-service mobile wireless networks with bandwidth asymmetry.

4.2.2 Proposed CAC Schemes

The mismatch of bandwidth allocation and asymmetric traffic load in the multi-service mobile cellular networks may result in a low bandwidth utilization. In order to improve the bandwidth utilization, the key of the proposed CAC schemes is to determine how much bandwidth can be used by RT calls and NRT calls while taking into account the handoff calls. This can be achieved by setting the specific bandwidth regions for the RT calls and the NRT calls. In the proposed CAC schemes, we divide total uplink channels into three regions. The first region is composed by a certain number of channels which are reserved as *guard channels* for handoff RT calls because of their highest priority. The second region is made up by the channels reserved for NRT calls and we name these reserved channels *NRT channels*. In our scheme, we set the size of the NRT channels equal to B_{NRT}, which can be obtained from (4.6). Besides the *guard channels* and the *NRT channels*, the remaining uplink channels compose the third region and we name these channels *common channels*, which are not reserved for any call classes. Thus there are three different classes of channels in the system: *guard channels*, *NRT channels* and *common channels*.

In this chapter, we propose two CAC schemes to address the issue of bandwidth asymmetry between uplink and downlink in multi-service mobile networks. Scheme 1 is a conservative scheme (Figure 4.2). The maximum bandwidth size that can be used by NRT calls on uplink is equal to B_{NRT}, which implies that when the NRT channels are used up, both the new and the handoff NRT calls will be blocked. When a call arrives, the system checks the downlink channels first. If there are no sufficient downlink channels, the call is blocked. Otherwise, the system examines the call

class. For a handoff RT call, the system checks the common channels. If the remaining common channels are sufficient, the call is accepted. Otherwise, the system checks the handoff channels. If the sum of remaining common channels and remaining handoff channels can satisfy the call's bandwidth requirement, the call can also be accepted. If the above conditions cannot be satisfied, the call is blocked. If the arrival call is a new RT call, the system checks the common channels only. The new RT call cannot be accepted if there are no sufficient free common channels in the system. For an NRT call (handoff or new), it is accepted if there are sufficient free NRT channels. Otherwise, the call is blocked. The pseudo code of the proposed Scheme 1 is shown in Figure 4.3.

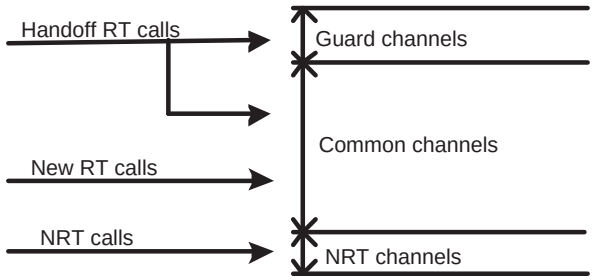

Figure 4.2: Illustration of the proposed Scheme 1.

```
If (RT call) then
 If (handoff call) then
  If (sufficient free guard channels or common channels)
      and (sufficient free downlink channels) then
      Accept the call
  Else
      Reject the call
 Else
  If (sufficient free common channels)
      and (sufficient free downlink channels) then
      Accept the call
  Else
      Reject the call

If (NRT call) then
 If (sufficient free NRT channels)
     and (sufficient free downlink channels) then
     Accept the call
 Else
     Reject the call
```

Figure 4.3: The pseudo code of Scheme 1.

Since handoff NRT calls have higher priority than new NRT calls and blocking handoff calls may waste the system resources unnecessarily, we propose Scheme 2 (Figure 4.4) for considering to decrease the dropping probability of handoff NRT calls. In this scheme, handoff NRT calls can

use the common channels and the NRT channels while new NRT calls are limited to use the NRT channels only. Without loss of generality, we assume that there are enough downlink channels. When an NRT call arrives, if there are sufficient NRT channels, the call (handoff or new) can be accepted. Otherwise, the new NRT call is blocked. For handoff NRT call, the system checks both the NRT channels and the common channels. If the sum of the remaining NRT channels and remaining common channels can satisfy the call's bandwidth requirement, the call is accepted. Otherwise, the call is blocked. The treatment to RT call in Scheme 2 is identical to that in Scheme 1. The pseudo code of the proposed Scheme 2 is shown in Figure 4.5.

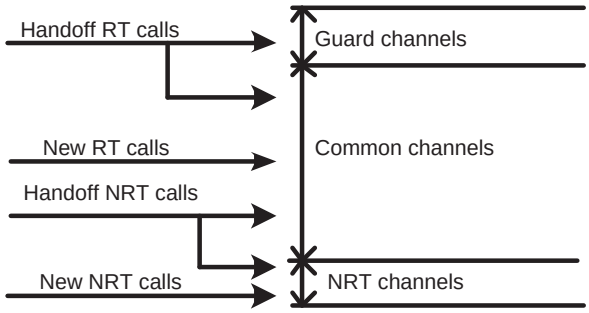

Figure 4.4: Illustration of the proposed Scheme 2.

4.3 Performance Analysis

In this section, we use Markov model to analyze the performance of the proposed CAC schemes in terms of call dropping and blocking probability and bandwidth utilization. In the analysis of this section, we consider a general model where there are multiple classes of RT and NRT calls with different bandwidth requirements. Assume that there are M sub-classes of RT calls and N sub-classes of NRT calls. Different RT call classes are labeled from 1 to M while NRT call classes are labeled from $M+1$ to $M+N$. We assume that call arrival process follows the Poisson distribution. Let λ_i and h_i denote the mean arrival rate of new call and handoff call of class $i(1 \leq i \leq (M+N))$, respectively. The service time of call class i is assumed to be exponentially distributed with mean $1/\mu_i$. In addition, we assume that the dwell time of call class i follows exponential distribution with mean $1/\nu_i$. Then the connection holding time of call class i is exponentially distributed with mean $1/(\mu_i + \nu_i)$.

```
If (RT call) then
 If (handoff call) then
  If(sufficient free guard channels or common channels)
      and (sufficient free downlink channels) then
     Accept the call
  Else
     Reject the call
 Else
  If (sufficient free common channels)
      and (sufficient free downlink channels) then
      Accept the call
  Else
      Reject the call

If (NRT call) then
 If (handoff call) then
  If (sufficient free NRT channels or free common channels)
      and (sufficient free downlink channels) then
      Accept the call
  Else
      Reject the call
  Else
   If(sufficient free NRT channels)
      and (sufficient free downlink channels) then
      Accept the call
   Else
      Reject the call
```

Figure 4.5: The pseudo code of Scheme 2.

In the analysis model, the system state is defined by a row vector π as

$$\pi = (n_1, n_2, \cdots, n_i, \cdots, n_{(M+N)}) \,, \tag{4.9}$$

where $n_i (1 \leq i \leq (M+N))$ denotes the number of the class i calls in process. Let b_i^u and b_i^d be the uplink and downlink bandwidth requirements of call class i, respectively. Let B_u and B_d denote the total amount of uplink bandwidth and downlink bandwidth, respectively. So the feasible state space, ψ, is

$$\psi = \left\{ \pi : (\sum_{i=1}^{M+N} n_i b_i^u \leq B_u) \; and \; (\sum_{i=1}^{M+N} n_i b_i^d \leq B_d) \right\} \,. \tag{4.10}$$

We use B_{GC}, B_{CC}, B_{NRT} to denote the capacity of the guard channels, the common channels and the NRT channels, respectively. Let $R_{h,i}^u(\pi)$ and $R_{n,i}^u(\pi)$ denote the remaining uplink bandwidth that could be used by the handoff calls and new calls of call class i respectively when the system state is π. Let $R_i^d(\pi)$ be the remaining downlink bandwidth that can be used by the calls of class i when the system state is π.

The computations of $R_{h,i}^u(\pi)$, $R_{n,i}^u(\pi)$ and $R_i^d(\pi)$ of Scheme 1 and Scheme 2 are shown in Table

4.1 and Table 4.2, respectively.

Table 4.1: $R^u_{h,i}(\pi)$, $R^u_{n,i}(\pi)$ and $R^d_i(\pi)$ of Scheme 1

$R^u_{h,i}(\pi)$	$B_u - B_{NRT} - \sum_{i=1}^{M} n_i b_i^u \quad (RT\ call)$
	$B_{NRT} - \sum_{i=M+1}^{M+N} n_i b_i^u \quad (NRT\ call)$
$R^u_{n,i}(\pi)$	$B_{CC} - \sum_{i=1}^{M} n_i b_i^u \quad (RT\ call)$
	$B_{NRT} - \sum_{i=M+1}^{M+N} n_i b_i^u \quad (NRT\ call)$
$R^d_i(\pi)$	$B_d - B_{NRT} - \sum_{i=1}^{M+N} n_i b_i^d$

Table 4.2: $R^u_{h,i}(\pi)$, $R^u_{n,i}(\pi)$ and $R^d_i(\pi)$ of Scheme 2

$R^u_{h,i}(\pi) \quad (RT\ call)$	$B_u - \sum_{i=1}^{M+N} n_i b_i^u \quad (\sum_{i=M+1}^{M+N} n_i b_i^u > B_{NRT})$
	$B_u - \sum_{i=1}^{M} n_i b_i^u - B_{NRT} \quad (\sum_{i=M+1}^{M+N} n_i b_i^u \le B_{NRT})$
$R^u_{h,i}(\pi) \quad (NRT\ call)$	$B_{NRT} - \sum_{i=M+1}^{M+N} n_i b_i^u \quad (\sum_{i=1}^{M} n_i b_i^u > B_{CC})$
	$B_u - B_{GC} - \sum_{i=1}^{M+N)} n_i b_i^u \quad (\sum_{i=1}^{M} n_i b_i^u \le B_{CC})$
$R^u_{n,i}(\pi) \quad (RT\ call)$	$B_u - B_{GC} - \sum_{i=1}^{M+N} n_i b_i^u \quad (\sum_{i=M+1}^{M+N} n_i b_i^u > B_{NRT})$
	$B_{CC} - \sum_{i=1}^{M} n_i b_i^u \quad (\sum_{i=M+1}^{M+N} n_i b_i^u \le B_{NRT})$
$R^u_{n,i}(\pi) \quad (NRT\ call)$	$B_{NRT} - \sum_{i=M+1}^{M+N} n_i b_i^u$
$R^d_i(\pi)$	$B_d - \sum_{i=1}^{M+N} n_i b_i^d$

In such a system, any state transition is caused by one of the following events:

1. Arrival of a handoff RT call or a handoff NRT call,

2. Arrival of a new RT call or a new NRT call,

3. Termination of a call,

4. Handoff of a call.

Then we can define two neighboring states of π, π_{i+} and π_{i-}, as

$$\pi_{i+} = (n_1, n_2, \cdots, n_i + 1, \cdots, n_{(M+N)}) \quad i \in [1, M+N] \tag{4.11}$$

$$\pi_{i-} = (n_1, n_2, \cdots, n_i - 1, \cdots, n_{(M+N)}) \quad i \in [1, M+N] \, . \tag{4.12}$$

Two events, the arrival of a class i handoff call and the arrival of a class i new call, will cause the system to transit from state π to π_{i+}. We use $q_i^h(\pi)$ and $q_i^n(\pi)$ to denote the transition rates when the state transition is triggered by the arrival of a class i handoff call and new call, respectively. $q_i^h(\pi)$ and $q_i^n(\pi)$ are expressed as

$$q_i^h(\pi) = I_{R_{h,i}^u(\pi) \geq b_i^u} \cdot I_{R_i^d(\pi) \geq b_i^d} \cdot h_i \tag{4.13}$$

and

$$q_i^n(\pi) = I_{R_{n,i}^u(\pi) \geq b_i^u} \cdot I_{R_i^d(\pi) \geq b_i^d} \cdot \lambda_i \, , \tag{4.14}$$

where I_c is a binary variable, which is equal to one if condition c is true or zero otherwise. Let us consider the system state transition from π to π_{i-}. This transition can be caused by two events: termination or handoff of a class i call. We use $p_i^t(\pi)$ and $p_i^h(\pi)$ to denote the state transition rates triggered by these two events. Then

$$p_i^t(\pi) = n_i \mu_i \tag{4.15}$$

$$p_i^h(\pi) = n_i \nu_i \, . \tag{4.16}$$

Let P_π denote the stationary probability of the state π. Then P_π should satisfy the following flow balance equation:

$$\begin{aligned} & \mathrm{P}_\pi \sum_{i=1}^{M+N} \left[q_i^h(\pi) + q_i^n(\pi) + p_i^t(\pi) + p_i^h(\pi) \right] \\ & = \sum_{i=1}^{M+N} I_{\pi_{i+} \in \psi} \mathrm{P}_{\pi_{i+}} \left[p_i^t(\pi_{i+}) + p_i^h(\pi_{i+}) \right] \\ & + \sum_{i=1}^{M+N} I_{n_i \geq 1} \mathrm{P}_{\pi_{i-}} \left[q_i^h(\pi_{i-}) + q_i^n(\pi_{i-}) \right] \quad \pi \in \psi \end{aligned} \, . \tag{4.17}$$

Note P_π should also satisfy the normalization equation:

$$\sum_{\pi \in \psi} \mathrm{P}_\pi = 1 \ . \tag{4.18}$$

Using the flow balance equation (4.17) and the normalization equation (4.18), we can obtain the stationary probability P_π when the system state is π ($\pi \in \psi$).

So far, we have obtained the flow balance equation and thus the stationary probability P_π, from which we can calculate the measures that we concern about in our schemes, which include the call dropping and blocking probability and the bandwidth utilization. Let P_h^i and P_n^i denote the call dropping probability of class i handoff calls and the call blocking probability of class i new calls, respectively. Let ξ_i denote the subset of the feasible state-space ψ when the class i handoff call cannot be accepted. Then

$$\xi_i = \left\{ \pi : (R_{h,i}^u(\pi) < b_i^u) \ or \ (R_i^d(\pi) < b_i^d) \quad \pi \in \psi \right\} \ . \tag{4.19}$$

The dropping probability of class i handoff call, P_h^i, is given by

$$P_h^i = \sum_{\pi \in \xi_i} \mathrm{P}_\pi \ . \tag{4.20}$$

Let η_i be the subset of the state-space ψ when the class i new call cannot be accepted. Then

$$\eta_i = \left\{ \pi : (R_{n,i}^u(\pi) < b_i^u) \ or \ (R_i^d(\pi) < b_i^d) \quad \pi \in \psi \right\} \ . \tag{4.21}$$

The new call blocking probability of call class i , P_n^i, is

$$P_n^i = \sum_{\pi \in \eta_i} \mathrm{P}_\pi \ . \tag{4.22}$$

Another important measure is bandwidth utilization, which is the ratio of used bandwidth over total system bandwidth. Let U_{up} and U_{down} denote the uplink and downlink bandwidth utilization,

respectively. Then U_{up} and U_{down} can be expressed as

$$U_{up} = \frac{\sum\limits_{\pi \in \psi} \mathrm{P}_\pi \sum\limits_{i=1}^{M+N} n_i b_i^u}{B_u} \tag{4.23}$$

and

$$U_{down} = \frac{\sum\limits_{\pi \in \psi} \mathrm{P}_\pi \sum\limits_{i=1}^{M+N} n_i b_i^d}{B_d} \ , \tag{4.24}$$

respectively. The total bandwidth utilization, U, is

$$U = \frac{\sum\limits_{\pi \in \psi} \mathrm{P}_\pi \sum\limits_{i=1}^{M+N} n_i b_i^u + \sum\limits_{\pi \in \psi} \mathrm{P}_\pi \sum\limits_{i=1}^{M+N} n_i b_i^d}{B_u + B_d} \ . \tag{4.25}$$

We use an experiment to verify the above analysis model. We assume that there are two types of calls, RT call and NRT call, and 80% of new calls are RT calls. There are total 100 channels in the system and 60 channels are allocated to downlink. An RT call requires one channel on both uplink and downlink while an NRT call requires 1 uplink channel and 5 downlink channels. Call arrival follows the Poison process and call serving time follows the exponential distribution. The mean serving time of RT calls and NRT calls are 120 seconds and 900 seconds, respectively. We use the analysis model to evaluate the performance of Scheme 1. The comparisons of handoff RT call dropping probability, new RT call blocking probability and NRT call blocking probability obtained from the analysis model and the simulation results are shown in Figure 4.6. From the results, we know that the handoff call dropping probability of RT calls is the lowest while the call blocking probability of NRT calls is the highest. The NRT call blocking probability increases rapidly with traffic load. In this extreme case, there are only 5 uplink channels can be used by NRT calls. When the traffic load is heavy, it is obvious that most of NRT calls will be blocked. The obtained results are as expected. Figure 4.7 shows bandwidth utilization of uplink and downlink. Both the uplink and the dowlink bandwidth utilization increase rapidly with the traffic load. We also find that the downlink bandwidth utilization is slightly higher than the uplink bandwidth utilization. As downlink bandwidth is 1.5 times of the uplink bandwidth while downlink bandwidth required by an NRT call is 5 times of uplink bandwidth, we can expect that the downlink bandwidth utilization

should be higher than the uplink bandwidth utilization. The numerical results also demonstrate that the simulation results match the results obtained from the above analytical model well.

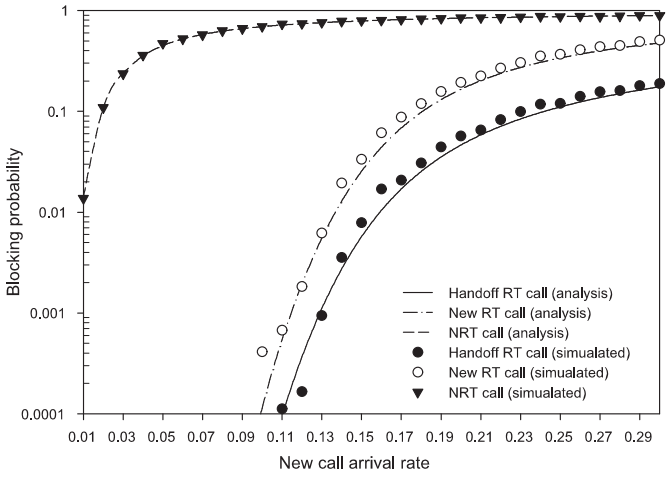

Figure 4.6: Comparisons of call blocking probabilities of analysis and simulation results.

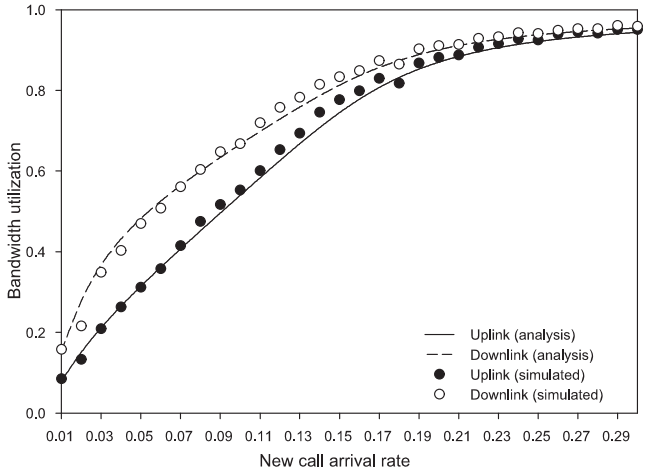

Figure 4.7: Comparisons of uplink and downlink bandwidth utilization of analysis and simulation results.

In the following section, we will use simulation experiments to compare the performance of the proposed CAC schemes with that of some existing CAC schemes.

4.4 Performance Evaluation

4.4.1 Traffic Model

In our simulation experiments, we use the traffic model based on the IMT-2000 system [90]. Table 4.3 lists the traffic parameters used in the simulation experiments. These parameters are also used in [8].

Table 4.3: Traffic Model

	RT call		NRT call	
	Uplink	Downlink	Uplink	Downlink
Information Rate,I	16kbps	16kbps	64kbps	384kbps
Activity Factor, α	0.5	0.5	0.00285	0.015
Effective Bandwidth, αI	8kbps	8kbps	182.4bps	5.76kbps
Mean Call Duration	120sec		3000sec	
Mean Cell Dwell Time	300sec		1200sec	
Service Example	voice		web access	

We assume that the downlink bandwidth is $2.7Mbps$ while the uplink bandwidth is $1.3Mbps$, which are also used in [8]. There are two types of calls, RT calls and NRT calls, in the system. The RT calls require symmetric bandwidth on uplink and downlink while the NRT calls require more downlink bandwidth than uplink bandwidth as shown in Table 4.3.

According to the derivation in Section 4.2, B_{NRT} used in our scheme can be found equal to $44kbps$. The arrival of the new calls and the handoff calls follows the Poisson process. Let q be the ratio of the new RT calls over all new calls. Then $(1-q)$ of the arrival new calls are the NRT calls. We also assume that 40% of the RT calls in the system are the handoff RT calls while 10% of the NRT calls are the handoff NRT calls. Note that in the simulation the call admission decision is made according to the following rules: (1) for an RT call, it can be accepted only if its information rate can be satisfied since it has more stringent QoS requirement than the NRT calls; (2) for an NRT call, it can be accepted if its effective bandwidth can be matched. The effective bandwidth means the minimum required bandwidth to provide a specific QoS given the traffic parameters of a call connection. it is the product of information rate and activity factor [8].

First, we compare the performance of Scheme 1 with that of Scheme 2. Next we chose the one which has better performances in the simulation, Scheme 2, to compare with the scheme which does not set threshold for the NRT calls such as such as DTBR scheme [21]. Then, we compare

the performance of Scheme 2 with that of Jeon's scheme [8] which also implements the bandwidth asymmetry. Last, we will show the performance of the proposed scheme when the asymmetry factor of the NRT calls (Γ_{NRT}) changes.

4.4.2 Comparison of Scheme 1 and Scheme 2

We compare the performance of Scheme 1 and Scheme 2 when q is equal to 80%. In this experiment scenario, most traffic load is brought by RT calls. Figure 4.8 shows the blocking probability of NRT calls. From this figure, we can find that Scheme 2 can achieve much lower dropping probability of handoff NRT call as we expected since the handoff NRT calls are able to use the common channels. Although the new NRT call blocking probability of Scheme 2 is slightly higher than that of Scheme 1, it is reasonable to make such a tradeoff between the low priority calls and the high priority calls. The simulation results show that the differences of the bandwidth utilization and the RT call blocking probability obtained from Scheme 1 and Scheme 2 are invisible. Thus we do not show them here. In this scenario, we can find that Scheme 2 outperforms Scheme 1. However, Scheme 2 may not always achieve better performance than Scheme 1. For example, in some hot spot areas in a mobile network, the arrival rate of handoff NRT call may be very high. In such environment, Scheme 1 may outperform Scheme 2 since it avoids the handoff NRT call to overuse the downlink bandwidth. It is necessary to consider both Scheme 1 and Scheme 2 and use them in different scenarios.

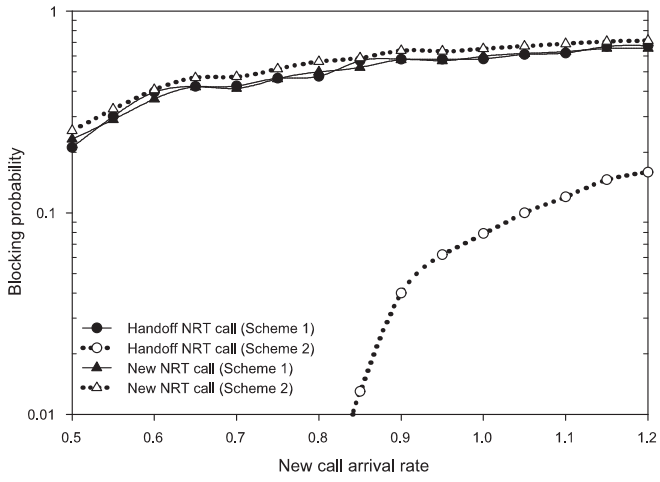

Figure 4.8: Comparisons of the NRT call blocking probabilities of Scheme 1 and Scheme 2

4.4.3 Comparison of Scheme 2 and DTBR Scheme

In this part, we compare the proposed Scheme 2 with the DTBR scheme [21]. How to determine the values of thresholds K_1 and K_2 is a difficult problem in the DTBR scheme. Since there is no detailed method to compute these values in the paper, we have done extensive experiments by setting different threshold values. Then we choose a set of better results to compare with our scheme.

We compare the uplink and downlink bandwidth utilization and the total bandwidth utilization of the proposed Scheme 2 and the DTBR scheme in Figure 4.9 and Figure 4.10 respectively when q is equal to 70%. From these figures we find that the uplink and downlink bandwidth utilization of the DTBR scheme changes dramatically. When traffic load is relatively low, the DTBR scheme admits too many NRT calls and the superfluous NRT calls use up the downlink bandwidth. As a result, certain amount of uplink bandwidth cannot be used and more RT calls are blocked, as shown in Figure 4.11. With the increase of the traffic load, the DTBR scheme may accept too many RT calls. The superfluous RT calls overuse the uplink bandwidth and thus certain amount of downlink bandwidth cannot be used and more NRT calls will be blocked, as shown in Figure 4.12. These results show that the proposed scheme 2 can achieve better bandwidth utilization on uplink and downlink when traffic load increases and call dropping probability of some high priority calls is also controlled at a reasonable low level. In addition, from this simulation experiment we realize that the fixed threshold values for the DTBR scheme cannot achieve satisfied performance in the asymmetric bandwidth allocation networks. However, how to dynamically adjust the values of two key parameters,K_1 and K_2, may be a complicated problem and it is not addressed in [21]. In our scheme, a certain number of channels are set as NRT channels, which can be computed from (4.6) readily. Without a dynamic adjusting strategy, our scheme can achieve good performance when traffic load changes.

4.4.4 Comparison of Scheme 2 and Jeon's Scheme

From the above results, we know that the bandwidth threshold for NRT calls is very desirable in multi-service mobile wireless networks. However, limiting only the number of NRT calls in the system cannot guarantee the high bandwidth utilization when traffic changes. In Jeon's scheme [8],

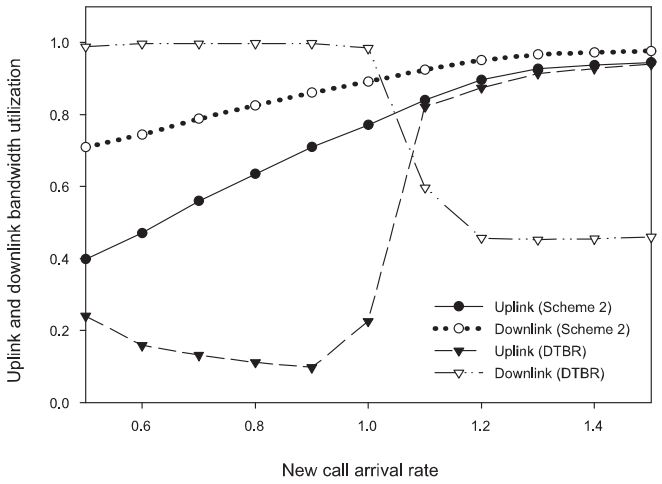

Figure 4.9: Comparisons of the uplink and downlink bandwidth utilization of Scheme 2 and the DTBR scheme when q=70%.

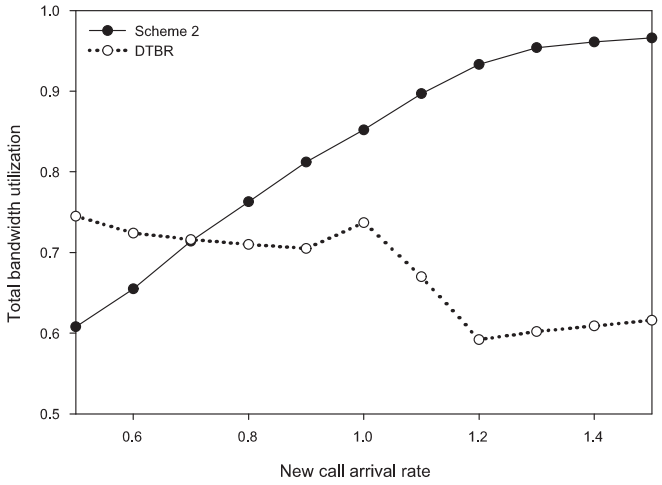

Figure 4.10: Comparison of the total bandwidth utilization of Scheme 2 and the DTBR scheme when q=70%.

the authors use multi-guard-channel to guarantee the QoS requirements of high priority calls. By setting different guard channel for different call class on uplink and downlink separately, the scheme prevents low priority calls from overusing the resources. The authors demonstrated the good performance in terms of bandwidth utilization and call blocking probability of Jeon's scheme when q is equal to 85%. However, when we increase q to 95%, we find that Jeon's scheme suffers.

Figure 4.13 shows the uplink and downlink bandwidth utilization when q is equal to 95%. In this scenario, most traffic are brought by RT calls. From the figure, we can find that with the increase of call arrival rate (i.e., the traffic load becomes heavier) the uplink bandwidth utilization

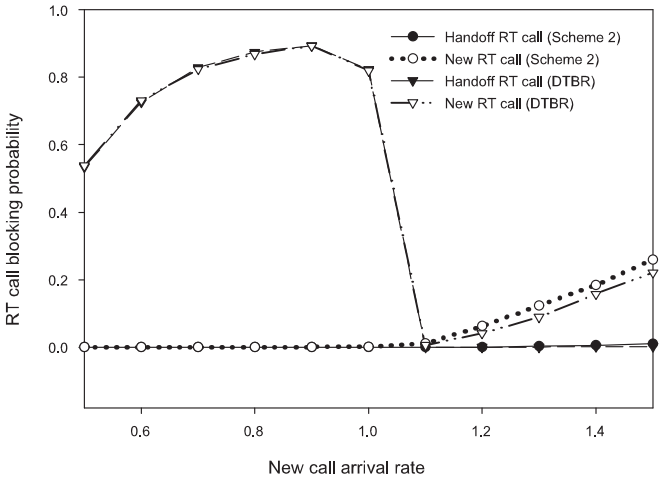

Figure 4.11: Comparisons of the RT call blocking probabilities of Scheme 2 and the DTBR scheme when q=70%.

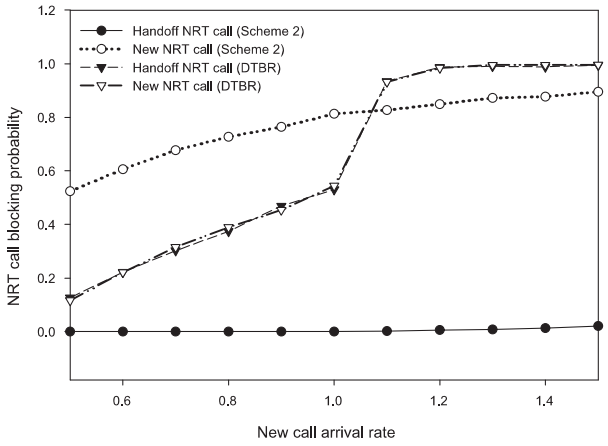

Figure 4.12: Comparisons of the NRT call blocking probabilities of Scheme 2 with the DTBR scheme when q=70%.

of both schemes increase fast and the increasing speeds are very close. However, the proposed scheme can achieve significantly higher downlink bandwidth utilization than Jeon's scheme. In this scenario, most incoming calls are RT calls. If no bandwidth is reserved for the NRT calls, too many RT calls will be accepted and thus the RT calls will consume almost all the uplink bandwidth. As a result, the downlink bandwidth is not sufficient to accept enough NRT calls to use the remaining bandwidth and a certain amount of downlink bandwidth is wasted.

Figure 4.14 and Figure 4.15 show the call blocking probability of RT calls and NRT calls, respectively. From these two figures, we can find that both the proposed scheme and Jeon's scheme can guarantee the handoff RT call blocking probability under a certain threshold (1%). Although

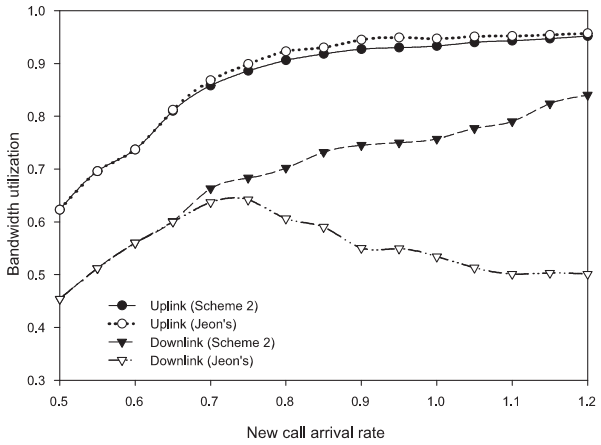

Figure 4.13: Comparisons of the uplink and downlink bandwidth utilization of Scheme 2 and Jeon's scheme when q=95%.

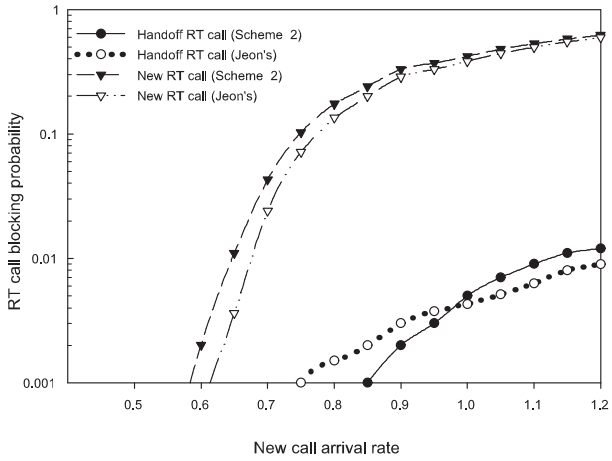

Figure 4.14: Comparisons of the RT call blocking probabilities of Scheme 2 and Jeon's scheme when q=95%.

the new RT call blocking probability of the proposed scheme is slightly higher than that of Jeon's scheme, both the dropping and blocking probability of NRT call of the proposed scheme are significantly lower than that of Jeon's scheme. In this case, Jeon's scheme accepts too many RT calls and almost all the uplink bandwidth is used by the RT calls. Thus more NRT calls are blocked. From these results, we know that in order to improve system performance it is necessary to reserve certain bandwidth bandwidth resources for NRT calls and thus balance the admission of RT calls and NRT calls in multi-service mobile networks with bandwidth asymmetry.

Next we examine the performance of the proposed scheme and Jeon's scheme under a set of scenarios with different q values (q varies from 75% to 95% with interval 5%). As it is more mean-

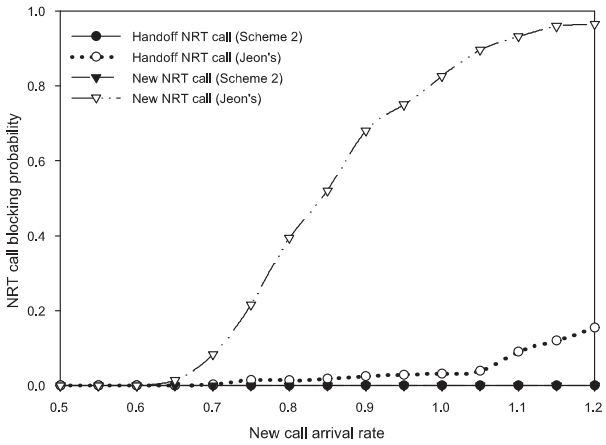

Figure 4.15: Comparisons of the NRT call blocking probabilities of Scheme 2 and Jeon's scheme when q=95%.

ingful to judge the bandwidth utilization when the system under the heavy traffic load condition, we fix the new call arrival rate at 1.0. Figure 4.16 and Figure 4.17 show the bandwidth utilization of Scheme 2 and Jeon's scheme with different q values. We can find that the proposed scheme can obtain stable bandwidth utilization (close to 100%) when q changes while the bandwidth utilization of Jeon's scheme changes dramatically. Regarding to the blocking probability, Figure 4.18 shows the call blocking probability when q changes. From this figure, we can find that the call dropping probability of both RT calls and NRT calls of the proposed scheme can be controlled at a reasonable low level. When the ratio of RT calls over all calls is low ($q < 85\%$), the bandwidth threshold limits the bandwidth which can be used by the new NRT calls. As the traffic is heavy in this scenario and the system is close to saturation, in order to obtain high bandwidth utilization and guarantee low blocking probability of high priority calls, blocking superfluous new NRT calls is reasonable. When q increases, the call blocking probability of new RT calls increases accordingly. By properly rejecting a certain number of new RT calls, the proposed scheme can guarantee the call dropping probability of handoff NRT calls at a reasonable low level and improves system bandwidth utilization.

In the above simulation experiments, we assume only one class of NRT calls in the system and the minimum bandwidth required by an NRT call is fixed. Thus the asymmetry factor of the NRT calls, Γ_{NRT}, is also fixed. Indeed, there may be more than one class of NRT calls in the system and different call classes may have different bandwidth requirements. As a result, Γ_{NRT} may change

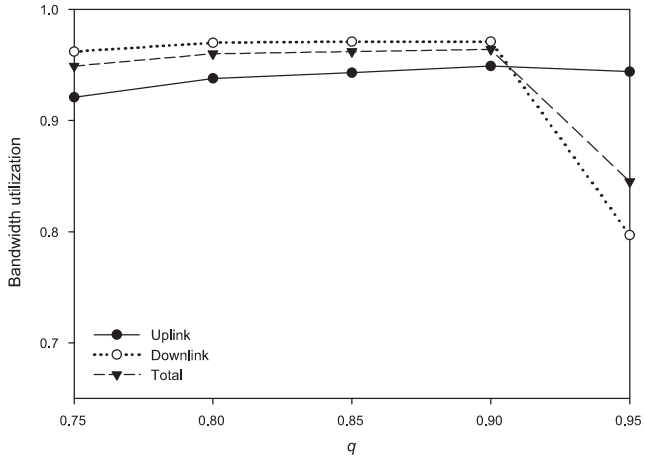

Figure 4.16: Bandwidth utilization of Scheme 2 under different values of q.

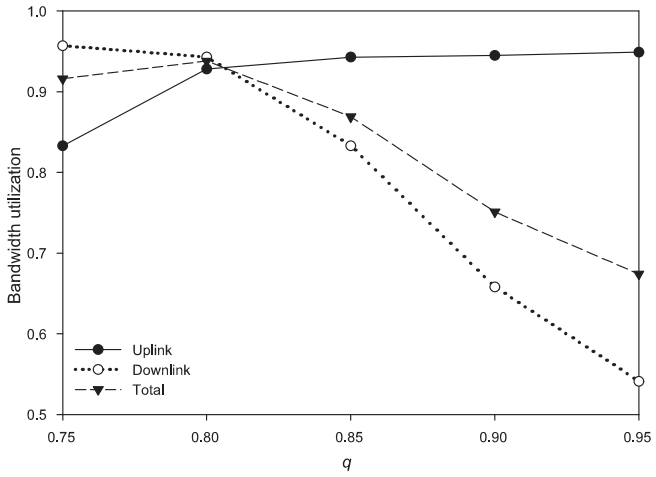

Figure 4.17: Bandwidth utilization of Jeon's scheme under different values of q.

with the arrival rates of NRT calls belonging to different call classes. Here we use simulation to evaluate the performance of the proposed scheme when Γ_{NRT} changes. In the experiment, the parameters are identical to those used in the above simulation experiments except that we assume two classes of NRT calls (class 1 and class 2) and they have same activity factors with different bandwidth requirements. We also assume that the arrivals of these two classes of NRT calls follow the Poisson distribution with rates λ_1 and λ_2, respectively. From (4.7), we know that we should obtain a_i for calculating Γ_{NRT}. In a statistical point of view, a_i can be rewritten as $\frac{\lambda_i}{\sum_{i=0}^{N-1} \lambda_i}$, where λ_i is the mean arrival rate of the class i NRT calls. In this experiment, a_1 and a_2 are $\frac{\lambda_1}{\lambda_1+\lambda_2}$ and $\frac{\lambda_2}{\lambda_1+\lambda_2}$ respectively and thus Γ_{NRT} is $\frac{a_1 \cdot B_1^d + a_2 \cdot B_2^d}{a_1 \cdot B_1^u + a_2 \cdot B_2^u}$. The value of Γ_{NRT} can be determined if the call

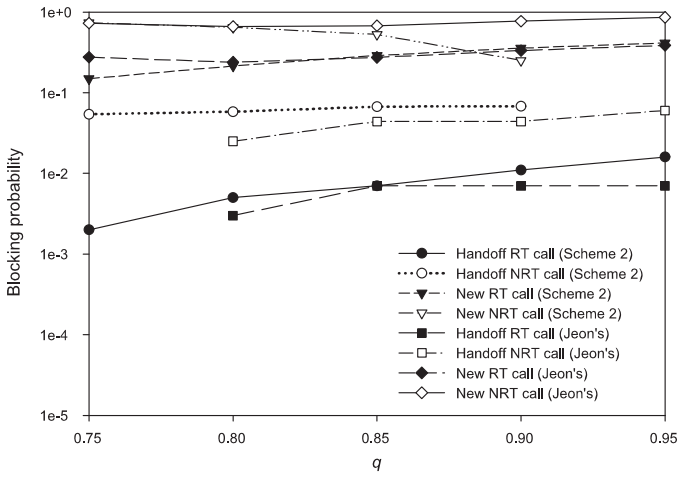

Figure 4.18: Comparisons of blocking probabilities of Scheme 2 and Jeon's scheme under different values of q.

arrival rate can be estimated. Since how to scale the call arrival rate is beyond the scope of our research, we will not discuss it further. We just assume that the average call arrival rate can be scaled. The parameters of the class 1 and the class 2 NRT calls are listed in Table 4.4. In the simulation, we set $q = 70\%$, which means that 70% of the arrival calls are the RT calls.

Table 4.4: Traffic model of the NRT calls

	NRT call			
	Class 1		Class 2	
	Uplink	Downlink	Uplink	Downlink
Information Rate,I	64kbps	384kbps	103kbps	546kbps
Activity Factor, α	0.00285	0.015	0.00285	0.015
Effective Bandwidth, αI	182.4bps	5.76kbps	293bps	8.19kbps
Mean Call Duration	3000sec		900sec	
Mean Cell Dwell Time	1200sec		600sec	

Figure 4.19 (a) and (b) show the uplink and downlink bandwidth utilization for different Γ_{NRT}. From these figures we find that the proposed scheme can achieve satisfactory bandwidth utilization on both uplink and downlink when traffic load increases. It avoids the possible problems of low bandwidth utilization even when Γ_{NRT} has different values. Figure 4.20 (a) to (f) show the call blocking probabilities of different call classes (handoff RT calls, handoff class 1 NRT calls, handoff class 2 NRT calls, new RT calls, new class 1 NRT calls and new class 2 NRT calls) with different Γ_{NRT} values. These figures illustrate that the dropping probabilities of handoff calls are controlled at a low level and the blocking probability of new RT calls is also retained at a reasonable low level

under different Γ_{NRT} values.

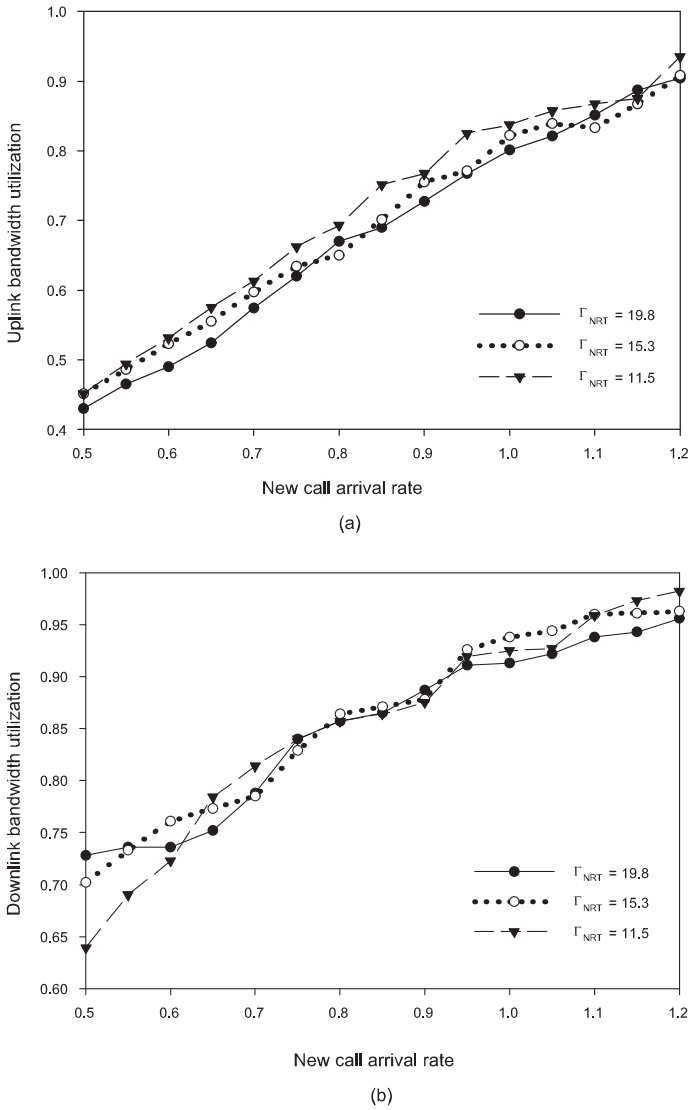

Figure 4.19: Uplink and downlink bandwidth utilization with different Γ_{NRT} values.

4.5 Summary

In this chapter, we have identified and analyzed the problems that may result in a low bandwidth utilization in bandwidth asymmetry mobile cellular networks and presented two schemes to address such problems. By setting the admissible bandwidth regions for RT calls and NRT calls, the proposed schemes determine the bandwidth that can be used by RT calls and NRT calls and thus prevent the calls of specific classes from overusing the bandwidth resources. The problems caused

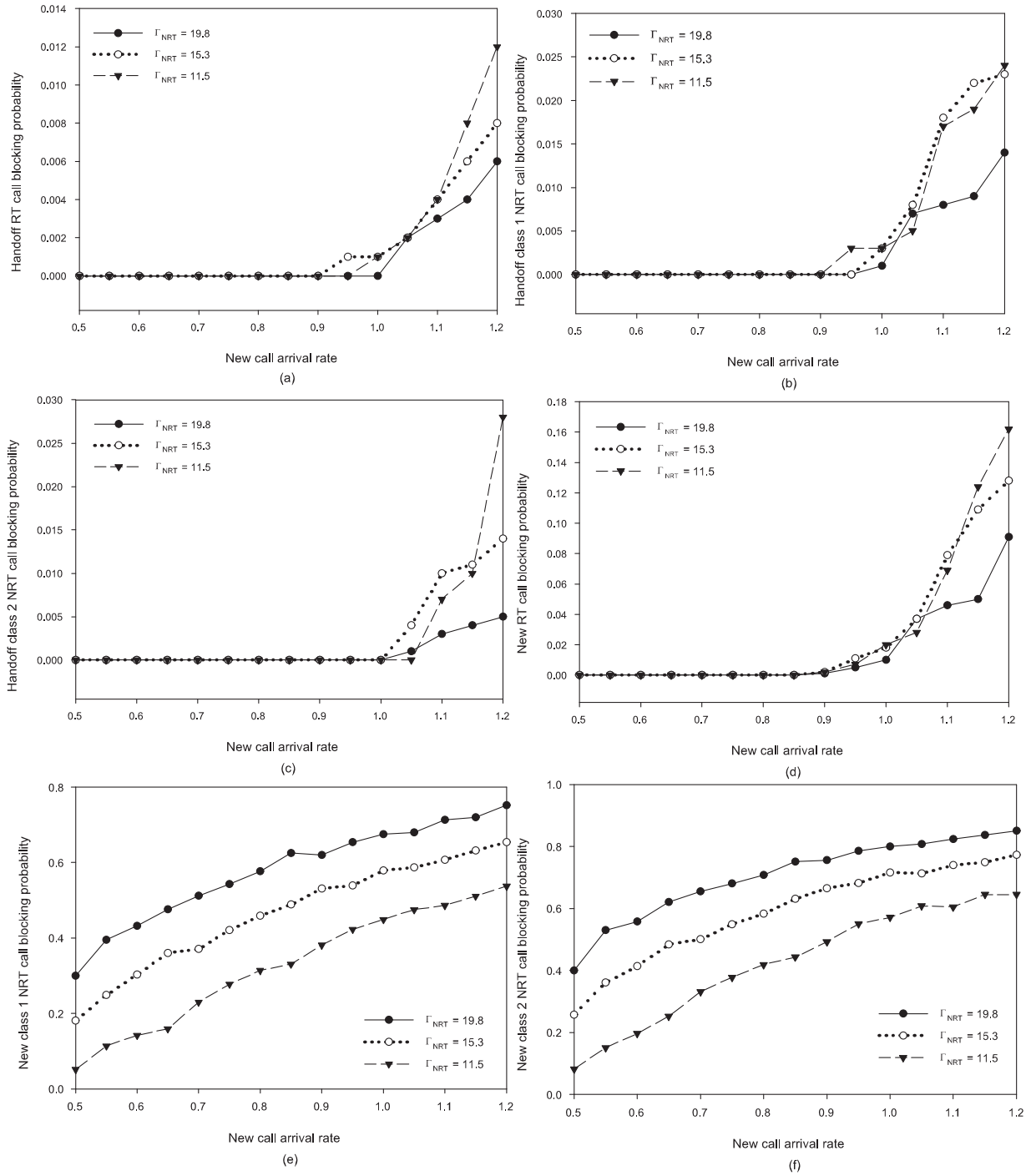

Figure 4.20: Call blocking probabilities with different Γ_{NRT} values.

by the mismatch of the bandwidth allocation and traffic changing are solved gracefully and system bandwidth utilization is also improved. In the proposed Scheme 1, the admissible bandwidth region for NRT calls is also the threshold for both handoff and new NRT calls. In Scheme 2, we set the

bandwidth threshold for new NRT calls only since handoff NRT calls have higher priority than new NRT calls. The simulation results demonstrate that both the proposed schemes can avoid the low bandwidth utilization problems in the bandwidth asymmetry networks while Scheme 2 can guarantee the dropping probability of handoff NRT calls at a low level without deteriorating the blocking probability of RT calls. Compared with some existing CAC schemes such as the DTBR scheme and Jeon's scheme, scheme 2 can achieve a higher bandwidth utilization when traffic changes in bandwidth asymmetry networks. At the same time, it guarantees the dropping probability of some high priority calls (handoff RT calls and handoff NRT calls) at a reasonable low level. A feature of our schemes is that the size of the bandwidth regions for RT calls and NRT calls is time-invariant. Such fixed size may not be optimal under some dynamic traffic conditions. It would be an interesting topic to design an algorithm to adjust the bandwidth regions according to actual traffic patterns and thus improve system performances under different traffic load environments.

Chapter 5

Minimizing Average Cost in Bandwidth Asymmetry Mobile Networks

Next generation mobile networks need to support multi-class services with asymmetric bandwidth allocation between uplink and downlink to match asymmetric traffic load brought by some data applications. For the design of call admission control policy in such networks, how to decrease average system cost is one of the key issues. In this chapter, we study the optimal admission policy for minimizing system cost. By modeling the admission control problem as a Markov decision process (MDP) and analyzing the corresponding value function, we obtain some monotonicity properties of the optimal policy. These properties suggest that the optimal admission control policy for the bandwidth asymmetry mobile networks have a threshold structure and the threshold specified for a call class may change with system states. Because of the prohibitively high complexity for computing the thresholds in a system with large state-space, we propose a heuristic CAC policy called Call-Rate-based Dynamic Threshold (CRDT) policy to approximate the theoretical optimal policy based on the insights we obtain from the modeling and the analytical study on the properties of the optimal policy. The CRDT policy is efficient and can be easily implemented. Numerical results show that the performance in terms of average system cost of the proposed CRDT policy is close to that of the optimal policy from the MDP model and is better than that of some known existing CAC schemes, including those performing well in bandwidth asymmetry mobile networks.

5.1 Introduction

One of the most prominent features of next generation mobile cellular networks is to support multi-service applications, such as voice, video, web browsing, file transmission, interactive gaming, etc. Since some data applications, such as web browsing and file downloading may bring more traffic load on downlink than on uplink, next generation mobile wireless networks are expected to present distinctive traffic asymmetry between uplink and downlink [5, 7–9, 89, 91]. In such environment, it is necessary to allocate different bandwidth between uplink and downlink in order to support asymmetric traffic load. It has been proven that the asymmetric bandwidth allocation outperforms the symmetric bandwidth allocation in such environment [7]. How to guarantee the QoS of different call classes and improve the system performance in such asymmetric bandwidth allocation mobile wireless networks is an attractive research topic in recent years [5, 7–9, 89, 91].

In Chapter 4, we studied the mismatch problem between asymmetric bandwidth allocation and dynamic traffic load in a system. We find that if too many bandwidth-symmetric calls such as RT calls are accepted, some downlink bandwidth resources might be wasted. On the other hand, if too many bandwidth-asymmetric calls such as NRT calls are accepted, some uplink bandwidth might be wasted. We proposed two new call admission control schemes to address this problem. The proposed schemes improve the bandwidth utilization in an asymmetric bandwidth allocation mobile network and guarantee the QoS of some high priority calls such as handoff RT calls. We categorize the problem we addressed in Chapter 4 as the MAXU problem, which is defined as maximizing system bandwidth utilization subject to constraints on the blocking probabilities of some high priority calls. Indeed, the MAXU problem in the symmetric bandwidth allocation wireless networks has been intensively studied in the literatures recently [17, 21, 92–96].

In this chapter, we study call admission control in mobile cellular networks with bandwidth asymmetry from another perspective. We consider call admission as a decision process which decides whether or not to accept an arrival call subject to the MINCost problem, which is defined as minimizing a linear objective cost function to minimize system average cost. To the best of our knowledge, there is few work which focuses on modeling and analysis of the MINCost problem especially in the bandwidth asymmetry mobile networks. For traditional mono-service networks, the MINOBJ problem, which is similar to MINCost problem, is studied in [19] and the GC scheme [16]

is proven be optimal. In [20], the authors studied maximizing reward problem, which is similar to the MINCost problem except that it is concerned about reward maximization instead of cost minimization. The authors demonstrated the sub-modularity for the 2-classes problem and established some properties of optimal policies for a resources-sharing system. These properties could be extended to the asymmetric mobile cellular networks as we consider in this chapter. Moreover, the authors formulated CAC problem of a resource-sharing system into a fluid model and study the optimal admission control for a large-capacity system. They showed that the trunk reservation policy is optimal when the calls in the system have identical service time. When the call duration does not depend on the call class, the system model is reduced to a one-dimensional state-space model. Indeed, such a one-dimensional state-space model has been studied in [19] when the number of call classes is two and a similar conclusion has been drawn, which indicates the guard channel policy is optimal. However, it is unrealistic to require that different call classes have identical service time in a multi-service environment. It is necessary to find a feasible dynamic scheme based on the obtained properties to handle the MINCost problem in dynamic traffic load system, especially in asymmetric traffic load system.

In this chapter, we focus on modeling and analysis of admission control subject to the MINCost problem in mobile networks with bandwidth asymmetry. By formulating CAC problem into a Markov decision process (MDP) model and analyzing the corresponding value function, we extend the properties in [20] and identify some monotonicity properties of a value function for bandwidth asymmetry networks. These properties suggest that the optimal policy in such environment have a threshold structure and the thresholds of different call classes may vary with system states. Because of prohibitively high complexity of computing the thresholds in a large system state-space, we propose a heuristic policy called Call-Rate-based Dynamic Threshold (CRDT) policy based on our insights obtained in the modeling and the analysis. The numerical results show that the average cost obtained from the CRDT policy is very close to that obtained by applying the policy from the MDP model in a dynamic traffic load system.

Our contribution is threefold: 1) We formulate the admission control for the MINCost problem in asymmetric bandwidth allocation mobile networks into an MDP model; 2) We prove some monotonicity properties of the optimal admission policy in the bandwidth asymmetry mobile networks.

These properties may imply certain monotonicity properties of the optimal admission policy, e.g., a threshold structure; and 3) We propose a heuristic policy, which can be readily implemented, and use numerical example to demonstrate the good performance of the proposed policy.

The rest of this chapter is organized as follows. In Section 5.2, we present the MDP formulation in detail. In Section 5.3, we analyze the corresponding value function. We show that the optimal CAC policy for the MINCost problem should have a threshold structure in the asymmetric bandwidth allocation mobile networks. In Section 5.4, we present the proposed CRDT admission control policy. The numerical results are given in Section 5.5. In this section, we compare the average cost of the proposed policy with that of the policy obtained from the MDP model and other known polices, which are also proposed for the bandwidth asymmetry mobile wireless networks. Finally, we conclude this chapter in Section 5.6.

5.2 MDP Formulation of CAC for the MINCost Problem

5.2.1 Problem Formulation

We consider a cell in a multi-service mobile wireless network with bandwidth asymmetry. Suppose calls from M classes share B_u and B_d units of bandwidth resources in a cell, where B_u and B_d denote the uplink bandwidth and the downlink bandwidth, respectively. Since blocking a handoff call may incur more cost than blocking a new call, we treat the handoff calls and the new calls as different call classes in our system model. Call requests of class i ($1 \leq i \leq M$) arrive according to the Poisson process with parameter λ_i. A call of class i ($1 \leq i \leq M$) demands b_i^u and b_i^d bandwidth on uplink and downlink, respectively. The connection holding time of the class i calls is exponentially distributed with mean $1/\mu_i$. The system state is composed of the number of each call class in the system and it is determined by the control decisions made by admission control policy and random events. The control decisions include call acceptance and call rejection, and the random events involve call arrival, call connection completion and call handoff. When a call arrives, the system needs to decide whether the call can be accepted or not according to a certain CAC policy based on current system state. Costs can be associated with the decisions. Thus the admission control problem can be viewed as a continuous time Markov decision process. A Markov

decision process is a sequential decision problem where the set of actions, rewards and transition probabilities depend only on the current state of the system and the current decision selected. The history of the problem has no effect on the current decision. By solving the MDP problem, we may find the optimal admission policy, which results in minimum average cost.

In the following, we formulate the admission control policy for the MINCost problem into an MDP model. The MINCost problem is to minimize a linear objective function to obtain the minimum average cost.

The basic ingredients of an MDP function include system states, actions, transitions, costs and an objective function. Let $\mathbf{x} = (x_1, \cdots, x_M)$ denote the system state, where x_i represents the number of class i calls in the system. The feasible system states should satisfy $\sum_{i=1}^{M} b_i^u x_i \leq B_u$ and $\sum_{i=1}^{M} b_i^d x_i \leq B_d$ simultaneously. Thus, the set of the feasible system states, denoted by S, is finite. Let W and w denote the set of random events and individual random event, respectively. There are two events in the system: call arrival (w_a) and call departure (w_d) and thus $W = \{w_a, w_d\}$. When a call arrives ($w = w_a$), a decision needs to be made to accept or reject the call. No decision is needed for the call departure event ($w = w_d$), which could be call completion in the cell under consideration or call handoff between cells. The set of control space Y is defined as $Y = \{y_a, y_r\}$, where y_a and y_r signify acceptance and rejection, respectively.

In an infinite Markov decision process with a finite state-space, state $\mathbf{x}$ ($\mathbf{x} \in S$) transits to state $\mathbf{x}'$ ($\mathbf{x}' \in S$) in a time interval with a given probability $P_{\mathbf{x}\mathbf{x}'}$, which depends on a decision from U at the current state. The time interval between state transitions is called "stage". During the k_{th} stage, the system is at the state $\mathbf{x}(t_k)$ ($\mathbf{x}(t_k) \in S$) and the control $y(t_k)$ ($y(t_k) \in Y$) is applied then the system transits to $\mathbf{x}(t_{k+1})$ ($\mathbf{x}(t_{k+1}) \in S$). During the transition from the k_{th} stage to the $(k+1)_{th}$ stage, the decision $y(t_k)$ ($y(t_k) \in Y$) may incur a cost $\int_{t_k}^{t_{k+1}} g(\mathbf{x}(t_k), y(t_k))dt$, where $g(\cdot)$ is a given cost function. Let y_k denote $y(t_k)$ for simplicity. Then the goal of our admission control problem is to find the optimal policy $\pi^* = (y_1^*, y_2^*, \cdots)$ to minimize the average cost. The objective average cost function can be formulated as

$$\min \lim_{N\to\infty} \frac{1}{E\{t_N\}} E\{\sum_{k=1}^{N} G_k\} , \tag{5.1}$$

where

$$G_k = \int_{t_k}^{t_{k+1}} g(\mathbf{x}(t_k), y(t_k))dt \tag{5.2}$$

is the cost of the k_{th} stage. The cost could be composed by the revenue (negative cost) of call's acceptance and the cost of call's rejection. The revenue may be associated with call duration and the cost may be determined by call class. As it is well recognized that the average duration of a specific call class is usually known, we assume that the cost is associated with the call class only for mathematical tractability. Thus the function $g(\cdot)$ does not depend on the length of time spent at a particular state. (5.1) is expressed as

$$\min \lim_{N\to\infty} \frac{1}{E\{t_N\}} E\{\sum_{k=1}^{N} g(\mathbf{x}(t_k), y(t_k))\} \ . \tag{5.3}$$

In (5.1) and (5.3), N is an arbitrary positive integer to denote the number of states that the system has experienced. In order to obtain the average cost of N states, we need to compute the mean total cost of N states and the mean time that the system spends on these states. Then we let N go to infinity and obtain the average system cost per unit time under a specific admission control policy.

Next we define the system state transition probabilities. Assume that there are total M call classes. The calls of class i $(1 \leq i \leq M)$ arrive according to the Poisson process with parameter λ_i and the connection holding time for the class i $(1 \leq i \leq M)$ calls is exponentially distributed with mean $1/\mu_i$. We define the rate of all events' occurrences starting from a state $\mathbf{x}$ as the overall rate $\Lambda_{\mathbf{x}}$, which is the sum of the rates of all possible events and is given by

$$\Lambda_{\mathbf{x}} = \sum_{i=1}^{M} (\lambda_i + x_i\mu_i) \ . \tag{5.4}$$

$\Lambda_{\mathbf{x}}$ can be regarded as the average rate that the system leaves state $\mathbf{x}$. Thus $1/\Lambda_{\mathbf{x}}$ is the average time that the system stays at state $\mathbf{x}$. In order to establish the optimization equation, we still need to obtain the average system transition time, which is defined as the average time that the system transits from state $\mathbf{x} = (x_1, \cdots, x_i, \cdots, x_M)$ $(\mathbf{x} \in S)$ to $\mathbf{x}' = (x'_1, \cdots, x'_i, \cdots, x'_M)$ $(\mathbf{x}' \in S)$ under control y $(y \in Y)$. We assume that the control decision takes effect immediately when the decision

is made. Thus, the average transition time is determined by the average time spent at state $\mathbf{x}'$, which is $1/\Lambda_{\mathbf{x}'}$. We use $\overline{\tau}_{\mathbf{x}}(y)$ to denote the average transition time from state $\mathbf{x}$ to state $\mathbf{x}'$. Thus

$$\overline{\tau}_{\mathbf{x}}(y) = \frac{1}{\Lambda_{\mathbf{x}'}} \ . \tag{5.5}$$

The system state transition probability under the control y ($y \in Y$) is given by

$$P_{\mathbf{x}\mathbf{x}'}(y) = \begin{cases} \lambda_i/\Lambda_{\mathbf{x}'}, & w = w_a, 1 \leq i \leq M \\ x_i'\mu_i/\Lambda_{\mathbf{x}'}, & w = w_d, x_i' > 0, 1 \leq i \leq M \end{cases} \ . \tag{5.6}$$

So far, we have formulated the admission control problem in the asymmetric bandwidth allocation mobile wireless networks as an average cost MDP problem. Next we solve the MDP problem to obtain the optimal policy.

Let v^* denote the optimal average cost. v^* should satisfy the Bellman's optimality equation

$$v^*\overline{\tau}_{\mathbf{x}}(y) + h(\mathbf{x}) = \min_{y \in Y} \left[g(\mathbf{x}, y) + \sum_{\mathbf{x}' \in S} P_{\mathbf{x}\mathbf{x}'}(y) h(\mathbf{x}') \right] \ \forall \mathbf{x} \in S \ , \tag{5.7}$$

where $h(\mathbf{x})$ is the corresponding differential cost and $\overline{\tau}_{\mathbf{x}}(y)$ is the expected value of transition time from state $\mathbf{x}$ to the next state under the control y. We may use the policy iteration to solve (5.7) to obtain v^* and at the same time to obtain the optimal policy $\pi^* = (y_1^*, y_2^*, \cdots)$. Since there are many existing methods to solve the MDP problem [97], we will not discuss the solving process further in this chapter.

5.3 Monotonicity Properties of Value Function

In Section 5.2, we have formulated the CAC problem as an MDP problem. In this section, we use event based dynamic programming [98] to derive some properties of value function.

5.3.1 Value Function

First we need to define the value function. Let $V_n(\mathbf{x})$ denote the minimum total cost over n stages from an initial state $\mathbf{x}$, which can be expressed as

$$V_n(\mathbf{x}) = \min E\left\{\sum_{k=1}^{n} G_k\right\} . \tag{5.8}$$

Then (5.1) could be rewritten as

$$\lim_{n\to\infty} \frac{1}{E\{t_n\}} V_n(\mathbf{x}) . \tag{5.9}$$

From (5.9), we know that the properties of the value function (5.8) decide the properties of the objective average cost function (5.1).

Let $\mathbf{x}_k$ and y_k denote $\mathbf{x}(t_k)$ and $y(t_k)$ respectively and we define the cost function as

$$g(\mathbf{x}_k, y_k) = \begin{cases} c_i & \textit{reject a class } i \textit{ call} \\ r_i & \textit{accept a class } i \textit{ call} \\ 0 & \textit{others} \end{cases} , \tag{5.10}$$

where c_i is the cost of rejecting a class i call and r_i is the cost of accepting a class i call (it can be interpreted as a reward equal to $-r_i$). Without loss of generality, we assume that $\sum_{i=1}^{M}(\lambda_i + \min(\left\lfloor B_u/b_i^u \right\rfloor, \left\lfloor B_d/b_i^d \right\rfloor)\mu_i) = 1$, where $\lfloor\delta\rfloor$ is the greatest integer smaller than δ ($\delta > 0$). Let L_i denote $\min(\left\lfloor B_u/b_i^u \right\rfloor, \left\lfloor B_d/b_i^d \right\rfloor)$. Then the optimal cost value function $V(\cdot)$ satisfies

$$\begin{aligned} V_n(\mathbf{x}) = & \sum_{i=1}^{M} \lambda_i \min(V_{n-1}(\mathbf{x}+\mathbf{e}_i) + r_i, V_{n-1}(\mathbf{x}) + c_i) \\ & + \sum_{i=1}^{M} x_i \mu_i V_{n-1}(\mathbf{x}-\mathbf{e}_i) + \sum_{i=1}^{M} (L_i - x_i)\mu_i V_{n-1}(\mathbf{x}) \end{aligned} , \tag{5.11}$$

where $\mathbf{e}_i$ is the i_{th} unity vector and is expressed as

$$\mathbf{e}_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} i_{th} \ . \tag{5.12}$$

$V_{n-1}(\mathbf{x})$ denotes the minimum total cost over $n-1$ stages from the initial state $\mathbf{x}$. Since the minimum total cost over n stages can be express as the summation of the minimum total cost over $n-1$ and the cost of the last stage, we can express $V_n(\mathbf{x})$ by $V_{n-1}(\mathbf{x})$ as shown in (5.11). In (5.11), the first term is the cost incurred by the arrival of a class i call. Here, there are two decision options. Accepting a class i call $(\mathbf{x}+\mathbf{e}_i)$ may incur a cost r_i while rejecting the call may incur a cost c_i. The second term is the contribution to the cost due to call completion or handoff. The last term is a consequence of the uniformization. In order to prevent the state from leaving the state space S, we assume that $V_n(\mathbf{x}) = \infty$ if $\mathbf{x} \notin S$.

5.3.2 Event-based Dynamic Programming

In the following, we extend the properties in [20] and employ the event-based dynamic programming approach [98] to deduce some properties of the value function (5.11) for the bandwidth asymmetry multi-service mobile networks.

Let operator $T_{AC(i)}$ model the admission decision on the arrival of a class i call. Then

$$T_{AC(i)}V_n(\mathbf{x}) = \min(r_i + V_n(\mathbf{x}+\mathbf{e}_i), c_i + V_n(\mathbf{x})) \ . \tag{5.13}$$

Let the operator $T_{D(i)}$ model the departure of a class i call, which is defined as

$$T^k_{D(i)}V_n(\mathbf{x}) = \begin{cases} V_n(\mathbf{x}-\mathbf{e}_i) & if\ x_i \geq k \\ V_n(\mathbf{x}) & others \end{cases} , \tag{5.14}$$

where k is the number of class i calls in the system and $k = 1, \cdots, min(\lfloor B_u/b_i^u \rfloor, \lfloor B_d/b_i^d \rfloor)$.

Thus (5.11) could be rewritten as

$$V_n(\mathbf{x}) = \sum_{i=1}^{M} \lambda_i T_{AC(i)} V_{n-1}(\mathbf{x}) + \sum_{i=1}^{M} \mu_i \sum_{k=1}^{L_i} T_{D(i)}^k V_{n-1}(\mathbf{x}) \tag{5.15}$$

and we define $V_0(\mathbf{x}) = 0$ ($\mathbf{x} \in S$). The following lemmas are needed to be established for the optimal policy of the MINCost problem. Note that the following lemmas and theorem are obtained based on stable traffic load conditions, which means that λ_i and μ_i in (5.15) do not change over time.

Lemma 1: For all $\mathbf{x} \in S$, $1 \leq j \leq M$ and $n \geq 0$, $V_n(\mathbf{x}) \leq V_n(\mathbf{x} + \mathbf{e}_j)$.

Proof: Obviously, $V_0(\mathbf{x}) \leq V_0(\mathbf{x} + \mathbf{e}_j)$. We need to prove that if $V_{n-1}(\mathbf{x})$ satisfies this inequality, so does $T_{AC(i)}V_{n-1}(\mathbf{x})$ and $T_{D(i)}V_{n-1}(\mathbf{x})$. Since the inequality is maintained under linear combinations, then the lemma can be proved directly by induction on n.

First, we consider $T_{AC(i)}V_{n-1}(\mathbf{x})$. Suppose that $V_{n-1}(\mathbf{x}) \leq V_{n-1}(\mathbf{x}+\mathbf{e}_j)$. From the definition of $T_{AC(i)}V_{n-1}(\mathbf{x})$, we know that $\min(r_i + V_{n-1}(\mathbf{x}+\mathbf{e}_i), c_i + V_{n-1}(\mathbf{x})) \leq \min(r_i + V_{n-1}(\mathbf{x}+\mathbf{e}_i+\mathbf{e}_j), c_i + V_{n-1}(\mathbf{x}+\mathbf{e}_j))$. Thus $T_{AC(i)}V_{n-1}(\mathbf{x})$ also satisfies the inequality. In terms of $T_{D(i)}V_{n-1}(\mathbf{x})$, it is easy to prove that $T_{D(i)}V_{n-1}(\mathbf{x}) \leq T_{D(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_j)$ from the definition (5.14). Thus we have proved that $V_n(\mathbf{x}) \leq V_n(\mathbf{x}+\mathbf{e}_j)$, which means that $V_n(\mathbf{x})$ is non-decreasing for all states $\mathbf{x} \in S$ for all j. ■

Lemma 2: For all n and $\mathbf{x} \in S$,

$$V_n(\mathbf{x} + \mathbf{e}_i) + V_n(\mathbf{x} + \mathbf{e}_j) \leq V_n(\mathbf{x}) + V_n(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)\ . \tag{5.16}$$

Proof: It is clear that $V_0(\cdot)$ satisfies the above inequality. We follow the same idea used in the proof of Lemma 1. If $V_{n-1}(\mathbf{x})$ satisfies the above inequality, so do $T_{AC(i)}V_{n-1}(\mathbf{x})$ and $T_{D(i)}V_{n-1}(\mathbf{x})$. Then the lemma follows directly by induction.

We consider $T_{AC(i)}V_{n-1}(\mathbf{x})$ first. Let y_1, y_2, y_3, and y_4 denote the access control decision made for $T_{AC(i)}V_{n-1}(\mathbf{x}+\mathbf{e}_i)$, $T_{AC(i)}V_{n-1}(\mathbf{x}+\mathbf{e}_j)$, $T_{AC(i)}V_{n-1}(\mathbf{x})$ and $T_{AC(i)}V_{n-1}(\mathbf{x}+\mathbf{e}_i+\mathbf{e}_j)$, respectively. Given that $V_{n-1}(\mathbf{x} + \mathbf{e}_i) + V_{n-1}(\mathbf{x} + \mathbf{e}_j) \leq V_{n-1}(\mathbf{x}) + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$,

a) If $y_1 = y_2 = y_a$,

$$\begin{aligned} &T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_j) = \\ &r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i) + r_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)\ . \end{aligned} \tag{5.17}$$

When $y_3 = y_4 = y_a$,

$(5.17) \leq r_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i) + r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i + \mathbf{e}_j) = T_{AC(i)}V_{n-1}(\mathbf{x}) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$.

When $y_3 = y_4 = y_r$,

$(5.17) \leq c_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i) + c_i + V_{n-1}(\mathbf{x} + \mathbf{e}_j) \leq c_i + V_{n-1}(\mathbf{x}) + c_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$

$= T_{AC(i)}V_{n-1}(\mathbf{x}) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$.

When $y_3 = y_a, y_4 = y_r$,

$(5.17) \leq c_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i) + r_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j) = T_{AC(i)}V_{n-1}(\mathbf{x}) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$.

When $y_3 = y_r, y_4 = y_a$, we need to combine $V_{n-1}(\mathbf{x} + 2\mathbf{e}_i) + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j) \leq V_{n-1}(\mathbf{x} + \mathbf{e}_i) + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i + \mathbf{e}_j)$ with $V_{n-1}(\mathbf{x} + \mathbf{e}_i) + V_{n-1}(\mathbf{x} + \mathbf{e}_j) \leq V_{n-1}(\mathbf{x}) + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$ together. Thus $V_{n-1}(\mathbf{x} + 2\mathbf{e}_i) + V_{n-1}(\mathbf{x} + \mathbf{e}_j) \leq V_{n-1}(\mathbf{x}) + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i + \mathbf{e}_j)$.

Then, $(5.17) \leq r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i) + c_i + V_{n-1}(\mathbf{x} + \mathbf{e}_j) \leq c_i + V_{n-1}(\mathbf{x}) + r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i + \mathbf{e}_j) = T_{AC(i)}V_{n-1}(\mathbf{x}) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$.

Following the similar way, we can prove that $T_{AC(i)}(V_{n-1}(\mathbf{x}))$ also satisfies (5.16) when $y_1 = y_2 = y_r$.

b) If $y_1 = y_a, y_2 = y_r$,

$$\begin{aligned} &T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_j) = \\ &r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i) + c_i + V_{n-1}(\mathbf{x} + \mathbf{e}_j)\ . \end{aligned} \tag{5.18}$$

When $y_3 = y_4 = y_a$,

$(5.18) \leq r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i) + r_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j) \leq r_i + V_{n-1}(\mathbf{x} + \mathbf{e}_i) + r_i + V_{n-1}(\mathbf{x} + 2\mathbf{e}_i + \mathbf{e}_j)$

$= T_{AC(i)}V_{n-1}(\mathbf{x}) + T_{AC(i)}V_{n-1}(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$.

Under other conditions ($y_3 = y_4 = y_r$, $y_3 = y_a, y_4 = y_r$ and $y_3 = y_r, y_4 = y_a$), the proof is similar to that of a). Thus we prove that $T_{AC(i)}(V_{n-1}(\mathbf{x}))$ satisfies inequality (5.16). Since we have

assumed that $V_{n-1}(\mathbf{x})$ satisfies (5.16), it is easy to prove that $T_{D(i)}(V_{n-1}(\mathbf{x}))$ also satisfies (5.16). Thus we have proved the value function $V_n(\mathbf{x})$ satisfies the inequality (5.16). ■

From Lemma 1 and Lemma 2, we can obtain the following theorem:

Theorem 1: To minimize the average cost of the CAC policy in the bandwidth asymmetry mobile wireless networks, a call of class i can be accepted if and only if $x_j < Th_j(x_1, \cdots, x_i, x_k, \cdots, x_M)$ $(j \neq i)$, where $Th_j(x_1, \cdots, x_i, x_k, \cdots, x_M)$ is a threshold of the class j calls when the system state is $\mathbf{x}$, $\mathbf{x} = (x_1, \cdots, x_i, x_j, x_k \cdots, x_M),\ \mathbf{x} \in S$.

Proof: Let us rewrite (5.16) as

$$V_n(\mathbf{x} + \mathbf{e}_i) - V_n(\mathbf{x}) \leq V_n(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j) - V_n(\mathbf{x} + \mathbf{e}_j)\ . \tag{5.19}$$

From (5.19) we know that if $V_n(\mathbf{x} + \mathbf{e}_i)$ is greater than $V_n(\mathbf{x})$, $V_n(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j)$ is also greater than $V_n(\mathbf{x}+\mathbf{e}_j)$. $V_n(\mathbf{x}+\mathbf{e}_i) > V_n(\mathbf{x})$ means that accepting a class i call will incur more cost than rejecting a class i call after n stages from the initial state $\mathbf{x}$ while $V_n(\mathbf{x} + \mathbf{e}_i + \mathbf{e}_j) > V_n(\mathbf{x} + \mathbf{e}_j)$ means that accepting a class i call will incur more cost than rejecting a class i call after n stages from the initial state $\mathbf{x}+\mathbf{e}_j$. From Lemma 1, we know that $V_n(\mathbf{x}+\mathbf{e}_i+a\cdot\mathbf{e}_j) \geq V_n(\mathbf{x}+\mathbf{e}_i+\mathbf{e}_j)$ where a is an arbitrary integer and $a \geq 1$. If a class i call is rejected at $\mathbf{x} = (x_1, \cdots, x_i, x_j, \cdots, x_M)$, it is also rejected at $\mathbf{x} + a \cdot \mathbf{e}_j = (x_1, \cdots, x_i, x_j + a, \cdots, x_M)$. Thus we can find a threshold $Th_j(x_1, \cdots, x_i, x_k, \cdots, x_M)$ at the system state $\mathbf{x}$, such that the class i call can be accepted, if the number of class j calls in the system is smaller than the threshold. Otherwise, the cost of accepting an arrival class i call will be greater than that of rejecting it, if $x_j \geq Th_j(x_1, \cdots, x_i, x_k, \cdots, x_M)$ at the system state $\mathbf{x}$. The call should be rejected. Therefore, we have proved Theorem 1. ■

5.3.3 Discussions

Next, we discuss the applications of Theorem 1 in different system models. Let us consider a simple system model first. We assume that there are two classes of calls ($M = 2$): handoff calls (class 1) and new calls (class 2), in the system. The channel holding time of both handoff calls and new calls are exponentially distributed with mean $1/\mu$ ($\mu_1 = \mu_2 = \mu$). The bandwidth requirements of a handoff call and a new call are identical and equal to b ($b_1^u = b_2^u = b$ *and* $b_1^d = b_2^d = b$).

In this simple model, the system state is determined by only the total number of calls in the system. Lemma 2 can be rewritten as

$$2V_n(x+1) \leq V_n(x) + V_n(x+2) \ , \tag{5.20}$$

where x denotes the number of calls in the system. This property is called convexity. We may change (5.20) as

$$V_n(x+1) - V_n(x) \leq V_n(x+2) - V_n(x+1) \ . \tag{5.21}$$

From (5.21), we may find that if an arrival call is rejected at state $\mathbf{x}$, which means $V_n(\mathbf{x}+\mathbf{e}_i) - V_n(\mathbf{x}) > c_i - r_i (i = 1 \text{ or } i = 2)$ (c_i and r_i are the same as that defined in (5.10)), the call should also be rejected at state $\mathbf{x}+e_1, \mathbf{x}+e_2, \cdots$. It is obvious that threshold policy could be the optimal policy for the MINCost problem in such system. Indeed, it has been proved that the GC scheme is the optimal policy for the MINCost problem in such simple environment in [19]. So our theorem matches the result of [19] in the simple system model.

Next, let us consider Theorem 1 in a multi-service mobile wireless network with asymmetric bandwidth allocation. We classify all the calls into two categories: RT calls (class 1) and NRT calls (class 2) , where an RT call requires the same bandwidth on uplink and downlink and an NRT call requires asymmetric bandwidth on uplink and downlink. The RT calls and the NRT calls have different connection holding time ($1/\mu_1 \neq 1/\mu_2$) and bandwidth requirements ($b_1^u \neq b_2^u \ and \ b_1^d \neq b_2^d$).

From Theorem 1, we may find that when system state is $\mathbf{x} = (x_1, x_2)$, an arrival RT (NRT) call can be accepted only if the number of NRT (RT) calls in the system does not exceed a certain threshold. This threshold may change with the system state $\mathbf{x}$. Thus the optimal policy for the MINCost problem in such asymmetric bandwidth allocation multi-service wireless networks should be a dynamic threshold policy. However, when the base of system states becomes large, the computational complexity for solving the Bellman equation (5.7) is prohibitively high and it may be very time-consuming to decide the corresponding threshold values. In a real system, both the RT calls and the NRT calls may have handoff attempts and this makes the procedure of finding the optimal solution more challenging. It is unlikely to design an optimal CAC policy according to the above analysis by on-line computing the dynamic thresholds.

To address the above-mentioned difficulty, we propose a new admission policy called Call-Rate-based Dynamic Threshold (CRDT) admission control policy, which aims at approximating the optimal CAC policy deduced from the analytical model for bandwidth asymmetry multi-service wireless networks. In order to design an effective and efficient policy, we need to analyze the system states and make the decisions based on the system states. We can divide all the system states into two sets. In some states, all calls can be accepted and we name these states "unsaturated states". While in some states, only the calls of some classes or no calls can be accepted and we name these states "saturated states". There are two main tasks for an admission policy: 1) Judging the current system state is unsaturated or saturated; 2) Deciding what policy could be used if the system is in the saturated states. Theorem 1 provides a rule to determine the optimal policy to solve the MINCost problem in asymmetric bandwidth networks when the system is at the saturated state. However, how to decide the system is at a unsaturated state or a saturated state and the corresponding thresholds depends on the complicated computation of solving the Bellman equation (5.7). In the proposed CRDT policy, the bandwidth used by the RT calls and the NRT calls respectively is used to decide the current system state. When the bandwidth used by the RT calls or the NRT calls reaches a pre-calculated threshold, we deem that the current system is at the saturated state. In light of Theorem 1, when system is at the saturated state, the decision made for an arrival RT (NRT) call is determined by the estimated arrival rate of the NRT (RT) calls. In stead of computing the threshold of the number of the RT calls or the NRT calls, we use a measurable parameter, the call arrival rate, to make the decision and thus decrease the computational complexity. When the bandwidth used by the RT (NRT) calls in the system reaches the bandwidth threshold set for the RT (NRT) calls on uplink and downlink, whether an arrival RT (NRT) call can be accepted or not is determined by the NRT (RT) call arrival rate. If the NRT (RT) call arrival rate is greater than a reference rate, the arrival RT (NRT) call is blocked. In the next section, we will describe in detail how to compute the bandwidth threshold and the reference rate value for a specific class of calls.

5.4 Call-Rate-based Dynamic Threshold (CRDT) Admission Control Policy

5.4.1 Computing Threshold

In the underlying multi-service mobile wireless networks, we assume that there are four classes of calls: handoff RT call, handoff NRT call, new RT call and new NRT call. An RT call requires same bandwidth on uplink and downlink while an NRT call requires asymmetric bandwidth on two links [8, 91]. The RT call arrival rate and the NRT call arrival rate follow the Poisson distribution with mean λ_{RT} and λ_{NRT}, respectively. The connection holding time of the RT calls and the NRT calls is exponentially distributed with mean $1/\mu_{RT}$ and $1/\mu_{NRT}$, respectively. The system asymmetry factor Γ_s and the NRT call asymmetry factor Γ_{NRT} are defined as $\Gamma_s = \frac{B_d}{B_u}$ and $\Gamma_{NRT} = \frac{b^d_{NRT}}{b^u_{NRT}}$, respectively.

Let us consider a system at steady states with heavy traffic load. From statistical point of view, if no bandwidth is wasted, the uplink bandwidth and the downlink bandwidth used by the RT calls and the NRT calls should satisfy

$$\rho_{RT} \times b^u_{RT} + \rho_{NRT} \times b^u_{NRT} = B_u \tag{5.22}$$

and

$$\rho_{RT} \times b^d_{RT} + \rho_{NRT} \times b^d_{NRT} = B_d \ , \tag{5.23}$$

where ρ_{RT} and ρ_{NRT} denote the traffic load brought by RT calls and NRT calls, respectively. b^u_{RT} (b^u_{NRT}) and b^d_{RT} (b^d_{NRT}) denote the uplink and downlink bandwidth requirements of each RT (NRT) call. Total uplink and downlink bandwidth of the system are represented as B_u and B_d, respectively. Let Γ_s and Γ_{NRT} denote the system asymmetry factor and the asymmetry factor of NRT calls, respectively. Given that $b^u_{RT} = b^u_{RT}$, $b^d_{NRT} = \Gamma_{NRT} b^u_{RT}$ and $B_d = \Gamma_s B_u$, (5.23) minusing (5.22) yields

$$\rho_{NRT} = \frac{\Gamma_s - 1}{\Gamma_{NRT} - 1} \times \frac{B_u}{b^u_{NRT}} \quad . \tag{5.24}$$

Since $\rho_{NRT} = \frac{\lambda_{NRT}}{\mu_{NRT}}$, we can obtain the average NRT call arrival rate at this system state as

$$\lambda_{NRT} = \frac{\Gamma_s - 1}{\Gamma_{NRT} - 1} \times \frac{B_u}{b^u_{NRT}} \times \mu_{NRT} \ . \tag{5.25}$$

The average RT call arrival rate at this system state can be obtained by combining (5.22) and (5.24) and it is shown as

$$\lambda_{RT} = \frac{\Gamma_{NRT} - \Gamma_s}{\Gamma_{NRT} - 1} \times \frac{B_u}{b^u_{RT}} \times \mu_{RT} \ . \tag{5.26}$$

Let us use $\overline{\lambda}_{RT}$ and $\overline{\lambda}_{NRT}$ to denote the value of λ_{RT} and λ_{NRT} at this system state. $\overline{\lambda}_{RT}$ and $\overline{\lambda}_{NRT}$ are used as the reference rate for the RT calls and the NRT calls, respectively. The meaning of $\overline{\lambda}_{RT}$ and $\overline{\lambda}_{NRT}$ are as follows. When the RT call arrival rate is $\overline{\lambda}_{RT}$ and the NRT call arrival rate is $\overline{\lambda}_{NRT}$, the bandwidth allocated to the uplink and the downlink is able to satisfy the traffic load requirements of the RT calls and the NRT calls exactly without bandwidth waste.

We use $\overline{B}^u_{RT}$ and $\overline{B}^d_{RT}$ to denote the bandwidth used by the RT calls on the uplink and the downlink respectively when the RT call arrival rate is $\overline{\lambda}_{RT}$. Thus $\overline{B}^u_{RT} = \overline{B}^d_{RT} = \frac{\Gamma_{NRT} - \Gamma_s}{\Gamma_{NRT} - 1} B_u$. Accordingly, let $\overline{B}^u_{NRT}$ and $\overline{B}^d_{NRT}$ denote the bandwidth used by the NRT calls on the uplink and the downlink respectively when the NRT call arrival rate is $\overline{\lambda}_{NRT}$. Thus $\overline{B}^u_{NRT}$ and $\overline{B}^d_{NRT}$ are equal to $\frac{\Gamma_s - 1}{\Gamma_{NRT} - 1} B_u$ and $\frac{\Gamma_s - 1}{\Gamma_{NRT} - 1} B_u \cdot \Gamma_{NRT}$, respectively. $\overline{B}^u_{RT}$, $\overline{B}^d_{RT}$, $\overline{B}^u_{NRT}$ and $\overline{B}^d_{NRT}$ are just four bandwidth thresholds set for the RT calls and the NRT calls in our policy.

5.4.2 Call Rate Estimation

Our policy is composed of two functional components: call rate estimation algorithm and admission control algorithm. Let us describe the call rate estimation algorithm first. The call rate estimation algorithm is based on the exponential smoothing method [99]. We define a certain period of time (T) as the time interval between two estimations. The call rate estimation is performed at the end of each time interval. For example, at the end of time interval N, the system scales the average call arrival rate λ_N of the current time interval and estimates the call arrival rate of the time interval $(N + 1)$ by using (5.27), where $\hat{\lambda}_N$ is the estimated call arrival rate obtained in the time interval $(N - 1)$ and α $(0 < \alpha < 1)$ is a parameter used to determine how fast the algorithm responds to the changes of the arrival rate. At the beginning, we can set $\hat{\lambda}_1 = \lambda_1$ as the initial value and then

use (5.27) recursively to estimate the call arrival rate of the next time interval.

$$\hat{\lambda}_{(N+1)} = \alpha\lambda_N + (1-\alpha)\hat{\lambda}_N \ . \tag{5.27}$$

5.4.3 CRDT Policy

Next, we present the proposed admission control policy, which needs to make use of above call rate estimation algorithm. In order to simplify the description of the proposed CRDT policy, we assume that there is sufficient uplink and downlink bandwidth to satisfy the call requests. If the remaining bandwidth on the uplink and/or the downlink cannot satisfy the bandwidth requirement of the arrival call, the call is blocked directly. Then it does not need to make a CAC decision in this case. The proposed CRDT policy can be described as follows.

When a handoff RT call arrives, it is accepted since there is sufficient bandwidth on uplink and downlink to satisfy the call bandwidth requirement. On the other hand, when a new RT call arrives, the system checks the uplink bandwidth and the downlink bandwidth occupied by the RT calls in the system ($\hat{B}^u_{RT}, \hat{B}^d_{RT}$). If accepting the call dose not cause the bandwidth used by the RT calls to exceed the threshold $\overline{B}^u_{RT}$ and $\overline{B}^d_{RT}$ on the uplink and the downlink respectively, the call can be accepted. Otherwise, the system checks the estimated NRT call arrival rate $\hat{\lambda}_{NRT}$ in the current time interval. If $\hat{\lambda}_{NRT} < \overline{\lambda}_{NRT}$, the arrival new RT call can be accepted; else, it is blocked.

When a handoff NRT call arrives, the system checks the uplink bandwidth and the downlink bandwidth occupied by the NRT calls in the system ($\hat{B}^u_{NRT}, \hat{B}^d_{NRT}$). If accepting the call does not cause the bandwidth used by the NRT calls to exceed the threshold $\overline{B}^u_{NRT}$ and $\overline{B}^d_{NRT}$ on the uplink and the downlink respectively, the call can be accepted. Otherwise, the system checks the estimated RT call arrival rate $\hat{\lambda}_{RT}$ in the current time interval. If $\hat{\lambda}_{RT} < \overline{\lambda}_{RT}$, the arrival handoff NRT call can be accepted; else, it is blocked.

The treatment to the new NRT calls is similar to that of the handoff NRT call except that only if $\hat{\lambda}_{RT} < \overline{\lambda}_{RT} \cdot \Delta$, the arrival new NRT call can be accepted, where Δ ($0 < \Delta < 1$) is a design parameter used to guarantee the priorities of the RT calls and the handoff NRT calls. Since the new NRT calls have lowest priority, it is necessary to limit the number of the new NRT calls in the system and thus avoid these low priority calls overusing system resources. We will discuss in detail

the effect of this parameter on the system performance in the next section. Figure 5.1 shows the pseudo code of the proposed algorithm.

```
if (enough uplink and downlink bandwidth)
  if (handoff RT call)
    accept

  if (new RT call)
    if ((B̂^u_RT + b^u_RT) < B̄^u_RT and (B̂^d_RT + b^d_RT) < B̄^d_RT)
      accept
    else if (λ̂_NRT < λ̄_NRT)
      accept
    else
      reject

  if (handoff NRT call)
    if ((B̂^u_NRT + b^u_NRT) < B̄^u_NRT and (B̂^d_NRT + b^d_NRT) < B̄^d_NRT)
      accept
    else if (λ̂_RT < λ̄_RT)
      accept
    else
      reject

  if (new NRT call)
    if ((B̂^u_NRT + b^u_NRT) < B̄^u_NRT and (B̂^d_NRT + b^d_NRT) < B̄^d_NRT)
      accept
    else if (λ̂_RT < λ̄_RT · Δ)
      accept
    else
      reject
else
  reject
```

Figure 5.1: Pseudo code of the proposed CRDT policy.

5.5 Performance Evaluation

In this section, we use simulation experiments to examine the performance of the CRDT policy and compare the average cost of the CRDT policy with that of some known CAC policies. We assume that the call arrival is according to the Poisson process and the call connection holding time is exponentially distributed. We assume that the system allocates 10 channels on uplink and 16 channels on downlink, respectively. The parameters used in the simulation are listed in Table 5.1.

In the simulation, we choose three policies as our comparison bases. The first is the policy

Table 5.1: Traffic Model

	RT call		NRT call	
	Uplink	Downlink	Uplink	Downlink
Number of channels required per call	1	1	1	3
Mean Call Duration	180sec		600sec	
Mean Cell Dwell Time	200sec		1200sec	
	Handoff	New	Handoff	New
Rejection cost	8	4	5	1
Acceptance cost	-2	-2	-4	-4

obtained from Bellman equation (5.7). As we mentioned in Section 5.2, we may use policy iteration to obtain the optimal policy from the Bellman equation (5.7) and we call this policy "calculated policy" in our simulations. The other two are Jeon's policy [8] and the Scheme 2 in [91] which is proposed by us and we call it "Yang's policy" in the simulations. Both of these two policies are designed for the asymmetric bandwidth allocation mobile networks and good performance in terms of call blocking probabilities and bandwidth utilization has been demonstrated.

This section is composed of two parts. In the first part, we examine how the parameters (i.e., α, T and Δ) used in the CRDT policy affect the system performance. In the second part, we compare the average cost of the proposed CRDT policy with that of other three policies under two scenarios. Let q be the ratio of the number of RT calls over the number of all arrival calls. In the first scenario, we assume a static traffic load environment, which means q does not change with time dynamically. While in the second scenario, q may change with time according to a given probability distribution. Compared with the first scenario, the second scenario assumes a more dynamic environment.

5.5.1 Setting Parameters

We first examine how the system average cost is affected by the parameters, α, T and Δ, in a dynamic traffic load environment. We assume that q varies with time according to the normal distribution with mean 0.7 and variance 0.2.

In Figure 5.2 (a) and (b), we compare the average cost of the CRDT policy with different α values as a function of the new call arrival rate when T is 1 minute and 10 minutes, respectively. From Figure 5.2 (a), we find that the average cost is sensitive to the value of α when T is small

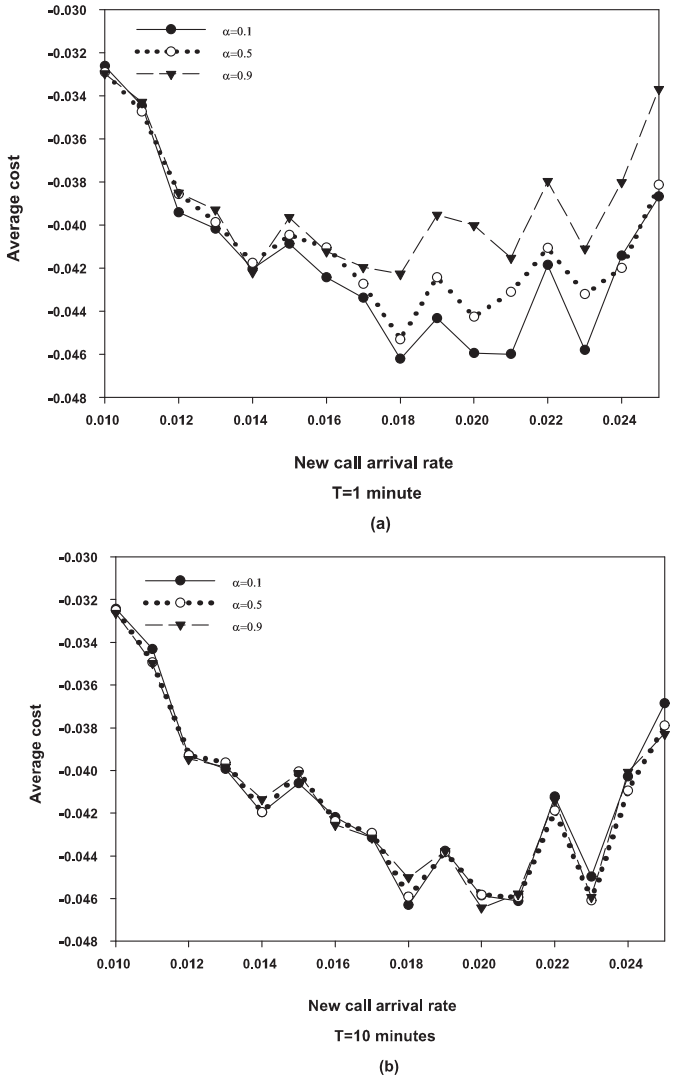

Figure 5.2: Average costs of the CRDT policy when $T = 1$ *minute* and $T = 10$ *minute*.

(1 minute) and a small value of α ($\alpha = 0.1$) results in a lower average cost. It is obvious that the estimated rate depends on the "past" estimation not the rate of the "current" time interval when T is small. When the time interval is large (10 minutes), we can find that the average costs of the CRDT policy with different α values are very close. Figure 5.3 compares the average costs when T is 1 minute, 10 minutes, 30 minutes and 1 hour, respectively. From this figure, we can find that the difference of the average costs is trivial. When the traffic load is light (new call arrival rate is smaller than 0.02), the small interval ($T <$ 1 hour) may obtain lower average cost. Thus in the

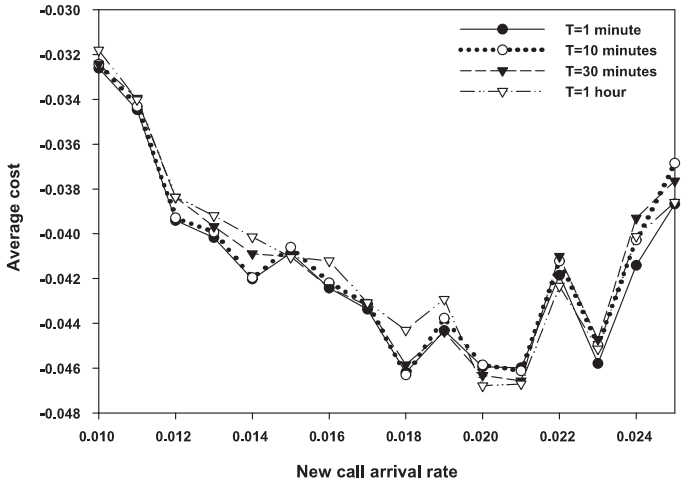

Figure 5.3: Average costs of the CRDT policy with different T.

following simulation experiments, we set α to be 0.1 and T to be 1 minute.

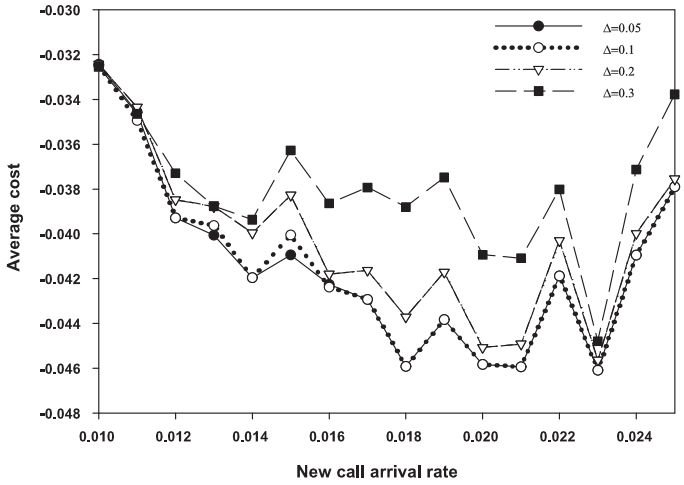

Figure 5.4: Average costs of the CRDT policy with different Δ.

Figure 5.4 shows the average costs of the CRDT policy with different Δ values. From the figure, we can observe that the average cost increases with the value of Δ. When Δ is smaller than 0.1, the difference is small. In the subsequent simulation experiments, we set Δ to be equal to 0.1.

We have conducted extensive simulation experiments for understanding the effects of different parameter settings. We show only some representative results in above figures. In the following part, we focus on performance evaluation and comparison.

5.5.2 Scenario 1

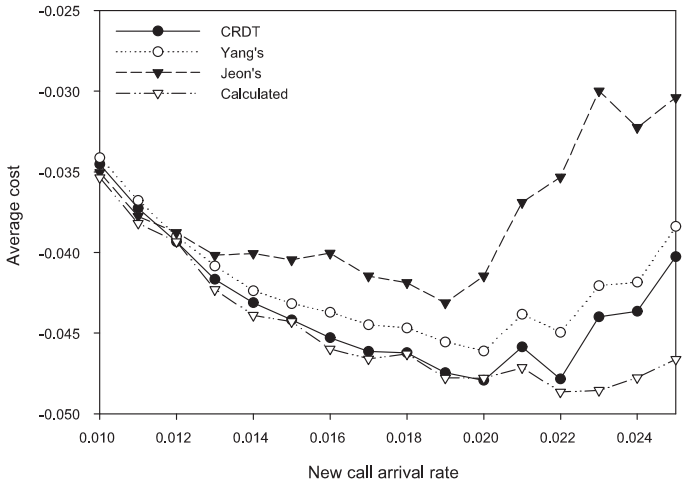

Figure 5.5: Average cost of the CAC policies when $q = 70\%$ in Scenario 1 ($T = 1\ minute, \alpha = 0.1, \Delta = 0.1$).

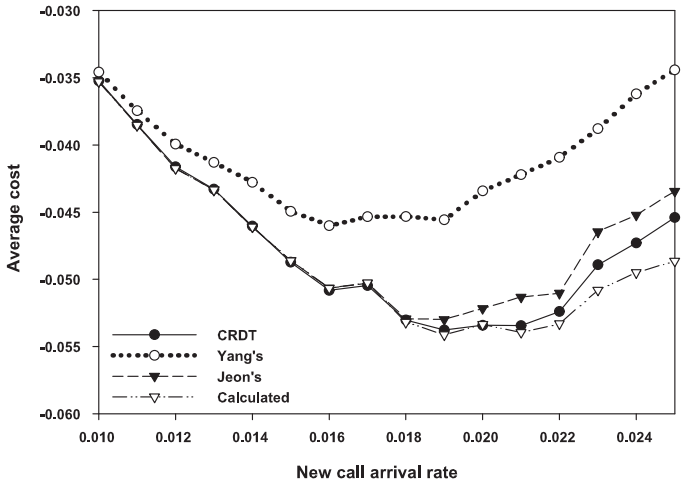

Figure 5.6: Average cost of the CAC policies when $q = 90\%$ in Scenario 1 ($T = 1\ minute, \alpha = 0.1, \Delta = 0.1$).

Figure 5.5 shows the average cost obtained from the proposed CRDT policy and other policies when $q = 70\%$. When the new call arrival rate is low, from the figure, we can observe that the average cost of the policies except Jeon's policy monotonically decreases with the new call arrival rate. The average costs obtained from the CRDT policy, the calculated policy and Yang's policy are very close and smaller than that of Jeon's policy. With the increase of the new call arrival rate,

the difference between the average cost of Yang's policy and that of the calculated policy becomes more evident while the average cost of the CRDT policy is also close to that of the calculated policy and is smaller than that of Yang's policy and Jeon's policy obviously. When the new call arrival rate is very high and the system is overloaded, the average cost of the proposed CRDT policy still smaller than that of Jeon's policy and Yang's policy.

Figure 5.6 shows the average cost of the CAC policies when $q = 90\%$. In this case, most traffic load in the system is generated by the RT calls. From the figure, we can find that the average cost of the proposed CRDT policy is very close to that of the calculated policy and is smaller than that of Yang's policy and Jeon's policy. In order to decrease the handoff call blocking probability, Yang's policy and Jeon's policy may reserve too much bandwidth for the handoff calls and thus blocking some new calls unnecessarily. With the increase of the new call arrival rate, the average costs of Yang's policy and Jeon's policy increase obviously. The proposed CRDT policy focuses on not only one specific class of calls but the average cost of the whole system and thus it can guarantee the low average cost and keeps the average cost close to that of the calculated policy.

5.5.3 Scenario 2

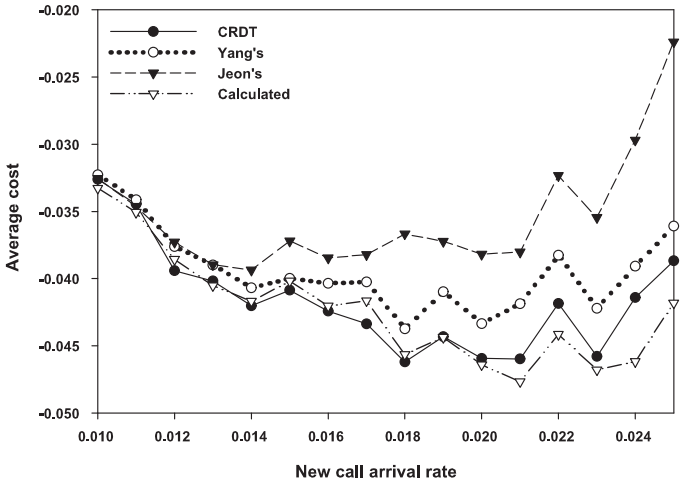

Figure 5.7: Average cost of the CAC policies when q changes with time ($T = 1\ minute, \alpha = 0.1, \Delta = 0.1$).

Figure 5.7 shows the average cost in a dynamic traffic load environment where q varies with time

according to the normal distribution with mean 0.7 and variance 0.2. We assume that the accurate mean call arrival rate can be obtained beforehand for the calculated policy and Jeon's policy. From the figure, we can find that the average cost obtained from the CRDT policy is close to that of the calculated policy and smaller than that of Yang's policy and Jeon's policy significantly when the new call arrival rate increases. Although the bandwidth thresholds are also defined for RT calls and NRT calls in Yang's policy to avoid a specific call class overusing the bandwidth, such policy with fixed thresholds may be inflexible in a dynamic traffic load environment, leading to deteriorated system performance. The average cost of Yang's policy is higher than that of the proposed CRDT policy. When the new call arrival rate is low, the average costs of Yang's policy and Jeon's policy are close to that of the CRDT policy and the calculated policy. When the new call arrival rate increases, the average cost of Yang's policy and Jeon's policy is higher than that of the proposed CRDT policy. From the simulation results, we also find that the calculated policy may not always achieve the minimum average cost. Since the optimal policy computed from (5.7) is based on fixed traffic load, it may not grantee the minimum average cost in a dynamic traffic environment. When the traffic load in the system varies over time, the proposed CRDT policy may obtain better performance since it dynamically controls the call admission according to the estimations of call arrival rate. In summary, the proposed CRDT policy provides a heuristic solution to the optimal policy for the MINCost problem in the bandwidth asymmetry mobile wireless networks.

5.6 Summary

In this chapter, we investigate the admission control policy for the MINCost problem in the bandwidth asymmetry mobile networks. By formulating the CAC problem into an MDP model and analyzing the corresponding value function, we find that the optimal admission policy for the MINCost problem in such asymmetric bandwidth allocation multi-service mobile wireless networks should have a threshold structure. The threshold specified for a class of calls may vary with the system state. Because of the prohibitively high computational complexity, it is hard to on-line calculate the threshold for each call class in a real-time system with a large system state-space. Based on the analysis, we propose a heuristic policy called Call-Rate-based Dynamic Threshold (CRDT) policy as a suboptimal solution to the MINCost problem for the bandwidth asymmetry mobile

wireless networks. The values of the thresholds in the CRDT policy can be computed readily. The numerical results show that the performance of the proposed CRDT policy is very close to that of the optimal policy obtained from the MDP model and better than that of other two known policies, which are also proposed for the multi-service mobile wireless networks with bandwidth asymmetry.

Chapter 6

Minimizing Call Blocking Probability in Multi-Service Mobile Networks

As one of the critical QoS measurements, handoff call dropping probability has drawn a lot of attention in the design of call level admission control of mobile cellular networks. In traditional mono-service mobile networks, Limited Fractional Guard Channel (LFGC) scheme has been proved be optimal for the MINBlock problem. In this chapter, we study the MINBlock problem in multi-service mobile networks and propose Distributed Multi-service Admission Control scheme (DMS-AC). By analyzing the relationship between the call admission of different classes, we decompose system overload states into overload states of individual call class and study the interrelationship of the admission of various call classes. Based on system states of local cell and information from neighboring cells, different thresholds are computed and set for each call class to prevent new calls from overusing system resources and control the number of potential handoff calls. We also conduct extensive experiments to verify the performance of DMS-AC. Numerical results show that DMS-AC is able to guarantee the handoff dropping probability of different call classes under hard constraints in a dynamic traffic load environment. Although more new NRT calls are blocked compared with another dynamic multi-guard-channel scheme, it is more reasonable to make the tradeoff between low and high priority calls and thus guarantee the QoS of high priority calls.

6.1 Introduction

With the rapid growth of mobile cellular networks, traditional simple voice and short message services cannot satisfy the increasing multimedia service requirements. Future mobile networks will provide more and more multimedia services such as audio, video, web browsing, on-line games and file transmission, etc. to the users on move. Since different applications have inherently different traffic characteristics, their QoS requirements may differ in terms of bandwidth, delay, and connection dropping probabilities. It is the networks' responsibility to fairly and efficiently allocate network resources among different users to satisfy such differentiated QoS requirements for each type of service [48].

In traditional mono-service mobile networks, only voice service is supported. Handoff calls and new calls share limited system resources. Handoff calls are assigned the highest priority since it is more undesirable to block an ongoing call than a new call. Handoff call dropping probability (P_h) is always used as a QoS measurement in the design of CAC scheme. In such resource sharing system, there usually be a tradeoff between the admission of handoff and new calls and the admission of new calls is limited in order to reserve system resources for handoff calls. How to minimize new call blocking probability (P_n) while keeping handoff call dropping probability under an acceptable low level (MINBlock) is a critical problem for the design of CAC in mobile networks. So far, many CAC schemes have been proposed to handle such call blocking probability minimization problem in mono-service mobile networks as we introduced in Chapter 2. In [19], Ramjee et al. proposed Limited Fractional Guard Channel (LFGC) scheme and proved that LFGC is an optimal solution in minimizing new call blocking probability with a hard constraint on handoff call dropping probability. LFGC uses two parameters, T and β, to control the admission of new calls. Since P_n (P_h) is proved be a monotonically decreasing (increasing) function of T and β, the authors used bisection method to find the appropriate values of these critical parameters.

Different from traditional mobile networks, not only voice service but also many data services are supported in multi-service mobile networks and both voice and data service have handoff attempts. For example, an game player may play an on-line game on a train and the train moves between different wireless communication cells in a mobile area during a certain period of time. For such users, they cannot tolerate recurrent disconnections during playing process. How to guarantee the

handoff dropping probability of different call classes below certain constraints and at the same time minimize new call blocking probability challenges the traditional CAC schemes. In [52], Chau et al. proposed multi-service admission control scheme based on LFGC scheme. Two call classes, voice and data, are considered in the literature. Emulating LFGC, multi-service LFGC uses two parameters to limit the admission of new calls of different call classes. Since bisection method is not applicable for such multi-service admission control, simulated annealing is employed to compute the thresholds for each call classes. Unfortunately, the author did not explain how to compute the critical parameters in detail. In [21], the authors proposed a Double Threshold Bandwidth Reservation (DTBR) scheme. In DTBR scheme, the total channels of each cell are divided into three regions by two bandwidth thresholds K_1 and K_2. The performance of DTBR scheme is totally determined by the parameters K_1 and K_2. However, the authors did not illustrated how to find appropriate values of these two critical parameters. Jeon et al. proposed a dynamic multi-guard-channel scheme in [8]. In Jeon's scheme, the asymmetric traffic load brought by NRT calls is considered and the size of guard channels for each traffic class on uplink and downlink is computed and set separately. The number of reserved channels is proportional to the call arrival rate, the mean call duration and the required bandwidth of each call class. This scheme tries to obtain the optimal guard channel size for each call class by estimating the call arrival rate of each call class. Although handoff calls are assigned higher priority than new calls, Jeon's scheme cannot guarantee the handoff call blocking probability of different call classes under certain constraints in a dynamic traffic load system.

In [17], the authors proposed distributed CAC scheme (DCA) for mono-service mobile networks. By using threshold to limit the admission of new calls, DCA guarantees the overload probability of the local cell and all neighboring cells under the upper bound and thus satisfies the QoS requirements of handoff calls. In this chapter, we extend DCA scheme and propose a Distributed Multi-service Admission Control (DMS-AC) to address the MINBlock problem in multi-service mobile networks. Based on the system states of local cell and information of neighboring cells, different threshold is computed and set for every call class to prevent new calls from overusing system resources and control the number of potential handoff calls from local cell to neighboring cells. In order to guarantee handoff call dropping probability of different call classes under some predefined hard

constraints, it is critical to find appropriate threshold values. Different from traditional mono-service networks, the admission of a call affects not only the handoff call dropping probability of this call class but also other call classes. Thus, the situation becomes more complicated to compute the threshold of each call class in multi-service networks. In our work, DMS-AC tries to find different thresholds for each call class according to the traffic pattern. By analyzing the relationship between the admission of different call classes, we decompose all system overload states into the overload states of individual call class and study how the calls of a specific class result in the overload states of other call classes. The details of finding appropriate thresholds are explained comprehensively in this chapter. We also conduct extensive experiments to verify the performance of the proposed DMS-AC scheme. We employ Jeon's scheme as the comparison base since it also considers asymmetric traffic load brought by NRT calls. The experiments' results show that DMS-AC can guarantee the handoff call dropping probabilities of different call classes under predefined constraints with the expense of blocking more new NRT calls in a dynamic traffic load environment. It is seasonable to make such a tradeoff to guarantee the QoS of higher priority calls.

The rest of this chapter is organized as follows. We illustrate the proposed DMS-AC scheme in Section 6.2. We first consider a simple two-cell system and present the computation process of thresholds in detail. Then, we extend the proposed scheme to a multi-cell system. Numerical results and analysis are given in Section 6.3. At last, we conclude this chapter in Section 6.4.

6.2 Distributed Multi-service Admission Control (DMS-AC)

The proposed Distributed Multi-service Admission Control (DMS-AC) scheme operates in a distributed manner. The information of system states, such as the number of calls of different call classes etc., could be exchanged between adjacent cells periodically. The base station of a cell makes an admission decision based on the state information of the cell itself (called observing cell) and its neighboring cells. DMS-AC uses threshold to limit the admission of new calls. When the number of calls of a specific class reaches the threshold of this class, new arrivals of this call class are rejected. Since the fixed thresholds may not be able to guarantee the QoS requirements when the offered traffic pattern changes, we design a dynamic threshold scheme and the threshold of a specific call class can be re-computed and reset periodically according to the change of traffic pat-

tern of the system. We define the interval between two threshold computing processes as a control period, which lasts T units of time, and the threshold of a specific call class is fixed in a control period. The duration of the control period should be associated with the dynamics of traffic load. Too long or too short interval may affect the behavior and performance of the proposed scheme. If the control interval is too short, such as few minutes, the scheme may be sensitive to the traffic burst. On the other hand, if it is too long such as several hours, the scheme may not adjust the thresholds promptly according to the traffic pattern. In this chapter, we assume that T could take the value between 15 and 60 minutes. In the rest part of this section, we first consider a simple system, which is composed of two cells, and then extend the proposed admission control scheme to a multi-cell system.

6.2.1 DMS-AC in a Two-cell System

The system we consider first is composed of two cells, denoted by C_r and C_l respectively, as shown in Figure 6.1. In the rest of the chapter, we use r and l in superscript or subscript of notations to denote the right cell C_r and the left cell C_l respectively, and use u and d in superscript or subscript of notations to denote uplink and downlink respectively. B_u^r (B_u^l) and B_d^r (B_d^l) units of bandwidth are allocated to uplink and downlink of the cell C_r (C_l) respectively. The total bandwidth in C_r (C_l) is denoted by B_r (B_l), where B_r is equal to $B_u^r + B_d^r$ ($B_l = B_u^l + B_d^l$). Without loss of generality, let C_r be the current observing cell and C_l be the neighboring cell.

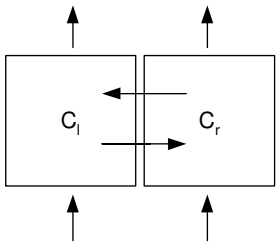

Figure 6.1: Two-cell system.

Before we present DMS-AC, we need to define the overload states of a specific call class in the multi-service system. In a mono-service system, the system is at the overload state when no more calls can be accepted. In multi-service networks, the set of overload states for different call class could be different. We use an example to illustrate this. Suppose that there are 10 downlink

channels and 5 uplink channels in a cell. Two call classes, class 1 and class 2, are supported. A class 1 call requires 1 channel on both uplink and downlink while a class 2 call requires 1 uplink channel and 3 downlink channels. A system state is denoted by (n_1, n_2), where n_1 and n_2 represent the number of class 1 calls and class 2 calls in the system, respectively. In Figure 6.2, we show all feasible states with dots. From the figure, we find that when the system is at states $(0, 3)$ and $(2, 2)$, no class 2 calls can be accepted while class 1 calls are still admissible. Thus these two states, $(0, 3)$ and $(2, 2)$, are the overload states of call class 2 (but not of call class 1). The solid dots in Figure 6.2 (a) and (b) are used to indicate the overload states of call class 1 and 2, respectively. From this example, we know that the set of overload states of call class 1 (Figure 6.2 (a)) are different from that of call class 2 (Figure 6.2 (b)). Generally, for the multi-service networks, the sets of overload states of different call classes may be different. We use φ_i to denote the probability that the system is at any one of the overload state of a specific call class i, which can also be regarded as the call dropping probability of call class i.

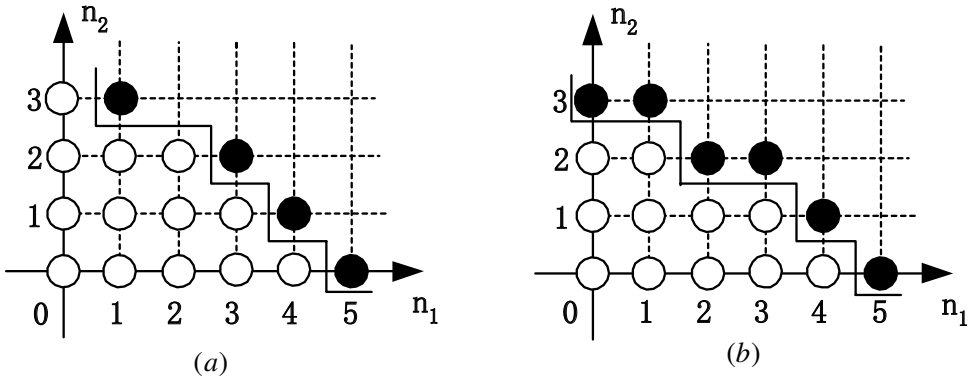

Figure 6.2: An example: (a) Overload states for call class 1; (b) Overload states for call class 2.

During a control period, the admission of a class i ($i \in [1, M]$) new call in the observing cell C_r should satisfy the following two admission conditions:

1) The admission of a new class i ($i \in [1, M]$) call in C_r cannot cause the call dropping probability of call class j in C_r, denoted by φ_j^r, to exceed η_j ($\forall j \in [1, M]$).

2) The admission of a new class i ($i \in [1, M]$) call in C_r cannot cause the call dropping probability of call class j in the neighboring cell C_l, denoted by φ_j^l, to exceed η_j ($\forall j \in [1, M]$).

The second condition is used to limit the number of potential handoff calls from C_r to C_l in order

to avoid superfluous handoff calls of a specific call class overusing the resources in C_l.

The key of DMS-AC is to determine the threshold of every call class in each cell. To this end, we need to compute φ_j^r and φ_j^l ($\forall j \in [1, M]$). Assume that there are r_i and l_i class i calls in C_r and C_l at the beginning of a control period, respectively. Our objective is to find the maximum value of r_i, denoted by Th_i^r, as the threshold of the new calls of class i during the current control period. We assume that a class i call in C_r remains in the same cell during the control period with probability $P_{i,r}^s$, and moves to C_l with probability $P_{i,rl}^m$. Accordingly, $P_{i,l}^s$ denotes the probability that a class i call remains in C_l and $P_{i,lr}^m$ denotes the probability of a class i call moving to C_r during the control period. Let λ_i^r (λ_i^l) and ν_i^r (ν_i^l) denote the mean new call and handoff call arrival rates of call class i in C_r (C_l) during the control period, respectively.

Let us consider the first admission condition. During a control period, the probability that x_i class i calls out of r_i calls stay in C_r has a binomial distribution given by

$$B(x_i, r_i, P_{i,r}^s) = \begin{pmatrix} x_i \\ r_i \end{pmatrix} (P_{i,r}^s)^{x_i} (1 - P_{i,r}^s)^{r_i - x_i} . \tag{6.1}$$

Similarly, the probability that y_i class i calls handoff to C_r from C_l during a control period is

$$B(y_i, \theta_i^l, P_{i,lr}^m) = \begin{pmatrix} y_i \\ \theta_i^l \end{pmatrix} (P_{i,lr}^m)^{y_i} (1 - P_{i,lr}^m)^{\theta_i^l - y_i} , \tag{6.2}$$

where θ_i^l is expressed as $l_i + (\hat{\nu}_i^l + \hat{\lambda}_i^l)T$. Since the new arrivals in C_l during the current control period, which include both new and handoff calls, may also handoff to C_r, we use θ_i^l instead of l_i in (6.2). The number of handoff and new calls that will be admitted during the control period is represented as $(\hat{\nu}_i^l + \hat{\lambda}_i^l)T$, where $\hat{\nu}_i^l$ and $\hat{\lambda}_i^l$ are equal to $(1 - \varphi_i^l)\nu_i^l$ and $(1 - \phi_i^l)\lambda_i^l$, respectively. We use φ_i^l and ϕ_i^l to denote the handoff call dropping probability and new call blocking probability of call class i in C_l during the current control period, respectively. The controller of C_l may compute φ_i^l and ϕ_i^l given certain bandwidth allocation and threshold values. The admission controller of C_r can obtain the values of φ_i^l and ϕ_i^l by exchanging information with C_l. Let $P_r(n_i)$ denote the probability that there are n_i class i calls in C_r during T units of time, where $n_i = x_i + y_i$. Thus $P_r(n_i)$ is the convolution sum of two binomial distributions $B(x_i, r_i, P_{i,r}^s)$ and $B(y_i, \theta_i^l, P_{i,lr}^m)$, where

x_i and y_i should satisfy $0 \leq x_i \leq r_i$ and $0 \leq y_i \leq \theta_i^l$, respectively. Since CAC is always used in a heavy traffic load system, we could approximate the binomial distribution $B(i, n, p)$ by a Gaussian distribution $G(m, \sigma)$ with mean $m = np$ and variance $\sigma = \sqrt{np(1-p)}$ [100]. Thus the number of class i calls in C_r during the control period also has a Gaussian distribution given by

$$\begin{aligned} P_r(n_i) &= B(x_i, r_i, P_{i,r}^s) \otimes B(y_i, \theta_i^l, P_{i,lr}^m) \\ &\simeq G\left(r_i P_{i,r}^s + \theta_i^l P_{i,lr}^m, \sqrt{r_i P_{i,r}^s (1 - P_{i,r}^s) + \theta_i^l P_{i,lr}^m (1 - P_{i,lr}^m)}\right) \end{aligned} \tag{6.3}$$

We know that a system stays at a feasible state at any time, which means that a state s, $s = (n_1, n_2, \ldots, n_M)$, should satisfy $\sum_{i=1}^{M} n_i b_i^u \leq B_u$ and $\sum_{i=1}^{M} n_i b_i^d \leq B_d$. n_i ($i \in [1, M]$) is the number of class i calls in the system. b_i^u and b_i^d are the bandwidth required by a class i call on uplink and downlink, respectively. B_u and B_d are the bandwidth allocated to uplink and downlink, respectively. Let $\mathbf{S}$ denote the set of all feasible states of a cell. Since the resources in the system are limited, the number of feasible states of the system is also limited. Let there be total q feasible states and $\mathbf{S}$ can be expressed as

$$\mathbf{S} = \begin{bmatrix} s_1 \\ \vdots \\ s_k \\ \vdots \\ s_q \end{bmatrix} = \begin{bmatrix} n_1^1, n_1^1, \cdots, n_M^1 \\ \vdots \\ n_1^k, n_1^k, \cdots, n_M^k \\ \vdots \\ n_1^q, n_1^q, \cdots, n_M^q \end{bmatrix}, \tag{6.4}$$

where s_k, $s_k = (n_1^k, n_1^k, \ldots, n_M^k)$, is the k_{th} state of $\mathbf{S}$ and n_i^k is the number of class i calls in the system when system state is s_k, which satisfies $\sum_{i=1}^{M} n_i^k b_i^u \leq B_u$ and $\sum_{i=1}^{M} n_i^k b_i^d \leq B_d$.

We define $\mathbf{S_{i,j}}$ ($\mathbf{S_{i,j}} \subseteq \mathbf{S}$) to be the set of states of call class j. When a system is at a state $s_k (s_k \in \mathbf{S_{i,j}})$, it can reach the overload states of call class j with the increase of the number of class i calls in the system. We continue the example used previously to explain the meaning of $\mathbf{S_{i,j}}$. From Figure 6.2, we find that when $n_1 = 0$ or $n_1 = 2$ the system cannot reach the overload states of call class 1 by only increasing the number of class 2 calls. When the system is at state $(0, 3)$ or $(2, 2)$, class 1 call still can be accepted although no class 2 calls can be accepted. On the other hand, when $n_1 = 1, 3, 4$, or 5 the system can reach the overload states of call class 1 by increasing n_2. Thus $\mathbf{S_{2,1}}$ is represented as shown in Figure 6.3. When the system is at a state in $\mathbf{S_{2,1}}$, the

system can reach the overload states of call class 1 by increasing the number of class 2 calls only. We also define $N_{i,j}(s_k)$ ($s_k \in \mathbf{S_{i,j}}$) to be the minimum number of class i calls that let the system enter the overload states of call class j when system is at state s_k. For example, we know that $N_{1,2}(0,2) = 2$ from Figure 6.3.

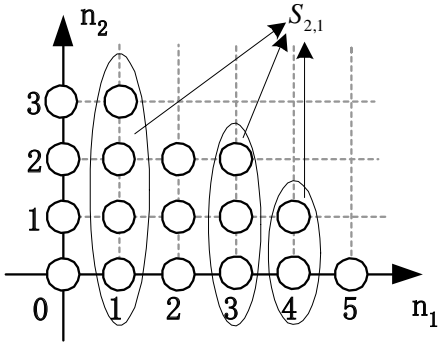

Figure 6.3: Illustration of $\mathbf{S_{i,j}}$.

Let $\varphi^r_{i,j}$ denote the probability that the cell C_r is at one of the overload states of call class j, which results due to the admission of class i calls. For the first admission condition, the handoff call dropping probability of class j calls in C_r can be expressed as

$$\varphi^r_j = \sum_{i=1}^{M} \varphi^r_{i,j} \tag{6.5}$$

and

$$\varphi^r_{i,j} = \sum_{\forall s^r_k \in \mathbf{S^r_{i,j}}} P(s^r = s^r_k) \sum_{n_i \geq N_{i,j}(s^r_k)} \{P_r(n_i) | (s^r = s^r_k)\} \,, \tag{6.6}$$

where s^r is a random variable which denotes the system state of C_r and s^r_k is a specific system state.

Next we consider the second admission condition, which states that the call dropping probability of class j in C_l incurred by the handoff class i calls from C_r must be smaller than or equal to η_j. Assume that there are r_i and l_i class i calls in C_r and C_l respectively at the beginning of a control period. Similar to that in the discussion of the first admission condition, the probability that x_i class i calls out of r_i calls handoff from C_r to C_l during the control period has a binomial distribution given by $B(x_i, r_i, P^m_{i,rl})$ and the probability that y_i class i calls out of $\theta_{i,l}$ stay in C_l during the control period is $B(y_i, \theta_{i,l}, P^s_{i,l})$. Thus the probability distribution of having n_i class i

calls in C_l during the control period, denoted by $P_l(n_i)$, is given by the convolution sum of two binomial distributions $B(x_i, r_i, P^m_{i,rl})$ and $B(y_i, \theta_{i,l}, P^s_{i,l})$ ($0 \le x_i \le r_i$, $0 \le y_i \le \theta_{i,l}$ and $x_i + y_i = n_i$). We approximate the binomial distribution by a Gaussian distribution with appropriate mean and variance. As a result, $P_l(n_i)$ is given as (6.7).

$$\begin{aligned} P_l(n_i) &= B(x_i, r_i, P^m_{i,rl}) \otimes B(y_i, \theta_{i,l}, P^s_{i,l}) \\ &\simeq G\left(r_i P^m_{i,rl} + \theta_{i,l} P^s_{i,l}, \sqrt{r_i P^m_{i,rl}(1 - P^m_{i,rl}) + \theta_{i,l} P^s_{i,l}(1 - P^s_{i,l})}\right) \end{aligned} \tag{6.7}$$

For the second admission condition, the handoff call dropping probability of call class j in C_l is expressed as

$$\varphi^l_j = \sum_{i=1}^{M} \varphi^l_{i,j} \tag{6.8}$$

and

$$\varphi^l_{i,j} = \sum_{\forall s^l_k \in \mathbf{S^l_{i,j}}} P(s^l = s^l_k) \sum_{n_i \ge N_{i,j}(s^l_k)} \{P_l(n_i) | (s^l = s^l_k)\} \ , \tag{6.9}$$

where s^l is a random variable which denotes the system state of C_l and s^l_k is a specific system state.

6.2.2 Derivation of Admission Thresholds

In the following, we derive the thresholds for the proposed DMS-AC. Let us consider the first admission condition, where φ^r_j is required to be smaller than or equal to η_j ($j \in [1, M]$). From (6.5), we know that φ^r_j can be expressed as the summation of $\varphi^r_{i,j}$ ($\forall i \in [1, M]$), which results due to the arrival of class i calls. Thus we can require

$$\varphi^r_{i,j} \le \frac{\lambda^r_i + \nu^r_i}{\sum_{k=1}^{M} (\lambda^r_k + \nu^r_k)} \cdot \eta_j \tag{6.10}$$

($\forall i, j \in [1, M]$). $\varphi^r_{i,j}$ can be expressed as that shown in (6.6). However, it is difficult to compute threshold of each call class from (6.6) directly. We use an indirect method to compute the thresholds. From (6.6) and (6.9), we know that the overload probabilities of the specific feasible states s^r_k and s^l_k can be expressed as

$$\varphi^r_{i,j}(s^r_k) = \sum_{n_i \ge N_{i,j}(s^r_k)} \{P_r(n_i) | (s^r = s^r_k)\} \simeq Q\left(\frac{N_{i,j}(s^r_k) - (r_i P^s_{i,r} + l_i P^m_{i,lr})}{\sqrt{r_i P^s_{i,r}(1 - P^s_{i,r}) + l_i P^m_{i,lr}(1 - P^m_{i,lr})}}\right) \tag{6.11}$$

and

$$\varphi_{i,j}^{l}(s_k^l) = \sum_{n_i \geq N_{i,j}(s_k^l)} \{P_l(n_i) | (s^l = s_k^l)\} \simeq Q(\frac{N_{i,j}(s_k^l) - (r_i P_{i,rl}^m + l_i P_{i,l}^s)}{\sqrt{r_i P_{i,rl}^m (1 - P_{i,rl}^m) + l_i P_{i,l}^s (1 - P_{i,l}^s)}}), \tag{6.12}$$

respectively. $Q(\cdot)$ is the integral over the tail of a Gaussian distribution which can be expressed in terms of the error function [22, 100]. We consider a conservative way to compute threshold by requiring the overload probability of the specific state ($\varphi_{i,j}^r(s_k^r)$ or $\varphi_{i,j}^r(s_k^r)$) to be smaller than certain constraint and thus obtain the lower bounds of thresholds. Then the thresholds will be tuned according to the call blocking probability as illustrated in Section 6.2.4.

When we require $\varphi_{i,j}^r(s_k^r) \leq \frac{\lambda_i^r + \nu_i^r}{\sum_{k=1}^M (\lambda_k^r + \nu_k^r)} \eta_j$, we can find a value, say $a_{i,j}^r$, to satisfy

$$\frac{\lambda_i^r + \nu_i^r}{\sum_{k=1}^M (\lambda_k^r + \nu_k^r)} \eta_j = Q(a_{i,j}^r) \; . \tag{6.13}$$

Thus we have

$$N_{i,j}(s_k^r) - (r_i P_{i,r}^s + l_i P_{i,lr}^m) - a_{i,j}^r \sqrt{r_i P_{i,r}^s (1 - P_{i,r}^s) + l_i P_{i,lr}^m (1 - P_{i,lr}^m)} = 0 \; . \tag{6.14}$$

By manipulating (6.14), we can obtain a value of r_i, which is regarded as the threshold of class i calls that satisfies (6.13) when the system is at a specific state s_k^r. We use $Th_{i,j}^1(s_k^r)$ to represent this value as

$$Th_{i,j}^1(s_k^r) = \begin{array}{l} \frac{1}{2P_{i,r}^s} \Big(2N_{i,j}(s_k^r) - 2\theta_{i,l} P_{i,lr}^m + (a_{i,j}^r)^2 (1 - P_{i,r}^s) \\ -a_{i,j}^r \sqrt{4N_{i,j}(s_k^r) \cdot (1 - P_{i,r}^s) + (a_{i,j}^r)^2 (1 - P_{i,r}^s)^2 + 4\theta_{i,l} P_{i,lr}^m (P_{i,r}^s - P_{i,lr}^m)} \Big) \end{array} \; . \tag{6.15}$$

Next let us consider (6.12). The second admission condition requires $\varphi_j^l \leq \eta_j$. We can obtain

$$\varphi_{i,j}^l \leq \frac{\lambda_i^l + \nu_i^l}{\sum_{k=1}^M (\lambda_k^l + \nu_k^l)} \cdot \eta_j \tag{6.16}$$

for all $i, j \in [1, M]$. If $\frac{\lambda_i^l + \nu_i^l}{\sum_{k=1}^M (\lambda_k^l + \nu_k^l)} \cdot \eta_j$ is required to be equal to $Q(a_{i,j}^l)$, we have

$$N_{i,j}(s_k^l) - (r_i P_{i,rl}^m + \theta_{i,l} P_{i,l}^s) - a_{i,j}^l \sqrt{r_i P_{i,rl}^m (1 - P_{i,rl}^m) + \theta_{i,l} P_{i,l}^s (1 - P_{i,l}^s)} = 0 \; . \tag{6.17}$$

From (6.17), we can obtain the threshold $Th^2_{i,j}(s^l_k)$ as

$$Th^2_{i,j}(s^l_k) = \begin{array}{l} \frac{1}{2P^m_{i,rl}} \Big(2N_{i,j}(s^l_k) - 2\theta_{i,l}P^s_{i,l} + (a^l_{i,j})^2(1-P^m_{i,rl}) \\ -a^l_{i,j}\sqrt{4N_{i,j}(s^l_k)\cdot(1-P^m_{i,rl}) + (a^l_{i,j})^2(1-P^m_{i,rl})^2 + 4\theta_{i,l}P^s_{i,l}(P^m_{i,rl}-P^s_{i,l})}\Big) \end{array} , \tag{6.18}$$

which satisfies (6.16).

From (6.15) and (6.18), we can obtain a series values of $Th^1_{i,j}(s^r_k)$ and $Th^2_{i,j}(s^l_k)$ for specific states s^r_k and s^l_k in C_r and C_l to satisfy (6.10) and (6.16), respectively. The admission thresholds of class i calls in C_r to satisfy the two admission conditions are given by:

$$Th^1_{i,j} = \sum_{\forall s^r_k \in \mathbf{S^r_{i,j}}} P(s^r_k)\cdot Th^1_{i,j}(s^r_k) \tag{6.19}$$

and

$$Th^2_{i,j} = \sum_{\forall s^l_k \in \mathbf{S^l_{i,j}}} P(s^l_k)\cdot Th^2_{i,j}(s^l_k)\ . \tag{6.20}$$

From (6.19) and (6.20), we can obtain a series values of the threshold of call class i to satisfy different QoS requirements of all call classes in the system. Let Th^1_i and Th^2_i denote the thresholds of call class i calls that satisfies the first and the second admission conditions, respectively. Thus, Th^1_i and Th^2_i can be expressed as

$$Th^1_i = \min_{\forall j\in[1,M]}(Th^1_{i,j}) \tag{6.21}$$

and

$$Th^2_i = \min_{\forall j\in[1,M]}\left(Th^2_{i,j}\right)\ . \tag{6.22}$$

The final admission threshold of call class i in C_r which satisfies all admission conditions is given by $Th^r_i = \min(Th^1_i, Th^2_i)$.

6.2.3 Extension to a Multi-cell System

In this sub-section, we extend the above distributed multi-service admission control policy for the two-cell system to a multi-cell system. We consider a system with seven hexagonal cells (denoted as $C_0, C_1, \cdots$ and C_7) as shown in Figure 6.4. Without loss of generality, let C_0 be the current

observing cell and C_1 to C_6 be the neighboring cells. During a control period, the admission of a class i ($i \in [1, M]$) call in C_0 should satisfy:

1) The admission of a new class i ($i \in [1, M]$) call in C_0 cannot cause the call dropping probability of call class j in C_0, denoted by φ_j^0, to exceed η_j ($\forall j \in [1, M]$).

2) The admission of a new class i ($i \in [1, M]$) call in C_0 cannot cause the call dropping probability of call class j in the neighboring cells to exceed η_j ($\forall j \in [1, M]$).

The procedure for computing the threshold of call class i is similar to that in the two-cell system.

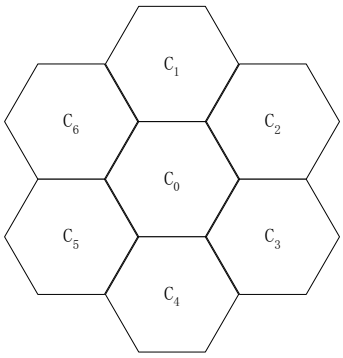

Figure 6.4: Seven-cell system.

Let us consider the first admission condition. At the beginning of a control period, there are $w_{i,h}$ class i calls in cell C_h, where $h = 0, \ldots, 6$. Let $\theta_{i,h}$ ($h = 1, \ldots, 6$) denote the number of calls in cell C_h ($h = 1, \ldots, 6$) during the control period and it is defined as $\theta_{i,h} = \min(w_{i,h} + \lambda_i^h T, Th_i^h)$, where λ_i^h is the average arrival rate of class i calls in C_h during the control period and Th_i^h is the admission threshold of class i calls in C_h where $h = 1, \ldots, 6$. The probability that a class i call stays in C_0 during the control period is denoted by $P_{i,0}^s$ and the probability that a class i call hands-off to C_0 from $C_1, \ldots, C_6$ is represented by $P_{i,h0}^m$ ($h = 1, \ldots, 6$). Thus the probability distribution of the number of class i calls in the cell C_0 during the control period is given by a convolution sum of seven binomial distributions $B(x_i, w_{i,0}, P_{i,0}^s)$ and $B(y_{i,h}, \theta_{i,h}, P_{i,h0}^m)$ where $h = (1, \ldots, 6)$. We approximate the binomial distributions by Gaussian distributions with appropriate means and variances. The probability that there are $n_i = x_i + \sum_{h=1}^{6} y_{j,h}$ class i calls in C_0 during the control period is represented as

$$P_{C_0}(n_i) \simeq G\left(n_{i,0}P_{i,0}^s + \sum_{h=1}^{6}\theta_{i,h}P_{i,h0}^m, \sqrt{n_{i,0}P_{i,0}^s(1-P_{i,0}^s) + \sum_{h=1}^{6}\theta_{i,h}P_{i,h0}^m(1-P_{i,h0}^m)}\right) . \quad (6.23)$$

Then, the overload probability of class j calls in C_0 is

$$\varphi_j^0 = \sum_{i=1}^{M} \left(\sum_{\forall s_k^0 \in \mathbf{S_{i,j}^0}} P(s^0 = s_k^0) \sum_{n_i \geq N_{i,j}(s_k^0)} \{P_{C_0}(n_i) | (s^0 = s_k^0)\} \right). \tag{6.24}$$

where $\mathbf{S_{i,j}^0}$ is the set of states for class j in C_0, such that for the system at a state $s_k \in \mathbf{S_{i,j}^0}$, it can reach the overload states for class j with the increase of the number of class i calls in C_0. s^0 is a random variable representing the system state while s_k^0 is the k_{th} state in C_0. $N_{i,j}(s_k^0)$ represents the minimum number of class i calls that let the system enter the overload state of call class j when C_0 is at s_k^0. By applying similar method used in (6.14), (6.15) and (6.19), we could find the threshold Th_{i,C_0}^0 of call class i ($i \in [1, M]$) in C_0 that satisfies the first admission condition.

Then we consider the second admission condition. Let Th_{i,C_h}^0 be the call admission threshold of call class i in C_0 to satisfy the second admission condition of cell C_h, where $h = (1, \ldots, 6)$. We show how to compute Th_{i,C_1}^0 as an example and $Th_{i,C_2}^0, \ldots, Th_{i,C_6}^0$ can be calculated in the similar manner. From Figure 6.4, we know that the neighboring cells of C_1 are C_0, C_2 and C_6. Following the similar way used in the two-cell system, let $P_{C_1}(n_i)$ denote the probability distribution of the number of class i calls in C_1 during the control period and it is given by

$$P_{C_1}(n_i) = B(y_i, \theta_{i,1}, P_{i,1}^s) \otimes B(x_i, w_{i,0}, P_{i,01}^m) \otimes B(y_i, \theta_{i,2}, P_{i,21}^m) \otimes B(y_i, \theta_{i,6}, P_{i,61}^m) \tag{6.25}$$

b $B(y_i, \theta_{i,1}, P_{i,1}^s)$ denotes the probability that y_i class i calls out of $\theta_{i,1}$ class i calls remain in C_1 during the control period, where $P_{i,1}^s$ represents the probability that a class i call stays in C_1 during the control period. $B(x_i, w_{i,0}, P_{i,01}^m)$ is the probability that x_i calls out of $w_{i,0}$ class i calls move from C_0 to C_1 during the control period and $P_{i,01}^m$ is the probability that a class i call hands-off from C_0 to C_1 during the control period. $B(y_i, \theta_{i,1}, P_{i,h1}^m)$ is the probability that y_i class i calls out of $\theta_{i,h}$ class i calls handoff from C_h to C_1, where $P_{i,h1}^m$ is the probability of a class i call hands-off from C_h to C_1 during the control period where $h = 2$ or $h = 6$. Thus we could obtain $P_{C_1}(n_i)$ and the overload probability by approximating the binomial distributions with Gaussian distributions with appropriate means and variances. By applying similar technical method as that used in (6.17), (6.20) and (6.18), we could obtain Th_{i,C_1}^0 for class i calls in C_0 to satisfy certain QoS limitations in C_1. Similarly, we could obtain Th_{i,C_h}^0 where $h = 2, \ldots, 6$. Finally, the threshold for class i calls in C_0 is $Th_i^0 = \min_{h=0}^{6}(Th_{i,C_h}^0)$.

6.2.4 Threshold-based Admission Control Policy

So far, we have described how to compute the thresholds of different call classes. In the above threshold computing process, $\varphi^r_{i,j}(s^r_k)$ and $\varphi^l_{i,j}(s^l_k)$ are set to be smaller than the criteria for each specific state in C_r and C_l. Indeed, this method is too conservative since it is not necessary to require $\varphi^r_{i,j}(s^r_k)$ and $\varphi^l_{i,j}(s^l_k)$ to be smaller than the criteria for every possible state, which may cause some new calls to be blocked unnecessarily. In the proposed admission control scheme, we use the computed thresholds as the lower bound and carefully tune the threshold values by increasing them until one or more of the following conditions are violated: (we consider C_r as an example and the similar conditions can be applied to C_l)

1. $\hat{\varphi}^r_i \leq \eta_i$ and $\hat{\varphi}^l_i \leq \eta_i$ $(\forall i \in [1, M])$;
2. $\hat{\phi}^r_i > \rho_i$;
3. $\hat{\phi}^r_j \leq \rho_j$ $(\forall j \in [1, i-1])$;
4. $\hat{\phi}^r_i \leq \rho_i$ $(\forall i \in [1, M])$;
5. $Th^r_i \leq \Delta^r_i$,

where $\hat{\varphi}_i$ and $\hat{\phi}_i$ are the computed blocking probabilities of handoff and new class i call, respectively. Δ^r_i is the maximum number of class i calls that can be admitted in C_r under a given bandwidth allocation. Without loss of generality, we sort call classes according to the values of upper bounds of the new call blocking probabilities of different call classes in ascending order from 1 to M. In other words, the class 1 has the lowest upper bound of new call blocking probability.

The pseudo code of the tuning process of the thresholds is shown in Figure 6.5. After we obtain Th^r_i $(\forall i \in [1, M])$, we need to check whether or not condition 1 can be satisfied for all call classes in the cell (C_r) and its neighboring cell (C_l). If it is satisfied, Th^r_i increases continually until the blocking probability of class i call is smaller than the predefined upper bound or Th^r_i reaches the maximum value. If Th^r_i does not reach the maximum value, Th^r_i still can be increased repeatedly until conditions 1 and 4 cannot be satisfied. By employing threshold tuning process, we can avoid sacrificing other call classes too much (e.g. let the new call blocking probability exceed the upper bound) and at the same time decrease the new call blocking probability. After obtaining Th_i, whether a new class i call can be accepted is determined by the number of the class i calls in the system. If the number reaches the threshold, no more new class i calls can be accepted.

```
for(i=1; i<=M; i++)
{
   if(i==1)
   {
     while(1 and 2 and 5)
     {
        Th_i^r++
     }
   }
   else
   {
     while(1 and 2 and 3 and 5)
     {
        Th_i^r++
     }
   }
}
for(i=1; i<=M; i++)
{
   while(1 and 4 and 5)
   {
      Th_i^r++
   }
}
```

Figure 6.5: Pseudo code of the process for tuning the thresholds.

6.3 Performance Evaluation

In this section, we evaluate the performance of the proposed DMS-AC scheme by comparing with Jeon's scheme [8], which also considers the asymmetric traffic load in multi-service mobile networks. Suppose that the simulation system is composed of C_r and C_l. We assume that there are total 100 channels in each cell and 50 channels are allocated to the uplink and the downlink in each cell. Two call classes, RT call and NRT call, are supported. We assume that RT calls has higher priority than NRT calls while the prioirty of handoff calls is higher than that of new calls. We can sort different call classes according their priority in descending order as: handoff RT call, handoff NRT call, new RT calls and new NRT calls. An RT call requires 1 channel on both uplink and downlink while an NRT call requires 1 uplink channel and 3 downlink channels. The highest tolerable handoff dropping probabilities of RT calls and NRT calls are 1% and 5%, respectively. We assume that the call arrivals follow Poison distribution. Let λ^r_{RT} (λ^l_{RT}) and λ^r_{NRT} (λ^l_{NRT}) denote the mean call arrival rate of new RT calls and new NRT calls in C_r (C_l), respectively. The service time of RT calls and NRT calls is exponentially distributed with mean 120 and 900 seconds, respectively. The probability of a new RT call moves from one cell to another is 0.4 and the handoff probability of a new NRT call is 0.2. For Jeon's scheme, we assume that the estimation of call arrival rate is accurate and the parameter Δ is assumed to be 0.03 as suggested in the literature.

In order to examine the performance of the proposed DMS-AC comprehensively, we conduct simulation experiments in five different scenarios. The changes of call arrival rates in the experiments are shown in Table 6.1. In the first two experiment scenarios, the call arrival rate of only one call class in a cell increases. In the subsequent three experiment scenarios, the call arrival rates of the same/different call classes in two cells increase simultaneously. We will examine the performance in terms of call dropping/blocking probability and system resource utilization in these experiments.

Table 6.1: Call arrival rates in experiment scenarios

Experiment scenarios	C_r		C_l	
	λ^r_{RT}	λ^r_{NRT}	λ^l_{RT}	λ^l_{NRT}
1	$0.05 \sim 0.12$	0.01	0.05	0.01
2	0.1	$0.005 \sim 0.012$	0.1	0.005
3	$0.05 \sim 0.12$	0.01	$0.05 \sim 0.12$	0.01
4	0.1	$0.005 \sim 0.012$	0.1	$0.005 \sim 0.012$
5	$0.05 \sim 0.12$	0.01	0.1	$0.005 \sim 0.012$

6.3.1 Experiment 1: λ^r_{RT} increases from 0.05 to 0.12

In the first experiment, λ^r_{RT} increases from 0.05 to 0.12 which means the increase of traffic load brought by RT calls. Figure 6.6 (a) and (b) show the RT and the NRT call dropping/blocking probability of C_r. With the increase of λ^r_{RT}, we can find that DMS-AC is able to guarantee the handoff dropping probabilities of both RT and NRT calls under the constraints 1% and 5%, respectively. Since our objective is to guarantee handoff call dropping probabilities under certain constraints but not to achieve as low as possible handoff dropping probability, DMS-AC does not try to achieve low handoff RT call dropping probability as that of Jeon's but obtains much lower new RT call blocking probability than Jeon's scheme when the QoS of handoff calls can be guaranteed. Although Jeon's scheme obtains lower new NRT call blocking probability, it cannot guarantee the dropping probability of handoff NRT call under 5% when λ^r_{RT} increases. Since more NRT calls are accepted and the NRT call consumes more downlink channels (3 channels), Jeon's scheme can achieve higher downlink bandwidth utilization and thus total bandwidth utilization as shown in Figure 6.7. However, the difference of the bandwidth utilization obtained by Jeon's scheme and DMS-AC is not significant. It is reasonable to block more low priority new NRT calls in order to

guarantee the QoS requirements of handoff calls, which has higher priority.

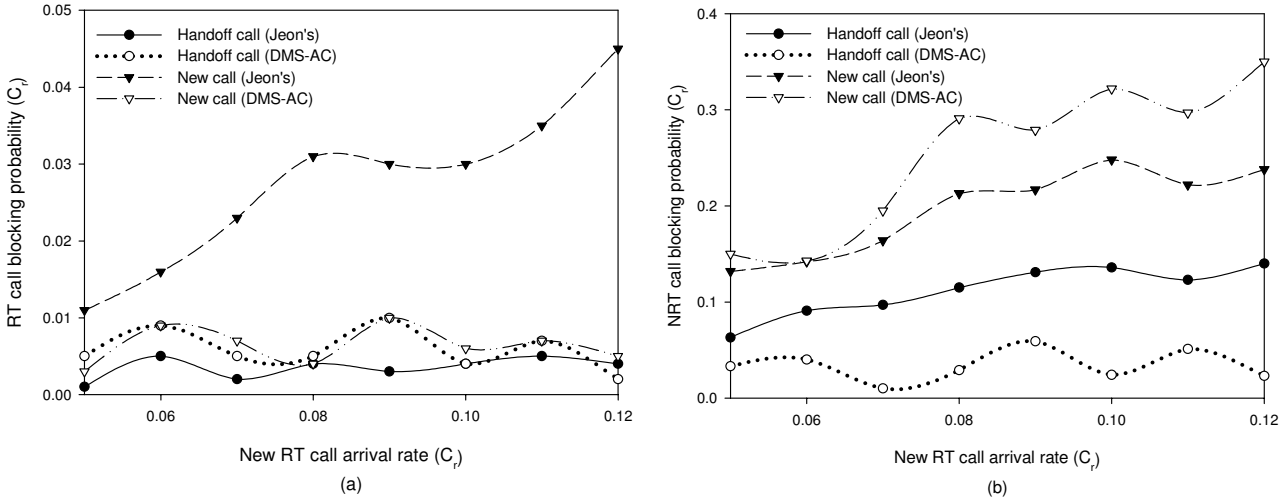

Figure 6.6: Call blocking probabilities of C_r when λ^r_{RT} increases from 0.05 to 0.12 (experiment 1).

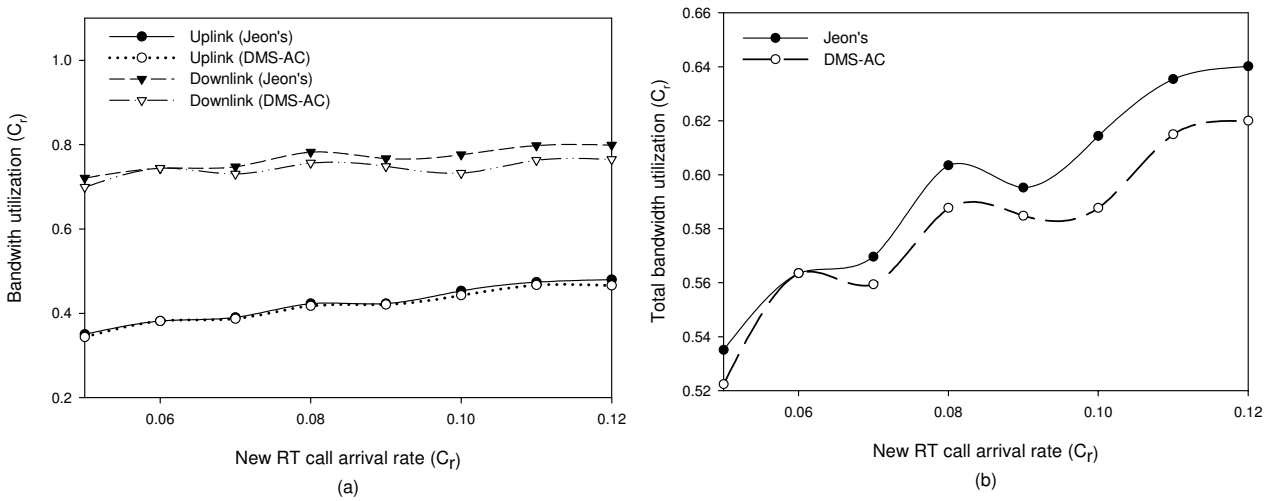

Figure 6.7: Bandwidth utilization when λ^r_{RT} increases from 0.05 to 0.12 (experiment 1).

6.3.2 Experiment 2: λ^r_{NRT} increases from 0.005 to 0.012

In the second experiment, λ^r_{NRT} increases from 0.005 to 0.012, which means that the traffic load brought by NRT calls becomes heavier. Since more new NRT calls are accepted, Jeon's scheme achieves higher downlink bandwidth utilization as shown in Figure 6.8. Although Jeon's scheme achieves low new NRT call blocking probability, it cannot guarantee the handoff NRT call dropping probability under the constraint (Figure 6.9 (b)), which is the premier QoS requirement in the

design of CAC. On the other hand, Jeon's scheme also sacrifice too many new RT calls (Figure 6.9 (a)), which have higher priority than new NRT calls. It is more reasonable to make the tradeoff between the admission of new RT and NRT calls as DMS-AC does. By limiting the admission of NRT calls, DMS-AC keeps the handoff dropping probabilities of RT calls and NRT calls under the constraints and at the same time achieves much lower new RT call blocking probability in this scenario.

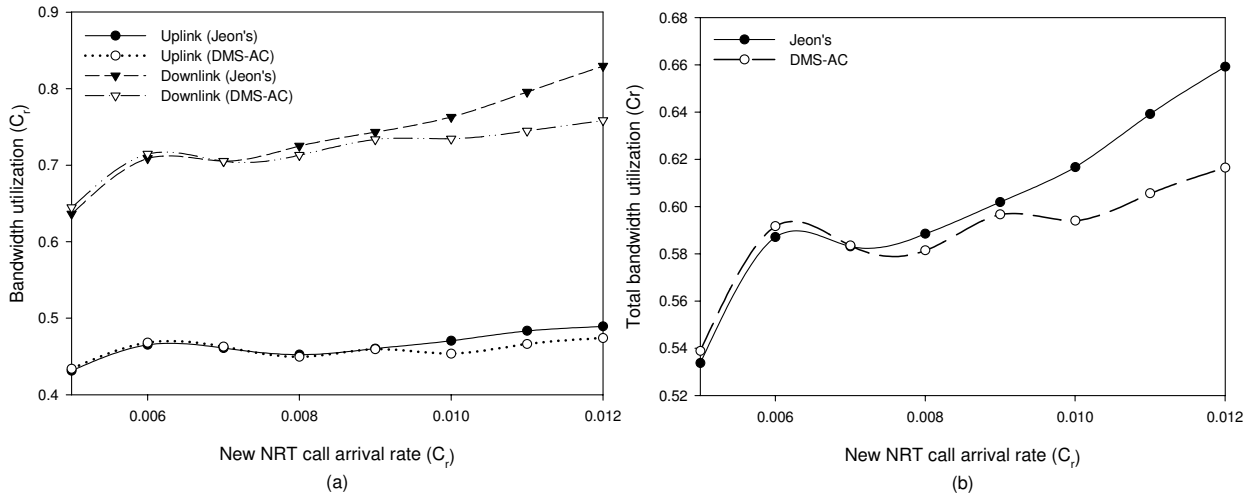

Figure 6.8: Bandwidth utilization when λ^r_{NRT} increases from 0.005 to 0.012 (experiment 2).

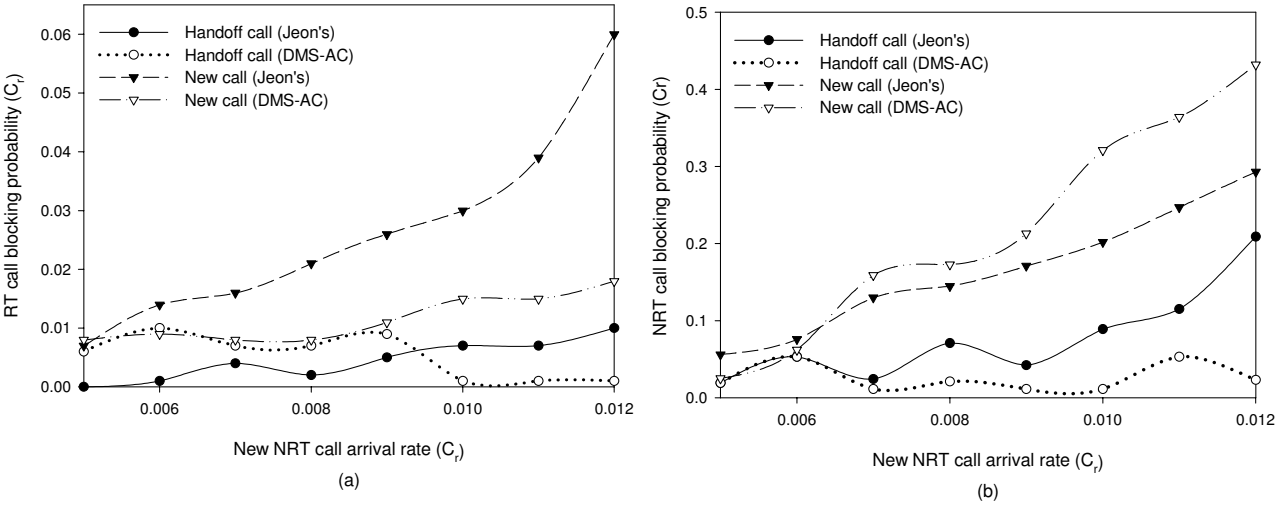

Figure 6.9: Call blocking probabilities when λ^r_{NRT} increases from 0.005 to 0.012 (experiment 2).

6.3.3 Experiment 3: λ^r_{RT} and λ^l_{RT} increase from 0.05 to 0.12 simultaneously

In the third experiment, both λ^r_{RT} and λ^l_{RT} increases from 0.05 to 0.12, which implies both new and handoff RT call arrival rate in C_r increase. Figure 6.10 shows the RT and NRT call dropping/blocking probability of C_r. From the figure, we find that DMS-AC guarantees the handoff call dropping probability of RT calls and NRT calls under the predefined constraints and also achieves lower new RT call blocking probability than Jeon's scheme by sacrificing more new NRT calls. Since more NRT calls are accepted, Jeon's scheme achieve a little bit higher downlink bandwidth utilization than DMS-AC as shown in Figure 6.11.

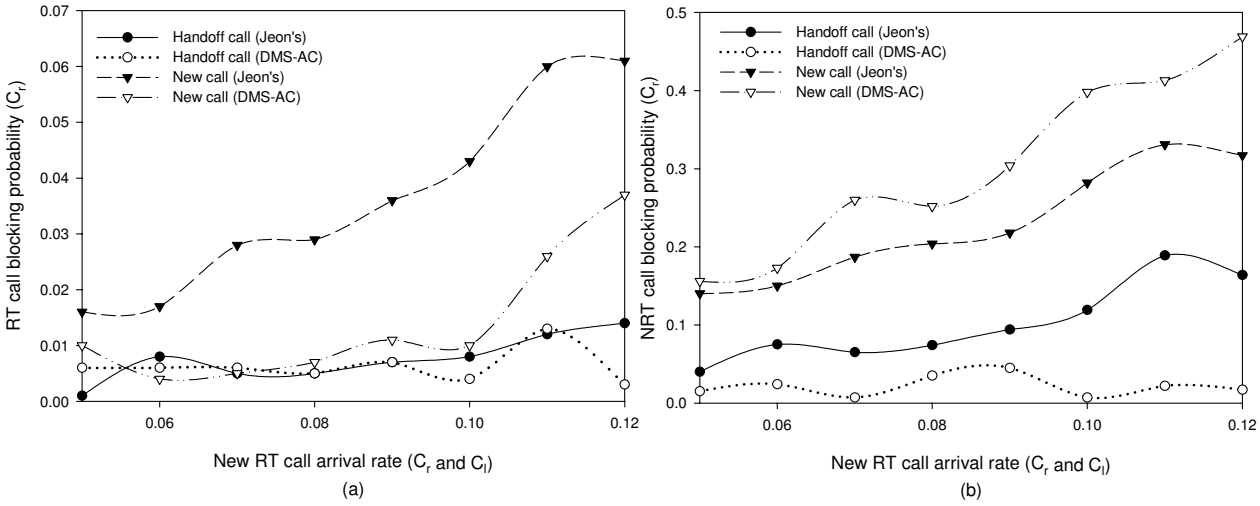

Figure 6.10: Call blocking probabilities when λ^r_{RT} and λ^l_{RT} increase from 0.05 to 0.12 simultaneously (experiment 3).

6.3.4 Experiment 4: λ^r_{NRT} and λ^l_{NRT} increase from 0.005 to 0.012 simultaneously

In this experiment, λ^r_{NRT} and λ^l_{NRT} increase from 0.005 to 0.012 simultaneously, which means that the traffic load in C_r brought by not only new NRT calls but also handoff NRT calls increases. In order to guarantee handoff RT and NRT call dropping probability under the constraints, DMS-AC blocks more new NRT calls to avoid new NRT calls overusing system resources. Figure 6.12 shows the RT and NRT call dropping/blocking probability of C_r. From the figure, we can find that DMS-AC is able to guarantee the handoff dropping probability of RT calls and NRT calls under the constraints and also achieves lower new RT call blocking probability than Jeon's scheme. Since more new NRT calls are blocked, the downlink bandwidth utilization of DMS-AC is lower than

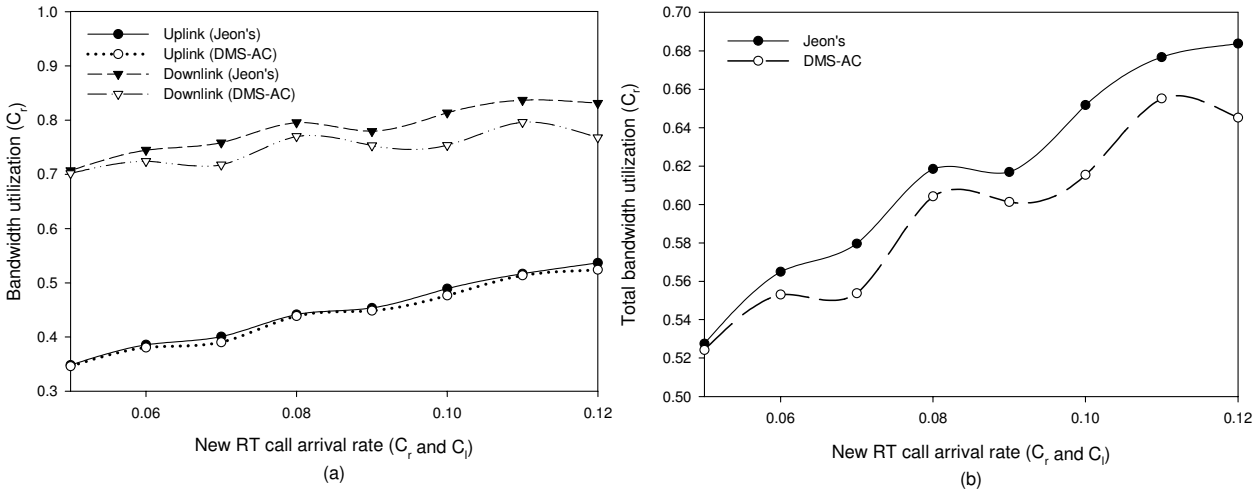

Figure 6.11: Bandwidth utilization when λ^r_{RT} and λ^l_{RT} increase from 0.05 to 0.12 simultaneously (experiment 3).

that of Jeon's scheme as shown in Figure 6.13.

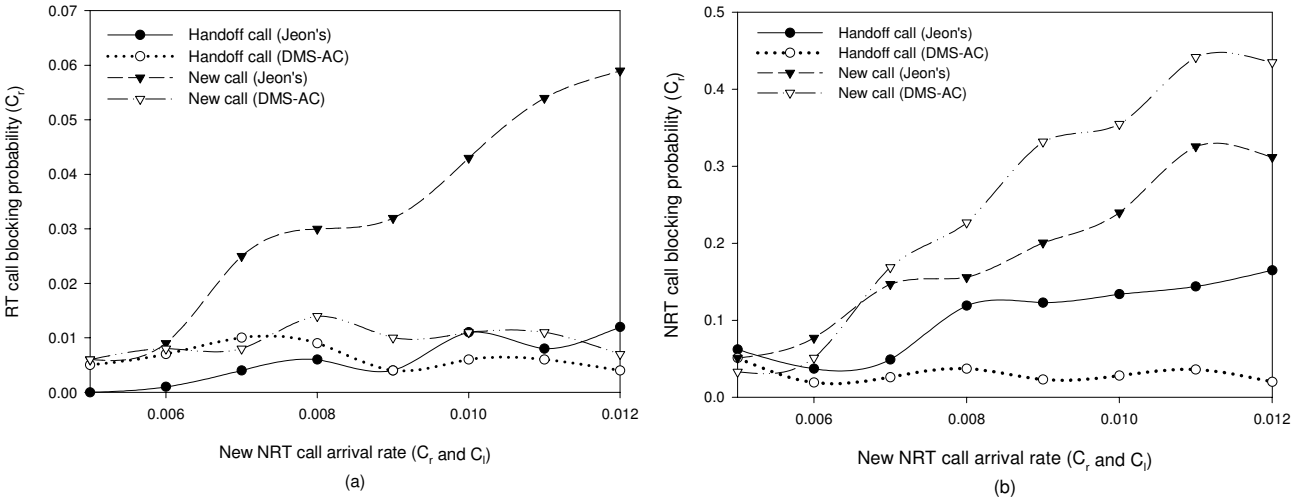

Figure 6.12: Call blocking probabilities when λ^r_{NRT} and λ^l_{NRT} increase from 0.005 to 0.012 simultaneously (experiment 4).

6.3.5 Experiment 5: λ^r_{RT} increases from 0.05 to 0.12 while λ^l_{NRT} increases from 0.005 to 0.012 simultaneously

In the last experiment, λ^r_{RT} increase from 0.05 to 0.12 and λ^l_{NRT} increase from 0.005 to 0.012 simultaneously, which means that the traffic load brought by new RT calls and handoff NRT calls increases simultaneously. Figure 6.14 and Figure 6.15 show the RT and the NRT call drop-

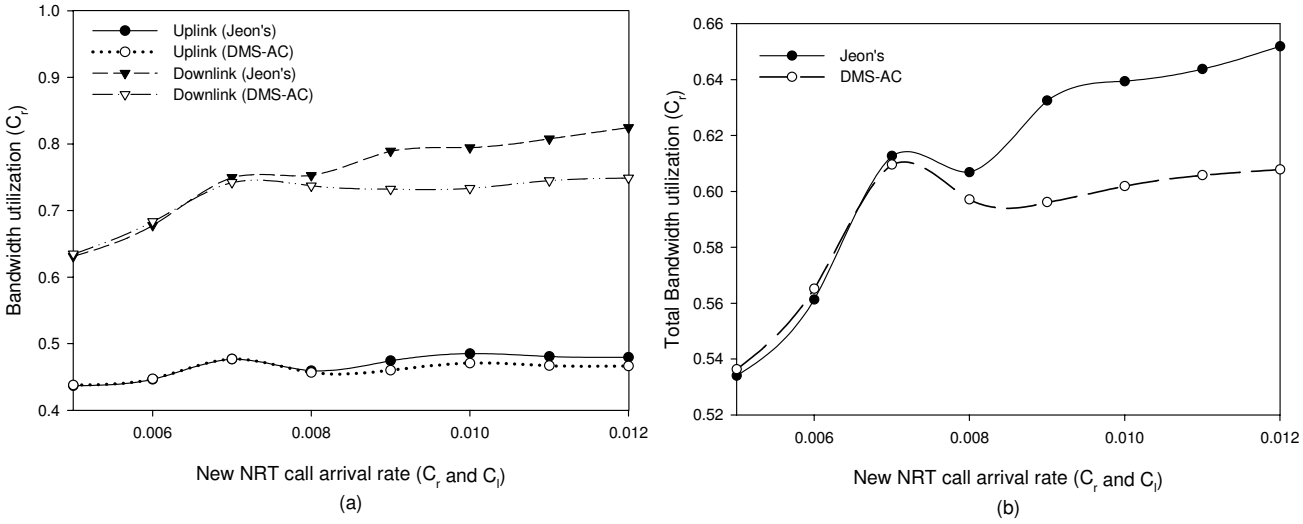

Figure 6.13: Bandwidth utilization when λ_{NRT}^{r} and λ_{NRT}^{l} increase from 0.005 to 0.012 simultaneously (experiment 4).

ping/blocking probabilities of C_r and C_l, respectively. From these figures, we find that DMS-AC is able to guarantee the handoff dropping probability of RT and NRT calls under the constraints and also achieves lower new RT call blocking probability than Jeon's scheme no matter in C_r or C_l. Although Jeon's scheme obtains lower new NRT call blocking probability than DMS-AC, it cannot guarantee the handoff NRT call dropping probability under the constraint in both C_r and C_l when the traffic load becomes heavier. Since more NRT calls are accepted, Jeon's can achieve higher downlink bandwidth utilization and thus total bandwidth utilization in two cells as shown in Figure 6.16 and 6.17, respectively.

From the above experiments, we can conclude that DMS-AC can guarantee handoff dropping probabilities of both RT and NRT calls under predefined constraints and achieve minimal new RT call blocking probability with the expense of blocking more new NRT calls. However, it is also undesirable to sacrifice too more new NRT calls. A complimentary method is necessary to improve the system performance. At the same time, the experiments' results also show that the asymmetry of bandwidth utilization of uplink and downlink is obviously and the utilization of downlink is much higher than that of uplink. In order to improve the system performance in such asymmetric traffic load environment, it is necessary to adjust the bandwidth allocation between uplink and downlink, which is the focus of next chapter.

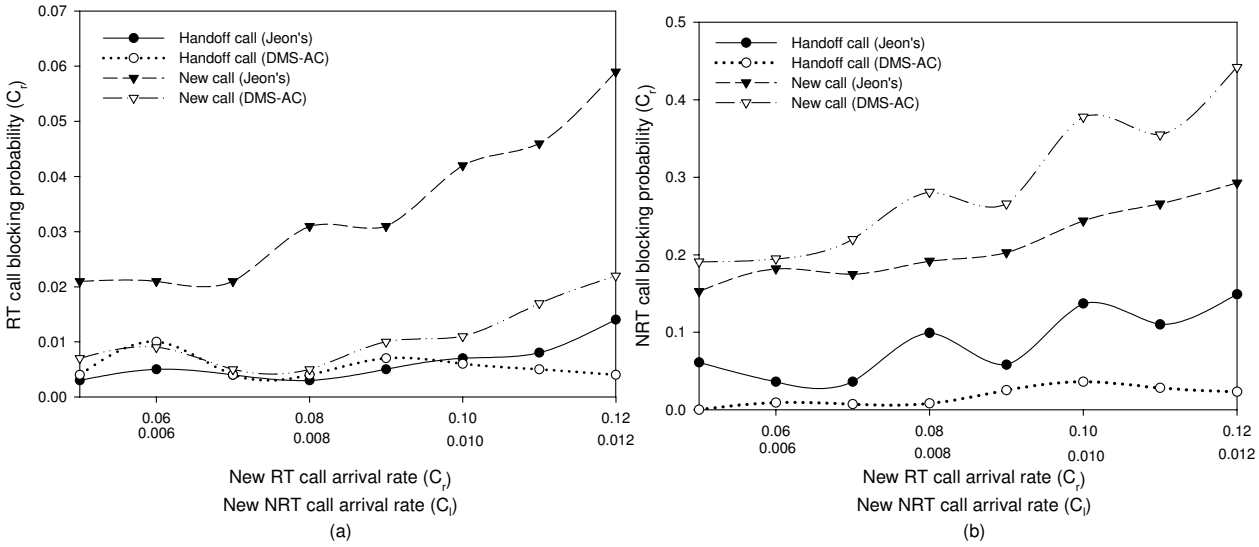

Figure 6.14: Call blocking probabilities of C_r when λ^r_{RT} increases from 0.05 to 0.12 while λ^l_{NRT} increases from 0.005 to 0.012 simultaneously (experiment 5).

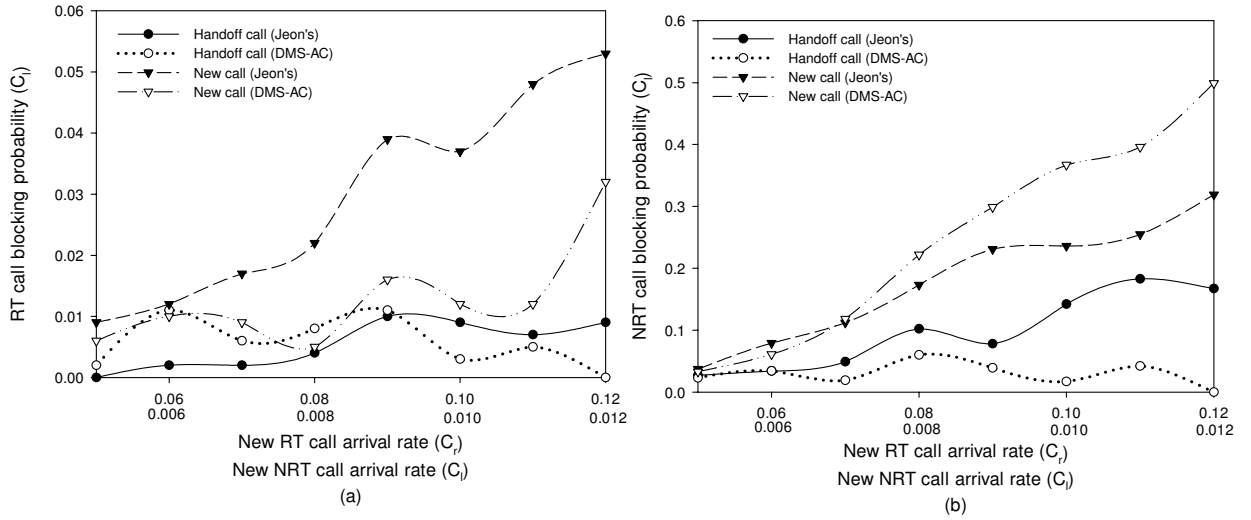

Figure 6.15: Call blocking probabilities of C_l when λ^r_{RT} increases from 0.05 to 0.12 while λ^l_{NRT} increases from 0.005 to 0.012 simultaneously (experiment 5).

6.4 Summary

Handoff call dropping probability is always a critical QoS measurement in the design of call admission control no matter in traditional mono-service networks or in future multi-service mobile celllular networks. Although LFGC scheme has been proposed and proved be optimal for the MINBlock problem in mono-service mobile networks, it is hard to extend LFGC to multi-service networks. In this chapter, we propose DMS-AC scheme to handle the MINBlock problem in multi-

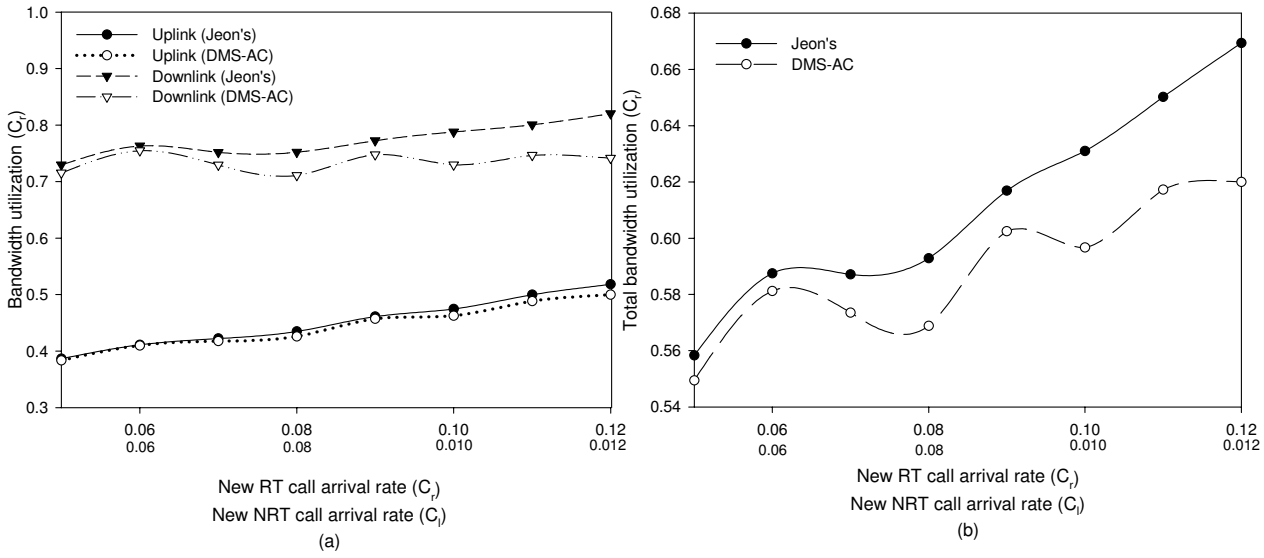

Figure 6.16: Bandwidth utilization of C_r when λ^r_{RT} increases from 0.05 to 0.12 while λ^l_{NRT} increases from 0.005 to 0.012 simultaneously (experiment 5).

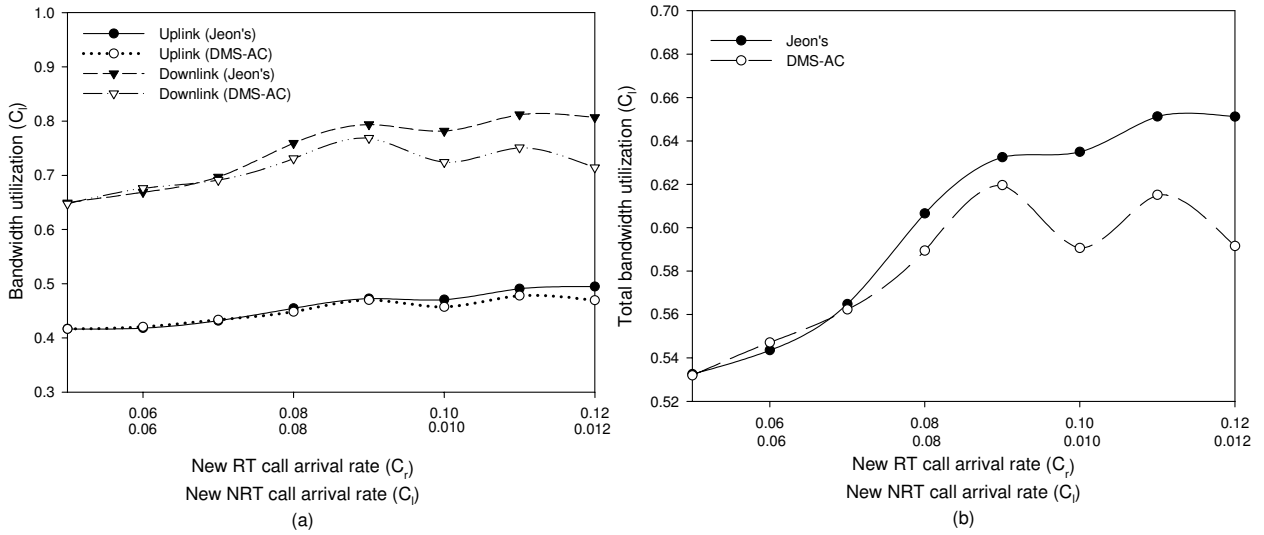

Figure 6.17: Bandwidth utilization of C_l when λ^r_{RT} increases from 0.05 to 0.12 while λ^l_{NRT} increases from 0.005 to 0.012 simultaneously (experiment 5).

service environment. By setting thresholds for different call classes, DMS-AC prevents the new calls from overusing system resources and at the same time reduce the number of potential handoff calls. In order to determine appropriate thresholds, we analyze the relationship between the admission of different call classes. We decompose all system overload states into the overload states of individual call class and study how the admission of calls from a specific class result in the system overload states of other call classes. Based on the system states of local cell and the information from

neighboring cells, DMS-AC is able to dynamically compute the thresholds for various call classes. Numerical results show that DMS-AC is able to guarantee the handoff call dropping probabilities of different call classes under certain constraints in a dynamic traffic load environment with the expense of blocking more new NRT calls, which have the lowest priority. From the experiments' results, we find that the discrepancy of the bandwidth utilization of uplink and downlink is evidence and downlink bandwidth utilization is much higher than that of uplink. It implies that the bandwidth allocation between uplink and downlink should be adjusted to satisfy the asymmetric bandwidth requirements in multi-service networks, which is just the focus of Chapter 7.

Chapter 7

Bandwidth Re-allocation for Bandwidth Asymmetry Mobile Networks

From Chapter 6, we know that in order to improve system performance, it is necessary to consider bandwidth re-allocation to collaborate with the employed CAC in multi-service mobile networks with dynamic traffic load. In this chapter, we address *when* and *how* to adjust bandwidth allocation on uplink and downlink in a multi-service mobile cellular network with bandwidth asymmetry under dynamic traffic load conditions. The design objective is to improve system bandwidth utilization while satisfying call-level QoS requirements of different call classes. The proposed Distributed Multi-service Admission Control (DMS-AC) scheme is used as the studying base for bandwidth re-allocation. When the traffic load brought by calls of some specific classes exceeds the control range of DMS-AC, the QoS of some call classes may not be guaranteed. In such situations, bandwidth re-allocation process is activated and DMS-AC will try to meet the QoS requirements under the adjusted bandwidth allocation. We explore the relationship between admission thresholds and bandwidth allocation by identifying certain constraints for verifying the feasibility of the adjusted bandwidth allocation. We conduct comprehensive simulation experiments to validate the effectiveness of the proposed bandwidth re-allocation scheme. Numerical results show that when traffic pattern with certain bandwidth asymmetry changes, the system can re-allocate the bandwidth on uplink and downlink adaptively. With the designed bandwidth re-allocation scheme in conjunction with the proposed DMS-AC scheme, the QoS requirements of different call classes can be guaranteed under dynamic traffic conditions and in the mean time the system bandwidth utilization is

improved significantly.

7.1 Introduction

With the rapid growth of multi-service mobile networks, many applications which are popular in wired networks are emerging in mobile environment. Since some data applications bring more traffic load on downlink than on uplink, next generation multi-service mobile networks are expected to present distinctive traffic asymmetry between uplink and downlink [5,7,23,89]. In such networks, in order to improve system bandwidth utilization, it is necessary to allocate different bandwidth on two links. For deterministic traffic parameters and mobility characteristics, fixed bandwidth allocation is able to provide an optimal solution for the resource allocation problem in mobile wireless networks with bandwidth asymmetry [23,24]. However, many emerging applications and services with bursty and variable bandwidth requirements call for new treatments of network resource management, in order to satisfy application needs and improve network resource utilization. Furthermore, in multi-service mobile networks, the traffic generated by some applications is time-dependent. For example, the bandwidth asymmetry caused by some data applications could be significantly higher than usual during peak hours in some particular cells. In addition, because of mobility, some users with certain applications may handoff from one cell to another causing the change of traffic load asymmetry in that cell. In such dynamic traffic load networks, there is no such an RM scheme that can satisfy the QoS requirements of different call classes all the time. From the experiments of Chapter 6, we find that more NRT calls are blocked in order to guarantee the QoS requirements of higher priority call classes and the system bandwidth utilization is also violated. Therefore, it is imperative to develop a dynamic bandwidth allocation scheme to collaborate with CAC in multi-service mobile networks with dynamic traffic conditions to provide desired QoS for different call classes and at the same time maximizing system bandwidth utilization.

In [22], the authors proved that the system with different time-slot allocations for different cells always outperforms that with the same time-slot allocation, if the time slots on uplink and downlink are properly allocated. However, there is only little known work in the literatures which addresses how to "properly" allocate bandwidth on uplink and downlink. On the other hand, since bandwidth re-allocation on uplink and downlink may affect all ongoing calls in the system [23],

we should limit the bandwidth re-allocation frequency and perform the bandwidth re-allocation when it is "necessary". Although it is suggested that the system allocate bandwidth to uplink and downlink according to traffic load [23, 25], we still do not know when a system needs to adjust the bandwidth allocation on uplink and downlink. To the best of our knowledge, there is no similar work in literatures that addresses the dynamic bandwidth allocations between uplink and downlink in bandwidth asymmetry mobile networks with changing traffic load and pattern. In this chapter, we explore *when* and *how* to adjust bandwidth allocation properly in a multi-service mobile network with bandwidth asymmetry. Our objective is to design a dynamic bandwidth allocation scheme to provide the desired QoS requirements for different call classes and in the mean time utilize the bandwidth resources in the best way.

As an indispensable component of system resource management framework, call admission control is always employed to guarantee the system QoS in terms of call blocking and dropping probability at call level. Although numerous admission control schemes have been proposed for mono/multi-service mobile networks, there is no such a CAC scheme that can guarantee the QoS of every call class under changing traffic load conditions. This motivates us to employ bandwidth re-allocation as a complementary strategy for admission control scheme to meet the QoS requirements of different call classes and maximize system resource utilization under changing traffic conditions. In Chapter 6, we proposed a Distributed Multi-service Admission Control (DMS-AC) to minimize new call blocking probabilities while maintaining handoff call dropping probabilities under certain constraints. In this chapter, we employ DMS-AC as the base for studying bandwidth re-allocation problem. By identifying certain admission conditions, DMS-AC tries to find different threshold for each call class according to the traffic pattern. If the feasible thresholds of some call classes cannot be found or the blocking probabilities of some new call classes exceed specific upper bounds, it indicates the QoS requirements of those call classes cannot be guaranteed. In such situation, the system may adjust the bandwidth allocation on uplink and downlink and re-compute the call admission thresholds until the proper thresholds are determined for each call class in the cell. By studying bandwidth re-allocation based on DMS-AC scheme, we find that the bandwidth allocated to uplink and downlink should not only be proportional to the traffic load as suggested in [23], but also satisfy certain constraints, which are obtained from the derivations of the thresholds. By

using these constraints to verify the feasibility of a bandwidth allocation, we link the admission control and the bandwidth allocation closely and provide a good solution to the problem- *when* and *how* to adjust bandwidth allocation to guarantee the QoS requirements of different call classes in a multi-service mobile network with bandwidth asymmetry between uplink and downlink.

The rest of this chapter is organized as follows. In Section 7.2, we study the bandwidth re-allocation problem based on the proposed admission control scheme. In this section, we address when and how to adjust the bandwidth allocation in a bandwidth asymmetry network. The proposed bandwidth re-allocation scheme is also presented in this section. Numerical results and analysis are given in Section 7.3. At last, we conclude this chapter in Section 7.4.

7.2 Bandwidth Re-allocation for Bandwidth Asymmetry Mobile Wireless Networks

In Chapter 6, we illustrate how to compute the threshold for each call class. In order to find feasible threshold, we studied how the admission of class i new calls affects the dropping probability of call class j. In this section, we present a bandwidth re-allocation scheme based on the proposed DMS-AC scheme.

Before discussion, let us define *feasible threshold* to be the threshold Th_i for call class i in a specific cell with value between 0 and Δ_i, where Δ_i is the maximum number of class i calls that can be admitted in the cell under a given bandwidth allocation. If the computed threshold is greater than Δ_i, we can set the threshold to be Δ_i since the threshold greater than Δ_i could guarantee the QoS requirements and thus the threshold which is equal to Δ_i. On the other hand, if the derived threshold is smaller than 0, it means that we cannot find a feasible threshold under current bandwidth allocation and the QoS requirements of one or more call classes cannot be satisfied.

When system is unable to determine the feasible thresholds for some call classes, that implies the traffic load brought by some call classes exceeding the control range of the admission control scheme. The bandwidth re-allocation function could be triggered to adjust the bandwidth allocation between uplink and downlink and then the call admission thresholds are re-computed until the feasible thresholds are found. In this chapter, we assume that the feasible thresholds can be found

by adjusting bandwidth allocation between uplink and downlink of the cell if the traffic load exceeds the control range of the employed admission control scheme. If the thresholds cannot be found under any possible bandwidth allocation in a cell, it means that traffic load has exceeded the sustainable capacity of the cell. We do not consider such situation in this chapter as the bandwidth re-allocation problem becomes trivial in this situation. On the other hand, we cannot sacrifice too many new calls in order to guarantee the QoS of handoff calls. Thus, there should be some upper bounds of blocking probabilities for the new calls of different call classes. When the feasible thresholds cannot be found or the new call blocking probability reaches the predefined upper bound, the bandwidth re-allocation process is trigged. Next, we will discuss how to find the feasible bandwidth allocation on uplink and downlink. Since the basic procedure used in the proposed admission control scheme in two-cell system and multi-cell system are similar, we discuss the bandwidth allocation based on the former for ease of discussion and the obtained results and algorithm can be readily extended to the multi-cell system.

From Chapter (6), we know that $Th^1_{i,j}$ depends on the state (s^r_k) of the observing cell (C_r) while $Th^2_{i,j}$ depends on the state (s^l_k) of the neighboring cell (C_l). There are two cases that may result in bandwidth re-allocation. In case 1, if we cannot find a feasible threshold from (6.15) and (6.19) to satisfy the first admission condition, it implies that the bandwidth allocation of C_r should be adjusted. In case 2, if a proper threshold from (6.18) and (6.20) cannot be found to satisfy the second admission condition, that means the QoS of some call classes in C_l may be violated under the current traffic condition and bandwidth re-allocation should be executed in C_l. In the following we discuss these two cases in detail.

7.2.1 Case 1

Let us examine the first case. In the beginning of a control period, if the admission control scheme cannot find the feasible thresholds for some call classes, the system needs to re-allocate the bandwidth on uplink and downlink of the cell. From (6.15), we can compute the threshold for class i calls for a given η_j, where $i, j \in [1, M]$. We can rewrite (6.15) as

$$Th^1_{i,j}(s^r_k) = Az^2 + Bz + C \ , \tag{7.1}$$

where z is shown in (7.2)

$$z = \sqrt{4N_{i,j}(s_k^r) \cdot (1 - P_{i,r}^s) + (a_{i,j}^r)^2(1 - P_{i,r}^s)^2 + 4\theta_{i,l}P_{i,lr}^m(P_{i,r}^s - P_{i,lr}^m)}, \tag{7.2}$$

$A = \frac{1}{4P_{i,r}^s(1-P_{i,r}^s)}$, $B = -\frac{a_{i,j}^r}{2P_{i,r}^s}$ and $C = \frac{1}{4P_{i,r}^s}(a_{i,j}^r)^2(1 - P_{i,r}^s) - \frac{1}{P_{i,r}^s(1-P_{i,r}^s)}\theta_{i,l}P_{i,lr}^m(1 - P_{i,lr}^m)$.
From (7.1) we can find that the threshold $Th_{i,j}^1(s_k^r)$ is a function of z while z increases monotonously with $N_{i,j}(s_k^r)$ as $z > 0$. The value of $N_{i,j}(s_k^r)$ could be $1, 2, \cdots, \Delta_i^r$. Thus the value of z lies between $[z_{min}, z_{max}]$, where z_{min} and z_{max} are given in (7.3) and (7.4), respectively. We obtain z_{min} and z_{max} by setting $N_{i,j}(s_k^r)$ to be 1 and Δ_i^r, respectively.

$$z_{min} = \sqrt{4(1 - P_{i,r}^s) + (a_{i,j}^r)^2(1 - P_{i,r}^s)^2 + 4\theta_{i,l}P_{i,lr}^m(P_{i,r}^s - P_{i,lr}^m)} \tag{7.3}$$

$$z_{max} = \sqrt{4\Delta_i^r(1 - P_{i,r}^s) + (a_{i,j}^r)^2(1 - P_{i,r}^s)^2 + 4\theta_{i,l}P_{i,lr}^m(P_{i,r}^s - P_{i,lr}^m)} \tag{7.4}$$

In order to obtain (7.1), $P_{i,r}^s$ cannot be equal to 0 and 1. Since $P_{i,r}^s$ is a statistical variable used to represent the probability that class i calls remain in C_r during T, it is reasonable that $P_{i,r}^s \neq 0, 1$ though the value of $P_{i,r}^s$ may be very close to 0 or 1. Similarly, $P_{i,lr}^m \neq 0, 1$.

From the definitions of A, B and C, we realize that $B^2 - 4AC = \frac{\theta_{i,l}P_{i,lr}^m(1-P_{i,lr}^m)}{(P_{i,r}^s(1-P_{i,r}^s))^2} \geq 0$. When $(B^2 - 4AC) > 0$ ($\theta_{i,l} \neq 0$), we can sketch the curve of $Th_{i,j}^1(s_k^r)$ as the function of z as shown in Figure7.1. Regardless $C > 0$ or $C \leq 0$, we can obtain $z_{1,2} = a_{i,j}^r(1 - P_{i,r}^s) \pm 2\sqrt{\theta_{i,l}P_{i,lr}^m(1 - P_{i,lr}^m)}$, which are the solutions to the equation $Az^2 + Bz + C = 0$.

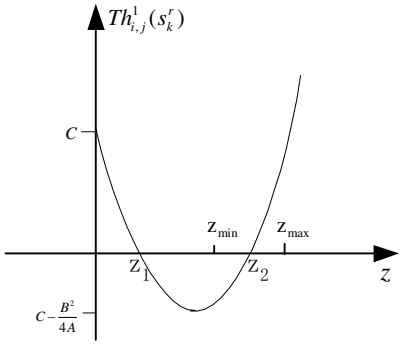

Figure 7.1: $Th_{i,j}^1(s_k^r)$ as a function of z when $\theta_{i,l} \neq 0$.

Obviously, $z_1 < z_{min} < z_{max}$, where $z_1 = a_{i,j}^r(1-P_{i,r}^s) - 2\sqrt{\theta_{i,l}P_{i,lr}^m(1 - P_{i,lr}^m)}$. We are concerned about whether or not $z_{max} > z_2$. If $z_{max} < z_2$, the values of $Th_{i,j}^1(s_k^r)$ are negative. In fact,

the negative threshold is infeasible, which means that the QoS of some call classes cannot be guaranteed although no class i calls can be admitted when the system is at some specific states. When $z_{max} < z_2$, we cannot find a feasible threshold for class i calls to satisfy the specific QoS requirement η_j of class j calls no matter which state that the system is at during the control period T. On the other hand, if there exists a feasible threshold, z_{max} must be greater than z_2. Let $z_{max} > z_2$ and we can obtain

$$\Delta_i^r > \theta_{i,l} P_{i,lr}^m + a_{i,j}^r \sqrt{\theta_{i,l} P_{i,lr}^m (1 - P_{i,lr}^m)} \tag{7.5}$$

where $\Delta_i^r = \min(\lfloor \frac{B_u^r}{b_i^u} \rfloor, \lfloor \frac{B_d^r}{b_i^d} \rfloor)$ is the maximum number of class i calls that can be admitted in C_r and it is totally determined by the bandwidth allocated to the uplink and the downlink of C_r.

Let $\alpha_i = \max\limits_{\forall j \in [1,M]} \left(\theta_{i,l} P_{i,lr}^m + a_{i,j}^r \sqrt{\theta_{i,l} P_{i,lr}^m (1 - P_{i,lr}^m)} \right)$. We obtain

$$\Delta_i^r > \alpha_i \ . \tag{7.6}$$

In order to find a feasible threshold, constraint (7.6) should be satisfied. Especially, when $\theta_{i,l} P_{i,lr}^m \gg a_{i,j}^r \sqrt{\theta_{i,l} P_{i,lr}^m (1 - P_{i,lr}^m)}$, i.e., $\theta_{i,l} \gg \frac{(a_{i,j}^r)^2 (1 - P_{i,lr}^m)}{P_{i,lr}^m}$, a feasible threshold for class i calls to satisfy the first admission condition exists only if the maximum admissible number of class i calls in the current observing cell is greater than the number of handoff class i calls from all neighboring cells, i.e., $\Delta_i^r > \theta_{i,l} P_{i,lr}^m$.

From the above analysis we know that the re-allocated bandwidth should satisfy (7.6). Since the total bandwidth in a cell is fixed and B_d^r can be obtained from $B_r - B_u^r$, we only show the relationship between Δ_i^r and B_u^r. From the definition of Δ_i^r and (7.6), we depict $\frac{B_r - B_u^r}{b_i^d}$ and $\frac{B_u^r}{b_i^u}$ as the function of B_u^r as shown in Figure 7.2 (a). We can see the curve of Δ_i^r consists of two segments represented by the solid lines in Figure 7.2 (a). In order to satisfy (7.6), the bandwidth of uplink (B_u^r) should be between $B_{\min}^i$ and $B_{\max}^i$, where $B_{\min}^i$ and $B_{\max}^i$ are the lower bound and the upper bound for B_u^r, respectively. When we consider multiple call classes in the cell, the feasible uplink bandwidth value should be between $(B_{\min}, B_{\max})$, where

$$B_{\min} = \max_{\forall i \in [1,M]} (B_{\min}^i) \tag{7.7}$$

and

$$B_{\max} = \min_{\forall i \in [1,M]} (B^i_{\max}) \ . \tag{7.8}$$

(B_{min}, B_{max}) is the common part of the ranges $(B^i_{\min}, B^i_{\max})$ for all $i \in [1, M]$. If $B_{\max} < B_{\min}$, it indicates the feasible bandwidth allocation under current traffic conditions cannot be found and we do not need to consider this situation as the problem becomes trivial. Figure 7.2 (b) shows an example when there are two call classes. According to (7.7) and (7.8), the feasible uplink bandwidth values should be between $B^2_{\min}$ and $B^1_{\max}$.

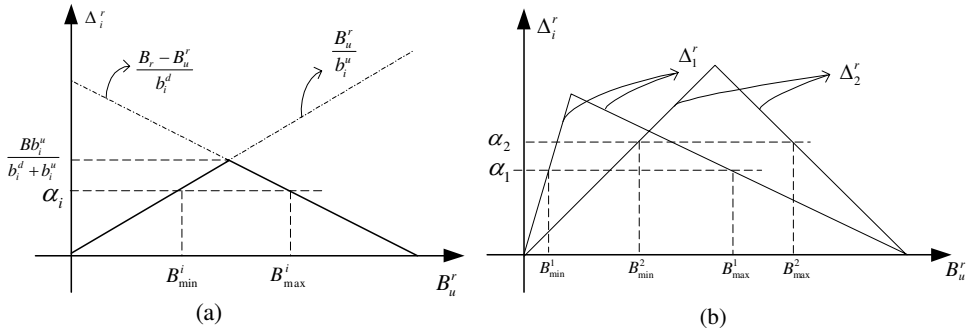

Figure 7.2: Δ_i^r as a function of B_u^r. (a) One call class. (b) Two call classes.

We regard the uplink bandwidth B_u^r between (B_{min}, B_{max}) as the feasible bandwidth. Accordingly, we can determine the feasible downlink bandwidth B_d^r. When the system has feasible bandwidth on both uplink and downlink, we regard the bandwidth allocation (B_u^r, B_d^r) as a feasible bandwidth allocation. Indeed, there could be multiple feasible bandwidth allocations of the cell. We select the one with minimal $|\gamma_u^r/\gamma_d^r - B_u^r/B_d^r|$ to maximize the system utilization as the solution, where γ_u and γ_d denote the time-average traffic load during a period on uplink and downlink respectively as that defined in [23]. Then we can try to find a threshold from (6.15) and (6.19) for class i based on the adjusted bandwidth allocation. If we still cannot find a feasible threshold, we should repeat the above process to find the new bandwidth allocation until a feasible threshold for class i calls is found. The details of the bandwidth re-allocation algorithm will be given in the subsequent sub-section.

When $B^2 - 4AC = 0$, i.e., $\theta_{i,l} = 0$, $z_1 = z_2 = a^r_{i,j}(1 - P^s_{i,r}) = z_0$. Obviously, $z_{min} > z_0$. We can depict the curve of $Th^1_{i,j}(s^r_k)$ as the function of z in this situation as shown in Figure 7.3. Since $\theta_{i,l} = 0$, the number of handoff calls from C_l during T is 0. In such extreme case, the threshold for

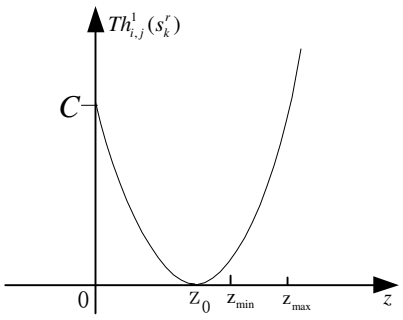

Figure 7.3: $Th^1_{i,j}(s^r_k)$ as a function of z when $\theta_{i,l} = 0$.

class i calls in C_r cannot be smaller than 0. Thus, there must be a feasible threshold for the class i calls.

7.2.2 Case 2

Next, let us consider the second case. The second admission condition requires that the number of the new class i calls should be limited in order to avoid the class i calls that handoff from C_r to C_l in the near future violating the QoS requirements of some higher priority call classes in C_l. From (6.18) and (6.20), we can find that Th^2_i is highly dependent on the system states of C_l. If we cannot find a feasible threshold for call class i to satisfy the second admission condition, it suggests that the bandwidth allocation in C_l be adjusted. By following the similar procedure used in case 1, we could obtain the following condition:

$$\Delta^l_i > \theta_{i,l} P^s_{i,l} + a^l_{i,j} \sqrt{\theta_{i,l} P^s_{i,l} (1 - P^s_{i,l})} \, . \tag{7.9}$$

If the maximum admissible number of class i calls in C_l does not satisfy the above condition, we cannot find a feasible threshold for class i calls in C_r to satisfy a specific QoS requirement in C_l no matter what system state C_l is at. If condition (7.9) cannot be satisfied or the feasible threshold cannot be found through (6.18) and (6.20), that means the bandwidth allocation in C_l needs to be adjusted. If we consider case 2 in cell C_l, (7.9) is changed to

$$\Delta^r_i > \theta_{i,r} P^s_{i,r} + a^r_{i,j} \sqrt{\theta_{i,r} P^s_{i,r} (1 - P^s_{i,r})} \, . \tag{7.10}$$

Let $\beta_i = \max_{\forall j \in [1,M]} \left(\theta_{i,r} P^s_{i,r} + a^r_{i,j} \sqrt{\theta_{i,r} P^s_{i,r} (1 - P^s_{i,r})} \right)$ and we have

$$\Delta^r_i > \beta_i \ , \tag{7.11}$$

which can be used to examine the feasibility of a bandwidth allocation by given a threshold for class i calls during a control period. At any time, the maximum admissible number of class i calls under a give bandwidth allocation should satisfy (7.11). Otherwise, we may not find feasible thresholds for some call classes in the neighboring cell C_l to satisfy the QoS requirements of some call classes in C_r.

7.2.3 Bandwidth Re-allocation Algorithm

Based on the above analysis of case 1 and case 2, we can describe the bandwidth re-allocation algorithm as follows: 1) At the beginning of a control period, if the admission control scheme cannot find the feasible thresholds for some call classes or the new call blocking probabilities of some call classes exceed the upper bounds, the bandwidth re-allocation function is triggered. 2) Then, the feasible bandwidth range $(B_{\min}, B_{\max})$ is calculated and the feasible bandwidth allocations can be obtained accordingly by using $B^r_d = B_r - B^r_u$. Next, we sorts all the feasible bandwidth allocations in ascending order according to the value of $|\gamma^r_u/\gamma^r_d - B^r_u/B^r_d|$. We select the first bandwidth allocation as the new bandwidth allocation for the system. 3) The thresholds are computed for each call class based on the new bandwidth allocation. If the feasible thresholds of some call classes cannot be found or the new call blocking probabilities of some call classes exceed the upper bounds, we select the second feasible bandwidth allocation as the new bandwidth allocation. Repeat this step until all feasible thresholds are found for every call class and the new call blocking probabilities are below the upper bounds. 4) Check whether the current bandwidth allocation and the threshold of class i calls satisfy (7.11) for all call classes. If (7.11) cannot be satisfied for some call classes, we need to find a new bandwidth allocation and repeat steps 3 and 4 until (7.11) is satisfied for all call classes. The pseudo code of the proposed bandwidth re-allocation algorithm is shown in Figure7.4.

```
BW_Reallocation{
  Find_Bmin_Bmax();          //find B_min and B_max
  B = B_min;
  while(B < Bmax){
     B = B + 1;
     U[i] = B;               //U[i]: feasible number of uplink channels
     i++;
  }
  BA_max = i;                //BA_max is the number of total feasible allocations
  BA[] = Sort(U[]);          //Sort bandwidth alloctions according to traffic load
  for(i=0; i<BA_max; i++)
  {
     //find threshold for every call class given a bandwidth allocation BA[i]
     Th[] = Threshold(BA[i]);
     //verify the feasibility of the thresholds
     for(j=0; j<M, j++)
     {
       if((Th[j] <= 0) || ( φ̂_i > ρ_i) ||(Δ_j <= β_j) )
         break;
       else if(j == M-1)
         return BA[i];
     }
  }
}
```

Figure 7.4: Pseudo code of bandwidth reallocation algorithm.

7.3 Performance Evaluation

In this section, we demonstrate the effectiveness of the proposed bandwidth re-allocation scheme. We consider a two-cell system which is composed of C_r and C_l and there are total 100 channels in each cell. Two call classes, RT call and NRT call, are considered. An RT call requires 1 channel on both uplink and downlink while an NRT call requires 1 channel on uplink and 3 channels on downlink. The highest tolerable handoff dropping probabilities of RT calls and NRT calls are 1% and 5%, respectively. The upper bounds of new call blocking probabilities for RT calls and NRT calls are 10% and 20%, respectively. We assume that the call arrivals follow Poison distribution and let λ^r_{RT} (λ^l_{RT}) and λ^r_{NRT} (λ^l_{NRT}) denote the mean call arrival rate of new RT calls and new NRT calls in C_r (C_l), respectively. The mean service time of RT calls and NRT calls is assumed to be 120 seconds and 900 seconds, respectively. The probability of a new RT call moves from one cell to another is 0.4 and the handoff probability of a new NRT call is 0.2. We also assume that the call will terminate in the target cell after it hands-off successfully.

We compare the performance of the system with DMS-AC only (termed "AC without BA") with

that of DMS-AC in conjunction with bandwidth re-allocation scheme (termed "AC with BA"). In order to examine the behaviors and the performance of the proposed approaches comprehensively, we conduct simulation experiments in five different scenarios. The changes of call arrival rates in the experiments are shown in Table 7.1. In the first two experiment scenarios, the call arrival rate of only one call class in a cell changes. In the subsequent three experiment scenarios, the call arrival rates of the same/different call classes in two cells change. We will examine the performance in terms of call blocking probability and resource utilization in these experiment scenarios.

Table 7.1: Call arrival rates in experiment scenarios

Experiment scenarios	C_r		C_l	
	λ^r_{RT}	λ^r_{NRT}	λ^l_{RT}	λ^l_{NRT}
1	$0.07 \sim 0.13$	0.01	0.05	0.01
2	0.1	$0.007 \sim 0.012$	0.1	0.005
3	$0.07 \sim 0.12$	0.01	$0.07 \sim 0.12$	0.01
4	0.1	$0.006 \sim 0.011$	0.1	$0.006 \sim 0.011$
5	$0.06 \sim 0.11$	0.01	0.1	$0.006 \sim 0.011$

7.3.1 Experiment 1: λ^r_{RT} increases from 0.07 to 0.13

In the first experiment, 30 channels are allocated to uplink and 70 channels are allocated to downlink in each cell initially. λ^r_{RT} increases from 0.07 to 0.13, which implies that the ratio of the number of arrival RT calls over all calls increases and thus the asymmetry degree of the traffic load between uplink and downlink decreases. When "AC with BA" is employed, the bandwidth re-allocation scheme increases the number of channels assigned to uplink of both cells C_r and C_l with λ^r_{RT} as shown in Figure 7.5. From the figure, we find that when the new RT call arrival rate is low ($\lambda^r_{RT} = 0.07$) the number of channels allocated to uplink and downlink of the system does not need to be adjusted. With the increase of λ^r_{RT}, the QoS of some call classes cannot be satisfied. In such situation, "AC with BA" allocates more channels to uplink. The increase of λ^r_{RT} results in more handoff RT calls from C_r to C_l. The number of channels allocated to uplink in C_l also increases but the increasing is not as fast as that of C_r. Since both "AC with BA" and "AC without BA" can keep the dropping probability of handoff NRT calls to be below 5%, we only show the handoff RT call dropping probability and the new RT/NRT call blocking probabilities in Figure 7.6 (a) and (b), respectively. These figures illustrate that "AC without BA" cannot guarantee the QoS

of handoff and new RT calls when λ_{RT}^{r} exceeds a certain level ($\lambda_{RT}^{r} \geq 0.08$). With bandwidth reallocation, the proposed admission control scheme satisfies the QoS requirement of handoff RT calls and guarantees the new RT/NRT call blocking probabilities to be below the upper bound. At the same time, the system bandwidth utilization is also notably improved by using bandwidth re-allocation as shown in Figure 7.7.

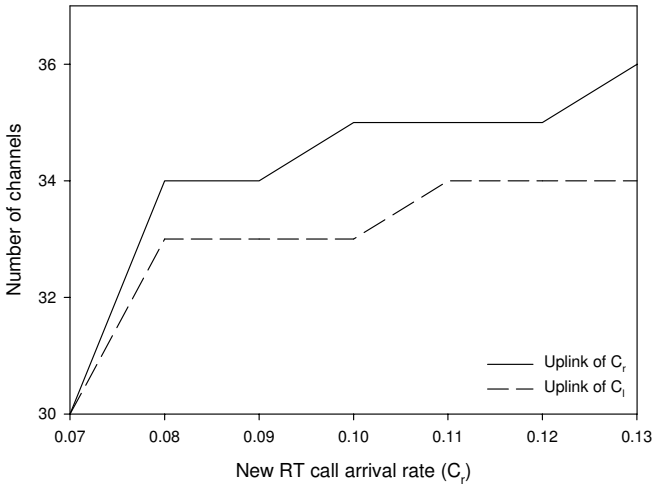

Figure 7.5: Change of the number of uplink channels when λ_{RT}^{r} increases from 0.07 to 0.13 (experiment 1).

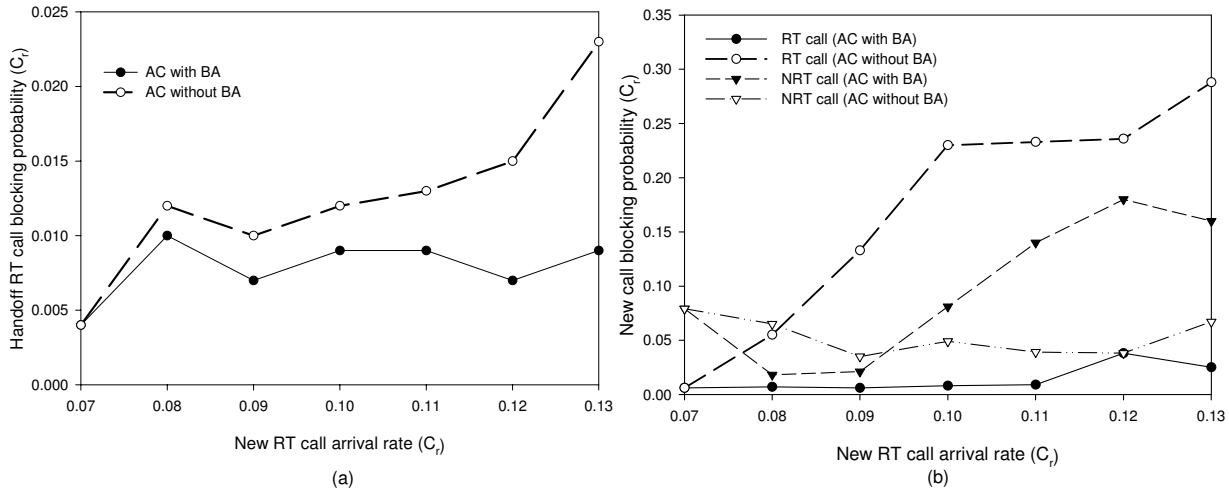

Figure 7.6: Call blocking probabilities of C_r when λ_{RT}^{r} increases from 0.07 to 0.13 (experiment 1). (a) Handoff RT call blocking probability. (b) New RT/NRT blocking probabilities.

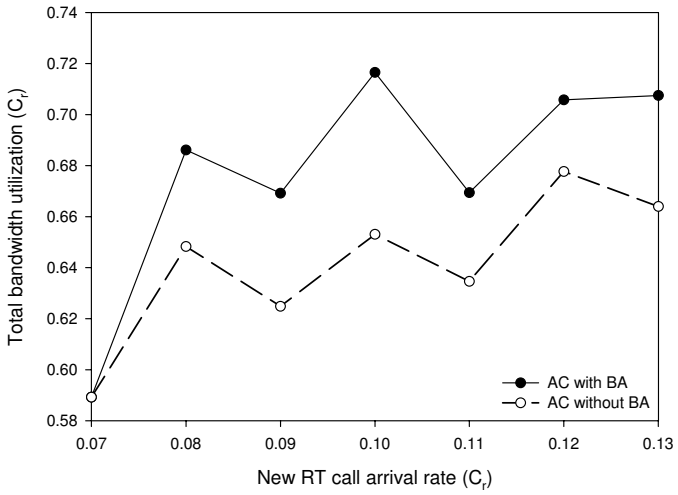

Figure 7.7: Total bandwidth utilization of C_r when λ^r_{RT} increases from 0.07 to 0.13 (experiment 1).

7.3.2 Experiment 2: λ^r_{NRT} increases from 0.007 to 0.012

In the second experiment, 50 channels are initially allocated to uplink and downlink respectively in each cell. Let λ^r_{NRT} increase from 0.007 to 0.012. With the increase of the NRT call arrival rate, the traffic load asymmetry between uplink and downlink becomes more evident and thus more channels should be allocated to downlink. We compare the performance of "AC without BA" and "AC with BA" in C_r. From the simulation results, we find that the call blocking probabilities of RT calls and handoff NRT calls of two schemes can be guaranteed below 1% and 5%, respectively. We show only the NRT blocking probability in Figure 7.8. Since "AC with BA" allocates more channels to downlink with the increase of λ^r_{NRT} as shown in Figure 7.9, more NRT calls can be accepted and thus the blocking probability of NRT calls is controlled below the upper bound. The total bandwidth utilization of "AC with BA" is also drastically higher than that of "AC without BA", as shown in Figure 7.10.

7.3.3 Experiment 3: λ^r_{RT} and λ^l_{RT} increase from 0.07 to 0.12 simultaneously

In this experiment, 30 channels are assigned to uplink in both cells initially. The average RT call arrival rates in both C_r and C_l increase from 0.07 to 0.12 simultaneously. With the increase of the RT call arrival rate in the system, "AC with BA" allocates more channels to uplink as shown in Figure 7.11. More RT calls could be accepted. Without bandwidth re-allocation, "AC without

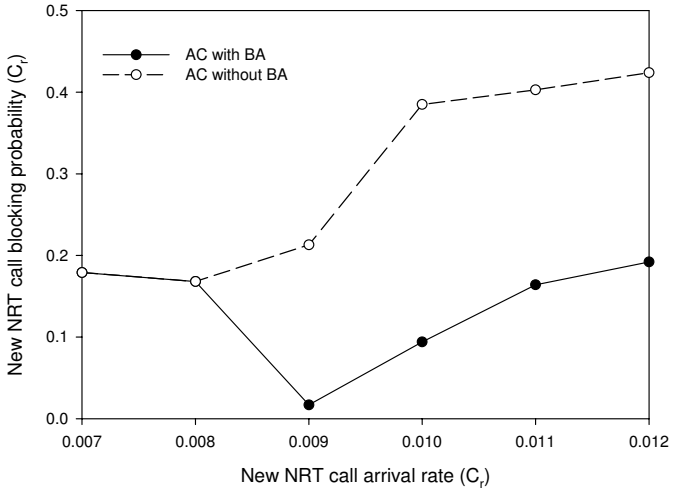

Figure 7.8: New NRT call blocking probability of C_r when λ^r_{NRT} increases from 0.007 to 0.012 (experiment 2).

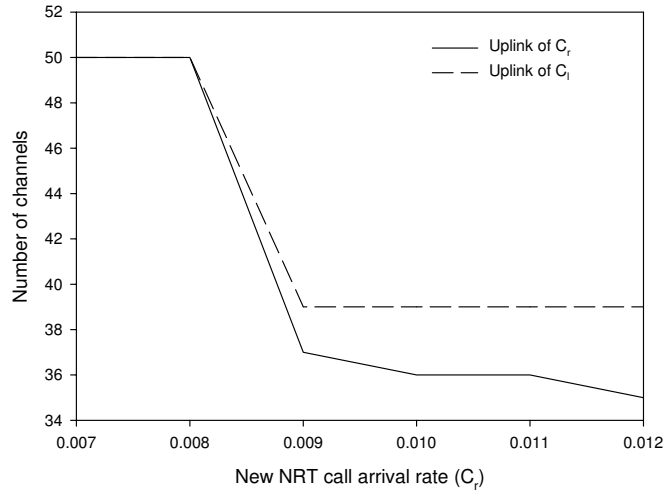

Figure 7.9: Change of the number of uplink channels when λ^r_{NRT} increases from 0.007 to 0.012 (experiment 2).

BA" blocks more RT calls in order to guarantee the QoS of handoff RT calls and thus cause the blocking probability of new RT calls to exceed the upper bound as shown in Figure 7.12. Obviously, "AC with BA" achieves much higher bandwidth utilization than "AC without BA" as shown in Figure 7.13.

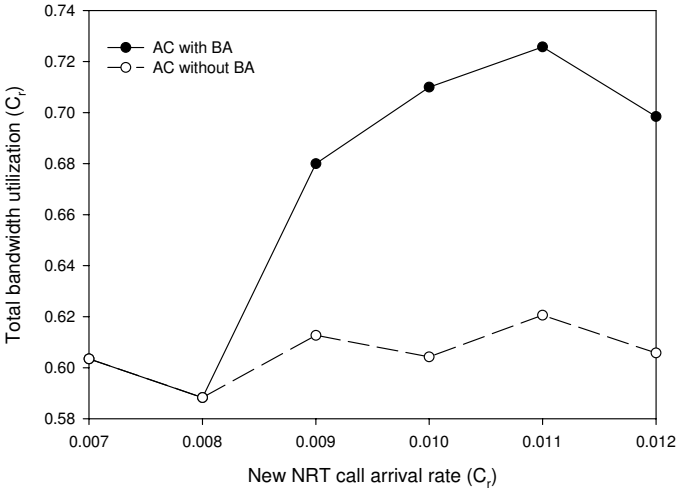

Figure 7.10: Total bandwidth utilization of C_r when λ^r_{NRT} increases from 0.007 to 0.012 (experiment 2).

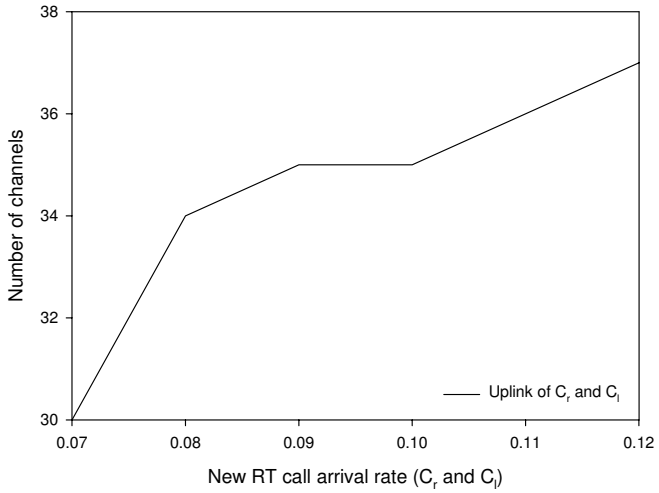

Figure 7.11: Change of the number of uplink channels when λ^r_{RT} and λ^l_{RT} increase from 0.07 to 0.12 simultaneously (experiment 3).

7.3.4 Experiment 4: λ^r_{NRT} and λ^l_{NRT} increase from 0.006 to 0.011 simultaneously

In this experiment, 50 channels are assigned to uplink in both cells initially. The average NRT call arrival rates in both C_r and C_l change from 0.006 to 0.011 simultaneously. With the increase of new NRT call arrival rate, the proposed bandwidth re-allocation scheme could assign more channels to downlink as shown in Figure 7.14 and thus "AC with BA" accepts more NRT calls than "AC without BA" while the QoS requirements of other call classes are also satisfied. Figure 7.15 compares the NRT call blocking probabilities of two schemes. "AC with BA" can guarantee

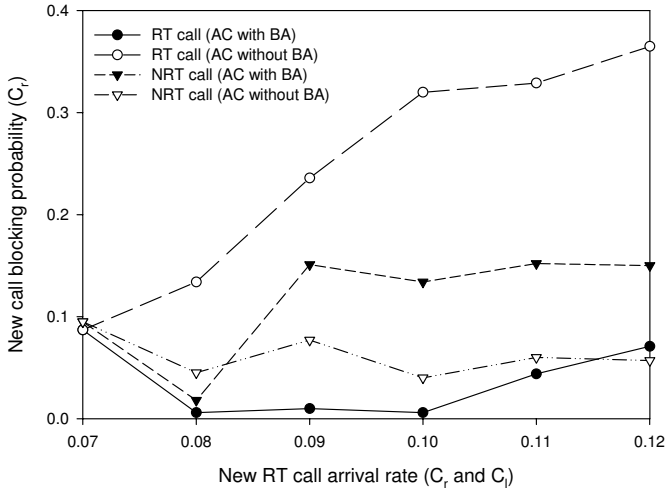

Figure 7.12: New call blocking probabilities of C_r when λ^r_{RT} and λ^l_{RT} increase from 0.07 to 0.12 simultaneously (experiment 3).

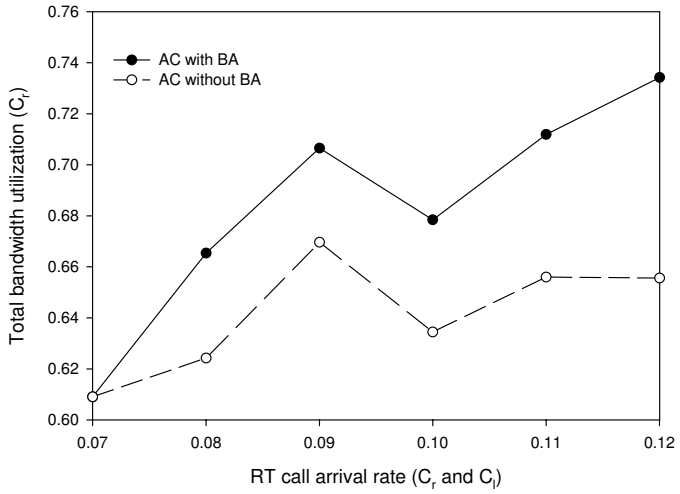

Figure 7.13: Total bandwidth utilization of C_r when λ^r_{RT} and λ^l_{RT} increase from 0.07 to 0.12 simultaneously (experiment 3).

the new NRT call blocking probability to be below the upper bound with the increase of average NRT call arrival rate. "AC with BA" also improves the system resource utilization significantly as shown in Figure 7.16.

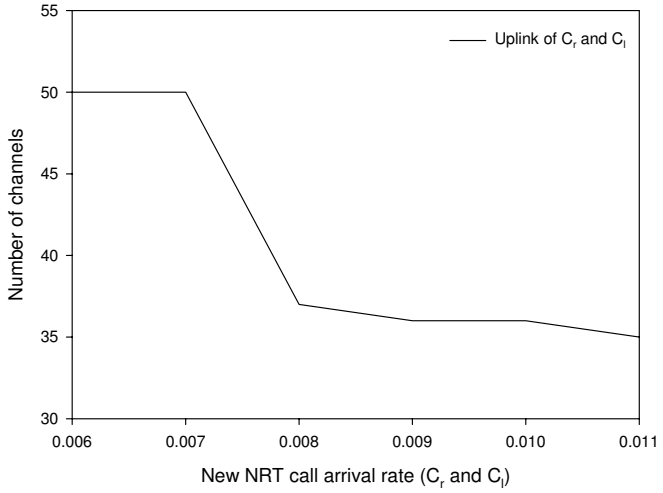

Figure 7.14: Change of the number of uplink channels when λ^r_{NRT} and λ^l_{NRT} increase from 0.006 to 0.011 simultaneously (experiment 4).

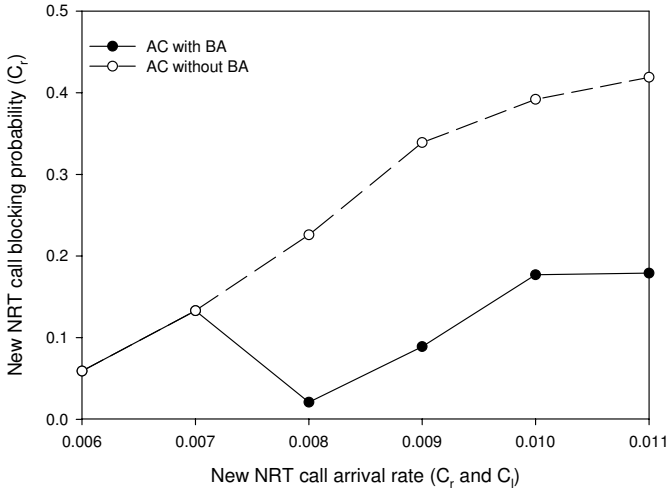

Figure 7.15: New NRT call blocking probability of C_r when λ^r_{NRT} and λ^l_{NRT} increase from 0.006 to 0.011 simultaneously (experiment 4).

7.3.5 Experiment 5: λ^r_{RT} increases from 0.06 to 0.11 while λ^l_{NRT} increases from 0.006 to 0.011 simultaneously

In this experiment, there are 50 channels on uplink of both cells initially. let λ^r_{RT} increase from 0.06 to 0.11 and λ^l_{NRT} increase from 0.006 to 0.011 simultaneously. This means that the traffic load asymmetry degree decreases in C_r but increases in C_l. Figure 7.17 shows the number of uplink channels assigned to uplink in both C_r and C_l when "AC with BA" is applied. From the figure, we find that the change of the uplink channels in C_r is more evident than that of C_l. Since RT calls

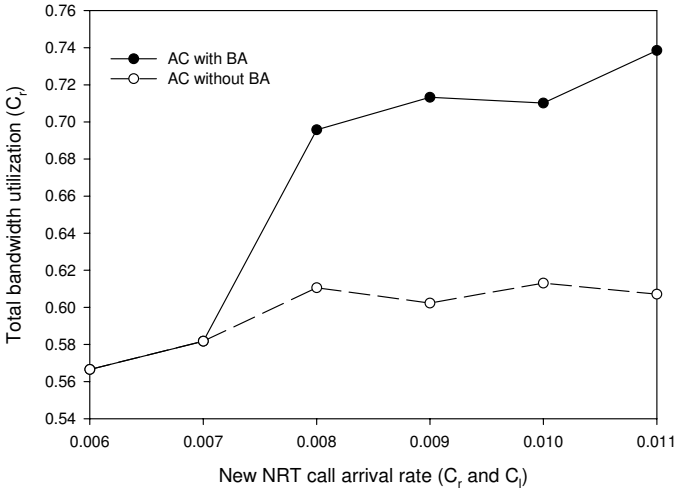

Figure 7.16: Total bandwidth utilization of C_r when λ^r_{NRT} and λ^l_{NRT} increase from 0.006 to 0.011 simultaneously (experiment 4).

have more stringent blocking probability requirements, the system is more sensitive to the change of the RT call arrival rate. In order to satisfy the QoS requirements of the high priority call classes, "AC without BA" blocks more new NRT calls in both C_r and C_l as shown in Figure 7.18 (a) and (b). Undoubtedly, "AC with BA" can achieve much higher bandwidth utilization in both cells as shown in Figure 7.19 (a) and (b).

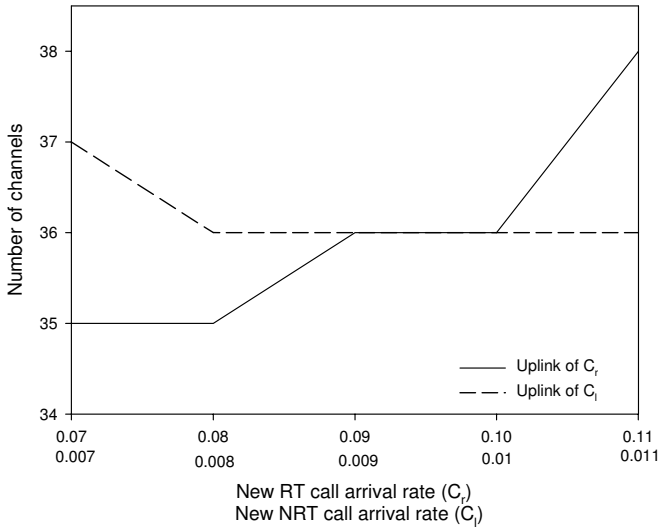

Figure 7.17: Change of the number of uplink channels when λ^r_{RT} increases from 0.06 to 0.11 while λ^l_{NRT} increases from 0.006 to 0.011 simultaneously (experiment 5).

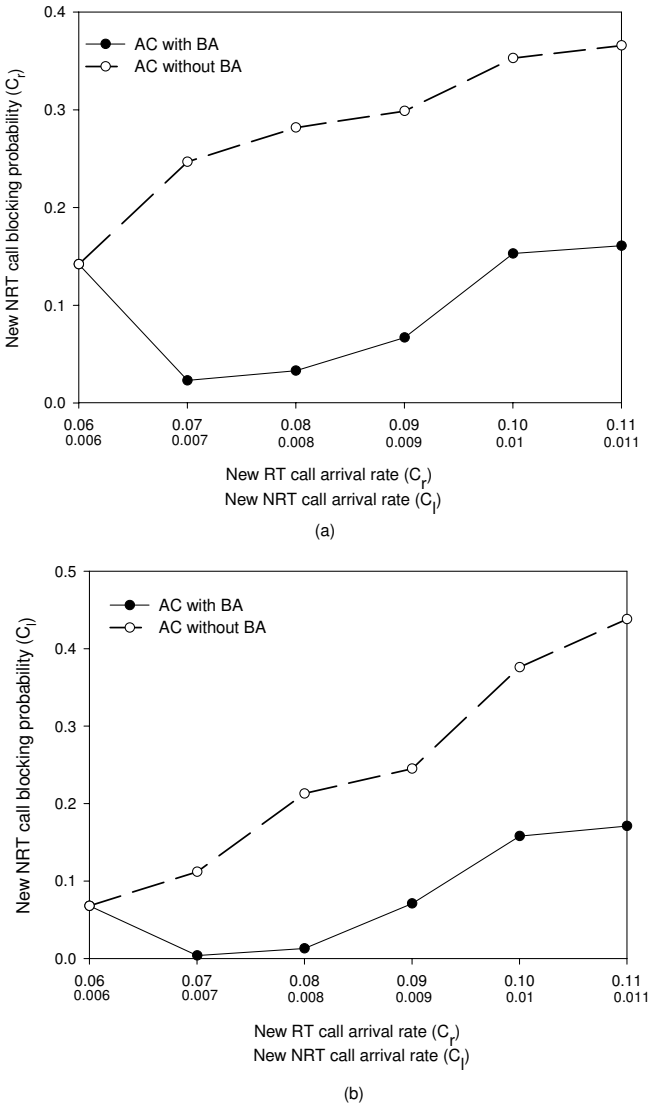

Figure 7.18: New NRT call blocking probabilities when λ^r_{RT} increases from 0.06 to 0.11 while λ^l_{NRT} increases from 0.006 to 0.011 simultaneously (experiment 5). (a) New NRT call blocking probability of C_r. (b) New NRT call blocking probability of C_l.

7.4 Summary

In multi-service mobile wireless networks, bandwidth allocation on uplink and downlink should be asymmetric to match the asymmetric traffic load. Under dynamic traffic load conditions, bandwidth asymmetry degree is changing accordingly. Thus bandwidth adjustment or re-allocation becomes an necessary mechanism to maximize the resource utilization while guaranteeing the QoS requirements of users. In this chapter, we study the problem- when and how to adjust bandwidth allocation between uplink and downlink under changing traffic load in multi-service wireless net-

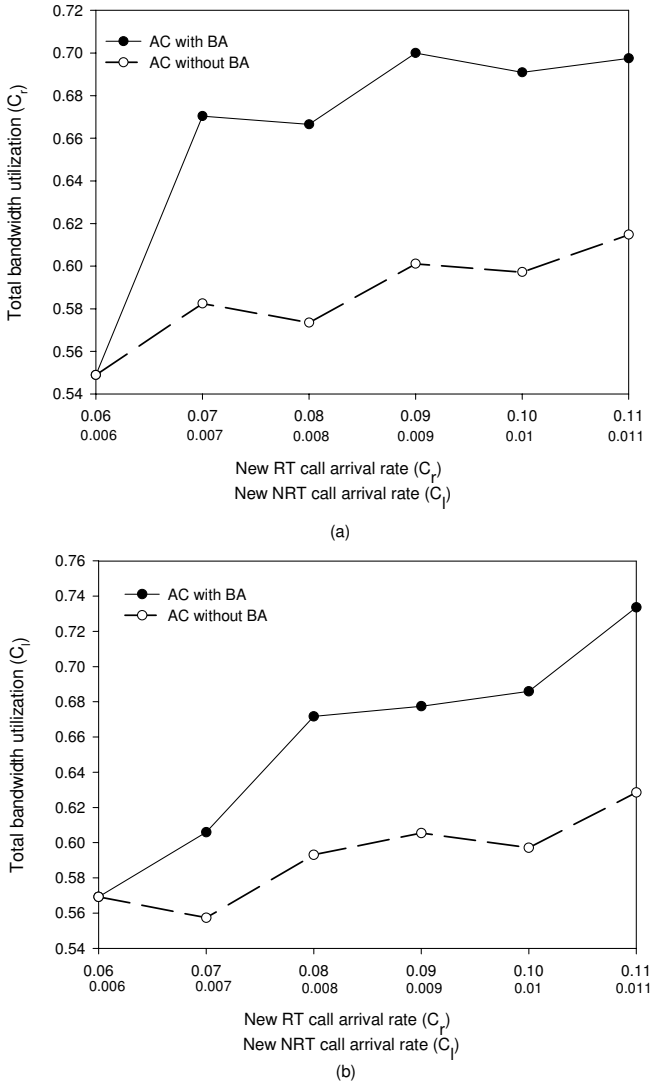

Figure 7.19: Total bandwidth utilization when λ_{RT}^{r} increases from 0.06 to 0.11 while λ_{NRT}^{l} increases from 0.006 to 0.011 simultaneously (experiment 5). (a) Total bandwidth utilization of C_r. (b) Total bandwidth utilization of C_l.

works. The design objective is to improve the system resource utilization while satisfying different QoS requirements of various call classes. We address the problem based on the proposed DMS-AC. When the traffic load brought by some call classes exceeds the control range of the employed admission control scheme and thus the QoS requirements of some call classes may not be guaranteed, the bandwidth re-allocation scheme is performed. Based on the proposed admission control scheme, we identify certain constraints that can be used to verify the feasibility of the bandwidth allocation of a cell. Numerical results show that the proposed admission control scheme in conjunction with

the bandwidth re-allocation scheme can guarantee the QoS of handoff calls and at the same time the new call blocking probabilities are maintained below some reasonable levels under dynamic traffic load conditions. Compared with that in static bandwidth allocation, the bandwidth utilization using our bandwidth re-allocation scheme under changing traffic load has been significantly improved.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis, we have addressed three main RM optimization problems: the MAXU problem, the MINCost problem and the MINBlock problem, in multi-service mobile cellular networks, especially in bandwidth asymmetry networks. By investigating two important RM issues, CAC and BA, we provided effective solutions to these problems.

In multi-service mobile networks with bandwidth asymmetry, the mismatch between dynamic traffic load and fixed bandwidth allocation results in low bandwidth utilization. In Chapter 4, we addressed the MAXU problem in bandwidth asymmetry mobile networks by proposing two CAC schemes. In bandwidth asymmetry mobile networks, inappropriate admission control scheme may accept "too many" RT/NRT calls and thus overuse uplink/dowlink bandwidth. We computed and set the admissible bandwidth regions for RT calls and NRT calls based on the traffic load of system to limit the admission of RT calls and NRT calls and thus prevent the calls of a specific class from overusing the limited bandwidth resources. The problems caused by the mismatch of bandwidth allocation and dynamic traffic load are solved and the system bandwidth utilization is also improved significantly. The simulation results show that the proposed schemes can avoid the low bandwidth utilization problems in the bandwidth asymmetry networks while the proposed Scheme 2 can guarantee the blocking probability of the handoff NRT calls at a low level without

deteriorating the blocking probability of RT calls when the arrival rate of handoff NRT call is not high. Compared with some existing CAC schemes such as GC scheme and Jeon's scheme, the proposed schemes can achieve a higher bandwidth utilization when traffic changes in bandwidth asymmetry networks. While the proposed Scheme 2 can guarantee the blocking probability of the high priority calls (the handoff RT calls and the handoff NRT calls) at a reasonable low level.

For the design of CAC in multi-service mobile networks, how to decrease the average system cost is one of critical issues. In Chapter 5, we explored the admission control policy for the MINCost problem in mobile networks with bandwidth asymmetry. We first formulated call admission decision into an MDP model and analyzed the corresponding value function. We find that the optimal admission policy for the MINCost problem in such asymmetric bandwidth allocation multi-service networks should have a threshold structure. The threshold specified for a call class may vary with system state. Due to the prohibitively high computational complexity, it is difficult to on-line compute the threshold for each call class in a real-time system with a large system state-space. Based on the analysis, we proposed a heuristic policy, CRDT policy, as a suboptimal solution to the MINCost problem for the bandwidth asymmetry mobile networks. The values of the thresholds in the CRDT policy can be computed readily. The numerical results show that the performance of the proposed CRDT policy is very close to that of the optimal policy obtained from the MDP model and better than that of other two known policies, which are also proposed for the bandwidth asymmetry multi-service mobile networks.

Handoff call blocking probability is always a critical QoS measurement in the design of call admission control no matter in traditional mono-service networks or in future multi-service mobile networks. In Chapter 6, we studied MINBlock problem in multi-service mobile networks and proposed DMS-AC scheme to address the MINBlock problem in multi-service environment. By setting thresholds for different call classes, DMS-AC prevents new calls from overusing system resources and reduces the number of potential handoff calls at the same time. The process of finding appropriate thresholds was illustrated comprehensively. First, we decomposed all system overload states into the overload states of individual call class and studied how the admission of calls from a specific class results in the overload states of other call classes. Based on the system states of local cell and the information from neighboring cells, DMS-AC can dynamically compute the

thresholds for various call classes. The numerical results show that DMS-AC is able to guarantee the handoff call blocking probabilities of different call classes under certain constraints in a dynamic traffic load environment with the expense of blocking more NRT calls. It is reasonable to make a tradeoff between low and high priority calls in order to guarantee the QoS of high priority calls. The experiment results also showed that the adjustment of bandwidth allocation is necessary to achieve better performance in asymmetric traffic load multi-service mobile networks.

In multi-service mobile cellular networks, bandwidth allocation on uplink and downlink should be asymmetric to match the traffic pattern/load. Under dynamic traffic load conditions, bandwidth asymmetry degree is changing accordingly. Thus bandwidth adjustment or re-allocation becomes an effective approach to maximize the resource utilization while guaranteeing the QoS of different call classes. In Chapter 7, we studied the problem–when and how to adjust bandwidth allocation between uplink and downlink under changing traffic load in multi-service mobile wireless networks. The design objective is to improve the system resource utilization while satisfying different QoS requirements of different call classes. We addressed the problem based on the DMS-AC scheme proposed in Chapter 6. When the traffic load brought by some call classes exceeds the control range of the employed admission control scheme and thus the QoS requirements of some call classes may not be satisfied, the bandwidth re-allocation is activated. Based on the proposed admission control scheme, we have identified certain constraints that can be used to verify the feasibility of the bandwidth allocation of a cell. Numerical results show that the proposed admission control scheme in conjunction with the bandwidth re-allocation scheme can guarantee the QoS of handoff calls and at the same time the new call blocking probabilities are controled below some predefined upper bounds under changing traffic load. Compared with that in static bandwidth allocation, the bandwidth utilization using our bandwidth re-allocation scheme under changing traffic load has been significantly improved.

8.2 Future Work

With the evolution of mobile wireless communication industry, diverse wireless technologies are proposed or applied nowadays. From 802.1X based networks, such as WiFi and WiMAX, to 3G, 4G or even future generation cellular mobile networks, people are making great effects to realize

the dream of always connection. Different wireless technology has its own advantages and disadvantages. For example, WLAN can support high bit-rate services but the coverage area is limited and handoff problem is still a research issue. While cellular networks may cover both metropolitan and country but the bandwidth is limited. No matter what physical technology is used, it has been widely accepted that future networks will be integrated at IP layer and eventually be evolved to *all-IP* networks [101–104]. In order to utilize all potential system resource and provide satisfied QoS to mobile users, "generalized mobility" is a key aspect of future networks, which provides mobile users seamless and transparent mechanisms for roaming between network operators and continual access to tailored services from a variety of environments while using a variety of terminals with varying capabilities [105]. Generalized mobility enables intelligent mobile devices to chose the most appropriate radio resources when several different physical resources are available. The generalized mobility process could be transparent to the end users and the users may/may not be involved into the resource selection process. Because of the complimentary in the coverage size of different radio resources, the end users will obtain desired QoS in diverse network environments and different traffic conditions and system resources are also utilized in a more efficient way. Due to the heterogeneous architecture of future all-IP networks, generalized mobility challenges the existing protocols and algorithms employed in the existing networks. In order to realize generalized mobility, many research issues should be addressed from physical layer to application layer and we present some possible research directions of our future work.

1) Mobility management in all-IP networks.

All-IP mobile networks accommodate diverse radio-access systems and offer multi-service among them in a seamless manner. Users will be able to choose the radio access system that offers the data speed, quality, and mobility best suited to the desired multimedia services. Mobility management will be one of the important factors in realizing seamless services over the all-IP wireless networks [104].

Mobility management includes location management and handoff management. Location management tracks and locates a terminal for delivering of incoming calls, while handoff management allows for an active connection to remain alive while the terminal roams. Location management

handles information concerning the mobile terminal, its original cell, the cell where it is currently located, and paths and routes toward the current location. So far, mobile IP [106] and its enhancement have been proposed for mobility location management in future mobile Internet along with the Internet architectural principles. According to the architectural principles of Internet, IETF RFC1958 [107] states that the goal of Internet is connectivity, the tool is the Internet Protocol, and the intelligence is end-to-end rather than hidden in the network. However, mobility management in cellular networks has been implemented as a network intelligence. That is, mobility management has been handled through collaboration between the network nodes within the mobile network. This confliction affects the performance of existing mobile IP strategies in future all-IP networks. For example, agent discovery is redundant for cellular networks since base station can detect users' handoff by supervising users' power level in conventional cellular networks. In addition, mobile IP takes about 1 second for a mobile host to be assigned the rerouting address for encapsulation (i.e., care-of address). This is too long for handoff in cellular systems. Rerouting over the all-IP mobile networks must be more efficient [104]. It is necessary to find more effective policies to solve the location management problem and provide the desired QoS for diverse users in future all-IP networks.

Handoff management in all-IP networks cares about the continuity of calls when users handoff between cells or roam in different radio access networks. Let us consider two possible handoff scenarios in future networks. We first consider multi-media conference scenario. Suppose a salesman should attend a group meeting when he is on a train. He needs to report the selling results and some market analysis to his manager and colleagues who attend the meeting. During his report, not only voice and video but also graphs, tables or even some power point files need to be shared between the participants of the meeting. When the salesman requests to start the call to join the meeting, his call includes multi-class sessions and each session has the same importance to his report. When such multi-session call hands-off, we cannot set the priority of different sessions according to the service type as we do in the traditional mobile wireless networks. We need to consider the relationship between different sessions belonging to a single call during the handoff procedure. With more and more different multi-media services provided in future mobile wireless networks, how to handle the multi-session handoff problem under different traffic conditions is a

interesting research topic that challenges the existing handoff management policies. In the second scenario, we assume that several young men play on-line game together in a coach during their trip from one city to another. These young men build up an ad hoc network and at the same time this ad hoc network connects to Internet through cellular network. When the coach moves between different cells between the two cities, game players need fast, fair and smooth handoff to guarantee their QoS. So far, few literatures study such "group handoff" case. With the increase of diverse applications provided in future heterogeneous mobile wireless networks, the above two scenarios will become common. In order to provide satisfied QoS to the users, the study of handoff management in future all-IP networks becomes a hot research topic and it is an interesting research direction in our future work.

2) Resource allocation and management in all-IP networks

Resource management, along with network planning and air interface design, determine QoS performance at the individual user level and network level as well [27]. In all-IP networks with heterogeneous architecture, different access technologies coexist. The mobile users in such networks will not just be pure sender or receiver but may play a important role in data transition as well as act as a cooperative agent for another user [108]. In such networks, it is necessary to consider resource allocation in different network domains across multiple layers in order to fully utilize potential system resource and provide desired QoS to end users. Because of the diversification of provided services and complicated heterogeneous network architecture, future IP-based mobile wireless networks require a more complex QoS model and more sophisticated management of scarce radio resources. QoS can be classified according to its implementation in the networks, based on a hierarchy of five different levels: bit, packet, session, call, and application. Transmission accuracy, system throughput, delay and delay jitter, fairness, and user perceived quality are the main considerations in this classification. To efficiently utilize scarce radio resources and achieve overall QoS satisfaction, cross-layer information is necessary. Since the link layer has statistical knowledge of the lower physical layer, such as the average channel capacity, it is better to jointly design the application layer or transport layer with link layer in order to guarantee the application-level QoS such as an acceptable visual quality of video services or a guaranteed TCP throughput of

data services [109]. We have to be very careful when we design cross-layer protocols and policies, since cross-layer design breaks the existing network layer architecture and may bring big troubles to whole system although it may enhance the performance in some sub-networks. Thus, it is necessary to delicately study cross-layer design for resource management in future all-IP networks.

Author's Publications

- Xun Yang and Gang Feng, "Bandwidth Re-allocation for Bandwidth Asymmetry Wireless Networks based on Distributed Multi-Service Admission Control", accepted by IEEE Transactions on Mobile Computing.

- Xun Yang and Gang Feng, "Dynamic Bandwidth Allocation for Bandwidth Asymmetry Wireless Networks based on Distributed Multi-Service Admission Control", in Proc. IEEE ICC 2007, GLASGOW, SCOTLAND, Jun. 2007.

- Xun Yang and Gang Feng, "Optimizing Admission Control for Multi-service Wireless Networks with Bandwidth Asymmetry between Uplink and Downlink," IEEE Trans. Veh. Technol., vol. 56, 907 - 917, Mar. 2007.

- Xun Yang and Gang Feng, "Cost Minimization for Admission Control in Bandwidth Asymmetry Wireless Networks", in Proc. IEEE ICC 2007, GLASGOW, SCOTLAND, Jun. 2007.

- Xun Yang, Gang Feng, and C. K. Siew, "Call Admission Control for Multi-service Wireless Networks with Bandwidth Asymmetry between Uplink and Downlink," IEEE Trans. Veh. Technol., vol. 55, pp. 360 - 368, Jan. 2006.

- Feng Xie, Gang Feng, and Xun Yang, "Optimizing Caching Policy for Loss Recovery in Reliable Multicast", in Proc. IEEE INFOCOM 2006, Barcelona, Spain, Apr. 2006.

- Xun Yang, Gang Feng, and Chee Kheong Siew, "Call Admission Control for Multi-Service Mobile Networks with Bandwidth Asymmetry between Uplink and Downlink", in Proceedings of IEEE GLOBECOM 2004, Dalas, USA, Nov./Dec. 2004.

- Gang Feng, Xun Yang and C. K. Siew, "Chapter 7: Call Admission Control for Multi-Service

Wireless Networks with Bandwidth Asymmetry between Uplink and Downlink", Resource Allocation in Next Generation Wireless Networks, pp.143-170, Publisher: Nova Biomedical (5 Aug. 2005).

Bibliography

[1] A. Jamalipour, *The wireless mobile Internet*. John wiley & Sons Incl, 2003.

[2] "Magic mobile future 2010-2020," Report No.37 From The UMTS Forum, 2005.

[3] J. Diederich and M. Zitterbart, "Handoff prioritization schemes using early blocking," *IEEE COMM. SURVEYS*, vol. 7, pp. 26 – 45, Second Quarter 2005.

[4] W. C. Y. Lee, "Smaller cells for greater performance," *IEEE Communications Magazine*, vol. 29, pp. 19 – 23, Nov. 1991.

[5] P. Chaudhury, "The 3GPP proposal for IMT-2000," *IEEE Communications Magazine*, vol. 37, pp. 72 – 81, Dec. 1999.

[6] "3G offered traffic characteristics," Report No.33 From The UMTS Forum, 2003.

[7] M. Zeng, A. Annamalai, and V. K. Bhargava, "Recent advances in cellular wireless communications," *IEEE Communications Magazine*, vol. 37, Sep. 1999.

[8] W. S. Jeon and D. G. Jeong, "Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink," *IEEE Transactions on Vehicular Technology*, vol. 50, pp. 59 – 66, Jan. 2001.

[9] D. G. Jeong and W. S. Jeon, "CDMA/TDD system for wireless multimedia services with traffic unbalance between uplink and downlink," *IEEE Journal on Selected Areas in Communications*, vol. 17, May 1999.

[10] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, pp. 968 – 981, 1991.

[11] A. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission of high speed networks," *IEEE/ACM Trans. on Networking*, vol. 1, pp. 329 – 343, 1993.

[12] K. Sohraby, "On the asymptotic behavior of heterogeneous statistical multiplexer with applications," *Proc. INFOCOM'92*, pp. 839 – 847, 1992.

[13] H. Saito, "Call admission control in an ATM network using upperbound of cell loss probability," *IEEE Transactions on Communications*, pp. 1512 – 1521, 1992.

[14] T. Faber and L. Landweber, "Dynamic time windows: packet admission control with feedback," *Proc. SIGCOM'92*, pp. 124 – 135, 1992.

[15] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, pp. 982 – 989, Sep. 1991.

[16] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radiotelephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, pp. 77 – 92, Aug. 1986.

[17] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, pp. 711 – 717, May 1996.

[18] C. Oliveira, J. B. Kim, and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 858 – 874, Aug. 1998.

[19] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," *Proc. INFOCOM'96*, vol. 1, pp. 43 – 50, Mar. 1996.

[20] E. Altman, T. Jimenez, and G. Koole, "On optimal call admission control in a resource-sharing system," *IEEE Transactions on Communications*, vol. 49, pp. 1659 – 1668, Sep. 2001.

[21] B.Li, L.Li, B.Li, K. Sivalingam, and X.R.Cao, "Call admission control for voice/data integrated cellular networks: performance analysis and comparative study," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 706 – 718, May 2004.

[22] W. S. Jeon and D. G. Jeong, "Comparison of time slot allocation strategies for CDMA/TDD system," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 1271 – 1278, Jul. 2000.

[23] D. G. Jeong and W. S. Jeon, "CDMA/TDD system for wireless multimedia services with traffic unbalance between uplink and downlink," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 939 – 946, May 1999.

[24] X. Yang and G. Feng, "Optimizing admission control for multi-service wireless networks with bandwidth asymmetry between uplink and downlink," *IEEE Transactions on Vehicular Technology*, vol. 56, pp. 907 – 917, Mar. 2007.

[25] J. Zhang, J. Huai, R. Y. Xiao, and B. Li, "Resource management in the next-generation DS-CDMA cellular networks," *IEEE Wireless Communications Magazine*, vol. 11, pp. 52 – 58, Aug. 2004.

[26] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," *University of Massachusetts Technical Report UM-CS-1995-064*, Jul. 1995.

[27] M. Ahmed, "Call admission control in wireless networks: a comprehensive survey," *Communications Surveys and Tutorials*, vol. 7, pp. 49 – 68, First Quarter 2005.

[28] A. Sutivong and J. M. Peha, "Performance comparisons of call admission control algorithms in cellular systems," *Proc. GLOBECOM'97*, vol. 3, pp. 1645 – 1649, Nov. 1997.

[29] K. Lee, "Adaptive network support for mobile multimedia," *Proc. MOBICOM'95*, pp. 62 – 74, 1995.

[30] I. F. A. David A. Levine and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 1 – 12, Feb. 1997.

[31] O. T. W. Yu and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless pcn," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1208 – 1225, Sep. 1997.

[32] C.-C. Chao and W. Che, "Connection admission control for mobile multiple-class personal communications networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1618 – 1626, Oct. 1997.

[33] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks," *Proc. SIGCOMM '98*, vol. 28, pp. 155 – 166, 1998.

[34] P. Ramanathan, K. M. Sivalingam, P. Agrawal, and S. Kishore, "Dynamic resource allocation scheme during handoff for mobile multimedia wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1270 – 1283, Jul. 1999.

[35] P. Ramanathanl, K. M. Sivalingam, P. Agrawal, and S. Kishore, "Resource allocation during handoff through dynamic schemes for mobile multimedia wireless networks," *Proc. INFOCOM'99*, vol. 3, pp. 1204 – 1211, Mar. 1999.

[36] 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, pp. 523 – 534, Mar. 2000.

[37] D. Zhao, X. Shen, and J. W. Mark, "Efficient call admission control for heterogeneous services in wireless mobile ATM networks," *IEEE Communications Magazine*, vol. 38, pp. 72 – 78, Oct. 2000.

[38] M.-H. 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, pp. 510 – 522, Mar. 2000.

[39] J. Mišić and T. Y. Bun, "Adaptive admission control in wireless multimedia networks under nonuniform traffic conditions," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 2429 – 2442, Nov. 2000.

[40] W.-S. Soh and H. S. Kim, "Dynamic guard bandwidth scheme for wireless broadband networks," *Proc. INFOCOM'01*, vol. 1, pp. 572 – 581, Apr. 2001.

[41] T. Zhang, E. van den Berg, J. Chennikara, P. Agrawal, J.-C. Chen, and T. Kodama, "Local predictive resource reservation for handoff in multimedia wireless IP networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 1931 – 1941, Oct. 2001.

[42] Y. Zhang and D. Liu;, "An adaptive algorithm for call admission control in wireless networks," *Proc. GLOBECOM'01*, vol. 6, pp. 3628 – 3632, Nov. 2001.

[43] W.-S. Soh and H. S. Kim, "QoS provisioning in cellular networks based on mobility prediction techniques," *IEEE Communications Magazine*, vol. 41, pp. 86 – 92, Jan. 2003.

[44] H. Zeng and I. Chlamtac, "Adaptive guard channel allocation and blocking probability estimation in PCS networks," *Computer Networks*, vol. 43, pp. 163 – 176, Oct. 2003.

[45] X. Chen and Y. Fang, "An adaptive bandwidth reservation scheme in multimedia wireless networks," *Proc. GLOBECOM'03*, vol. 5, pp. 2830 – 2834, Dec. 2003.

[46] J. Y. Lee, J.-G. Choi, K. Park, and S. Bahk, "Realistic cell-oriented adaptive admission control for QoS support in wireless multimedia networks," *IEEE Transactions on Vehicular Technology*, vol. 52, pp. 512 – 524, May 2003.

[47] C. Quevedo-Lodi and J. de Marca, "Performance of a dynamic multiple class admission control strategy for wireless systems," *Proc. VTC'04*, vol. 2, pp. 1083 – 1087, Sep. 2004.

[48] X. Chen, B. Li, and Y. Fang, "A dynamic multiple-threshold bandwidth reservation (DMTBR) scheme for QoS provisioning in multimedia wireless networks," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 583 – 592, Mar. 2005.

[49] G. Schembra, "A resource managements strategy for multimedia adaptive-rate traffic in a wireless network with TDMA access," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 65 – 78, Jan. 2005.

[50] M. Salamah and H. Lababidi, "Dynamically adaptive channel reservation scheme for cellular networks," *Computer Networks*, vol. 49, pp. 787 – 796, Mar. 2005.

[51] H. B. Kim, "An adaptive bandwidth reservation scheme for multimedia mobile cellular networks," *Proc. ICC'05*, vol. 5, pp. 3088 – 3094, May 2005.

[52] T.-C. Chau, K. Y. M. Wong, and B. Li, "Optimal call admission control with QoS guarantee in a voice/data integrated cellular network," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 1133 – 1141, May 2006.

[53] W.-S. Soh and H. S. Kim, "A predictive bandwidth reservation scheme using mobile positioning and road topology information," *IEEE/ACM Transactions on Networking*, vol. 14, pp. 1078 – 1091, Oct. 2006.

[54] O. Yu, E. Saric, and A. Li, "Fairly adjusted multimode dynamic guard bandwidth admission control over CDMA systems," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 579 – 592, Mar. 2006.

[55] A. Aljadhai and T. Znati, "A predictive, adaptive scheme to support QoS guarantees in multimedia wireless networks," *Proc. ICC'99*, vol. 1, pp. 221 – 225, Jun. 1999.

[56] ——, "Predictive mobility support for QoS provisioning in mobile wireless environments," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 1915 – 1930, Oct. 2001.

[57] J. Diedericha and M. Zitterbart, "A simple and scalable handoff prioritization scheme," *Computer Communications*, vol. 28, pp. 773 – 789, May 2005.

[58] M. Oliver and J. Borras, "Performance evaluation of variable reservation policies for hand-off prioritization in mobile networks," *Proc. INFOCOM'99*, vol. 3, pp. 1187 – 1194, Mar. 1999.

[59] T. S and B. J. Jabbari, "A measurement-based prioritization scheme for handovers in mobile cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 10, pp. 1343 – 1350, Oct. 1992.

[60] J. N. Daigle and N. Jain, "A queueing system with two arrival streams and reserved servers with application to cellular telephone," *Proc. INFOCOM'92*, vol. 3, pp. 2161 – 2167, May 1992.

[61] R. Guérin, "Queueing-blocking system with two arrival streams and guard channels," *IEEE Transactions on Communications*, vol. 36, pp. 153 – 163, Feb. 1998.

[62] C. H. Yoon and C. K. Un, "Performance of personal portable radio telephone systems with and without guard channels," *IEEE Journal on Selected Areas in Communications*, vol. 11, pp. 911 – 917, Aug. 1993.

[63] A. E. Xhafa and O. K. Tonguz, "A new queuing scheme for handoffs in 3G wireless networks," *Proc. VTC'01*, vol. 2, pp. 738 – 742, Oct. 2001.

[64] O. K. Tonguz and A. E. Xhafa, "Improving handover performance in wireless networks: dynamic priority queuing versus guard channel method," *Electronics Letters*, vol. 38, pp. 338 – 339, Mar. 2002.

[65] A. E. Xhafa and O. K. Tonguz, "Dynamic priority queueing of handover calls in wireless networks: An analytical framework," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 904 – 916, Jun. 2004.

[66] X. Tian and C. Ji, "QoS provisioning with distributed call admission control in wireless networks," *Proc. ICC'98*, vol. 2, pp. 797 – 801, Jun. 1998.

[67] ——, "Bounding the performance of dynamic channel allocation with QoS provisioning for distributed admission control in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 50, pp. 388 – 397, Mar. 2001.

[68] S. G. Jiang, B. Li, X. Y. Luo, and D. H. K. Tsang, "A modified distributed call admission control scheme and its performance," *Wireless Networks*, vol. 7, pp. 127 – 138, 2001.

[69] Y. Iraqi and R. Boutaba, "A novel distributed call admission control for wireless mobile multimedia networks," *Proc. WOWMOM'00*, pp. 21 – 27, Aug. 2000.

[70] ——, "When is it worth involving several cells in the call admission control process for multimedia cellular networks?" *Proc. ICC'01*, pp. 336 – 340, Jun. 2001.

[71] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 257 – 271, Apr. 2002.

[72] K. W. Ross and D. H. K. Tsang, "The stochastic knapsack problem," *IEEE Transactions on Communications*, vol. 37, pp. 740 – 747, Jul. 1989.

[73] M. Gavious and Z. Rosberg, "A restricted complete sharing policy for a stochastic knapsack problem in B-ISDN," *IEEE Transactions on Communications*, vol. 42, p. 2375 C 2379, Jul. 1994.

[74] E. A. Feinberg and M. I. Reiman, "Optimality of randomized trunk reservation," *Probability in the Engineering and Informational Sciences*, vol. 8, pp. 463 – 489, 1994.

[75] J. Choi, T. Kwon, Y. Choi, and M. Naghshineh, "Call admission control for multimedia services in mobile cellular networks: a markov decision approach," *Proc. ISCC'2000*, pp. 594 – 599, Jul. 2000.

[76] R. Akl, M. Hegde, M. Naraghi-Pour, and P. Min, "Call admission control scheme for arbitrary traffic distribution in CDMA cellular systems," *Proc. WCNC'2000*, vol. 1, pp. 465 – 470, Sep. 2000.

[77] H. Tong and T. X. Brown, "Adaptive call admission control under quality of service constraints: A reinforcement learning solution," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 209 – 221, Feb. 2000.

[78] J. Hou, J. Yang, and S. Papavassiliou, "Integration of pricing with call admission control to meet QoS requirements in cellular netowrks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, pp. 11 – 29, Sep. 2003.

[79] W. Ibrahim, J. Chinnek, and S. Periyalwar, "A QoS-based charging and resource allocation framework for next generation wireless networks," *Wireless Communications and mobile Computing*, vol. 3, pp. 895 – 906, Nov. 2003.

[80] J. Ni and S. Tatikonda, "Revenue optimization via call admission control and pricing for mobile cellular systems," *Proc. ICC'05*, vol. 5, pp. 3359 – 3364, May 2005.

[81] F. Yu, V. Wong, and V. Leung, "A new QoS provisioning method for adaptive multimedia in cellular wireless networks," *Proc. INFOCOM'04*, vol. 3, pp. 2130 – 2141, Mar. 2004.

[82] K. Kuppuswamy and D. C. Lee, "On subscription admission control for network service provision," *IEEE Communications Letters*, vol. 9, pp. 66 – 68, Jan. 2005.

[83] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[84] M. Haardt, A. Klein, R. Koehn, S. Oestreich, M. Purat, V. Sommer, and T. Ulrich, "The TD-CDMA based UTRA TDD mode," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 1375 – 1385, Aug. 2000.

[85] W. Jeong and M. Kavehrad, "Dynamic-TDD and optimum slot-allocation in fixed cellular systems," *Proc. VTC'01*, vol. 1, pp. 37 – 41, 2001.

[86] F. Nazzarri and R. Ormondroyd, "An effective dynamic slot allocation strategy based on zone division in WCDMA/TDD systems," *Proc. VTC'02*, vol. 2, pp. 646 – 650, Sep. 2002.

[87] J. Nasreddine and X. Lagrange, "Time slot allocation based on a path gain division scheme for TD-CDMA TDD systems," *Proc. VTC'03*, vol. 2, pp. 1410 – 1414, Apr. 2003.

[88] R. Berezdivin, R. Breinig, and R. Topp, "Next-generation wireless communications concepts and technologies," *IEEE Communications Magazine*, vol. 40, pp. 108 – 116, Mar. 2002.

[89] D. G. Jeong and W. S. Jeon, "Time slot allocation in CDMA/TDD systems for mobile multimedia services," *IEEE Communications Letters*, vol. 4, Feb. 2000.

[90] *Spectrum requirements for IMT-2000*, ITU-R IMT-SPEC, Rev. ITU-R Report M.2023.

[91] X. Yang, G. Feng, and C. K. Siew, "Call admission control for multi-service wireless networks with bandwidth asymmetry between uplink and downlink," *IEEE Transactions on Vehicular Technology*, vol. 55, pp. 360 – 368, Jan. 2006.

[92] Y. G. Fang, "Thinning scheme for call admission control in wireless networks," *IEEE Transactions on Computers*, vol. 52, pp. 685 – 687, May 2003.

[93] L.Huang, S.Kumar, and C.-C. J. Kuo, "Adaptive resource allocation for multimedia QoS management in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 53, pp. 547 – 558, Mar. 2004.

[94] S.Kim and P. K. Varshney, "An integrated adaptive bandwidth-management framework for QoS-sensitive multimedia cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 53, pp. 835 – 846, May 2004.

[95] F.Hu and N. K. Sharma, "Priority-determined multiclass handoff scheme with guaranteed mobile QoS in wireless multimedia networks," *IEEE Transactions on Vehicular Technology*, vol. 53, pp. 118 – 135, Jan. 2004.

[96] F. A. Cruz-Perez and L. Ortigoza-Guerrero, "Flexible resource allocation strategies for class-based QoS provisioning in mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 53, pp. 805 – 819, May 2004.

[97] D. P. Bertsekas, *Dynamic programming: deterministic and stochastic models.* Prentice-hall, INC, 1987.

[98] G. Koole, "Structural results for control of queueing systems using event-based dynamic programming," *Queueing System*, vol. 30, pp. 323 – 339, 1998.

[99] C. Chatfield, *The analysis of time series: an introduction.* Chapman and Hall, 1984.

[100] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1991.

[101] L. Bos and S. Leroy, "Toward an all-IP-based UMTS system architecture," *IEEE Network*, vol. 15, pp. 36 – 45, Jan. 2001.

[102] P. Newman, "In search of the all-IP mobile network," *IEEE Communications Magazine*, vol. 42, pp. S3 – S8, Dec. 2004.

[103] I. Akyildiz, J. Xie, and S. Mohanty, "A survey of mobility management in next-generation all-IP-based wireless systems," *IEEE Wireless Communications Magazine*, vol. 11, pp. 16 – 28, Aug. 2004.

[104] M. Yabusaki, T. Okagawa, and K. Imai, "Mobility management in all-IP mobile network: end-to-end intelligence or network intelligence?" *IEEE Communications Magazine*, vol. 43, pp. supl.16–supl.24, Dec. 2005.

[105] M. Carugi, B. Hirschman, and A. Narita, "Introduction to the ITU-T NGN focus group release 1: target environment, services, and capabilities," *IEEE Communications Magazine*, vol. 43, pp. 42 – 48, Oct. 2005.

[106] C. Perkins, "IP mobility support," *RFC 2002*, Oct. 1996.

[107] B. Carpenter, "Architectural principles of the Internet," *RFC 1958*, Jun. 1996.

[108] A. Nosratinia and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Communications Magazine*, vol. 42, pp. 74 – 80, Oct. 2004.

[109] H. Jiang, W. Zhuang, and X. Shen, "Cross-layer design for resource allocation in 3G wireless networks and beyond," *IEEE Communications Magazine*, vol. 43, pp. 120 – 126, Dec. 2005.