CHANNEL ALLOCATION WITH QOS GUARANTEE IN MOBILE WIRELESS NETWORKS: SINGLE AND MULTIPLE TYPES OF TRAFFIC

BY

YIN LI

A Thesis Presented to
The Hong Kong University of Science and Technology
In Partial Fulfillment
of the Requirements for
the Degree of Master of Philosophy
in Computer Science

Hong Kong, December 2000

Copyright © by Yin Li 2000
Authorisation

I hereby declare that I am the sole author of the thesis.

I authorise The Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

[Signature]

I further authorise The Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

[Signature]
CHANNEL ALLOCATION WITH QOS GUARANTEE IN MOBILE WIRELESS NETWORKS: SINGLE AND MULTIPLE TYPES OF TRAFFIC

BY

YIN LI

APPROVED:

DR. BO LI, SUPERVISOR

PROF. ROLAND T. CHIN, HEAD OF DEPARTMENT

Department of Computer Science

27 December 2000
ACKNOWLEDGEMENTS

First, I would like to express my truly appreciation to my supervisor Dr. Bo Li for his unreserved guidance and help. With Dr. Li’s guidance, I learned wireless communication networks systematically, published my first paper and found a new research area in front of me. Without his encouragement and valuable suggestions, this thesis cannot be accomplished.

I am also grateful to Prof. Wong from the Department of Physics. It is my pleasure to work with him and his student Isaac. The valuable discussions with them provided me with a deeper understanding of my research and broader knowledge of mathematical methods.

Specifically, I would like to take this chance to express my sincere thanks to my thesis committee members: Dr. Manhoi Choy, Dr. Jelena Misic and Dr. Jogesh K. Muppala for their valuable suggestions in my proposal and thesis.

Also, I would like to take this opportunity to thank all from whom I have learned, both teachers and my classmates. Specific thanks go to my friends Chen Feng, Cheng Hui, Fang Huirong and Zhang Chuqian for their accompany and support.

Finally, I would like to take this chance to say thanks to my family for their unreserved encouragement and support. Their love is the most important drive that supported me in pursuing this degree.
# TABLE OF CONTENTS

Title Page                               i
Authorisation Page                      ii
Signature Page                          iii
Acknowledgements                        iv
Table of Contents                       v
List of Figures                         vii
Abstract                                ix

1 Introduction                          1
   1.1 Development of Mobile Communication  2
   1.2 Key Concepts
       1.2.1 Connection-level Quality of Service  5
       1.2.2 Channel Allocation Strategies  6
   1.3 Objective  7

2 Literature Survey                     9
   2.1 Guarded Channel Scheme and Its Variations  10
   2.2 Motion Prediction  11
   2.3 Distributed Call Admission Control  13
   2.4 Handoff Support for Multi-Traffic  14

3 An Efficient and Adaptive Call Admission Control Scheme Using An On-Line Local Estimation Technique  17
   3.1 System Model and Basic Assumption  18
   3.2 Control Algorithm  19
3.2.1 The Survival Probability for Uniform Handoff Rate 20
3.2.2 The Computation of Mean ($n_i(t)$) and Variance ($\sigma_i(t)^2$) of the Occupancy Distribution 23
3.2.3 The Derivation of the Time-dependent Dropping Probability 24
3.3 The On-line Estimation Algorithm 24
3.4 Simulation Results 28
  3.4.1 Comparison with SDCA 28
  3.4.2 Periodically Changing Traffic 30
  3.4.3 Fluctuating Traffic 31
  3.4.4 Impact of Estimation of the Coefficients $\alpha_i$ 33
  3.4.5 Estimated Handoff Rates 35
3.5 Conclusions 37

4 A Dual-Threshold Reservation Model for Voice/Data Integrated Mobile Wireless Networks 39
  4.1 DTR Scheme 40
  4.2 Analytical Model 41
    4.2.1 Transition Diagram and Balance Equation 41
    4.2.2 Calculation of State Probability $P_{ij}$ 43
  4.3 Numerical Results 46
    4.3.1 Increasing $K_1$ Under a Constant $K_2$ 46
    4.3.2 Increasing $K_2$ Under a Constant $K_1$ 48
    4.3.3 Increasing $K_1$ and $K_2$ Simultaneously 49
  4.4 Algorithms to Set $K_1$ and $K_2$ for Given Objective Functions 50
    4.4.1 Algorithm I 52
    4.4.2 Algorithm II 57
  4.5 Conclusions 62

5 Conclusions and Future Work 64

References 66
LIST OF FIGURES

1 Architecture of the Cellular Network 3
2 A 19-cell hexagonal cellular network with wrap-around connection 19
3 Flow chart of the derivation of the acceptance ratio 27
4 (a) A 19-cell hexagonal cellular network with wrap-around connection, (b) Topology of paths connecting $k = i$. 28
5 The comparison of call dropping probability with the SDCA 29
6 The comparison of channel utilization with the SDCA 29
7 The evolution of the traffic input scaled by the factor $P$ 30
8 The dynamic call dropping probability for each control period at $P = 2$ s$^{-1}$ 30
9 The traffic input with scaling factor $P$ increased from 0.4 s$^{-1}$ to 12 s$^{-1}$ 31
10 The call dropping probability vs. changing $P$ 32
11 The channel utilization vs. changing $P$ 32
12 The new call blocking probability vs. changing $P$ 32
13 The traffic input with periodic change 33
14 The call dropping probability vs. adjusted $\alpha$ values 33
15 The dynamic call dropping probability for periodic traffic input 34
16 The call dropping probability based on local estimation of the handoff rates 35
17 The call dropping probability with estimated handoff rates 35
18 The estimated handoff rate vs. time 36
19 The call dropping probability for periodic handoff rate 36
20 The architecture of cells 41
21 State transition diagram 42
22 Voice handoff dropping probability vs. K1 46
23 Voice blocking probability vs. K1 47
24 Data blocking probability vs. K1 47
25 Voice handoff dropping probability vs. K1 48
26 Voice blocking probability vs. K1 48
27 Data blocking probability vs. K1 49
28 Voice handoff dropping probability vs. K1 49
29 Voice blocking probability vs. K1 50
30 Data blocking probability vs. K1 50
31 Voice dropping probability under different $K_1$ and $K_2$ 51
32 Voice blocking probability under different $K_1$ and $K_2$ 51
33 Data blocking probability under different $K_1$ and $K_2$ 52
CHANNEL ALLOCATION WITH QOS GUARANTEE IN MOBILE WIRELESS NETWORKS: SINGLE AND MULTIPLE TYPES OF TRAFFIC

BY

YIN LI

A Thesis Presented to
The Hong Kong University of Science and Technology
in Partial Fulfillment
of the Requirements for
the Degree of Master of Philosophy

Hong Kong, December 2000

ABSTRACT

Mobile communication system has evolved remarkably since its first deployment. The tremendous growth in the mobile subscribers and the introduction of new services in the next generation of wireless networks have posed challenging problems in providing quality-of-service support for various of applications. In this thesis research, we first propose an efficient adaptive call admission control scheme which can guarantee the QoS requirement without any prior knowledge of traffic parameters and signaling overhead. The novelty of this scheme lies in its ability of dynamically obtaining traffic parameters and status information through local on-line estimation. Simulation results demonstrate the satisfactory system performance of this on-line estimation algorithm under different system setups. The simplicity and efficiency of this scheme makes it possible to be put into actual deployment.

Another scheme proposed is the dual-threshold reservation scheme (DTR) for integrated voice/data wireless networks. The next generation of mobile networks
is expected to carry multimedia traffic. Bandwidth allocation and management is of great importance due to spectrum limitation. DTR is proposed to provide effective bandwidth management while satisfying the QoS requirement in terms of voice handoff dropping probability, voice blocking probability and data loss probability. An analytical model based on this scheme is developed. In addition, for the DTR model, the selection of optimal thresholds is of great importance to maximize spectrum utilization while guaranteeing the QoS requirements. Based on the observation of the relationship between thresholds and system performance, we propose two algorithms to reduce the high computational complexity in finding the optimal thresholds.
CHAPTER 1

INTRODUCTION

The public mobile communication systems were first deployed shortly after the World War II in the United States. After that, the mobile communication systems have evolved remarkably from the first generation of networks to the second generation and now they are heading for the third generation of networks. The tremendous growth in the number of subscribers, the enhanced handoffs and the introduction of new services such as multimedia and data delivery have posed challenging problems in providing the quality-of-service (QoS) for a variety of applications.

One of the key generic QoS parameters is the call dropping probability, which has to be maintained at a predefined level independent of the traffic condition. In the presence of bursty data and the emerging multimedia traffic, an adaptive and dynamic bandwidth allocation is essential in ensuring this QoS. The paradox, however, is that all existing dynamic bandwidth allocation schemes, require the prior knowledge of all traffic parameters and/or user mobility parameters [5], [10], [20], [25]. In addition, most proposed schemes require extensive status information exchange among cells in order to dynamically re-adjust the control parameters, thus making them difficult to be used in actual deployment. This thesis studies a novel adaptive bandwidth allocation scheme which can dynamically obtain the changing traffic parameters through local on-line estimation. Simulations are performed to investigate the system performance of the call admission control scheme.

Another posed challenge in the QoS support is caused by the existence of multiple traffic with diverse bandwidth and QoS requirement in the next generation of wireless networks. Various issues related to this carrying multimedia
traffic have to be carefully examined before such systems can be realized [19]. Given the wireless spectrum remains primary limited, therefore, it is necessary to develop a mechanism that can provide effective bandwidth management while satisfying the QoS requirement for both traffic. This thesis proposes a scheme for the investigation of bandwidth allocation for a voice/data integrated mobile wireless network. Based on the scheme, an analytical model is developed and two algorithms are proposed.

This thesis is organized as follows: In the following part of this chapter, fundamental background knowledge and key concepts are introduced, then the objective of this thesis is addressed. In chapter 2, a literature survey of related work is presented. In chapter 3, an efficient and adaptive bandwidth allocation scheme based on on-line local parameter estimation is proposed and simulation results are discussed. The dual-threshold reservation scheme for integrated voice/data wireless network is proposed in chapter 4. Finally, conclusions and future works are given in chapter 5.

1.1 Development of Mobile Communication

Before the cellular concept was developed, early mobile communication systems used a single, high-powered transmitter and large tower to cover a large area in the market. This kind of system can only serve very few users because of resource limitation. The early FM (frequency modulation) push-to-talk telephone systems use 120kHz of RF bandwidth in a half-duplex mode to transmit only 3kHz of telephone-grade speech. The bandwidth requirement was decreased to 60kHz in 1950s, then to 30kHz in 1960s as a result of improvement in technology. In the 1950s and 1960s, automatic channel trunking was developed to increase spectrum efficiency. The trunking technique, instead of assigning each user a dedicated channel, allows a large number of users to share a number of channels in a channel pool. By adopting the trunking technique, the mobile communication system can accommodate more users. However, these improvements in spectrum efficiency was too limited to satisfy the growing demand of the market. By 1976, New York
City could serve only 543 paying customers while with a waiting list of over 3,700 people [1].

In order to further increase spectrum efficiency, the cellular concept was developed by Bell Laboratories in the 1960s and 1970s. With the development of highly reliable, miniature, solid-state frequency hardware in the 1970s, it is possible to implement the cellular mobile systems. The basic idea of cellular networks is to replace the single, high power transmitter with many low power transmitters, in another word, the cellular networks breaks the coverage zone into small areas, known as cells. Each cell is assigned a portion of channels available to the entire systems and neighboring cells are assigned different groups of channels to minimize interference.

![Diagram of Cellular Network](image)

**Figure 1: Architecture of the Cellular Network**

Fig. 1 shows a basic cellular system architecture which consists of mobile stations, base stations and a mobile telephone switching center (MSC). Each mobile communicates via radio with one of the base stations and may be handed-off to other base stations during its lifetime. The base station serves as a bridge between all mobile users in the cell and connects the mobile calls to the MSC. The MSC is responsible for coordinating the activities of all of the base stations and
connecting the entire cellular system to the Public Switched Telephone Network (PSTN). In the cellular networks, channel can be reused as many times as necessary as long as co-channel base stations (base stations using the same set of channels) are separated far enough to keep the interference acceptably low. By increasing the number of base stations, cellular networks can provide additional radio capacity with no additional requirement in radio spectrum. The cellular concept enables a fixed number of channels to serve an arbitrarily large number of subscribers by reusing the channels throughout the coverage area.

In 1979, the world’s first cellular system was deployed in Japan, by the Nippon Telephone and Telegraph company. After that, cellular systems sprouted up quickly throughout the world in 1980s. In the late 1991, the first U.S. Digital Cellular (USDC) system hardware was installed in major U.S cities. By adopting the USDC standard, the single-user analog channels is replaced by digital channels which can accommodate three users in the same 30kHz spectrum.

However, these cellular systems are generally incompatible with each other due to the adoption of different frequency and communication protocols. To overcome the barriers between incompatible system, the Pan European Digital Cellular standard Global System for Mobile (GSM) was first deployed in 1991. GSM is the first universal digital cellular system with modern network features extended to each mobile user and is gaining worldwide acceptance.

In 1989, the concept of Personal Communication Services (PCS) was first originated in the United Kingdom. PCS incorporates more network features and is more personalized than existing cellular radio systems. The third generation of PCS aims to create a seamless network infrastructure to provide not only mobile communication services, but also broadband and multimedia services, with quality comparable to the fixed networks and enable mobile users to exchange all kinds of information at any time, anywhere.
1.2 Key Concepts

1.2.1 Connection-level Quality of Service

When a mobile subscriber requests a new call, the call attempt may be rejected because of the lack of available channels, such event is called a "block". Otherwise, the call will be set up successfully. After a call enters the system, due to its roaming tendency, it may move from one cell to another before it is terminated. This operation is called handoff. This handoff operation not only involves identifying a new base station, but also requires bandwidth (channel) switching. If there is no channel available in the new base station, the call is forced to be terminated (dropped).

The probability of dropping an on-going call due to a handoff failure (handoff dropping probability) or the probability of blocking a new call due to the lack of available channels are often used as the criteria for performance evaluation of mobile systems. In addition, quality of service (QoS) in a wireless cellular network is often characterized in terms of call blocking probability (CBP) and call dropping probability (CDP). In order to differentiate the QoS referred to a mobile cellular network and the QoS terminology used in ATM networks (which focus more on parameters such as transmission rate and call loss rate), many researchers use the term "connection-level QoS" to evaluate performance in cellular systems.

With the development of wireless communication networks, the support of QoS is facing new challenges, mainly in the following two aspects:

1. With the exponential growth in the number of mobile subscribers and the introduction of new services, the cell size tends to be smaller and smaller to provide more capacity. As a result, the rate of handoffs will increase tremendously because even when the users are moving at a very low speed (walking) the probability to cross the cell boundary during the lifetime of a typical call will be higher. Another effect of having smaller cells is the rapid changes in the network traffic conditions. This smaller cell tendency poses new challenges in providing QoS support.
2. Another challenge comes from the emergence of new services in mobile networks. Recently, there has been increasing concerns in supporting multimedia traffic using allocated radio spectrum. Therefore, the support of heterogeneous traffic and resource sharing among different service classes need to be addressed in the design of QoS support.

1.2.2 Channel Allocation Strategies

Channel allocation scheme is a frequency reuse scheme, whose objective is to increase capacity and minimize interference in order to utilize the limited radio spectrum efficiently. Channel allocation schemes can be classified into three categories, namely, fixed, dynamic and hybrid. Channel allocation strategies have important impact on system performance [26].

In a fixed channel assignment scheme (FCA), each cell is permanently allocated a set of nominal channels for its exclusive use. Any call attempt within the cell can only be served by the unused channels associated with that particular cell. If all channels in that cell are occupied, the call has to be blocked and the subscriber receives no service. One variation of the fixed channel allocation scheme is the channel borrowing scheme, in which a cell is allowed to borrow channels from a neighboring cell if no channel is available in itself under the constraint that the borrowing of a channel does not disrupt or interfere with any of the calls in progress in the donor cell. The fixed channel allocation strategy has been widely used by cellular service providers in their frequency plan because of its simplicity.

The dynamic channel assignment strategy (DCA) is proposed recently. With DCA, channels are not allocated to any cells permanently. Instead, each channel is accessible to a cluster of cells. Every time a call request is made, the MSC allocates a channel to the requested cell following some algorithms. The allocation of a channel takes into account the likelihood of future blocking within the cell, the frequency of the candidate channel, the reuse distance of the channel, and other cost functions. Apparently, DCA can achieve better channel utilization at
the expense of the increase of system storage and computation load.

Based on these two strategies, a mixture strategy of FCA and DCA called \textit{hybrid channel assignment strategy (HCA)}, is proposed. In HCA, a portion of channels are permanently assigned to each cell as in FCA and at the same time, some channels are reserved to be accessible by a cluster of cells as in DCA. The ratio of fixed to dynamic channels has a significant impact on the system performance.

The schemes proposed in this thesis are applicable to a fixed channel assignment scenario where the capacity of a deterministic system is predetermined.

1.3 Objective

The objective of this thesis is to study the call admission control (CAC) scheme for single and multiple traffic. The CAC algorithm is expected to fulfill the following requirements:

- Admit calls with QoS guarantee on call dropping probability independent of changing traffic. It is the service providers’ responsibility to provide satisfactory QoS against call dropping to subscribers.

- Adaptive to changing traffic. The CAC algorithm should be able to guarantee the QoS requirement in the presence of bursty data and multimedia traffic.

- Efficient channel utilization. The CAC algorithm must try to maximize spectrum efficiency while maintaining the target handoff dropping probability.

- Simplicity for implementation including the computation and communication complexity. A good CAC scheme must be possible for practical implementation, since handoff support is primarily an engineering problem arising from practical systems.
• Serve multiple traffic with diverse QoS support according to their different QoS requirements. This requirement is presented specifically for the CAC scheme for multiple traffic.
CHAPTER 2

LITERATURE SURVEY

In chapter 1, the basic channel allocation schemes are discussed. However, the channel allocation schemes did not take into account the effect of handoffs in the performance of the system. However, when the cell size is relatively small, the handoff procedure has an important effect on system performance, therefore, the guarantee of handoff probability under some predefined level is of great importance.

Different systems have different policies and methods for managing handoff requests. Some systems treat handoff requests in the same way that handle new call. In such systems, the handoff dropping probability and blocking probability are roughly the same. However, from the point of view of a mobile user, a forced termination of an on-going call is less desirable than blocking a new call. To improve the quality of service as perceived by the users, an intuitive way is to reduce the chances of unsuccessful handoffs by prioritizing handoff request over new call request when allocating channels.

Handoff prioritizing schemes provide improved service at the expense of a reduction in the total channel utilization and an increase in the blocking probability. Recently, various of wireless call admission control schemes (CAC) are proposed to limit the handoff dropping probability at a predefined level.

CAC is the process of making an admission decision on a new call request. One of the key design goal is to guarantee the QoS in terms of call dropping probability, which, however, usually comes at the expense of potentially higher blocking probability, therefore results in a poor channel utilization. Given that the radio channels are considered to be primary scarce resource in mobile radio networks, the main challenge in the design of an efficient call admission control
scheme is to achieve better channel utilization while maintaining the handoff dropping probability at a predefined level. In the following sections, a literature survey of proposed CAC schemes are presented.

2.1 Guarded Channel Scheme and Its Variations

The so-called guarded channel concept has long been used in telecommunication systems. In the mid-'80s, it was introduced for mobile systems. The guarded channel policy (GC), also referred as trunk reservation policy, offers a generic means of improving the handoff success probability by simply reserving a number of channels exclusively for handoff. The remaining channels can be shared equally between handoffs and new calls. The number of reserved guard channels is determined according to the long term call arrival rate and handoff arrival rate. Both locally originated calls and handoff calls from neighboring cells are assumed to follow a Poisson process. The call duration time and channel holding time within a cell are also assumed to be exponentially distributed. Similar assumptions are made in other schemes likewise. Guarded channel schemes can offer lower handoff dropping probability at the expense of higher call blocking probability.

There are many variations of GC are proposed. One approach, called fractional guarded channel (FGC) policy, is proposed. Instead of rejecting new call request once the channel occupancy is over the given threshold, the FGC admits new calls with a certain probability which depends on the current channel occupancy. The FGC scheme has been proved to be optimal for the objective to minimize the call blocking probability subject to a hard constraint on the handoff dropping probability [31].

To further reduce the handoff dropping probability, a modification of GC was proposed in [8], in which handoff request is queued instead of dropped immediately when no available channels. The queueing of handoff calls is made possible by the existence of the time interval that the mobile is physically communicable with both current and next base station. The basic queueing discipline is first-
in-first-out (FIFO). To improve the performance of the handoff queuing scheme, in [35], a non-preemptive queueing discipline based on a mobile's subscriber measurement was adopted for queueing handoff calls. The handoff queueing is dynamically re-ordered according to the most recently submitted power level of mobile terminals. This scheme is shown to offer lower handoff dropping probability and better spectrum efficiency. However, this handoff queueing may not suitable in a micro/pico environment due to high handoff rate and the short handoff interval in it.

Another modification of GC is new call queueing which is adopted to release the reduction in the total carried traffic [33]. This new call queueing is more feasible than handoff request queueing due to the delay insensitivity of new calls. It is shown that the handoff dropping decreases much faster than the queueing probability of new calls increases. In addition, there is an increase in the carried traffic since less handoff call are dropped and queued new call will get service eventually.

Generally, GC schemes is simple and easy to put into practical use. However, it can not adapt to the network traffic changing due to its "static" nature. Also, the performance of GC is highly related to the accuracy of traffic parameters and channel occupancy distribution. If the channels are dimensioned based on wrong traffic parameters, GC will suffer high call blocking probability and inefficient spectrum utilization due to resource over-provision.

2.2 Motion Prediction

This category of CAC scheme make admission decision by predicting the motion pattern of mobile users.

The shadow cluster CAC proposed by Levine et al [10] is one good example of this category. The shadow cluster mechanism is 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 the travel". The coverage of a shadow cluster for a given
active mobile mainly consists of the cell where the mobile is currently present and all its adjacent cells along the direction of the travel. This area changes when the mobile call is handed off to other cells, thus a tentative shadow cluster needs to be implemented for every new call as well as every handoff call.

The call admission algorithm requires every base station exchange information with other cells in its shadow cluster periodically. With this information, base stations make prediction on future motion and channel reservation accordingly, and admit those mobile terminals with adequate support. Therefore, the decision process involves all base stations in the mobile terminal’s tentative shadow cluster. Generally, when a new call arrives, the original base station predicts the probability that the mobile terminal will visit cells within its shadow cluster in consecutive time periods. By exchanging these probability information, the survivability of the call is evaluated. Based on this evaluation, the admission decision is made. When a call handoff to its neighboring cells, the tentative shadow cluster changes accordingly. The new base stations will take the responsibility to maintain the new shadow cluster for the mobile terminal.

Simulations show that the shadow cluster mechanism is able to reduce the dropping probability in a controlled fashion. The efficiency of this scheme, however, depends on the accuracy of prediction of the future motion pattern, which makes it most suitable for strong directional environment such as highway. One practical drawback of this scheme is the potentially high computation and communication complexities in re-calcultating the shadow cluster for each new calls as well as each handoff.

Instead of predicting the mobile movement periodically, to reduce the computational complexity, an event-triggered adaptive control scheme was proposed by Mišić et al. in [20]. In this scheme, The probability of a call originated from a given cell will visit a given target cell within its lifetime is defined as the spatial activity factor of a call. The spatial activity factor of a call depends on the relative distance of target cell from its originating cell, the call duration time and call dwelling time, and the directional characteristics of the terrain. Instead of estimating the resource every regular time intervals, the resource estimations are
triggered by call handoff, originating and termination here.

In summary, this kind of schemes requires cooperation among base stations to exchange information which is a high computation and communication load to the system. More importantly, the performance heavily relies on the accuracy of motion prediction.

2.3 Distributed Call Admission Control

One deficiency of GC and its variation is its “blind” channel reservation for handoff without taking into account the neighboring environment. Unlike in macro cellular network, in micro/pico network with smaller cell size and higher handoff rate, the availability of channel in the local cell does not necessarily guarantee its survivability in all cells it visits during its life time. Therefore, it will be helpful to consider both local and neighboring cell information when making admission decision. A distributed CAC (DCAC) scheme proposed in [25] is one good example based on this principle.

In DCAC, every base station makes an admission decision by exchanging state information with adjacent cells periodically. The basic constraint of DCAC is that the admitting of a new call will not violate the desired QoS of any existing calls in the system and the new call can be guaranteed the desired QoS.

In DCAC, the base stations exchange the number of active calls every $T$ seconds, where $T$ is predetermined. Based on this information, each base station estimate the total future resource usage after time $T$. The future resource usage is composed of three parts, namely, the number of calls remaining in the cell, new calls generated, handoff calls arrived. In the period of $T$, the resource usage is approximated by a Gaussian distribution. Based on this estimation of future resource usage, together with the conditions “not to overload the system” in both itself and its adjacent cells, the base station derives the admission threshold for new calls.

One of the main features of DCAC is its simplicity in that the admission decision can be made in real time and does not require much computational effort.
In addition, the DCAC mechanism does not require the directional environment that shadow cluster needs which makes it more suitable for urban areas. However, the DCAC requires a frequent information exchange among base stations which increases the system signaling load. Also, the choice of control period $T$ remains to be solved. A long control period $T$ may result in an inaccurate estimation and resource over-provision while a small $T$ will increase the system load due to more frequent information exchange.

Another dynamic call admission control algorithm based on stochastic control, called stable dynamic call admission control (SDCA), was recently proposed to precisely guarantee the target handoff dropping probability while at the same time maximizing the channel utilization [30]. The precision of SDCA is obtained by taking into consideration the time-dependent dropping probability, the finite capacity and the influences from non-neighboring cells. Also, in SDCA, an acceptance ratio $\alpha_i$ is derived based on the QoS requirement on dropping probability, but different from DCA, where the acceptance ratio is used to determine an admission policy, in SDCA, the new calls are uniformly spreaded over the whole control period and the admission of new calls is stochastically determined based on this $\alpha_i$. This stochastic admission policy is shown to result in a more effective and stable control. Simulations show SDCA can steadily satisfy the hard constraint on call dropping probability while maintaining an efficiency utilization of spectrum. However, SDCA also requires periodic status information exchange among different cells.

2.4 Handoff Support for Multi-Traffic

The development of multimedia applications in the next generation of mobile networks necessitate the support of different classes of traffic with diverse QoS requirements which need to be guaranteed by the transmission network. Various issues related to this future wireless mobile networks carrying multimedia traffic have to be carefully examined before such systems can be realized [19]. An appropriately designed channel allocation strategy is one of the challenges.
In [32], based on the multi-dimensional birth-death processes, a framework that is useful for teletraffic performance analysis and modeling for a broad class of problems arisen in cellular communication networks has been proposed. In [3] and [4], the resource allocation for two classes of traffic were discussed. Three types of resource sharing policies, namely, completely sharing (CS), complete partitioning (CP), and hybrid allocation were investigated.

In [24], the resource management in integrated wireless networks with real-time and non-realtime traffic are decomposed into two independent sub-problems: (1) a class-based call admission control. (2) wireless bandwidth allocation to each admitted connection according to some resource sharing discipline based on its QoS requirement and QoS requirement of other connections sharing the same base station. Particularly, in order to adopt to changing traffic and to efficiently utilize the scarce wireless resource, the admission threshold for different classes of traffic is recalculated periodically. In addition, the resource sharing algorithm is performed in a decentralized fashion and ensures that base station capacity is efficiently shared among different classes of connections and reacts to the frequent change of the number of mobile terminals within each radio cell. Analysis results show that the combination of the call admission control and the resource sharing schemes guarantee a predefined QoS for each class of traffic. One major advantage of this approach is its distributed fashion. The limitation of this approach is the non-prioritized handoff where the channel assignment to the new and the handoff request are not distinguishable.

In [14], Li et. all proposed a generic model, called hybrid cutoff priority scheme which can handle multiple streams of traffic, each having potentially different QoS requirements, thus requiring potentially different reservation thresholds. In the scheme, each type of traffic can be assigned a different cutoff threshold in which higher priority is given to the handoffs while lower priority new arrival calls will be served only if enough channels are available. The performance analysis was carried out using Stochastic Petri Net (SPN) along with decomposition and iteration techniques. New call and handoff call blocking probabilities, channel utilization and call forced termination probability for each type of traffic are
calculated. The choice of cutoff priority threshold of each traffic, however, remains a problem to be solved.

Recently, an analytical model was proposed to investigate the performance of an integrated voice/data mobile network with finite data buffer in terms of voice-call blocking probability, data loss probability and mean data delay [36]. The model is based on the movable-boundary schemes which dynamically adjusts the number of channels for voice and data traffic. The model is suitable to be used to analyzed the impact of hot-spot traffic in the heterogeneous PCS networks, in which the traffic parameters can be varied. In addition, based on this proposed model, an iterative algorithm is proposed to determine the handoff traffic within polynomial-bounded time. The limitation of this model is also the non-prioritized handoff.
CHAPTER 3

AN EFFICIENT AND ADAPTIVE CALL ADMISSION CONTROL SCHEME USING AN ON-LINE LOCAL ESTIMATION TECHNIQUE

In the next generation of wireless networks, there are a number of unique aspects that an effective admission control scheme needs to take into consideration:

- The employment of smaller cells (micro/pico-cells) will increase the number of handoffs during a call’s lifetime. In addition, there is also an increased influence from neighboring and even next neighboring cells.

- Multiple QoS requirement for multi-class of calls and potentially more stringent QoS requirements of individual calls mandate a highly precise resource allocation.

- Existence of diversified traffic load requires an adaptive and dynamic call admission control to ensure this QoS requirement.

Lots of call admission control scheme are proposed to efficiently utilized the channel while guaranteeing the QoS requirement in the presence of bursty data and changing traffic. The paradox is, however, from previous discussion, all existing dynamic call admission control schemes require the prior knowledge of all traffic parameters or/and mobility parameters. Under such conditions, the fixed CAC strategy based on guarded channel scheme can yield optimal solutions for steady state [31], but can not adapt to changing traffic conditions. In addition, most proposed dynamic allocation schemes requires extensive status information exchange among cells in order to dynamically re-adjust the control parameters, thus making them difficult to be put into practical deployment.
Our major motivation is to design an adaptive and dynamic call admission control scheme that can overcome these two deficiencies. The main advantages of the proposed algorithm are:

- Simplify the implementation by eliminating the signaling overhead. The status information exchange among cells is made unnecessary by local estimation on all parameters required by the control algorithm.

- Effectively adaptive to the changing traffic condition, which is achieved by doing on-line and periodically estimation. This effective adaptability makes our algorithm particularly suitable for the emerging multimedia traffic in which the statistical behavior of the traffic is either not available or is difficult to obtain.

- Efficient statistical multiplexing bandwidth allocation. The bandwidth allocation mechanism based on a probabilistic model is first introduced in SDCA [30], which can reserve the bandwidth in a more efficiently statistical multiplexing manner. This eliminates the needs to channel reservation for each call set up. In addition, this can spread new call evenly over a control period, thus leading to a more effective and stable control.

The following part of this chapter is organized as follows: In the first part, the system model and basic assumption is introduced. Then the control algorithm adopted is presented. After that, the on-line estimation algorithm is proposed. In section 4, simulation results are discussed. Finally, a conclusion is given.

### 3.1 System Model and Basic Assumption

We consider a cellular network consisting of closed packed hexagonal cells (shown in Fig. 2) and using a fixed channel allocation scheme. Each cell has a capacity of $N$ channels. New calls arrive in cell $i$ at a rate of $\lambda_i$. The channel holding time of each call is exponentially distributed with a mean of $\mu$. In addition, calls will handoff from cell $k$ to cell $i$ with a rate of $h_{ik}$. And $h_k = \sum_i h_{ik}$ be the handoff rate out of cell $k$. 
Figure 2: A 19-cell hexagonal cellular network with wrap-around connection

3.2 Control Algorithm

We adopt the control algorithm in SDCA because of its precision. Apparently, the idea of on-line estimation can work on other control algorithm which needs traffic parameter input.

The control algorithm is executed in a distributed and periodic manner [5]. Each cell executes the identical algorithm based on the local estimation. The length of the control period is given as $T$. At the beginning of each control period, the scheme determines the amount of channels reserved in the next period for the particular cell by taking into consideration of the network traffic condition. Notice that this reservation consider all potential handoff in the coming control period, therefore, for each handoff call request, no need to do any channel reservation. Before go into details of the control algorithm, the key features of it is summarized as follows:

- The algorithm guarantee the QoS requirement in terms of handoff dropping probability, denoted by $P_{QoS}$. For each acceptance ratio $a_i$, which is defined as the maximum fraction of new traffic can be admitted into cell $i$ in next control period, an expression of the dropping probability is calculated. Based on this calculation, the maximum $a_i$ which can satisfy the hard constraint on the handoff dropping probability, $P_{QoS}$, is derived. Instead of using this obtained ratio $a_i$ to determine the admission threshold as in GC, we stochastically admit the new call requests uniformly over the whole con-
trol period. This stochastical call admission avoids a sudden overload at the beginning of the control period due to network congestion, leading to a more effective and stable control.

- The handoff dropping probability is derived by taking into consideration both the time-dependency and the finite capacity. The derivation is based on solution to the evolution equation of the occupancy distribution. It greatly improves over the Gaussian approximation commonly used [18], [25]. Then an average dropping probability $\bar{D}_i$ over a control period is calculated. This calculation increases the precision over a single-value approximation within the control period.

The key to the control algorithm is to derive the acceptance ratio $a_i$ for cell $i$ periodically. To derive this $a_i$, first, the mean $(n_i(t))$ and variance $(\sigma_i(t)^2)$ of the time-dependent occupancy distribution in each cell is calculated based on the calculation of the inter-cellular transition probability $f_{ik}(t)$. Next a diffusion equation whose solution describes the evolution of the occupancy distribution is derived, then based on these two result, the handoff dropping probability is obtained by a mean-rate approximation. Major steps of the computation is described in following sections:

### 3.2.1 The Survival Probability for Uniform Handoff Rate

In this subsection, we present the derivation for the survival probability for uniform handoff rate $(h_{ik} = h/6)$. Specifically, we use the estimated value of $h = h^{(e)}$ obtained at the beginning of each control period.

Consider the single-call transition probability $f_{ik}(t)$ that an ongoing call in cell $k$ at the beginning of the control period ($t = 0$) is located in cell $i$ at time $t$. In particular, $f_{ii}(t)$ is the survival probability of a call in cell $i$ at time $t$. For an effective control enforcing dropping probabilities of the order $10^{-3}$ to $10^{-2}$, we assume that essentially all calls hand off successfully, resulting in the evolution
equation

\[
\frac{df_{ik}(t)}{dt} = -\sum_{j} J_{ij} f_{jk}(t) \quad \text{and} \quad f_{ik}(0) = \delta_{ik},
\]  

(3.1)

where \(J_{ik}\) is the transition matrix given by \(J_{ii} = h_i + \mu\) and \(J_{ik} = -h_{ik}\) for \(i \neq k\). The solution to (3.1) is

\[
f_{ik}(t) = [\exp(-Jt)]_{ik}.
\]  

(3.2)

The computational complexity of this matrix operation can be reduced by considering the off-diagonal terms as perturbations to the diagonal part of \(J\). Each term in the resultant perturbation series of \(f_{ik}(t)\) corresponds to the contribution of a path connecting \(k\) and \(i\) by cell hopping. While the perturbation technique is applicable to non-uniform handoff rates in general [30], results for the case of uniform rates is particularly illustrating. Notice that the matrix \(J\) can be written as

\[
J = J_0 + J_1,
\]  

(3.3)

where

\[
(J_0)_{ik} = (h_k + \mu)\delta_{ik},
\]

(3.4)

\[
(J_1)_{ik} = \begin{cases} -h_{ik} & k, i = \text{nearest neighbors}, \\ 0 & \text{otherwise}. \end{cases}
\]

For homogeneous handoff rates, \(h_k = h\) and \(h_{ik} = h/6\). Considering \(J_1\) as the perturbation, we have

\[
[\exp(-Jt)]_{ik} = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} \left( J_0^{-1} J_1 + \cdots + J_1 J_0^{-1} \right) \cdots \right)_{ik}. 
\]  

(3.5)

The zeroth order term consists of those terms in (3.5) which contain no \(J_1\). Hence the zeroth order contribution goes only to \(i = k\), with

\[
q_0(t) = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} (J_0^{-1})_{ii} = \exp[-(h + \mu)t].
\]  

(3.6)

It corresponds to the case that no handoff events take place between time 0 and \(t\).
The first order terms consist of one and only one \( J_1 \). Since the elements of \( J_1 \) are nonzero only for neighboring cells, the first order contributions go only to neighboring cells \( i \) and \( k \), with

\[
q_1(t) = \sum_{r=0}^{\infty} \frac{(-ht)^r}{r!} [(h + \mu)^{r-1} \frac{h}{6} + \cdots + \frac{h}{6} (h + \mu)^{r-1}] 
\]

It corresponds to the event that a call is handed off from cell \( k \) to \( i \).

Higher order contributions can be evaluated similarly. For a path with \( n \) hops along the path from \( k \) to \( i \), we obtain

\[
q_n(t) = \frac{1}{n!} \left( \frac{ht}{6} \right)^n \exp[-(h + \mu)t].
\]

This equation can be intuitively interpreted by counting the number of handoff events along the path. Since all hopping and termination events are assumed to take place independently at the same rate, the occurrence of \( n \) such events in time \( t \) obeys a Poisson distribution with mean \((h + \mu)t\). A path with \( n \) hops requires each of the \( n \) events to be a handoff to a specified neighboring cell, excluding the other five neighbors and the termination event. Hence the probability \( q_n(t) \) is given by \([h/6(h+\mu)]^n p_n\), where the Poisson distribution \( p_n = [(n+\mu)t]^n \exp[-(h+\mu)t]/n!\), resulting in Eq. (3.8).

Hence \( f_{ik}(t) \) is obtained by summing over all possible paths between \( k \) and \( i \). For the cellular network in Fig. 4(a), Fig. 4(b) shows the example of \( k = i \), in which each diagram represents the topology of a path connecting \( i \) to itself, with vertices and edges representing cells and paths respectively. Hence there are one path of 0 hop, no path of 1 hop, six paths of 2 hops and twelve paths of 3 hops, leading to

\[
f_{ii}(t) = q_0(t) + 6q_2(t) + 12q_3(t) + \cdots.
\]

Since \( ht \) is the average number of hops in time \( t \), the resultant perturbation series is rapidly converging for \( ht \) up to \( O(1) \). For a handoff rate \( h \) as high as 0.05 s\(^{-1}\) and \( t = 20 \) s, \( \mu = 0.005 \) s\(^{-1}\), the computed values for \( f_{ii}(t) \) are lower than the true values by 1% up to 2 hops, and 0.3% up to 3 hops.
The survival probability averaged over time $t$ is given by

$$s_i(t) = \frac{1}{t} \int_0^t dt' f_{ii}(t - t'),$$

(3.10)

whose closed form can be easily obtained using integration by parts.

### 3.2.2 The Computation of Mean ($\langle n_i(t) \rangle$) and Variance ($\sigma_i(t)^2$) of the Occupancy Distribution

The channel occupancy distribution is determined superposition of three parts, namely, the new calls admitted to the cell, the handoffs (in and out), and the terminations of existing calls. Specifically, the mean of the occupancy distribution $p_{n_i}(t)$ in cell $i$ at time $t$ is given by

$$\langle n_i(t) \rangle = \sum_k f_{ik}(t) n_{k0} + \sum_k g_{ik}(t) a_k \lambda_k$$

(3.11)

and the variance is

$$\sigma_i(t)^2 = \sum_k f_{ik}(t) [1 - f_{ik}(t)] n_{k0} + \sum_k g_{ik}(t) a_k \lambda_k$$

(3.12)

Where the $f_{ik}(t)$ is the transition probability that an ongoing call in cell $k$ at the beginning of the control period ($t=0$) is located in cell $i$ at time $t$. $g_{ik}(t)$ is the integrated transition probability, which captures the probability that a new call arrives in cell $k$ at time $t'$ ($0 < t' < t$) and is located in cell $i$ at time $t$. According to our control algorithm, such calls are uniformly spreaded over the control period, therefore, the number of new calls in cell $i$ at time $t$ follows a Poisson distribution with mean $\sum_k g_{ik}(t) a_k \lambda_k$, and $g_{ik}(t)$ is obtained by

$$g_{ik}(t) = \int_0^t dt' f_{ik}(t - t')$$

(3.13)

The computation of $f_{ik}(t)$ is done off-line and the detailed derivation can be found in the Appendices A and B in [30].
3.2.3 The Derivation of the Time-dependent Dropping Probability

The time-dependent call dropping probability $D_i(t)$ for cell $i$ can be expressed in terms of the quantities $n_{i0}$ (channel occupancy status at the beginning of a control period), $\langle n_i(t) \rangle$ (the mean of the channel occupancy distribution) and $\sigma_i(t)$ (the variance of the channel occupancy distribution).

$$D_i(t) = 2 \frac{\exp[-\xi_i(t)^2/2]}{\sqrt{2\pi\sigma_i(t)^2}} + 2 \frac{\langle n_i(t) \rangle - n_{i0}}{\sigma_i(t)^2} H(\xi_i(t)), \quad (3.14)$$

where $\xi_i(t) \equiv (N - \langle n_i(t) \rangle)/\sigma_i(t)$ is the normalized vacancy in cell $i$ at time $t$, with $N$ being the capacity of cell $i$, and $H(x)$ is related to the complementary error function via $H(x) = \text{erfc}(x/\sqrt{2})/2$ [27]. The $\langle n_i(t) \rangle$ and $\sigma_i(t)^2$ are the mean and variance of the channel occupancy distribution.

The average dropping probability over a control period is obtained by

$$\bar{D}_i = \frac{1}{T} \int_0^T dt \ D_i(t). \quad (3.15)$$

For an on-line periodic control, the complexity of the integration could be very high. However, since our control is based on a probabilistic model, the precision for integration needs not be high. We found that it is sufficient to use a 7-point Simpson rule [27]. The acceptance ratio $a_i$ can then be easily obtained by solving numerically

$$\bar{D}_i = P_{QoS}. \quad (3.16)$$

At low traffic, it may happen that $\bar{D}_i < P_{QoS}$ even for $a_i = 1$. Then $a_i$ is set to 1. Similarly, at high traffic, $a_i$ is set to 0 if $\bar{D}_i > P_{QoS}$ even for $a_i = 0$. The maximum $a_i$ can be derived by changing the value of $a_i$ bi-sectionally and calculating the corresponding handoff dropping probability for the $a_i$.

3.3 The On-line Estimation Algorithm

Like some of other dynamic call admission control scheme, SDCA requires a significant amount of signaling to obtain the status information required by the
control algorithm, such as the channel occupancy $n_{i0}$, $\lambda_i$, and $a_i$ from its neighboring and even next neighboring cells. One potential solution for this heavy system load in communication is to enlarge the control period $T$ to reduce the signaling frequency. However, this may result in the inaccuracy in the estimation of the future occupancy distribution because the statistical uncertainty grow with time. More seriously, if the traffic condition changes very fast in any cell, the control algorithm will not be able to adapt its control parameter in time and result in a wrong estimation. These two factors might hinder the deployment of such algorithms in actual implementation.

To overcome this limitation, we develop an estimation-based version of our control algorithm, in which both arrival and new traffic rates are obtained using strictly local on-line estimations. A similar technique was adopted in TCP adaptive retransmission to estimate the round-trip time (RTT) [12].

Concretely, let $\lambda_i^{(o)}(j)$ be the observed arrival rate in cell $i$ for the $j^{th}$ control period. This value is needed and is available at the beginning of the $(j+1)^{th}$ control period (i.e., the end of $j^{th}$ control period). Let $\lambda_i^{(e)}(j)$ to be the estimated arrivals for the $j^{th}$ control period (at the beginning of the $j^{th}$ control period). Using exponential smoothing, we have

$$\lambda_i^{(e)}(j+1) = \alpha_1 \lambda_i^{(e)}(j) + (1 - \alpha_1) \lambda_i^{(o)}(j),$$  \hspace{1cm} (3.17)

Under the uniform handoff rate case, the handoff rate $h$ can also be obtained similarly:

$$h_i^{(e)}(j+1) = \alpha_2 h_i^{(e)}(j) + (1 - \alpha_2) h_i^{(o)}(j),$$  \hspace{1cm} (3.18)

For the observed channel occupancy $n_{i0}(j+1)$ at the beginning of the $(j+1)^{th}$ control period, we note that it consists of two components. First, the controllable component consists of the channels occupied by calls admitted to cell $i$ during the $j^{th}$ control period. It is directly controlled by the admission actions of the local cell, and is approximated by $a_i(j) \lambda_i^{(o)}(j)T$. Secondly, the background component consists of the channels occupied by all the ongoing calls which were admitted in the previous control periods, and handoff calls which entered the cell during
the $j^{th}$ period or beforehand. The background occupancies cannot be controlled directly by the actions of the local cell, but it is expected to exhibit some long term statistical behavior given the traffic does not change too rapidly. Letting $N^{(o)}(j)$ be the observed background channel occupancy at the end of the $j^{th}$ control period, we have

$$N^{(o)}(j) = n_{i0}(j + 1) - s_i(T)a_i(j)\lambda_i^{(o)}(j)T,$$  

(3.19)

where $s_i(t)$ is the survival probability of a call in cell $i$ averaged over a time interval $t$, the expression with $t = T$ being averaged over the entire $j^{th}$ control period. $s_i(t)$ is computed in previous section 3.2.1.

The estimated background channel occupancy for the $(j + 1)^{th}$ control period is given by

$$N^{(e)}(j + 1) = \alpha_3 N^{(e)}(j) + (1 - \alpha_3)N^{(o)}(j).$$  

(3.20)

Based on this estimation, the mean of the occupancy distribution in cell $i$ time $t$ of the $(j + 1)^{th}$ control period consists of both the background component and controllable components, given by

$$\langle n_i(t) \rangle = N^{(e)}(j + 1)(j + 1) + s_i(t)a_i(j + 1)\lambda_i^{(e)}(j + 1)t$$  

(3.21)

Where the background component $N^{(e)}(j + 1)$ is obtained in Eq. (3.20), $\lambda_i^{(e)}(j + 1)$ is the estimated new call arrival in the next control period by Eq. (3.17), and $a_i(j + 1)$ is the acceptance ratio for the next control period that needs to be computed. Similarly, we can obtain the estimation for the variance $\delta_i(t)^2$ of the channel occupancy distribution at time $t$. As it turns out that the channel occupancy distribution can be approximated by a Poisson one, we take the variance to be the same as the mean $\langle n_i(t) \rangle$.

After obtaining the mean and variance of the occupancy distribution, the time-dependent call dropping probability $D_i(t)$ can be calculated based on Eq.(3.14). In addition, by using bi-section algorithm, the maximum acceptance ratio $a_i$ can be derived. The flow chart is given in Fig. 3.

Notice that the coefficients $\alpha_i$ ($i = 1, 2, 3$) used in Eqs. (3.17), (3.18) and (3.20) have to be properly selected to smooth all the estimated values. In general,
Figure 3: Flow chart of the derivation of the acceptance ratio
a small value of $\alpha_i$ (thus a large value of $1 - \alpha_i$) can keep track of the changes more accurately, but is perhaps too heavily influenced by temporary fluctuations. On the other hand, a large value of $\alpha_i$ is more stable but could be too slow in adapting to the real traffic changes. In our experiment, we find the setting of $\alpha_1$ and $\alpha_2$ between 0.6 and 0.7, and $\alpha_3$ between 0.8 and 0.9 are adequate for the estimation.

### 3.4 Simulation Results

![Diagram](image)

Figure 4: (a) A 19-cell hexagonal cellular network with wrap-around connection, (b) Topology of paths connecting $k = i$.

Simulations were performed on a hexagonal cluster of 19 cells given in Fig. 4. To alleviate finite size effects, we implement periodic connections on the 3 pairs of opposite sides of the cluster (wrap-around). The parameters used in the simulation are: $N = 100$, $\mu = 0.005$ s$^{-1}$, $h_i = h = 0.01$ s$^{-1}$, $T = 20$ s, and $P_{QoS} = 0.01$. Under such a setting, a connection lasts on average 200 s and the mobile hands off twice during its life time. The coefficients $\alpha_1$ and $\alpha_2$ used are 0.6, unless specified otherwise. And except for the Fig. 16 and 17, the handoff rate is assumed to be given.

### 3.4.1 Comparison with SDCA

We first compare the result with that of SDCA proposed in [30]. Fig. 5 shows that both schemes can guarantee the target call dropping probability ($P_{QoS} =$
Figure 5: The comparison of call dropping probability with the SDCA

Figure 6: The comparison of channel utilization with the SDCA
\( \tilde{D}_i = 0.01 \), but the scheme based on local parameter estimation has a slight over-provision of the bandwidth, thus yielding a call dropping probability slightly lower than the target. This is caused by the conservative estimation in the control algorithm, partly due to the Poisson approximation for the variance. Notice that this is also evident from the utilization curve shown in the Fig. 6.

### 3.4.2 Periodically Changing Traffic

![Figure 7](image1)

Figure 7: The evolution of the traffic input scaled by the factor \( P \)

![Figure 8](image2)

Figure 8: The dynamic call dropping probability for each control period at \( P = 2 \, s^{-1} \)

We next present the results when the traffic condition changes, by considering the scenario that the traffic input changes periodically, as is best reflected from daily telephone operations. Specifically, the traffic input evolves as the staircase function shown in Fig. 7, in which each step of the staircase is 40 control periods
or 800 s. The parameter $P$ is the scaling factor, and the traffic changes in steps of $0.2P$.

Fig. 8 demonstrates the dynamic behavior of the call dropping probability for each control period when the network is under saturation loading $P = 2 \text{ s}^{-1}$. In this case, the long term call dropping probability is maintained at 0.0085 (below the target dropping probability 0.01). The utilization obtained is 0.81. The new call blocking probability is 0.67 caused by the overloading condition. One observes from Fig. 8 that the transient behavior of the call dropping probability is also periodic, well matching the periodic changes of the traffic input. In addition, the target call dropping probability is violated for the first few control periods when the traffic $\lambda_i$ changes, due to inaccurate estimations of the traffic. Once the estimation becomes stable, the target dropping probability is guaranteed for those control periods.

3.4.3 Fluctuating Traffic

![Fluctuating Traffic Graph](image_url)

Figure 9: The traffic input with scaling factor $P$ increased from 0.4 $\text{s}^{-1}$ to 12 $\text{s}^{-1}$

We next investigate the performance under fluctuating traffic conditions. We consider a similar traffic pattern but with increasing overall traffic intensity given in Fig. 9. This is the same as previous one shown in Fig. 8 except that the scaling parameter $P$ increases after every long cycle of 360 control periods. Fig. 10 demonstrates that the target call dropping probability is still well maintained. The call dropping probability for each cycle (i.e., for a fixed $P$) is also presented in Fig. 10. Notice that the cumulative average of the call dropping probability is...
Figure 10: The call dropping probability vs. changing $P$

Figure 11: The channel utilization vs. changing $P$

Figure 12: The new call blocking probability vs. changing $P$
considerably lower than the individual measurements, but the two converges when the scaling factor $P$ becomes significantly large as expected. The corresponding utilization and new call blocking probability are plotted in Fig. 11 and 12 respectively. The high blocking probability for new calls is a consequence of the overloaded traffic, which is the range of the interest. This clearly demonstrates the robustness and stability of our bandwidth allocation mechanism. Under such heavy loading, the channel utilization is maintained at over 80%. This also illustrates the fact that on average about 20% of bandwidth is reserved for handoff in order to maintain the target call dropping probability.

3.4.4 Impact of Estimation of the Coefficients $\alpha_i$

![Figure 13: The traffic input with periodic change](image)

![Figure 14: The call dropping probability vs. adjusted $\alpha$ values](image)

We next consider more volatile traffic conditions, and study the impact on the call dropping probability of different estimation coefficients $\alpha_i$ (i.e., $\alpha_1$ and $\alpha_2$ since
the handoff rate is assumed to be given). The input traffic follows a periodic change given in Fig. 13, with \( n \) being the number of control periods. Note that under the given system parameters, the system saturates when the average traffic input is \( 0.5 \, \text{s}^{-1} \), therefore under both low input \( (\lambda = 1 \, \text{s}^{-1}) \) and high input \( (\lambda = 4 \, \text{s}^{-1}) \), the system is under saturation. Fig. 14 describes the behavior of the call dropping probability under a variety of traffic input and different values of \( \alpha_i \). Specifically, Curve 1 shows the case that the \( n = 150 \) (i.e., the traffic is changed every 150 control periods) and \( \alpha_i = 0.6 \). In this case the target call dropping probability is well maintained around \( P_{QoS} = 0.01 \). Curve 2 presents a similar scenario with the same \( \alpha_i \), but the traffic is changing more frequently, i.e., the traffic is changed every \( n = 40 \) control periods. The result from Curve 2 shows that the target dropping probability cannot be satisfied. The major reason is that the chosen value of \( \alpha_i = 0.6 \) is not adequate for keeping track of such frequent traffic change. Adjusting the \( \alpha_i \) value can clearly improve the performance guarantee. The result is illustrated in Curve 3 of Fig. 14 with \( \alpha_i \) set to 0.1, and the target dropping probability is indeed guaranteed. However, the target cannot be met under more frequent traffic updates such as Curve 4 in the same figure, in which the traffic is updated every \( n = 10 \) control periods.

![Figure 15: The dynamic call dropping probability for periodic traffic input](image)

The call dropping probability obtained reflects the periodic change of the input traffic every \( n = 40 \) control periods. More importantly, at the beginning of the traffic change from low input \( (\lambda = 1 \, \text{s}^{-1}) \) to high input \( (\lambda = 4 \, \text{s}^{-1}) \), i.e., every 80 control periods, the instantaneous call dropping probability is increased
significantly, as much as about 10–15 times the target call dropping probability \( P_{\text{QoS}} = 0.01 \) in Fig. 15. This is caused by an excessive underestimation of the input traffic during the initial control periods when traffic increases. Such an impact can be leveraged over longer control periods, such as \( n = 150 \) when \( \alpha_i = 0.6 \) and \( n = 40 \) when \( \alpha_i = 0.1 \), but clearly cannot be compensated under \( n = 10 \) with any \( \alpha_i \) setting since the call dropping probability in the single control period immediately after the hike in call rate can account for more than 10 periods' target call dropping probability. Therefore, the target call dropping probability cannot be guaranteed when \( n = 10 \) shown in Curve 4 in Fig. 14. This can also be better observed from the dynamical behaviors shown in Fig. 15.

### 3.4.5 Estimated Handoff Rates

![Graph showing estimated handoff rates](image)

Figure 16: The call dropping probability based on local estimation of the handoff rates

![Graph showing call dropping probability](image)

Figure 17: The call dropping probability with estimated handoff rates
Figure 18: The estimated handoff rate vs. time

Figure 19: The call dropping probability for periodic handoff rate
Finally, we are interested in the system performance when the handoff rates are estimated. First we adjust the traffic input and use the estimation algorithm to trace the handoff changes accordingly. Fig. 16 essentially re-captures the call dropping probability shown in Fig. 5. The only difference is that in Fig. 16 the handoff rate is on-line periodically estimated as in Eq. (3.18), thus the survival probability $f_{in}(t)$ is computed according to Eq. (3.9). It shows that the cumulative (target) call dropping probability can be guaranteed. Fig. 17 presents the results for the traffic input given in Fig. 9, and is similar to those presented in Fig. 10. Next we assume that the user mobility pattern is periodically changed, specifically, the user handoff rate $h$ is oscillated between 0.05 and 0.1 every 200 control periods. Fig. 18 illustrates the handoff rate obtained by the estimation algorithm, which accurately reflects the real changes. The corresponding call dropping probability is on target and plotted in Fig. 19.

### 3.5 Conclusions

In this chapter, we introduce a novel adaptive bandwidth allocation scheme for mobile wireless networks based on local on-line parameter estimations. The novelties of the proposed scheme are:

1. Both estimation and bandwidth allocation are carried out periodically, thus can effectively adapt to the changing traffic condition;

2. The estimation is restricted to the local cell, thus eliminating the signaling overhead often required by all existing bandwidth allocation schemes;

3. The allocation algorithm is based on a stochastic control, which results in a more efficient use of the bandwidth and leads to a more effective and stable control. The results demonstrate that the proposed adaptive bandwidth allocation can guarantee the pre-defined bound on call dropping probability under changing traffic conditions, while at the same time achieving high bandwidth utilization.
One limitation is that the control algorithm still relies on the assumptions that the arrival obeys a Poisson process and call durations follow an exponential distribution, other distributions such as Pareto-distribution for data traffic [16] and Hyper-Erlang distribution [11] will be considered in the future work. In addition, we are investigating other QoS parameters guarantee such as end-to-end delay and considering multiple types of traffic [13, 14, 29].
CHAPTER 4

A DUAL-THRESHOLD RESERVATION MODEL FOR VOICE/DATA INTEGRATED MOBILE WIRELESS NETWORKS

There has been a rapid development in wireless cellular communications. The next generation of networks are expected to eventually carry multimedia traffic - voice, video, images, data, or combinations of them. Various issues related to this future wireless mobile network carrying multimedia traffic have to be carefully examined before such systems can be realized. The wireless spectrum remains as the prime limited resource in integrated voice/data mobile networks, therefore, it is necessary to develop mechanisms that can provide effective bandwidth management that can furnish effective and dynamic allocation of the bandwidth to satisfy diverse QoS requirements.

In integrated voice/data mobile networks, several schemes have been proposed based on the movable-boundary strategy with or without reservation for voice or data traffic. For example, in [36], Haung, Lin and Ho have discussed a movable-boundary scheme which can dynamically adjust the number of channels for each type of traffic. The limitation is, however, the non-prioritized handoff. In [14], Li, Lin and Chanson have proposed a generic model which can handle multiple streams of traffic with different reservation thresholds, however, the determination of cutoff remains a problem to be solved.

In this chapter, we propose a bandwidth allocation model for integrated voice/data mobile wireless networks. The main focus is on the QoS guarantee in terms of voice handoff dropping probability, voice blocking probability and data loss probability. Our model is based on the well-known Guarded Channel (GC) policy which is employed in the traditional voice-centric cellular networks.
The basic idea of GC is to reserve a fixed number of channels in each cell for handoff use exclusively. We extend the idea to a two threshold model, referred as a dual-threshold reservation (DTR) scheme. In addition, we develop an analytical model which can be used to achieve the algorithm's performance by using the iterative technique. Based on the analytical model, two algorithms are proposed to achieve optimal system performance.

4.1 DTR Scheme

We consider an integrated voice/data mobile system under a fixed channel allocation scheme. Each cell has a capacity of $C$ channels. Following the conventional assumptions, we assume that the arrivals of voice calls (new calls and handoffs) and data traffic (new and handoff) are Poisson processes, with rates $\lambda_{vn}, \lambda_{vh}, \lambda_{dn}, \lambda_{dh}$, respectively. Although the cell residence times are typically non-exponential in a particular mobile system, the analysis based on the simplified exponential assumption has been widely used ([25], [31], [37], [38]) and does provide useful mean value information for the output measures. Therefore, the residence time of voice and data are assumed to follow an exponential distribution with means $1/\mu_{vr}$ and $1/\mu_{dr}$. For simplicity, we assume the bandwidth requirement of voice and data are the same, while this assumption can be easily relaxed.

The rationale behind the design of the new model is that voice calls, both handoffs and new calls, are expected certain QoS guarantee. They are usually measured by the voice call dropping probability and new call blocking probability, which often need to be maintained at a pre-defined level. Data traffic, on the other hand, can be adaptive, yet its loss probability is an important system performance indication. Based on this observation, we propose a dual-threshold reservation algorithm, which is a natural extension of the well-know GC policy.

The key idea of the DTR is to reserve certain channels for both handoff and arrivals of voice traffic, respectively. The model is illustrated in Fig. 4, where there are two thresholds, $K_1$ and $K_2$. Data can be accepted only if the system
occupancy is less than $K_2$, while a new voice call will be rejected whenever the channel occupancy exceeds the threshold $K_1$.

![Cell Diagram](image)

Figure 20: The architecture of cells

### 4.2 Analytical Model

#### 4.2.1 Transition Diagram and Balance Equation

$P_{ij}$ is defined as the steady state probability with $i$ voice calls and $j$ data packets in the cell. The DTR scheme can be modeled as a two-dimensional Markov chain. The corresponding state transition diagram is drawn in Fig. 21, where $\lambda_v = \lambda_{vn} + \lambda_{vh}$ represents the total voice traffic arrival rate, $\lambda_d = \lambda_{dn} + \lambda_{dh}$ is the total arrival rate of data, $\mu_v = \mu_{vr} + \mu_{vh}$ is the service rate of voice and $\mu_d = \mu_{dr} + \mu_{dh}$ is the service rate of data.

According to Fig. 21, the balance equations are derived as follow:
Figure 21: State transition diagram

\[
\begin{align*}
\text{[If } j \neq K_2]\quad P_{ij} &= \\
&= \begin{cases} 
0 & \text{if } i + j > C \\
\frac{\lambda_{th} P_{(i-1),j}}{\mu_{th} + j \mu_d} & \text{if } i + j = C \\
\frac{(i+1) \mu_r P_{(i+1),j} + \lambda_{th} P_{(i-1),j} + (j+1) \mu_d P_{i,j+1}}{\lambda_{th} + (i+1) \mu_r + j \mu_d} & \text{if } K_1 < i + j < C \\
\frac{\lambda_{rh} P_{(i-1),j} + (j+1) \mu_d P_{i,j+1}}{\lambda_{rh} + j \mu_d} & \text{if } K_1 < i + j < K_1 \\
\frac{(i+1) \mu_r P_{(i+1),j} + \lambda_{rh} P_{(i-1),j} + (j+1) \mu_d P_{i,j+1} + \lambda_d P_{i,j-1}}{\lambda_r + (i+1) \mu_r + j \mu_d + \lambda_d} & \text{if } K_2 < i + j < K_2 \\
\frac{(i+1) \mu_r P_{(i+1),j} + \lambda_r P_{(i-1),j} + (j+1) \mu_d P_{i,j+1} + \lambda_d P_{i,j-1}}{\lambda_r + (i+1) \mu_r + j \mu_d + \lambda_d} & \text{if } 0 < i + j < K_2 \\
\frac{(i+1) \mu_r P_{i,j} + \mu_d P_{j+1}}{\lambda_r + \lambda_d} & \text{if } i = 0 \text{ and } j = 0 \\
0 & \text{if } i < 0 \text{ or } j < 0
\end{cases}
\end{align*}
\]
Let $P_{ud}$ denote the voice dropping probability, $P_{vb}$ be the new voice call blocking probability and $P_{d}$ as the data lost probability\(^1\). For given $K_1$ and $K_2$, $P_{ud}$, $P_{vb}$ and $P_{d}$ are given by:

$$P_{ud} = \sum_{i+j=C} P_{ij}$$  \hspace{1cm} (4.3)

$$P_{vb} = \sum_{i+j=K_1} P_{ij}$$  \hspace{1cm} (4.4)

$$P_{d} = \sum_{i+j=K_2} P_{ij}$$  \hspace{1cm} (4.5)

### 4.2.2 Calculation of State Probability $P_{ij}$

As shown in Eqs. (4.3), (4.4) and (4.5), the key to computing $P_{ud}$, $P_{vb}$ and $P_{d}$ is to derive the steady state probability $P_{ij}$. $P_{ij}$ can be solved by adopting the recursive technique first proposed by Herzog, Woo and Chandy in [7]. The algorithm is based on the inherent feature of Champan-Kolmogoroff system equations which

\(^1\)We do not distinguish the performance between handoff and new arrivals of data traffic.
state that there exist a set of state probabilities, called boundaries and all states can be expressed as a function of boundaries. The recursive technique has been shown to be suitable for solving a wide class of queuing problems, the key steps are summarized as follow:

1. Select proper state as boundaries.

2. Express all remaining state probabilities as functions of boundaries.

3. Calculate boundaries.

4. Solve all remaining state probabilities based on boundary values.

Concretely, in our specific case, \( P_{ij} \) is computed as follow:

1. Selection of boundaries.

   In our formulation, we choose states \( P_{i0} (i = 0, \cdots, C) \) as boundaries. Accordingly, all state probabilities can be re-written as functions of boundaries:

   \[
   P_{ij} = \sum_{k=0}^{C} C_{i0}^{k} P_{k0} \quad (4.6)
   \]

   where coefficient \( C_{i0}^{k} \) is defined as:

   \[
   C_{i0}^{k} = \begin{cases} 
   1 & \text{if } i = k \\
   0 & \text{if } i \neq k 
   \end{cases} \quad (4.7)
   \]

   and coefficients \( C_{ij} (j \neq 0) \) are to be solved.

2. Computation of coefficients \( C_{ij}^{k} \).

   From Eqs. (4.1) and (4.2), \( P_{ij} \) can be denoted by:

   \[
   P_{ij} = \frac{A_{ij1}P_{i-1,j} + A_{ij2}P_{i+1,j} + A_{ij3}P_{i,j-1} + A_{ij4}P_{i,j+1}}{B_{ij}} \quad (4.8)
   \]

   where \( B_{ij}, A_{ij1}, A_{ij2}, A_{ij3} \) and \( A_{ij4} \) are corresponding coefficients shown in Eq. (4.1). Moreover, Eq. (4.8) can further be written as:

   \[
   P_{ij} = \frac{P_{i,j+1} - P_{i,j-1} - P_{i+1,j}A_{ij1}A_{ij2}A_{ij3} - P_{i+1,j}A_{ij2}A_{ij3}}{A_{ij1}A_{ij2}A_{ij3}} \quad (4.9)
   \]

44
For the $ith$ column in the state transition diagram in Fig. 21 ($i$ from 0 to $K_2$), substitute all state probabilities in the first $C - i - 1$ row in Eq. (4.9) to functions of boundaries (given in Eq.(4.6)), we obtain,

$$
\sum_{k=0}^{C} C_{ij}^k P_{k0} = \frac{\sum_{k=0}^{C} C_{i(j-1)}^k P_{00} B_{ij} - \sum_{k=0}^{C} C_{i(j-1), (j-1)}^k P_{00} A_{i(j-1), (j-1)}^k}{A_{i(j-1), (j-1), (j-1)}^0} - \frac{\sum_{k=0}^{C} C_{i(j+1), (j-1)}^k P_{00} A_{i(j-1), (j-1)}^k + \sum_{k=0}^{C} C_{i(j-2), (j-1)}^k P_{00} A_{i(j-1), (j-1)}^k}{A_{i(j-1), (j-1)}^0}
$$

(4.10)

For each particular $k$ ($k$ from 0 to C), the Eq. (4.10) is to be held under the special case: $P_{k0} = 1$ and $P_{00} = 0$ ($i \neq k$). Substitute all $P_{k0}$, Eq. (4.10) leads to

$$
C_{ij}^k = \frac{C_{i(j-1)}^k B_{ij} - C_{(i-1), (j-1)}^k A_{i(j-1), (j-1), (j-1)}^1 - C_{(i+1), (j-1)}^k A_{i(j-1), (j-1), (j-1)}^2 - C_{i(j-2), (j-1)}^k A_{i(j-1), (j-1), (j-1)}^3}{A_{i(j-1), (j-1)}^0}
$$

(4.11)

For $i \neq k$ and given $k$ ($k = 0, \cdots, C$), $C_{ij}^k$ can be determined by solving a system of $CK_2 - \frac{K_2(K_2-1)}{2}$ linear equations recursively.

3. Calculation of boundaries.

After knowing all coefficients $C_{ij}^k$, boundaries can be determined by solving the remaining $C$ independent equations along with the normalizing condition.

$$
\sum_{i} \sum_{j} P_{ij} = 1
$$

Notice that $P_{ij}$ is a function of boundaries, therefore, $P_{ij}$ is known once the boundaries are solved.

The complexity of solving of $P_{ij}$ is significantly reduced by adopting the recursive technique.

After computing all state probabilities, the voice blocking probability, voice dropping probability and data dropping probability are to be calculated based on Eqs. (4.3), (4.4) and (4.5), respectively.
4.3 Numerical Results

In the formulation, we have adopted the widely-used assumption in literature that \( \lambda_{vh} \) is given and fixed \(^2\). The parameters are set as: channel size \( C \) is 20, \( \lambda_{vn} = \lambda_{vh} = 0.07 \), \( \mu_{vr} = 0.005 \, S^{-1} \), \( \mu_{vh} = 0.015 \, S^{-1} \), \( \mu_{dr} = 0.0005 \, S^{-1} \), \( \mu_{dh} = 0.0001 \, S^{-1} \), \( \lambda_{dn} = 0.007 \), and \( \lambda_{dh} = 0.0007 \).

Eqs. (4.3)-(4.5) clearly show different \( K_1 \) and \( K_2 \) configurations result in different system performance in terms of \( P_{vd}, P_{vb} \) and \( P_d \). To investigate the impact of \( K_1 \) and \( K_2 \) on system performance, we consider three scenarios, 1) increasing \( K_1 \) under a constant \( K_2 \), 2) increasing \( K_2 \) under a constant \( K_1 \) and 3) increasing \( K_1 \) and \( K_2 \) simultaneously.

4.3.1 Increasing \( K_1 \) Under a Constant \( K_2 \)

The relationships between \( K_1 \) and the obtained call dropping probability, call blocking probability and data blocking probability are plotted in Fig. 22, 23 and 24, respectively. Each curve demonstrates the system performance for a particular \( K_2 \) value. Moreover, for each \( K_2 \) value, \( K_1 \) is increased from \( K_2 + 1 \) to \( C \).

![Figure 22: Voice handoff dropping probability vs. K1](image)

Fig. 22 shows that the voice dropping probability grows with \( K_1 \). This is caused by the diminution of dedicated channels for handoff.

\(^2\)Noticeably, under a single traffic type, the handoff rate can be derived from other parameters using the iterative method, e.g., shown in [6], [17]. However, this poses considerable difficulty in the integrated voice/data systems, and there is no known mechanism that can be applied.
Fig. 23 demonstrates the impact of increasing $K_1$ on the voice blocking probability. The increase of $K_1$ implies that less channels are reserved for voice handoff calls, which benefits new calls and results in a decrease of voice blocking probability. Notice that Fig. 23 also demonstrates that the increment of accepted new calls causes more handoff calls being generated, consequently, the total traffic is increased.

We next investigate the performance of data blocking probability. It can be observed from Fig. 24 that data blocking probability increases when $K_1$ is increased. More total voice traffic due to $K_1$’s increase means more voice calls are competing for the limited channels with data, as a result, more data are lost in the system.
4.3.2 Increasing $K_2$ Under a Constant $K_1$

Fig. 25-27 demonstrate the performance of voice dropping probability, voice blocking probability and data blocking probability for various $K_2$ under a constant $K_1$. Each curve in the figures illustrates the performance for a particular $K_1$ value and for each given $K_1$, $K_2$ is increased from 0 to $K_1 - 1$.

![Figure 25: Voice handoff dropping probability vs. K1](image)

![Figure 26: Voice blocking probability vs. K1](image)

When $K_2$ is small, the voice dropping probability is almost stable with a little fluctuation. This is because the number of channels that are shared by voice and data are too few to cause any significant effect on the voice dropping and blocking probabilities. With $K_2$ increasing, the voice dropping probability is increased accordingly as a result of more traffic being injected into the system, which further causes more handoff calls to compete for the reserved channels.

Fig. 26 illustrates the performance of voice blocking probability. The increase of $K_2$ permits data to occupy more channels. In another word, less channels are left for specific voice use and more voice calls are rejected by the system.
Fig. 27 shows the results of data blocking probability. The increase of $K_2$ obviously benefits data packets and result in the decrease of data blocking probability.

4.3.3 Increasing $K_1$ and $K_2$ Simultaneously

Calculation results are shown in Fig. 28-30. Each curve represents the performance with constant distance between $K_1$ and $K_2$.

Voice dropping probability increases as a result of both $K_1$ and $K_2$ increase as shown in Fig. 28.

Fig. 29 shows that the voice blocking probability decreases while the distance between $K_1$ and $K_2$ is increased. The increase of $K_1$ obviously benefits new voice calls while the increase of $K_2$ will permit more data enter the system which may block more voice calls. However, the fact that there are more voice traffic than data traffic determines that the impact of $K_1$'s increase will show greater effect on
the performance of voice blocking probability than the $K_2$'s increase. As a result, the voice blocking probability will decrease when $K_1$ and $K_2$ increase together.

Fig. 30 demonstrates the decrease of data blocking probability caused by the increment of channels shared by voice and data.

Previous analysis and results demonstrate that the effect of $K_1$ and $K_2$ variance is determinable. The monotonic of $P_{vo}$, $P_{vb}$ and $P_d$ with the increase of $K_1$ and $K_2$ makes it possible to propose some algorithms to simplify the process of obtaining the optimal $K_1$ and $K_2$ threshold solution.

4.4 Algorithms to Set $K_1$ and $K_2$ for Given Objective Functions

In this section, we propose two algorithms to achieve the optimal thresholds $K_1$ and $K_2$ under different pre-defined objective functions. One aims to minimize the data blocking probability while satisfying the QoS requirement against voice
handoff dropping probability and voice new call blocking probability. The other objective function is to minimize the voice blocking probability while maintaining the QoS requirement of voice dropping probability and data blocking probability. According to the discussion in previous section 4.3, the impact of $K_1$ and $K_2$ on voice handoff dropping probability, voice call blocking probability and data blocking probability can be summarized in Fig. 31-33. Adopting the special monotonic shown in the figures, it is possible to reduce the feasible solution area while searching the optimal threshold. The fundamental idea is to reject impossible results directly according to the obtained monotonical property.

Figure 31: Voice dropping probability under different $K_1$ and $K_2$

![Diagram](image)

Figure 32: Voice blocking probability under different $K_1$ and $K_2$

![Diagram](image)

The input parameter set includes $\lambda_{vm}$, $\lambda_{vh}$, $\lambda_{dn}$, $\lambda_{dh}$, $\mu_{vh}$, $\mu_{vr}$, $\mu_d$, $C$. The algorithm output $K_1$ and $K_2$ as a result.
4.4.1 Algorithm I

Algorithm I aims to achieve the optimal $K_1$ and $K_2$ to minimize the data blocking probability while maintaining the required QoS target on voice dropping probability and voice blocking probability.

Algorithm Description

Considering the threshold $K_1$, it is obviously that the optimal $K_1$ can not be larger than the optimal threshold if no data is permitted, denoted by $K^*$, therefore, searching optimal $K_1$ from $K^*$ can achieve the same optimal result as starting from $C$. Moreover, as demonstrated in Fig. 31, 32 and 33, for each $K_1$ value, sets of $K_2$ can guarantee the target QoS requirements against the voice dropping probability. The largest feasible $K_2$ value within this set can achieve the least data blocking probability for this given $K_1$. With decrease in $K_1$, this largest feasible $K_2$ will increase according to the properties. The increment of $K_2$ and decrement of $K_1$ both cause the voice blocking probability to increase. There is a special point at which a decrease of $K_1$ or an increase of $K_2$ will violate the target voice blocking probability QoS requirements. Obviously, this $K_2$ value is the largest $K_2$ value which can meet the target QoS constraints on both voice dropping probability and voice blocking probability. In another word, the $K_2$ and $K_1$ value are the optimal solution subjected to the objective function: minimize
the data blocking probability while maintaining the QoS requirement against
voice dropping probability and voice blocking probability.

The algorithm is described as follows:

[Step 1] Calculate the optimal threshold with no data packets input under certain
QoS requirement against voice handoff dropping probability. Represent it
by $K^*$.  

[Step 2] Initialize $K_2$ as 0 and $K_1$ as $K^*$. 

[Step 3] Increase $K_2$ until it violates the target voice dropping probability $P_{vd(QoS)}$
or the target voice blocking probability $P_{vb(QoS)}$. 

[Step 4] If $P_{vb(QoS)}$ is violated, stop. Otherwise go to Step (5). 

[Step 5] Decrease $K_1$ until the target $P_{vd(QoS)}$ is satisfied again or $K_1 = K_2$.
If $K_1 > K_2$, go to Step (3). Otherwise, this QoS requirement can not be
satisfied under current network configuration.

Proof and Analysis

Let $P_*(i,j)$ be the probability of $P_*$ when the threshold is set to(i,j), where * can be voice blocking probability, voice dropping probability or data blocking probability. From the algorithm, we have:

$$P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}. \quad /*\text{Monotonic of } P_{vd} */$$

$$P_{vb}(K_1, K_2 + 1) > P_{vb(QoS)}. \quad /*\text{Stop condition. } */$$

Assume there is another set of threshold, denoted by $(K_1^*, K_2^*)$, which can achieve better or no worse performance than the one obtained by the algorithm. We analyze all possibilities case by case.

[Case I] $K_1^* > K_1$ and $K_2^* > K_2$

This case contains three possibilities.
1. \( K_1^* - K_1 > K_2^* - K_2 \)

From the monotonic of \( P_{vd} \), we have

\[
P_{vd}(K_1^*, K_2^*) \geq P_{vd}(K_1^* + (K_2 - K_2^*), K_2^* + (K_2 - K_2^*))
= P_{vd}(K_1^* + (K_2 - K_2^*), K_2)
\]  

(4.12)

From \( K_1^* - K_1 > K_2^* - K_2 \) We have

\[
K_1^* + (K_2 - K_2^*) \geq K_1 + 1
\]

Moreover, from the obtained monotonic of \( P_{vd} \) shown in Fig. 31, we have

\[
P_{vd}(K_1^* + (K_2 - K_2^*), K_2) \geq P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}
\]

Therefore,

\[
P_{vd}(K_1^*, K_2^*) \geq P_{vd}(K_1^* + (K_2 - K_2^*), K_2) > P_{vd(QoS)}
\]

This obviously contradicts with the assumption that \( (K_1^*, K_2^*) \) is another set of threshold which can satisfy both specific QoS requirement of voice dropping probability and voice blocking probability.

2. \( K_1^* - K_1 < K_2^* - K_2 \)

Similarly, we have

\[
P_{vd}(K_1^*, K_2^*) \geq P_{vd}(K_1^* + (K_1 + 1 - K_1^*), K_2^* + (K_1 + 1 - K_1^*))
= P_{vd}(K_1 + 1, K_2^* + (K_1 + 1 - K_1^*))
\]  

(4.13)

From condition \( K_1^* - K_1 < K_2^* - K_2 \), we have

\[
K_2^* + (K_1 + 1 - K_1^*) \leq K_2
\]

Moreover, from the obtained monotonic of \( P_{vd} \) shown in Fig. 31, we have

\[
P_{vd}(K_1 + 1, K_2^* + (K_1 + 1 - K_1^*)) \geq P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}
\]

Therefore,

\[
P_{vd}(K_1^*, K_2^*) \geq P_{vd}(K_1 + 1, K_2^* + (K_1 + 1 - K_1^*)) > P_{vd(QoS)}
\]

This results in the contradiction again.
3. $K_1^* - K_1 = K_2^* - K_2$

From the monotonic of $P_{vd}$, we have

$$P_{vd}(K_1^*, K_2^*) \geq P_{vd}(K_1 + 1, K_2 + 1) \geq P_{vd}(K_1 + 1, K_2) > P_{vd}(QoS)$$

This result leads to the contradiction again.

Therefore, it is impossible to find another optimal threshold solution $K_1^*$ and $K_2^*$ that can satisfy $K_1^* > K_1$ and $K_2^* > K_2$.

[Case II] $K_1^* > K_1$ and $K_2^* < K_2$

From the shown monotonic of $P_d$ in Fig. 33, we have

$$P_d(K_1^*, K_2^*) > P_d(K_1, K_2^*) > P_d(K_1, K_2)$$

This means under $(K_1^*, K_2^*)$, the data blocking probability is higher than $(K_1, K_2)$, which contradicts the assumption that the $(K_1^*, K_2^*)$ can achieve better or no worse performance than $(K_1, K_2)$. Therefore there does not exist another optimal threshold $K_1^*$ and $K_2^*$ that can satisfy $K_1^* > K_1$ and $K_2^* < K_2$.

[Case III] $K_1^* < K_1$ and $K_2^* > K_2$

The obtained monotonic of $P_{vb}$ shown in Fig. 32 leads to

$$P_{vb}(K_1^*, K_2^*) > P_{vb}(K_1, K_2^*) > P_{vb}(K_1, K_2 + 1) > P_{vb}(QoS)$$

This result contradicts the assumption that thresholds $(K_1^*, K_2^*)$ will not violate the target QoS requirement, therefore, it is impossible that there is another set of optimal $K_1^*$ and $K_2^*$ with $K_1^* < K_1$ and $K_2^* > K_2$.

[Case IV] $K_1^* < K_1$ and $K_2^* < K_2$

This case can be divided into two sub-cases.
1. $K_1 - K_1^* \leq K_2 - K_2^*$

From the monotonic of $P_d$ shown in Fig. 33, we have

$$P_d(K_1^*, K_2^*) > P_d(K_1, K_2) \geq P_d(K_1, K_2 + (K_1 - K_1^*))$$

This violates the assumption that $(K_1^*, K_2^*)$ will achieve better or no worse performance than $(K_1, K_2)$.

2. $K_1 - K_1^* > K_2 - K_2^*$

From Fig. 32, we have

$$P_v(K_1^*, K_2^*) \geq P_v(K_1^* + (K_2 - K_2^* + 1), K_2 + 1) \geq P_v(K_1, K_2 + 1) \geq P_v(QoS)$$

This result violates the target QoS requirements. Therefore, it contradicts that $(K_1^*, K_2^*)$ is another set of optimal solution which can satisfy the target QoS requirements.

[Case V] $K_1^* = K_1$

This also contains two cases,

1. $K_1^* = K_1$ and $K_2^* > K_2$. Obviously,

$$P_v(K_1^*, K_2^*) \geq P_v(K_1, K_2 + 1) > P_v(QoS)$$

Similarly, it violates the pre-assumption that $(K_1^*, K_2^*)$ can satisfy the target QoS requirements.

2. $K_1^* = K_1$ and $K_2^* < K_2$. Obviously,

$$P_d(K_1^*, K_2^*) > P_d(K_1, K_2)$$

This result contradicts the assumption that $(K_1^*, K_2^*)$ can achieve better or no worse performance than $(K_1, K_2)$.
[Case VI] \( K_2^* = K_2 \)

It can be analyzed by two sub-cases,

1. \( K_1^* > K_1 \) and \( K_2^* = K_2 \). Obviously,

\[
P_{vd}(K_1^*, K_2^*) \geq P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}
\]

This violates the target QoS requirements, therefore, leads to a contradiction.

2. \( K_1^* < K_1 \) and \( K_2^* = K_2 \). Obviously,

\[
P_{vb}(K_1^*, K_2^*) \geq P_{vb}(K_1 - 1, K_2) > P_{vb}(K_1, K_2 + 1) > P_{vb(QoS)}
\]

Similarly, it violates the target QoS requirements and contradicts with the pre-assumption.

From the previous analysis, we know that the result obtained from the algorithm I is the optimal and the only optimal solution to the objective function.

### 4.4.2 Algorithm II

This algorithm is for the objective function which aims to minimize the voice blocking probability while satisfying the QoS requirement of voice dropping probability and data blocking probability.

**Algorithm Description**

It is always true that the optimal \( K_2 \) can not be less than the largest threshold by setting \( K_1 \) to \( K_2 \), denoted by \( K' \) because when \( K_1 \) is set equal to \( K_2 \), less voice calls compete for the limited channels with data, with \( K_1 \) increased, more voice calls enter the system to compete with data which causes the data blocking probability to increase. Thus to meet the target requirement of data blocking probability, \( K_2 \) must be increased. Therefore, searching \( K_2 \) from \( K' \) can obtain the same optimal result as starting from 0. Moreover, as shown in Fig. 31, 32
and 33, for each $K_2$ value, a set of $K_1$ can guarantee the target QoS requirements against the data blocking probability. The largest feasible $K_1$ value within this set can achieve the least voice blocking probability for this given $K_2$. With $K_2$ increases, this largest feasible $K_1$ will increase accordingly. The increment of $K_1$ and increment of $K_2$ both cause voice dropping probability to increase. Therefore, by moving $K_1$ and $K_2$ in the same direction, at some point, the increase of $K_1$ or $K_2$ will violate the target voice dropping probability QoS requirement. These $K_1$ and $K_2$ value can meet our objective function: minimize the voice blocking probability while maintaining the QoS requirement against voice dropping probability and data blocking probability.

The algorithm is described as follow:

[Step 1] Set $K_1$ equal to $K_2$, increase $K_2$ unit by unit until the QoS requirement against data blocking probability is satisfied, denote this $K_2$ by $K'$. 

[Step 2] Initialize $K_2$ and $K_1$ both to the pre-obtained $K'$. 

[Step 3] Increase $K_1$ until it violates the target voice dropping probability $P_{vd(QoS)}$ or the target data blocking probability $P_{db(QoS)}$. 

[Step 4] If $P_{vd(QoS)}$ is violated, stop. Otherwise, go to Step (5). 

[Step 5] Increase $K_2$ until the target $P_{d(QoS)}$ is satisfied again or $K_1 = K_2$. If $K_1 > K_2$, go back to Step (3). Otherwise, this QoS requirement cannot be satisfied under the current network configuration.

**Proof and Analysis**

Let $P_{*(i,j)}$ be the corresponding probability when the threshold is set to (i,j) where * means voice blocking probability, voice dropping probability or data blocking probability, respectively.

From the algorithm we have

1) $P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}$ /* Stop condition */
2) \( P_d(K_1, K_2 - 1) > P_d(QoS) \)

The second statement can be proven by considering the corresponding \( K_1 \) value at which \( K_2 \) is increased from \( K_2 - 1 \) to the current \( K_2 \), denoted as \( K^1 \). From the algorithm, we have \( P_d(K^1, K_2 - 1) > P_d(QoS) \). Moreover, from Fig. 33, we have \( P_d(K_1, K_2 - 1) > P_d(K^1, K_2 - 1) > P_d(QoS) \). Thus, we have \( P_d(K_1, K_2 - 1) > P_d(QoS) \).

Now assume there is another set of threshold, denoted by \( (K^*_1, K^*_2) \) which can achieve better or no worse performance than the one obtained using the algorithm without violating the target QoS requirements. All possibilities are analyzed case by case.

[Case I] \( K^*_1 > K_1 \) and \( K^*_2 > K_2 \)

Let us analyze this case by three sub-cases,

1. \( K^*_1 - K_1 > K^*_2 - K_2 \).

   We have
   \[
   P_{vd}(K^*_1, K^*_2) > P_{vd}(K^*_1 + (K_2 - K^*_2), K^*_2 + (K_2 - K^*_2)) = P_{vd}(K^*_1 + (K_2 - K^*_2), K_2)
   \]  \hspace{1cm} (4.14)

   From \( K^*_1 - K_1 > K^*_2 - K_2 \), we have
   \[ K^*_1 + (K_2 - K^*_2) \geq K_1 + 1 \]

   Moreover, from Fig. 31, we have
   \[ P_{vd}(K^*_1 + (K_2 - K^*_2), K_2) \geq P_{vd}(K_1 + 1, K_2) > P_{vd}(QoS) \]

   Therefore,
   \[ P_{vd}(K^*_1, K^*_2) > P_{vd}(QoS) \]

   This obviously contradicts the assumption that \( (K^*_1, K^*_2) \) is another set of threshold which can satisfy both specific QoS requirement of voice dropping probability and data blocking probability.
2. \( K^*_1 - K_1 < K^*_2 - K_2 \)

We have

\[
P_{vd}(K^*_1, K^*_2) \geq P_{vd}(K^*_1 + (K_1 + 1 - K^*_1), K^*_2 + (K_1 + 1 - K^*_1)) = P_{vd}(K^*_1 + 1, K^*_2 + (K_1 + 1 - K^*_1)) \tag{4.15}
\]

From \( K^*_1 - K_1 < K^*_2 - K_2 \), we have

\[
K^*_2 + (K_1 + 1 - K^*_1) \leq K_2
\]

In addition, from Fig. 31, we have

\[
P_{vd}(K_1 + 1, K^*_2 + (K_1 + 1 - K^*_1)) \geq P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}
\]

Finally,

\[
P_{vd}(K^*_1, K^*_2) > P_{vd(QoS)}
\]

This results in the same contradiction again.

3. \( K^*_1 - K_1 = K^*_2 - K_2 \). From the monotonic of \( P_{vd} \), we have

\[
P_{vd}(K^*_1, K^*_2) \geq P_{vd}(K_1 + 1, K_2 + 1) \geq P_{vd}(K_1 + 1, K_2) > P_{vd(QoS)}
\]

This leads to the contradiction again.

Therefore, we just show that it is impossible for another set of solution with \( K^*_1 \geq K_1 \) and \( K^*_2 > K_2 \) to exist.

[Case II] \( K^*_1 > K_1 \) and \( K^*_2 < K_2 \)

From Fig. 33, we have

\[
P_d(K^*_1, K^*_2) > P_d(K_1, K^*_2) > P_d(K_1, K_2 - 1) > P_{d(QoS)}
\]

This result contradicts the assumption that \((K^*_1, K^*_2)\) can satisfy the required data blocking probability. Hence, this case is impossible.
[Case III] $K_1^* > K_1$ and $K_2^* > K_2$

From Fig. 32, we derive the following results

$$P_{vb}(K_1^*, K_2^*) > P_{vb}(K_1, K_2^*) > P_{vb}(K_1, K_2)$$

This contradicts the assumption that $(K_1^*, K_2^*)$ can achieve better or at least no worse performance than $(K_1, K_2)$.

[Case IV] $K_1^* < K_1$ and $K_2^* \leq K_2$

This case contains two sub-cases.

1. $K_1 - K_1^* \leq K_2 - K_2^*$.

   From Fig. 33, we get,

   $$P_d(K_1^*, K_2^*) \geq P_d(K_1^* + (K_1 - K_1^*), K_2 + (K_1 - K_1^*))$$
   $$\geq P_d(K_1, K_2 - 1) > P_d(QoS)$$

   (4.16)

   This violates the assumption that $(K_1^*, K_2^*)$ can guarantee the QoS requirement.

2. $K_1 - K_1^* > K_2 - K_2^*$

   From Fig. 32, we have,

   $$P_{vb}(K_1^*, K_2^*) \geq P_{vb}(K_1^* + (K_2 - K_2^*), K + 2^* + (K_2 - K_2^*))$$
   $$\geq P_{vb}(K_1^* + (K_2 - K_2^*), K_2) \geq P_{vb}(K_1, K_2)$$

   (4.17)

   This also contradicts with the assumption that $(K_1^*, K_2^*)$ can achieve better or no worse performance than $(K_1, K_2)$.

[Case V] $K_1^* = K_1$

It can be divided into two sub-cases.
1. $K_1^* = K_1$ and $K_2^* > K_2$. Obviously

$$P_{ub}(K_1^*, K_2^*) > P_{ub}(K_1, K_2)$$

It contradicts the assumption that $(K_1^*, K_2^*)$ can achieve better or no worse performance than $(K_1, K_2)$.

2. $K_1^* = K_1$ and $K_2^* < K_2$. Obviously,

$$P_d(K_1^*, K_2^*) \geq P_d(K_1, K_2 - 1) > P_d(QoS)$$

This result contradicts to the assumption that $(K_1^*, K_2^*)$ can maintain the QoS requirements at a pre-defined level.

[Case VI] $K_2^* = K_2$

We can analyze it into two sub-cases,

1. $K_1^* > K_1$ and $K_2^* = K_2$. Obviously,

$$P_{ud}(K_1^*, K_2^*) \geq P_{ud}(K_1 + 1, K_2) > P_{ud}(QoS)$$

It contradicts the target QoS requirements.

2. $K_1^* < K_1$ and $K_2^* = K_2$. Obviously,

$$P_{ub}(K_1^*, K_2^*) > P_{ub}(K_1, K_2)$$

This violates the assumption that the $(K_1^*, K_2^*)$ can achieve better or no worse performance than $(K_1, K_2)$.

The previous analysis has proved that the algorithm is able to obtain the optimal solution.

### 4.5 Conclusions

In this chapter, we have proposed a new bandwidth allocation scheme called *dual-threshold reservation (DTR) scheme*, which is a natural extension from the
well-known Guarded Channel (GC) scheme used in cellular networks supporting voice traffic. The basic idea is to use two thresholds, one for reserving channels for voice handoffs, while the other is used to block data traffic into the network in order to preserve the voice performance in terms of handoff dropping and call blocking probabilities. We further develop an analytical model which can obtain DTR's performance by adopting an iterative technique. Results obtained from the analysis are used to illustrate the system tradeoff.

The key observation from the obtained results demonstrates the monotonic of the performance measures with respect to the selection of the two thresholds. This enables us to derive the corresponding $K_1$ and $K_2$ values under a given QoS requirement of voice dropping and blocking probabilities, similar to the way the single optimal threshold is determined in [31]. Based on the observation of the relationship between thresholds and system performance, we have proposed two algorithms to reduce the high computational complexity in finding the optimal thresholds.
CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In the next generation of networks, due to the existence of bursty data, multimedia traffic and high handoff rates, new challenges are posed in providing QoS support for a variety of applications. Channel allocation and management is of great importance given bandwidth is considered as the prime scarce resource in wireless networks.

In this thesis, first, an efficient adaptive call admission control scheme is proposed to eliminate the requirement of prior knowledge and information exchange among cells of most existing schemes. The novelty of this scheme lies in its ability in adopting a local on-line estimation periodically to obtain the traffic parameters and channel resource occupancy parameters which are required by most existing control scheme. This scheme is adaptive to the changing traffic, therefore it can guarantee the QoS requirements in terms of handoff dropping probability in the presence of bursty data and changing traffic. In addition, the local on-line estimation eliminates the requirements of prior knowledge of traffic parameters and reduces the signaling overhead in exchanging status information among cells, therefore, makes it simple and effective to be put into actual deployment. Simulations are performed to investigate the performance of the scheme under various traffic environments and control parameter inputs. It is shown from the simulations that this algorithm can guarantee the QoS requirements while eliminating the requirements of traffic parameters input and signaling overhead. The limitation of the algorithm is that the control algorithm still assumes that the call arrivals follow a Poisson process and call durations follow an exponential distribution. In the future work, other distributions such as the Pareto-distribution for data traffic and Hyper-Erlang distribution should be considered. In addition,
methods to deploy the algorithm in networks with multiple types of traffic should be investigated.

Another major part of the thesis discusses the DTR scheme, which is proposed for integrated voice/data wireless communication networks. The DTR scheme is a natural extension of the well-know GC policy. The DTR scheme aims to provide guaranteed QoS support in terms of voice handoff dropping probability, voice blocking probability and data loss probability. Based on the scheme, an analytical model is developed. In addition, for the DTR model, the selection of optimal thresholds is necessary to maximize the spectrum utilization while guaranteeing the QoS requirements. Based on the observation of the relationship between thresholds and system performance, we have proposed two algorithms to reduce the high computational complexity in finding the optimal thresholds. One limitation of the work is that it assumes no buffer is available in the system. While in some real systems, buffers are available and have great impact on the system performance. In future work, we would like to investigate the system performance under different threshold configurations.
REFERENCES


