UNIVERSITY OF CINCINNATI

March 16 , 20 01

I, Jingao Wang ,
hereby submit this as part of the requirements for the degree of:

Master of Science

in:

Computer Engineering

It is entitled:

Performance Analysis of Hybrid Reservation Handoff Scheme
for Integrated Wireless Mobile Network

Approved by:

Dharma P. Agrawal
Samir R. Das
Carla C. Purdy
Qingan Zeng

Performance Analysis of Hybrid Reservation Handoff Scheme for Integrated Wireless Mobile Network

A thesis submitted to the

Division of Research and Advanced Studies
of the University of Cincinnati

in partial fulfillment of the
requirements of the degree of

MASTER OF SCIENCE

In the Department of Electrical and Computer Engineering and Computer
Science of the College of Engineering
2001

by

Jingao Wang

B.S., Beijing Univerisity, China, 1995

Committee Chair: Dr. Dharma P. Agrawal

Abstract

The next generation of wireless mobile network has been designed to support a combination of both real-time service (such as voice, video transmission) and non real-time service (such as data transmission). To support such a wide range of traffic, Quality of Service (QoS) has to be taken into consideration while designing the infrastructure itself for the wireless mobile networks. One of the central issues characterizing the performance is how the handoff is handled. Two hybrid reservation handoff scheme, one is priority reservation and the other with preemptive procedure, for integrated real-time and non real-time service wireless mobile network have been proposed and analyzed in this dissertation. These handoff schemes are service dependent. The system is modeled by a multi-dimensional Markov chain and a numerical analysis is presented to estimate blocking probabilities of originating calls, forced termination probability of real-time service handoff requests calls, and average transmission delay of non real-time service calls. This scheme is also simulated using extensive runs and the results are observed to agree fairly closely. In the proposed scheme, forced termination probability of real-time service calls is seen to be significantly decreased. The probability of packet loss of non real-time transmission is made to be negligibly small, as a non real-time service handoff request in the queue can be transferred from the queue of current base station to another one.

Acknowledgments

First and foremost I would like to thank Dr. Dharma P. Agrawal for all of his invaluable advice and guidance. During the past two years, he always was helping me to get through the various problems in my research work. I am also very appreciated for the help that I received from Dr. Qing-An Zeng. He always gave me assistance as soon as possible whenever I have questions. Without his help, I couldn’t have continued the progress on my project so smoothly.

Acknowledgements must always be made to my committee members, Dr. Samir R. Das and Dr. Carla C. Perduy, who have read the draft of this dissertation thoroughly and gave me so many valuable suggestions.

It is also necessary to acknowledge to my friends and colleagues in center of distributed and mobile computing, who give me warmth and selfless help whenever I needed.

Last but not least, I want to express my deepest appreciation to my wife, Li Gao, who has always been there with me. Without her support and love, this work will be impossible. Thanks.

Jingao Wang

List of Figures

Chapter 1.

Introduction

1.1 Wireless Mobile Network

In the late 1970s, the first Advanced Mobile Phone Service (AMPS) test system was established in Chicago and Washington, D.C. Over the last two decades, the evolution of core radio and mobile network technologies, from Advanced Mobile Phone Service (AMPS), Global System for Mobile Communications (GSM) to EIA/TIA IS-95 Digital Cellular System (CDMA) and third-generation Mobile System (WCDMA, CDMA2000), made wireless network penetrating people's common life and enabled the development of the ubiquitous wireless mobile network that can provide the mobile user with integrated voice, data, and multimedia service at any time, any place, and in any format [1].

In order to overcome the restriction of limited radio frequency resources all of these systems are designed around the concepts of cells so that radio frequency can be reused (shown in Figure 1.1). Basically, the idea is to split geographic area into small cells and cover each geographical area by a low-power transmitter. Since the signals transmitted by the low-power base station are distance-limited, signals from the noncontiguous cells using the same frequency band do not interfere with each other. And hence frequency can be reused by cells as long as some minimum physical distance could be maintained by two cells using the same channel.

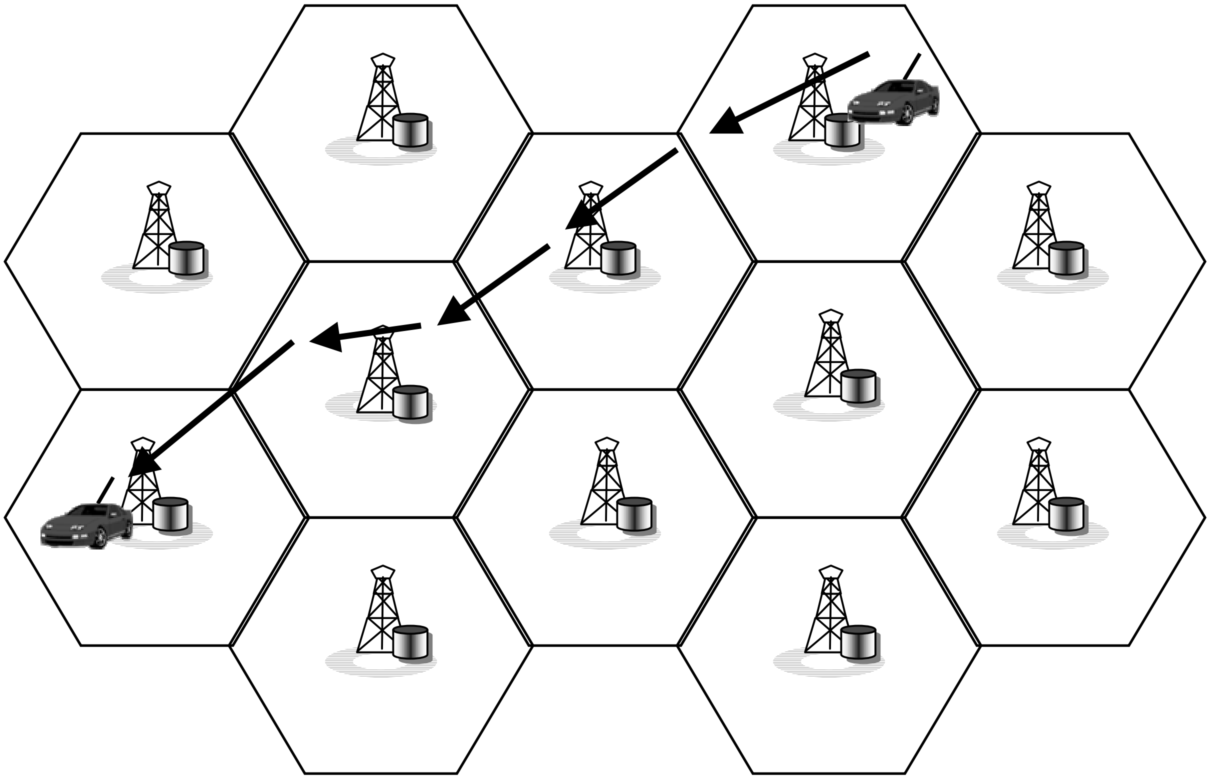


Figure 1.1 Cell and mobility support in a wireless network

1.2 Handoff Function in Wireless Mobile Network

1.2.1 Handoff Procedure

When a mobile user is engaged in conversation, the mobile station (MS) is connected to the base station (BS) via a radio link. If the mobile user moves to the coverage area of another BS, that means it moves out of current cell to another adjacent cell, the radio link to the prior BS has to be eventually disconnected and a new radio link to the new BS need to be established to continue the communication. Handoff represents such a process of changing the channel (frequency, time slot, spreading code, or combination of them) associated with the current connection while a call is still in progress [2]. To continue the communication, when an MS moves from one cell to another cell, the handoff procedure must be completed while the MS is in the overlap

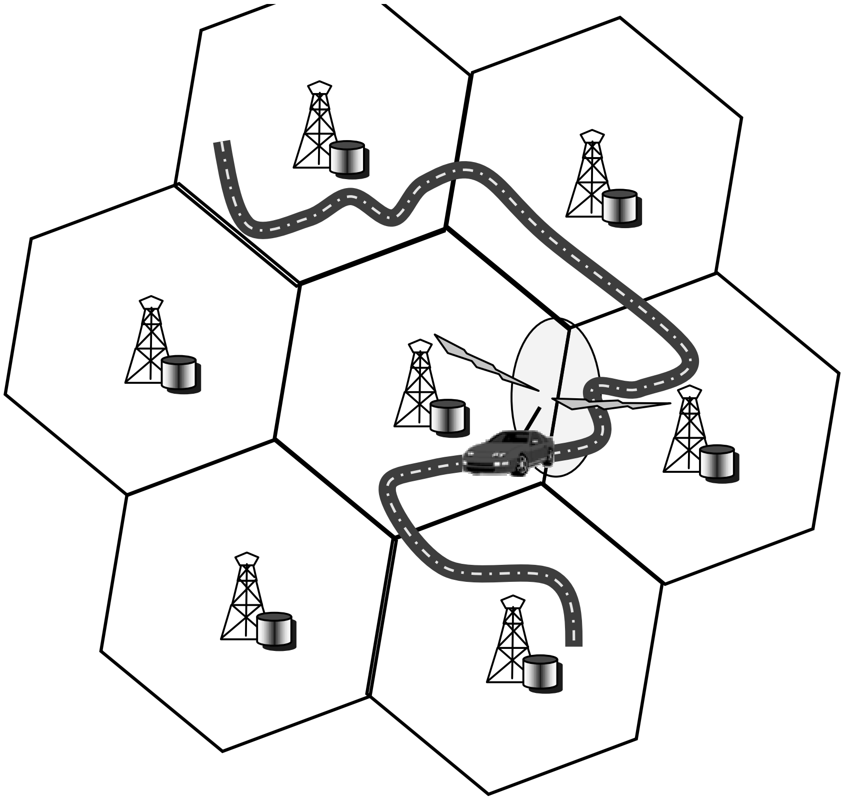


Figure 1.2 Handoff and handoff area

region. As the MS moves toward the edge of the cell, the signal strength and quality begin to deteriorate. At some point, the signal from a neighboring BS (the new BS) becomes stronger than the signal from the serving BS (the old BS). Additionally, the new BS receives a stronger signal from the MS than that received by the old BS. The radio connection needs to be handed over to the new BS before the link between the old BS and the MS becomes unusable. Otherwise, the call could be disconnected or dropped.

1.2.2 Handoff Detection

The handoff process usually consists of two phases. One is the handoff initialization phase and the other is the handoff execution phase.

To initiate a handoff, the quality of current communication channel is monitored in order to decide when to trigger the handoff. Handoff is a very rigorous process, so unnecessary handoffs should be avoided. If the handoff criteria are not chosen carefully, the call might be handed back and forth several times between two neighbor BS, especially when mobile user is moving around the overlapping region between the two BS coverage area boundaries. If the criteria are too conservative, then the call may be lost before the handoff can take place.

Handoff detection is based on measurements on the link status. The measurement process determines the need for handoff and the target new cell for transfer. Since the propagation condition between the BS and the MS is made up of the direct radio propagation paths (direct, reflection, refraction), various types of handoff initiating criteria have been proposed [3][4]. Such as:

- Word error indicator: Metric that indicates whether the current burst was demodulated properly in the MS.
- Received signal strength indication: Measure of received signal strength. The Received signal strength indicator metric has a large useful dynamic range, typically between 80 to 100 dB.
- Quality indicator: Estimate of the “eye opening” of a radio signal, which related to the signal to interference and noise ratio, including the effects of dispersion. Quality indicator has a narrow range (relating to the range of S/I ratio from 5 dB to 25 dB).

The handoff detection can be initiated either by MS (i.e., Mobile-controlled handoff) or by BS (i.e., Network-controlled handoff) or by both (i.e., Mobile-assisted handoff).

1.2.3 Handoff Execution

After the handoff detection phase decides to initiate handoff, the handoff execution phase begins. In handoff execution phase, the Message Switching Center (MSC) controlling the new BS assigns new channel (frequency, time slot, spreading code, or combination of them) to the handoff call. Channel assignment schemes attempt to archive a high degree of spectrum utilization for a given grade of service. There could be a failure in the handoff execution phase if:

- No channel is available on selected BS,
- Handoff is denied by the network for reasons such as the MS has exceeded some limit on the number of handoffs that may be attempted in a given period of time,
- It has taken too long for the network to set up the handoff after its initiation, or
- During the execution of handoff, the target link fails in some way [1].

When a mobile user initializes a call in current cell, an originating call is generated in the current cell. Since handoff requests and originating calls compete for radio resources, new call or handoff request could fail if a BS is too busy. The originating call attempts that fail because there are no channels available is called blocking call. The handoff requests for existing calls that must be turned down because there are no channels available are called forced terminations. Obviously, from user’s point of view, forced termination of an ongoing call is much more undesirable than blocking of new call attempts.

1.3 QoS in Next Generation Wireless Mobile Network

The next generation of wireless mobile network has been designed to support true convergence of both real-time service (such as voice, video transmission) and non real-time service (such as data traffic). In order to support such a wide range of traffic, the Quality of Service (QoS) has to be taken into consideration during the design phase of the wireless mobile network infrastructure itself.

From the viewpoint of end users, QoS should be provided based on an end-to-end basis. QoS attributes should be general enough but simple. From the viewpoint of the network, QoS can be defined with a set of parameters. The classes QoS defined in wireless mobile network are very different from fixed networks due to the restrictions and limitations of the air interface [1]. Based on delay sensitivity, two QoS classes have been defined for next generation wireless mobile network. Real-time service is defined for the most delay-sensitive applications, such as the traditional voice call, VoIP, video-on-demand, etc. The real-time service needs better channel coding, retransmission scheme and handoff design to meet the required QoS. Non real-time service is defined for delay-insensitive applications, such as Telnet, e-mail, FTP, etc. These services can tolerate some degree of transmission delays without affecting the QoS substantially. Several QoS parameters have been defined in 3G TR 23.907, including maximum, minimum, and guaranteed bit rates, delivery order, maximum packet size and reliability [1][5].

One of the central issues in the performance characterization is how the handoff is handled. In the design of a general handoff scheme, it is expected that the blocking probability for calls originated in a cell need to be minimized. However, from the user's point of view, traffic with service of handoff request is more important, as forced termination of an ongoing call is much more disastrous than blocking of new calls. On the other hand, attempt should be made to decrease transmission delay of non real-time service calls and maximize channel utilization. Unfortunately, all simultaneous requirements cannot be optimized and there are some tradeoffs between various parameters. From the reliability of transmission viewpoint, minimization of forced termination probability of real-time service and the average transmission delay of non real-time service need to be paid more attention.

Poorly designed handoff schemes tend to generate very heavy signaling traffic and thereby, a dramatic decrease in QoS. With the penetration of next generation wireless mobile network and personal communication system, the micro cell and hybrid cell (macro-, micro-, pico-) structures are being exploited to support the drastically increased demand. However, smaller the size of cell is, more frequent handoffs will be due to the variable propagation condition of signals [6]. Therefore, the handoff strategy should be given a careful consideration in the next generation wireless mobile networks.

It should be observed that the focus of this dissertation is about the handoff execution phase, and it is assumed that the handoff request detection and initiation procedures are perfect for real-time application (i.e., all valid requests are detected and no invalid requests activate the handoff procedure).

1.4 Organization of the Dissertation

This paper is organized as following: Chapter 2 gives a brief introduction of previous work. Chapter 3 proposes two hybrid reservation handoff schemes. Chapter 4 presents the analytical model of hybrid reservation handoff schemes, starting from traffic analytical model, the analytical model for proposed scheme has been obtained using Multi-dimension Markov Chain. Simulation and numerical results are presented in Chapter 5. Finally, Chapter 6 concludes the thesis, with some discussion on potential future work.

Chapter 2.

Related Work

The study of handoff is not a new topic to the wireless communication world. In order to provide better service for mobile user with limited frequency spectrum resource, various types of handoff scheme have been proposed. A brief introduction about those handoff schemes will be made in this chapter.

2.1 Handoff Scheme in Voice Cellular Network

2.1.1 Non-prioritized Handoff Scheme

A non-prioritized handoff scheme is employed by most first generation wireless mobile network and is simplest one to implement. In this scheme the BS handles a handoff request call in exactly the same manner as a new call. That is, the handoff request call is blocked immediately if no channels are available. Since no priority is given to handoff request calls over originating calls, the forced termination probability is relatively higher than normally anticipated [1].

2.1.2 Guard Channel Handoff Scheme

The use of guard channels for handoff has been commonly employed by voice cellular networks. Guard channel handoff scheme is similar to non-prioritized handoff scheme except that a number of channels in each BS are exclusively reserved for handoff request calls. Therefore, the total number of channels are divided into two groups: the normal channels, which serve both originating call and handoff request calls, and the reserved channels, which serves handoff request calls exclusively. By this way, there is a built in priority for handoff request call over originating call as long as channels within Guard channels are still available. The system performance is better than non-prioritized handoff scheme.

2.1.3 Queueing Priority Handoff Scheme

The queueing priority handoff scheme is based on the fact that there are overlapped areas between adjacent cells in a wireless mobile network. This area is called the handoff area where a cell can be handled by either BS of adjacent cells. The time that an MS spends in the handoff area is referred as the handoff area dwell time. In queueing priority handoff scheme, each BS have one or more queueing buffers for all incoming calls. When a call arrives the BS, it will check whether there is channel available. The call can be serviced immediately if there is channel available in BS. However, even if no channel is available when a call arrives, the call will not be blocked or dropped as far as there is free space in the queue for this kind of service. The incoming call is kept put in the queue to wait for the next channel available. Whenever a channel is released, the BS first check whether there are any waiting calls in the queue. If there is, the released channel is assigned to a waiting call in the queue and is usually done on a first-in-first-out basis from the queue.

In queueing priority handoff scheme, there are two issues that is of major concern. How many queues should BS have and which kind of service call should be included in to the queue? If handoff request call can be queued, the queued handoff request calls can keep communication with the old BS as long as it is still in handoff area, so that the forced termination probability can be decreased. If originating call can be queued, the blocking probability of originating call can be decreased. In [7], system with queues only for voice handoff requests are studied. In [8], queues are allowed only for the originating voice calls. Both the originating calls and handoff requests are allowed to be queued in [9]. In [10][11], the handoff schemes with two-level priority reservation have been proposed. Cellular communication systems that support a mixture type of platform are considered in [12][13]. However, in all of the above studies, only cut-off priority for real-time user is considered while the multiple types of services have not been taken into consideration.

2.2 Handoff Scheme in Integrated Service Wireless Network

With the development of wireless mobile system, non real-time service has to be incorporated and taken into consideration [14]. Thus, a handoff scheme in non real-time wireless network has been studied in [15]. However, future wireless networks will be required to support multiple types of services, such as voice, video and data traffic. In order to meet the future demands, the handoff strategy needs to take different features of these services into account, which implies that the ideal handoff processes is service-dependent. For example, real-time service transmission is very sensitive to interruptions. On the other hand, transmission delay of non real-time service does not have much impact on the non real-time service performance (delay insensitive). Therefore, a successful handoff without interruption is very important to real-time service, but not so critical to non real-time service.

2.2.1 Two Dimensional Model

F.-N. Pavlidou proposed a special two-dimensional model for real-time cellular mobile systems [16]. This model gave preemptive priority to real-time service calls. However, in this model, no distinction is made between originating real-time service calls and handoff requests. Since, from the user's point of view, forced termination of ongoing real-time service calls is more annoying than blocking of originating calls, much more priority should be given to real-time service handoff calls.

2.2.2 Priority Reservation Handoff

Q. A. Zeng proposed a handoff scheme in real-time cellular systems with priority reservation for real-time service handoff requests in [17]. In this scheme, some number of channels is reserved for real-time service handoff requests. Queues are allowed for real-time service handoff requests and non real-time service handoff requests. Moreover, a non real-time service handoff request in the queue can be transferred to another queue in an adjacent cell when the mobile user moves out of cell before getting a channel.

2.2.3 Priority Reservation with preemptive priority Handoff

In [18], Q. A. Zeng proposed a number reservation handoff scheme with preemptive procedure. In this scheme, calls are divided into three different classes: originating calls, real-time service handoff calls, and non real-time service handoff calls.

The real-time and non real-time service handoff requests make their own queues. A given number of channels in each cell are reserved exclusively for handoff request calls. Out of these channels, some are reserved exclusively for the real-time service handoff request calls. The remaining channels are shared by both the originating and handoff request calls. The real-time service handoff requests have priority over non real-time service handoff requests and all the handoff requests have priority over originating calls. The right to preempt the service of non real-time service is given to real-time service handoff calls if no channels available when real-time service handoff call on arrival. The interrupted non real-time service call will be returned to the last position of the non real-time service queue and waits for some channels to be available.

These handoff schemes are helpful to reduce the forced termination probability of real-time service handoff requests. However, if the traffic load is heavy, the non real-time service calls (both originating and handoff calls accepted by the system) may be keeping waiting or being preempted due to its lower priority. It will lead to the starvation problem for non real-time service requests. What is more important is that all of these handoff schemes are based on the number of reserved channels. All the channels are open to every class of calls, but the maximum number of channels used by lower priority class of calls ought to be limited. There is another important reservation method in wireless mobile system, which is the channel priority reservation. Every class of calls has a part of channels reserved exclusively for their use. When a call arrives, it only checks for its own class of channels and sees whether there is an idle channel or not. The implementation of channel reservation handoff scheme is simpler and faster than that of number reservation handoff scheme. To get a comprehensive result for handoff scheme design, the service-depended channel priority reservation handoff scheme ought to be studied.

Chapter 3.

Hybrid Reservation Queueing Handoff Scheme

In this chapter, two hybrid reservation queueing handoff schemes for integrated service wireless mobile network are proposed: One is priority reservation only and the other is with preemptive procedure.

3.1 System Model

For simplicity, we consider a system with many homogenous cells with a fixed channel assignment scheme and a set of S channels is permanently assigned to each cell. In such a homogenous system, we focus our attention on a single cell, which we call as the referenced cell in the paper. When a mobile user generates a call in the referenced cell, we donate it as an originating call. When a real-time service mobile user holding a channel approaches toward the referenced cell from a neighboring cell and enters the handoff area of the referenced cell, a handoff request of real-time service is generated in the base station of the referenced cell. There is no handoff area for non real-time service mobile users. Instead, we use the cell boundary, which is defined by the points where the received signal strength between two adjacent cells is equal. When a mobile user holding a channel approaches the referenced cell and crosses the cell boundary, a handoff request is generated.

The system model for the referenced cell is shown in Figure 3.1. In the figure, the following notations are used:

λ_{OR}: The arrival rate of originating real-time service calls;

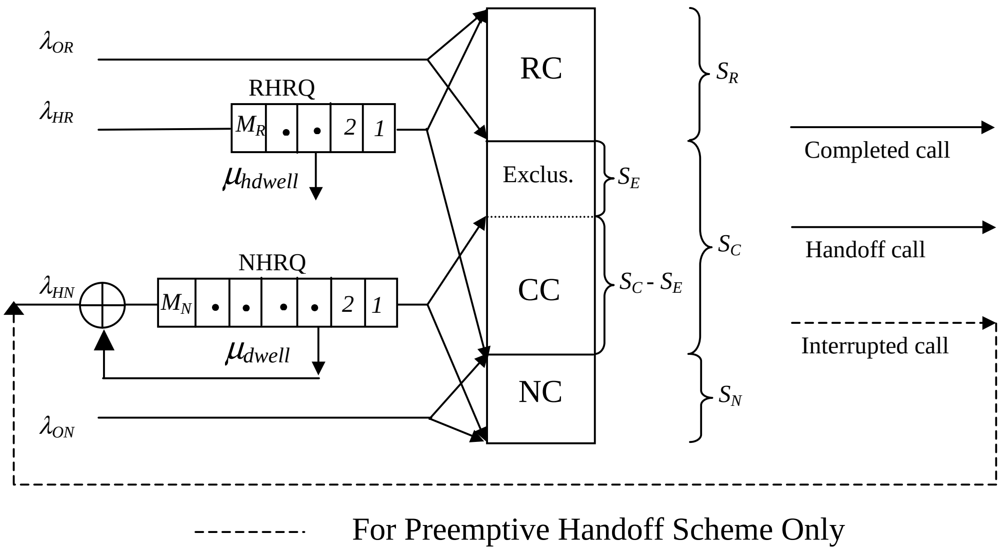


Figure 3.1. System model

λ_{HR}: The arrival rate of handoff request calls of real-time service;

λ_{ON}: The arrival rate of originating non real-time service calls;

λ_{HN}: The arrival rate of handoff request calls of non real-time service;

μ_{hdwell}: The dropping rate of the queueing real-time service handoff requests;

μ_{dwell}: The transferring rate of the queueing non real-time service handoff requests.

The total channels S of referenced cell are divided into three sections, namely, Real-time service Channel (RC) section, Non real-time service Channel (NC) section, and Common Channel (CC) section with the channel capacity S_R, S_N, and S_C respectively. RC are reserved for real-time service calls (including both real-time originating and handoff request calls) only and NC are reserved for non real-time service call (including both non real-time originating and handoff request calls). CC are reserved for the overflow from the real-time service or the non real-time service handoff request calls from RC or CC respectively. Out of CC, a certain number of channels is reserved exclusively for the real-time service handoff request calls so that real-time service handoff requests are given higher priority over non real-time service. We use S_E to mark the number of free channel reserved before non real-time service handoff request calls could be served in S_C. There are two queues in the referenced cell, real-time service handoff request queue (RHRQ) and non real-time service handoff request queue (NHRQ). RHRQ with finite capacity M_R for real-time service handoff request calls and NHRQ with finite capacity M_N for non real-time service handoff request calls. The originating service calls do not have their own queue. The departuring rate of serving channels (RC, CC and NC) are marked as the rate of completed calls, the rate of handoff calls (and the the rate of preemptive calls in the preemptive hybrid channel reservation handoff shceme).

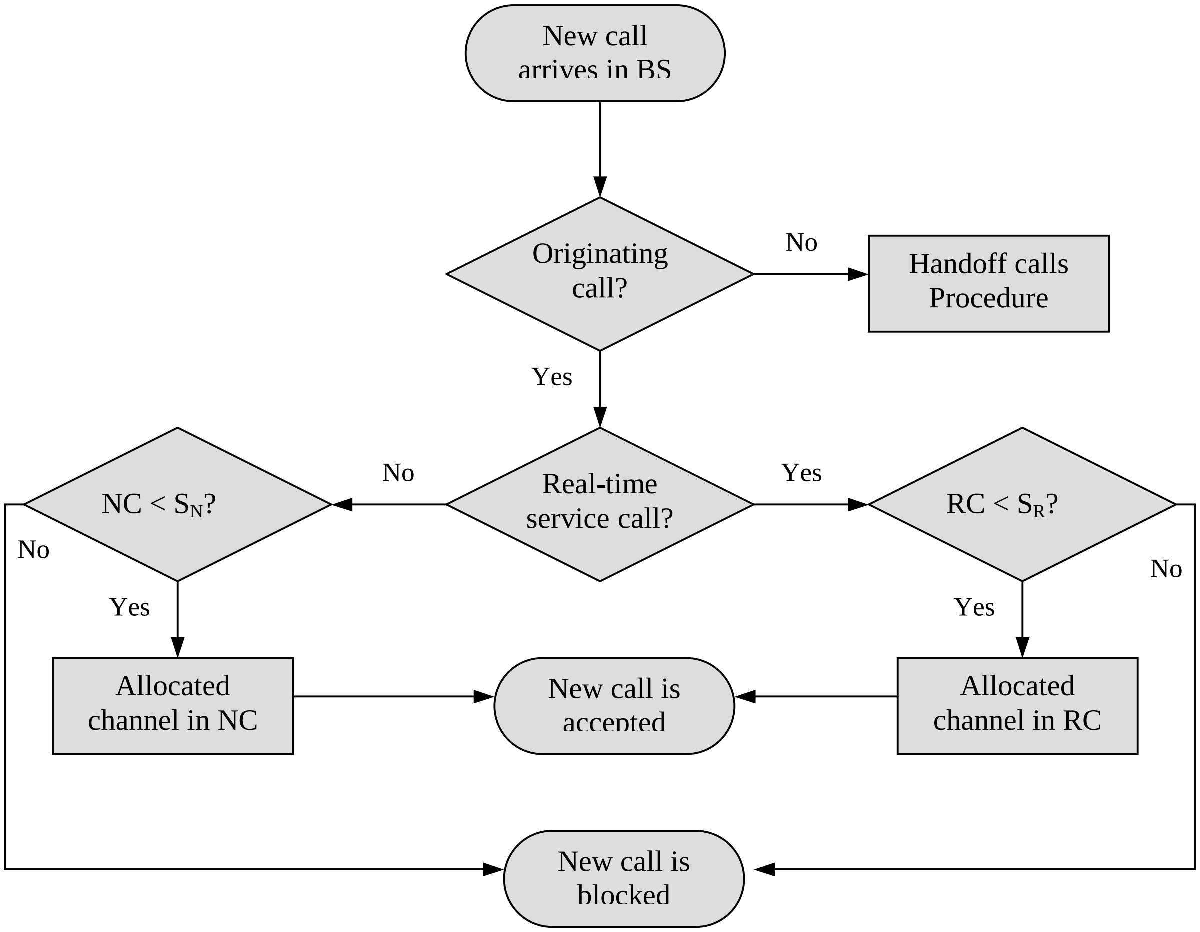
3.2 Priority Hybrid Reservation Queueing handoff Scheme

When an originating real-time service request arrives, it can be served only if there are channels available in RC. Similarly, an originating non real-time service request can be served only if there are idle channels in NC. An originating real-time service call (or an originating non real-time service call) will be blocked if it finds no channels in the RC (or NC). Figure 3.2 shows the flow diagram of handling call in priority hybrid channel reservation scheme.

When a real-time service mobile user holding a channel approaches toward the referenced cell from a neighboring cell and enters the handoff area of the referenced cell, a handoff request of real-time service is generated in the base station of the referenced cell. The real-time service handoff request call will be served if there are idle channels in

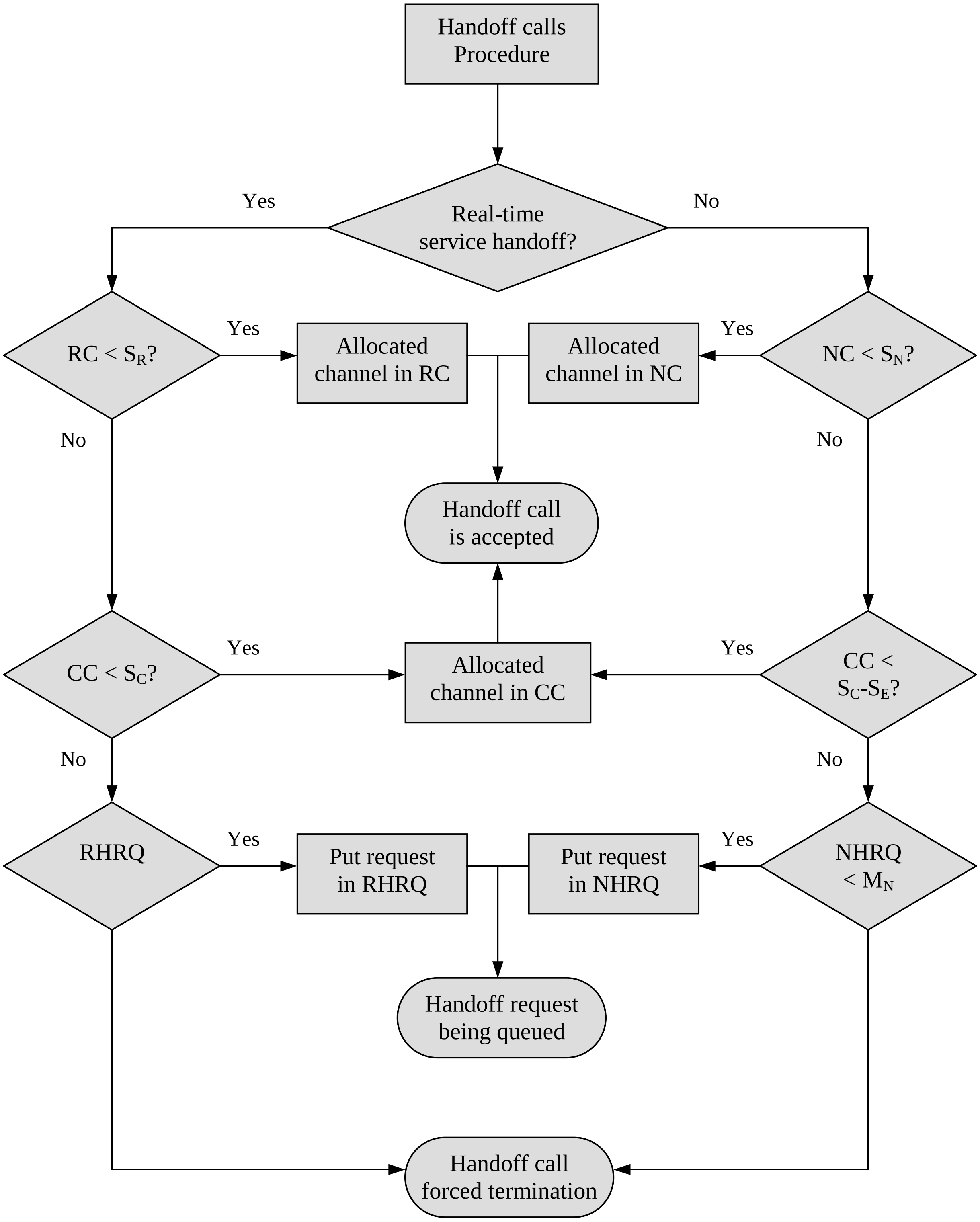
RC on arrival. If RC is full on arrival, it will check whether there are idle channels in CC. A real-time service handoff request call is queued in RHRQ when it finds there is no available channel in both RC and CC. If RHRQ is full, it will be blocked. The real-time service handoff requests waiting in the queue could be dropped if the mobile user moves out of the handoff area before getting service. Allowing waiting time for the real-time service handoff requests in RHRQ is the dwell time of real-time service handoff requests in the handoff area. A blocked real-time service handoff request call will still keep the communication via the current base station until the received signal strength goes below the receiving threshold. However, in this dissertation, we neglect the probability that a blocked real-time service handoff request call completes the communication before the received signal strength goes below the threshold of received signals. A blocked handoff request can repeat trial handoff until the received signal strength goes below the threshold. However, in the proposed schemes, we do not consider repeated trials of blocked handoff requests.

When a non real-time service mobile user holding a channel enters the referenced cell and the strength of the received signal of the current base station goes below that of the base station of the referenced cell, a handoff request of non real-time service is generated. The non real-time service handoff request call is served in NC if there is any channel available in NC. If NC is full, it can be served in CC if the number of idles channels in CC are larger than S_E. Otherwise, it is put into NHRQ. A non real-time service handoff request is blocked on arrival if NHRQ is full. The non real-time service handoff requests in NHRQ can be transferred from the referenced cell to one of the target cells when the mobile user moves out of the referenced cell before it gets service.



(a)

Figure 3.2. (a) A flow diagram in priority hybrid reservation handoff scheme for originating calls



(b)

Figure 3.2. (b) A flow diagram in priority hybrid reservation handoff scheme for handoff request calls

Therefore, if the size of NHRQ is to be large enough to hold all the non real-time handoff requests, the forced termination of non real-time service would never happen. Indeed, we do not need an infinite buffer to hold the non real-time service handoff request queue. The maximum possible waiting time of non real-time service handoff request in NHRQ is just the dwell time of non real-time service subscribers in the referenced cell, even if the non real-time service user does not get service in the referenced cell.

3.3 Preemptive Hybrid Reservation Queueing Handoff Scheme

In preemptive hybrid reservation handoff scheme, the handling of coming originating real-time service calls, originating non real-time service calls, and non real-time service handoff request calls are same as the previous priority reservation handoff scheme arrives (as we have introduced in the previous section). The difference between preemptive hybrid channel reservation handoff scheme and the priority hybrid channel reservation handoff scheme is in the procedure for real-time service handoff request calls. A flow diagram of handling handoff request calls in preemptive hybrid channel reservation scheme is shown in Figure 3.3.

When a real-time service user holding a channel enters the handoff area of the referenced cell, a handoff request of real-time service is generated. The real-time service handoff request call can be served if there is any channel available in RC or CC. If both RC and CC have idle channels, the free channel in RC will be assigned to the handoff request call firstly. However, even there is no idle channel in either RC or CC, the real-time service handoff request calls can still be served the service of non real-time handoff request call in CC can be preempted (we assume there is ongoing non real-time service call in CC and NHRQ is not full). The interrupted non real-time service call returns back to NHRQ and waits for an idle channel to be served based on first-in-first-out rule. A real-time service handoff request is queued in RHRQ when all the channels of RC and CC are occupied by prior calls and NHRQ is full. This means that non real-time service calls in service can not be preempted by real-time service handoff calls when NHRQ is full. We could possibly consider preempting non real-time service calls irrespective of non real-time service handoff queue being full or not. However, this does not cause any difference with a very large non real-time service handoff queue. The real-time service handoff requests waiting in RHRQ could be removed when the mobile user moves out of handoff area even if the mobile user does not get service in the handoff area. The maximum allowed waiting time for the real-time service handoff requests in RHRQ is the dwell time of real-time service handoff requests in the handoff area and is discussed in the next chapter.

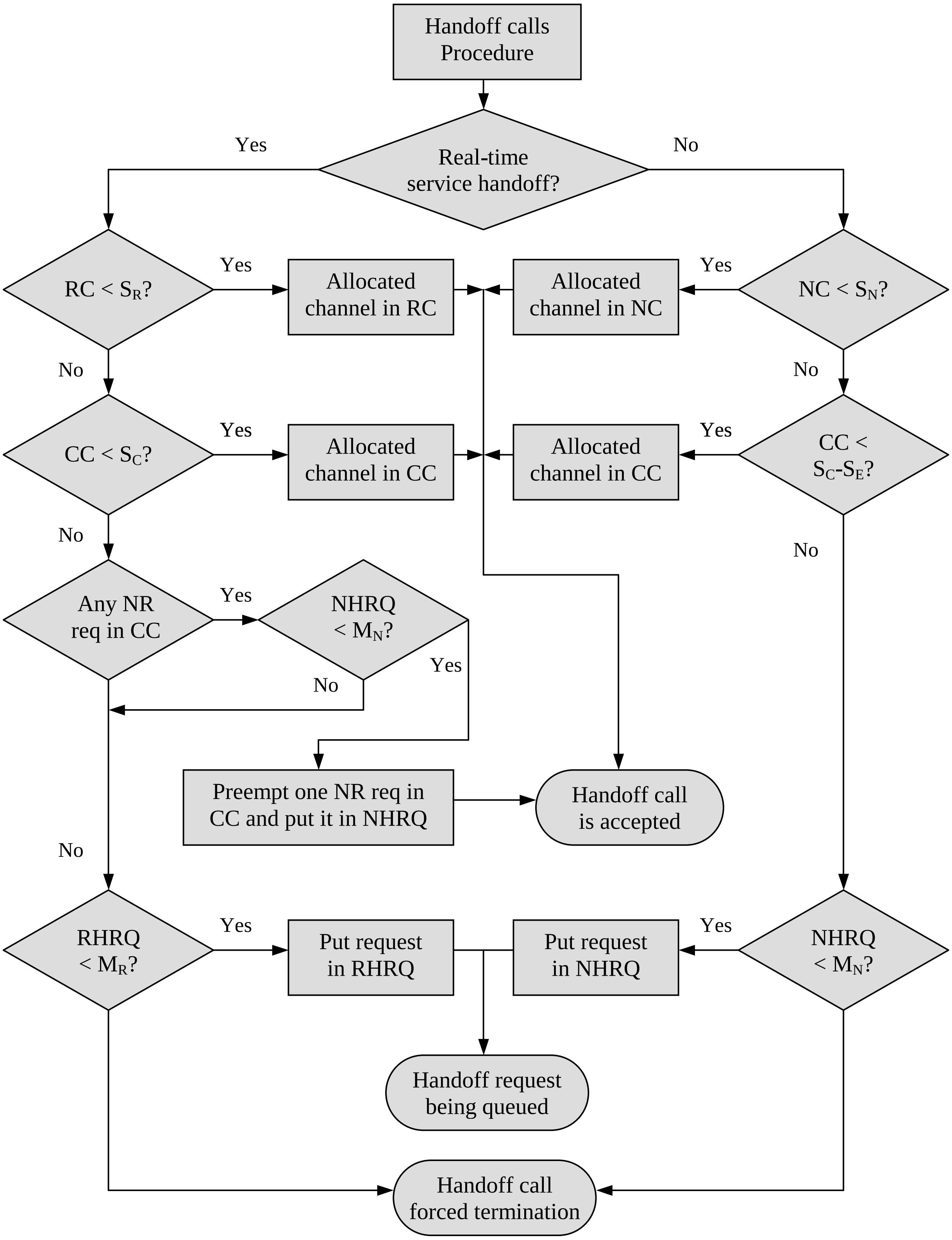


Figure 3.3. A flow diagram of handoff request calls in preemptive hybrid reservation handoff scheme

Chapter 4.

Analytical Model

4.1. Traffic Model

4.1.1. Cell Dwell Time

We choose the fluid flow model [19] as the mobility model of mobile users. However, our proposed method can be easily used to other mobility models as well. The model assumes a uniform density of users throughout the area and also assumes that a user is equally likely to move in any direction with respect to the cell boundary. Let $f_V(v)$ be the probability density function (pdf) of the random variable V of the speed of all mobile users in the following section.

We assume that both the real-time and non real-time service subscribers have the same probability distributions of moving speed. Let $f_V(v)$ be the probability density function (pdf) of the random variable V with mean $E[V]$ being the speed of mobile users. For two-dimensional fluid flow model, the average outgoing rate μ_{dwell} is given by

$$\mu_{dwell} = \frac{E[V]L}{\pi A}, \tag{4.1}$$

where L is the length of the perimeter of a cell with arbitrary shape and A is the area of the cell. We assume that the cell dwell time T_{dwell} has a random exponential distribution with mean $1/\mu_{dwell}$, then the average cell dwell time is given by

$$E[T_{dwell}] = \frac{\pi A}{E[V]L}. \tag{4.2}$$

4.1.2. Handoff Area Dwell Time

A real-time service handoff request calls is going to be put in the RHRQ if the BS in the target cell can allocate channel resource to it. If a channel is assigned to the request during its stay in the handoff area, the handoff is successful. If received signal strength reaches the receiver threshold prior to the assignment of a channel (that means mobile user moves out the handoff area before it gets channel resource), the call is forced to terminate. The maximum allowable waiting time of a real-time service handoff request in RHRQ is equal to the dwell time in the handoff area. In the following, we find the average dwell time of a handoff call in the handoff area.

We consider the new random variable V^*, which is the speed of mobile users crossing the cell's boundary. As is mentioned in [23], the pdf $f_{V^*}(v)$ is different from the above pdfs $f_V(v)$. $f_{V^*}(v)$ is given by

$$f_{V^*}(v) = \frac{v f_V(v)}{E[V]}. \tag{4.3}$$

where E[V] is the average of the random variable V.

The random variable T_h is defined as the time spent in the handoff area of real-time service mobile user. It is given by

$$T_h = \frac{D}{V^*}, \tag{4.4}$$

where random variable D is the length of moving path of the mobile users in handoff area.

From Equation 4.3, we have

$$E[\frac{1}{V_H^*}] = \int_0^\infty \frac{1}{v} f_{V_H^*}(v)dv = \int_0^\infty \frac{f_{V_H^*}(v)}{E[V_H]} dv = \frac{1}{E[V_H]}. \tag{4.5}$$

Assuming that the path length and velocity of mobile users are independent, we can get the average dwell time $E[T_h]$ using equations 4.4 and 4.5 as

$$E[T_h] = E[\frac{D}{V*}] = E[\frac{1}{V*}]E[D] = \frac{E[D]}{E[V*]}, \tag{4.6}$$

The random variable T_h is assumed to have an exponential distribution with the above average. Therefore, we use μ_{hdwell} to present the leaving rate of mobile user of handoff area and the average value of μ_{hdwell} is given by

$$E[\mu_{hdwell}] = E\frac{1}{E[T_h]}. \tag{4.7}$$

4.1.3. Arrival Process of Service Calls

We assume that the arrival processes of the originating real-time service and the non real-time service call in a cell are Poisson. The arrival rates of originating real-time service and non real-time service calls are donated by λ_{OR} and λ_{ON}, respectively.

For the arrival rate of real-time service handoff requests λ_{HR}, we can not arbitrarily assume that it does not relate to λ_{OR}. We have to compute it by the following method. If the call holding time of a mobile user is greater than the dwell time in the cell, a handoff request is initialized in the neighboring cell. Since we assume an equilibrium homogeneous mobility pattern, the mean number of incoming users into a cell is equal to that of outgoing ones from the cell. Therefore, the arrival rate of real-time service handoff request calls in the referenced cell is equal to the departure rate of real-time service handoff calls from the cell. For the outgoing rate λ_{OUTR} of real-time handoff request in a referenced cell, we have

$$\lambda_{OUT\,R} = E[C_R]\mu_{dwell} \quad , \tag{4.8}$$

where $E[C_R]$ is the average number of real-time service calls holding channels in a cell. Therefore, the average arrival rate of real-time service handoff requests λ_{HR} is given by

$$\lambda_{HR} = \lambda_{OUT\,R} = E[C_R]\mu_{dwell}\,. \tag{4.9}$$

Similarly, for the arrival rate of non real-time service handoff requests λ_{HN}, we can assume it as an arbitrary value neither. The arrival rate of non real-time service handoff request calls in the referenced cell is equal to the departure rate of the non real-time service handoff calls from the cell. For non real-time service calls, not only the ongoing non real-time service call but also the request waiting in the NHRQ can cause handoff requests in neighbor cells. If the rate of calls going out of a cell without completing communication for non real-time service users, is $E[C_N]\mu_{dwell}$, where $E[C_N]$ is the average number of non real-time service users holding channels in a cell. Therefore, the first part of the handoff rate of non real-time service handoff request calls λ_{HN1}, which is caused by ongoing non real-time service call, is given by

$$\lambda_{HN1} = E[C_N]\mu_{dwell}\,. \tag{4.10}$$

The second part of part of the handoff rate of non real-time service handoff request calls λ_{HN2} is caused by waiting non real-time service request in NRHQ. The non real-time service handoff request in the queue of the current cell is transferred to the queue of target cell from when the non real-time service mobile user moves out of the cell before getting a channel. Thus, it given by

$$\lambda_{HN2} = E[L_N]\mu_{dwell}, \tag{4.11}$$

where $E[L_N]$ is the average length of NHRQ. We approximate the arrival process of each type of calls above by a Poisson process with above rate. Therefore, we can get the arrival rate of non real-time service handoff requests λ_{HN} by

$$\lambda_{HN} = E[N_N]\mu_{dwell}, \tag{4.12}$$

where $E[N_N]$ is the average number of both non real-time service requests and calls in each cell,

$$E[N_N] = E[C_N] + E[L_N]. \tag{4.13}$$

4.1.4. Channel Holding Time

A channel held by a user will be released by either of the following reasons, that is, by the completion of the conversation or by the handing off the call to a neighboring cell (or failure of it, i.e., forced termination). Thus, the channel service is decomposed into two rates.

We assume that the call holding time T_{CR} of real-time service calls has an exponential distribution with mean

$$E[T_{CR}] = \frac{1}{\mu_{CR}}. \tag{4.14}$$

Therefore, the channel holding time T_R of a real-time service call is equal to the smaller one between T_{dwell} and T_{CR}. Since the random variable T_{dwell} and T_{CR} have exponential distributions with mean $1/\mu_{dwell}$ and $1/\mu_{CR}$, the average channel holding time of real-time service call can be given by the memory-less property of the exponential pdfs

$$E[T_R] = \frac{1}{\mu_C} = \frac{1}{\mu_{CR} + \mu_{dwell}}, \tag{4.15}$$

Similarly, We assume that the call holding time T_{CN} of non real-time service calls has an exponential distribution with mean

$$E[T_{CN}] = \frac{1}{\mu_{CN}}, \tag{4.16}$$

and we can get the average channel holding time of non real-time service call

$$E[T_N] = \frac{1}{\mu_C} = \frac{1}{\mu_{CN} + \mu_{dwell}}, \tag{4.17}$$

And the channel holding time T_R and T_N are exponentially distributed with mean $E[T_R]$ and $E[T_N]$ respectively.

4.2 Performance Analysis

4.2.1. Multi-dimension Markov Chain

Queueing theory is the key analytical modeling technique used for handoff procedure performance analysis. Let's give some definitions in Queueing theory to help analyzing large system.

Stochastic Chain: A random function of time or sequences, in which the number of possible value of states if finite or countable, is called a discrete-state process. A discrete-state process is also called a Stochastic Chain [20].

Markov Chain: In the future states of a stochastic chain are independent of the past and depend only on the present, the stochastic chain is called a Markov Chain [20].

Birth-Death Markov Chain: A Markov Chain in which the transitions are restricted to neighboring states only is called Birth-Death Markov Chain [20].

The priority handoff scheme and preemptive handoff scheme queueing system can be specified by following characters:

- Arrival Process is given by Arrival Process of Service Calls in Chapter 4.1.3
- Service Time Distribution is given by Channel Holding Time in Chapter 4.1.4
- Number of Server is the number of total channels in RC, NC and CC
- System Capacity is the sum of number of channels and capacity of RHRQ and NHRQ.
- Population Size is the number of mobile users, we assume it is a very large number.
- Service Discipline is described in System Model in Chapter 3.2 and Chapter 3.3.

Based on those characteristics of our handoff scheme, the handoff system can be presented as a Multi-dimension Birth-Death Markov Chain. The state of the referenced cell can be marked as five-tuple of non-negative integers (i, j, k, l, m), where

i is the number of channels used by real-time service calls (including originating calls and handoff request calls) in RC,

j is the sum of the number of real-time service handoff request calls in CC and RHRQ,

k is the number of non real-time service handoff request calls in CC,

l is the number of non real-time service calls (including originating calls and handoff request calls) in NC,

m is the number of non real-time service handoff requests waiting in NHRQ.

The value ranges of discrete parameters are

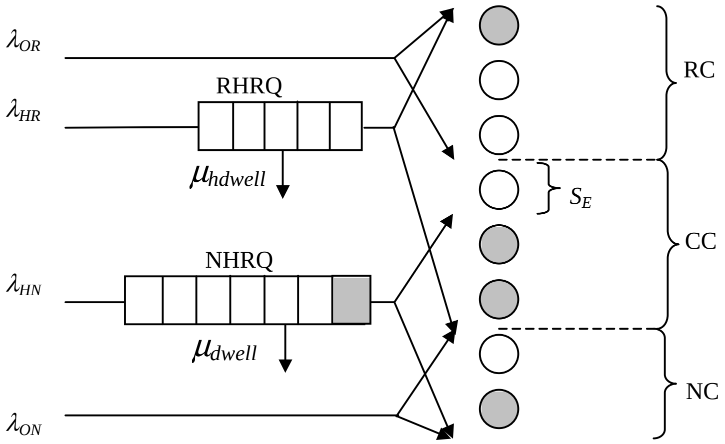
$i \in [0, S_R]$,

$j \in [0, S_C+M_R]$,

$k \in [0, S_C - S_E]$,

$l \in [0, S_N]$,

$m \in [0, M_N]$.



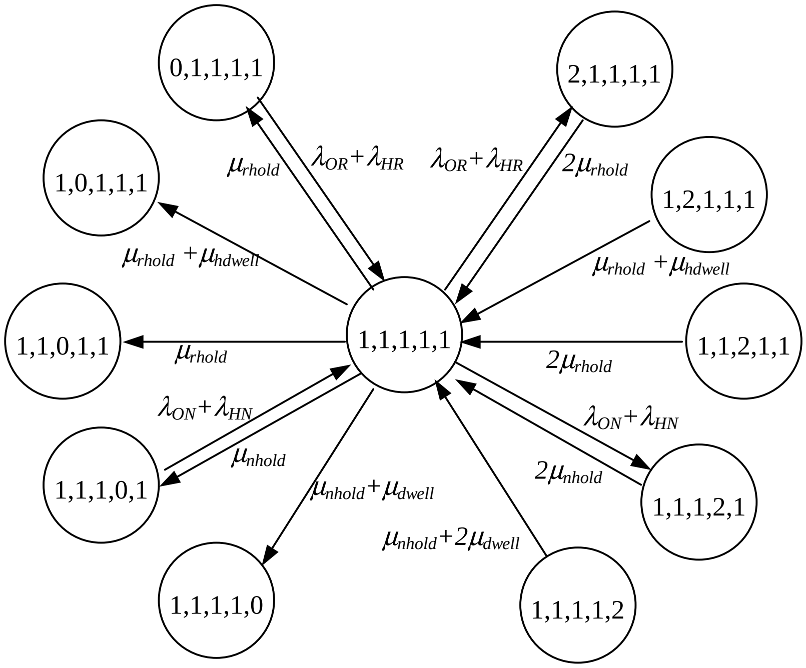


Figure 4.1. State transition diagram for (1, 1, 1, 1, 1)

The total number of states in Markov chain is

$$\begin{aligned} N_T &= M_N \left(S_C - S_E + 1\right) \left(M_R + S_R S_E + S_R + S_E + 1\right) \\ &+ \frac{1}{2} \left(S_N + 1\right) \left(S_C - S_E + 1\right) \left(2M_R + \left(S_R + 1\right) \left(S_C + S_E + 2\right)\right). \end{aligned} \tag{4.18}$$

As an example, the state transition diagram of state (1, 1, 1, 1, 1) is given in Figure 4.1.

We can get N_T balance equations through state transition diagram. Equilibrium probabilities *P(i, j, k, l, m)*s are related to each other through the state balance equations. However, notice that any one of these balance equations can be obtained from other $N_T - 1$ equations because we have that the sum of all state probabilities is equal to 1, i.e., normalizing condition. The normalizing equation is given by

$$\begin{aligned} &\sum_{i=0}^{S_R} \sum_{k=0}^{S_C - S_E} \sum_{j=0}^{S_C - k} \sum_{l=0}^{S_N} P(i, j, k, l, 0) + \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - k + 1}^{S_C + M_R - k} \sum_{l=0}^{S_N} P(S_R, j, k, l, 0) \\ &+ \sum_{i=0}^{S_R} \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - S_E - k}^{S_C - k} \sum_{m=1}^{M_N} P(i, j, k, S_N, m) + \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - k + 1}^{S_C + M_R - k} \sum_{m=1}^{M_N} P(S_R, j, k, S_N, m) = 1. \end{aligned} \tag{4.19}$$

Adding the normalizing equation (11), we can obtain N_T independent equations. In these equations, λ_{HR} and λ_{HN} are two unknown variables. Considering the relations (6) and (7), we can get N_T+2 nonlinear independent simultaneous equations for N_T+2 unknown variables.

4.2.2 SOR and Iteration method

As N_T is rather a large number, we use the Successive Over-Relaxation (SOR) iteration method [21] in the form of pseudo-code to solve N_T+2 independent nonlinear equations and compute all the state probabilities *P(i, j, k, l, m)*s.

Step1: Select arbitrary initial (positive) values for λ_{HR} and $\lambda_{HN.}$

Step2: Compute all the probability P(i, j, k, l m)s using SOR method.

Step3: Compute all average numbers of real-time service calls holding channels E[C_R] and non real-time service waiting and holding calls E[N_N] using the following formulas:

where

$$E[C_R] = \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=0}^{S_C-k} (i+j) \sum_{l=0}^{S_R} P(i,j,k,l,0) + \sum_{k=0}^{S_C-S_E} (S_R + S_C - k) \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{l=0}^{S_N} P(S_R,j,k,l,0)$$
$$+ \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-S_E-k}^{S_C-k} (i+j) \sum_{m=1}^{M_N} P(i,j,k,S_N,m) + \sum_{k=0}^{S_C-S_E} (S_R + S_C - k) \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{m=1}^{M_N} P(S_R,j,k,S_N,m), \tag{4.20}$$

and

$$E[N_N] = \sum_{l=0}^{S_N} \sum_{k=0}^{S_C-S_E} (k+l) \sum_{j=0}^{S_C-k} \sum_{i=0}^{S_R} P(i,j,k,l,0) + \sum_{k=0}^{S_C-S_E} \sum_{j=S_c-k+1}^{S_C+M_R-k} \sum_{l=0}^{S_N} (k+l) P(S_R,j,k,l,0)$$
$$+ \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-S_E-k}^{S_C-k} \sum_{m=1}^{M_N} (S_N + k + m) P(i,j,k,S_N,m) + \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{m=1}^{M_N} (S_N + k + m) P(S_R,j,k,S_N,m). \tag{4.21}$$

Step4: Compute new λ_{HR} and λ_{HN} substituting with Equation 4.20 and 4.21. If $|new\ \lambda_{HR} - old\ \lambda_{HR}| <= \varepsilon$ *and* $|new\ \lambda_{HN} - old\ \lambda_{HN}| <= \varepsilon$, stop. Otherwise, go to step2.

And ε is a small positive number to check the convergence.

```
$\lambda_{HR}$ and $\lambda_{HN}$
Initialize and normalize all probability P and P'
do
{
        P=P'
        P'(i)=W*state balance equation of P(i) + (1-W)P(i)
        /* Compute the new probability P' where $W \in (1, 2)$ is the convergence adjustment
     parameter*/
}while ($\sum|P'(i)-P(i)|/|P'(i)+P(i)| > \varepsilon$)
/* $\varepsilon$ is a small positive convergence criteria */
Output all $N_T$ probability P
```

Figure 4.2 SOR method in computing state probability in Markov Chain

The SOR method to compute all the probabilities P(i, j, k, l, m)s is given in the pseudo code in Figure 4.2.

4.2.3 Blocking Probability

Based on the above probability set *P(i, j, k, l, m)*s, the system performances can be obtained. When originating real-time service call arrives, the call is blocked if there is no free channel in RC since there is no queue for originating calls in our system. Therefore, the blocking probability B_{OR} of originating real-time service call is the sum of probabilities of states in which the number of channels used by real-time service calls (including originating calls and handoff request calls) in RC is equal to the capacity of RC, i.e. $i = S_R$.

$$B_{OR} = 1 - \sum_{i=0}^{S_R-1} \sum_{k=0}^{S_C-S_E} \sum_{j=0}^{S_C-k} \sum_{l=0}^{S_N} P(i, j, k, l, 0) - \sum_{i=0}^{S_R-1} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-S_E-k}^{S_C-k} \sum_{m=1}^{M_N} P(i, j, k, S_N, m). \quad (4.22)$$

Similarly, the blocking probability of originating non real-time service call is the sum of probabilities of states in which the number of channels used by non real-time service calls (including originating calls and handoff request calls) in NC is equal to the capacity of NC, i.e. $m = S_N$.

$$B_{ON} = 1 - \sum_{l=0}^{S_N-1} \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=0}^{S_C-k} P(i, j, k, l, 0) - \sum_{l=0}^{S_N-1} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} P(S_R, j, k, l, 0). \quad (4.23)$$

4.2.4 Forced Termination Probability

Forced termination probability F_{HR} of real-time service handoff request calls in the market cell consists two parts, i.e. B_{HR} and D_R. B_{HR} is the blocking probability of real-time service handoff request calls, that is the probability of real-time service handoff request call is dropped when there is no space available in RHRQ, i.e. $j+k = S_C+M_R$. For non preemptive handoff scheme, this formula means all the CC and RHRQ are full. For preemptive handoff scheme, this formula also implies that NHRQ is full, i.e., $m=M_N$, or $k=0$ otherwise the non real-time service calls in CC will be preempted by arriving real-time service handoff request calls. However, in both schemes, the B_{HR}, the blocking probability of real-time service handoff request calls, can be given by

$$B_{HR} = \sum_{k=0}^{S_C-S_E} \sum_{l=0}^{S_N} P(S_R, S_C + M_R - k, k, l, 0) + \sum_{k=0}^{S_C-S_E} \sum_{m=0}^{M_N} P(S_R, S_C + M_R - k, k, S_N, m), \qquad (4.25)$$

D_R is the dropping probability of real-time service handoff requests in RHRQ. The real-time service handoff requests waiting in RHRQ could be removed when the mobile user moves out of handoff area even if the mobile user does not get service in the handoff area. D_R represents the time-out probability of real-time service handoff request in RHRQ. The average number of real-time service handoff request calls arrived in unit time is equal to $(1-B_{HR})\lambda_{HR}$. The average number of real-time service handoff request removed in unit time from RHRQ is equal to $\mu_{hdwell}*E[L_R]$, where $E[L_R]$ is the average

$$\begin{aligned} E[L_R] = &\sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} (j + k - S_C) \sum_{l=0}^{S_N} P(S_R, j, k, l, 0) \\ &+ \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} (j + k - S_C) \sum_{m=1}^{M_N} P(S_R, j, k, S_N, m). \end{aligned} \qquad (4.26)$$

Therefore, the time-out probability of real-time service handoff request in RHRQ

$$D_R = \frac{\mu_{hdwell} E[L_R]}{(1-B_{HR})\lambda_{HR}}. \qquad (4.27)$$

The forced termination probability F_{HR} of real-time service handoff request calls in the market cell, which is combination of B_{HR} and D_R, is given by

$$F_{HR} = B_{HR} + (1 - B_{HR}) D_R, \qquad (4.28)$$

For the non real-time service handoff requests in NHRQ, it can be transferred from the referenced cell to one of the target cells when the mobile user moves out of the referenced cell before it gets service. If the size of NHRQ were to be large enough to hold all the non real-time handoff requests, the forced termination of non real-time service would never happen. Therefore, forced termination probability of non real-time service call is not given here.

4.2.5 System Utilization

The average utilization of RC is given by:

$$\eta_{RC} = \frac{E[C_{RC}]}{S_R}. \tag{4.29}$$

where $E[C_{RC}]$, the average number of calls in RC, is given by

$$\begin{aligned} E[C_{RC}] = & \sum_{i=0}^{S_R} i * \sum_{k=0}^{S_C - S_E} \sum_{j=0}^{S_C - k} \sum_{l=0}^{S_R} P(i, j, k, l, 0) + S_R * \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - k+1}^{S_C + M_R - k} \sum_{l=0}^{S_N} P(S_R, j, k, l, 0) \\ & + \sum_{i=0}^{S_R} i * \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - S_E - k}^{S_C - k} \sum_{m=1}^{M_N} P(i, j, k, S_N, m) + S_R * \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - k+1}^{S_C + M_R - k} \sum_{m=1}^{M_N} P(S_R, j, k, S_N, m), \end{aligned} \tag{4.30}$$

The average utilization of CC is given by:

$$\eta_{CC} = \frac{E[C_{CC}]}{S_C}. \tag{4.31}$$

where $E[C_{CC}]$, the average number of calls in CC, is given by

$$\begin{aligned} E[C_{CC}] = & \sum_{i=0}^{S_R} \sum_{k=0}^{S_C - S_E} \sum_{j=0}^{S_C - k} (j + k) \sum_{l=0}^{S_R} P(i, j, k, l, 0) + S_C * \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - k+1}^{S_C + M_R - k} \sum_{l=0}^{S_N} P(S_R, j, k, l, 0) \\ & + \sum_{i=0}^{S_R} \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - S_E - k}^{S_C - k} (j + k) \sum_{m=1}^{M_N} P(i, j, k, S_N, m) + S_C * \sum_{k=0}^{S_C - S_E} \sum_{j=S_C - k+1}^{S_C + M_R - k} \sum_{m=1}^{M_N} P(S_R, j, k, S_N, m), \end{aligned} \tag{4.32}$$

The average utilization of NC is given by:

$$\eta_{NC} = \frac{E[C_{NC}]}{S_N}. \tag{4.33}$$

where $E[C_{NC}]$, the average number of calls in NC, is given by

$$\begin{aligned} E[C_{NC}] = {} & \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=0}^{S_C-k} \sum_{l=0}^{S_R} l * P(i, j, k, l, 0) + \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{l=0}^{S_N} l * P(S_R, j, k, l, 0) \\ & + S_N * \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-S_E-k}^{S_C-k} \sum_{m=1}^{M_N} P(i, j, k, S_N, m) + S_N * \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{m=1}^{M_N} P(S_R, j, k, S_N, m), \end{aligned} \tag{4.34}$$

The average channel utilization is given by

$$\eta = \frac{E[C_{RC}] + E[C_{CC}] + E[C_{NC}]}{S}. \tag{4.35}$$

The average utilization of RHRQ is given by:

$$\eta_{RHRQ} = \frac{E[L_R]}{M_R}. \tag{4.36}$$

where $E[L_R]$, the average length of RHRQ, is given by Equation 4.26.

The average utilization of NHRQ is given by:

$$\eta_{NHRQ} = \frac{E[L_N]}{M_N}. \tag{4.37}$$

where $E[L_N]$, the average length of NHRQ, is given by

$$E[L_N] = \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-S_E-k}^{S_C-k} \sum_{m=1}^{M_N} m * P(i, j, k, S_N, m) + \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{m=1}^{M_N} m * P(S_R, j, k, S_N, m), \tag{4.38}$$

4.2.5 Forced Termination Probability of Real-time Service in Lifetime

The forced termination probability P_{hf} of real-time service calls is defined as the probability that a real-time service call accepted by the system is forced to terminate during its lifetime. The probability P_h that a real-time service call triggers handoff request in the referenced cell is the probability that the call holding time exceeds the dwell time of the real-time service user in the referenced cell. Thus, we have

$$P_h = \Pr\left\{ T_{CR} > T_{dwell} \right\} \tag{4.39}$$

We assume T_{CR} and T_{dwell} are independent, we can get

$$P_h = \frac{\mu_{dwell}}{\mu_{CR} + \mu_{dwell}}. \tag{4.40}$$

Therefore, the P_{hf} of real-time service calls can be expressed as

$$P_{hf} = \sum_{l=1}^{\infty} P_h F_{HR} [(1-F_{HR})P_h]^{l-1} = \frac{P_h F_{HR}}{1-P_h(1-F_{HR})}. \tag{4.41}$$

4.2.6 Average Transmission Delay of Non Real-time Service in Lifetime

In the proposed scheme, non real-time service handoff requests can be transferred from the queue of the referenced cell to the queue of another one if it cannot be served in the referenced cell. The transmission delay T_N of non real-time service is the sum of transmission delay T_W of non real-time service in every cell during its lifetime. Using the Little's formula, the average waiting time E[T_W] of non real-time service handoff requests in the NHRQ is given by

$$E[T_w] = \frac{E[L_N]}{(1-B_{HN})\lambda_{HN}}, \tag{4.42}$$

where

$$\begin{aligned} E[L_N] &= \sum_{i=0}^{S_R} \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-S_E-k}^{S_C-k} \sum_{m=1}^{M_N} mP(i,j,k,S_N,m) \\ &\quad + \sum_{k=0}^{S_C-S_E} \sum_{j=S_C-k+1}^{S_C+M_R-k} \sum_{m=1}^{M_N} mP(S_R,j,k,S_N,m). \end{aligned} \tag{4.43}$$

The average serving time E[T_S] of non real-time service calls is given by

$$E[T_S] = \frac{N_N}{(1-B_{ON})\lambda_{ON} + (1-B_{HN})\lambda_{HN}}. \tag{4.44}$$

We define N_h as the average number of handoff per a non real-time service handoff request during its lifetime. We have

$$N_h = \frac{N_h E[T_w] + E[T_{CN}]}{E[T_S]}, \tag{4.45}$$

From equation (42) we can get

$$N_h = \frac{E[T_{CN}]}{E[T_S] - E[T_w]}. \tag{4.46}$$

Therefore, the average transmission delay (expect average non real-time service time) $E[T_N]$ of non real-time service is

$$T_N = N_h E[T_W]. \tag{4.47}$$

Chapter 5.

Simulation and Numerical Results

5.1 Simulation Platform

5.1.1 Discrete Event Simulation

In order to check the correction of our analytical model and closed formula for system performance, a disserted-event simulation environment has been developed and established. Disserted-event system change state at discrete points in time, as opposed to continuous systems, which change state over time [22].

In modeling our handoff system, we need to describe its dynamics composition, the way it accomplishes the handoff procedure. Three entities – activities, processes, and events – are the constructs used to describe the dynamic behavior of discrete handoff system. An activity is the smallest unit of work in our view of system. A process is a logically related set of activities. A process may, in turn, be viewed as an activity of a higher-level process. An event is a change of state of system entity or time. This change of state results from the action of an activity. The initiation of activities is triggered by events. Our system is viewed dynamically as a collection of interacting processes, with the interactions controlled and coordinated by the occurrence of events. Especially, this process view is a hierarchical one, a system is described at a given level of abstraction by a set of process descriptions, each specifying the activities of that process, together with those of the previous level, form the expanded description of the system.

5.1.2 Random Number Generation

Determining and generating the distribution to use to represent a model variable is one of the most important aspects of simulation modeling. Input data for the simulation is generated probabilistically within the simulation program. Basically, there are two ways to generate random number, the hardware generator and software pseudo-random number generator. Modern statistical methods seem to be requiring huge numbers of random numbers and are using them in more demanding ways, and various researches have been done in this area and some methods are discussed in [23][24]. Various testing programs, such as George Marsaglia's Diehard tests, can be used to test pseudo-random number generators. In our simulation, we choose the ANSI C math library to generate a uniform distributed random number. Although the pseudo-random number generator can not be deemed as the same as a real random number generator, it does work well in our situations. Based on the uniform distribution random number, we can generate an arbitrary random distribution based on the CDF of that random distribution, as shown in exponential inverse CDF transformation algorithm in figure 5.1.

```
//Calculate the random number of exponential distribution
double expntl(double lamda){
        long l;
        while((l=rand())==0);
        double r = ((double)l)/RAND_MAX;
        return (-1*log(r)/lamda);
}
```

Figure 5.1 Exponential distribution generator

5.1.3 Criteria of Termination of Simulation

In order to achieve a specified accuracy in simulation, we should set the criteria of termination of simulation. It should be noted that our handoff system is represented by a steady-state system. That means output values should be independent of initial conditions (the length of simulation is not). When the simulation is allowed to run infinity, the mean will be a true mean of the underlying distribution. Practically, run lengths are finite and simulation provides only a sample set of value from the distributions [22]. Our task is get sample mean from finite-length simulation close enough representative to the true mean.

During our implementation of simulation, batch mean method [25] has been used to achieve this goal. The batch means method divides one long run into a set of k subruns of length m, called batches, computing a separate sample mean for each batch, and using these batch mean to computer the grand mean and the confidence interval. This method provides quicker convergence of the half width to the desired value for a sequential procedure.

5.1.4 Simulator Results

Figure 5.2 and figure 5.3 show a comparison of numerical results with analytical model and results of our discrete event driven simulation for Priority Hybrid Reservation Handoff scheme. In our numerical examples, we assume that the shape of the cell is circular with radius r, and two kind of mobile users as pedestrian users. Parameters are set as follows: r=0.1 Km, $E[D]$=0.1r, $E[V]$=0.5 meter/second, $E[T_{Cv}]$=120 seconds, $E[T_{Cd}]$=60 seconds, $S=S_R+S_C+S_N$=10, S_R=4, S_C=4, S_N=2, M_R=5, S_E=2, M_N=50, and

$\varepsilon=10^{-8}$. The ratio of originating real-time service calls and non real-time service calls $\lambda_{OR}/\lambda_{ON}$ is set to 1. The simulation stop criteria set to $\varepsilon=10^{-8}$. Both arrival events are generated by exponential function with pseudo random number generator. Figure 5.2 shows the comparison about the average utilization of common channels, the average utilization of real-time service channel, the average utilization of non real-time service channel, average length of real-time service queue, and average length of non real-time service queue. Figure 5.3 illustrate a comparison about blocking probability of originating real-time service calls (B_{OR}), blocking probability of originating non real-time service calls (B_{ON}), forced termination probability of real-time service call in life time (Phf), and transmission delay of non real-time service call in life time (T_N). From the figures, we can see that the results of simulation and analytical closed formula match each other very well and are consistent.

Similarly, for preemptive hybrid reservation handoff scheme, Figure 5.4 and figure 5.5 shows a comparison of numerical results with analytical model and results of our discrete event driven simulation for non-preemptive handoff scheme. The parameters are set as the one of the priority reservation handoff scheme before. The ratio of originating real-time service calls and non real-time service calls $\lambda_{OR}/\lambda_{ON}$ is set to 1. The simulation stop criteria is set to $\varepsilon=10^{-8}$. Both arrival events are generated by exponential function with pseudo random number generator. Figure 5.4 illustrates the comparison about the average utilization of common channels, the average utilization of real-time service channel, the average utilization of non real-time service channel, average length of real-time service queue, and average length of non real-time service queue. Figure 5.5 shows the comparison about blocking probability of originating calls, forced termination

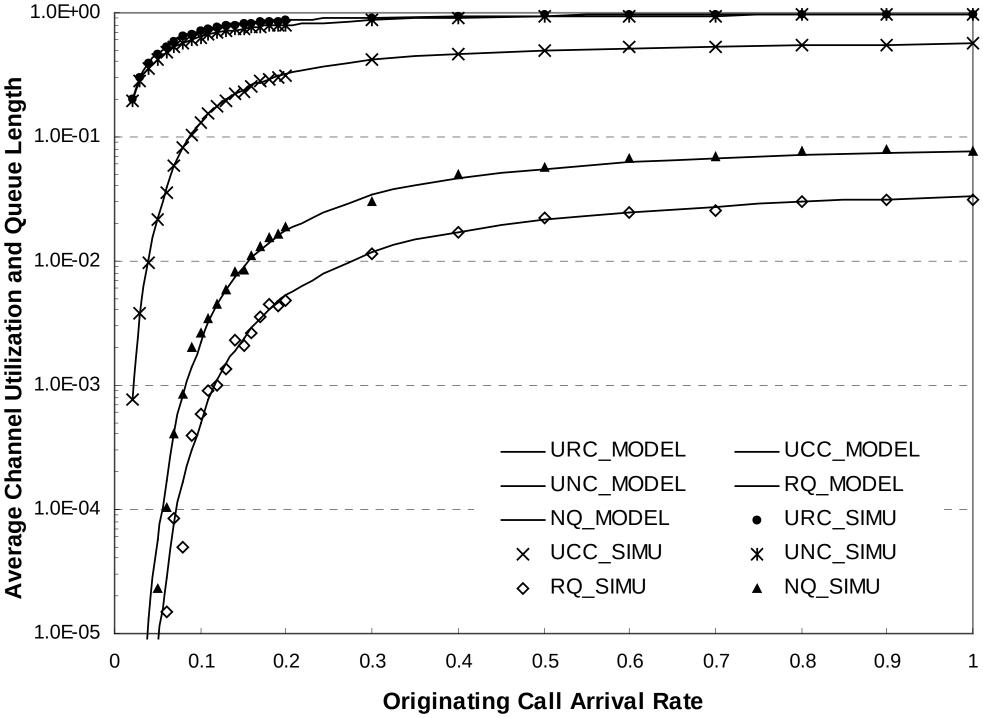


Figure 5.2 Comparison of average channel utilization and queue length of analytical model and simulation for priority handoff scheme

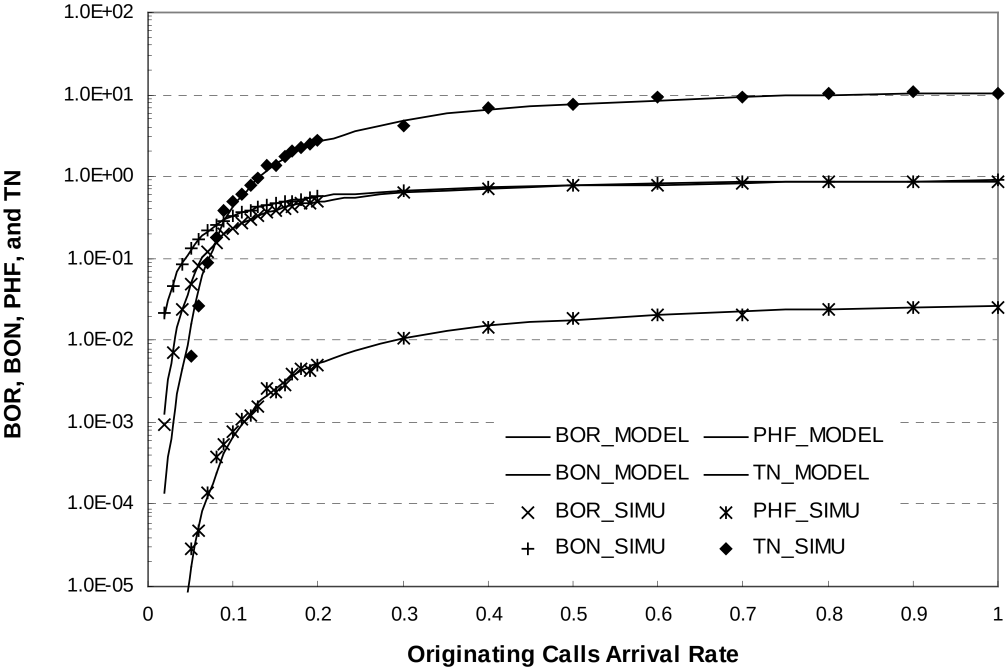


Figure 5.3 Comparison of B_{OR}, B_{ON}, Phf and T_N between analytical model and simulation for priority handoff scheme

probability of real-time service call in life time and transmission delay of non real-time service call in life time. From the figures, we can see that the results of simulation and analytical closed formula match each other very well and are consistent.

These observations substantiate the accuracy and correctness of our analytical model for both priority handoff scheme and preemptive handoff scheme.

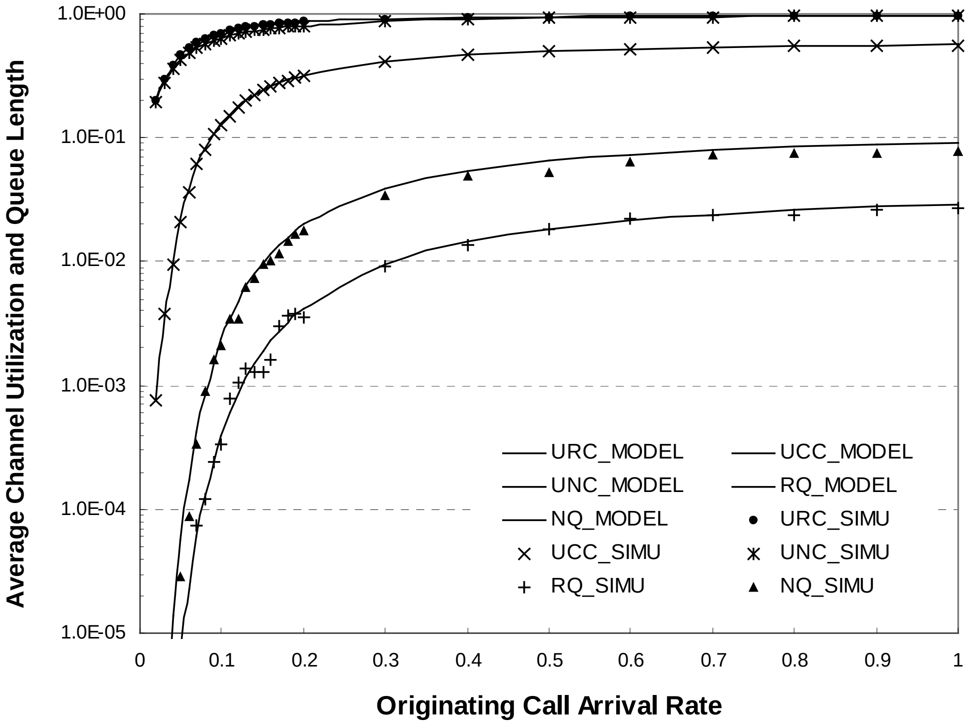


Figure 5.4 Comparison of average channel utilization and queue length of analytical model and simulation for preemptive handoff scheme

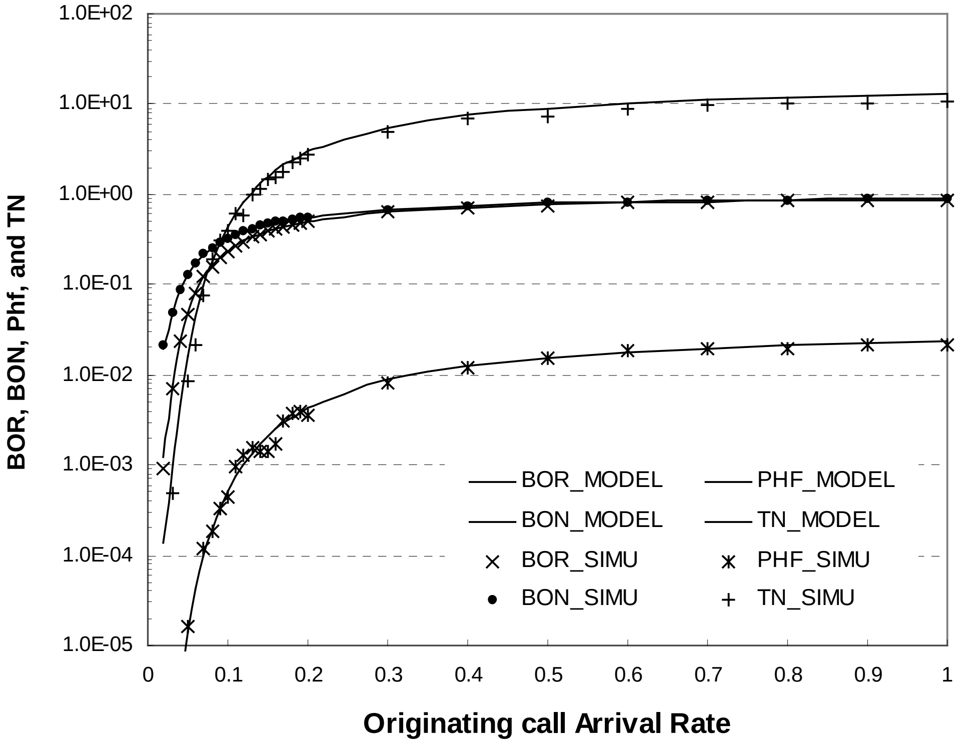


Figure 5.5 Comparison of B_{OR}, B_{ON}, Phf and T_N between analytical model and simulation for preemptive handoff scheme

5.2 Numerical Results and Discussion

In this section, system characteristics are numerically evaluated using performance analytical model developed in Chapter 4. In our numerical examples, we assume that the shape of the cell is circular with radius r, and two kind of mobile users as pedestrian users. Parameters are set as follows r=0.1 Km, $E[D]$=0.1r, $E[V]$=0.5 meter/second, $E[T_{Cv}]$=120 seconds, $E[T_{Cd}]$=60 seconds, $S=S_R+S_C+S_N=10$, $S_R=4$, $S_C=4$, $S_N=2$, $M_R=5$, $S_E=2$, $M_N=50$, and $\varepsilon=10^{-8}$.

5.2.1 Priority Hybrid Reservation Queueing Handoff Scheme

Figure 5.6 shows the blocking probabilities B_{OR} of originating real-time service calls and the force termination probability P_{hf} of real-time service calls versus offered traffic λ_O, where λ_O is the arrival rate of all originating calls in a cell. In the figure, Sr donates the given reservation number of channel in CC. When Sr is zero, the handoff scheme is normal guard channel scheme. From the figure, we can see that the blocking probability B_{OR} and the forced termination probability P_{hf} increase as the offered traffic increasing. However, the forced termination probability P_{hf} with reserved channel scheme is one order smaller than that one of without reserved channel scheme and the blocking probability is almost unchanged. Therefore, we can conclude that the proposed scheme with priority reservations is effective in decreasing the forced termination probability of real-time service calls.

Figure 5.7 shows the average transmission delay T_d versus offered traffic λ_O when $\lambda_{Ov}/\lambda_{Od}=1$. We can see that T_d decrease when the number of reserved channels increases.

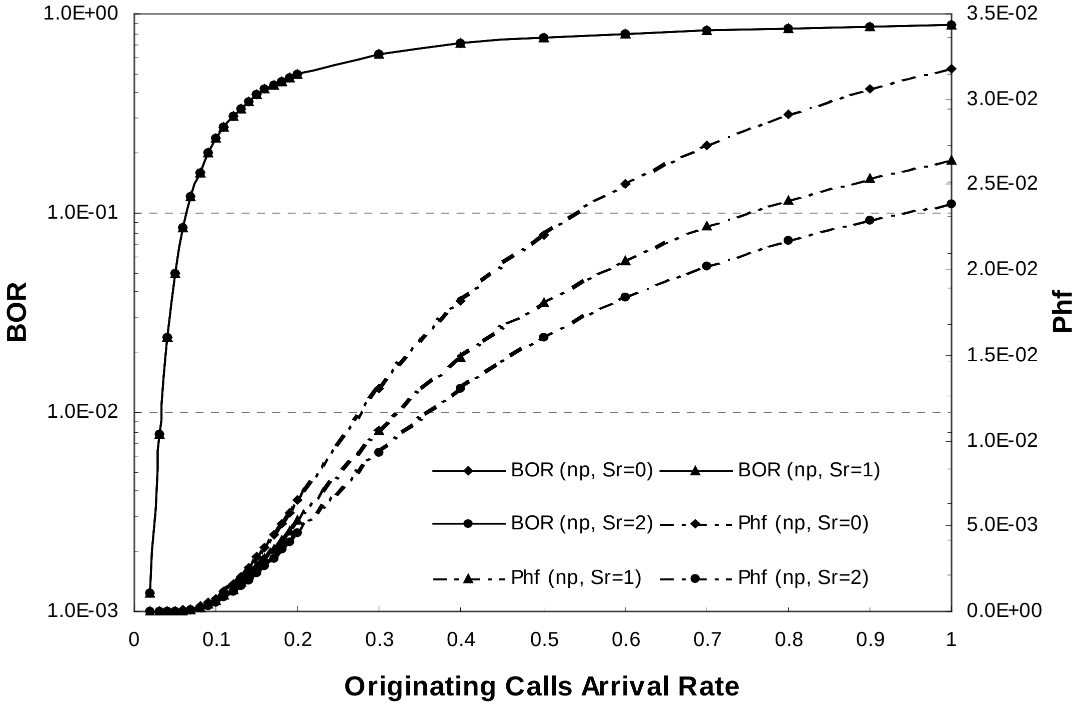


Figure 5.6. Blocking probability and forced termination probability versus offered traffic in priority handoff scheme

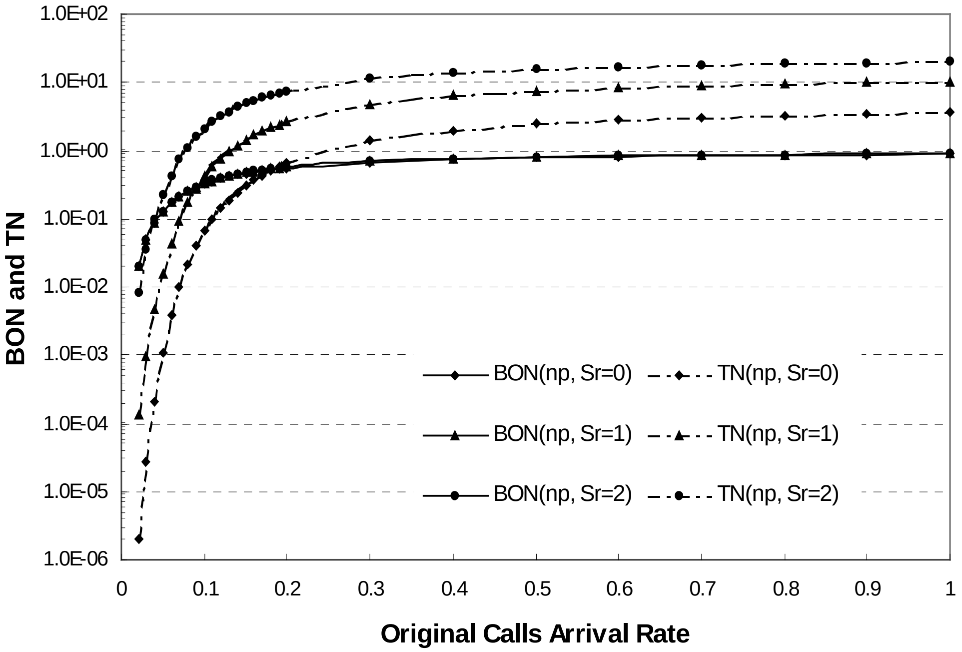


Figure 5.7. Blocking probability and transmission delay of non real-time service versus offered traffic in priority handoff scheme

This is because the number of channels available decreases for non real-time service handoff requests when S_R increases. Since real-time service transmission is very sensitive to interruption, decreasing forced termination probability is more important than decreasing transmission delay of non real-time service.

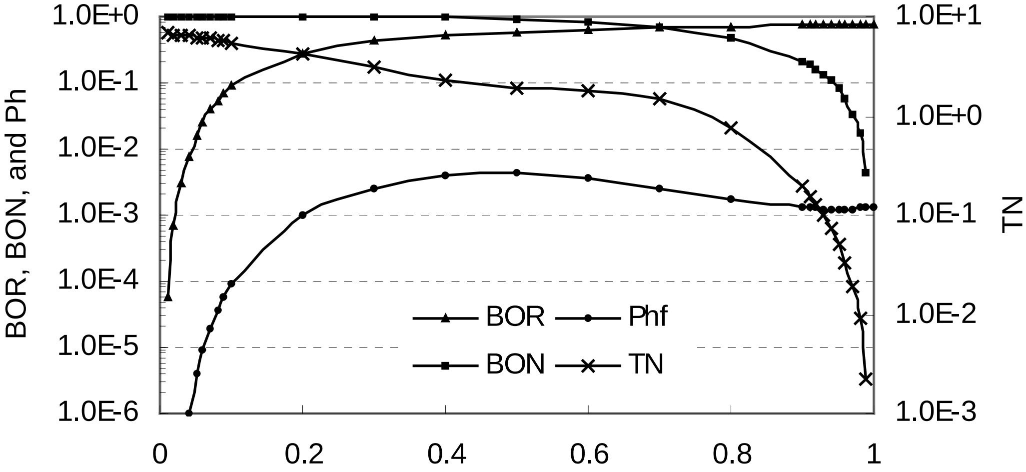


Figure 5.8. The loss probability and the average transmission delay versus the ratio of real-time traffic / total traffic in priority handoff

Figure 5.8 shows the loss probability and the average transmission delay versus the ratio of real-time service over total offered traffic. We choose offer traffic λ_O = 5. From the figure, we can see that the blocking probability B_{ov} of originating real-time service calls increases and the blocking probability B_{oh} of originating non real-time service calls decreases when the real-time traffic ratio rises. For the forced termination probability P_{hf} of real-time service call, it increases at a fast rate when real-time traffic ratio increases from 0 to 0.5 and reaches a summit when the traffic ratio is 0.5. P_{hf} decrease slightly when the real-time traffic ratio keeps increasing. The reason is, when real-time traffic ratio increases over 0.5, the non real-time service calls decreases and can be served by the reserved channels for non real-time service calls, so more reserved channels in S_R can be used by real-time service handoff requests.

5.2.2 Preemptive Hybrid Reservation Queueing Handoff Scheme

Figure 5.9 shows the blocking probabilities B_{OR} of originating real-time service calls and the force termination probability P_{hf} of real-time service calls versus offered traffic λ_O, where λ_O is the arrival rate of all originating calls in a cell. The ratio of originating real-time service calls and non real-time service calls $\lambda_{OR} / \lambda_{ON}$ is set to 1. Figure 5.9 also shows the difference between scheme with and without preemptive channel reservation scheme with S_E as the threshold number of free channel reserved before non real-time service handoff request calls could be served by S_C. From the figure, we can see that the blocking probability B_{OR} and the forced termination probability P_{hf} increase as the offered traffic is increased. However, for the non-preemptive handoff scheme, the forced termination probability P_{hf} with reserved channel scheme is an order smaller than that one of without reserved channel scheme. For the handoff scheme without reserved channels, we can see that the forced termination probability P_{hf} with preemptive scheme is much smaller than the one with non-preemptive scheme. Specifically, the preemptive handoff scheme without reserved channels performs even better than non-preemptive scheme with reserved channels. On the other hand, the blocking probability of originating real-time service calls remains almost unchanged. Therefore, we can conclude that the proposed scheme with preemptive priority reservations is very effective in decreasing forced termination probability of real-time service calls.

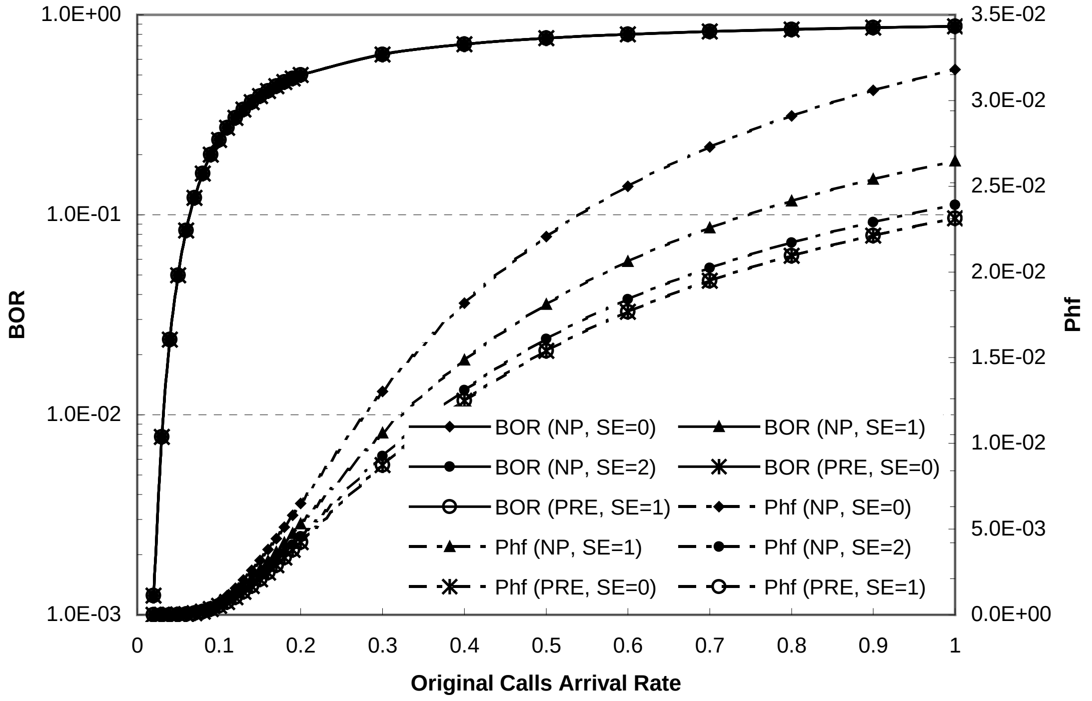


Figure 5.9 Blocking probability and forced termination probability versus offered traffic in preemptive handoff scheme

Figure 5.10. shows the average transmission delay T_N versus offered traffic λ_O when the ratio of originating real-time service calls and non real-time service calls λ_{OR} /λ_{ON} is set to 1. We can see that T_N decrease when the number of reserved channels increase. This is because the number of channels available decreases for non real-time service handoff requests when S_E increases. Since real-time service transmission is very sensitive to interruption, decreasing forced termination probability is more important than decreasing transmission delay of non real-time service. We also see that the average non real-time traffic delay of preemptive handoff scheme without reserved channels is much smaller than those of preemptive handoff scheme with reserved channels, but the forced termination probabilities P_{hf} of real-time service calls of those are almost the same. Therefore, we can conclude that the reserved channels will not help the QoS of wireless

mobile network when preemptive handoff scheme is implemented in the system. This implies the preemptive handoff scheme without reserved channels is better than other handoff schemes.

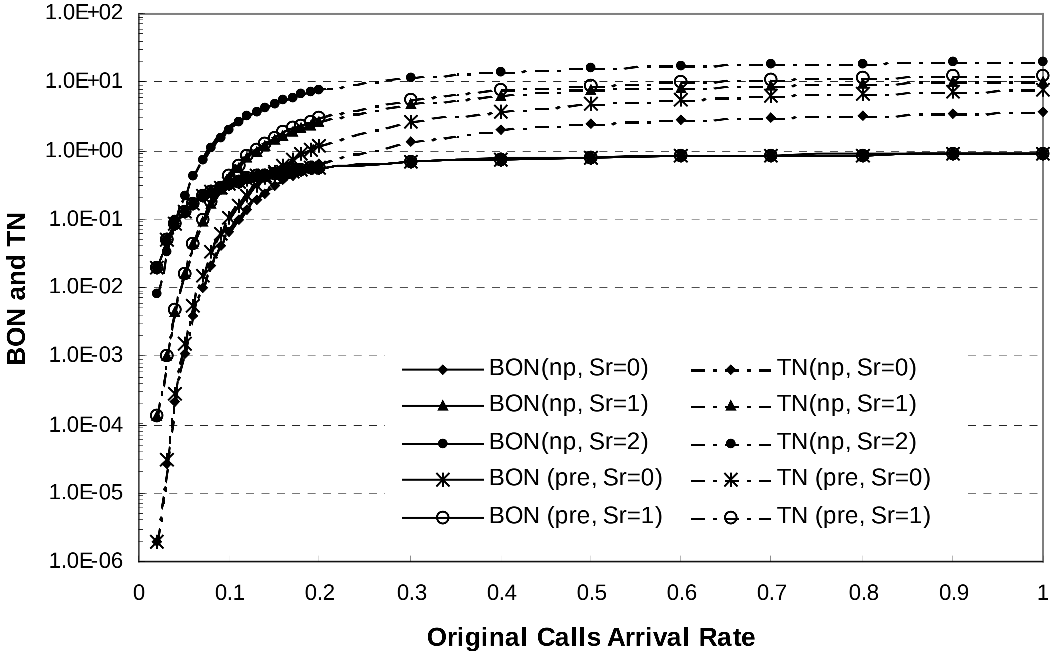


Figure 5.10 Blocking probability and average transmission delay of non real-time service versus offered traffic in preemptive handoff scheme

Chapter 6.

Conclusions

The design of handoff scheme is an important consideration for QoS in the integrated wireless mobile network system with real-time and non real-time services. The hybrid reservation handoff schemes (without and with preemptive procedure) have been proposed in this dissertation. An analytical model for the system performance has been presented. Our extensive simulation results are also matches with the analytical evaluations. Blocking probability of originating calls, forced termination probability of real-time service calls, and average transmission delay of non real-time service has been evaluated. It is observed that forced termination probability of real-time service handoff request calls and average transmission delay of non real-time service users could be decreased by our preemptive priority reservation handoff scheme. There is no transmission failure of non real-time service handoff requests except for a negligible small increase in blocking probability as a non real-time service handoff request can be transferred from the queue of one base station to another.

Some future developments of the work done so far are as follows.

- Mobility pattern of mobile user is a key character to model the system. How about the system performance if we choose a mobility pattern other than the Fluid Flood Model?
- Based on our analytical model, the performance of system can be evaluated if the number of channels and other system parameters are given. However, during the

system design phase, following question should be answered. Given a set parameters of desired QoS, how to compute the minimum number of channel necessary and how to choose the best reservation number?

- In next generation wireless network, the type of traffic may be divided into multi-classes instead of just two classes (real-time service and non real-time service). Every class of service has different requirement of QoS and maybe has different charge police. Therefore, extending our work to support the wireless mobile network integrated multi-class traffic is one of the future work area.

References:

[1] Y.-B. Lin, and I. Chlamtac, “Wireless and Mobile Network Architecture,” *John Wiley and Sons, Inc.* , 2001.

[2] U. Black, “Mobile and Wireless Network”, *Prentice-Hall*, 1996.

[3] M. D. Austin and G. L. Stuber, "Direction biased handoff algorithms for urban microcells," *Proceedings of the IEEE VTC-94*, pp.101-105, June. 1994.

[4] A. Murase, I. C. Symington, and E. Green, "Handover criterion for macro and microcellular systems," *Proceedings of the IEEE VTC-91*, pp. 524-530, June. 1991.

[5] S. Sen, A. Arunachalam, K. Basu, and M. Wernik, “A QoS management framework for 3G wireless mobile network,” *Proceedings of IEEE Wireless Communications and Networking Conference*, pp. 155-162, Sept. 1999.

[6] S. Tekinay, and B. Jabbari, “Handover and channel assignment in mobile cellular networks,” *IEEE Communication Magazine*, Vol. 29, No. 11, pp. 42-46, Nov. 1991.

[7] D. Hong, and S. S. Rappaport, “Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures,” *IEEE Transactions on Vehicular Technology*, Vol. VT-35, No. 3, pp. 521-520, Aug., 1986.

[8] R. Guerin, “Queuing-blocking system with two arrival streams and guard channels,” *IEEE Transactions on Communication*, Vol. 36, No. 2, pp. 153-163, Feb. 1998.

[9] Q-A. Zeng, K. Mukumoto, and A. Fukuda, “Performance analysis of mobile cellular radio system with priority reservation handoff procedures,” *Proceedings of the IEEE Vehicular Technology Conference*, Vol. 3, pp. 1829-1833, June 1994.

[10] D. Hong, and S. S. Rappaport, “Priority oriented channels access for cellular systems serving vehicular and porable radio telephones,” *IEE Proceedings I, CSV-136*, No. 5, pp. 339-346, 1989.

[11] Q-A. Zeng, K. Mukumoto, and A. Fukuda, “Performance analysis of mobile cellular radio system with two-level priority reservation handoff procedure,” *IEICE Transactions on Communication*, Vol. E80-B, No. 4, pp. 598-607, 1997.

[12] S. S. Rappaport, “The multiple-cell handoff problem in high-capacity communications system,” *IEEE Transactions on Vehicular Technology,* Vol. 40, No. 3, pp. 546-557, 1991.

[13] C. Purzynski, and S. S. Rappaport, “Multiple call handoff problem with queued handoffs and mixed platform types,” *IEEE Proceeding I*, *CSV-142*, No. 1, pp. 31-39, 1995.

[14] D. J. Goodman, "Trends in cellular and cordless communication," *IEEE Communication magzine,* Vol. 29, no.6, pp.31-40, Jun. 1991.

[15] Q-A Zeng, and D. P. Agrawal, “Performance analysis of a handoff scheme in wireless data networks,” *Submitted to publication*, Jan. 2000.

[16] F. N. Pavlidou, “Two-dimensional traffic models for cellular mobile systems,” *IEEE Transactions on Communication*, Vol. 42, No. 2/3/4, pp. 1505-1511, Feb./Mar./Apr., 1994.

[17] Q-A. Zeng, and D. P. Agrawal, "Performance Analysis of a Handoff Scheme in Integrated Voice/Data Wireless Networks," *Proceedings of IEEE Vehicular Technology Conference,* pp 845-851, Sep. 2000.

[18] Q-A. Zeng, and D. P. Agrawal, "An Analytical Modeling of Handoff for Integrated Voice/Data Wireless Networks with Priority Reservation and Preemptive Priority Procedures," *Proceedings of Workshop on Wireless Networks and Mobile Computing in conjunction with ICPP2000*, pp. 523-529, Aug. 2000.

[19] H. Xie, and S. Kuek, “Priority Handoff analysis”, *Proceedings IEEE Vehicular Technology Conference,* pp. 182-190, Aug. 1993.

[20] R. Jain, “The art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling”, *Wiley*, 1991.

[21] H. Akimaru, and R. B. Copper, “Teletraffic Engineering,” *Ohm*, 1985.

[22] M. H. MacDougall, “Simulation Computer Systems Techniques and Tools*”*, *The MIT Press*, 1987.

[23] J. Banks and J.S. Carson, “Discrete-Event System Simulation”, *Prentice-Hall*, 1984.

[24] P. Bratley, B.L. Fox, and L.E. Schrage, “A Guide to simulation”, *Springer-vrlag*, 1983.

[25] H. T. Odum, E. C. Odum, “Modeling for all scales: an introduction to system simulation”, *Academic Press*, 2000