An Adaptive Multi-Class Call Admission Control For Multimedia Wireless Networks

Ebrahim Abdulrahman Ismail

CENG 97-16

Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089-2562
(213) 740-4579

October 1997

An Adaptive Multi-Class Call Admission Control For Multimedia Wireless Networks

By

Ebrahim Abdulrahman Ismail

A Dissertation Presented to the

FACULTY OF THE GRADUATE SCHOOL

UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

(Computer Engineering)

December 1997

To my father, mother, wife and children,

for their love, encouragement, support, and patience

UNIVERSITY OF SOUTHERN CALIFORNIA
THE GRADUATE SCHOOL
UNIVERSITY PARK
LOS ANGELES, CALIFORNIA 90007

This dissertation, written by

EBRAHIM ISMAIL

under the direction of his....... *Dissertation Committee, and approved by all its members, has been presented to and accepted by The Graduate School, in partial fulfillment of requirements for the degree of*

DOCTOR OF PHILOSOPHY

....
Dean of Graduate Studies

Date ..October 21, 1997..........

DISSERTATION COMMITTEE

.... *Chairperson*

....

....

Acknowledgment

First and foremost all thanks and praise goes to God for the strength, patience, and knowledge He granted me which made this life-long dream a reality. I would like to thank my committee chairman and advisor Dr. John Silvester for his guidance, insightful comments, and valuable discussions, which made me a better researcher. I am grateful to Dr. Estrin, Dr. Chugg, Dr. Gupta, and Dr. Gaudiot for the time they spared by becoming members of my defense and qualifying committees. I am forever in debt to Tess Reyes for her valuable efforts in arranging my meetings with Dr. Silvester. My thanks also go Diane Dimetrias and Mary Zittercob for their help and advice. To my office mates Dr. Gilberto Mayor and Tien-Chien Yu I offer my heart felt gratitude for their warm friendship. I also would like to acknowledge members of our research group Jinsung Choi, Hasik Jeen, Yu-Shun Wang, Yongsuk Park, and Antonia Boadi for the valuable discussions we shared on many occasions. I also thank my dear friend Dr. Mohammed Arafeh for his help, support and friendship. Last but not least, special thanks goes to my Kuwait Embassy advisor Suzanne Dadzie, whose help was invaluable to me and to my family. To all the people I mentioned above, thanks for making my Ph.D. experience an enjoyable one.

Abstract

We classify handoff traffic into two priority classes, realtime and non-realtime. To provide the different priority classes with the quality of service required, we extended the basic guard-channel scheme to include multi-level guard-channel reservation for the different priority classes. An adaptive multi-guard channel scheme that adapts the number of the guard channels to the changing network congestion is introduced to better utilize the network resources. The adaptive scheme performs as well as the deterministic optimal solution of this system, i.e. performance of the different classes when the guard-channel sizes are optimized for best low priority class performance.

To support data communications, which can require multiple channels per call, the adaptive scheme was extended to use a best effort technique. Best effort is a technique where calls requiring multiple channels are allocated a smaller number of channels than requested based upon availability of channels. When compared with an adaptive scheme with a number of channels reserved exclusively for the multi-channel class, the adaptive scheme with best effort extension provided a better performance for all the priority classes at a cost of a longer channel holding time. Employing a user-based priority classification instead of call-based classification, the call admission-control problem transforms into a multiple priority-class problem with multi-rate calls. The adaptive scheme is compared to three best effort extensions, best effort for all classes, best effort for the highest two classes, and best effort for the highest priority class only. Although there is a performance gain from using the best effort extension, this gain is not substantial.

Table of Contents

Table of Figures

List of Tables

Chapter 1

Introduction

The increasing popularity of wireless services and wireless networks can be attributed to the users' demand for increased connectivity and mobility. This demand is forecasted to increase even more in the next decade especially as market place competition drives unit prices and call rates down and also as new multi-media services like video telephony, video conferencing, and Internet connections are introduced in the wireless and mobile networks. But the increase in demand is hampered by the limited capacities of wireless networks, which limits the number of users the network can support. Although progress in increasing the capacity of the radio channel has been achieved, the scarce resources of the wireless spectrum are still very challenging.

1.1 Cellular Networks

Wireless networks are composed of numerous cells each controlled by an associated base station connected to an antenna. To increase the capacity of the cellular network, micro-cellular design is suggested and in some cases, as in Japan's PHS, Personal Handy-phone System, is implemented. The micro-cellular concept is based on the premise that a cell can be divided into several micro-cells, each of which is assigned a subset of the original cell's frequencies. These frequencies are assigned to these micro-cells such that the minimum frequency re-use distance is satisfied to eliminate any co-channel interference. Hence, the micro-cellular system can yield larger capacity than a cellular system. On the other hand, the micro-cellular design suffers from a profound problem, called the handoff problem.

1.2 Handoffs

As the sizes of cells are reduced, the frequency of a mobile user crossing cell boundaries also increases. A handoff is caused by an ongoing call that crosses cell boundaries. Handoffs pose an interesting and fundamental problem in mobility, that is the problem of provision of an uninterrupted seamless communication during the lifetime of a call. As a mobile user moves away from the base station, the signal strength of the connection received by the mobile decreases. A handoff is triggered when the level of signal strength of the connection to the base station is equal to the signal strength level received from a neighboring cell. Other handoff triggering criteria have also been proposed in the literature, but they are still a derivative of this basic concept. As the mobile user continues to move away from the original base station, the signal strength level received drops below the minimum level required to carry on a communication connection, at which point the mobile's connection to the original base station is dropped. The time interval between when the handoff is first triggered to the time the connection is dropped is called the handoff area or dwell time. When a mobile requests a handoff to a neighboring cell, it may or may not find a channel available in that cell due to the heavy state of congestion, causing the call to be interrupted. How to manage handoffs, and call admission control, especially when there are different call types is the main focus of our research.

1.3 Call Admission Control in Wireless Networks

There are many call admission control and management issues related to wireless networks especially handoff calls: Should a base station give priority to handoffs over new calls? Should a base station queue handoffs until a channel is available? Should it queue new calls if no channel is available also? Should a base station treat all handoffs the same? Are there priority criteria to prioritize handoffs? Are there priority criteria to prioritize calls?

All these questions are asked in the context of under what network load conditions a handoff call or a new call should be accepted, i.e. allocated a channel. This is the basic question which call admission control policies proposed in the literature try to answer. We can classify admission control schemes on the following criteria:

1) What criterion does a base station use to classify calls into priority classes?

2) How are channels assigned to calls such that higher priority calls have better perception of quality of service and lower priority calls do not have a drastically worse perception of quality of service?

1.4 Previous Work

Several call admission control schemes have been proposed for wireless networks in the literature. They take different approaches to solving the problem ranging from the simple FIFO channel allocation to sophisticated call classification and priority classes.

In [1], Hong and Rappaport considered two prioritized handoff procedures. In the first scheme, a set of channels is used exclusively for handoff calls, these are called guard channels, while the remaining channels are used for both new and handoff calls, and blocked calls are cleared from the system immediately. In the second scheme, handoff call attempts can be queued for the time duration in which the mobile dwells in the handoff area between cells until a channel frees up in which case the handoff call is admitted and allocated the channel. The channels are shared in the same way as in the previous scheme. The authors found that the probability of dropping handoff calls is lower for the second scheme where queueing of handoffs is employed while there is no difference in the probability of blocking of new calls. This work introduced the guard channel method to wireless call admission schemes. This demonstrates the effectiveness of guard channels in providing a better quality of service to handoff calls, but it was a simple scheme that used FIFO discipline within the handoff queue. With FIFO, handoffs are prioritized according to their arrival time not according to their delay constraints. Handoff calls have different delay constraints and FIFO does not accommodate those calls with tighter delay constraints than others.

Roch Guerin in [2] presented a method that minimized the blocking of handoff calls and increased the total carried traffic by queueing new calls. In this scheme, a mobile that wants to initiate a call sends a call request to the system on the control channel, and waits for an answer from the system, during which time the request is queued waiting for a channel to become available. The mobile automatically resends the call request after a specified timeout period. Although this method increases the carried load in a cell by queueing new calls, it does not solve the larger problem of handoffs dropping. Handoff dropping is an important quality of service criterion from the user's perspective, since a new call blocking is more tolerable by a user than an ongoing call dropping, i.e. handoff forced termination.

Tekinay and Jabbari in [3] introduced a scheme that prioritized handoff requests. New calls are dropped if no channel is available, and handoffs are queued such that as soon as a channel is available it is offered to the mobile with the signal strength measurement result closest to the minimum acceptable power level for communication. This scheme reduced the probability of handoff dropping due to this prioritization criterion. The premise of this work is that handoffs are prioritized according to who is going to lose the channel next, although a better question might have been who is going to lose more, content wise, when losing the channel. A transaction application would not lose much, content wise, when the channel is dropped compared to a realtime voice call.

Keilson and Ibe in [4] presented a guard channel scheme in which a cell has two arrival streams, new and handoff calls. Arriving calls are indiscriminately assigned channels as long as the number of busy channels is below a pre-defined threshold. When this number is greater or equal to the threshold, only arriving handoff calls are assigned channels immediately; arriving new calls are queued. When all channels are busy, handoff calls are dropped while new calls continue to be queued. This scheme although gives priority to handoffs by using a guard channel scheme, when all channels are busy it reverses that priority in favor of new calls. It is important that handoff calls have priority at all time to provide seamless mobile communication and to uphold the user's perception of quality of service.

McMillan in [5] analyzed another guard channel scheme where when all channels are busy, handoff calls are queued while new calls are blocked. He also assumes that the waiting time is small enough that no queued call would leave the queue before being assigned a channel. This assumption is not a valid one in all cases, since for a highly mobile user the time spent in the handoff area is very short and if the mean channel holding time is high, e.g. > 4 minutes, handoffs will defect from the queue and be dropped.

Chen and Schwartz in [6] proposed a two-tier scheme, where multiple cells pool some of their channels in a shared pool. Traffic into a cell is divided into three classes, new narrowband, handoff narrowband, and wideband calls. New narrowband calls are queued if no channels are available, while narrowband handoffs and wideband calls are dropped if no channels are available. A dropped narrowband handoff or wideband call would overflow their request to the shared channel pool. The authors further assume that wideband calls will not handoff due to their assumed short channel holding time. Epstein and Schwartz in [7] used the same classification mentioned in [8] to analyze three admission control policies, complete sharing (CS), complete partitioning (CP), and reservation or guard channels. They concluded that complete sharing (CS) favors channel throughput at the expense of providing quality of service. Also, CS gives no priority to handoff traffic and favors calls with small bandwidth requirement. On the other hand, they found that the reservation policy provided the high priority class with a better quality of service at the expense of lower total system throughput. The question of how to classify calls in a cell is an important one to answer. As mentioned here, the authors elected to classify the calls according to their bandwidth requirements, while this might be a valid classification, it does not consider the time constraints that a handoff call might have.

Hong and Rappaport in [8] classified traffic into a cell into handoffs, vehicular new calls, and portable new calls. Their admission scheme gives priority to handoffs over new calls and to vehicular new calls over portable new calls. The scheme uses the guard channels method to reserve some channels to handoff calls, with handoff calls queued if no channel is available.

Naghshineh and Acampora in [9] classify traffic into a cell into realtime and non-realtime calls. They investigated three admission control schemes, complete partitioning (CP), complete access (CA), and restricted access (RA). In complete partitioning, the bandwidth is completely partitioned between the two classes of traffic. In complete access, realtime calls can use up to the total bandwidth of the cell with pre-emptive priority over non-realtime calls. At any given time, the capacity not utilized by realtime calls is equally shared between all non-realtime calls. In restricted access, realtime calls can use up to and no more than some threshold C with pre-emption priority over non-realtime calls. The remaining bandwidth B-C is dedicated to non-realtime calls. Again, the capacity not utilized by realtime calls is equally shared by non-realtime calls. The authors concluded from their analysis of the three schemes that only restricted access scheme can satisfy the quality of service requirements of the traffic classes.

Olviera, Suda, and Bae in [10] proposed a reservation scheme for handoffs by which once a call is accepted in a cell it reserves the required channels in all neighboring cells. This scheme although it gives a channel guarantee to a call before that calls handoffs to a neighboring cell, reserving in all neighboring cells is wasteful of resources, since only one channel is actually used when, if at all, the reserving call does handoff to a neighboring cell out of all channels reserved.

Yu and Leung proposed in [11] a dynamic guard channel scheme, which dynamically varies the number of guard channels in a cell with a current estimate of the handoff traffic, given by the number of ongoing calls in neighboring cells. This provides a good framework to adapt to network congestion to keep the high priority class's quality of service level constant. The down side of the scheme is that it uses a lot of inter-cell communication to update each cell with the current congestion condition of its neighbors.

1.5 Discussion

Most works on wireless admission control have concentrated on classifying calls into a cell to two priority classes, handoffs and new calls, except for some of the works sited above, where a classification based on time constraints were used and a sub-classification of new calls based on bandwidth requirements is used. From our discussion in the previous section we conclude that giving priority to handoffs over new calls is essential to ensure seamless mobile communication and to the user's perception of quality of service. We also conclude that different handoffs have different quality of service requirements and different constraints, and hence there is a need to sub-divide handoffs into several sub-classes. The most important constraint on handoff calls is the time constraint, where dropping a non-realtime handoff call is less catastrophic than dropping a realtime handoff call. Hence, sub-division of the handoff call class into realtime and non-realtime is warranted, with realtime handoffs having priority over non-realtime handoffs. Moreover, new calls should have their own low priority class and should not be sub-divided into sub-classes or dissolve into the handoff classes as one paper suggested.

1.5.1 Realtime vs. Non-Realtime Calls

Traffic incoming to a network can be divided into two major classes based on their delay constraints to realtime and non-realtime calls. Realtime calls have a strict delay constraint that needs to be satisfied to guarantee the useful exchange of information. Realtime calls are expressed in ATM terms as continuous media calls and classified as CBR and VBR traffic. Examples of realtime calls would be voice and video calls and any other delay sensitive traffic. Non-realtime calls, on the other hand, are delay tolerant, but are loss sensitive or have a reliability constraint. Non-realtime calls include but not limited to data communications like file transfer and HTTP traffic.

Although in a wireless environment loss takes on the meaning of a call drop that is undesirable by all calls, some types of calls are more tolerant to call drops than other types when it comes to information recovery. Information recovery is the ability to recover the lost information due to a call drop, this can be by starting the communication over or continuing from the point at which the call was dropped. For example a voice call would be classified as a realtime call since a call drop would mean an unrecoverable loss of information, however, a file transfer would be classified as a non-realtime call since the file can re-downloaded once a connection is established again. In this dissertation we classify voice and video communication as realtime calls and data communication as non-realtime calls.

1.5.2 Call Classifications

Based on the criterion of classification, priority is given to some classes over other classes. This priority is needed to satisfy these classes' quality of service constraint.

In the literature, traffic incoming to a wireless network is usually classified into handoffs and new calls. Since the dropping of an ongoing call is less tolerable to the user than the blocking of a new call, handoff calls have been given priority over new calls. There are several additional criteria to classify calls and award priority upon. These criteria can be used in addition to the handoff-new call classification or as a separate classification scheme. The two criteria we will focus on in this dissertation are the delay and the user.

1.5.2.1 Delay-Based Priority

Calls can be classified as discussed above based on their delay tolerance into realtime and non-realtime calls. Since realtime calls are more sensitive than non-realtime calls to delay, priority should be given awarded to the realtime calls class over the non-realtime class. This criterion can be employed in addition to the call request location criterion, i.e. handoff-new call criterion. In this case, the traffic into the network would be classified into three classes, realtime handoffs, non-realtime handoffs, and new calls. There is no need to divide the new calls class into realtime and non-realtime since a call that has not been initiated does not have any constraints. The problem of satisfying the quality of service constraints of these classes is the principal objective of this dissertation.

1.5.2.2 User-Based Priority

It is more practical from the service provider's perspective to assign priorities according to user categories. Users can subscribe to one of several prioritized service plans where the higher the priority of the service plan the more expensive its premium. In this paradigm, a high priority user needs all his/her calls to be treated as high priority whether the calls are realtime or non-realtime, handoffs or new calls. In this case a priority classification based upon users subscription choices is better suited to provide users of each class with their subscribed quality of service.

1.6 Dissertation Outline

In this dissertation we will consider the call admission control problem in wireless networks. In chapter two, the multi guard-channel scheme that divides incoming traffic into priority classes and associate guard-channels with the higher priority classes is presented. In chapter three, an adaptive extension of the multi guard-channel scheme is developed. This adaptive scheme dynamically changes the sizes of the guard-channels based on the current state of congestion experienced by the different classes. Chapter four investigates the case when the second highest priority class is of a multi-channel request type. A best effort extension to the adaptive scheme is evaluated. In chapter five, a user-based priority classification which results in a multi-rate call requests is investigated. The use of different levels of the best effort technique is compared with the pure adaptive scheme. In chapter six we summarize the results and the contributions of this dissertation and discuss some future directions in this research area.

Chapter 2

Class-Based Call Admission

In this chapter we propose a class-based call admission scheme, which is motivated by the need to classify traffic into three priority classes, realtime handoffs, non-realtime handoffs and new calls. Each of these priority classes has its own quality of service requirement, although in this work the lowest priority class' quality of service requirement is not considered as an objective. This scheme generalizes the guard channels method proposed in the literature, to provide each priority class with an appropriate quality of service. Although we propose only three classes in this chapter, the generalized guard channel method can be extended to a larger number of priority classes depending on the classification method used.

2.1 Guard Channels Method

In this method, when calls into a cell are classified into two priority classes, such as handoffs and new calls, a set of channels are reserved for the high priority class, called guard channels. When the number of available channels is greater than the number of guard channels, all traffic classes are admitted in the cell and allocated a channel. But when the number of available channels is less than the number of guard channels, then only the high priority class calls are admitted and allocated channels.

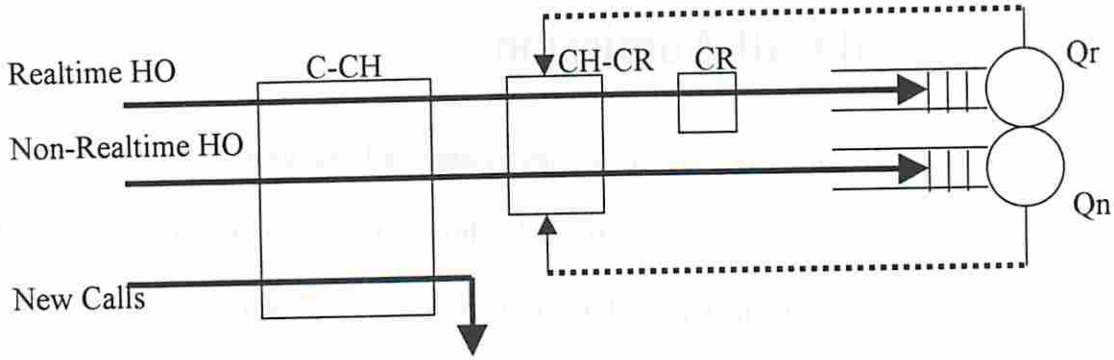

Figure 1- Multi Guard-Channel Scheme

To accommodate three priority classes, we can generalize this method by introducing two threshold points CH and CR. CH is the number of channels reserved for the highest two classes, while CR is the number of reserved channels for the highest class. In our proposed class-based admission control scheme, there are three classes of calls. The three priority classes in descending order are realtime handoffs, non-realtime handoffs, and new calls. To reduce the dropping rate of handoff calls, the scheme queues calls of both handoff classes when channels are not available. Let Available Channels denote the number of available channels in the cell. The scheme works as follows:

Available Channels	Realtime Calls	Non-Realtime Calls	New Calls
$> CH$	Admit	Admit	Admit
$> CR$ and $\leq CH$	Admit	Admit	Block
$\leq CR$	Admit	Queue	Block
$= 0$	Queue	Queue	Block

Table 1 - Admission Decision

1) If Available Channels $>$ CH :

   All calls are admitted and allocated channels.

2) If CH $\geq$ Available Channels $>$ CR :

   The highest two classes, realtime and non-realtime handoffs are admitted and allocated channels. New calls are blocked.

3) If Available Channels $\leq$ CR :

   Only the highest priority class, realtime handoffs, is admitted and allocated channels. Non-realtime handoff calls are queued. New calls are blocked.

4) If Available Channel = 0, i.e. all channels are busy :

   Both realtime and non-realtime handoff calls are queued. New calls are blocked.

5) In all cases new calls are never queued.

The reason that new calls are never queued is because of the valid assumption that mobile users usually re-dial when a channel is not allocated within a short period of time. The dwell time is the time interval during which the mobile can wait for an allocation of a channel from the base station, without being disconnected from the original base station. Therefore, handoff requests can be queued waiting for a channel for a time period of maximum length equal to the mobiles' dwell time, before losing their connection with the original base station and hence dropping from the queue. The dwell time's value depends mainly on the velocity of the mobile and the size of the cell. Faster mobiles or smaller cell sizes, or both, result in shorter dwell times for the mobiles, and vice versa.

2.2 Mathematical Solution

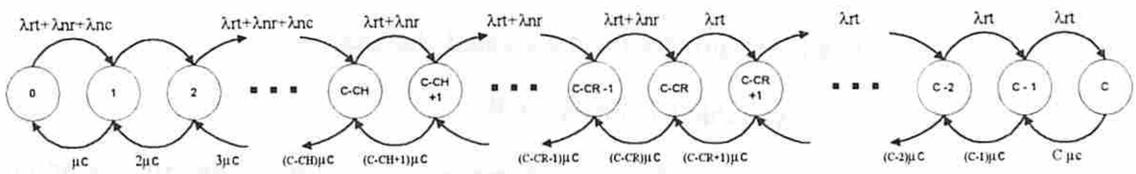

Figure 2 - Markov chain model of the multi guard-channel scheme

A mathematical analysis of system models, when feasible, can provide an insight on the dynamics of that system and results in better understanding of its behavior. It also provides a set of expressions to directly obtain the values of different system performance metrics such as dropping rate. Mathematical analysis can also be compared with simulation results to verify the correctness of the simulation results.

The multiple guard-channel scheme proposed can be modeled as a M/M/m queue with two priority cutoffs as shown in the Markov chain in figure (2). To arrive at the expressions for the dropping rate for each of the three classes, we apply queueing theory techniques such as those explained in [47] to this Markov chain. The dropping rate expressions for each priority class obtained by mathematical analysis are given below.

Although the multiple guard-channel provides queueing to the high priority handoff calls when channels are not available, we did not incorporate this queueing in the model presented by this Markov chain and its associate mathematical analysis, this is left for future work.

$$P_0=\left[\sum_{k=0}^{C-C_H}\frac{1}{k!}\left(\frac{\lambda_1}{\mu}\right)^k+\sum_{k=C-C_H+1}^{C-C_R}\frac{1}{k!}\left(\frac{\lambda_2}{\mu}\right)^{k-(C-C_H)}\left(\frac{\lambda_1}{\mu}\right)^{C-C_H}+\sum_{k=C-C_R+1}^{C}\frac{1}{k!}\left(\frac{\lambda_3}{\mu}\right)^{k-(C-C_R)}\left(\frac{\lambda_2}{\mu}\right)^{C_H-C_R}\left(\frac{\lambda_1}{\mu}\right)^{C-C_H}\right]^{-1}$$

$$P_{nc}=1-P_0\sum_{k=0}^{C-C_H-1}\frac{\left(\frac{\lambda_1}{\mu}\right)^k}{k!}$$

$$P_{nrt}=\frac{(\lambda_1)^{C-C_H}(\lambda_2)^{C_H-C_R}}{(\lambda_3)^{C-C_R}}P_0\sum_{k=C-C_R}^{C}\frac{1}{k!}\left(\frac{\lambda_3}{\mu}\right)^k$$

$$P_{rt}=\frac{1}{C!}P_0\left(\frac{\lambda_3}{\mu}\right)^{C_R}\left(\frac{\lambda_2}{\mu}\right)^{C_H-C_R}\left(\frac{\lambda_1}{\mu}\right)^{C-C_H}$$

The various symbols of these equations are defined in table (2). These expressions are the foundation of the deterministic optimal reserved-channel assignment solution that will be developed in chapter three. These performance metrics, i.e. the dropping rates, are a valuable measures of the performance of each class for monitoring the quality of service of each class under different network and system conditions. These metrics can also be helpful in providing an insight into the network performance by comparing the metrics of the different classes relative to the other classes in the network.

Symbol	Description
λ_1	The aggregate mean arrival rate to the cell
λ_2	The aggregate mean arrival rate for both handoff classes
λ_3	The mean arrival rate for the realtime handoffs class
μ	The mean service rate
P_{nc}	The dropping probability of the new calls class
P_{nrt}	The dropping probability of the non-realtime handoff class
P_{rt}	The dropping probability of the realtime handoff class
C	The total number of channels in the cell
C_H	The number of channels reserved for both handoff classes
C_R	The number of channels reserved exclusively for realtime handoffs

Table 2 - Mathematical symbol legend

2.3 Network Model

In our model, we consider a target cell in the network with incoming calls classified into three priority classes, realtime handoffs, non-realtime handoffs and new calls. The three arrival processes of the three priority classes are assumed to be Poisson arrival processes with mean rates λ_{rt}, λ_{nr}, and λ_{nc}. An admitted call is allocated a channel with an exponentially distributed channel holding time of mean rate μc. If a handoff call is queued due to the unavailability of channels, the time it spends in the queue, i.e. the dwell time, is also assumed to be exponentially distributed with mean rate μq. The cell also has two separate queues for realtime handoffs and non-realtime handoffs calls with sizes Q_r and Q_n respectively.

The method used in allocating channels and frequencies to the different cells in the network is irrelevant to our work, since the cell uses this admission control scheme after it has been assigned a set of channels. However, the number of channels assigned to the cells is of great importance and we assume it to be fixed. The incoming calls to the cell are assumed to require one channel per call, which is a reasonable assumption since currently voice and data are both carried on 32 Kbps channels. In the following chapters the more general case where calls request different number of channels are considered and analyzed. We also assume that the type of a handoff call, whether realtime or non-realtime, is known from its handoff request.

Several performance metrics are of interest when analyzing this scheme. The probability of dropping realtime handoffs, which is the fraction of the arriving realtime handoff calls that are dropped either due to full queue or due to long waiting in the queue. The same can be said about the probability of dropping non-realtime handoffs. Another metric of interest is the blocking of new calls, which is the fraction of the new calls that arrive and find the number of available channels is less than CH, and hence are blocked.

2.4 Simulation Results

To obtain the performance values mentioned above, we simulated a cell using the class-based admission control proposed here. We used the following values for our simulation:

Simulation Parameters	*Value*
Cell Channel Capacity	35
Number of Handoffs (both classes) guard channels CH	15
Number of Realtime Handoffs Guard Channels CR	5
Realtime Queue Size	10
Non-Realtime Queue Size	10
Mean Channel Holding Time	125 sec
Mean Realtime Queue Dwell Time	10 sec
Mean Non-Realtime Queue Dwell Time	10 sec

Table 3 - Simulation Parameter values

The parameter values chosen are used to reflect a more realistic condition of a cellular network. In our simulation, we varied these parameter values to study the various tradeoffs that exists among them and their effects on each class' performance, i.e. the dropping rates.

In figure (2), this scheme provides the three different priority classes with quality of service consistent with their priority. As can be seen from the figure, the highest priority class, the realtime handoffs, enjoys a lower probability of dropping than that of the second highest priority class, the non-realtime handoffs, and with the lowest priority class experiencing the highest blocking probability.

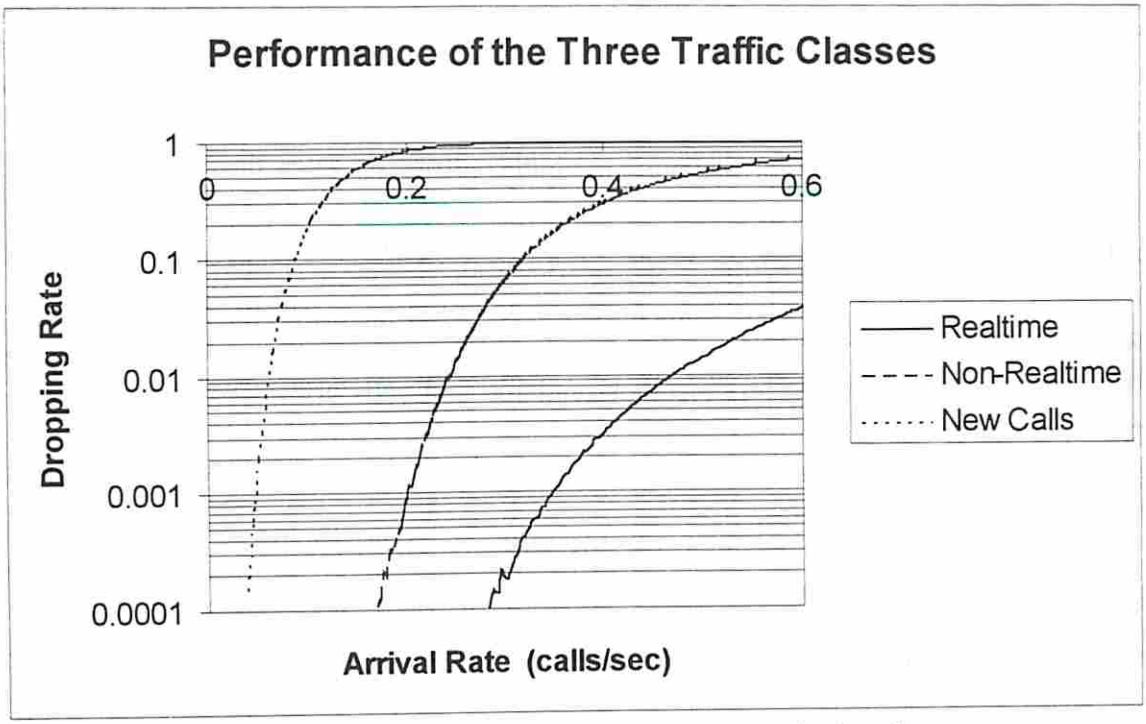

Figure 3 - Dropping rate for the priority classes

The different dropping rates experienced by the different classes is due to the guard channels scheme used which provides the different classes with different number of channels. When the number of available channels is greater than CH , no priority is given to any particular class over the others. But as the offered load to the cell increases, more calls are admitted, hence, the number of available channels decreases. As the number of available channels reaches the threshold point CH, new calls are blocked while both handoff classes' calls continue to be admitted. So the effective number of channels available to new calls is only C-CH, where C is the maximum capacity of the cell. For this reason new calls experience higher blocking probability than the other two classes. As the number of busy channels in the cell increases, both of the handoff classes are admitted and allocated channels until the number of busy channels reaches the second threshold CR. At this point non-realtime handoffs are not admitted but rather queued in the non-realtime handoff queue, while the realtime handoff calls are admitted. For this reason non-realtime handoffs experience higher dropping rates than realtime handoffs. Realtime handoff calls continue to be admitted and allocated channels as long as there are channels available. Once all channels are busy, realtime handoffs are queued in their associate queue. If there is a channel available and calls are queued, the scheme first checks if the realtime handoff queue is not empty. If the queue is not empty, the head of the queue is admitted and allocated the channel, otherwise the non-realtime handoff queue is checked and if not empty a queued call is allocated the channel.

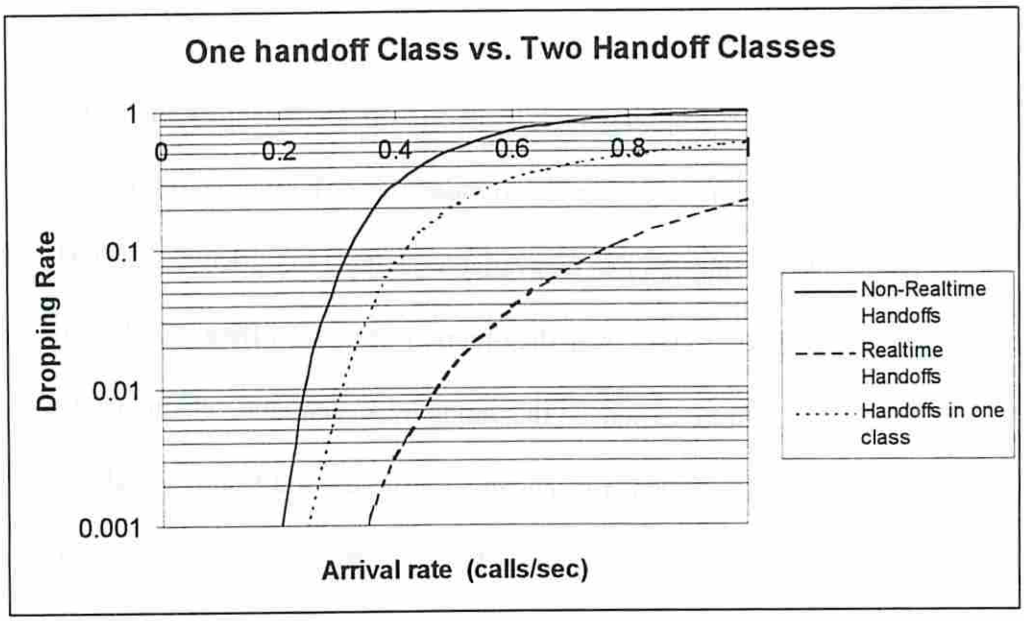

Figure 4 - Performance of handoff calls

A comparison between this scheme with the scheme presented in [5] which uses only one class for handoff calls is shown in figure (3). In this figure, dividing handoff calls into two separate classes yields a lower dropping probability for the highest priority class, the realtime handoffs than in the case of a single handoff class. This gain comes at the expense of a higher dropping probability for non-realtime handoffs when compared to the one class case. The improvement in realtime performance as seen from figure (3) is much higher than the degradation of the performance non-realtime handoffs. This amount of improvement or degradation is dependent on the choice of the guard channels threshold CH and CR.

2.4.1 Effect of Guard Channel Thresholds

The performance of the priority classes is dependent on the values of the thresholds. The more channels that are reserved for a class the better its performance. Since realtime and non-realtime handoffs share a reserved pool of channels when the available channels are below a threshold value CH, this threshold value profoundly effects the performance of the two classes. But out of that shared pool, CR channels are reserved exclusively for the realtime handoffs. It is clear that the greater the value of CR the better the performance of the realtime handoffs. On the other hand, the greater the value of CR the worse the performance of the non-realtime handoffs. Results of figures (4-5) confirm this intuition. When the value of CR is increased from 1 to 5 to 14 while the value of CH remains the same, the dropping rate of realtime handoffs is decreased by an order of magnitude. As the threshold value CR is increased, more channels are reserved for realtime handoff calls. On the other hand, this increase in the value of CR

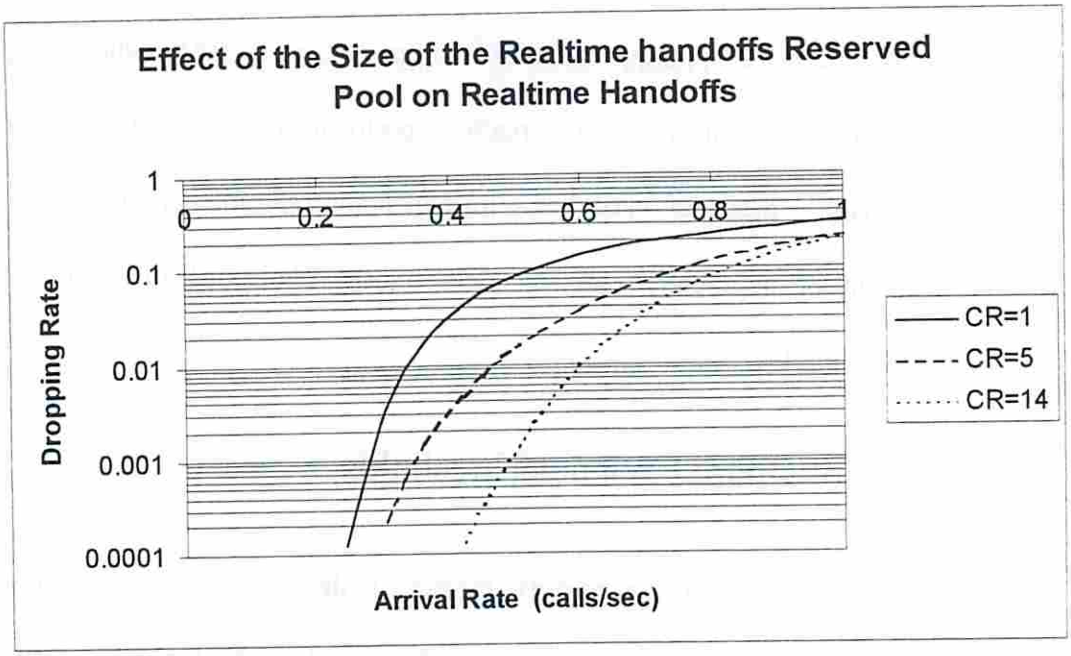

Figure 5 - Realtime performance as a function of CR

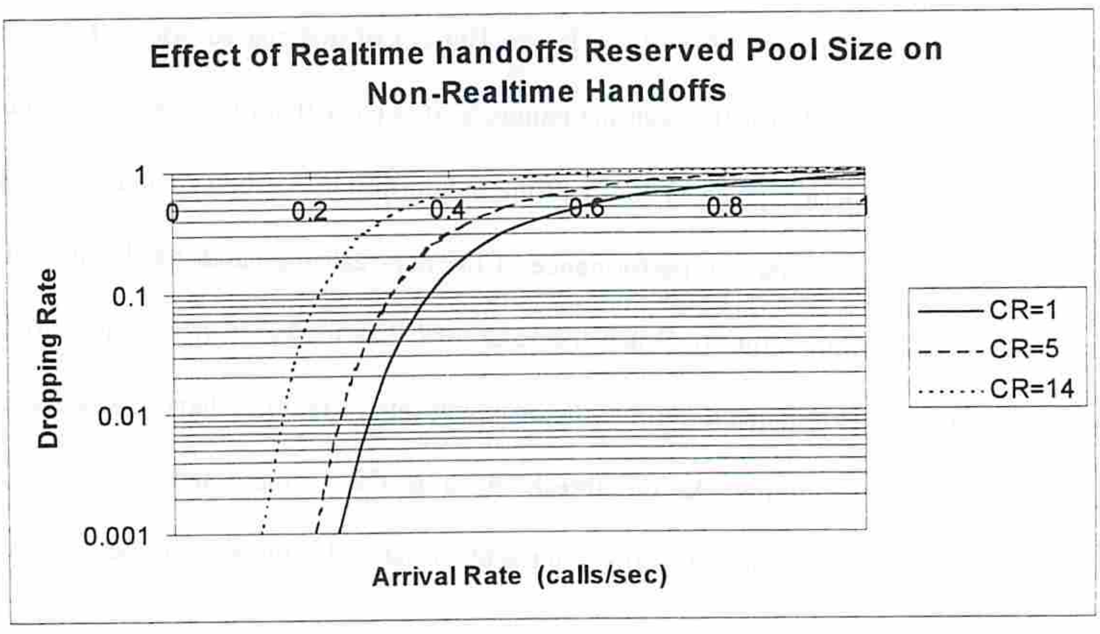

Figure 6 - Non-realtime performance as a function of CR

has the opposite effect on non-realtime handoffs. The more channels are reserved for realtime handoff calls, the less channels that are available to non-realtime handoffs, hence the higher dropping probability experienced by them.

Following the same reasoning, increasing the value of CH while CR remains the same would lower both of the handoffs dropping probability since they are allocated more channels. This gain is at the expense of the new calls, which will see a reduced pool of channels and hence suffers from a higher blocking probability. The situation is reversed when CH is decrease; new calls' blocking probability is decreased since the shared pool of channels has increased.

2.4.2 Effect of Channel Holding Time

The longer the time a call holds a channel, the longer another call has to wait to get that channel, when the system is congested. The shorter the channel holding time, the faster the channels would be available for other waiting calls. This would, intuitively, increase the carried load of the cell and hence reduces the dropping probabilities. In figure (6-8), we used three values for the channel holding times, 125, 167, and 250. The results confirm our intuition, the longer the holding time the higher the dropping probability. This gives an insight to what happens when the mobiles are highly mobile and fast, since faster mobiles tend to leave faster. A freeway cell for example, with fast users, would experience this lower dropping probability. But faster users as we mentioned before would require shorter queue dwell times too. So there is a tradeoff here between shorter queue dwell times and shorter channel holding times

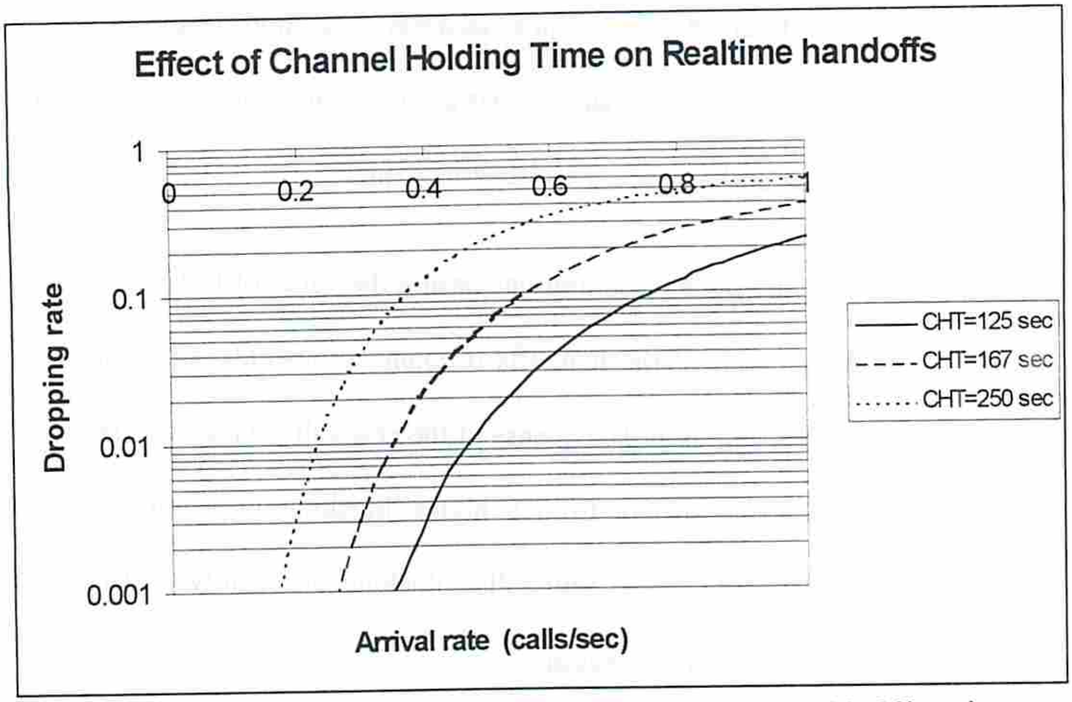

Figure 7 - Performance of realtime as a function of channel holding time

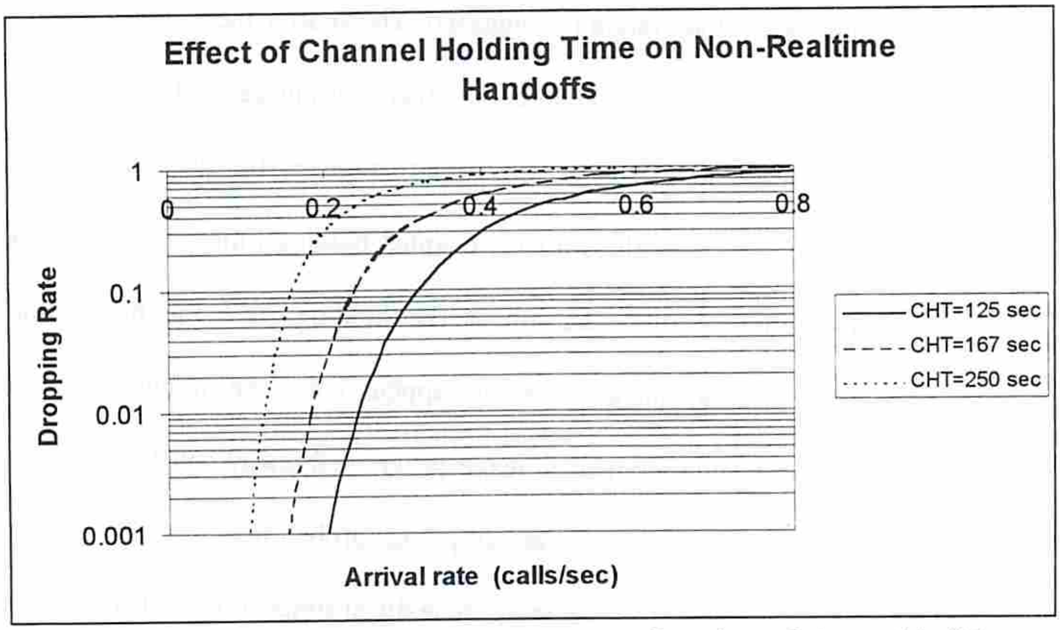

Figure 8 - Performance of non-realtime as a function of channel holding

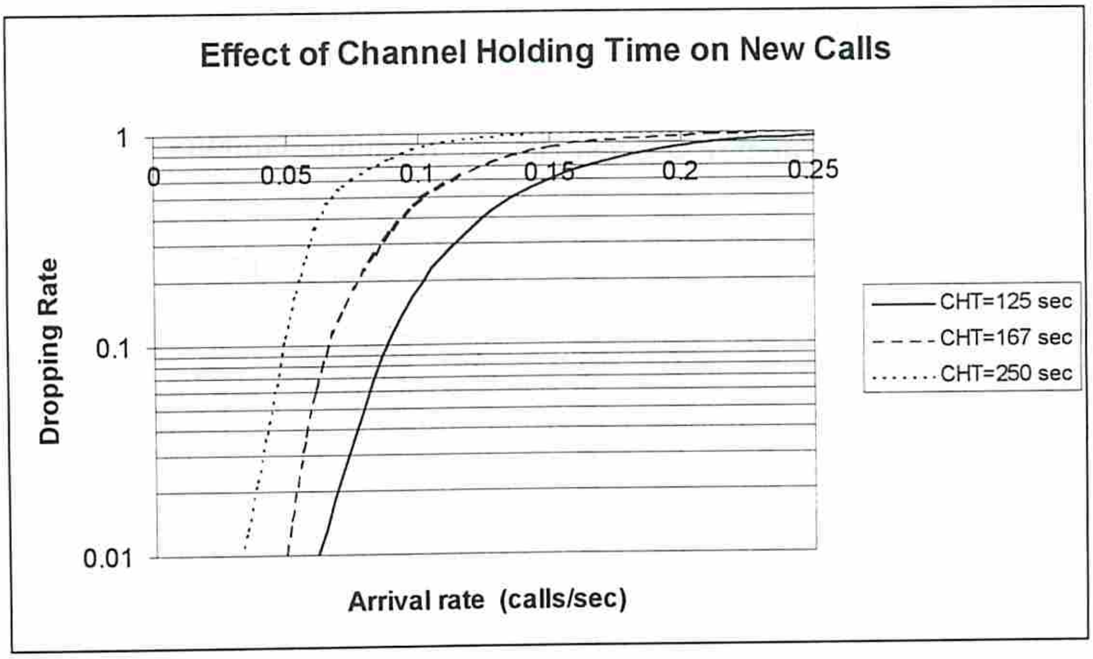

Figure 9 - Performance of New calls as a function of channel holding time

2.4.3 Effect of Dwell Time

A highly mobile user, fast moving user, would traverse the handoff area in short time; hence, the dwell time in the queue is short. In the case where the channel holding time is long, handoffs entering the queue would soon drop out since they would be out of the handoff area before a channel becomes available. Limiting the speed of the mobiles, or in other words, increasing the queue dwell time for the handoffs would intuitively increase the performance and the total carried load. In figures (9-10), we increase the dwell time from 10 seconds to 10000 seconds to exaggerate the situation for purpose of understanding the effect of dwell time on the performance, and also increase the queue size from 10 to 1000. This results in a large improvement as expected but only when the cell is lightly loaded. When the cell is heavily congested, e.g. lambda=0.4 fig 5, more calls are dropped due to a full queue and hence the high dropping probability.

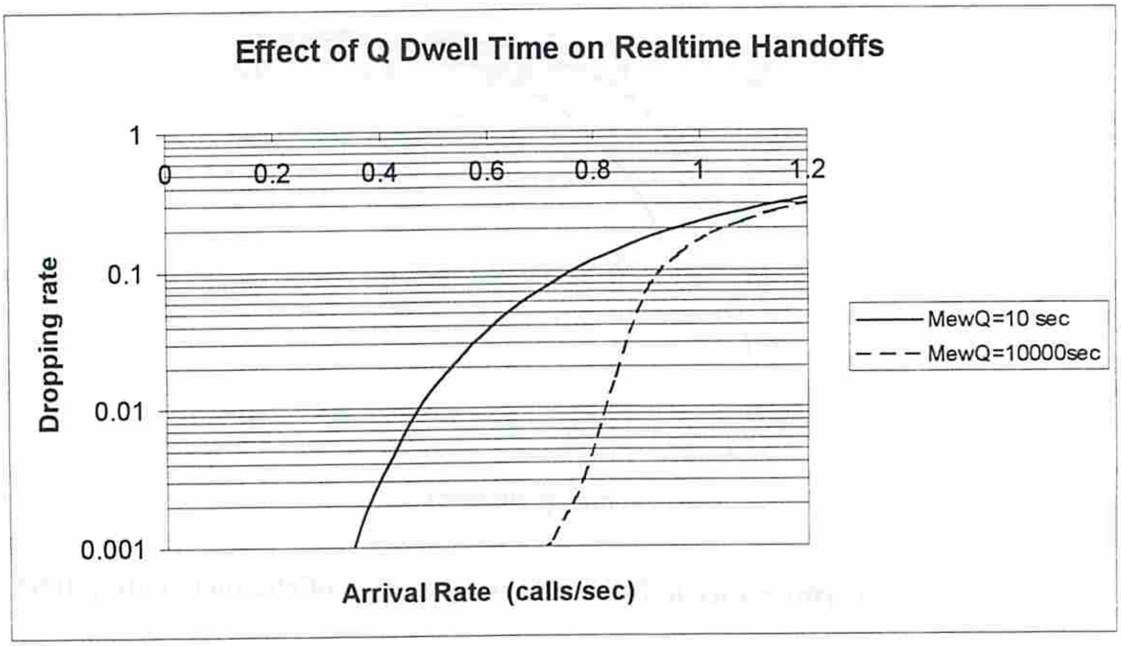

Figure 10 - Realtime performance as a function of queue dwell time

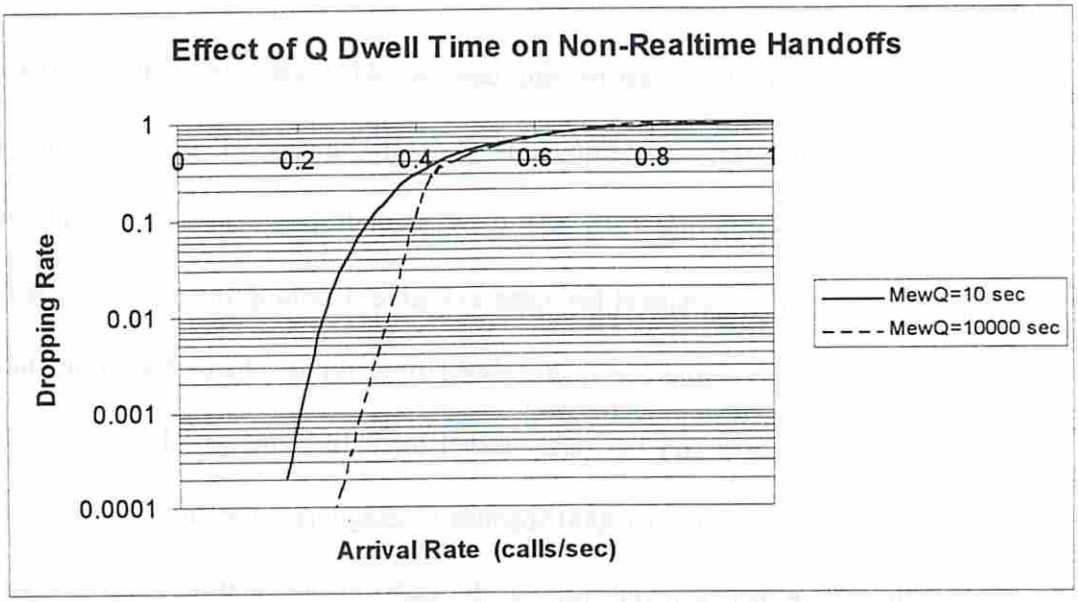

Figure 11- Non-realtime performance as a function of queue dwell time

2.5 Conclusion

Due to the sacristy of resources in the wireless environment, better management of these resources is required to increase the utilization of the network and to provide users with better performance. An admission control policy is needed to allocate channels of a wireless network cell to incoming calls such that the highest priority calls would experience better quality of service. This brings forth the need to classify incoming calls to a cell into priority classes. The classical way of classifying calls was to have two classes, handoffs and new calls. We have argued in this paper that a better classification of calls is what is based on the time constraints of the calls. Realtime handoffs are the one with the tightest time constraint and with more to lose in case of a channel drop. On the other hand, new calls cannot be classified based on time constraints since it is difficult to know whether a new call is realtime or non-realtime before assigning it a channel. From the point of view of a user, it is more tolerable to block a new call than to drop a handoff call. For these reasons, new call would be the lowest priority class.

We proposed a scheme based on the guard channel method to reserve channels to high priority classes. This generalized guard channel scheme when simulated resulted in a lower dropping probability for the highest priority class as compared with the one guard channel case proposed in the literature. This improvement came at an expense of a slightly higher dropping probability for the non-realtime handoff calls. Although in this paper we presented a case of three priority classes, a more general case accommodating N classes can be easily supported.

Chapter 3

Adaptive Guard Channel Scheme

3.1 Introduction

The nature of traffic loads in cellular networks is that it changes with time. We would expect more calls in morning as the commuters go to work, and in the evening as the commuters go home. We would also expect higher traffic loads when freeways are congested due to accidents or social events. This time variability of traffic loads provides network planners with an opportunity to take an advantage of so as to provide better service to users. This better service manifest itself as a quality of service metric that has to be met for each individual class of calls.

The driving objective behind the methods we propose in this chapter is to employ a scheme that adapts the number of reserved channels in the multi-guard channel scheme, introduced in chapter three, to traffic loads. In this adaptive scheme more resources are reserved for higher priority classes of calls when traffic loads are heavy and there is high competition among the several call classes to acquire these resources. On the other hand, more resources are made available to lower priority call classes when traffic loads are light and there is an abundance of resources available to high priority call classes. This adaptivity better utilizes the resources of the network to provide better overall performance for all call classes.

This chapter is organized as the following. Section II discusses the adaptive guard channel scheme. Section III discusses the deterministic optimal solution to the cellular system. Section IV discusses the simulation parameters and system assumptions. Section V compares the adaptive and deterministic optimal approaches. Section VI provides a summary and conclusion of this chapter.

3.2 Adaptive Guard Channel Scheme

The main objective of this scheme is to take advantage of the time-variant characteristic of the traffic loads to provide better performance for all call classes. This is achieved by adding more channels in the reserved pool of the higher priority call classes when traffic loads are heavy, thus providing these call classes with exclusive access to these additional channel which yields a lower dropping rate for these high priority calls. However, when traffic loads are light, channels are removed from the reserved pool of the high priority call classes and put in the shared pool of channels, thus lower priority call classes have access to more channels which yields a lower dropping rate for these low priority calls.

At the center of this scheme are two components, the metric by which to determine whether or not to activate the adaptive scheme and the window of time to measure this component. The proposed scheme presented here use the call-dropping rate as the performance metric measured over a sliding window time frame. Another important aspect of this adaptive scheme is the number of channels removed from the reserved channel pool when traffic loads are light and the number added when traffic loads are heavy. We investigate four such techniques, additive, ADD, multiplicative, MULT, additive increase multiplicative decrease, AIMD, and multiplicative increase additive decrease, MIAD.

3.2.1 Dropping Rate Metric

The main function of the performance metric in this adaptive scheme is to signal the occurrence of congestion or lack thereof in the network, and hence the necessity of increasing or decreasing the number of reserved channels for a certain call class. Although the arrival rate can be used as a metric to signify congestion in the network, the dropping rate provides a better estimate since it is the direct result of congestion. When the dropping rate is computed, the result provides a clear general measure of the current severity of the congestion in the network and a measure of the performance of the priority class under the present network conditions. The dropping rate also has the added advantage of being the quality of service criterion for the individual priority classes with which the current performance of the priority class should be compared to. In summary, the dropping rate as a metric provides the advantages of being an indication of congestion in the network and the performance of each class, which can readily be compared to the quality of service requirement which the adaptive scheme try to satisfy.

3.2.2 The sliding Window

Since the dropping rate is a vital component in the adaptive scheme, the measurement time window should result in measurements that correctly reflect the performance of the different priority classes. To correctly measure the dropping rate, the measurement interval or window, should not be too long so as to be slow in adapting to the changes in the traffic, nor be too short so as to be vulnerable to transient call drops. For example, in a measurement window of size 100 sec with mean arrival rate of 0.5, the mean number of calls arriving during that window is 50, therefore, any one call drop would be enough to trigger the adaptive scheme with quality of service constraint of 0.01. This single drop may be a transient event rather than signifying a state of congestion in the network. At the other extreme is the case where the measurement window is too long. Consider a window of size 10000 seconds with the same mean arrival rate and quality of service constraint. The mean number of calls arriving during this window is 5000 calls and hence 49 call drops will not trigger the adaptive scheme even if they occur during the last 100 seconds.

From this discussion we conclude that the measurement window should be large enough to be insensitive to transient call drops and small enough to quickly adapt to changes in the state of congestion in the network. The method we used for the adaptive scheme is a method where the dropping rate is measured over a measurement window of size W. This measurement window slides with time every time interval T. Each time the window slides, the dropping rate over the entire window is computed and a decision to change the number of reserved channels is made.

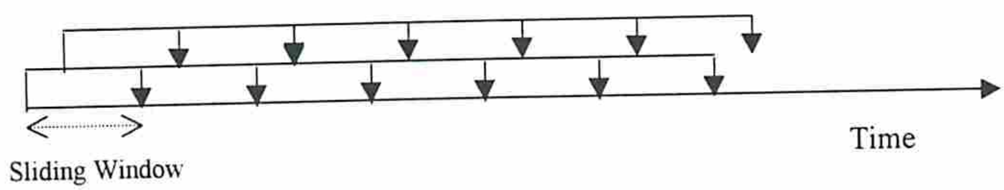

Figure 12 - Sliding window

This sliding aspect of the sliding window method provides the mechanism to detect the changes in the state of congestion in the network early, provided the sliding interval is chosen small enough.

3.2.3 Adaptive Techniques

When the measured dropping rate in a measurement window is either greater or less than the quality of service requirement, the adaptive scheme should react accordingly. When the dropping rate is greater than the quality of service constraint, this signifies a state of congestion in the network and the resources are scarce. In this case the adaptive scheme should react by reserving more channels for the high priority call class for whom the quality of service constraint has been violated. On the other hand, when the measured dropping rate for a class of calls is less than the quality of service constraint, this signifies the absence of congestion and the abundance of resources for this priority class. While this is the perceived status of the network as seen by this high priority class, the lower priority classes have a different perception. Lower priority classes experience higher dropping rate because of the limited number of resources available to these classes due to the number of reserved channels allocated for the high priority class. In this case, the adaptive scheme reduces the number of reserved channels for the high priority class and hence increasing the number of channels available to the lower priority classes.

There are several techniques that can be employed in the adaptive scheme to determine the number of channels to be removed from or added to the reserved pool each time the dropping rate of a class violates or satisfies the quality of service requirement. The techniques proposed in this chapter use two approaches to accomplish this task, the additive or linear approach and the multiplicative or exponential approach. When the additive approach is used, the adaptive scheme changes the number of reserved channel by adding or removing a fixed number of channels. On the other hand, the multiplicative approach adds double or removes half of the current number of channels in the reserved pool. We investigated the four possible techniques that employ the two approaches, the additive, multiplicative, additive increase multiplicative decrease, AIMD, and multiplicative increase additive decrease, MIAD.

3.2.3.1 Additive

The additive technique adds one channel to the reserved pool when the dropping rate is higher than the quality of service requirement, while removing one channel from the pool when the dropping rate is lower, and hence the additive name.

```
If   DropRate ≥ QoS   Then
          ReservedPool = ReservedPool + 1
Else   ReservedPool = ReservedPool - 1
```

This technique is a conservative one in adding and removing channels; thus it is slow to react to a quickly building up congestion in the network. On the other hand, this additive nature reserves as few channels as possible to higher priority classes, and hence provides the lower priority classes with access to more resources.

3.2.3.2 Multiplicative

The multiplicative technique doubles the number of channel in the reserved pool when the dropping rate is higher than the quality of service constraint, while reducing the number of channels in the pool by half when the dropping rate is lower, and hence the multiplicative name. This technique assumes that initially the number of channels in the reserved pool is at least one.

If DropRate $\geq$ QoS Then

ReservedPool = ReservedPool $\times$ 2

Else ReservedPool = ReservedPool $\div$ 2

This multiplicative nature of this technique makes it very fast in reacting to the changes in the state of congestion of the network.

3.2.3.3 AIMD

The additive increase multiplicative decrease AIMD technique combines the slow reaction to the increase in congestion in the network from the additive technique with the fast reaction to a decrease in congestion from the multiplicative technique.

If DropRate $\geq$ QoS Then

ReservedPool = ReservedPool + 1

Else ReservedPool = ReservedPool $\div$ 2

3.2.3.4 MIAD

The multiplicative increase additive decrease MIAD technique combines the fast reaction to an increase in congestion in the network from the multiplicative technique with the slow reaction to a decrease in congestion from the additive technique.

If DropRate $\geq$ QoS Then

ReservedPool = ReservedPool $\times$ 2

Else ReservedPool = ReservedPool - 1

3.3 The Deterministic Optimal Solution

The multi-guard channel scheme described in chapter three reserves a number of channels exclusively for the high priority class, and reserves a number of channels for the two highest-priority classes. When the arrival process is stationary, the minimal number of channels reserved in the two reservation pools for that specific arrival rate, which satisfies the two quality of service constraints for the two highest-priority classes, yields the optimal performance for the lowest priority class. This minimal reservation assignment provides the lowest priority class with access to the most number of channels and hence results in the lowest dropping rate for that particular arrival rate.

The mathematical formulization of this optimization problem is as follows:

CH is the total number of reserved channels

CR is the total number of reserved channels for the realtime class

C is the total number of channels in the cell

QoS_{rt} is the quality of service constraint for the realtime class

QoS_{nr} is the quality of service constraint for the non-realtime class

Minimize (CH) Subject to :

$$P_{rt} \leq QoS_{rt}$$

$$P_{nr} \leq QoS_{nr}$$

$$0 \leq CR \leq C$$

$$0 \leq CH \leq C$$

$$CR \leq CH$$

3.4 Analysis

3.4.1 System Assumptions and Simulation Parameters

We investigated a cell with a capacity of 35 channels. As described in chapter three, the multi-guard channel scheme is used. There are three priority classes, the realtime handoff calls class, which is the highest priority class, the non-realtime handoff calls class, and the new calls class, which is the lowest priority class. The call arrival process is assumed to be poisson with mean arrival rate equal for all priority classes. The channel holding times for all classes are assumed to be exponential with mean 100 seconds. All call requests are assumed to require one channel. The quality of service constraint for the two high-priority classes, the realtime and non-realtime handoffs, are chosen to be equal to 0.01.

3.4.2 Analysis

3.4.2.1 The adaptive techniques

Dropping Rate	ADD	AIMD	MIAD	MULT
> QoS	+ 1	+ 1	× 2	× 2
≤ QoS	- 1	÷ 2	- 1	÷ 2

Table 4 - Channel addition and removal by different techniques

Table (3) compares the reserved pool control approaches among the different proposed adaptive techniques. The different techniques impact the performance of the three priority classes differently. Figure (12) shows the performance of the realtime handoff calls class under the four adaptive techniques. All techniques result in two periods of performance improvement for the realtime handoff calls class. The first period starts around mean arrival rate of 0.23 and the second around mean arrival rate of 0.4. The first period is the result of the non-realtime handoff quality of service requirement violation, which triggers the adaptive scheme to add more channels into the handoffs reserved pool, i.e. increases CH. Although this increase in the size of the reserved pool was done to satisfy the quality of service requirement of the non-realtime handoffs, the realtime handoff calls benefit as well, resulting in the first performance improvement

period. The second improvement period is a result of the realtime handoff class quality of

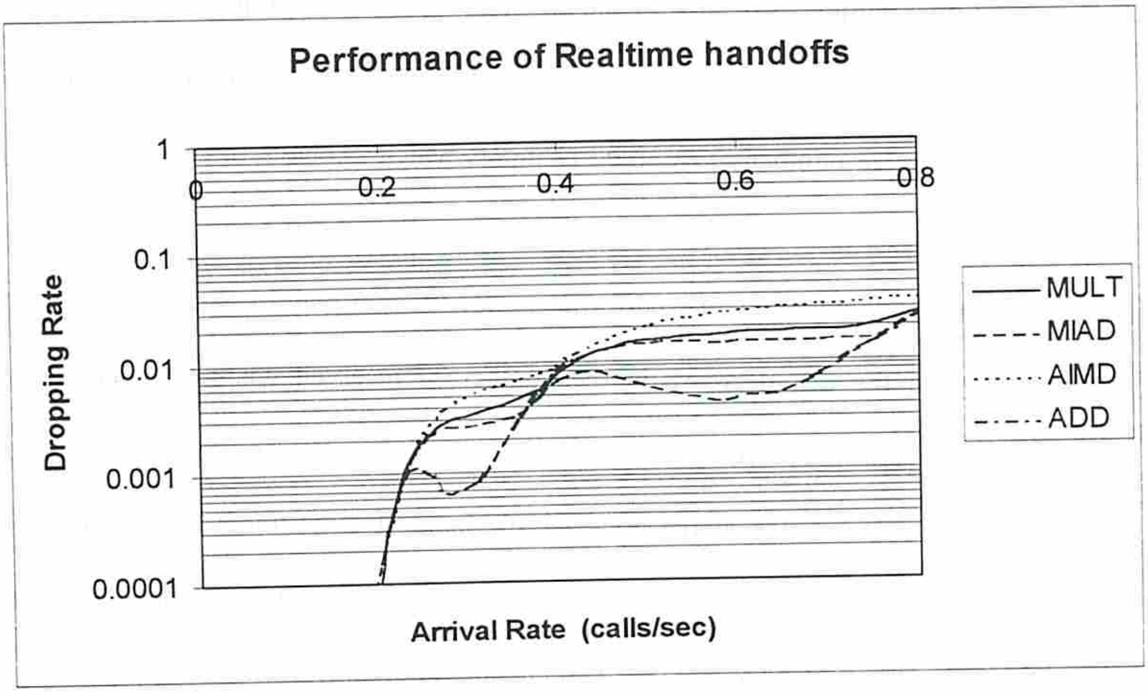

Figure 13- Performance of realtime under different techniques

service requirement violation, which triggers the adaptive scheme to add more channels in the realtime handoffs exclusive reserved channel pool, i.e. increases CR. This increase in the number of exclusively available channels reduces the number of call drops and hence a lower dropping rate.

The different adaptive techniques result in difference dropping rates for all classes. In the case of the realtime handoff class, the multiplicative increase additive decrease technique, MIAD, yields the best performance, while the additive increase multiplicative decrease, AIMD, yields the worst. Since the MIAD technique doubles the current number of channels in the reserved pool when the quality of service requirement is violated, the result might be more than the number of channels needed to lower the dropping rate below the required value. Moreover, the MIAD technique employs an additive decrease approach that is slow in reducing the number of reserved channels to the minimum number that satisfies the quality of service requirement. The slow reduction and the fast increase of the number of reserved channels in the MIAD technique result in a higher average number of reserved channels which explains the low dropping rate as compared to the other techniques. The analysis can be reversed for the case of the AIMD technique that yields the worst performance for the non-realtime handoff calls. The slow increase and the fast decrease in the number of reserved channels result in a lower average number of reserved channels and a higher dropping rate.

The performance of the non-realtime handoff class under the different adaptive techniques is shown in figure (13). This class has only one performance improvement period, since there is only one reserved channel pool which this class can draw channels from. The figure shows two distinct behaviors for the dropping rate under the four

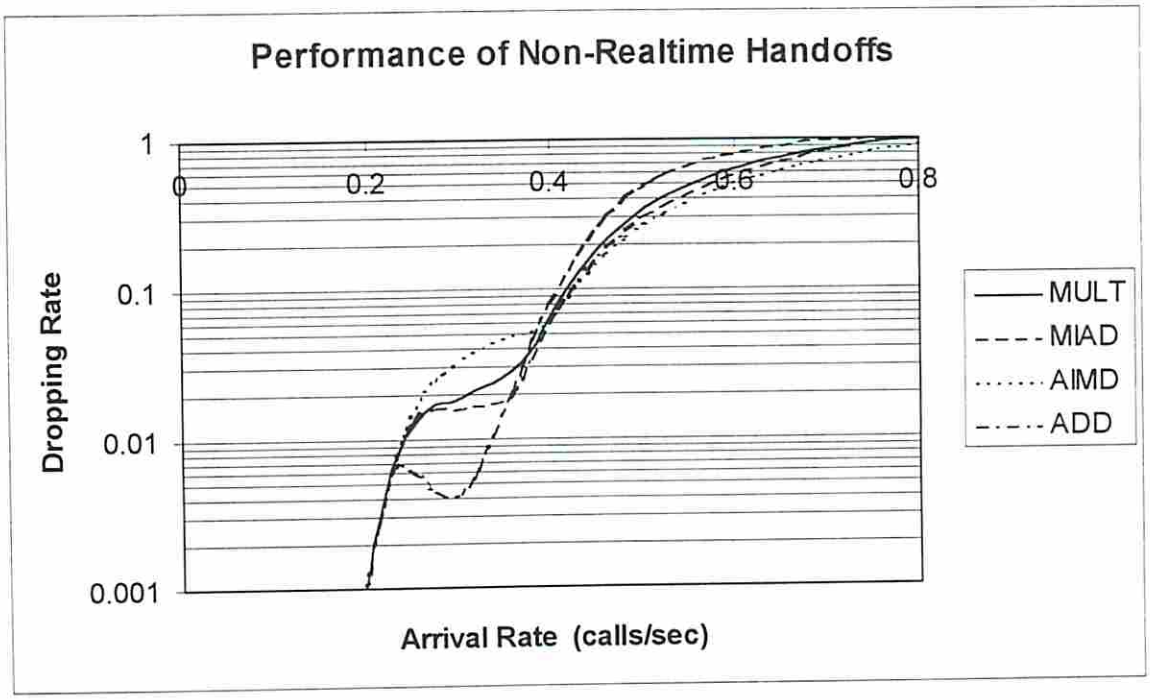

Figure 14 - Performance of non-realtime under different adaptive techniques

schemes when the mean arrival rates less and greater than 0.4. The performance analysis for this class for the first behavior, i.e. mean arrival rate less than 0.4, follows the same lines as the one done for the realtime handoff class. The MIAD technique provides the best performance since it yields a higher average number of channels reserved, while the AIMD the worst since it yields a lower average number of channels reserved. The second behavior of the dropping rate for the non-realtime handoff class is the opposite of the first one. When the mean arrival rate is greater than 0.4, the quality of service requirement for the realtime handoff class is not satisfied triggering the adaptive scheme to add more channels to the realtime handoffs exclusive pool. The larger the size of this pool, the smaller the number of channels that are available to non-realtime handoffs and new calls.

Therefore, the MIAD which reserves a higher average number of channels for the realtime handoff calls would result in a higher dropping rate for the non-realtime handoffs as shown in figure (13). The AIMD has the opposite effect since it reserves a smaller average number of channels for the realtime handoffs, which results in a lower dropping rate for the non-realtime handoff calls.

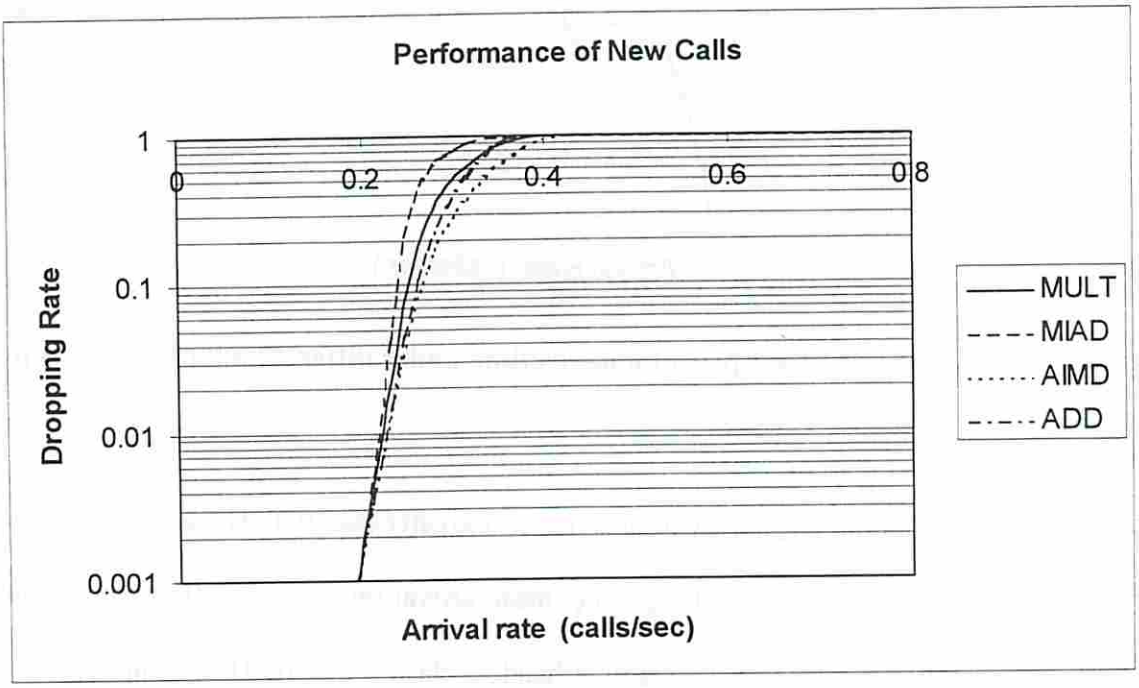

Figure 15- Performance of new calls under different adaptive techniques

The performance of the new call class has a similar analysis as that of the non-realtime handoffs in their second behavior when, the adaptive scheme is providing the realtime handoffs with a better performance at the lower priority classes' expense. When the mean arrival rate is between 0.2 and 0.4, the quality of service requirement for the non-realtime handoff class is not satisfied triggering the adaptive scheme to add more channels to the handoffs reserved pool. The larger the size of this pool, the smaller the number of channels that are available to new calls. Therefore, the MIAD which reserves a higher average number of channels for the handoff calls would result in a higher dropping rate for the new calls as shown in figure (14). The AIMD has the opposite effect since it reserves a smaller average number of channels for the handoffs classes, which results in a lower dropping rate for the new calls class.

The multiplicative and additive techniques result in a close performance in each class, although there is a performance advantage for the additive technique. The reason for this performance similarity is the balance between the approach for the increase and the decrease of the number of reserved channels. The use of the same approach for the increase and decrease results in an average number of reserved channels that is similar in both techniques. The performance advantage, which the additive technique enjoys over the multiplicative technique, is due to the fact that we chose the fixed number of channels increased or decreased by the additive technique to be one, which allows the additive technique to reserve up to all channels in the cell. On the other hand, the multiplicative technique can only reserve to the largest power of two number equal to or less than the total number of channels in the cell. For example, the maximum number the multiplicative technique can reserve in our model is 32, leaving 3 channels available for other classes to have access to.

3.4.2.2 The Adaptive Scheme vs. The Deterministic Optimal Solution

Since the optimal solution we are seeking is for a stationary process, it is easy to find it using an exhaustive search in a matrix of all possible solutions, especially since the number of channels in a cell is usually small. The optimal solution was obtained by numerically finding each class' dropping probability for all possible reserved channels assignments using the solution of the markov chain shown earlier. Using exhaustive search we could find the reserved channel assignment that would satisfy the optimization problem formulated earlier. The solution of this optimization problem was done for a range of arrival rates that corresponds to the adaptive scheme simulation results. As the arrival rate is increased, only the realtime call's quality of service constraint can be satisfied, at which point a solution is found for a second optimization problem with only one quality of service constraint, that is of the realtime class.

The adaptive scheme used in figures (15-17) is a multiplicative increase additive decrease, MIAD, with sliding window of size 3000 time slots and a sliding interval of 1000 time slots. The performance of the realtime class in the adaptive scheme shows a slight improvement over that of the optimal solution. That is due to the fact that the adaptive scheme takes advantage of the random characteristics of the arrival process and adapts accordingly to yield a better dropping rate. The multiplicative increase additive decrease technique, MIAD, also contributes to the dropping rate improvement by way of adding double the number of channels currently in the reserved channel pool when the quality of service constraint is not satisfied. The number of channels newly reserved may be more than what is needed to satisfy the quality of service constraint, hence an improvement in the dropping rate.

The figure also shows a leveling off effect around mean arrival rate of 0.2 and 0.3. The first leveling off is due to the increase in the reserved channel pool of both handoff classes, i.e., an increase in CH. At mean arrival rate of 0.2, the congestion in the cell is such that the quality of service constraint for the non-realtime handoff class is violated and hence an increase in the number of channels reserved in the handoff reserved pool is warranted. Recall from Chapter three that the multi-guard channel scheme divides the channels in a cell into three categories. (C - CH) channels are available for all classes of calls, (C - CR) channels are available for the non-realtime handoff calls, and all the channels are available for the realtime handoff calls. Therefore, when the dropping rate of the non-realtime handoff class violates the quality of service constraint CH is increased. This increase increases the number of channels available to the two handoff calls classes, and reduces the number of channels available to the new call class, hence improving the dropping rate for the realtime handoff class, and leveling off the dropping rate for both handoff classes.

The second leveling off starts at 0.33, at which point the quality of service constraint for the realtime handoff class is violated due to the increased load in the cell. This violation prompts the increase in the number of reserved channels for the realtime handoff class, which in turn reduces the number of available channels for both the non-realtime handoff and the new call classes. As the mean arrival rate increases so does the number of reserved channels for the realtime handoff class. The figure shows that at mean arrival rate of 0.7 the cell is in a heavy state of congestion with all channels reserved to the realtime handoff calls.It is important to note that the main advantage of the adaptive scheme over the deterministic solution is not in the dropping rate improvement shown, although the results show some improvement there for the high priority class. The main advantage is rather in the fact that the optimal solution is based on the assumption of poisson arrival process which the solution of the markov chain is based on. The adaptive scheme is arrival process independent and hence can be used with an arrival process of any distribution. This is especially useful when the arrival process is unknown, as it is the case in the many cellular networks.

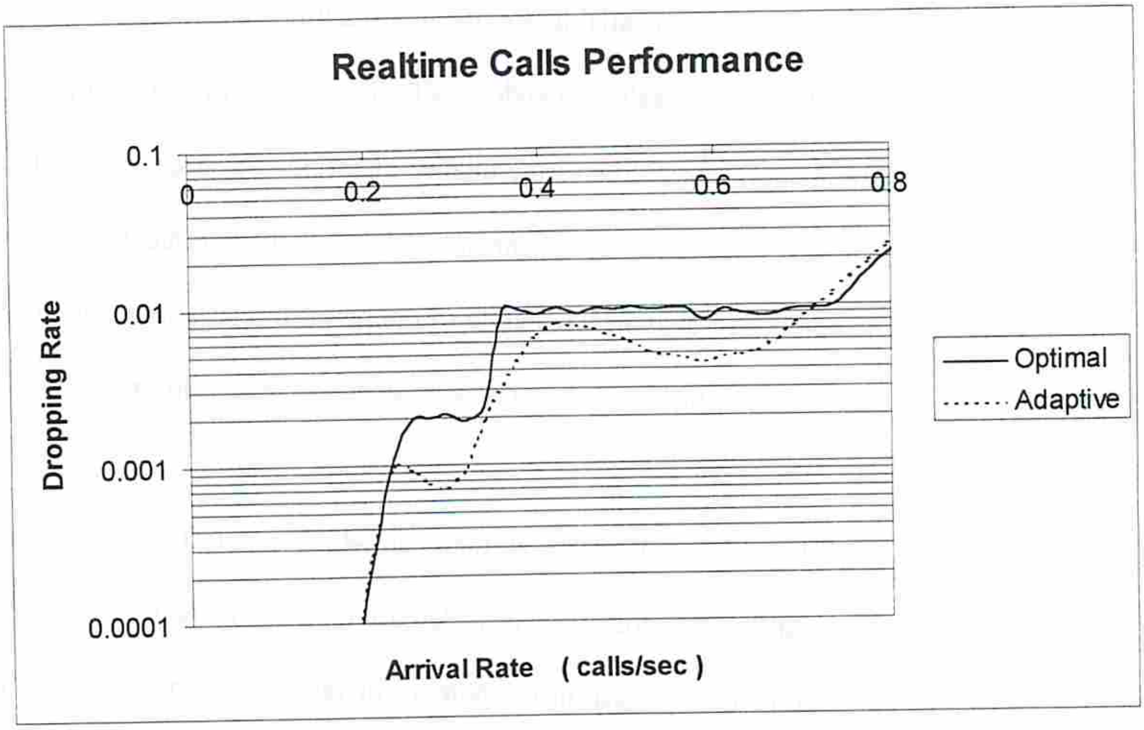

Figure 16- Adaptive vs. optimal realtime performance

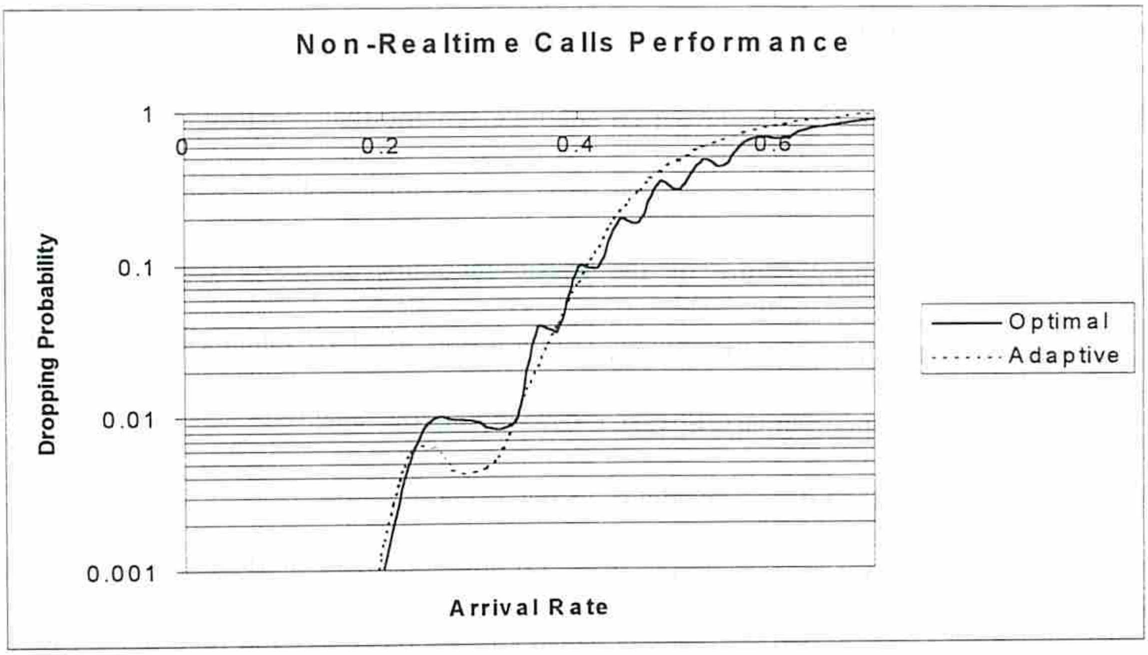

Figure 17- Adaptive vs. optimal non-realtime performance

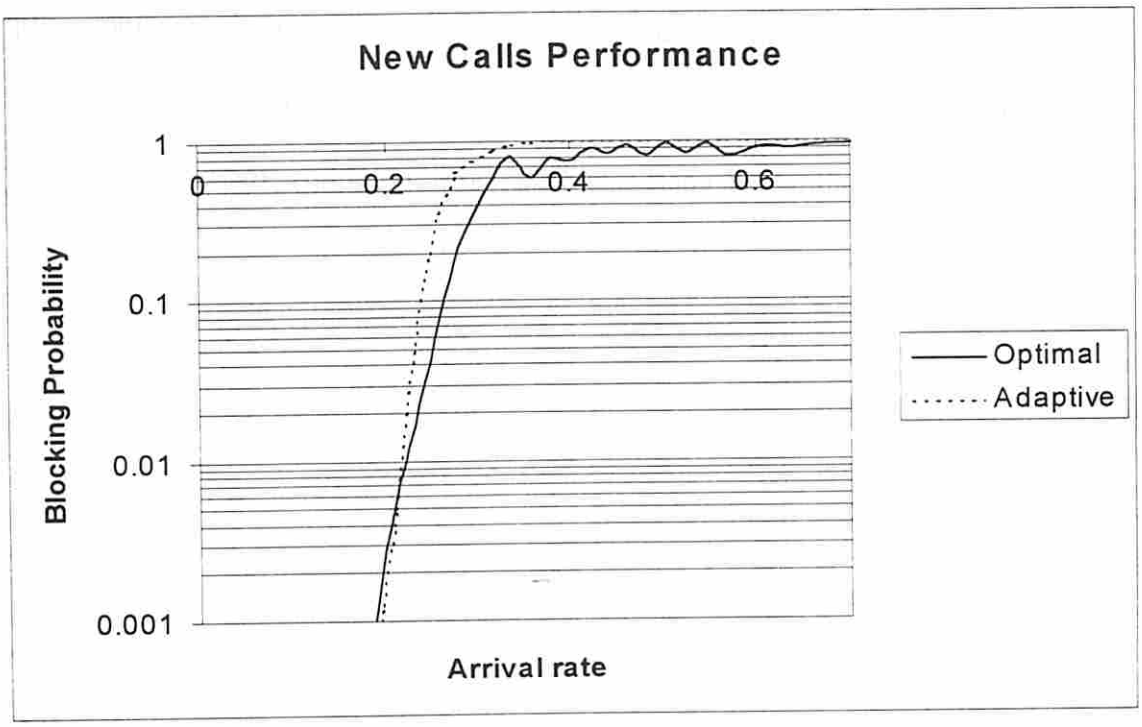

Figure 18- Adaptive vs. optimal new call performance

3.5 Conclusion

In this chapter we proposed an adaptive scheme that is based on the multi-guard channel scheme developed in chapter three. The adaptive scheme would take advantage of the non-stationary nature of the traffic incoming to a wireless network. When compared to an optimal deterministic solution, the adaptive scheme performed as well, although in a non-stationary arrival environment, the adaptive scheme is expected to perform better due to its independence of the arrival process.

Several adaptive techniques were proposed to control the number of channels added to and removed from the reserved pool. We proposed and investigated four such techniques, additive, multiplicative, additive increase multiplicative decrease, and multiplicative increase additive decrease. When we compared the performance of the different priority classes under these techniques, results show than the multiplicative increase additive decrease yield the best performance since it is fast to react to heavy congestion by adding channels and very slow to remove these extra channels when congestion has eased.

Chapter 4

Multi-Channel Call Admission

4.1 Introduction

With the increasing demand for communication and connectivity, more applications are being designed to incorporate video, voice, and data. The increase in the demand for multimedia applications and connectivity will highly impact wireless networks due to the increasing popularity of tetherless communication. Multimedia applications are known for their high bandwidth requirements, and therefore would require several channels in the wireless networks to meet their quality of service requirements. Data communications with high-speed modems can also require several radio channels to yield a transmission speeds compatible with that of the modem, as is the case of the Personal Handy-Phone System network, PHS, which currently provides data communications at a speed of 32 Kbps.

These multi-channel requirement calls, in a cellular network with a limited capacity of channels, can suffer from starvation if no new scheme is employed to help meet their quality of service constraint, i.e. their maximum dropping rate. This case of starvation is caused by the network's bias against multi-channel calls in favor of single-channel calls in call admission. This can be explained as follows, as the arrival rate increases more calls are accepted, and therefore few channels are available. At high arrival rates, when these available channels are less than the multi-channel calls' requirements, only the single-channel calls are accepted and the multi-channel calls are dropped. At these high arrival rates, once a call terminates and its channel becomes available, a single-channel call can be immediately accepted while a multi-channel call has to wait for a number of channels to be available. Therefore, single-channel calls continue to be accepted and multi-channel calls continue to be dropped.

4.2 Multi-Channel Calls Admission Control

Schemes are needed to reduce the bias against multi-channel calls and provide a quality of service to this class. We define the quality of service for the multi-channel call class as the maximum dropping rate this class can tolerate. Multi-channel calls can be classified based upon the number of channel requested per call into two classes, multi-rate calls and fixed multi-channel calls. Multi-rate calls are calls that request a number of channels from a range of possible values. Each multi-rate call can request a different number of channels based upon the client application and the type of call it generates, and the speed of the communication hardware available to the mobile unit. Fixed multi-channel calls are calls that request a fixed number of channels for the entire multi-channel call class. In this work, we are not considering multi-rate calls, it will be considered in chapter six, but rather our focus will be on fixed multi-channel calls. For the remainder of this chapter we will use multi-channel calls to refer to fixed multi-channel calls.

There are two types of multi-channel calls in a cellular network, data connections requiring higher bandwidths like FTP file transfer and video communications. In this chapter, we consider a multi-channel call class that is of the first type. We assumed the multi-channel calls are non-realtime data communication connections both handoffs and new calls. We did not consider the case of both the realtime and the non-realtime calls being multi-channel type calls because in that case the problem is simply a scaled version of what we have considered in chapter four.

4.2.1 Adaptive

The adaptive scheme introduced in chapter four can be useful in improving the dropping rate of the multi-channel call class. In this scheme, the cell has two pools of reserved channels. One reserved channel pool is exclusively reserved for the single-channel realtime handoffs, and the other is shared by the two classes of handoff calls, the multi-channel non-realtime handoff calls and the single-channel realtime handoff calls. As the traffic load increases, the dropping rate of the multi-channel call class also increases, violating its quality of service constraint. At this point the adaptive scheme is triggered to increase the number of channels reserved in the shared pool. This increase in the number of reserved channel lowers the dropping rate to the level demanded by the quality of service constraint.

The performance improvement from this adaptive scheme is limited by the number of channels required by each multi-channel call and the channel capacity of the cell. The larger the number of channels requested by each multi-channel call the smaller is the expected improvement in dropping rate. The reason for this is that the larger the number of requested channels the stronger the bias against multi-channel calls.

4.2.2 Reserved Channels

One technique to provide multi-channel calls with a lower dropping rate is the guard channel technique. In this technique, a number of channels are reserved exclusively for the multi-channel calls' class. When the traffic load is heavy and there are only few channels available for use, the multi-channel calls can use channels from their exclusive reserved pool. For this technique to have a major impact on the multi-channel class' dropping rate, the reserved channel pool has to be large enough in proportion to the number of channels the multi-channel calls request. For example, if the multi-channel calls require 8 channel per call, reserving 2 channels for the multi-channel class will have minimal impact. On the other hand, reserving a large number of channels would result in an increase in the number of accepted calls from this class and subsequently a decrease in the number of call drops, hence a lower dropping rate.

The main disadvantage of reserving a large number of channels for one traffic class is that it would have a huge negative impact on the dropping rates of the other traffic classes in the network. Since a large number of channels are reserved exclusively for one class, less channels are available and accessible to these other classes resulting in an increased dropping rates. Therefore, the size of the reserved pool of channels for the multi-channel call class needs to be large enough to have an impact on the dropping rate yet small enough not to increase the dropping rate of the other call classes in the network.

Reserved channel technique can be used as an extension to the adaptive scheme to investigate its impact on the performance of the other priority classes. A comparison with pure adaptive scheme would give an insight to the relative effectiveness of this scheme in providing the multi-channel class with better performance while not degrading the performance of other classes. A reserved channel extension of the adaptive scheme would reserve a number of channels in addition to the two reserved pools of channels for realtime and handoff calls as in the pure adaptive scheme.

4.2.3 Best Effort

At higher arrival rates, there are few available channels for incoming calls to use. Multi-channel calls not finding enough channels to satisfy their requests will have to be dropped. In some cases multi-channels calls are dropped while there are available channels but not enough to satisfy the multi-channel request. The best effort scheme targets this situation. When a multi-channel call arrives to the cell requesting M channels, the cell grants the call all M channels it requests if the number of available channels is larger than M, and grants the call all available channels if the number of available channels is less than M.

There are two types of multi-channel calls in a cellular network, data connections requiring higher bandwidths like FTP file transfer and video communications. In the case of high bandwidth data communication like file transfer, the best effort technique only effects the channel holding times for the calls. In other words, if a file transfer call takes 5 minutes using 32 Kbps or 4 channels, it would take 10 minutes to transfer the same file at 16 Kbps or two channels. Therefore, although the best effort technique improves the performance of the multi-channel class, it causes the mean channel holding time to increase. This increase in the mean channel holding time could have a negative effect on the user who will pay more for this prolonged call, assuming a charging structure based upon call duration.

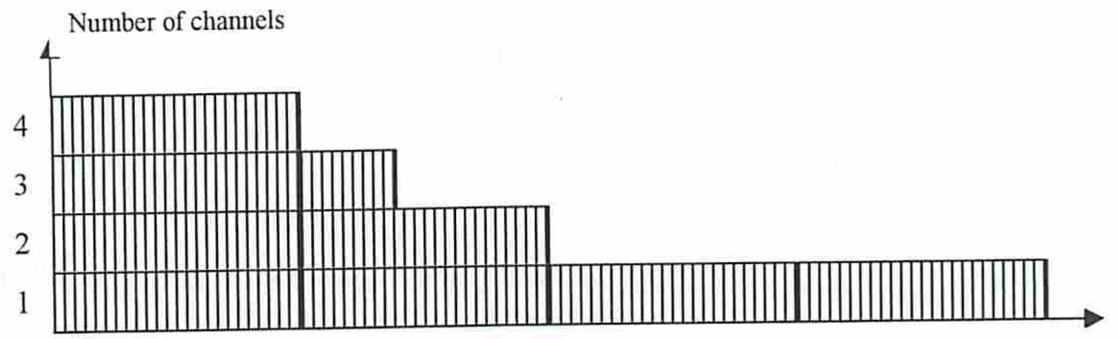

Figure 19 - Number of channels vs. channel holding time in Best effort

The second type of multi-channel calls is the video-based calls like video telephony. This type of calls can be coded in a tiered coding, where each level of resolution can be transmitted over a separate channel. This coding scheme facilitates the use of the best effort technique by dropping the higher resolution codes when there are not enough channels for the call. Therefore, the best effort technique effects the resolution of the transmitted video and not the channel holding time.

4.3 Analysis

4.3.1 System Assumptions and Simulation Parameters

We considered a cell in a cellular network with capacity of 35 channels. Four traffic classes arrive to the cell following a poisson arrival process, realtime handoff calls, non-realtime handoff calls, single-channel new calls, and multi-channel new calls. The mean arrival rate for the realtime handoffs, non-realtime handoffs and the aggregate new calls are set to be equal. We chose the realtime handoff calls to be single-channel calls, while the non-realtime handoff calls be multi-channel calls. The multi-channel calls, whether non-realtime handoffs or new calls request 4 channel per call. The channel holding time for all calls is assumed to be exponentially distributed with mean 100 seconds. The multi-channel guard channels were set at 4 channels.

As stated earlier, in this work we chose to focus on the data-communication-type multi-channel calls. Since employing best effort with this type of calls increases the channel holding time, in our simulation we adjust a call's channel holding time when the call is admitted using best effort by increasing it proportionally to the number of channels requested to channel granted. For example, if a multi-channel call requesting four channels with a channel holding time of 100 seconds is granted 2 channels, then the adjusted channel holding time is 200 seconds.

4.3.2 Results

Using the network model described above, we studied the effect of the multi-channel calls on the performance of the different call classes in the cell. For this comparison we use the adaptive multi-class priority scheme developed in chapter four as the admission control policy. The results of this study are analyzed in sub-section 1. We also compare three adaptive scheme variations for multi-channel non-realtime handoff calls' performance improvement. These three schemes are pure adaptive, adaptive reserved and adaptive best effort. Adaptive reserved is the adaptive scheme with a number of channels reserved exclusively for the multi-channel non-realtime handoff class. Adaptive best effort is the adaptive scheme with the best effort technique extension.

4.3.2.1 Multi-Channel vs. Single-Channel

As discussed earlier, the multi-channel nature of the non-realtime handoff calls effect the performance of all other classes in the network. This is evident from the figures below. The non-realtime handoff class' performance degrades noticeably, where the dropping rate is several orders of magnitudes worse. This can be explained by the bias phenomenon, where the cell favors the single-channel request calls over the multi-channel requests. The increase in the multi-channel dropping rate triggers the adaptive scheme to reserve more channels for the two-handoff classes. Because of the bias against multi-channels calls, the adaptive scheme does not have an impact in improving its performance. As the traffic load increases and as the number of channel requests increases, fewer channels are available resulting in worse performance for the multi-channel handoff class. The distributions of number of reserved channels for the two-handoff classes, CH, in the single-channel and the multi-channel cases shown in figures (20 - 21) support this explanation. The value of CH is increased by the adaptive scheme at a lower mean arrival rate, 0.042, for the multi-channel case than that of the single-channel case, 0.189, due to the higher dropping rate of the multi-channel case which violated the quality of service constraint at that point.

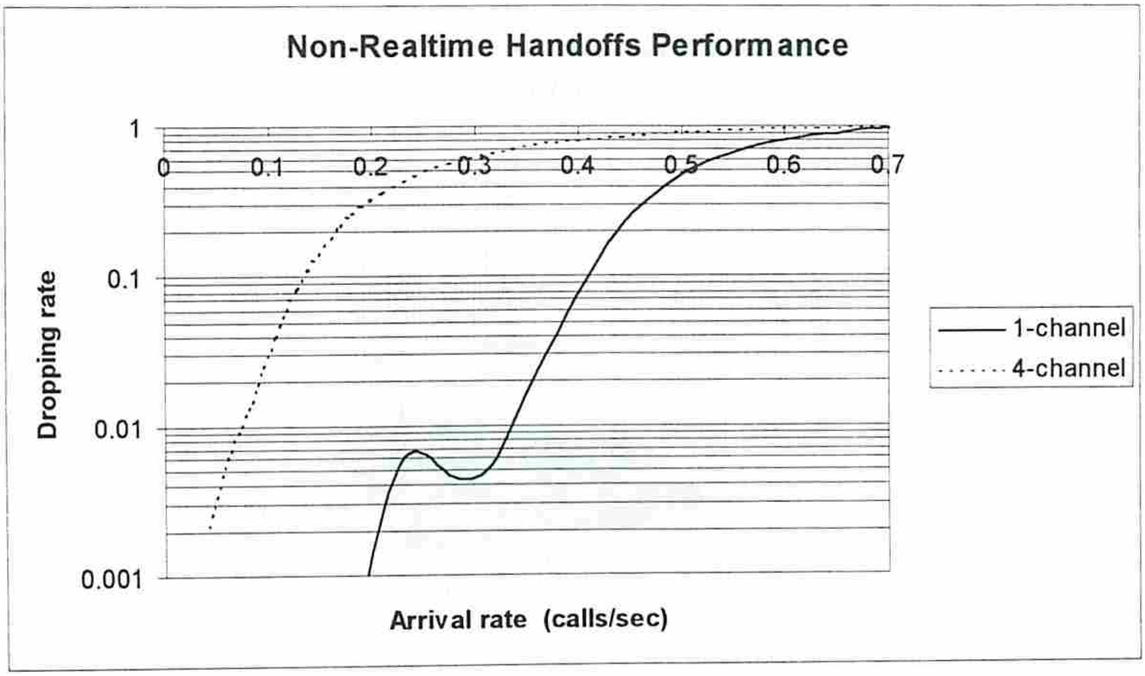

Figure 20 - single vs. multi-channel non-realtime performance

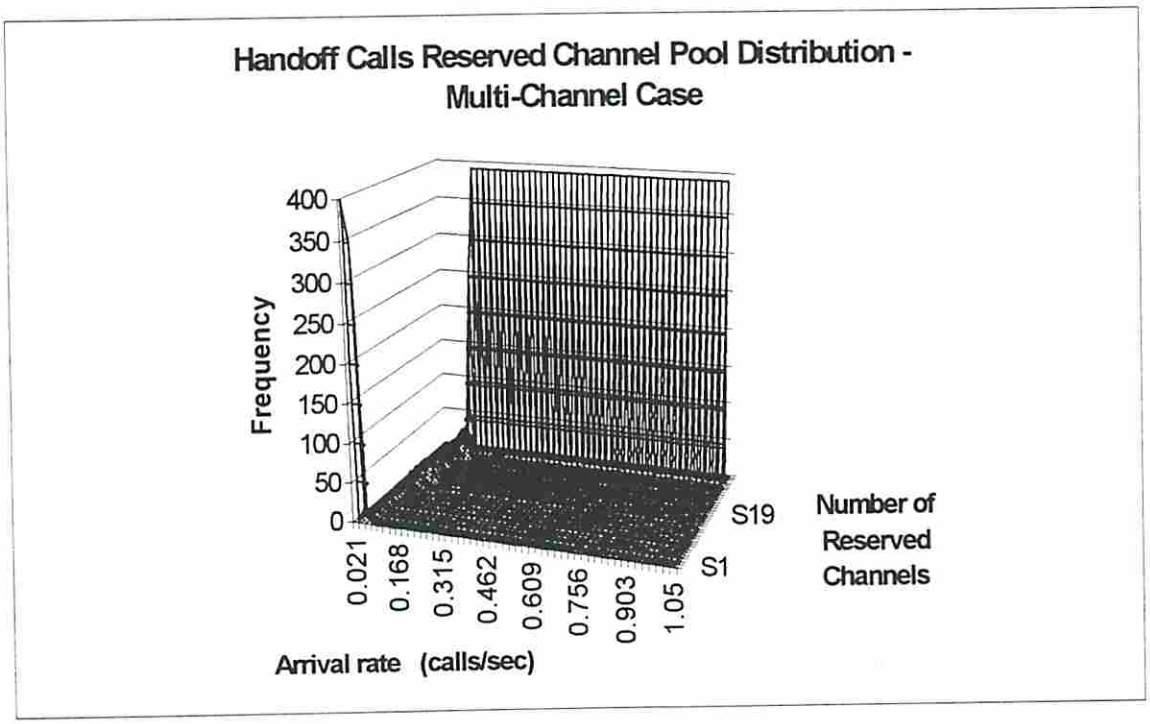

Figure 21- Distribution of CH in multi-channel case

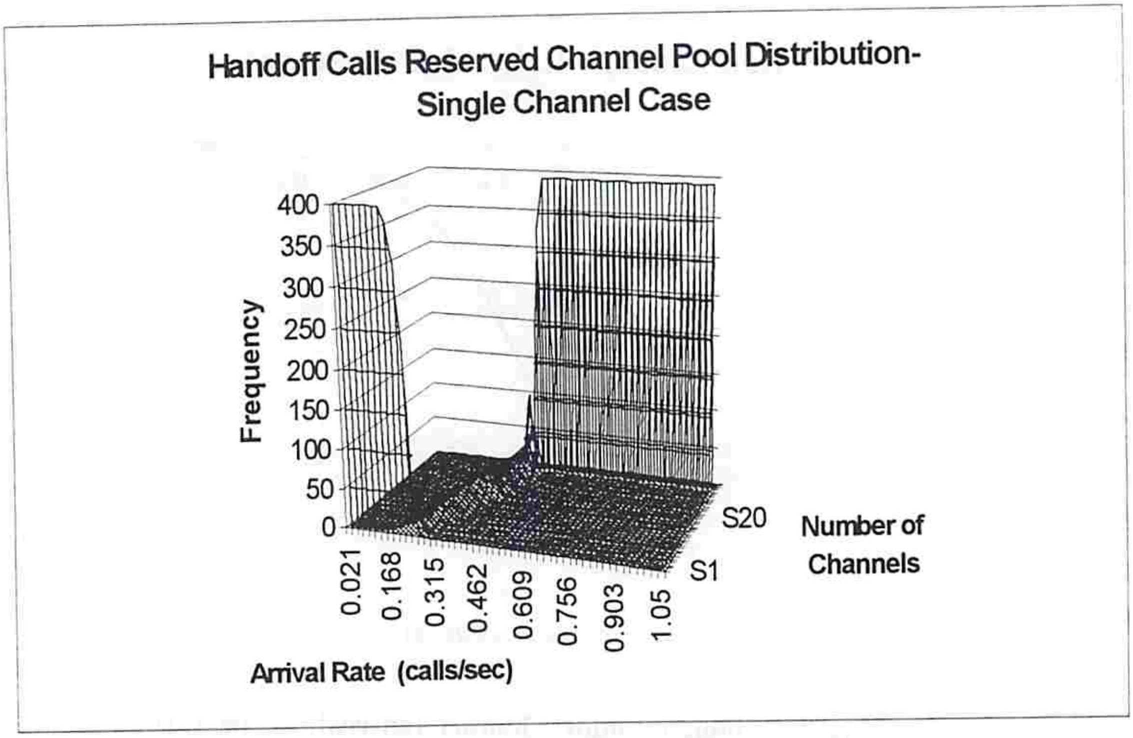

Figure 22 - Distribution of CH in single-channel case

Figure (22) shows the dropping rate of the realtime handoff call class. At lower arrival rates, the highest priority class, the realtime handoff class, experiences higher dropping rates when the non-realtime class is multi-channeled compared to the dropping rate when the non-realtime class is single-channeled. This is due to the fact that multi-channel calls that are admitted into the cell utilize more channels than in the case of the single-channel calls, therefore fewer channels are available for the realtime handoff calls to use and hence the higher dropping rate. When the non-realtime handoff calls violate their quality of service constraint, the adaptive scheme increases the number of reserved channels for both the handoff classes, CH. Increasing CH provides the realtime class with more channels to utilize even though the class' quality of service constraint is satisfied. The distribution of the number of channels reserved exclusively for the realtime handoff class, CR, in figures (23 - 24) shows that the adaptive scheme is increasing the size of CR due to a dropping rate violation of the quality of service constraint when the mean arrival rate around 0.2 for the case multi-channel and around 0.4 for the single-channel case. At these two points figure (22) shows a dropping rate below the quality of service constraint as a result of the increase in the CR. The adaptive scheme continues to increase the size of CR as the mean arrival rate increases until the maximum value of CR has been reached, at which point, mean arrival rate 0.6, the dropping rates of the multi-channel and the single-channel cases are identical. At this point the multi-channel calls are dropped and are a non-factor in the performance of the realtime handoff calls class.

Between mean arrival rates 0.2 and 0.4 the dropping rate of the multi-channel case is lower than that of the single-channel case. This interesting result takes place due to the fact that in the multi-channel case the number of reserved channels needed is greater than that needed by the single-channel case. This means that the adaptive scheme would go through several quality of service constraint violations to reserve channels enough to satisfy the quality of service constraint. Since the adaptive technique used is a multiple increase additive decrease, MIAD, the multiplicity characteristic of this technique results in reserving more channels than is needed to satisfy the quality of service constraint Hence, the number of channels reserved in the multi-channel case in the mean arrival rate interval 0.2 to 0.4, is greater than that for the single-channel case. The distribution of CR shown in figures (23- 24) supports this result.

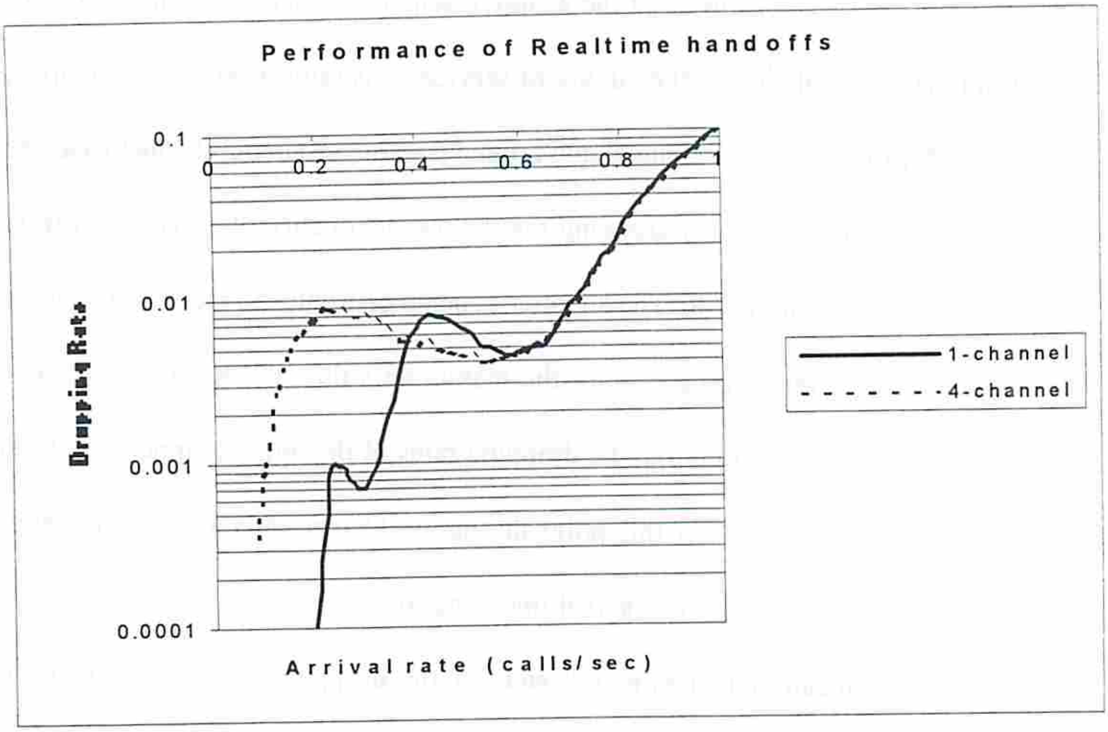

Figure 23– Single vs. Multi-channel realtime performance

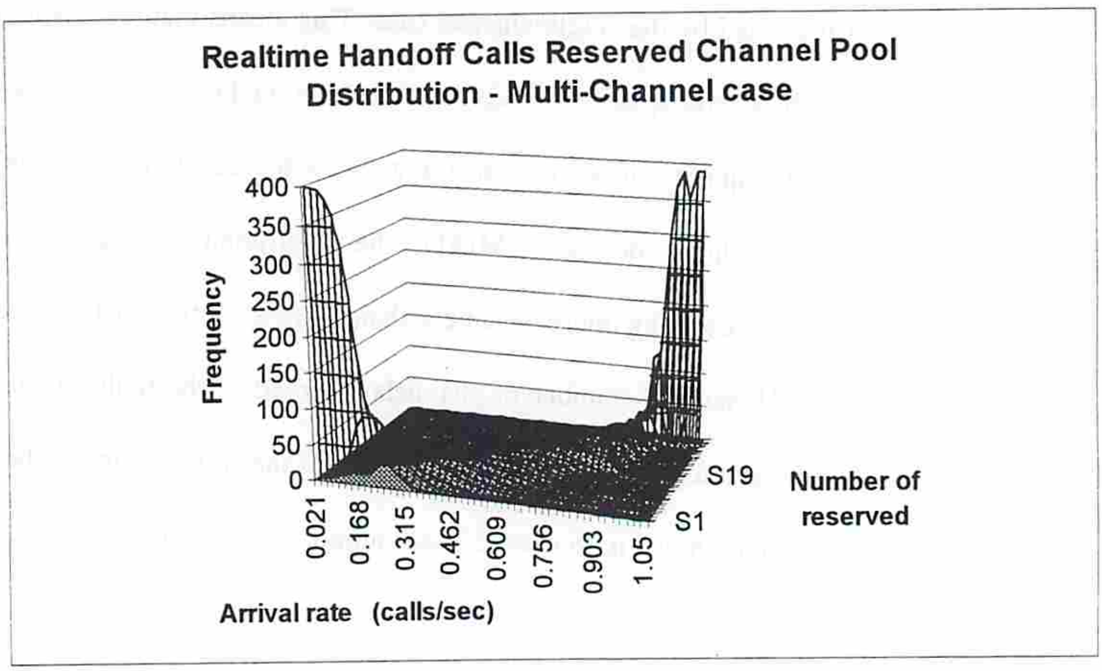

Figure 24– Distribution of CR for multi-channel case

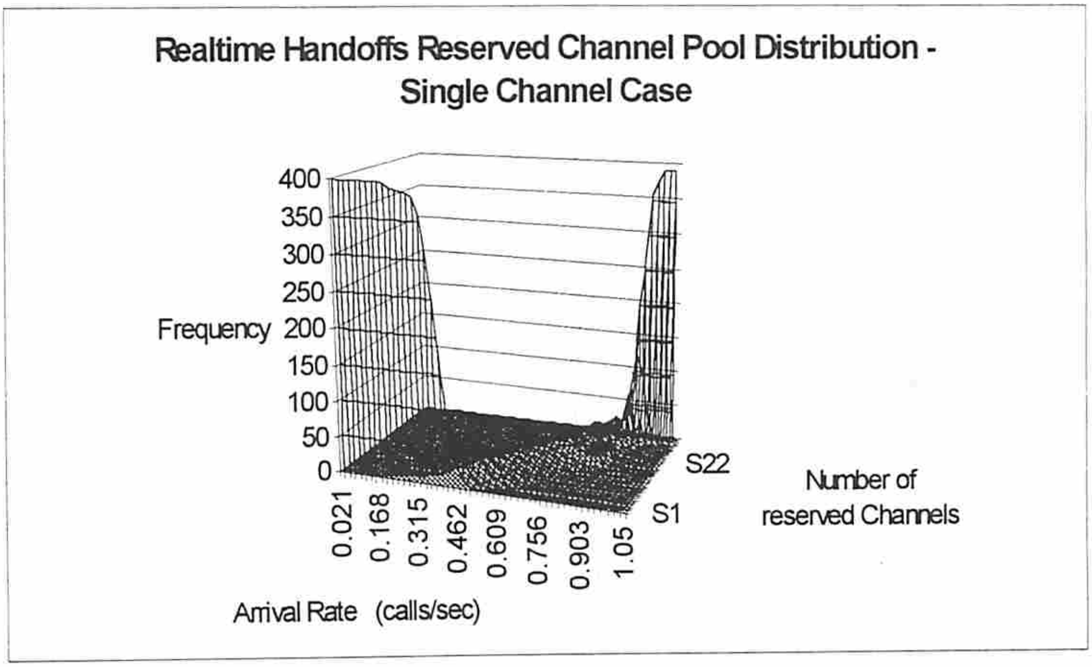

Figure 25 - Distribution of CR in single-channel case

The increasing demand for the limited available channels introduced by the multi-channel calls impacts the performance of the new calls class negatively. Figure (25) shows new calls dropping rates several orders of magnitude worse than the case of single-channel requests. This poor performance is amplified by the use of the adaptive scheme which in the case of the multi-channel calls reserves all of the channels of the cell for the handoff call classes at lower mean arrival rates than does the single-channel case.

From the above discussion and results, the impact of multi-channel non-realtime calls on the performance of the low priority class is severe. Moreover, the performance of the multi-channel class itself does not satisfy the quality of service constraint. This necessitates the employment of new techniques in order to satisfy the multi-channel non-realtime handoff calls' quality of service constraint and at the same time provide performance improvement to the other classes of calls. In the following section we

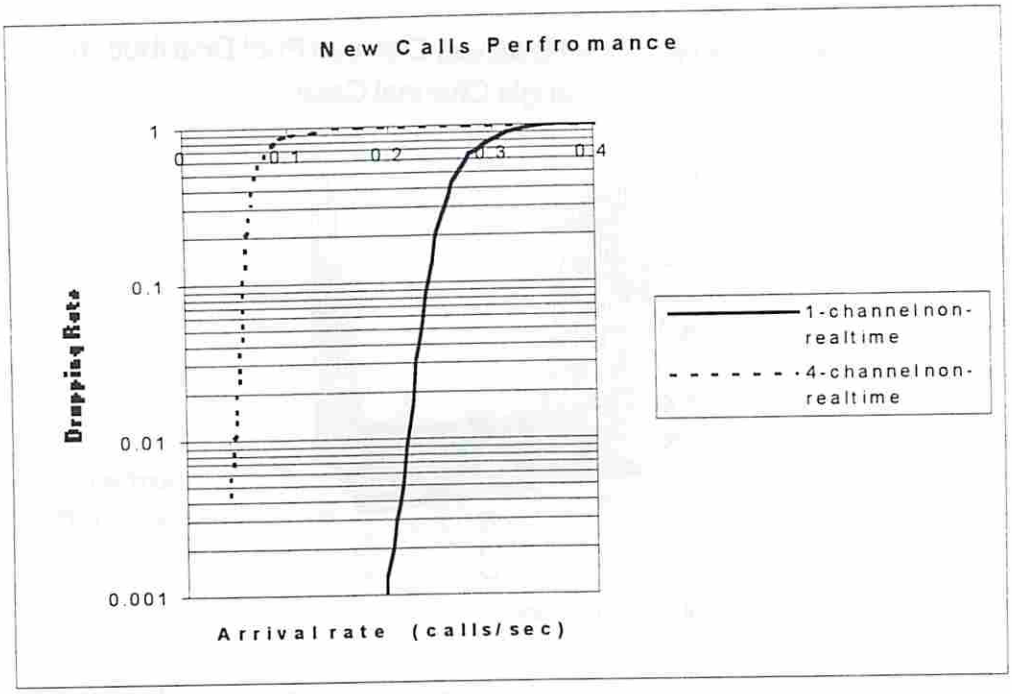

Figure 26 - Single vs. Multi-channel new call performance

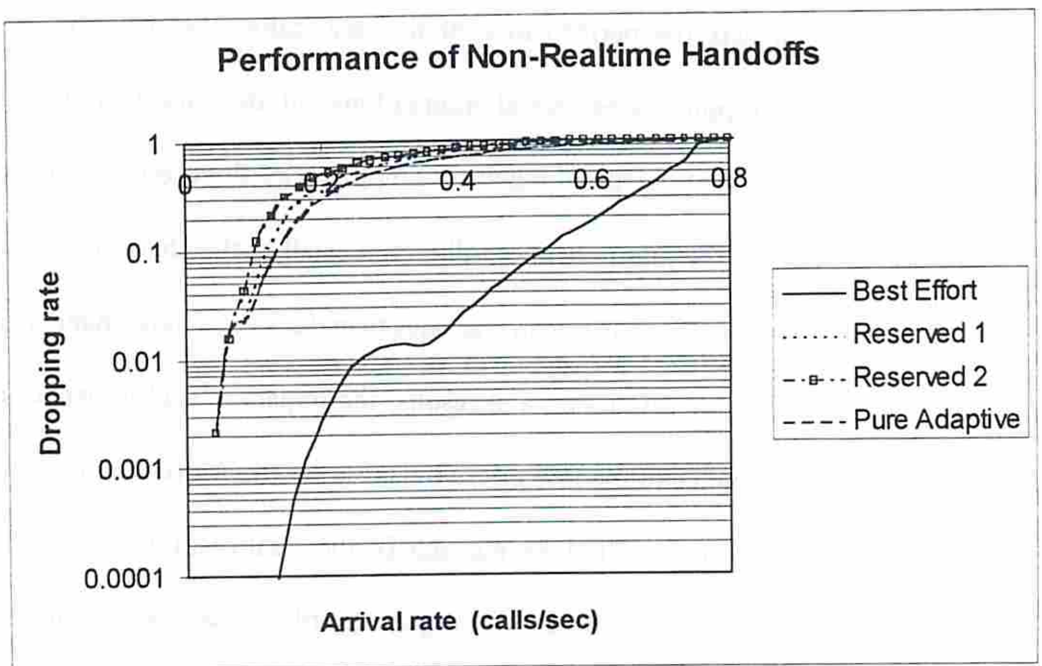

Figure 27- Performance of non-realtime under different multi-channel schemes

investigate two extensions to the adaptive scheme of chapter four, the reserved channels and the best effort techniques.

4.3.2.2 Adaptive Reserved vs. Adaptive Best Effort

The multi-channel non-realtime handoff calls enjoy a better performance when under adaptive best effort scheme compared to the other investigated schemes. At heavy traffic loads, an incoming multi-channel call requesting channels may be dropped if there are not enough channels available to satisfy it request if the scheme does not employ the best effort technique. While using the best effort technique, a multi-channel call is accepted as long as there is at least on channel available. This best effort technique reduces the number of multi-channel call drops and therefore provides the non-realtime handoff class with a lower dropping rate.

Figure (26) also shows an interesting result about the effectiveness of reservation technique for multi-channel calls. We compared the performance of the multi-channel calls under two adaptive reservation schemes, the first reserved 4 channels exclusively for the multi-channel calls, and the other reserved 8 channels. These two schemes resulted in an identical performance at lower arrival rates and that is because of the abundance of available channels which made the reservation of an extra 4 channels as in the second scheme unnecessary. At higher arrival rates, although the second scheme with 8 reserved channels provides the non-realtime handoff class with more channels, it also reduces the number of channels available for the high priority class which results in an

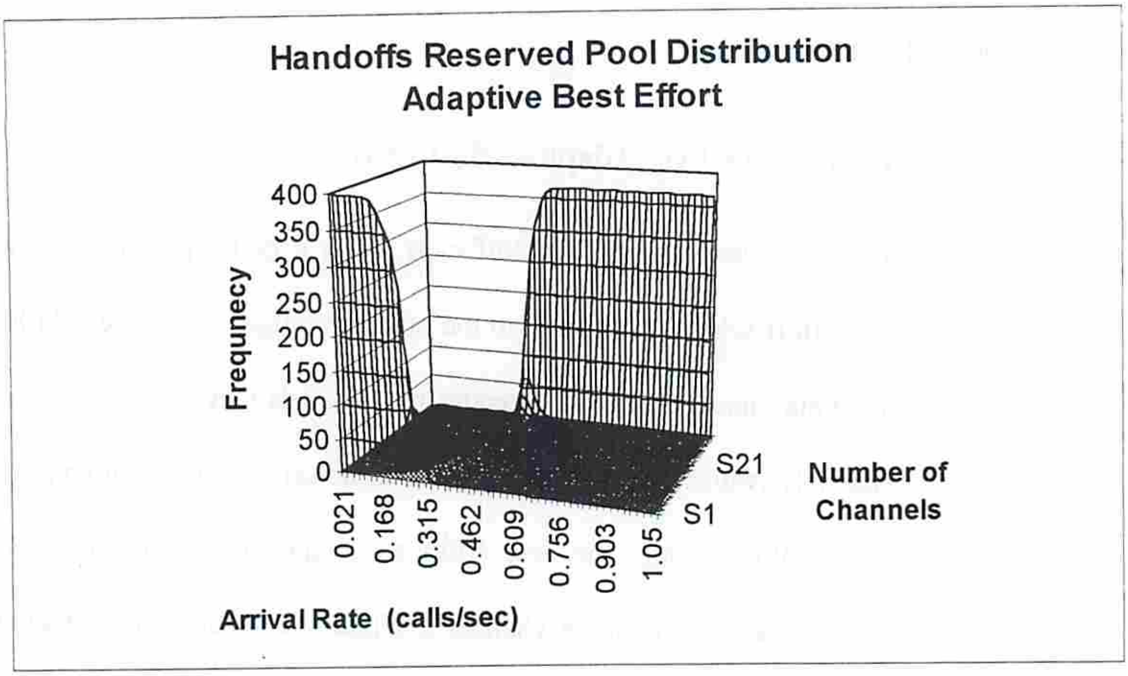

Figure 28- Distribution of CH under adaptive best effort

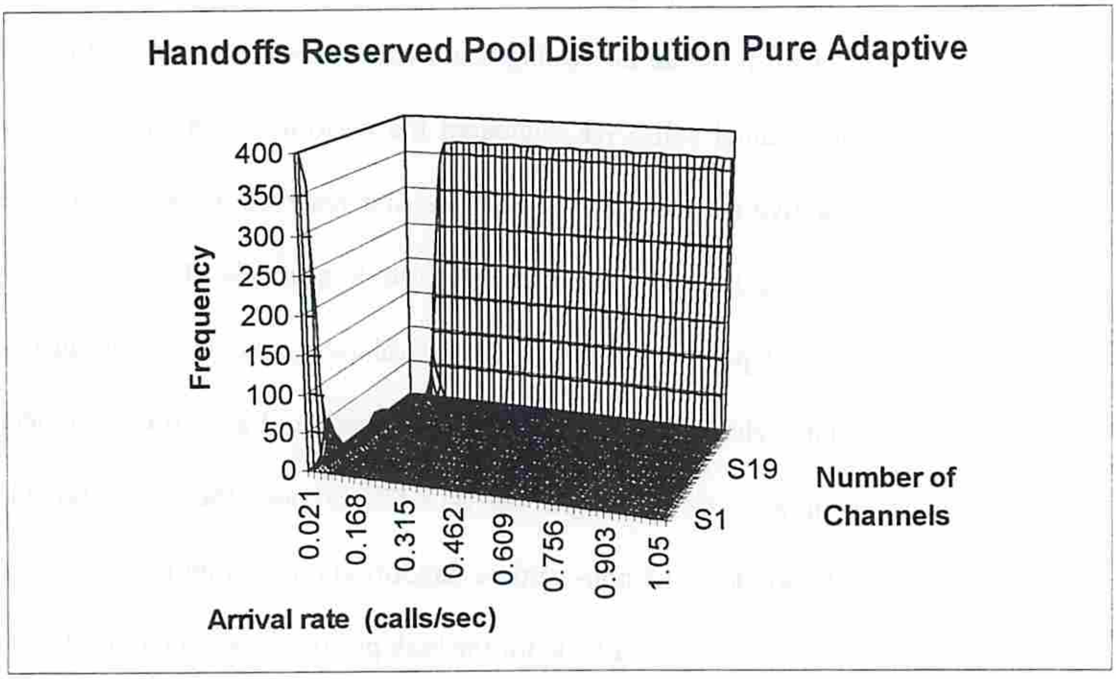

Figure 29- Distribution of CH under pure adaptive

increase in its dropping rate. A higher dropping rate which violates the quality of service constraint of the realtime handoff class triggers the adaptive scheme to increase the number of reserved channels for it. Subsequently, less channels are now available for the non-realtime handoff class and hence a higher dropping rate. The greater the number of reserved channels for the multi-channel class, the higher the dropping rate when the congestion is heavy.

The performance of the high priority class, the realtime handoffs, is also improved under the adaptive best effort scheme over the other considered schemes as shown in figure (30). Under the adaptive best effort scheme, the single-channel realtime handoff calls find more channels available for use than under the other cases when the cell is

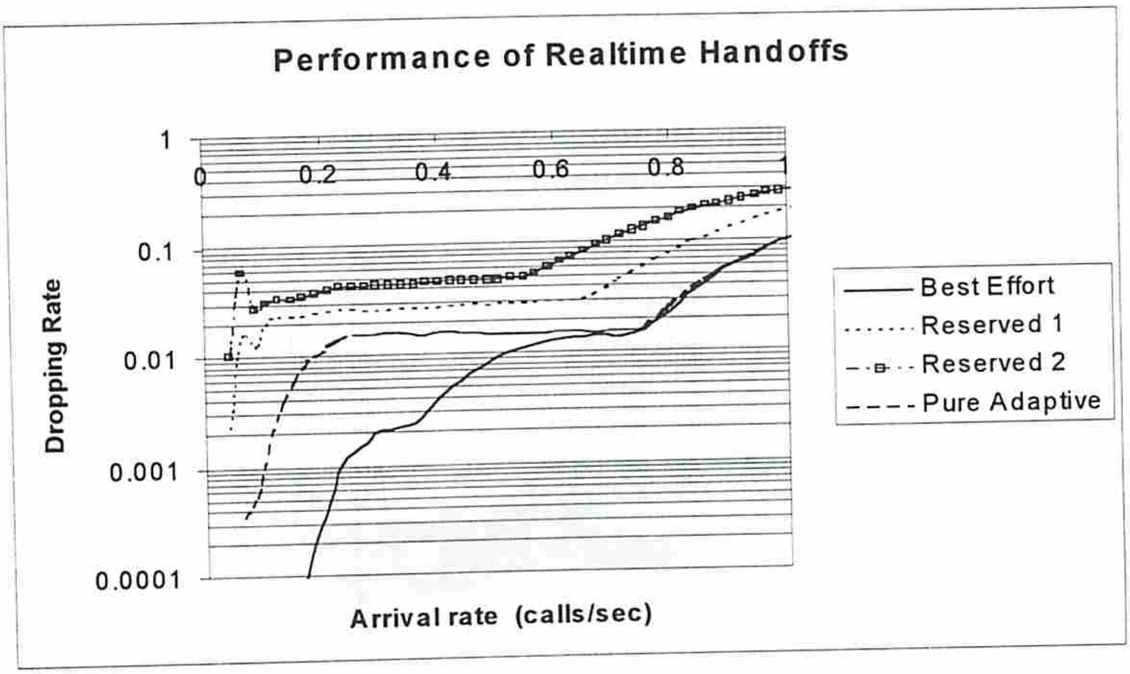

Figure 30- Performance of realtime under different multi-channel schemes

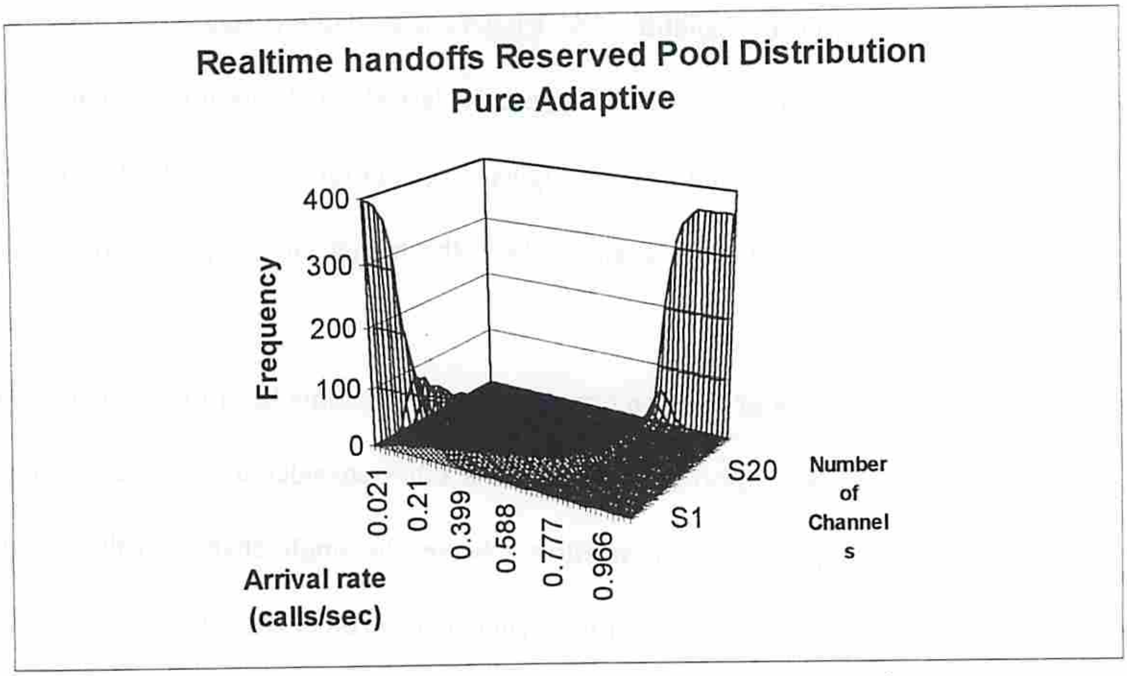

Figure 31- Distribution of CR under pure adaptive

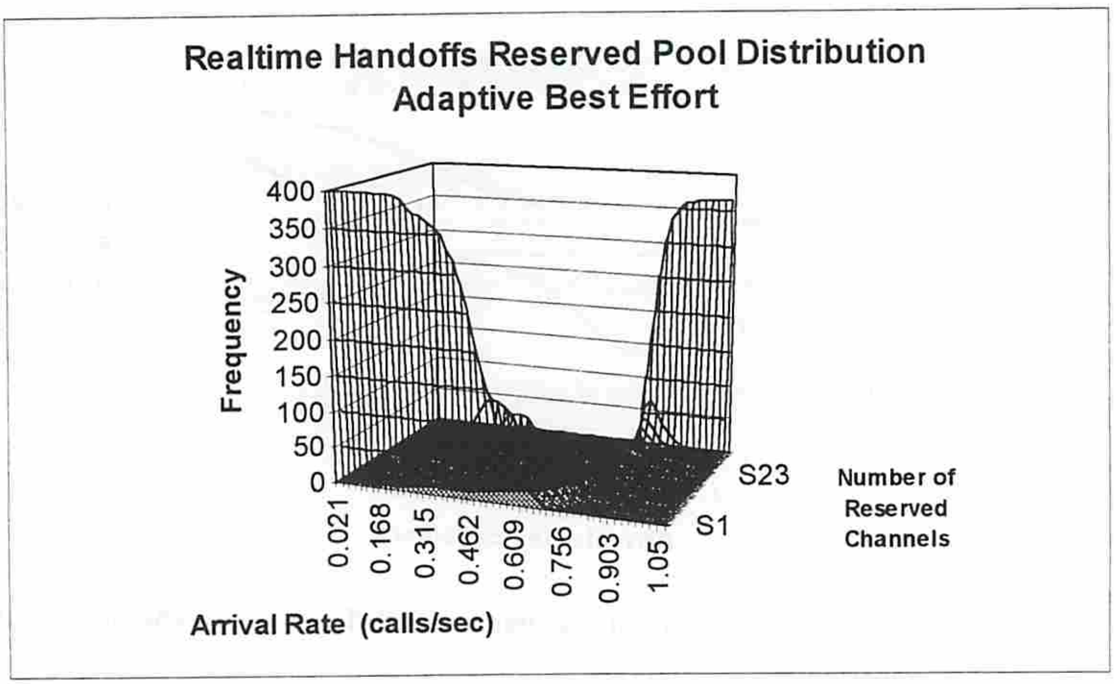

Figure 32- Distribution of CR under adaptive best effort

under heavy congestion. This is due to the admission of the multi-channel calls under the best effort technique which under congestion can admit multi-channel calls with less than requested number of channels, leaving more channels for the realtime handoff calls to use. The two adaptive reserved schemes performed as expected. The scheme with the greater number of reserved channels resulted in a higher dropping rate which is a reflection of the decrease in the number of available channels for the non-realtime handoff calls class.

An interesting point shown in figure (29) is the value of the mean arrival rate at which the adaptive scheme is triggered to add more channels to the handoffs exclusive reservation pool. The figure shows that the schemes have different arrival rates at which the adaptive scheme is triggered. These different arrival rates give an insight on the schemes' relative success in lowering the dropping rate of the non-realtime handoff class.

The performance of the new calls class shown in figure (32) indicate an improvement when the adaptive best effort scheme is used over the other schemes. This improvement is caused by two factors, first of which is the fact that new calls are of two types, single-channel and multi-channel calls. The use of the best effort technique reduces the number of multi-channel call drops, as discussed earlier. Therefore, the best effort technique reduces the number of multi-channel calls dropped from the new call class, which results in a lower dropping rate. The second factor is the use of best effort to admit non-realtime handoff calls, which increases the available channels for the other classes.

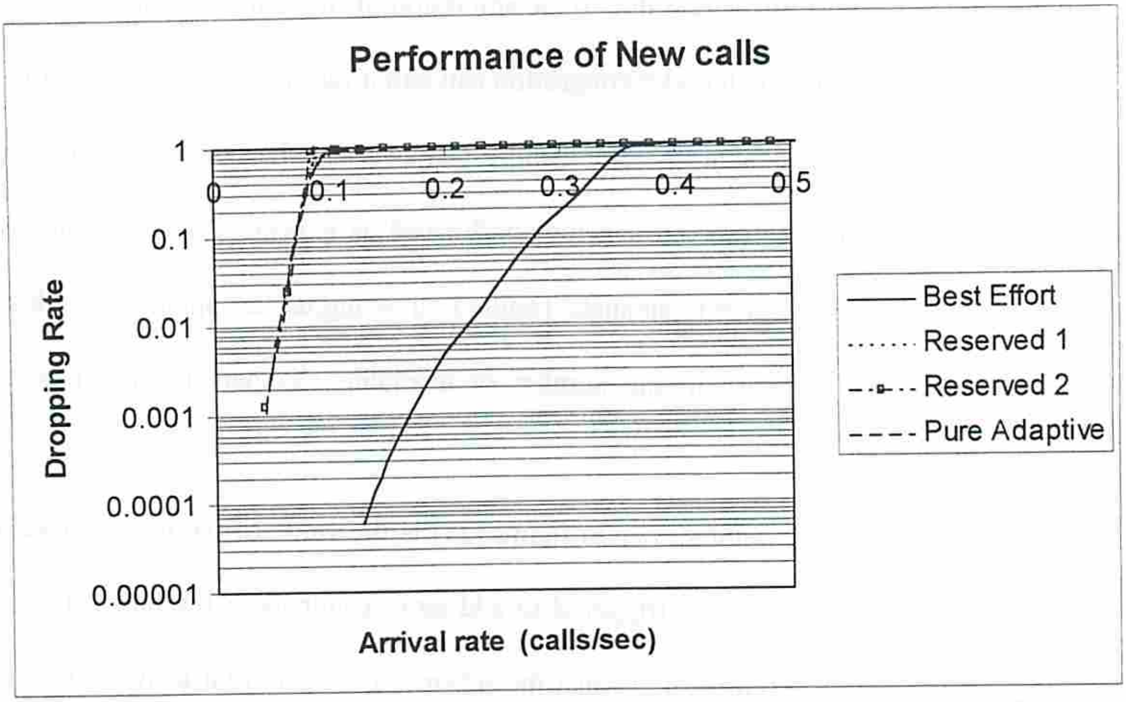

Figure 33 - Performance of new calls under different multi-channel schemes

4.4 Conclusion

In this chapter, we formulated a call admission control problem in which the incoming calls to the network are of two types, single-channel and multi-channel request calls. This problem corresponds to the case where the network supports voice communication utilizing one channel per call, and data communication utilizing a fixed number of channels per call. We assumed a homogeneous channel requests for the multi-channel calls.

Three call admission control schemes were proposed and compared. The three schemes employed the same adaptive scheme developed in chapter four with extensions in two of these schemes. One scheme had the reserved channel extension, in which a number of channels are reserved exclusively for the multi-channel calls. The other extension involved the use of best effort technique, in which a multi-channel call is accepted with a number of channels equal or less than the number requested is granted depending on the channel availability.

The results of the performance comparison among these three schemes showed the adaptive best effort scheme resulted in a superior performance for all the priority classes over the other two schemes.

Chapter 5

Multi-Rate Call Admission

5.1 Introduction

In the previous chapters, we discussed a call-based multi-class priority call admission scheme. In that paradigm, calls are classified according to their delay constraints whether realtime or non-realtime and according to their request position whether within the cell as new calls or from outside the cell as handoffs. We made some assumptions about the number of channels requested by each class. We first assumed all calls requested a single channel in chapter four, then we assumed that the non-realtime handoff calls requested fixed number of channels. But with the increasing popularity of multimedia communications and with the different communication speeds of wireless modems, it is reasonable to assume that calls from every priority class would request any number of channels, from a range of possible values, that would satisfy the call's quality of service requirement.

We can also make a paradigm shift and classify the priority classes based upon the particular users of the service. A user may belong to the highest priority class and hence all of his/her calls are treated as such regardless of its delay constraint or position at the time of admission request. It is possible in this classification that a new call of a higher priority class is admitted while at the same time a realtime handoff call of a lower priority class is dropped.

A service provider can implement this user-based priority classification through service plans. A service provider may offer several priority service plans, to which a customer may subscribe. Although a customer might pay a higher premium for a higher priority plan, the customer would enjoy a lower dropping rate for his/her calls. While the pricing of such services and plans is an interesting problem it is outside the scope of this dissertation.

In this chapter we will focus on formulating the multi-rate multi-class priority call admission problem and apply the framework established in the previous chapters to this problem.

5.2 Multi-Rate Call-Based Classification

In previous chapters we classified calls based upon their delay constraints or position when requesting admission. So calls were classified into realtime handoffs, non-realtime handoffs, and new calls. We investigated the cases when all call requests were single-channels, and when the non-realtime handoff calls were multi-channels, i.e. requesting a fixed number of channels. We can generalize this problem by taking into account multimedia realtime calls that can request multiple channels, and high speed wireless data modems that require multiple channels to transfer data at higher speeds. Therefore, the argument for call-based priority classification is still valid but the calls from the same priority class have heterogeneous number of channel requests. This heterogeneity in the requested channels by calls from the same class transforms the problem into a multi-rate multi-class priority admission control problem.

5.3 User-Based Priority Classification

In some network setups it is more suitable to classify priority classes by users than by call types. For example, user-based priority classification can be used in a network where users can choose their own priority class in the form of a subscription to a priority service plan. It can also be used in a setting like a corporate campus where employees are divided into priority levels based on their position in the company. For example, top executives in the company are in the highest priority class while the workers in the mailroom are in the lowest priority class. User-based priority classification can be used in any setting where the calls of a group of users are to be given priority over the calls of others.

Since each user in any one of these priority classes can make a call that requires a number of channels from a range of possible values, we can classify these calls as multi-rate calls. Therefore, the user-based priority admission problem is just a multi-rate multi-class priority admission control problem.

5.4 Multi-Rate Multi-Class Priority Call Admission

The problem that needs to be addressed here is how to satisfy the quality of service requirements of these different priority classes. When the cell is lightly loaded, satisfying the different classes with their corresponding quality of service is the simple task of granting every call the number of channels it requests. On the other hand, when the cell is congested, we cannot satisfy every class' quality of service requirement. In this situation the priority classes are the determining factor. Resources are allocated to calls such that higher priority classes' quality of service requirements are satisfied first then those of the lower priority classes.

To satisfy the different priority classes' quality of service requirements, we can use the adaptive multi-guard channel scheme developed in chapter four. Using this scheme, we create a pool of reserved channels for high priority classes to use when the cell is congested. We add channels to this pool of reserved channels when the dropping rate is greater than the quality of service requirement. When the cell is lightly loaded or the quality of service requirements of the higher priority classes are satisfied channels are removed from the reserved pool and made available for lower priority classes.

A useful technique to use in order to lower the dropping rates of the different priority classes is best effort service. The best effort service technique developed in chapter five allows a call requesting more than one channel, when there are not enough channels to satisfy this request, to be admitted and granted less than the requested number of channels. As discussed in chapter five, a call admitted using best effort would result in a longer channel holding time if it is a data communication call, or a lower resolution if it is a multi-layer coded video call.

In this chapter we will focus on two types of calls voice calls and data communication calls. We assume voice calls to require only one channel at a rate of 8 Kbps, while the data communication can request any number of channels from one to four for a minimum data rate of 8 kbps to a maximum of 32 kbps. The reason for excluding the video calls is that dropping rate, which is the metric we use to track the performance of the priority classes, provides misleading information about the performance of this class of calls. Consider N video calls arriving in an interval t requesting four channels each. Using the best effort technique when the cell is congested, these calls are granted one channel each, would still count as 0 dropping rate for that interval even though the quality of these calls are low due to the dropping of the higher resolution coded data. A new metric, which takes into account the effects of the best effort technique on the quality of the transmitted video is needed in order to evaluate the effectiveness of the best effort technique.

On the other hand, the dropping rate provides a good picture of the performance of the multi-rate data calls. When the cell is congested and best effort admits a data communication call with one channel instead of the requested four channels, the speed of the data transmission is reduced to one fourth of the requested speed and hence the length of transmission is four times that of the requested. This prolonging of the data communication call has two effects, one on the user and the other on incoming calls to the cell. A longer call would translate into a greater cost to the user, which can be rectified by charging per channel used. A longer call would mean a longer channel holding time and hence a longer time of unavailability for this channel. This would cause incoming calls to find less available channels and result in more drops. Therefore, the best effort technique effects the dropping rate of all the priority classes and hence can be a good performance metric. For these reasons, we will assume that calls incoming to the cell are either voice calls or data communication calls.

We will divide the cell into three priority classes as we did in the previous chapters, high, middle, and low priority. We will investigate the performance advantages of the adaptive scheme with the best effort extension as compared with the only the adaptive scheme. It is also of interest to investigate the performance of the three priority classes under three best effort cases. The first case is when the best effort technique is applied to all priority classes since all classes' calls are of multi-rate nature. The second case is to apply the best effort technique only to the two highest priority classes, the high and the middle. The last case is to apply he best effort technique only to the high priority class. We call these cases the 3 best effort, 2 best effort and 1 best effort cases respectively.

5.5 Results

5.5.1 Simulation Parameters and Assumptions

For this simulation we assumed a cell with capacity of 70 channels. There are three priority classes with equal mean arrival rates. The arrival process is assumed to be poisson, and the channel holding time is assumed to be exponentially distributed, with mean 100 seconds. The adaptive technique used is the multiplicative increase additive decrease, MIAD, with sliding window size of 3000 seconds and update interval of 1000 seconds. The channel holding time is increased in proportion to the number of channel requested to the number of channels granted. This will simulate the prolonged call time due to the best effort technique. An incoming call to any class is assumed to be equally likely to request 1, 2, 3, or 4 channels. The quality of service requirement for the high and middle priority classes are assumed to be 0.01 dropping rate, and the lower priority class has no quality of service requirement.

5.5.2 Analysis

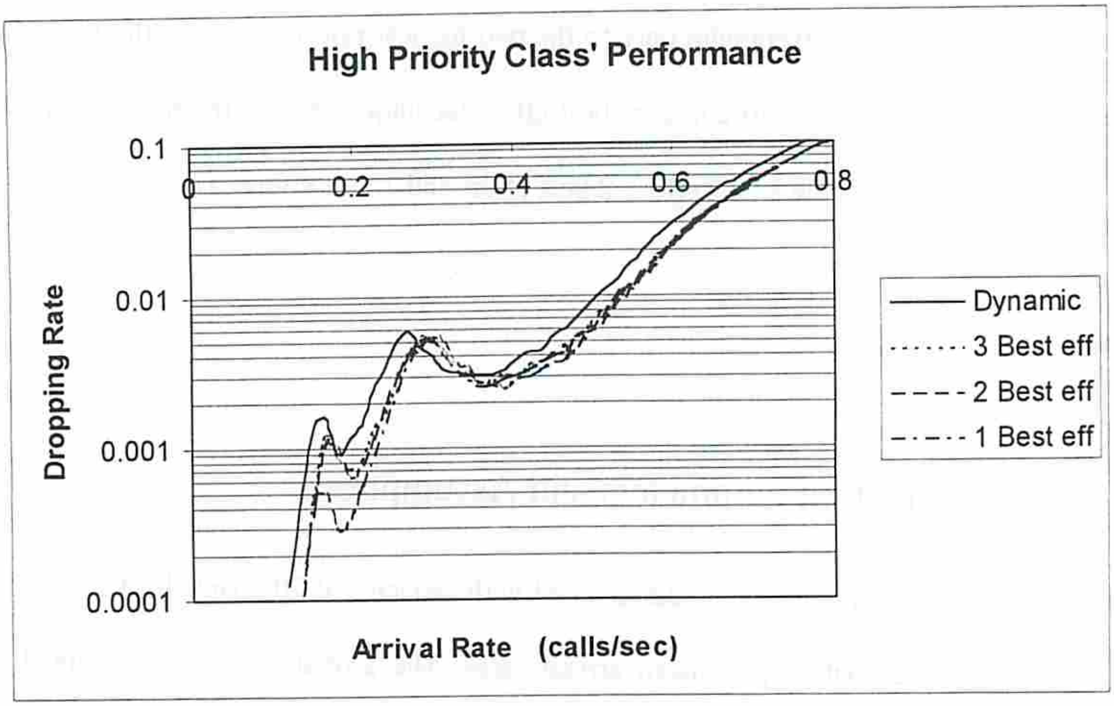

Figure 34 - High priority class performance under different multi-rate schemes

Figure (34 - 36) show the performance of the three different priority classes under the four adaptive cases. In figure (34) which shows the performance of the high priority class, it is evident that the best effort technique provides lower dropping rates compared with the pure adaptive case. It is shown in this figure that the 1 best effort scheme, where best effort is only applied to the high priority class, provides lower dropping rate than the other best effort variations. This is due to the fact that more high-class calls are admitted through the best effort technique while the other classes' calls are dropped when there are not enough available channels, leaving what ever is available to the high class' calls to utilize. The situation is different when best effort is also applied to middle priority or to both middle and lower priority classes as is the case in 2 best effort and 3 best effort respectively. In these cases, the added competition over the channels from the other priority classes' calls increases and result in the dropping of more calls from the high priority class. It is interesting to note that at mean arrival rate of 0.5, all the different variations of the best effort technique result in the same dropping rate. This is the result of the adaptive scheme reserving more channels for the high priority class when the quality of service requirement is not satisfied. At the point where all channels are reserved for the high priority class, all variations of the best effort are essentially identical with all other priority classes are extremely limited to their access to an available channel.

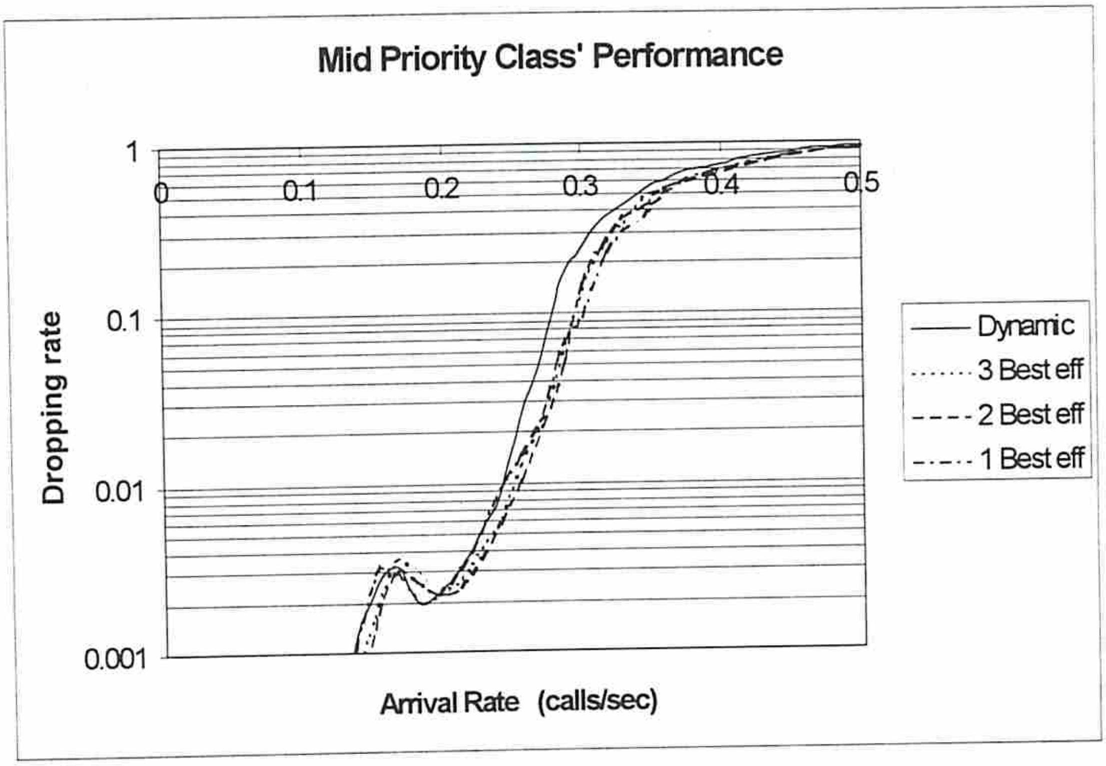

Figure 35 - Mid priority class performance under different multi-rate

Figure (35) shows an interesting result which is that the performance of the middle priority class is improved in the case of the best effort extensions to the adaptive scheme over the pure adaptive, although no major advantage is shown for one best effort technique over the other. There is no apparent improvement in the performance of the middle priority class whether best effort is applied to all classes, to the highest two classes, or only to the highest priority class. This is due to the fact that applying best effort to the middle priority class would result in more calls of this class being admitted, hence less available channels for the high priority class and higher dropping rate. The higher dropping rate of the high priority class triggers the adaptive scheme to reserve more channels for it, reducing the number of channels available for the middle priority class, hence a higher dropping rate. Therefore, the performance improvement gained from applying best effort to the middle priority class is offset by the performance deterioration resulting from its effect on the high priority class.

This previous result supports the choice of the dropping rate as the performance metric for best effort techniques. The dropping rate succeeded in this case to provide a global performance result that showed an indifference to any particular best effort technique, and gave an insight on the dynamics and the performance dependencies among the classes, where other metrics would have failed to show. The effects of the best effort technique go beyond the class on which it is applied, but also on other classes in the cell.

Figure (38) shows the performance of the low priority class under the four cases. The figure shows an improvement in the low class' performance under the 2 best effort and 3 best effort schemes over the pure adaptive and the 1 best effort schemes. The

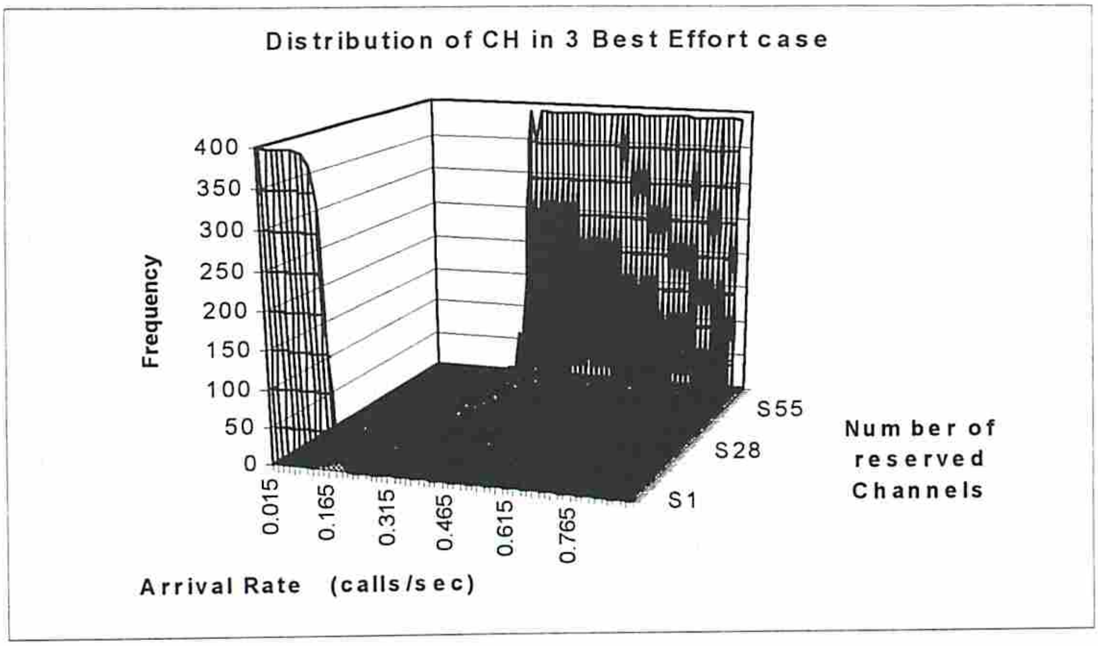

Figure 36 - Distribution of CH under 3 best effort case

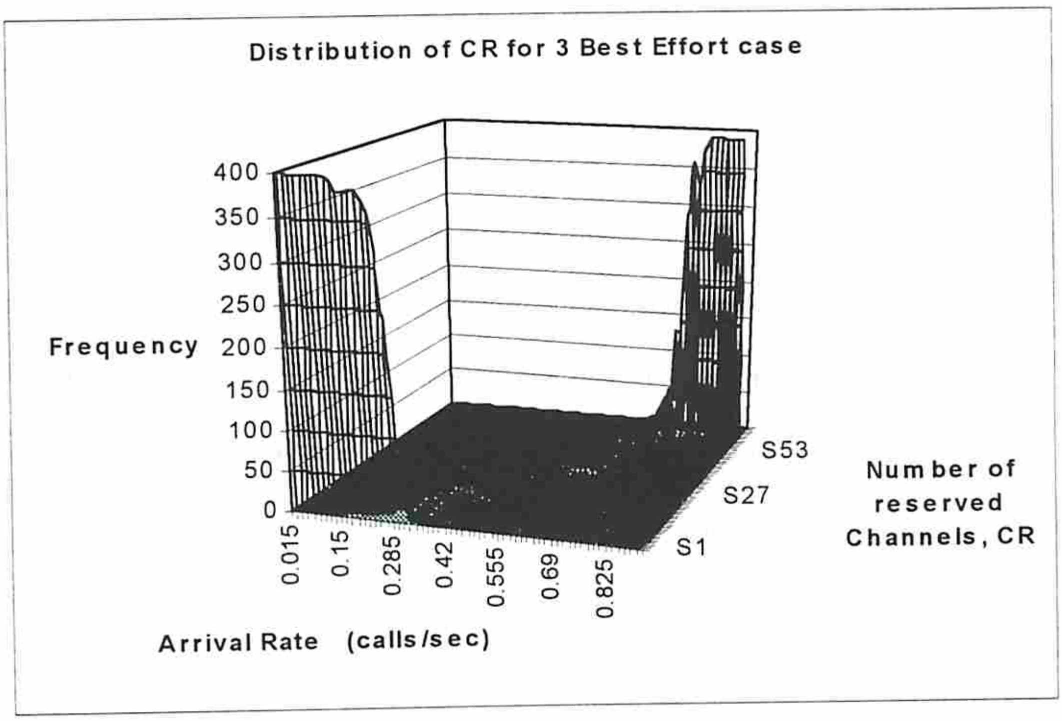

Figure 37 - Distribution of CR under 3 best effort scheme

improvement comes from the fact that the best effort technique is applied to the middle priority class which results in a decrease in the number of dropped middle calls, hence a decrease in dropping rate compared to the pure adaptive or 1 best effort schemes. A decrease in the dropping rate for the middle priority class results in the adaptive scheme being triggered, i.e. channels reserved for the two high and middle priority classes, at a higher mean arrival rate. Therefore, more channels are available for the low priority class in the 2 best effort and 3 best effort schemes than in the pure adaptive and 1 best effort schemes, which translates into a lower dropping rate for the low priority class.

At lower mean arrival rates, the 3 best effort scheme decreases the number of dropped lower priority calls as compared to the 2 best effort scheme. This improvement in dropping rate is due to the fact that best effort is applied to all classes in 3 best effort scheme whereas it is only applied to high and middle classes in 2 best effort scheme. As the arrival rate increases the middle priority class' dropping rate reaches a point where it does not satisfy the quality of service requirement and hence the adaptive scheme is triggered resulting in more channels reserved. These reserved channels reduce the number of available channels to the lower priority class. At high arrival rates, the adaptive scheme reserves all possible channels, resulting in a minimal effect of the best effort scheme on the lower priority class, and therefore an equivalent performance to that of the 2 best effort scheme.

The results obtained from the different schemes, which were discussed above, can be summarized into two main points. The first is that best effort technique when used with the adaptive scheme developed in chapter four can provide the multi-rate multi-class priority system of this problem with a better performance, albeit not very large

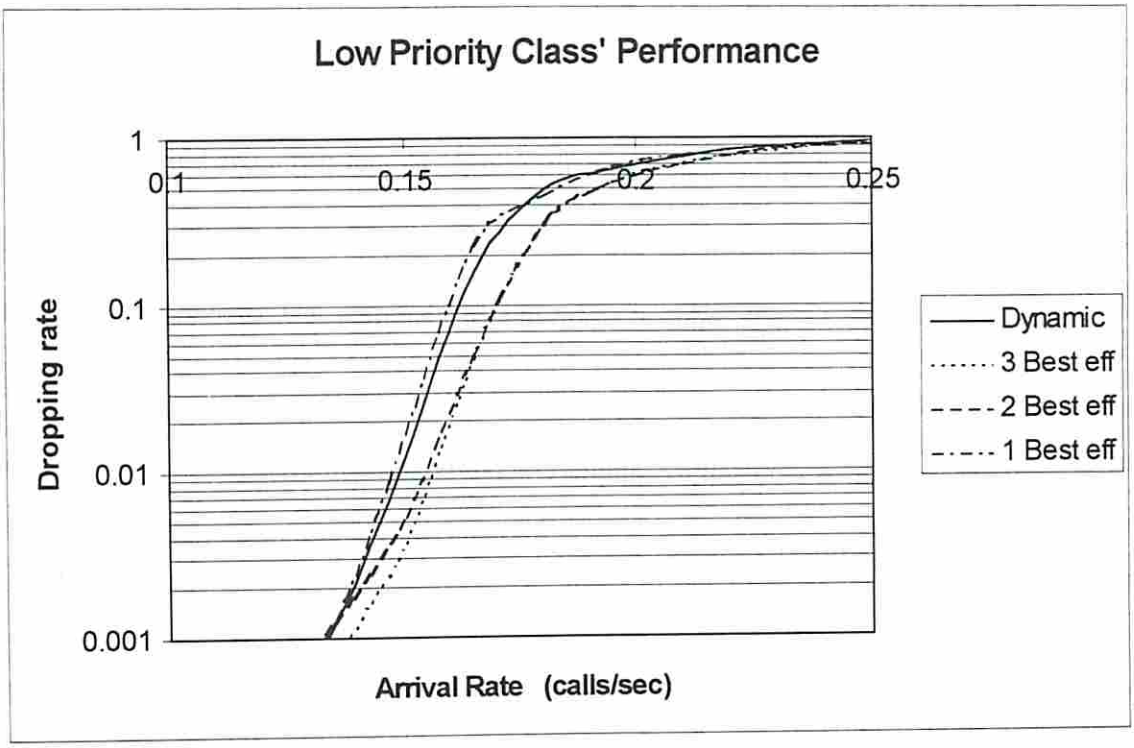

Figure 38 - Performance of low priority class under different multi-rate schemes

improvement. The second point is that the performance of the middle priority class is a non-factor in determining which variation in the best effort technique, out of the three proposed in this chapter, to use. The high priority and the low priority classes rather are the determining factors. The 1 best effort and the 3 best effort schemes provide better performance at low arrival rates for the high and the low priority classes respectively. Nevertheless, the 3 best effort scheme yields a higher performance gain for the low priority class than that the 1 best effort scheme yields for the high priority class.

5.6 Conclusion

In this chapter we motivated the problem of multi-rate multi-class priority call admission and applied the adaptive multi-guard channel framework developed in chapter four to it. We further extended our framework to include best effort service which we define as granting a call request a number of channels less or equal to the number of channels requested based on the number of channels available for the call's class. We investigated four schemes, pure adaptive, adaptive and all classes are best effort, adaptive and two highest priority classes are best effort, and adaptive and only high priority class best effort. A comparison among the dropping rates of all three classes showed that the adaptive and all classes are best effort provided the best performance for the low priority class, while the adaptive and only the high priority class best effort resulted in the best performance for the high priority class. The different schemes of best effort provided a performance improvement for the middle priority class over the pure adaptive scheme; however, no significant improvement resulted when compared to each other.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

In this dissertation we have addressed the problem of priority call admission control in wireless networks. We argued for the classification of incoming traffic into priority classes since calls have different quality of service requirements. We also argued for the classification of calls not only on the basis of the location of call request, i.e. handoff or new call, but also on the delay and loss constraints of the calls, thus we classified calls into realtime handoffs, non-realtime handoffs, and new calls.

To provide these different priority classes with their requested quality of service, we proposed in chapter three a generalized form of the guard channels concept that we called the multi-guard channel. This scheme provides a multi-level reserved channel pools for the high priority classes. Simulation results showed an enhanced performance for the high priority class when this scheme is employed when compared with a traditional handoff-new call classification. We also investigated the several parameters that effect the performance of the different classes. Our contribution in this chapter was the application of the multi-cutoff priority queueing system to the problem of call admission control in wireless networks. This queueing system is the foundation upon which the multi-guard channel scheme was developed.

In chapter four, the multi-guard channel scheme was modified to take advantage of the non-stationary nature of wireless call traffic. We proposed an adaptive version of the multi-guard channel scheme where the number of reserved channels in the reserved pools changes as the network congestion changes. This adaptive scheme is controlled by two parameters, the dropping rate of the call class in a last measurement period and the call class' target dropping rate, i.e. the class' quality of service requirement. When the measured dropping rate is greater than the quality of service requirement, channels are added to the reserved pool, and when it is less than the quality of service requirement channels are removed from the reserved pool. We showed that this adaptive scheme performs as well as the deterministic optimal solution, and since it is independent of the arrival process, we argued that it would perform better when the arrival process is non-stationary. Several techniques to add and remove channels from the reserved pools were proposed and analyzed. The adaptive extension to the multi-guard channel scheme and its related techniques would constitute our contribution to the call admission control problem in wireless networks.

We relaxed our assumption that all calls request one channel in chapter five. A case where the non-realtime calls require a fixed number of channels was investigated. This multi-channel call class severely impacted the performance of other call classes in the network. To reduce this impact and to this multi-channel class with an adequate performance, we compared two approaches both are extensions of the adaptive scheme developed in chapter four, reserved channels and best effort approach. The reserved channel approach is basically reserving a number of channels exclusively for the multi-channel call class, while the best effort approach is to accept a multi-channel call with any number of channels equal or less than the requested. The best effort extension to the adaptive scheme resulted in a substantial improvement when compared to pure adaptive or reserved channel schemes. Our contribution in this chapter is the best effort extension to the adaptive scheme that was effective in balancing the quality of service requirements of the multi-channel call class and other classes.

There are other classification criteria besides classifying calls based on their quality of service requirements. In chapter six, we considered classifying calls based on the user's subscribed priority level. In this situation, a user subscribes to one of several priority service plans and the calls are admitted according to the priority level subscribed to. This classification method results in priority classes with heterogeneous channel request calls that are known as multi-rate calls. We applied our best effort adaptive multi-guard channel framework to this problem. We compared three variations of this framework, all classes use best effort, only the two highest priority classes use best effort, and only the highest priority class use best effort. Results show that the best overall performance is achieved when all classes use best effort. The contribution of this chapter is the formulation of the problem using user-based classification with multi-rate calls and the application of our best effort adaptive multi-guard channel framework to solve it.

The adaptive scheme proposed and developed in this dissertation was applied to three network scenarios. The first scenario was a network with a fixed single channel request from all call classes as discussed in chapter three. The second was a network with one class of calls requiring a fixed number of channels while the other class requires one channel only. The third scenario was a network where all classes have variable channel request calls, i.e. multi-rate calls. We applied several variations of the adaptive scheme to these network scenarios and found that not all these schemes result in an acceptable performance for all scenarios. In the case of similar channel request calls in all classes, as in the case of single channel call request in chapter three or multi-rate calls in chapter five, the pure adaptive scheme is an effective scheme in providing the quality of service required by the classes. The best effort technique is not applicable in the single channel request case, and it yields little or no improvement in the multi-rate case. The ineffectiveness of best effort in these scenarios is due to the homogeneity of the call channel requests among all the priority classes. The best effort technique, however, is very effective when there is a narrowband-broadband class distinction in the network, which we referred to as single-channel vs. multi-channel classification. In this case, best effort can provide the broadband class with an improved performance due to admitting more calls with fewer channels than requested.

6.2 Future Directions

Although several aspects of the call admission control problem for wireless networks have been addressed in this dissertation, there are several interesting issues to investigate as an extension to this work or future directions for this research. There are also some interesting and related issues not addressed in this work. In what follows we will attempt to enumerate some of them with a brief discussion.

Analytical Model

Analytical models are extremely helpful in understanding the dynamics of the different variables controlling the behavior of a system. It is also a reliable tool to confirm the results obtained by simulations. For these purposes, modeling the admission control scheme proposed here is important in understanding its behavior. As a future work, a two-dimensional Markov chain model of the multi-guard channel scheme with queues could be constructed with numerical results obtained and checked against the results obtained from our computer simulations.

Multi-Cell Assumption

We can expand the underlying assumption of one cell to multi-cell setting where the dynamics of a network can be viewed and the performance of the traffic classes over the entire network can be observed under the different proposed schemes. When the network is in a heavy state of congestion, some flow control scheme can be employed to feedback congestion information to down-stream cells so as to reduce traffic. The effectiveness of such flow control schemes and their impact on the performance of the different classes in the network are of interest.

Preemption Priority

A variation on the framework developed here would be to allow an incoming high priority call to either preempt an ongoing lower priority call or use one of the channels this lower priority call is using, thus reducing the number of channels a multi-channel lower priority call has. When an incoming high priority call finds no channels available, it uses a channel from a lower priority multi-channel call, hence reducing the number of channels allocated to the lower priority call while admitting the high priority call. We call this scheme channel stealing. Channel stealing can be employed among the priority classes or within a priority class. The effectiveness of this scheme and the tradeoff between providing higher priority classes with such a service and the effects it has on the performance of the lower priority classes needs to be addressed and analyzed.

Multi-rate Optimal Solution

It is of interest to obtain the optimal solution for the problem formulated in chapter six so as to satisfy the constraints of the highest two priority classes and minimize the dropping rate of the lowest priority class. With this optimal solution found a comparison between the optimal results and the results obtained from the proposed schemes in chapter six can be insightful on the degree of effectiveness of these schemes.

Pricing-Based Classification of Calls

It is of interest to the service providers to classify calls in a wireless network into pricing-based priority classes rather than time constraints or bandwidth requirement criteria. A service provider might offer 3 priority classes each with a different price tag. The higher the priority the higher the price tag. The service provider might give priority within a class according to time constraint or bandwidth requirement. Although this problem uses the same generalized guard channel scheme proposed in this dissertation, different and interesting issues arise. Two metrics are of interest to measure in this scenario, the quality of service as seen by the user, which is the dropping or blocking probability, and the quality of service as seen by the service provider, mainly the revenues generated from this scheme. An interesting and insightful comparison is the tradeoff between these two metrics as a function of guard channels sizes. Another interesting comparison is by employing the dynamic guard channels method to this scheme. The adaptability of the guard channels to congestion effects the quality of service as seen by the user and the service provider. Therefore, it is of interest to find an optimal measurement window to adapt guard channels to congestion subject to the constraint of yielding the highest revenues or the best overall quality of service to all classes.

Bibliography

[1] D. Hong and S. S. Rappaport," Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Non-prioritized Handoff Procedures", IEEE Transactions on Vehicular Technology, August 1986.

[2] R. Guerin," Queueing-Blocking System with Two Arrival Streams and Guard Channels", IEEE Transactions on Communications, February 1988.

[3] S. Tekinay and B. Jabbari," A Measurement-Based Prioritization Scheme for Handovers in Mobile Cellular Networks", IEEE JSAC, October 1992.

[4] J. Keilson and O. C. Ibe," Cutoff Priority Scheduling in Mobile Cellular Communication Systems", IEEE Transaction on Communications, February/March/April 1995.

[5] D. McMillan," Delay Analysis of a Cellular Mobile Priority Queueing System", IEEE/ACM Transactions on Networking, June 1995.

[6] J. Chen and M. Schwartz," Two-tier Resource Allocation for a Multimedia Micro-Cellular Mobile System: Performance Summary", IEEE PIMRC1995.

[7] B. Epstein and M. Schwartz," Reservation Strategies for Multi-Media Traffic in a Wireless Environment ", IEEE 1995 Conference on Vehicular Technologies.

[8] D. Hong and S. Rappaport," Priority Oriented Channel Access for Cellular Systems Serving Vehicular and Portable Radio Telephones ", IEE Proceedings, October 1989.

[9] M. Naghsinsh and A. Acampora," QoS provisioning in micro-cellular networks supporting multiple classes of traffic ", Wireless Networks, August 1996.

[10] C. Oliveira, J. Bae Kim and T. Suda," Quality-of-Service Guarantee in high-Speed Multimedia Wireless Networks", International Conference on Communications 1996.

[11] O.T.W. Yu and V.C.M. Leung," Self-Tuning Prioritized Call Handling Mechanism with Dynamic Guard Channels for Mobile Cellular Systems ", IEEE Conference on Vehicular Technology 1996.

[12] M. Naghshineh and M. Schwartz," Distributed Call Admission Control in Mobile/Wireless Networks", IEEE JSAC, May 1996.

[13] S. S. Rappaport and C. Purzynski," Prioritized Resource Assignment for Mobile Cellular Communication Systems with Mixed Services and platform Types", IEEE Transactions on Vehicular Technology, August 1996.

[14] J. Lee, Y. Mun and B. Kim," Channel Allocation and Handover Schemes For Personal Communications Systems", IEEE Conference on Vehicular Technology, May 1996.

[15] C. Schaack and R. C. Larson," An N-Server Cutoff Priority Queue", Operation Research, March-April 1986.

[16] M. Schwartz," Network Management and Control Issues in Multimedia Wireless Networks", IEEE Personal Communications, June 1995.

[17] S. Tekinay and B. Jabbari," Handover and Channel Assignment in Mobile Cellular Networks", IEEE Communications Magazine, November 1991.

[18] G. P. Pollini," Trends in Handover Design", IEEE Communications Magazine, March 1996.

[19] R. Cohen, B. Patel, and A. Segall," Handover in a Micro-cell Packet Switched Mobile Network", Wireless Networks, March 1996.

[20] Y. Lin, S. Mohan and A. Norpel," Queueing Priority Channel Assignment Strategies for PCS Hand-Off and Initial Access", IEEE Transactions on Vehicular Technology, August 1994.

[21] C. Chang, T. Su, and Y. Chiang," Analysis of a Cutoff Priority Cellular Radio System with Finite Queueing and Reneging/Dropping", IEEE/ACM Transactions on Networking, April 1994.

[22] R. Schehrer," On a Cut-off Priority Queueing System with Hysteresis and Unlimited Waiting Room", Computer Networks and ISDN Systems, 1990.

[23] F. Pavlidou," Two-Dimensional Traffic Models for Cellular Mobile Systems", IEEE Transactions on Communications, February/March/April 1994.

[24] Y. Lin and W. Chen," Call Request Buffering in a PCS Network", Infocom 1994.

[25] H. Morikawa, M. Inoue, M. Hatori, and M. Mizumachi," Prioritized Wireless Access Protocols for Real-time VBR Traffic", PIMRC 1996.

[26] M. Saquib and R. Yates," Optimal Call Admission to a Mobile Cellular Network", IEEE Vehicular Technology Conference, 1995.

[27] M. Woo, N. Prabhu, and A. Ghafoor," Dynamic Resource Allocation for Mutimedia Services in Mobile Communication Environments", IEEE JSAC, June 1995.

[28] L. Cheng and H. Hughes," Call Admission Control Issues in a Wireless ATM Network", ICCCN 1995.

[29] T. Tamaki, H. Aida, and T. Saito," A Proposal of Reserved Channel Dynamic Channel Assignment Algorithm for Multimedia Mobile Communication Systems", PIRMC 1996.

[30] Y. Lin and I. Chlamtac," Effects of Erlang Call Holding Times on PCS Call Completion", Infocom 1997.

[31] I. Katzela and M. Naghshineh," Channel Assignment Schemes for Cellular Mobile Telecommunication Systems: A Comprehensive Survey", IEEE Personal Communications, June 1996.

[32] G. Immovilli and M. Merani," Priority Traffic in Wireless PCS Systems: Transmission and Traffic Parameters Trade-off", IEEE Vehicular Technologies Conference '96.

[33] M. Ivanovich, M. Zukerman, P. Fitzpatrick and M. Gitlits," Performance Analysis of Circuit Allocation Schemes for Half and Full Rate Connections in GSM", IEEE Vehicular Technologies Conference '96.

[34] C. Jedrzycki and V. Leung," Probability Distribution of Channel Holding Time in Cellular Telephony Systems", IEEE Vehicular Technologies Conference '96.

[35] R. Nelson and D. Towsley," Approximating the Mean Time in System in a Multiple-Server Queue that Uses Threshold Scheduling", Operation Research, Vol. 35, No. 3, May-June 1987.

[36] M. Sidi and D. Starobinski," New Call Blocking versus Handoff Blocking in Cellular networks", Wireless Network, Vol. 3, No. 1, March 1997.

[37] R. Ramjee, D. Towsley and R. Nagarajan," On Optimal Call Admission Control in Cellular Networks", Wireless Network, Vol. 3, No. 1, March 1997.

[38] P. Pax, T. Hirose and N. Kumahara," Introduction of Multi-Rate Services into Digital Mobile Communication Systems", PIMRC '95.

[39] S. Maric and A. Kolarov," Threshold Based Admission Policies for Multi-rate Services in the DECT System", PIMRC '95.

[40] M. Ruggieri, D. Giancristofaro, F. Graziosi and F. Santucci," An Optimizable Guard-Channel-Based Handover Procedure for Mobile Microcellular Systems", PIMRC '95.

[41] P. Mermelstein and S. Kandala," Capacity Estimates for Mixed-Rate Traffic on the Integrated Wireless Access Network", PIMRC '95.

[42] A. Acampora and M. Naghshineh," An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks", IEEE JSAC, October 1994.

[43] R. Cocchi, S. Shenker, D. Estrin and L. Zhang," Pricing in Computer Networks: Motivation, Formulation, and Example ", IEEE /ACM Transactions on Networking, December 1993.

[44] J. K. Mackie-Mason and H. R. Varian," Pricing Congestible Network Resources", IEEE Journal on Selected Areas in Communications, September 1995.

[45] A. Orda and N. Shimkin," Incentive Pricing in Multi-Class Communication Networks", Infocom 1997.

[46] W. LEE," Mobile Communications Design Fundamentals", John Wiley and sons, 1993.

[47] L. Kleinrock," Queueing Systems Volume 1", John Wiley and sons, 1976.

[48] L. Klienrock," Queueing Systems Volume 2", John Wiley and sons, 1976.

[49] C. Toh," Wireless ATM and Ad-Hoc Networks", Kluwer Academic Publishers, 1997.

[50] J. Walrand and P. Varaiya," High Performance Communication Networks", Morgan Kaufmann Publishers, 1996.

[51] R. Nelson," Probability, Stochastic Processes, and Queueing Theory", Springer-Verlag Publishers, 1995.

[52] R. Jain,"The Art of Computer Systems Performance Analysis", John Wiley and sons, 1991.

[53] M. Taylor, W. Wang and M. Banan," Internetwork Mobility", Prentice Hall, 1997.

[54] F. Ricci," Personal Communications Systems Applications", Prentice Hall, 1997.

[55] S. Keshav," An Engineering Approach to Computer Networking", Addidon-Wesley, 1997.

[56] U. Black," Mobile and Wireless Networks", Prentice Hall, 1996.

[57] R. Kuruppillai, M. Dontamsetti and F. Cosentino," Wireless PCS", McGraw Hill, 1997.

[58] T. Rappaport," Wireless Communications", Prentice Hall, 1996.

[59] D. Pierre," Optimization Theory and Application", Dover, 1986.