Resource Allocation Algorithms for Multi-Class Wireless Networks

Bracha M. Epstein

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

Columbia University

1999

ABSTRACT

Resource Allocation Algorithms for Multi-Class Wireless Networks

Bracha M. Epstein

In this thesis, we examine the multi-class admission control problem in a mobile wireless environment. Users originate in a particular cell and may migrate over the period of the call to other cells in the network. Each traffic class in the network has its own quality of service (QoS) requirements which include both call and handoff dropping probabilities and a call blocking probability profile and its own properties including call length, mobility characteristics, and bandwidth requirements. We introduce two sets of QoS performance measures which are used to evaluate performance. They require that users of each traffic class are not dropped during service with appropriate probability and that different traffic classes are blocked based on a pre-determined profile.

We explore three different approaches, the first of which is static and the others of which are dynamic, to solve the problem.

The first algorithm is a static reservation-based approach. It is a multi-dimensional generalization of the trunk reservation algorithm. It reserves a fixed number of basic bandwidth units (BBUs) for each traffic type and does not adapt to changes in traffic composition or load. At low loads, it performs similarly to the complete sharing (CS) algorithm which maximizes throughput. It also improves over the complete partitioning (CP) algorithm. At high loads it outperforms both the CS and CP algorithms.

The second approach is based on a one-step prediction mechanism. Decisions are made autonomously in each cell based on the current occupancy levels of the different traffic classes in each of the home and neighboring cells. The algorithm family contains several different variants which we analyze. The algorithms provide guarantees on the maximum probability of being dropped on handoff independent of load or traffic composition.

The third approach is a completely distributed measurement based algorithm. Partitions in each cell are periodically updated based on measurements of the call and handoff statistics in the cell to conform to the pre-specified requirements. New and handoff partitions are adjusted independently leading to both improvement in performance and algorithm simplicity.

Simulation results and analysis and comparison to other algorithms where appropriate are used as performance benchmarks.

Contents

List of Figures

Acknowledgements

I am indebted to my thesis advisor Professor Mischa Schwartz for his understanding of both problems and people, encouragement, optimism, and meticulous attention to all aspects of my work. Aside from guiding me through this project, he has given me valuable insight into the art and science of problem solving. I would like to acknowledge the generous support of the National Science Foundation and the Army Research Office.
On a more personal note, I would like to thank my family for always encouraging me to do my best and strive for excellence. I am also indebted to my fiancee Moshe Simon without whom this dissertation would never have been finished. His encouragement, support, and technical guidance enabled me to ultimately achieve this milestone. *Tovim hashnayim min ha'echad.*

Yom Yerushalayim 5758

To my loving family and family to be

Chapter 1

Introduction

1.1 Background and Motivation

The past few decades have seen dramatic changes in the telecommunications arena. Introduction of wireless communications services, multi-media services, and technologies such as the computer and the internet have fundamentally changed the method, means, and content of communications. Expectations of ubiquitous telephone service as well as an ever burgeoning array of multi-media services are increasing at fantastic rates. Though the notion of using a single network for all forms of communication was initially seen as a way of achieving more efficient network utilization, current consumer culture is forming expectations of that soon-to-be eventuality. As time goes by, the nature and performance of these services is also metamorphosing. Consumers expect seamless end-to-end services of all kinds using all media. The quality of service (QoS) demands of the constituent traffic classes on the network vary both by medium and by class.

Though the consumer expects to be able to use a single device for all forms of communication, the QoS the user expects for any particular service or application is predicated on the service being provided and to a lesser degree on the medium being used to provide that service.

In our work, we focus on the nexus of provision of multi-class or broadband services in a mobile wireless network. This problem may alternately seen as an extension of the provision of multi-class calls on a wired network to a wireless network or the provision of a single traffic class using a wireless network to provision of multi-class calls on that network.

The demands of this problem are complex and multi-dimensional in nature. The QoS requirements are many and include, among other things, probability of call blocking for each class, probability of successful call completion, quality of the connection which may include factors predicated on the physical quality of the connection such as slow and fast fading (related to signal to noise ratio (SNR) and the carrier to interference ratio (C/I)) as well as higher level concerns such as allocated bandwidth (BW) (mean, maximum, and minimum), average delay and jitter (in packet based systems). Although the different requirements are not completely independent, it is necessary to choose one area for closer study.

There are many different kinds of wireless services which are being developed, many of which are already deployed. As time goes by, both the nature of the services as well as the applications using these services are becoming more complex and demanding more efficient uses of the system.

All wireless systems involve the transmission of information over a wireless channel. Applications which include the wireless local area network (W-LAN) assume that there are no other systems which are competing for use of the same bandwidth and allocate all available bandwidth to devices which communicate with each other. This first group assumes that there is sufficient bandwidth to go around which all devices may share. Additionally, the distance between the different transmitters and receivers are sufficiently close making direct communication between the devices practical. The second group of systems assumes (effectively) stationary users. Due to limited system bandwidth, users are broken up into groups and communicate only within that group. They never move from one group to another. These systems reuse the bandwidth and are based on the notion of a cell. However, mobility is not an issue. Examples of such systems/applications include wireless local loop technology and broadcast television. A third group of systems assumes mobile users which freely roam over the entire area of the system without regard to cell boundaries. As users traverse "cell boundaries", they are seamlessly "handed off" from one "cell" to another. Two very different kinds of systems which apply this notion are traditional cellular systems and ad-hoc wireless networks.

We consider the traditional cellular model. In this model, the base stations form the backbone of the network. They are connected to the rest of the wired network. Each mobile user is assumed to be under the aegis of a particular cell at any given time and is allocated bandwidth via the base station. As users move around from cell to cell, they are handed off from one cell to another.

Though the work we describe in this thesis is applicable to all different kinds of wireless networks and different kinds of services, we narrow the focus to a small subset of this area to generate useful results which may then be reapplied to the area as a whole. The network we consider in defining the algorithms is of arbitrary shape and topology. One typical example is given in Figure 1-1. Each cell is assumed

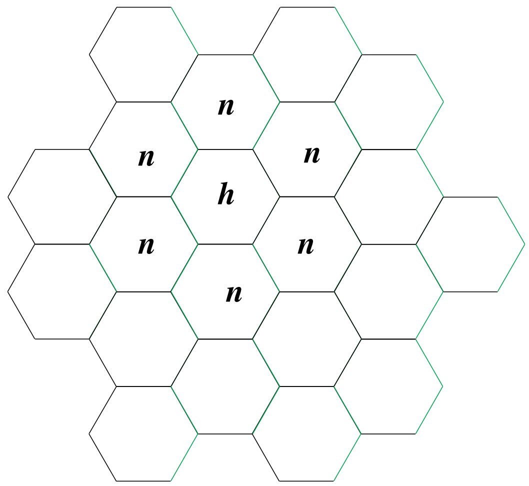


Figure 1-1: Typical hexagonal cellular topology

to have r neighbors. In practice, we simulate either a single cell or a ring of cells as shown in Figure 1-2. Each cell has a fixed bandwidth of N basic bandwidth units (BBUs) also referred to as "channels." We define the *home* cell for an arriving request to be the cell at which that call arrives and is denoted by cell h in Figures 1-1 and 1-2. *Neighbor* cells are the cells directly adjacent to the *home* cell and are denoted by n in these figures. The mobile users enter the network as new users and move from cell to cell throughout the course of a call. User movement is homogeneous. The average time until handoff is the same in all cells for a given traffic class. Movement to any given adjacent cell is equally likely and is equal to $1/r$. Each time a user traverses a cell boundary, it generates a handoff request at the cell it is about to enter. Arriving users not admitted are blocked and handoff calls dropped. We look at the two traffic class case. Each traffic class has an average call length, mobility parameter characterized by the average time until handoff,

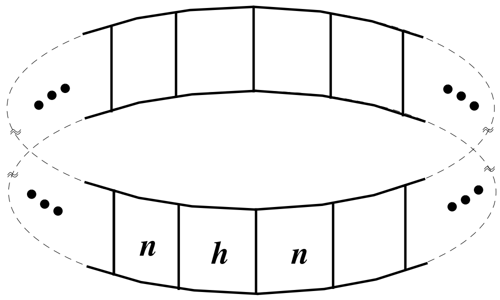


Figure 1-2: A ring of cells

bandwidth requirement, and QoS requirements. The QoS requirements of the user are call dropping probability and call blocking probability. The system sees these requirements as costs and would like to maximize throughput or carried traffic.

In the following chapters, we consider three different approaches in solving the mobile multi-class admission control problem. They are: static reservation-based control, predictive control, and measurement-based control.

In Chapter 2, we define a static reservation-based algorithm which is a multi-class reservation-based algorithm. This algorithm uses pre- and post-reservation. We develop a cost function which measures how close performance is to system defined priorities.

Next, in Chapter 3, we define a family of prediction algorithms for single and multi-class traffic. At the heart of these algorithms is the one-step prediction condition. We predict the capacity required in each of the *home* and *neighbor* cells such that the probability of being dropped is less than or equal to the maximum dropping probabilities. This calculation is done on a class by class basis with the sum of the per class capacity requirements being the per cell requirement. Given the number of class c users in each of the *home* and *neighbor* cells, we predict the capacity that would be required one time-step into the future such that given that capacity, the probability of a class c handoff user being dropped is less than the maximum allowed handoff dropping probability for that class. We call it the *cell i, class c capacity requirement.* We compute the required capacity for each class in each of the *home* and *neighbor* cells. This is just the sum of the per class required capacities in each cell. If each of the total required capacities in the *home* and *neighbor* cells is less than or equal to the respective total capacity in those cells, the one-step prediction condition has been met. The blocking probability mechanism is a measurement-based condition which assigns partitions on a per class basis to ensure that a pre-defined blocking probability profile is being met and no traffic class is being unfairly blocked from the system. We develop four prediction-based algorithms. The first two, OSPRED and MMOSPRED, are single and multi-class algorithms which admit new users assuming that the prediction condition has been met and that there is currently sufficient bandwidth available in the cell. Handoff users are admitted assuming that there is sufficient bandwidth available to admit the user. The second two algorithms, IMOSP-CS and IMOSP-RES, are both multi-class algorithms which implement both the prediction condition and the blocking probability mechanism. New users are admitted if both the above conditions are met and there is sufficient bandwidth available to admit the call. The algorithms differ in how they admit handoff users. IMOSP-CS admits handoff users if there is sufficient available bandwidth. IMOSP-RES users the *cell i, class c capacity requirement* defined above on new users arrival to set pre-reservation handoff partitions. These partitions assure each traffic class a minimum bandwidth available on handoff arrival in accordance with the predicted capacity. We compare these algorithms to each other and to another single class prediction algorithm.

We finally consider a measurement-based reservation algorithm described in Chapter 4. While this algorithm is also dynamic, it differs from the prediction algorithm family in two ways. First, while the prediction algorithms require state information from adjacent cells, the measurement-based algorithm MMDR is completely distributed and collects all data within the *home* cell. Second, MMDR adjusts reservation partitions based on past behavior while the prediction algorithms do so based on predicted future behavior. The QoS requirements are the same as for the prediction algorithms. The dropping probability requirements are threshold requirements while the blocking probability ones are relative profile-based ones. Though all of the traffic shares the same bandwidth, the different kinds of requirements require two separate approaches. We accomplish this using two-tiers of partitions. This both enables us to independently adjust each partition type and reduce problem dimensionality (and hence computational complexity). The first tier of partitions addresses the call dropping threshold requirements. It includes adjustment of handoff partitions for the handoff traffic and a base partition which is added to all new user partitions. It controls the amount of traffic in the system as a whole and is thus related to attainment of the threshold requirements. The relative call blocking partitions are adjusted independently and concern the relative amounts of traffic allowed in to the home cell of each class.

1.2 Previous Work

The multi-class admission control problem may be seen either as an extension of work done on multi-class wired networks, such as [13, 29, 32, 40, 41, 42, 10, 44, 52, 8, 36, 35, 60] to wireless networks or of proposed algorithms for single class networks in [22, 37, 23, 26, 27, 49, 50, 4, 82, 66, 68, 61, 67, 19, 24, 51, 25, 2, 80] to multiple classes of users.

The wireless admission control problem is multi-faceted. Due to factors such as mobility and frequency reuse, the boundaries of the admission control problem have been expanded and are concerned with other functions as well. Whereas call admission control in wired networks assumes a fixed start and end-point for the duration of the call, mobile wireless users move from cell to cell and change the network access point. Each time the user traverses a cell boundary, the handoff generates a bandwidth request to the system. Though this request is generated in mid-call, it is essentially a part of the admission control problem. As a result, the QoS requirements for handoff admission are stricter than for new user admission. Mobility analysis and channel holding time in a given cell profoundly impact performance and have been studied extensively. Some examples include [24, 46, 64, 20, 19, 51, 25, 86].

Spectrum is a precious resource and we assume that a fixed bandwidth is allocated to a system. Capacity may be increased by sub-dividing cells into smaller cells such as micro- and pico-cells [45, 74, 65, 70]. However, this results in a larger number of handoffs per call which in turn demands higher priority to maintain the required call dropping probabilities. Hierarchical systems such as [56, 30, 31] which overlay a layer of macrocells over smaller microcells are one solution which bridge the gap.

Unlike wired networks where the bandwidth available in any location is a function of the quantity of the media located between the two points, the quantity of bandwidth available in a particular cell may be a function of the bandwidth available to the system as a whole, other traffic carried in the network in adjacent cells, and resource allocation techniques. Dynamic channel allocation (DCA) techniques which adaptively allocate bandwidth to individual cells may increase the bandwidth available to a given cell at a particular time. Analysis of some DCA algorithms have been done in [18, 87, 34]. Other analysis [85] looks at the impact of traffic hotspots on performance. Algorithms such as [5, 59, 21, 79, 69, 48] look at these techniques and incorporate handoff admission into the DCA process. In [3, 57], work has been proposed to minimize the complexity and cost of admission of handoff users in the individual cell and integrates the backbone call admission process with the cell admission process.

Other work, such as [1, 7] looks at the multi-class admission control in wireless LANs where users are not mobile and thus no handoffs occur. In [9], admission control in ad-hoc networks is studied. This differs from other mobile networks since there are no base stations.

In [73], there is a discussion of issues in network management and control in wireless multi-media networks. The multi-class admission control problem is discussed more recently in [33, 83, 14, 78, 55, 39, 76] among others.

Packet reservation for users in service has been discussed by [53, 81] and reservation as a means for giving priority to handoff users in [38, 43]. A dynamic measurement-based algorithm for giving priority to handoff users was discussed in

[77] and predictive reservation techniques are discussed by [12, 63].

Yu and Leung discuss dynamic trunk or guard channel reservation for a single traffic class in wireless systems in [84]. In that algorithm, performance is enhanced by the system's ability to dynamically adapt to the changing load and mobility parameters. Naghshineh and Schwartz develop a distributed wireless admission control algorithm in [58] for a single traffic class. It takes into account call and mobility parameters as well as the call occupancy in the current and adjacent cells in making the admission control decision. In [75], Sutivong and Peha compare six different admission control algorithms (including [58]) for a single traffic class in a homogeneous wireless network under different conditions of load. We first extended the trunk reservation (or channel guard concept) to multi-media traffic in wireless networks in [15]. In the algorithms proposed there, the reservation partitions are static and do not vary with the offered load. Chao and Chen develop a numerical method for analyzing the performance of a class of reservation algorithms in a multi-media wireless environment in [11]. Two different multi-media reservation-based admission control algorithms were developed in [54] and [62]. They ensure the provision of adequate QoS to users in the system as defined by the probability of being dropped during service. The reservations are done in a manner, however, which does not take into account the differing QoS requirements of the individual traffic classes. Finally, a very complicated prediction-based algorithm is proposed by Levine, et. al. in [47]. It is based on the shadow cluster concept which assumes detailed knowledge of the users' routes in the system as a basis for predicting future load requirements. This prediction is done once for each call on being admitted into the system and is, in essence, a kind of reservation algorithm.

1.3 Thesis Contributions

The primary focus of this dissertation is the establishment of a mechanism which may be used to define objective QoS criteria in a multi-class wireless mobile network. These criteria are then used to assess new admission control algorithms in such systems.

The static reservation based algorithm is an extension of work done by Kraimeche and Schwartz in [40, 41, 42] and by Hong and Rappaport in [26, 27]. It was one of the first multi-class algorithms for wireless networks. It extended single class reservation for handoff users in mobile wireless networks to multiple traffic classes (here for the case of narrowband and wideband traffic) using some ideas proposed by Kraimeche and Schwartz [40, 41, 42] for multi-class users in wired networks. We distinguish between the concepts of pre- and post-reservation and consider the impact that they have on the system. Using the QoS criteria defined in this chapter, we see that by choosing the right algorithm parameters for a particular traffic mix, the system cost is minimized at both low and high loads performing close to the complete sharing algorithm (CS) at low loads and improving over the complete partitioning algorithm (CP) at high loads. The static algorithm thus maximizes throughput at low loads and maintains QoS demands at high loads while continuing to maximize throughput for that case. Thus, performance follows the complete sharing algorithm at low loads and improves over the complete partitioning algorithm at high loads. We also discuss the variation of other parameters on the system and analyze the impact that varying the size of the wideband calls has on system performance. The computational complexity is low due to the static nature of the algorithm and its simple implementation. This comes at the expense of inflexibility to changes in the traffic parameters such as total offered load and composition mix.

In Chapter 2, we develop QoS measures which are based on a cost function which compares the call blocking and handoff dropping probabilities to each other. A basic profile is developed which measures the relative importance of the quantities. It measures conformance to the desired profile as a function of the average and deviation while trying to maximize system throughput.

In Chapters 3 and 4, we develop another set of QoS performance measures. This measure assumes that each of the traffic classes has its own objective requirements defined by absolute asymptotic maximal dropping probabilities even at very high overloads and the imposition of a call blocking probability profile which may be used to enforce desired priorities among the different traffic classes. This ensures that greedy services don't crowd out other traffic classes during periods of overload. Since mobile wireless services are characterized by relatively small bandwidth availability, the overload region is frequently encountered and performance in this region more important than in wired networks. Additionally, since call continuity is directly tied into the admission process and the average number of handoffs per call is increasing as cell sizes shrink to accommodate the increase in demand, the need to enforce arbitrary maximum call/handoff dropping probabilities increases. The QoS requirements we have defined for use as a measure in evaluating the admission control algorithms address precisely these issues.

The one-step prediction algorithm family introduced in Chapter 3 is a dynamic set of algorithms which conform to the aforementioned QoS measures. They predict performance one time step into the future given the state of the home and neighboring cells. All of the algorithms meet the handoff dropping threshold requirement. IMOSP-CS and IMOSP-RES meet the call blocking requirement as well. We compare a single-class version of the algorithm, OSPRED, to a single class prediction algorithm developed by Naghshineh and Schwartz [58]. We simulate performance of the single and multi-class algorithms and analyze how parameter variation impacts on it. The algorithms automatically adjust to changes in total offered load, traffic mix, hotspots, and average time until handoff. They may additionally be set for any set of handoff dropping and call blocking probabilities. By pre-computing the probabilities for each traffic class, we minimize algorithm complexity.

In Chapter 4, we develop a completely distributed measurement-based algorithm, MMDR. This algorithm adjusts two-tier partitions to meet the absolute handoff dropping and relative call blocking requirements. MMDR improves on the long-term average performance of the one-step prediction algorithm family developed in Chapter 3. The QoS requirements are met exactly for all traffic classes. The dynamic nature of the algorithm automatically adjusts itself to variation in parameters.

1.4 Thesis Organization

Following is a short summary of the remainder of the dissertation.

Chapter 2 contains a description of the static reservation algorithm and a discussion of the QoS measures which are used to assess its performance. Analysis of the algorithm and its performance, including a comparison to the complete sharing (CS) and complete partitioning (CP) algorithms, is contained as well.

In Chapter 3, we discuss the one-step prediction algorithms. We develop the second QoS measure and introduce the one-step prediction mechanism and the blocking probability measurement function. A description of the algorithms in the family is followed by simulation and a discussion of the results.

The multi-media dynamic reservation algorithm (MMDR) is introduced in Chapter 4. A thorough discussion of the performance includes variation of algorithm parameters, system parameters, and a comparison to the one-step prediction algorithm, IMOSP-RES from Chapter 3.

Finally, Chapter 5 contains some conclusions and suggestions of areas for further study.

Chapter 2

Static Multi-Class Cellular Reservation

2.1 Introduction

As mentioned in Chapter 1, the focus of this dissertation is the application of different methodologies to the problem of multi-class admission control in mobile wireless networks. In this chapter, we develop a static multi-class reservation algorithm which has a very low level of complexity and whose admission conditions remain static at all times. We compare it to the complete sharing (CS) and complete partitioning (CP) algorithms. These algorithms may be viewed as a generalization of multi-class approaches for wired networks proposed by Kraimeche and Schwartz [40, 41, 42] applied to wireless networks. Alternatively, they may be seen as a generalization of wireless voice admission control developed by Hong and Rappaport [26, 27] and Tekinay and Jabbari [77] to include additional broadband services. These algorithms are simple in the sense that during the course of system operation, the precise rules for new and handoff user admission remain the same, independent of the total offered load or nature of the traffic currently in the network.

In this chapter, we consider a single cell in a wireless network which has an unchanging bandwidth and provides services to different classes of users. These users may request service either as a new user or a handoff user. Our analysis limits the QoS discussion to the issues of call acceptance and dropping in order to minimize the dimensionality of the problem. The resultant system model is thus essentially a circuit switched one. These requirements dictate that different types of users be accorded different levels of priority. To further simplify analysis and problem dimensionality, we do not allow queueing. We then consider several different priority access policies which span the range between complete sharing (CS) and complete partitioning (CP), focusing on hybrid policies incorporating aspects of both techniques.

The CS policy allows all users equal access to the bandwidth available at all times. In non-degenerate systems, this results in maximum usage of the available bandwidth, a goal of the network provider. However, at the same time, it does not differentiate between users of different priority. In all of the algorithms we consider in this thesis, no differentiation is made in bandwidth allocation between users requesting service and those already in service. Thus, when implementing the CS algorithm, the probability that a user is dropped from service on cell handoff is equal to the probability that the same user type is refused entry into the system. This is very problematic from the user's QoS perspective since a user would much prefer to be refused admission into the system than to be dropped in the midst of service. Additionally, the narrower-bandwidth users achieve access to the system with higher probability than wider-bandwidth users which is also unfair to the wider bandwidth users assuming that the users require the same a-priori priority.

The CP policy, on the other hand, divides up the available bandwidth into non-overlapping sub-pools according to user type. Thus, a user of type i is given access to the system provided that there are fewer than the maximum number of type i users already in service. This policy allows for more control of the relative blocking/dropping probabilities at the expense of overall usage of the network.

The hybrid policies discussed follow work done in both the wired [40, 41, 42] and wireless [66, 77] environments. They provide a compromise between the different policies by subdividing the available bandwidth into sections. Part of the bandwidth is completely shared and the other part is completely partitioned, thus giving different user types dedicated bandwidth. This allows more flexibility in catering to the QoS requirements of the different user types while maintaining higher network usage.

In this chapter, we provide a means for comparing different policies at different loads with an arbitrary number of user types. This tool allows for the easy comparison of different policies at the same time. We show the development and the use of this method in the analysis of a system consisting of two traffic types, narrowband voice users and wideband images, for two different traffic patterns over a range of different loads.

The chapter is organized as follows. Section 2.2 contains a description of the traffic model and assumptions which are used in the subsequent simulations and analysis. A description of the different access control strategies is contained in Section 2.3. This is followed, in Section 2.4, by a discussion of the performance objective and a comparative analysis of the admission control schemes. Finally, Section 2.5 contains some concluding remarks.

2.2 Basic Traffic Model and Assumptions

We consider a single cell within a wireless network with a fixed bandwidth containing a total of N Basic Bandwidth Units (BBUs). We assume that there are K distinct traffic types, each of which generates new or handoff call attempts for "channels" requiring M_i BBUs according to mutually independent Poisson processes with rates $\lambda_{NU,i}$ or $\lambda_{HO,i}$ and have channel holding times which are exponentially distributed with mean $1/\mu_i$. We note that the channel holding time of a call is dependent on both the call holding time as well as the motion of the vehicle through the cell. In spite of the fact that a user may traverse several cells during the course of service, the exponential model is still seen to be accurate [61]. We assume that both new and handover users of the same kind of traffic have the same service rate. Thus, no distinction is made between the new and handoff (HO) callers once they are connected. This is in concert with the memoryless property of the exponential distribution (and results in simplification of the model).

In general, there are many different parameters which define the QoS. These include properties such as call acceptance rate, call dropping rate, delay, jitter, and packet dropping rate [72]. In an attempt to simplify the analysis, we consider a circuit switched model which only takes into account the blocking/dropping rates of the classes of traffic considered.

Broadband systems of the future will accommodate many different kinds of traffic with different bandwidth profiles and varying QoS. Although our model could easily handle these different traffic types simultaneously, each possessing both new and handover users as well as a more detailed QoS profile, we decided to consider two different kinds of services which are representative of the traffic in the network. We therefore chose narrowband (NB) voice calls and wideband (WB) traffic such as low-bit rate video or images.

Without loss of generality, we assume that the NB traffic occupies a single BBU. For the sake of clarity we assume that the WB traffic occupies M BBUs. These traffic types were chosen since they are two of the most common kinds of traffic, especially in the third-generation mobile environment under consideration. Although the algorithms apply to any kind of NB or WB traffic, in this chapter we look at WB image traffic which is bursty in nature. A typical application would be downloading or communicating over the internet. This provides some first order results while minimizing complexity and confusion due to the interaction and inter-relationship between many different traffic classes using a network. Analyzing these two traffic types in a mobile environment as discussed would ordinarily require four streams of traffic. In an attempt to decrease the dimensionality of the system, we therefore model the wideband image traffic as a single traffic stream. This simplification is warranted for the kind of results that we are interested in for several reasons. First, the wideband traffic is bursty in nature, occasionally requiring large chunks of data for the transmission of a single image (or large file). Thus, we model each image request as a "separate" call request taken to be a new call. For the cases we consider, the image transmission time, $(1/\mu_{WB})$, is between one and two orders of magnitude less than the call holding time, $(1/\mu_{NB})$. The primary reason for a handover stream is to ensure that call continuity is maintained. Since the transmission duration of each broadband image is assumed to be small and cells overlap creating amorphous boundaries, we assume that the WB users will not have to be handed over to an adjacent cell. Thus, the existence of a WB handover traffic class will not contribute to the first order results we discuss. Lastly, since we only model call connections, if we were to model each WB user as a single call the average bandwidth transmission rate of the WB traffic would be very low. But since the peak transmission rate is often relatively large, the number of BBUs required to handle each WB channel, M, may be a sizeable portion of the total number of channels available, which would result in very inefficient usage of the available bandwidth.

This model differs from that used in Chapters 3 and 4, where the wideband and narrowband call times are comparable. In those cases, we therefore must consider handoff users for both the wideband and narrowband traffic.

2.3 Different Admission Control Policies

In the following sections, we consider three different types of admission control policies, all shown in Figure 2-1: complete sharing (CS), complete partitioning (CP), and reservation policies. The reservation policies lie on the scale between the CS and CP policies and are viewed as hybrid policies. A detailed discussion of the different policies is contained in the following sections.

As mentioned above, we assume that the system services three classes of traffic: NB new users (NU), NB handoff users (HO), and wideband new users (WB). Loosely stated, the goal is to maximize channel usage while insuring that NB calls are not dropped on cell handoff and new users are assured access to the system. The level of relative prioritization of the different users is specified by the relative blocking/dropping probabilities discussed in the previous section.

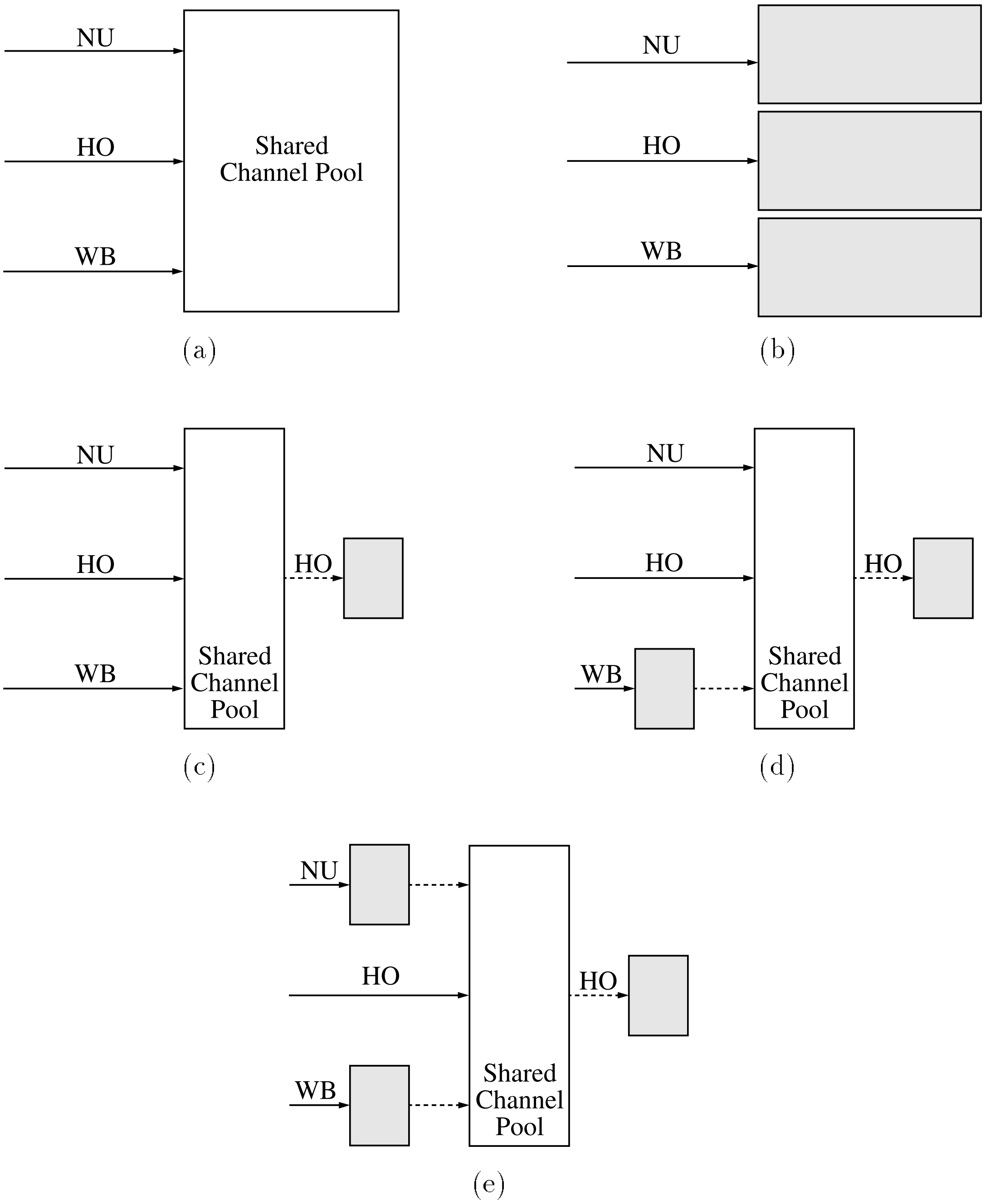


Figure 2-1: This figure contains block diagrams of all the different policies. An arriving user is accepted if at least one of the pools to which it has access is not full. Note that the shaded boxes indicate reserved channels and that traffic proceeds along the dotted arrow only when the box from which it is arriving is full on arrival (a) CS policy (b) CP policy (c) Reservation policy I (d) Reservation policy II (e) Reservation policy III

Since all of the admission control policies discussed admit users to the system based on the number of users in the system at the time of the arrival, the system is Markov. In general, we can encapsulate the behavior of each of the policies in terms of a multi-dimensional Markov chain, thus obtaining analytical results. When the Markov chain transitions can only occur between neighboring states, then the Markov process is a birth-death process and possesses special properties [6]. When we characterize the NB and WB traffic streams separately, yielding two-dimensional Markov chains, the result is a birth-death process.

Since the CS and CP policies additionally satisfy the time reversibility constraints [6], we are able to attain generalized closed form product solutions for those policies.

2.3.1 Model and Analysis of the Complete Sharing Policy

In the CS policy, shown in Figure 2-1(a), an incoming user requesting service is allocated the appropriate number of BBUs if available and is serviced on a first come first serve (FCFS) basis. This is illustrated by Figure 2-1(a). Since no distinction is made between the handoff and new users, the number of traffic classes drops to two: NB and WB. Thus, the probability of dropping a NB user on cell handoff (HO) is identical to the probability of blocking a new NB user (NU). Given that the system is in state (i, j) where i is the number of NB users in the system and j is the number of WB users in the system, an arriving NB user is admitted if $(i + 1) + Mj \leq N$, and an arriving WB user is admitted if $i + M(j + 1) \leq N$. Thus $p_{i,j}$ is defined as the probability that there are i NB users in the system and j WB users in the system. A state which does not admit users of a particular class is a blocking state for that class. When $M > 1$, there are more WB blocking states than NB blocking (in the case of a new user) or dropping (in the case of a handoff user) states. The Markov transition state diagram is two-dimensional in nature taking into account the NB and WB traffic types, and is shown in Figure 2-2.

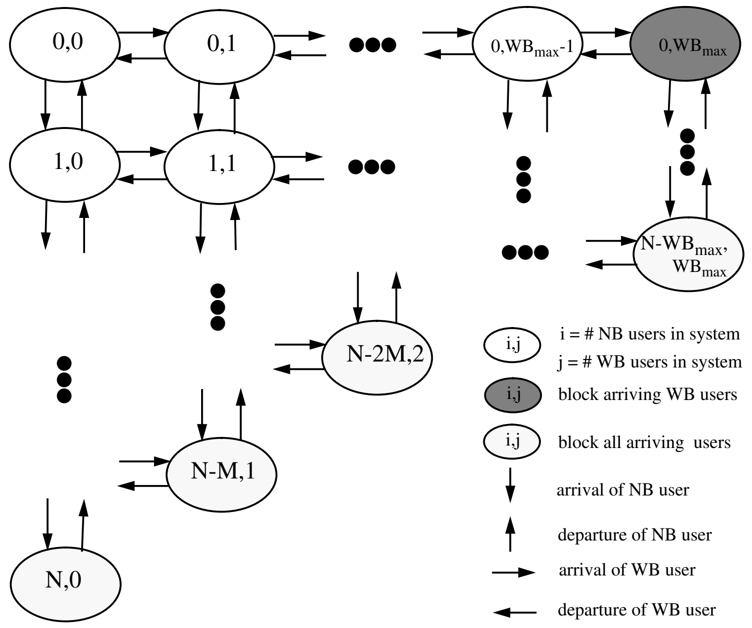


Figure 2-2: This figure contains a two-dimensional Markov figure of the CS scheme where the NU and HO classes were combined to form the NB class. The arrival rate of the NB class is $\lambda_{NU} + \lambda_{HO}$ and the NB departure rate from the state i, j is equal to $i \cdot \mu_{NB}$ according to the $M/M/m/m$ queueing scheme. The maximum number of NB users in the system is N. Similarly, the WB arrival rate is λ_{WB} and its departure rate $j \cdot \mu_{WB}$ from the state i, j. The maximum number of WB users in the system is given by $WB_{\max}$ which is equal to $\lfloor N/M \rfloor$.

The probability that a user of a given class is blocked/dropped is equal to the probability that a user of that class arrives while the system is in one of that class' blocking/dropping states. Specifically, a NB user is blocked/dropped whenever all N channels are in use. A WB user is blocked whenever there are more than $N - M$ channels in use. The values for the blocking/dropping probabilities for the NB and WB users are given below where $P_{B/D,NB}$ is the probability that a NB user is blocked/dropped and $P_{B,WB}$ is the probability that a WB user is blocked. They are achieved by summing over all the blocking states.

$$P_{B/D,NB} = \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} p_{N-j,j} \tag{2.1}$$

$$P_{B,WB} = \sum_{k=0}^{M-1} \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} p_{(N-k)-j,j} \tag{2.2}$$

Since the CS algorithm satisfies time reversibility, the blocking/dropping probabilities are product form in nature. These equations are direct multi-dimensional extensions of the $M/M/m/m$ queueing scheme.

$$P_{B/D,NB} = \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} \frac{1}{j!} \left(\frac{\lambda_{WB}}{\mu_{WB}} \right)^{j} \frac{1}{(N-jM)!} \left(\frac{\lambda_{NB}}{\mu_{NB}} \right)^{(N-jM)} p_{0,0} \tag{2.3}$$

$$P_{B,WB} = \sum_{k=0}^{M-1} \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} \frac{1}{j!} \left(\frac{\lambda_{WB}}{\mu_{WB}} \right)^{j} \frac{1}{((N-k)-jM)!} \left(\frac{\lambda_{NB}}{\mu_{NB}} \right)^{((N-k)-jM)} p_{0,0} \tag{2.4}$$

$$p_{0,0}^{-1} = \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} \frac{1}{j!} \left(\frac{\lambda_{WB}}{\mu_{WB}} \right)^{j} \sum_{i=0}^{N-jM} \frac{1}{i!} \left(\frac{\lambda_{NB}}{\mu_{NB}} \right)^{i} \tag{2.5}$$

given that λ_{NB} is the average NB arrival rate, λ_{WB} the average WB arrival rate, μ_{NB} is the average NB service rate per user, μ_{WB} the average WB service rate per user. Additionally, the normalized channel throughput is given by:

$$\text{normalized channel throughput} = \frac{\mathcal{E}\{n_{NB}\} + M \cdot \mathcal{E}\{n_{WB}\}}{N} \tag{2.6}$$

$$\mathcal{E}\{n_{NB}\} = \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} \left(\frac{1}{j!} \left(\frac{\lambda_{WB}}{\mu_{WB}} \right)^j \sum_{i=0}^{N-jM} \frac{1}{(i-1)!} \left(\frac{\lambda_{NB}}{\mu_{NB}} \right)^i \right) p_{0,0} \tag{2.7}$$

$$\mathcal{E}\{n_{WB}\} = \sum_{j=0}^{\lfloor \frac{N}{M} \rfloor} \left(\frac{1}{(j-1)!} \left(\frac{\lambda_{WB}}{\mu_{WB}} \right)^j \sum_{i=0}^{N-jM} \frac{1}{i!} \left(\frac{\lambda_{NB}}{\mu_{NB}} \right)^i \right) p_{0,0} \tag{2.8}$$

A Markov transition state diagram for the case where $N = 7$ and $M = 3$ is shown in Figure 2-3. In this example, the WB blocking states are $(0,2)$, $(1,2)$, $(2,1)$, $(3,1)$, $(4,1)$, $(5,0)$, $(6,0)$, and $(7,0)$, and the NB blocking/dropping states are $(1,2)$, $(4,1)$, and $(7,0)$.

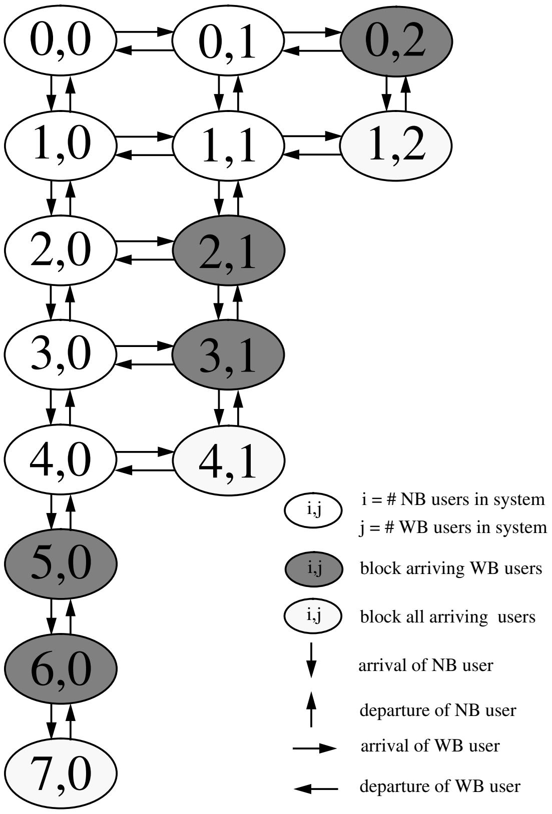


Figure 2-3: Markov state transition diagram for the CS scheme where $N = 7$ and $M = 3$.

We note that the probability of blocking a WB user is greater than blocking/dropping a NB user because there are more WB blocking states than NB blocking/dropping states, therefore providing a-posteriori greater priority to the NB users. This result is shown graphically in Figure 2-4.

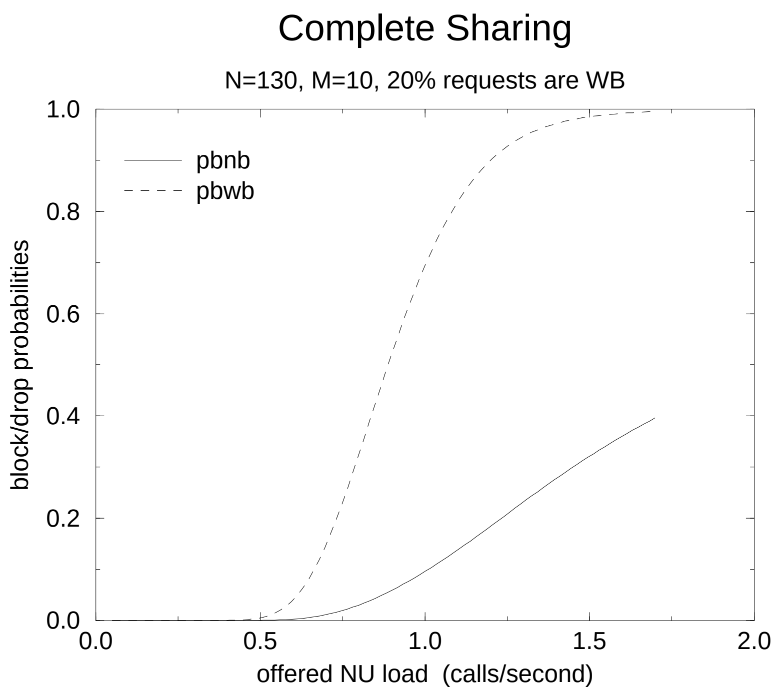


Figure 2-4: In the CS policy there are no adjustable parameters. As the load gets heavier, the policy begins to heavily favor the more NB traffic as is shown above.

Most importantly, we note that this algorithm has no variable parameters. As a result, the overall user performance, derived from the QoS parameters, may suffer. A typical case which exhibits this problem is when two traffic classes differing only in priority are serviced. Both classes of users will have the same blocking/dropping profiles even though their required priority may differ by orders of magnitude.

In conclusion, the CS policy favors channel throughput at the expense of providing a single fixed QoS profile which gives no priority to the handoff users and favors the NB users over the WB users.

2.3.2 Model and Analysis of the Complete Partitioning Scheme

In the CP policy, shown in Figure 2-1(b), the entire available bandwidth is partitioned into pools. Each pool is dedicated to a particular traffic class satisfying:

$$\sum_{i=1}^{K} M_i N_i \leq N \tag{2.9}$$

where M_i is the number of BBUs allocated to each traffic class and N_i is the number of "channels" allocated to each traffic class. Thus, a user is admitted if there is an available "channel" in the appropriate pool. Using this algorithm, the different traffic classes are completely independent of each other due to the decoupling of the systems. In our case, there are three such classes with the new and handoff users occupying a single BBU and the WB users occupying M BBUs, as noted previously.

Therefore, the equations for this system are the same as three independent $M/M/m/m$ queues where m equals the respective N_i. The Markov diagram representation may alternately be represented as a three dimensional extension of the $M/M/m/m$ queue or as three separate single dimensional $M/M/m/m$ queues. The blocking and dropping probabilities of each class are given below.

$$P_{B,NU} = \frac{1}{N_2!}\left(\frac{\lambda_{NU}}{\mu_{NB}}\right)^{N_2} \sum_{j=0}^{N_1} \frac{1}{j!}\left(\frac{\lambda_{HO}}{\mu_{NB}}\right)^{j} \sum_{l=0}^{N_3} \frac{1}{l!}\left(\frac{\lambda_{WB}}{\mu_{WB}}\right)^{l} p_{0,0,0} \tag{2.10}$$

$$P_{D,HO} = \frac{1}{N_1!}\left(\frac{\lambda_{HO}}{\mu_{NB}}\right)^{N_1} \sum_{k=0}^{N_2} \frac{1}{k!}\left(\frac{\lambda_{NU}}{\mu_{NB}}\right)^{k} \sum_{l=0}^{N_3} \frac{1}{l!}\left(\frac{\lambda_{WB}}{\mu_{WB}}\right)^{l} p_{0,0,0} \tag{2.11}$$

$$P_{B,WB} = \frac{1}{N_3!}\left(\frac{\lambda_{WB}}{\mu_{WB}}\right)^{N_3} \sum_{j=0}^{N_1} \frac{1}{j!}\left(\frac{\lambda_{HO}}{\mu_{NB}}\right)^{j} \sum_{k=0}^{N_2} \frac{1}{k!}\left(\frac{\lambda_{NU}}{\mu_{NB}}\right)^{k} p_{0,0,0} \tag{2.12}$$

$$p_{0,0,0}^{-1} = \sum_{j=0}^{N_1} \frac{1}{j!} \left(\frac{\lambda_{HO}}{\mu_{NB}}\right)^j \sum_{k=0}^{N_2} \frac{1}{k!} \left(\frac{\lambda_{NU}}{\mu_{NB}}\right)^k \sum_{l=0}^{N_3} \frac{1}{l!} \left(\frac{\lambda_{WB}}{\mu_{WB}}\right)^l \tag{2.13}$$

where $P_{B,NU}$ is the probability that a new user is blocked, $P_{D,HO}$ is the probability that a handoff user is dropped, and $P_{B,WB}$ is the probability that a wideband user is blocked. λ_{NU} is the average new user arrival rate, and λ_{HO} is the average handoff user arrival rate. N_1 is the number of "channels" allocated to new users, N_2 the number of "channels" allocated to handoff users, N_3, assumed to be equal to $\lfloor [N - (N_1 + N_2)]/M \rfloor$, is the number of "channels" allocated to wideband users, and $p_{i,j,k}$ is the probability there are i new users, j handoff users, and k wideband users in the system. The other terms are as above. Similarly, the normalized channel throughput equations are given by:

$$\text{normalized channel throughput} = \frac{\mathcal{E}\{n_{HO}\} + \mathcal{E}\{n_{NU}\} + M \cdot \mathcal{E}\{n_{WB}\}}{N} \tag{2.14}$$

$$\mathcal{E}\{n_{HO}\} = \frac{\sum_{j=0}^{N_1} \frac{1}{(j-1)!} \left(\frac{\lambda_{HO}}{\mu_{NB}}\right)^j}{\sum_{k=0}^{N_1} \frac{1}{k!} \left(\frac{\lambda_{HO}}{\mu_{NB}}\right)^k} \tag{2.15}$$

$$\mathcal{E}\{n_{NU}\} = \frac{\sum_{k=0}^{N_2} \frac{1}{(k-1)!} \left(\frac{\lambda_{NU}}{\mu_{NB}}\right)^k}{\sum_{l=0}^{N_2} \frac{1}{l!} \left(\frac{\lambda_{NU}}{\mu_{NB}}\right)^l} \tag{2.16}$$

$$\mathcal{E}\{n_{WB}\} = \frac{\sum_{l=0}^{N_3} \frac{1}{(l-1)!} \left(\frac{\lambda_{WB}}{\mu_{WB}}\right)^l}{\sum_{m=0}^{N_3} \frac{1}{m!} \left(\frac{\lambda_{WB}}{\mu_{WB}}\right)^m} \tag{2.17}$$

We note that by fixing the pool divisions, different levels of priority are accorded to the different classes of users. This is achieved at the expense of the total system throughput.

Due to the control which may be exercised in choosing the subdivisions, the QoS constraints may be more closely modeled. However, since the channel pool is not shared among the different user types, the user throughput is lower than when the CS policy is used, as a user may be blocked/dropped from service even though the entire channel pool is not in use.

2.3.3 Model and Analysis of the Reservation Schemes

We propose three hybrid policies, shown in Figure 2-1, which serve as intermediary policies between the CS and CP policies. Although they may be represented as Markovian birth-death processes, we are not aware of any generalized closed form solutions in multi-dimensional systems due to the violation of the time reversibility constraints. However, a closed form solution may be obtained by analyzing the multi-dimensional matrix produced by a particular given example.

The three policies are combinations of the CS and CP policies such that the available bandwidth is subdivided into two parts. Part of the bandwidth is allocated along the paradigm of the CS policy in the form of a shared channel pool and the remainder is allocated for CP operation as reserved channel pools. These policies reduce to the CS and CP policies in the extreme.

In the policies we discuss, we deal with two kinds of reservation: pre-reservation and post-reservation. For simplicity's sake, we focus on two abstract classes, 1 and 2. Later, we return to the case of the new user, handoff user, and wideband user classes. In pre-reservation, shown in Figure 2-5(a), an arriving class 1 user will be admitted into the reserved channel pool. In the event that the channel pool is full, the class 1 user will then contend with the arriving class 2 users for a channel in the shared channel pool. This ensures that even under heavy loads, a certain minimal traffic flow of class 1 users will be admitted. In post-reservation, shown in Figure 2-5(b), arriving class 1 and class 2 users both contend for admission into the shared channel pool. In the event it is full, class 1 users may additionally contend for admission into the post-reserved channel pool. Thus, the channels reserved for the class 1 users ensure that even under relatively heavy loads, extra priority is given to the class 1 users. Post-reservation has been used by others such as Rappaport in [66] and is similar to trunk reservation in the wired telephone world [71].

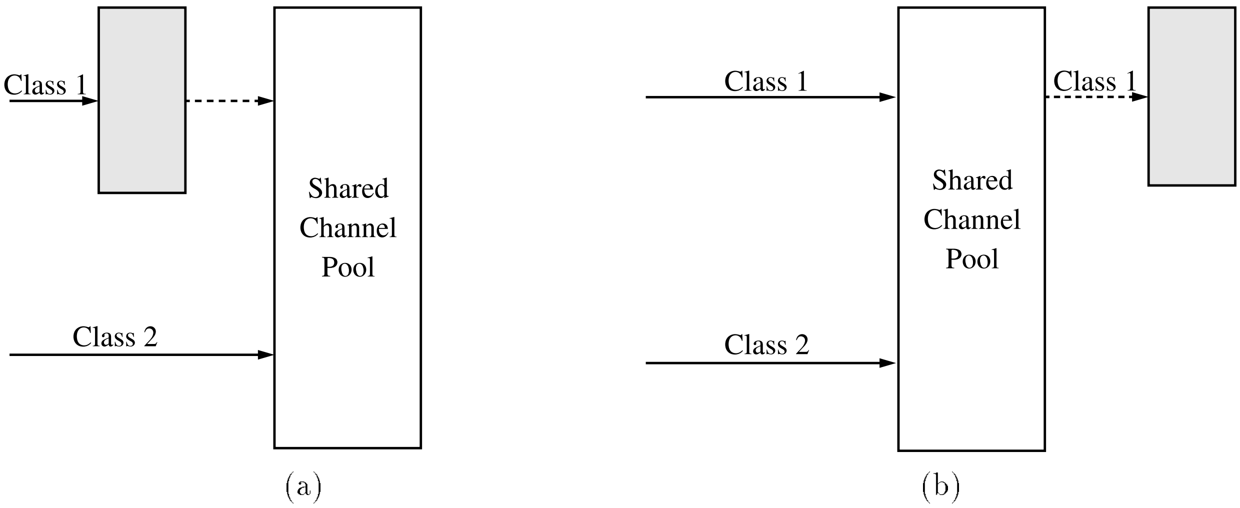


Figure 2-5: This figure contains the block diagrams for the reserved admission control issues discussed in Example 1. It may be viewed in conjunction with Figure 2-6 with one-to-one correspondence. In (a) pre-reservation is used for class 1 users and in (b) post-reservation is used for class 1 users. Shaded boxes correspond to reserved channels. Users proceed along the dotted arrow only when the box from which they are arriving is full.

The following example, example 1, graphically depicted in Figure 2-5 and analytically described by Figure 2-6, clarifies the motivation and understanding of the differences between pre- and post- reservation. We assume that we have a system with two classes of traffic, each of which requires one BBU and has identical arrival and departure rates. The system has four channels to allocate between the two user classes.

In the first scenario, we allocate one pre-reservation channel to class 1 and the remainder to a shared channel pool. See Figures 2-5(a) and 2-6(a) for pictorial and Markov diagrams.

In the second scenario, shown in Figures 2-5(b) and 2-6(b), we assign one post-reserved channel to the class 1 traffic and the remainder to the shared channel pool. In the Markov chain in Figure 2-6(b) note that there are three states $[(3,1)$, $(2,2)$,

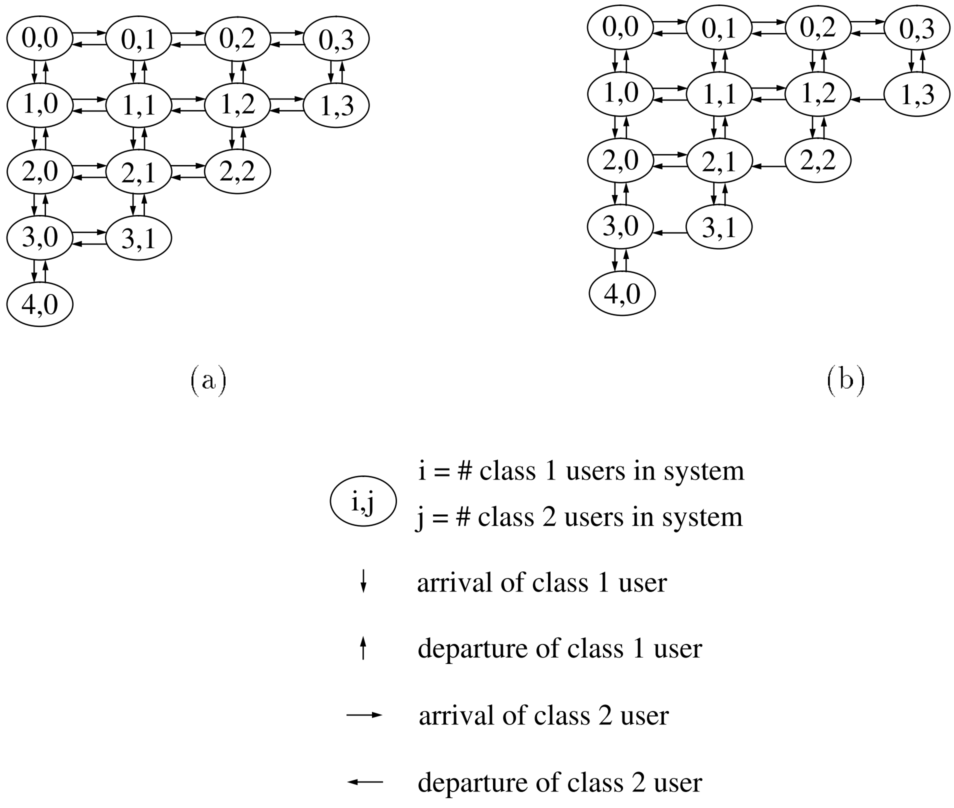


Figure 2-6: This figure contains the Markov chain diagrams associated with Figure 2-5, with one-to-one correspondence. There are 4 channels in the system. In (a) one channel is pre-reserved for class 1. In (b) one channel is post-reserved for class 1. Although the above figures may be reduced to a single dimension in this case, the diagrams are depicted in two dimensions to motivate understanding in the general case.

and $(1, 3)$] which may not be reached by the arrival of a class 2 user, giving more priority to class 1 users under heavy loads than was given in the first pre-reservation scenario to class 2 users. As a result of this characteristic the process is no longer reversible and we are thus unable to specify generalized closed form solutions for systems utilizing post-reservation.

We thus note that all things being equal, post-reservation is more powerful than pre-reservation. At the same time, it is also important to mention that it is possible to generalize the notion of pre-reservation to a multi-dimensional system and simultaneously pre-reserve channels for many different traffic types. However, the notion of post-reservation (in an exclusive sense) is not directly generalizable to a multi-dimensional system and may only be used for a single traffic type within a given system.

The three policies described in the next section are comprised of a shared channel pool together with pre- and post- reservation and are shown in Figure 2-1(c), 2-1(d), and 2-1(e). Each of the schemes discussed is a generalization of the one preceding it.

2.3.3.1 Reservation Scheme I

The first problem we noted in the CS policy was that although the blocking/dropping probability requirements of the new user and handoff traffic were very different, the probability of blocking/dropping the new user and handoff users was the same. In order to ensure that greater priority is given to handoff calls over new user calls, we use post-reservation for the handoff users [66]. In this policy, shown in Figure 2-1(c), all new callers (both new user and wideband) requesting service must rely on competition with the handoff users for the shared channels. The handoff users, on the other hand, may resort to using the reserved channels when all the shared channels are already taken. This serves to lower the dropping probability of the HO users at the expense of the new user and wideband calls. This policy provides priority to the handoff users while minimizing the overall throughput degradation we noted in the CP implementation.

2.3.3.2 Reservation Scheme II

The second problem that we noted in the CS policy was that even though the same a-priori priority was given to the narrowband and wideband users, the probability of blocking a wideband user was greater. This tendency becomes more pronounced as the overall traffic load increases and under heavy loads tends to completely exclude the wider band traffic (assuming that the occupancy distribution of the narrower band traffic is not very small compared to that of the wider band traffic). This is shown in Figure 2-4, an example with $N = 130$ channels, wideband users requiring 10 channels each. Here the wideband traffic is blocked with probability 84% at $\lambda_{NU} =$ 1.13 calls/second, while the narrowband traffic is only blocked with probability 15%. To combat this problem, the second reservation policy additionally allocates several channels using pre-reservation for wideband users since only a single traffic type in each scheme may employ post-reservation.

Reservation policy II is shown in Figure 2-1(d) and is a generalization of reservation policy I (which provides post-reservation for the handoff users) where channels are additionally pre-reserved for the wideband user.

2.3.3.3 Reservation Scheme III

Finally, we generalize the second reservation policy to additionally provide pre-reservation for the new users. This may be necessary depending on the relative arrival rates of the different traffic policies. This policy, as shown in Figure 2-1(e), provides post-reservation to the handoff users and pre-reservation to both types of new calls (new users and wideband users). This policy provides the ultimate in flexibility in allowing for the tailoring of the boundaries in consonance with the demands required by the network.

2.3.4 Notation

In the following sections we compare the performance of the CS, CP, and reservation policies. The specific CP scheme discussed is referred to as CP(i, j, k) where i, j, and k are the number of channels allocated to new users, handoff users, and wideband users respectively. Since the third reservation policy, R3, is a generalization of the first two policies, R1 and R2, with the appropriate variables set to zero, it performs at least as well as the other two reservation schemes. We therefore decided to compare only reservation policy R3 to the CS and CP policies. The specific reservation scheme discussed in the following sections is referred to as R3(i, j, k) where i, j, and k are as above.

2.4 Comparative Analysis of the Different Strategies

In the following sections, we compare specific implementations of the different policies. We analyze general trends in the exhibited behavior. This is done for several reasons. The information which is most valuable is general in nature. Also, the performance depends in large part not only on the policy (CS, CP, or reservation), but also on the implementation. It should be noted that in most cases the performance of a particular policy is very sensitive to the choice of parameters. Additionally, a brute force search (pruned heuristically) of all possible schemes would have to be conducted at every value to provide exact information due to the non-linear behavior of the policies. We also note that the graphical presentation of the results is generally the most insightful.

An event driven simulation was written to compare the performance of the different CS, CP, and reservation policies. We compared the results we got from the simulation to those produced by the equations for the CS and CP detailed in the previous section. An illustrative example is given in Figure 2-7 where we show that the simulation model correctly predicts the behavior of the $N = 130$, $M = 10$, CP$(40, 60, 3)$ system.

In analyzing the data, we assume that the ratio of the different traffic streams remains constant. As such, the data is plotted against the offered new user arrival rate or λ_{NU} which is termed offered new user load and is measured in calls/second. Another common measure is normalized utilization which is the percentage of the channel in use on average. The cost, to be defined in the following section, is a function of the call blocking rates of new voice (NU) and image (WB) users and the call dropping rate of the handoff voice users (HO). Other abbreviations include *pdho* for probability that a handoff call is dropped, *pbnu* for probability a new user is blocked, and *pbwb* for probability a wideband user is blocked.

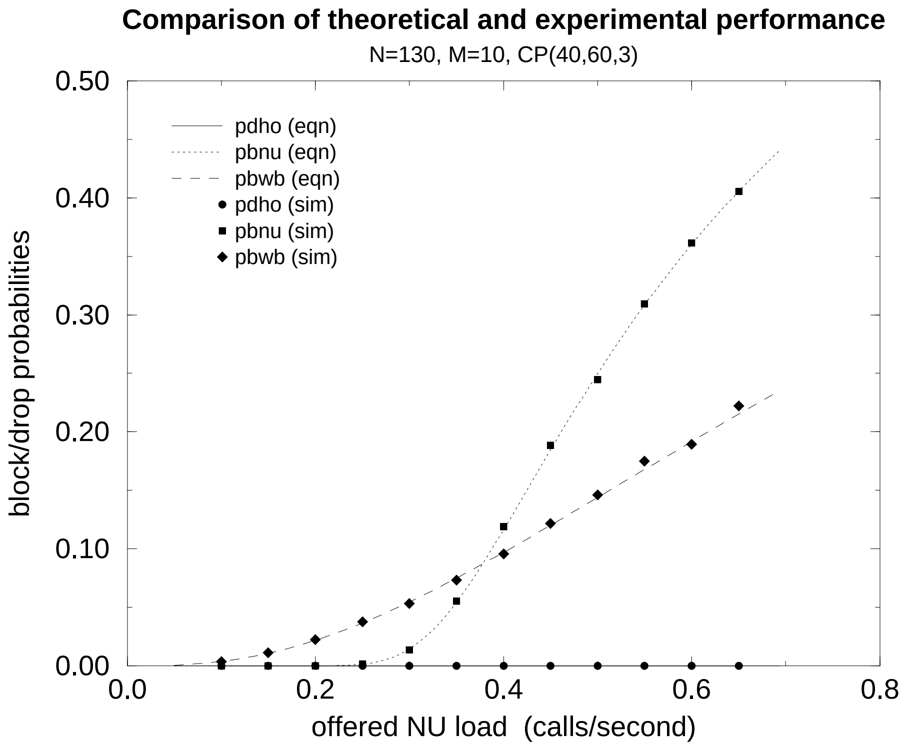


Figure 2-7: The simulation closely models the expected behavior of the CP system calculated using the CP equations (9) through (12). The ratio of new user, hand-off user, and wideband traffic load remained constant throughout, with 20% of all requests due to wideband traffic and 20% of all narrowband traffic due to call hand-off. The probability of dropping handoff traffic is much less than the probability of blocking new user or wideband traffic and thus appears to be zero, though it is not.

2.4.1 Cost Measure

As we previously discussed, the system seeks to maximize the system throughput while satisfying the QoS requirements of the different user classes. Since we analyze the system over a large range of loads, it is not meaningful to discuss a static QoS (i.e. probability blocking/dropping a user of class $i < k_i$). We rather define the QoS as a relative measure (i.e. probability of blocking/dropping a user of class $i < \theta k_i$) which we would like to conform to over the entire range of loads that we study. Because of this reasoning, we do not impose a single blocking/dropping probability for each class. Instead, we define a profile of relative blocking/dropping weights which we would like to adhere to. This profile is defined in terms of relative probability measures of the form:

$$P_{B/D,i} \propto \alpha_i \tag{2.18}$$

where $P_{B/D,i}$ is the probability that traffic of class i is blocked/dropped and the α_i are the relative weights. This is used both in assessing the strengths of the different policies discussed at any offered load and in defining a measure which is used in quantifying the relative adherence to the desired profile at a particular load.

It is very difficult, if not impossible, to monitor the adherence of a particular scheme to the QoS requirements much less compare it to the performance of another scheme by analyzing the curves themselves. We therefore define the following user-oriented cost measure which simplifies the process.

$$\begin{aligned} \text{cost measure} &= p_{avg} + \gamma p_{lsq} && (2.19) \\ p_{avg} &= \frac{1}{K}\sum_{i=1}^{K} p'_i && (2.20) \\ p_{lsq} &= \left(\sum_{i=1}^{K} |p'_i - p_{avg}|^{\beta}\right)^{1/\beta} && (2.21) \end{aligned}$$

where the $p'_i = \alpha_i p_i$ are the weighted blocking/dropping probabilities of the traffic streams mentioned above with the α_is the weighting ratios and K the number of traffic streams, (which is three in this case). As discussed previously, the QoS parameters dictate that the $P_{D,HO}$ be much less than the $P_{B,NB}$ and that $P_{B,NB}$ be the same as $P_{B,WB}$. This characterization is taken into account by the α_i. This cost measure is comprised of two components which are added together, weighting the significance of each by modifying γ. For our purposes, we assume that $\gamma = 1$, thereby giving equal weight to both components. The first component, p_{avg}, measures the weighted average of the blocking probabilities, while the second component, p_{lsq}, measures the deviation of the resulting profile from the desired profile. When $\beta = 2$, as we assume it to be, this is a least squares measure.

It should be noted that this measure seems to produce the results that we expect as will be seen for example in Figure 2-8(a) and (b), and is thus relied upon. Although we do not ascribe physical significance to the actual numerical costs, the results are illustrative of those expected in practice. A more detailed understanding of what is going on may be achieved when the cost analysis is coupled with the individual blocking/dropping probability curves.

2.4.2 System Analysis

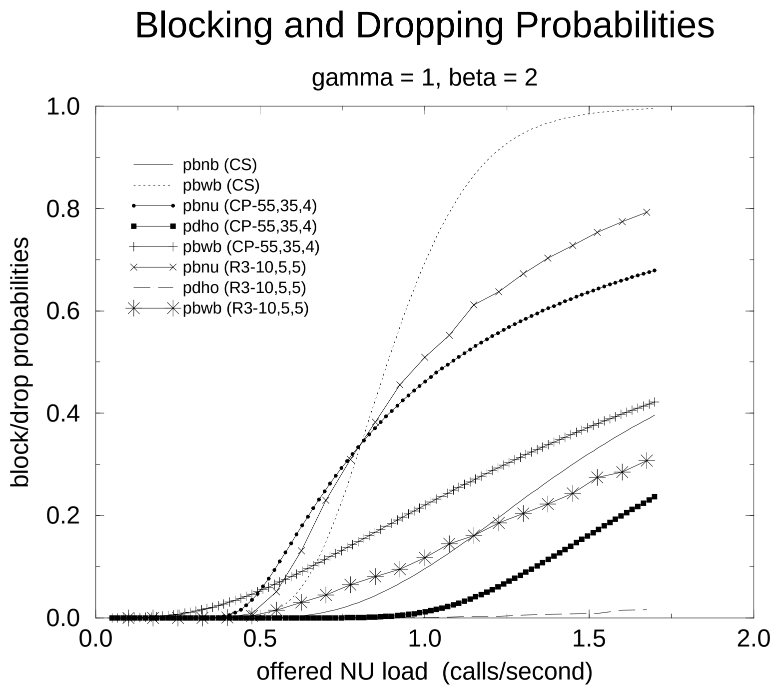
The following section contains an analysis of the CS, CP, and reservation policies discussed above in a single cell of a mobile system. We assume that the number of channels available per cell is 130 BBUs or narrowband channels, the average voice call lasts about 100 seconds – being either terminated or handed over to an adjacent cell, and that the average high quality image takes approximately eight seconds to transmit over a wideband channel 10 BBUs in size.

2.4.2.1 Case I – Macrocell Example

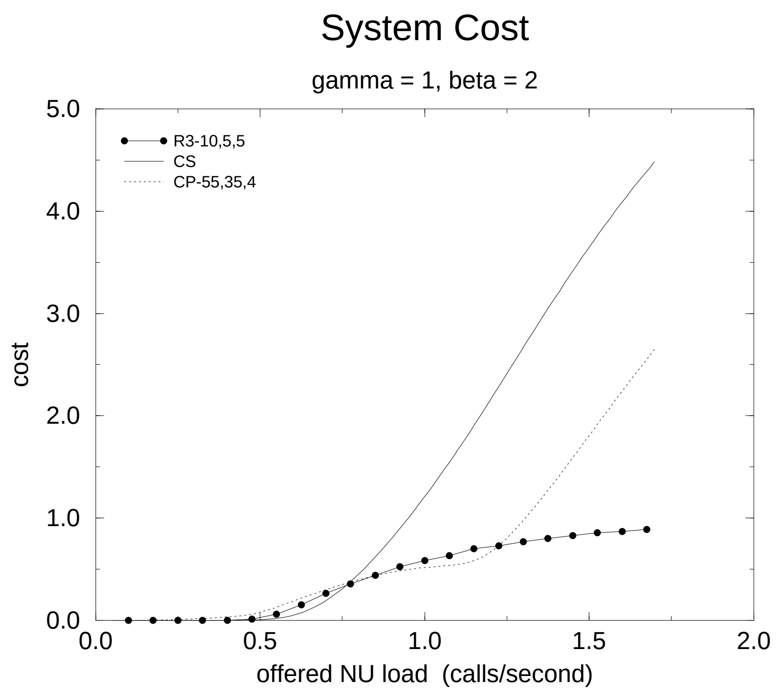
In the first case we studied, we assumed that 20% of all requests are due to wideband traffic and that 20% of the narrowband traffic requests are for handoff calls. These values are typical of traditional macrocell systems, where the cells are large in size and most traffic seeking admission into the cell is due to new users seeking access to the system. The relative weighting ratios, α_i, of the NU:HO:WB was $1 : 10 : 1$. This means that ideally the chance that a call in progress is dropped on handoff is one tenth as likely as either a new user or wideband user being refused service.

We compare the CS, CP$(55, 35, 4)$ [55 new user, 35 handoff, and 4 wideband reserved channels], and R3$(10, 5, 5)$ [10 new user, 5 handoff, and 5 wideband reserved channels] schemes. These three schemes are representative of the kinds of results that may be achieved with the CS, CP, and R3 policies as is determined by the cost function. The results that we produced are summarized in Figures 2-8 and 2-9.

The blocking and dropping probabilities (b/d) for the policies discussed above are shown in Figure 2-8(a). As expected, we note that at light loads, the b/d of all policies are small. In comparing the b/d of the CS policy, we note that beginning at about $\lambda_{NU} = .5$ calls/second the $pbnb$ and the $pbwb$ begin to diverge widely with wideband traffic being blocked more than 90% of the time at about $\lambda_{NU} = 1.2$ calls/second while only 19% of the narrowband traffic is blocked. This is a result of the a-posteriori priority achieved by the narrowband users over the wideband users in the CS policy. This phenomenon was first discussed in Sections 2.3.1 and 2.3.3.2 and shown in Figure 2-4. The CP$(55, 35, 4)$ scheme improves on this by sacrificing channel utilization and $pbnu$. Finally, we note the results achieved with the R3$(10, 5, 5)$ scheme. What differentiates this scheme from the others is the fact



(a)



(b)

Figure 2-8: Behavior of case I schemes. (a) individual b/d curves and (b) corresponding system cost.

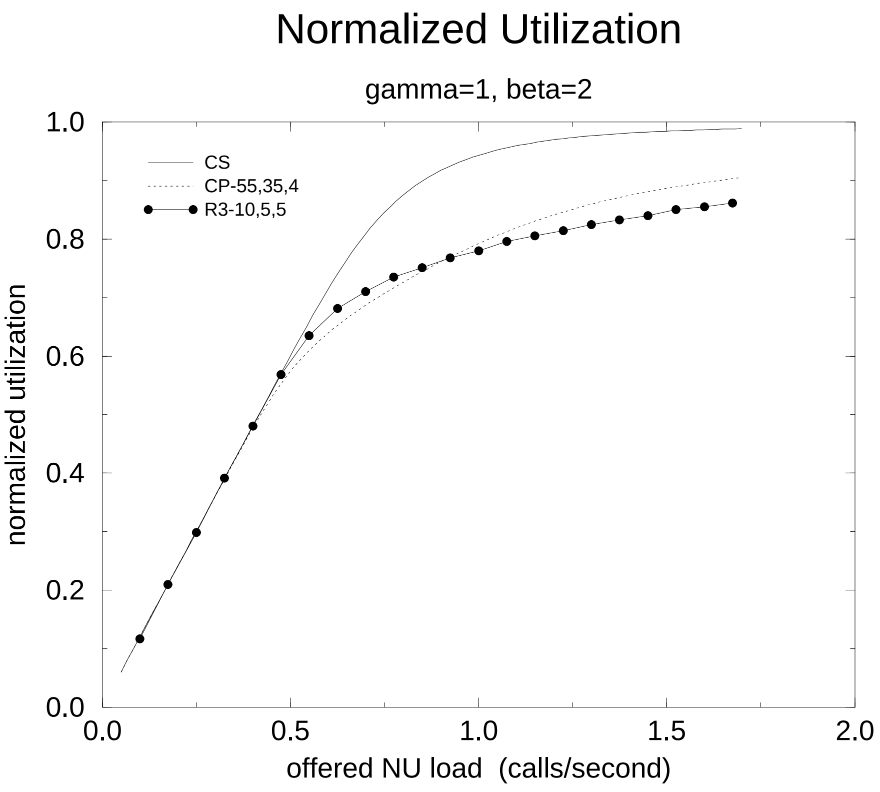


Figure 2-9: Normalized utilization curves for case I.

that the pdho remains small even at heavy loads.

We now compare the overall performance (as determined by the cost measure) of the different schemes using Figure 2-8(b). As expected, at very low loads there is no difference between the schemes. Over the next range of loads, the CS scheme marginally achieves the lowest cost. This is due to the fact that the higher utilization more than compensates for the b/d disparities. Over the next range the CP$(55, 35, 4)$ scheme achieves the best performance. Finally, at very heavy loads the R3$(10, 5, 5)$ scheme achieves the best performance. The most important thing to note in this experiment is that the cost of the R3$(10, 5, 5)$ scheme increased at worst linearly, seeming to almost level off at heavy loads. On the other hand the costs of the CS and CP$(55, 35, 4)$ schemes increased rapidly. Additionally, the R3$(10, 5, 5)$ scheme appeared to provide close to the lowest, or the lowest cost, at all values of offered load.

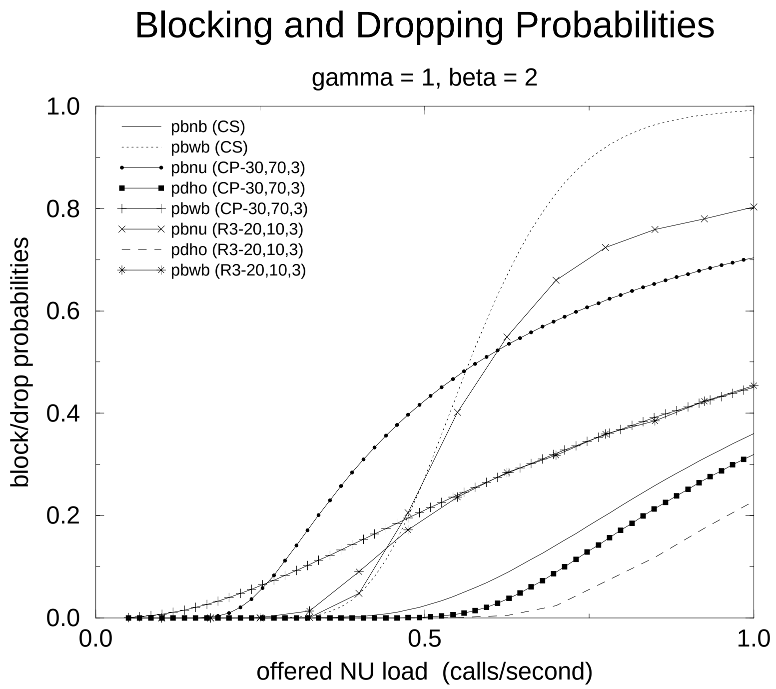
Lastly, we examine the system utilization over the same range of offered load. We note in Figure 2-9 that when the system load is light, the utilization of the different policies is essentially the same. As the load becomes heavier, we note that the CS policy outperforms the other two in channel utilization as expected and asymptotically approaches complete channel utilization. The sacrifice in channel utilization in the reservation policy is seen to be minor in comparison to the cost gain achieved by that policy over the CS and CP policies.

2.4.2.2 Case II – Microcell Example

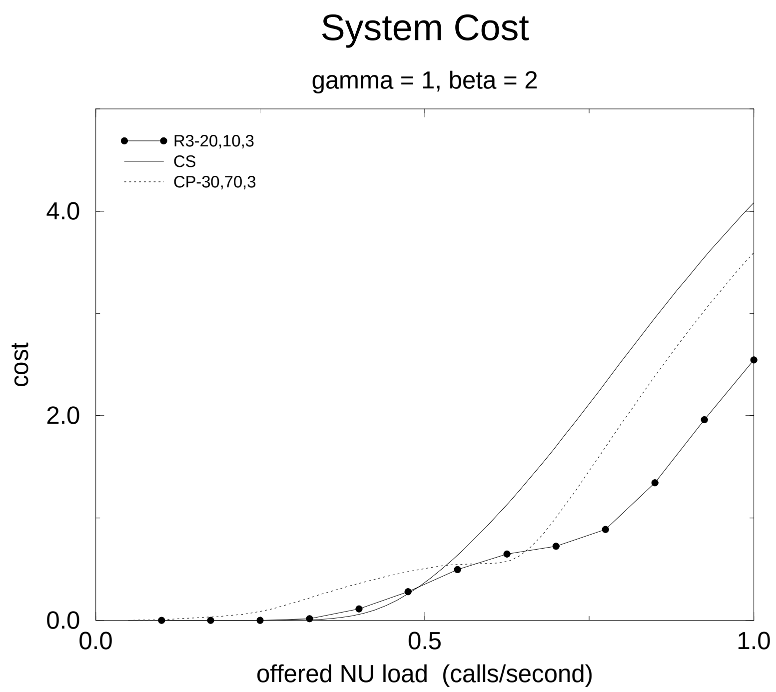
In the second case, we varied the ratio of the new user to narrowband call requests such that 50% of all narrowband call requests were handoff calls and as before 20% of all requests were wideband. In general, systems which are comprised of microcells have higher ratios of handoff to new user traffic, thus motivating this analysis. The relative weighting ratios remain as above at $1:10:1$ for the NU:HO:WB.

We compare the CS, CP$(30,70,3)$, and R3$(20,10,3)$ schemes. These three schemes are representative of the kinds of profiles that may be achieved under these circumstances. The results that we produced are given in Figures 2-10 and 2-11. Note that due to the manner in which this case was studied, the operational region occurs at a lower range of *offered new user load.*

As in case I, we note that under very light loads, the utilization of the different schemes is essentially the same. At heavier loads, the CS scheme achieves the highest utilization followed by the R3$(10,5,5)$ and CP$(30,70,3)$ schemes, as was expected.



(a)



(b)

Figure 2-10: Behavior of case II schemes. (a) individual b/d curves and (b) corresponding system cost.

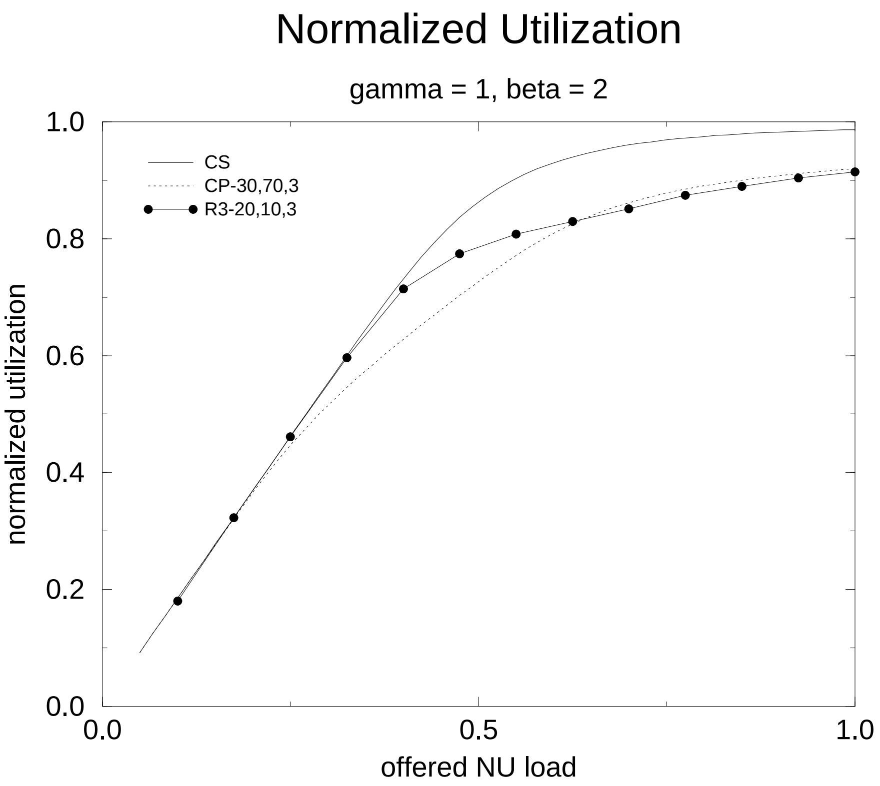


Figure 2-11: Normalized utilization curves for case II.

An analysis similar to case I may be performed on the data, providing the same kind of results with a twist which is easily seen in Figure 2-10(b). Whereas in case I, we noted that at the heaviest loads shown the cost of the reservation scheme increased at worst linearly, we note here that the reservation scheme cost in this case increases linearly at first, but eventually follows the slope of the other two schemes shown. This is attributed to the fact that a much higher percentage of the offered narrowband load (50% versus 20%) is comprised of handoff traffic. Additionally, the cost function is very sensitive (factor of 10) to fluctuations in the handoff traffic. However, the normally operable region of these schemes occurs in the region where the reservation cost is linear. Thus, once again, the usage of the reservation policy achieves qualitatively superior results over the range of normal (and even heavier than normal) operation.

2.4.2.3 Case III – Variation of Wideband Bandwidth

In the last case we studied, we varied the bandwidth of the wideband channel, M, between 2 and 20 BBUs while appropriately adjusting the wideband service rate, μ_{WB}, so that the average size of the transmitted wideband image remained constant. In other words, the quantity M/μ_{WB} didn't change. As M gets bigger, the information is transmitted to the user over a shorter period of time. As M shrinks, the information takes longer to transmit. All other parameters were left as in case I where 20% of all requests were due to wideband traffic and 20% of wideband requests were for handoff calls and the relative weighting ratios, α_i of the NU:HO:WB was $1:10:1$.

We then compared the performance of the following reservation schemes: $M = 10$, R3$(10,5,5)$; $M = 20$, R3$(10,5,2)$ and R3$(10,5,3)$; $M = 15$, R3$(10,5,3)$ and R3$(10,5,4)$; $M = 5$, R3$(10,5,10)$; and $M = 2$, R3$(10,5,25)$ where R3(i,j,k) represents i pre-reserved new user channels, j post-reserved handoff channels, and k pre-reserved wideband channels each of M BBUs. The number of reserved wideband channels was varied with M such that the number of reserved wideband channels was equal to either the floor or the ceiling of $50/M_i$ so that the number of BBUs reserved for wideband traffic was as close to 50 as possible. A comparison of the system cost and utilization of the above schemes is shown in Figures 2-12 and 2-13. The following discussion is based primarily on the results plotted there.

In cases where $50/M_i$ produced no remainder (i.e. $M = 10$, $M = 5$, and $M = 2$), the cost remained essentially constant. We divide the cases where $50/M_i$ had a

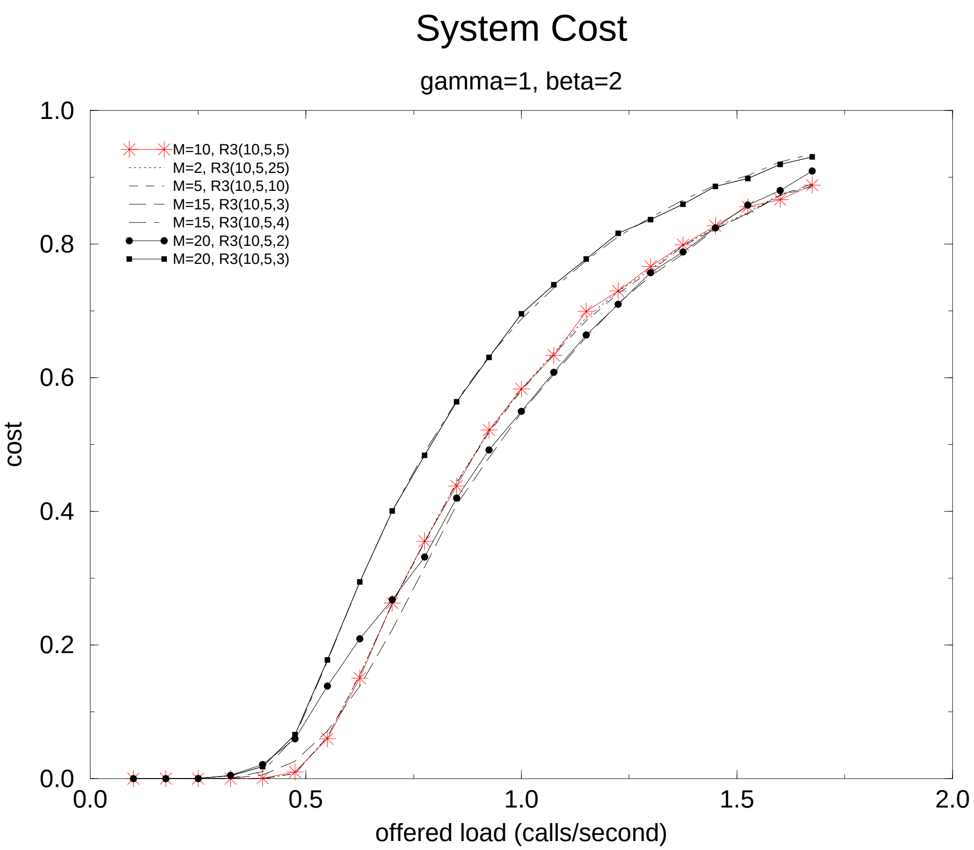


Figure 2-12: Cost curves for case III.

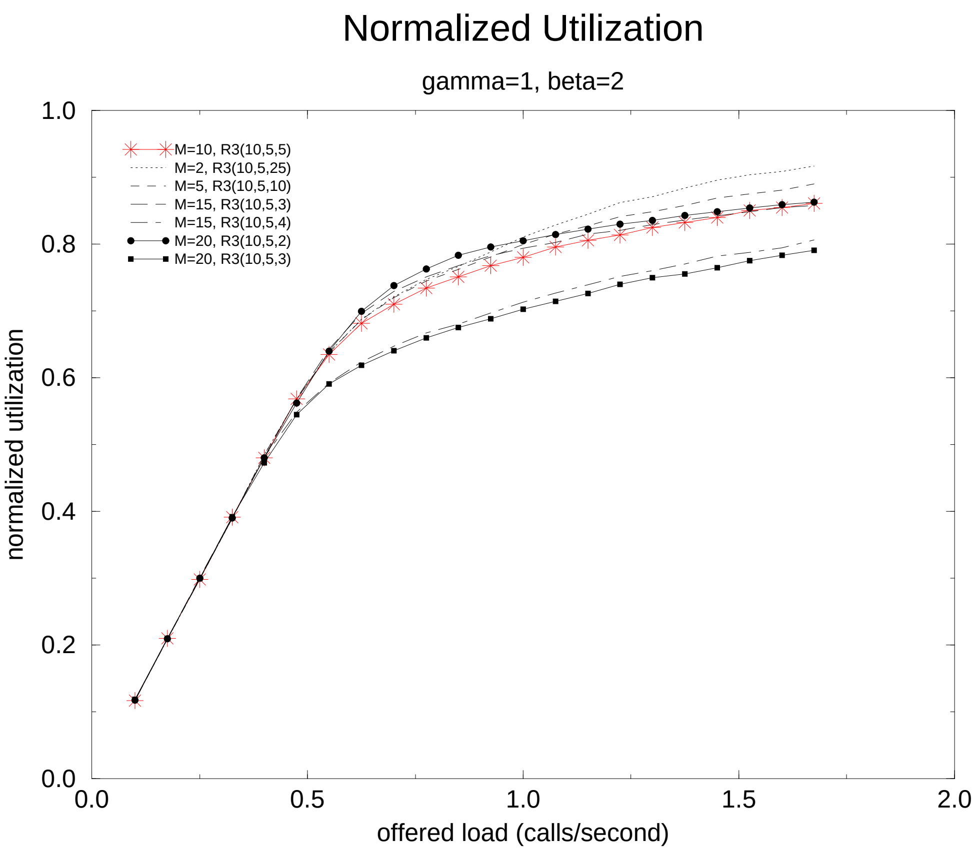


Figure 2-13: Normalized utilization curves for case III.

remainder (i.e. $M = 20$ and $M = 15$) into two groups. In the first group ($M = 15$, R3(10, 5, 4)), 60 BBUs were reserved for wideband use. In both cases, the cost was higher than when 50 BBUs (or less) were reserved. This was expected since the extra channels reserved for wideband traffic resulted in fewer shared channels available and thus raised the cost to the system. In the second group ($M = 15$, R3(10, 5, 3) and $M = 20$, R3(10, 5, 2)), 45 and 40 BBUs respectively were reserved for WB channels. Under heavy loads, this group performed similarly to the cases where 50 BBUs were reserved. However, note that for the $M = 20$, R3(10, 5, 2) case, the cost at lower loads is similar to the cases where 60 BBUs are reserved for wideband traffic. This is attributed to the fact that at low loads reserving less bandwidth for wideband traffic results in wideband traffic being blocked which more than offsets the cost gain due to the extra shared channels available which may additionally be used by the new user and handoff traffic. The channel utilization curves corroborated the above analysis, since the schemes which reserved 60 BBUs achieved lower utilization than the others. One additional observation we noticed here is that while under lighter loads, the $M = 20$, R3(10, 5, 2) scheme achieved the best utilization, under heavy loads, the smaller the wideband bandwidth, M, the greater the channel utilization with the highest utilization achieved by the $M = 2$, R3(10, 5, 25). This again was what we expected from the previous discussion in section 2.4.2.1. The *pbwb* was a little bit smaller for smaller bandwidth channels for the same reason. The *pbnu* and *pdho* quantities did not change however, as the total traffic in the system remained essentially the same.

We then compared the results of the $M = 2$, R3(10, 5, 20) scheme to the $M = 10$, R3(10, 5, 5). Although less bandwidth is reserved for the wideband channels in the first case, the $pbwb$, the cost, and the channel utilization are all better than the results of the $M = 10$, R3$(10, 5, 5)$. This is attributed to the fact that at lower bandwidths, allocations are more precise resulting in greater channel utilization and lower costs. This points to another method which could be used in channel allocation. Under light loads, transmit at full speed. However, under heavy loads – or in cases where there are not enough channels available transmit at a reduced rate. Adaptive methods such as these have been proposed for circuit switched wired networks [40]. The cost in schemes such as these is in the additional time delay the user experiences in receiving the requested information.

2.5 Conclusion

In this chapter, we discussed static reservation algorithms for systems with a given constant traffic mix. Under these assumptions, we note that the static reservation algorithms are best able to adapt the desired probability profile that is provided thereby achieving the lowest cost for the best utilization. Under light loads, the performance is close to the CS algorithm, and under heavy loads, it outperforms the other algorithms. Additionally, we are able to achieve a stable cost function as the load gets progressively heavier. This is important because the traffic load in cellular systems tends to drift a lot, especially as the complete bandwidth of the system may be only an order of magnitude greater than the constituent traffic types. Because of this, the system is extremely sensitive to all the input parameters and analyses which takes advantage of the law of large numbers arguments are often not valid. The results we discuss are applicable both to macrocellular and microcellular systems as these parallel the first and second cases of the last section.

We additionally showed that for reservation scheme R3, as the bandwidth of the wideband channel increases, the information may be transmitted to the user with less delay at the cost of throughput at heavier loads.

The results of the simulations that we ran are typical of results that may be achieved. These results are illustrative in nature and seem similar to ones that would be encountered in practice.

The results, however, are sensitive to the reservation parameters. Given that the number of users in any particular cell is small, the traffic mix is likely to vary from time to time and from cell to cell even when on average there are no long term variations in traffic. Additionally, the traffic load varies both in time and as a function of location. Though the QoS requirements remain the same, the partitions which will minimize cost will therefore vary with time and location. Improved performance is achieved with the static reservation algorithms through manual adjustment to reservation parameters on a cell by cell basis. This is impractical to implement.

We see this algorithm as proof of concept which indicates that the QoS demands may fundamentally be met. As such, in the chapters following, we consider algorithms which dynamically adjust to network parameters and conditions in a distributed fashion on a cell by cell basis. We note, however, that the model of wideband traffic is different in later chapters as is the definition of the QoS requirements.

Chapter 3

One-Step Prediction for Multi-Class Wireless Admission Control

3.1 Introduction

In Chapter 2, we discussed a static reservation approach used to service multi-class wireless networks. This approach confirmed the notion that it is possible to share the resources among the different traffic classes while at the same time provide independent QoS to all of the users in the network. Due to its simplicity, the static approach is limited in its ability to adapt to changing network parameters.

In this chapter, we discuss the characteristics and performance of four dynamic algorithms based on one-step prediction for call admission. The basic admission condition was originally developed for a single class of traffic and was extended to multiple traffic classes [17]. This one-step prediction condition is the sole admission condition in the single traffic class one-step prediction (OSPRED) algorithm discussed next.

When a new user arrives, the base station controller in the home cell and each of the adjacent cells predicts the amount of bandwidth needed to satisfy the demand in that cell a specified time interval ahead. This assumes that the number of users in the cell (including the newly arrived user) and its neighbors is known and an arbitrary maximum handoff dropping probability requirement is given. If the projected bandwidth requirement in each of the home and neighboring cells is less than the bandwidth available in each of those cells, the handoff dropping probability requirement is met. If there is sufficient bandwidth available in the home cell at the time of the new user arrival and the handoff dropping probability requirement is met, the new user is admitted. Otherwise, the new user is blocked. A handoff user is admitted assuming that there is sufficient bandwidth to accommodate that user.

The multi-media one-step prediction (MMOSPRED) algorithm extends OSPRED to systems with multiple traffic classes. Each traffic class is characterized by its average call length, mobility characteristics (average time until handoff), and bandwidth (BW) requirement. It also has a per class maximum handoff dropping probability criterion which is chosen independently for each traffic class. When a new user arrives in the system, the number of basic bandwidth units (BBUs) needed for each traffic class in the home and neighboring cells is predicted. If the total predicted demand in each of the home and neighbor cells is less than the total available bandwidth in each of those cells, and the total bandwidth requirement in the home cell including the arriving user is less than or equal to the bandwidth in the home cell, the new user is admitted. Otherwise, the user is blocked. Handoff users of any traffic class are admitted assuming that there is sufficient bandwidth available in the home cell.

In competing for admission into a cell, wider bandwidth calls tend to be blocked by more narrowband calls as was discussed previously in section 2.3.1. This bias is apparent to different degrees whenever the competing classes have different requirements. The more heavily a system is loaded, the more this effect is magnified. To reduce this inherent bias, we introduce two other algorithms which are modifications of the MMOSPRED algorithm.

The modified algorithms (IMOSP-CS and IMOSP-RES) provide the system with the ability to assign relative priority to each of the traffic classes so that the admission control algorithm will fairly apportion available bandwidth between the users of different classes while maintaining the handoff dropping probability bounds given previously. This is done using a form of dynamic reservation and is described in more detail in section 3.4.1.2. The updated algorithms differ from each other in the way they admit handoff users into the system. In the first algorithm, IMOSP-CS, handoff users completely share the available bandwidth while in the second, dynamic reservation partitions based on predictive demand are used to ensure that the dropping probabilities of each class are met independently.

The multi-class algorithms discussed in this chapter are implemented for the case of two traffic classes. However, we define the algorithms for an arbitrary number of traffic classes. All of the algorithms are simulated and results analyzed for a single-dimensional network for a number of different cases which illuminate the properties of the algorithms. The algorithms are equally applicable to networks of multiple dimensions. We expect that the results in those cases will be similar to the one-dimensional case.

The rest of this chapter is organized as follows. Section 3.2 contains a description of the traffic model. This is followed by a delineation of the QoS criteria in Section 3.3 and a description of the different algorithms in Section 3.4. Section 3.5 describes the simulation parameters used and is followed by some results and analysis of performance in Section 3.6, including a comparison with other proposed admission control algorithms. We end the chapter with some conclusions contained in Section 3.7.

3.2 Traffic Model

The first area that we discuss is the traffic model of the system. For the sake of simplicity, we consider a one-dimensional ring of cells. as is shown in Figure 3-1. The number of cells in the system is given by S. If S is sufficiently large, this system

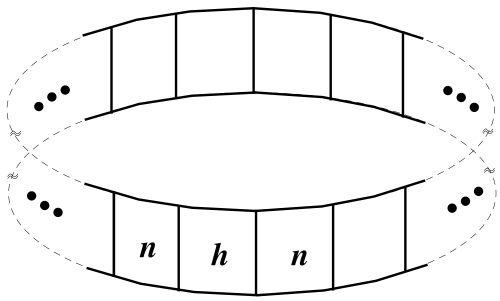


Figure 3-1: One-dimensional ring of cells.

is equivalent to a one-dimensional system. We consider the system to be shared by multiple traffic classes.

In a mobile system, three components must be considered for every user of class c: movement of the mobile within the system, call statistics, and bandwidth requirement. We assume that these components are independent and that the cells are uniform (i.e. the transitions into and out of every cell are the same). Users of each class are modelled by exponential service times (parameter $1/\mu_c$). Mobile movement is modelled by an exponential process with average handoff time $1/h_c$.

We then define the following quantities:

$$\begin{aligned} p_{e,c}(t) &\equiv \text{probability that class } c \text{ service completes in } t \text{ seconds} \\ &= 1 - e^{-\mu_c t} \qquad (3.1) \\ p_{m,c}(t) &\equiv \text{probability that a class } c \text{ user hands off within } t \text{ seconds} \\ &= 1 - e^{-h_c t} \qquad (3.2) \\ p_{s,c}(t) &\equiv \text{probability that a class } c \text{ user does not hand off within } t \text{ seconds} \\ &= 1 - p_{m_c}(t) \qquad (3.3) \end{aligned}$$

This model corresponds to a uniform Markovian system where the future state of the system is dependent only on the current state. These assumptions are appropriate for any system where the users' routes are not assumed to be directional or fixed. In our simulations, we do consider some heterogeneous cases, in particular one involving traffic hotspots in some cells. The results of these cases are discussed briefly in the results section.

3.3 QoS Criteria

In Chapters 3 and 4, we consider two different QoS criteria for each traffic class in the system. The first criterion is defined to be the expected probability of a call being dropped over the course of the call and is given by $p_{dc,c}$ for a user of class c. The second criterion is the relative priority of the different traffic classes as is manifested by the ratios of the call blocking probabilities, $p_{b,c}$.

Assuming that the network is uniform (i.e. the probability of transitioning into and out of every cell is the same) and that all cells in the network are equally loaded, $p_{dc,c}$ is related to the probability of dropping a class c call on handoff, $p_{d,c}$, in the following way:

$$\begin{aligned} p_{dc,c} &= p_{ho,c}p_{d,c} + p_{ho,c}^2 p_{sh,c}p_{d,c} + p_{ho,c}^3 p_{sh,c}^2 p_{d,c} + \cdots \\ &= \frac{p_{ho,c}p_{d,c}}{1 - p_{ho,c}(1 - p_{d,c})} \end{aligned} \tag{3.4}$$

where $p_{ho,c}$ is the probability that a class c call will be handed off before the call ends and is equal to $\frac{\mu_c}{\mu_c + h_c}$ where $1/\mu_c$ is the average length of a class c call, $1/h_c$ is the average time until a class c handoff occurs, and the call length and time until handoff are modeled by exponential random variables, and $p_{sh,c}$ is the probability that a class c handoff is successful and is equal to $1 - p_{d,c}$. This is analogous to Hong's formulation in [26]. His formulation replaces the $p_{ho,c}$ in the numerator with the probability that a new call admitted to the system hands off to an adjacent cell before terminating service.

For a given value of $p_{ho,c}$, there is a one-to-one correspondence between $p_{d,c}$ and $p_{dc,c}$. Therefore, we choose to consider the handoff dropping probability $p_{d,c}$ in analyzing the system as it is the quantity which is directly measured in system simulation.

If we assume that all cells in the system experience the same average load over time (as is the case in homogeneously loaded systems), we need only take into account cells in the neighborhood of the home cell to compute $p_{d,c}$. If the projected values of $p_{d,c}$ are below the mandated thresholds, the user is admitted into the system. Additionally, in systems where users move relatively slowly and therefore handoff very few times over the course of the call, there is no need to consider further neighbors even in heterogeneously loaded systems. However, in cases where users are highly mobile and the system load is heterogeneous, the traffic in additional cells would need to be considered in order to accurately predict $p_{d,c}$ and by extension $p_{dc,c}$. In heterogeneously loaded microcellular systems populated by fast-moving users, algorithms which take the traffic in fewer neighboring cells into account will perform more poorly than those which take the traffic in a larger area into account. Thus, the level of complexity or smartness needed to achieve optimal results is directly correlated both to the degree of traffic homogeneity within cells in a neighborhood as well as the average number of expected handoffs over the lifetime of a call.

The dropping probability QoS criterion we speak of in Chapters 3 and 4 is a threshold criterion. As such, it is satisfied so long as the handoff dropping probabilities do not exceed the maximum criteria set. The one-step prediction algorithms only consider users which are currently part of the system and not users which may arrive in the system in the future. We assert that this condition is sufficient for guaranteeing the call dropping QoS criterion of all the users in the system over all time assuming that the state of the cell neighborhood is known at time t. We expect this to be true since at no point in time is a user admitted which will (probabilistically speaking) violate the call dropping QoS criterion of the system in the future. Each admission provides guarantees for all previously admitted calls in addition to the current call being considered for admission. We also note that this criterion does not give any preference to users in any particular cell and does not take into account the statistics of the call arrival process.

The second criterion considers the level of priority of a class c call relative to calls of other classes in terms of call admission. When no control is used and all traffic classes are given equal access to the channel, classes which require less resources as is manifested by demands on the system in terms of absolute bandwidth per connection, handoff dropping probability, mobility statistics and the like (or some combination thereof) usurp a larger share of the resources leaving less for the other traffic classes. These classes are considered to be "over-privileged." "Under-privileged" classes are defined analogously to be those classes which are blocked more frequently than the "over-privileged" classes.

By using a measurement-based control in the form of reserved basic bandwidth units (BBUs) seen by incoming users, we shape the relative prioritization of the incoming users to reflect the desired blocking probability profile. This partition serves to block users of an "over-privileged" class in order to accommodate users of "under-privileged" classes. In so doing, the resulting throughput of the system is diminished. Understanding for this tradoff is motivated further in Section 3.6.

These two criteria therefore allow us to admit users of different classes onto the same network while providing the herein defined QoS to each class independent of the other traffic also using the network.

3.4 Admission Control Algorithms

In the first half of this section, we discuss the prediction and reservation-based mechanisms defined in this chapter. We first describe the prediction algorithm used to implement the QoS dropping probability condition which is the only condition in the OSPRED and MMOSPRED algorithms. This is followed by the measurement-based mechanism used in implementing the relative call blocking probability QoS criterion. The second half then describes all of the algorithms. For the sake of simplicity, the multi-class algorithms discussed in this section are initially given for two classes of traffic. They are then generalized to networks with three or more traffic classes.

3.4.1 QoS Condition Mechanisms

3.4.1.1 Predicting dropping probabilities

The heuristic predictive condition used as the basis for ensuring that the dropping probabilities of the single or multi-class traffic do not exceed pre-defined limits is applicable to mobile wireless systems of arbitrary shape and dimension. It is motivated by the following argument which serves as the basis for its definition. This condition is at the heart of all the algorithms described in this chapter.

We assume that the movement of the mobile and the call duration are modelled by independent exponential random variables. Each user is independent. We focus on the time of arrival of a user. The complete system is Markov as the future state of the system is completely determined by the current state of the system. Projecting one step of duration T_c seconds into the future, each user currently in the system may either remain in the cell it is in or move to a neighboring cell. It may therefore be modelled by a binomial random variable. In making calculations, we assume that the joint behavior of binomial random variables is Normal [28]. We additionally neglect the possibility of users having moved a distance of two or more cells.

The algorithm as formulated here applies to a two or three dimensional network topology of arbitrary shape (ex. rectangular or hexagonal grid, building). We assume here (as is typically done) that the network is uniform, i.e. the movement of the mobile is independent of location and direction. As such, the probability of handing off to or from any cell is the same as any other cell. Additionally, we assume nothing about the load offered in different cells or about the average number of users in any particular cell. Each cell has r neighbor cells and a total available bandwidth of N BBUs.

We do not assume anything about the relative mix of traffic or about the relative requirements of each class. A class c is defined by its handoff parameter h_c, its call length parameter μ_c, its bandwidth requirement in terms of basic bandwidth units (BBU) BW_c, the time step parameter T_c, and the quality of service parameter q_c. The parameter q_c, as will be seen below, is defined to be an upper bound on the handoff dropping probability $p_{d,c}$.

Say a user arrives at cell j at time t. The vector $f_c(t)$ is the class c channel occupancy vector at the time the user arrives. The entry $f_c(t)[j]$ reflects the number of class c users in the home cell including the arriving user.

We compute $N_{i,c}$, the minimum number of class c channels required in each of the home and neighboring cells i to satisfy the predicted one-step demand due to class c.

$$\min(N_{i,c}) \tag{3.5}$$

$$\text{s.t. } p_{d,i,c} \leq q_c \tag{3.6}$$

where $p_{d,i,c}$ is the handoff dropping probability of a class c call in cell i and is approximated by:

$$p_{d,i,c} \approx \frac{1}{2}\,\mathrm{erfc}\left(\frac{N_{i,c} - m_{i,c}(t+T_c)}{v_{i,c}(t+T_c)}\right) \tag{3.7}$$

erfc(x) is the complementary error function defined as $1 - \frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-y^2/2}dy$, $N_{i,c}$ is an integer greater than or equal to 0. The quantities $m_{i,c}$ and $v_{i,c}$ represent, respectively, the mean and variance of the number of class c users which will be in a given cell T_c seconds in the future assuming that no new users arrive or depart, neglecting the possibility of users travelling a distance of two or more cells during a time interval of length T_c. They are given by:

$$m_{i,c}(t+T_c) = f_c(t)[i]p_{s,c}(T_c) + \left[\sum_{s=1}^{r} f_c(t)[s]\right] p_{m,c}(T_c)/r \tag{3.8}$$

$$v_{i,c}(t+T_c) = \sqrt{f_c(t)[i]v_{s,c}(T_c) + \left[\sum_{s=1}^{r} f_c(t)[s]\right] v_{m,c}(T_c)} \tag{3.9}$$

where $f_c(t)[i]$ is the number of class c users in the center cell at time t, and $f_c(t)[s]$ is the number of class c users in neighbor s at time t (with the values including the arriving user), $p_{m,c}(T_c)$ is the binomial probability that a class c user will move to an adjacent cell in the next T_c seconds and $p_{s,c}(T_c) = 1 - p_{m,c}(T_c)$ is the probability that a class c user will remain in the home cell during the next T_c seconds. The variances $v_{m,c}(T_c)$ and $v_{s,c}(T_c)$ are given by the binomial parameters:

$$v_{m,c}(T_c) = \frac{p_{m,c}(T_c)}{r}\left(1 - \frac{p_{m,c}(T_c)}{r}\right) \tag{3.10}$$

$$v_{s,c}(T_c) = p_{s,c}(T_c)(1 - p_{s,c}(T_c)) \tag{3.11}$$

If the total predicted bandwidth required by each of the home and neighboring cells, N_i, is less than the total number of BBUs, N, the one-step dropping probability criterion is met. These conditions are given by the following equations:

$$N_i = BW_1 N_{i,1} + BW_2 N_{i,2} + \cdots + BW_K N_{i,K} \tag{3.12}$$

$$\text{if } \max(N_i) \leq N \ \forall \text{ cells } s.t. i \in \text{home, neighbor cell, admit user} \tag{3.13}$$

We call this completely partitioned multi-media one-step prediction since the predicted required capacity is computed for each traffic class independently using a single prediction time instant. This is fundamentally different than other complete partitioning schemes such as the CP algorithm described in Chapter 2 (see also [15]). This is because the partitions are dynamically adjusted based on the instantaneous traffic mix in the cell and the partitions vary from cell to cell.

3.4.1.1.1 Implementation Complexity

Since we assume that the network is uniform, predictions are independent of the mobile location. (This is independent of the degree of homogeneity or heterogeneity of the load offered to different cells in the network and discussed in more detail in Section 3.6). Additionally, since all cells are "identical", we need only compute the projected demand of a given class c in cell j T_c seconds in the future as a function of two variables: the number of class c calls in service in cell j and the sum of the class c calls in the adjacent cells. Since the projected movement between adjacent cells in the network is independent of the cell within which one is located, all of the calls in the adjacent cells may be represented by the sum of the class c calls in those cells. The projected required demand needed to meet the class c QoS criterion q_c is thus a function of two variables: the number of class c users in the home cell and the sum of the class c users in neighboring cells. This computation is done off-line for every possible combination of class c users for each of the K classes prior to system startup and downloaded to each base station controller in the form of a two-dimensional array. The prediction process is thereby converted to K simple lookups in each of the home and r neighboring cells. Assuming that each cell has the cell occupancy values for all K classes in all r neighbor cells, the computational complexity in each cell is as follows. We assume that finding the maximum of two values and comparing two values to each other are each equivalent to a single addition. Each decision then requires $Kr(r+1)$ additions, $K(r+1)$ table lookups, and the transfer of r integers to the home cell (one from each neighbor cell). If we were to remove the uniformity requirement (relating to mobility within the system), we would have to either compute a multi-dimensional lookup table instead of the two-dimensional table or perform the calculations on-line.

3.4.1.2 Measurement-based blocking probability criteria

We next describe the mechanism used to implement the blocking probability criteria of the IMOSP-CS and IMOSP-RES algorithms. This mechanism uses dynamically adjusted reservation partitions to control the blocking probability profile. At the heart of this process is a performance measurement function which is updated periodically.

The blocking probability measurement function (BPMF) controls the number of channels which are reserved for the traffic classes which have a higher probability of being blocked. Three parameters are input into the algorithm at startup: an update parameter UP, a blocking threshold parameter BT, and a desired blocking probability ratio PB_R. Roughly speaking, one wants the ratio of the blocking probabilities of the different traffic classes to be some specific value. Whenever UP new class c calls have arrived at cell j, the cell j class c call blocking probability function is updated. When the difference in BPMFs for the different classes adjusted by PB_R is greater than BT, the BPMFs are considered to be different and the cell j channel reservation is altered. At update status times, each class i in cell j is assigned a reservation partition equal to $R_{j,i}$ BBUs.

The reservation partition serves as a minimum channel pool which reserves $R_{j,i}$ BBUs in cell j expressly for use by traffic class i. These channels may only be used by class i traffic (i.e. which entered the cell either as new or handoff traffic). When a new class c user arrives in cell j, we calculate g_i, the number of BBUs reserved and used by each class, where:

$$g_i = \begin{cases} \max(f_{j,i}, R_{j,i}) & i \neq c \\ f_{j,c} & i = c \end{cases} \tag{3.14}$$

where $f_{j,i}$ is the class i BBU occupancy in cell j and $f_{j,c}$ includes the newly arrived user. If the total demand in the cell, $\sum g_i$, is less than the cell capacity N, the call blocking criterion has been met. This condition, together with the dropping probability criterion defined by equations (3.12) and (3.13) are two admission conditions which are used in admitting users into the system. Although the prediction condition must be met for all algorithms in this chapter, only IMOSP-CS and IMOSP-RES must meet the measurement-based blocking probability condition.

In order to keep track of the relative priority given to the two different traffic classes, a blocking probability measurement function (BPMF) is used for each class. The function is a weighted sum of the function from the previous interval added to the blocking probability from the current interval and is computed independently for each traffic class in each cell. We next describe the process used to update the BPMFs.

During system operation, counters in each cell keep track of the number of new users of each class that have arrived and the number admitted. Every UP class c arrivals in cell j, the algorithm enters the update status routine and updates the cell j class c BPMF, $d_{j,c}(n_{j,c})$, at update time $n_{j,c}$. First, $d_{j,c}(n_{j,c})$ is computed for cell j class c. It is given by:

$$d_{j,c}(n_{j,c}) = (1 - x_{j,c}(n_{j,c}))d_{j,c}(n_{j,c} - 1) + x_{j,c}(n_{j,c})\frac{s_{j,c}(n_{j,c} - 1, n_{j,c})}{r_{j,c}(n_{j,c} - 1, n_{j,c})} \tag{3.15}$$

where $s_{j,c}(n_{j,c}-1, n_{j,c})$ is the number of class c calls that were blocked in the interval $(n_{j,c}-1, n_{j,c})$ just concluded, $x_{j,c}(n_{j,c})$ is equal to $r_{j,c}(n_{j,c}-1, n_{j,c})/M$, and $r_{j,c}(n_{j,c}-1, n_{j,c})$ is the number of class c arrivals in period $(n_{j,c} - 1, n_{j,c})$ (which is equal to UP). An alternate way of expressing $d_{j,c}(n_{j,c})$ is given by:

$$d_{j,c}(n_{j,c}) = \frac{(M - r_{j,c}(n_{j,c} - 1, n_{j,c}))d_{j,c}(n_{j,c} - 1) + s_{j,c}(n_{j,c} - 1, n_{j,c})}{M} \tag{3.16}$$

We can gain further understanding from this equation as follows. M is the number of points that are needed for the function to measure an adequate level of significance and is determined using the input variables. Assume a blocking threshold BT which defines the level of resolution to which we would like to match the call blocking probabilities of the different traffic classes. M is then given by:

$$M = 10\left(\frac{1}{BT}\right) \tag{3.17}$$

This is the equivalent number of "events" which are needed to show significance for that level of blocking threshold. For example, if we would like to measure disparities between blocking probabilities of the two classes that exceed $BT = 1\%$, we would need ten times that number of events, or $M = 1000$ events in order for the blocking probabilities to have significance at a 1% threshold. The term $s_{j,c}(n_{j,c} - 1, n_{j,c})$

in the numerator of equation (3.15) is the number of calls blocked during the last interval. It accounts though for $r_{j,c}(n_{j,c}-1, n_{j,c})$ events as that is the number of attempted admissions as is indicated by the factor $x_{j,c}(n_{j,c})$. In order to achieve the requisite equivalent number of admission events, M, we weight the previous blocking probability function by the balance of events needed to achieve significance, which is equal to $M - r_{j,c}(n_{j,c}-1, n_{j,c})$. Thus, $d_{j,c}(n_{j,c})$ is a function which gives the blocking probability information from the last interval, $(n_{j,c}-1, n_{j,c})$, the maximum possible weight while weighting the function at the previous interval, $d_{j,c}(n_{j,c}-1)$, just enough so that the significance of the blocking probability is maintained. We note here that $d_{j,c}(n_{j,c})$ is not the blocking probability of class c during any time interval. It is, instead, a function which is related to the blocking probability of class c.

We parenthetically remark that during the startup phase (before the total number of events exceeds M), we do not compute $d_{j,c}(n_{j,c})$ as above since the total number of events in the interval $(0, n_{j,c})$ is smaller than M. Instead, the value of $d_{j,c}(n_{j,c})$ is given by:

$$d_{j,c}(n_{j,c}) = \frac{s_{j,c}(0, n_{j,c})}{r_{j,c}(0, n_{j,c})} \tag{3.18}$$

While the individual blocking probabilities are not as precise as that which is required to indicate statistical significance at the desired level, equation (3.18) is an adequate approximation in the startup phase.

3.4.1.2.1 Two Traffic Class Case

For the sake of simplicity, we first consider the $K = 2$ traffic class case. This is followed by a description of the algorithm for three or more traffic classes. After updating the appropriate BPMF, we compare the values of the adjusted BPMFs, $d_{j,I}$ and $d_{j,II}$, to each other. The blocking probability profile defines the desired relationship between the blocking probability function of the traffic classes as:

$$\frac{d_{j,I}}{d_{j,II}} = PB\_R \tag{3.19}$$

If the absolute value of the difference between the adjusted BPMFs is less than BT (resolution or significance factor) ($|d_{j,I} - PB\_R d_{j,II}| < BT$), nothing is done since the difference between the adjusted BPMFs of the two classes is numerically equivalent as defined by the significance level BT. If it is greater than BT, the reservation partitions ($R_{j,I}$ and $R_{j,II}$) must be adjusted. The pseudo-code in Figure 3-2 describes that adjustment. The reservation bounds are adaptively adjusted to reserve more

```
if |d_{j,I} - PB_R d_{j,II}| > BT
    if d_{j,I} > PB_R d_{j,II}                /* class I p_b is greater than adjusted class II */
        if R_{j,I} > 0 and R_{j,I} < N
            R_{j,I} ++
        else if R_{j,II} > 0
            R_{j,II} --
        else if R_{j,I} < N                  /* R_{j,I} = R_{j,II} = 0 */
            R_{j,I} ++
    else                                     /* adjusted class II p_b is greater than class I */
        if R_{j,II} > 0 and R_{j,II} < N
            R_{j,II} ++
        else if R_{j,I} > 0
            R_{j,I} --
        else if R_{j,II} < N                 /* R_{j,I} = R_{j,II} = 0 */
            R_{j,II} ++
```

Figure 3-2: Pseudo-code for adaptation control of reservation bounds bandwidth for the class with a larger adjusted BPMF at a rate which will maintain the desired call blocking probability profile. The reservation partitions are adjusted by increments and decrements of a single BBU independent of the number of BBUs allotted to each traffic class and the gap between the adjusted BPMFs. (We remark that we arbitrarily chose to increment and decrement by a single BBU. We expect that a more complex adjustment mechanism may be used for faster convergence to the desired values. We leave that as an area for further study.) We note that the reservation partitions are never allowed to exceed the number of BBUs allocated to the cell as this is meaningless and will lead to the entering of a degenerate state where the algorithm may break down. (When the partition is equal to or exceeds the number of BBUs in the cell, no new users of that class may enter the cell. If this occurs in all of the cells in the system, no new users of that class will ever be able to enter the system. If there is no new data from new users, the BPMFs will never change and all traffic of that class will be blocked from the system for all time. In addition, to block all new users of a particular class from entering the system in a given cell, it is sufficient to set the partition equal to the maximum number of BBUs available to the cell.) As with the dropping probability condition, the blocking probability condition involves a small number of simple computations.

The computations are split between those done on call arrival and those done each time the update procedure is invoked. On call arrival, the blocking probability condition requires two additions for the two class case (K for the K class case) given that taking the maximum of two numbers is equivalent to an addition. For reservation update, we update one BPMF and do the adaptation control. The total complexity of the condition is thus dependent on the value of UP. Typically, UP

is large enough that the update complexity adds little cost to the per admission complexity. Updating each BPMF requires two additions and two multiplications. At each update interval, we only adjust the BPMF of the class that triggered the update. Assuming PB_R is not equal to unity and comparisons are equivalent to additions, the adaptation control for two traffic classes requires one multiplication and up to seven additions.

3.4.1.2.2 Extensions to Three or More Traffic Classes

We now extend the call blocking condition described to three or more traffic classes. We consider the case where there are K traffic classes and K is greater than or equal to two. The only extension required for $K > 2$ involves a method of comparing the adjusted BPMFs of the different classes to each other as a basis for adjusting the reservation partitions. All other values described above do not change when there are more than three traffic classes.

We first define a set of desired blocking probability ratios, PB_R_c. We compare all of the different classes to class I. Thus, these ratios are defined to be:

$$\frac{d_{j,I}}{d_{j,c}} = PB\_R_c \quad \text{c} = 1, \ldots, \text{K} \tag{3.20}$$

where $d_{j,c}$ is the value of the blocking probability function in cell j for class c and $PB\_R_I = 1$. Using the transitive property, the equivalent blocking probability ratio between class r and class s is then

$$\frac{d_{j,r}}{d_{j,s}} = \frac{PB\_R_s}{PB\_R_r} \tag{3.21}$$

Next, we define the notion of a fundamental traffic class (FTC). The FTC in cell j is the class i with the smallest adjusted BPMF in cell j, $PB\_R_i\, d_{j,i}$, that has a partition $R_{j,i}$ equal to zero. Given the mechanics of the algorithm, at least one partition will always be equal to zero.

When UP calls of a given class c have arrived at cell j, the blocking probability of that class $d_{j,c}$ is updated. The FTC is then selected from the traffic classes with partition $R_{j,c}$ equal to zero. The partitions are then updated in comparison with the FTC by comparing the absolute value of the difference between the adjusted BPMFs of the other traffic classes to the adjusted BPMF of the FTC. If the difference is less than BT, the corresponding partition is not changed since the adjusted BPMFs are numerically indistinguishable (see above). If it is greater than BT, the reservation partitions are adjusted as shown in the pseudo-code in Figure 3-3. Assume that we are describing adjustments in cell j to class c. Note that adjustments are made in terms of a single BBU independent of the number of BBUs required by a given traffic class. We note here that by defining an FTC, we eliminate many of the conditions found in the case with two traffic classes. This is a direct outgrowth of that choice since by definition $R_{j,FTC} = 0$ and if $R_{j,c} = 0$ $c \neq FTC$, $PB\_R_{FTC}\, d_{j,FTC} < PB\_R_c\, d_{j,c}$.

We then need to ensure that the sum of the partitions is less than or equal to the bandwidth of the cell, i.e. $\sum_{c=1}^{K} R_{j,c} \leq N$, so that the algorithm cannot

```
if |PB_R_FTC d_{j,FTC} − PB_R_c d_{j,c}| > BT
    if PB_R_FTC d_{j,FTC} > PB_R_c d_{j,c}      /* FTC's adjusted p_b > adjusted class c */
        R_{j,c} − −
    else                                         /* adjusted class c p_b > adjusted FTC */
        R_{j,c} + +
```

Figure 3-3: Pseudo-code for adaptation control of reservation bounds in the general case

enter a degenerate state where all calls are blocked. In the case that this condition is met, the algorithm given by the pseudo-code in Figure 3-4 is followed. This

```
s = (∑_{c=1}^{K} R_{j,c}) − N
c = 0
while s > 0
    if |PB_R_FTC d_{j,FTC} − PB_R_c d_{j,c}| > BT and
        if PB_R_FTC d_{j,FTC} < PB_R_c d_{j,c}
            R_{j,c} − −
    c + +
    s − −
```

Figure 3-4: Pseudo-code for ensuring sum of reservation bounds remains within bound

will guarantee that the sum of the partitions will never exceed N since before any partition adjustment the sum was smaller or equal to N and the above algorithm will, if necessary, undo all the partitions which were increased.

3.4.2 Algorithm Descriptions

The following section describes the admission control algorithms using the QoS mechanisms described previously. We begin with a short description of the OSPRED and MMOSPRED algorithms which are based solely on the prediction mechanism. This is followed by a description of the IMOSP-CS and IMOSP-RES algorithms which also take into account the call blocking QoS criteria of equations (3.19) and (3.21). These algorithms differ from each other in the manner in which they admit multi-class handoff users into the network. These differences are based on the notions of complete sharing and reservation discussed in Chapter 2 and first described in [15].

3.4.2.1 One-Step Prediction and Multi-Media One-Step Prediction

The one-step single traffic class prediction (OSPRED) algorithm and multi-media one-step prediction (MMOSPRED) algorithm first described in [17] are completely based on the one-step prediction mechanism.

Networks with a single traffic class use OSPRED to regulate admission of new and handoff users. New users are admitted into the system assuming that there is currently enough bandwidth available in the home cell to accommodate their demand and that one-step into the future the approximate probability that there will be overload which would precipitate a drop computed as above is below the required QoS parameter q in each of the home and neighbor cells. Handoff users are admitted assuming that there is currently sufficient bandwidth to accommodate them.

MMOSPRED admission just extends this to multi-class networks. New users are admitted assuming both that there is sufficient bandwidth available in the home cell to accommodate them and that the QoS dropping probability criterion is met. Likewise, handoff users are admitted assuming sufficient bandwidth available in the home cell at the time of arrival.

3.4.2.2 Completely Shared One-Step Multi-Class Prediction

We next consider the independent multi-class one step prediction, complete sharing variant or IMOSP-CS algorithm. New users are admitted into the system assuming that the prediction-based dropping probability and the measurement-based blocking probability profile of (3.21) or (3.19) is met. Handoff users then completely share the available bandwidth and are admitted into the cell, assuming that there is sufficient bandwidth to accommodate the call. This method is simple and potentially allows maximum usage of the available bandwidth among all handoff users, as is generally the case with the complete sharing algorithm.

3.4.2.3 Reservation/Partition One-Step Multi-Class Prediction

The final algorithm is the independent multi-class one-step prediction, reservation variant or IMOSP-RES algorithm. This algorithm is based on the recognition that completely sharing bandwidth among users of different priorities as in IMOSP-CS accords those users different levels of QoS, as was discussed previously with respect to new user admission and the BPMF function.

When a new user is admitted into the system, the predictive handoff dropping probability condition dictates that calculations are made for the home and each of the neighbor cells regarding the expected capacity demand in each of these cells one step into the future assuming that no new users are admitted into the system. The calculation in the case of IMOSP-RES involves the independent prediction of the demand of each class. At that time, pre-reservation handoff partitions (first described in Chapter 2) are reset in the home cell. If the new user is being admitted into the system, the handoff partitions for each class are equal to the predicted required capacity of the corresponding class in the home cell just computed. If not, we set the pre-reservation handoff partitions based on the home cell predictions given the current number of users of each class in each of the home and neighbor cells.

When a class i handoff user arrives at a cell, we compute the following. For each traffic class aside from the class of the arriving user, we take the maximum of the BBUs currently in use for that traffic class, $BW_j n_j$, where BW_j is the number of BBUs for each class j call and n_j is the number of class j users in the cell, and the pre-reservation handoff partition of that class in the given home cell, k_j. If the sum of those maxima over the various traffic classes plus the number of BBUs needed to accommodate the arriving handoff user is less than or equal to N, the total number of BBUs alloted to each cell, the handoff user is admitted. This may be expressed by the following equation:

$$\text{if } BW_i n_i + \sum_{j, j \neq i} \max(BW_j n_j, k_j) \leq N, \text{ admit} \tag{3.22}$$

This handoff admission algorithm ensures that there is a minimum number of BBUs reserved for each class based on the predicted need of that class in that cell at that time. Any bandwidth beyond the reserved BBUs is then completely shared among the users of different classes. This process guarantees the dropping probability QoS while completely sharing the bandwidth among all classes which serves to further maximize system throughput.

3.5 Simulation Parameters

A ring consisting of ten cells was constructed as in [58, 17]. The probability of a user handing off to any given neighbor is equally likely. Given the mobility parameters of the traffic studied, a ring of size 10 is equivalent to a line of cells. A ring of twenty cells was used to analyze system behavior in the heterogeneous traffic case. The total channel bandwidth of each cell, N, is given as 50 BBUs. We focus here on the two-traffic (narrowband (NB) and wideband (WB)) traffic case. All narrowband users and users in single class simulations occupy 1 BBU and wideband users occupy 5 BBUs. The narrowband calls may, for example, be voice and the wideband calls low-rate video. The handoff and call times are assumed to be exponential. Both traffic classes were assumed to have the same mobility parameters. The average handoff time chosen was 100 seconds unless otherwise noted and the average call holding time 500 seconds as in [58, 17]. This is assumed to model an average call in a macrocellular system. The estimation interval is set to 100 seconds unless otherwise noted. This choice was shown to be appropriate in [17].

Traffic arrivals are Poisson and are the same for all cells in the network. The offered traffic parameter for K classes of traffic in units of BBU/second is then given generally as follows:

$$\lambda_T = BW_1\lambda_1 + BW_2\lambda_2 + \cdots + BW_K\lambda_K \tag{3.23}$$

Unless otherwise indicated, we assume that 75% of the traffic is due to narrowband traffic in BBU/second requiring 1 BBU and 25% of the traffic to the wideband traffic requiring 5 BBUs. In that case, this gives us

$$\lambda_T = \lambda_1 + 5\lambda_2 \tag{3.24}$$

which may be reduced to yield $\lambda_1 = 15\lambda_2$. The offered load which is the abcissa of most of the plots is λ_T. The average amount of bandwidth offered to each cell is thus the same for all cases and may be compared.

3.6 Results and Analysis

In this section, we discuss the family of one-step prediction algorithms. We first comment on the use of the handoff dropping probability instead of the call blocking QoS criterion. This is followed by a discussion of the single class algorithm OSPRED and the multi-class algorithms MMOSPRED, IMOSP-CS, and IMOSP-RES. We provide both performance analysis of each algorithm separately together with a discussion of the sensitivity to choice of algorithm parameters, and comparison of the algorithms to each other. We additionally compare OSPRED to a similar prediction algorithm proposed by Naghshineh and Schwartz in [58] and abbreviated NPRED.

3.6.1 QoS Criteria

As discussed previously, the measures of concern to the network provider are the call blocking and call dropping probabilities and system throughput. The call blocking probability determines the operating point of the system while the call dropping probability determines the QoS provided by the system. In developing both the OS-

PRED and MMOSPRED algorithms, we predict the probability that a call already in service will be dropped on handoff in both the home and neighboring cells.

Equation (3.4) provides an analytical connection between the handoff and call dropping probabilities which is valid assuming that the probability of handoff and the probability of being dropped on handoff remain constant for all calls of the given class in the network. In Figure 3-5 belowr, we see that given the handoff dropping probability, the predicted call dropping probability indeed is the same as the measured call dropping probability.

However, we note that this holds true on a per call basis for the case where all cells in the system are equally loaded and the probability of handing off is the same in all cells. The above would not hold true by definition for those cases where either of these conditions is violated even though the long term averages over all calls in the network may show this to be true.

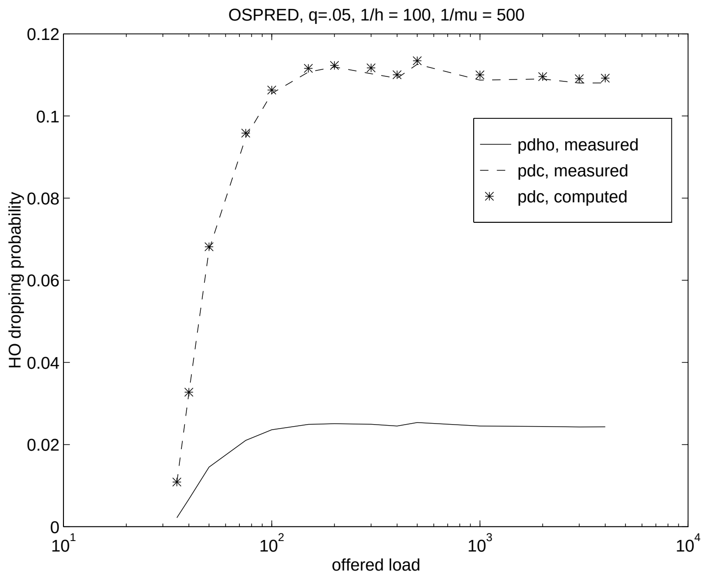


Figure 3-5: OSPRED: call dropping probability computed using p_{ho} (equal to .8345) and measured p_d where $N = 50$, $1/h = 100$, $1/\mu = 500$, $T = 100$, and $q = .05$

In the rest of the chapter, we therefore consider the handoff dropping probability as the quantity of interest.

3.6.2 OSPRED Performance

In the following sections, we analyze the performance of the OSPRED algorithm. We look at the sensitivity of the choice of the parameters T and q on the behavior of the algorithm as well as the ability of the algorithm to meet the previously defined QoS criterion. We additionally compare the results that we achieve using OSPRED to the NPRED algorithm proposed in [58].

We first make some comments regarding the general behavior of the OSPRED algorithm. The handoff dropping probability is approximately constant in the overload region. As is shown in Figure 3-6 and verified later in Figure 3-8, the overload dropping probability may be selected by proper choice of the QoS parameter q of equation (3.6). Once selected, the choice of a maximum call blocking probability then determines the nominal operating load in Erlangs. As an example, in Figure 3-6, the nominal load for a call blocking probability of 1% is about 35 Erlangs.

The algorithm maintains handoff dropping probability and hence call dropping probability constant over a wide range of loads, – something most other algorithms are incapable of doing. Note also that the handoff dropping probability (and hence call dropping probability) varies monotonically with q. As such, different dropping probability values may be achieved by varying q. It is because of this property that we call q the QoS parameter. Figure 3-6 contains simulation results of two typical systems. The operating point of the systems shown falls at the extreme left hand side of the plots as it is in this region that the call blocking probability is

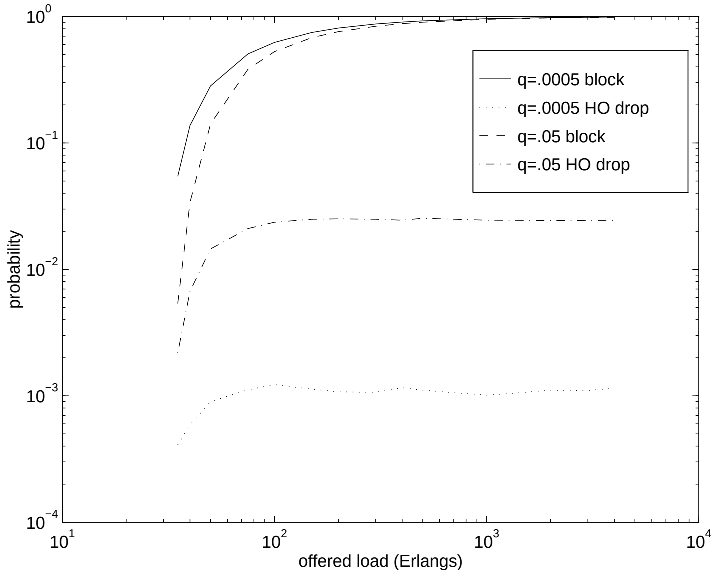


Figure 3-6: OSPRED: impact of variation of q on blocking and dropping probabilities.

within a reasonable range. We are most interested, however, in the overload cases as it is there that most algorithms are not able to maintain the guaranteed QoS. By reducing q, virtually any reduction in desired call dropping probabilities may be attained. This is compensated for by a (much smaller) increase in the call blocking probability, p_b.

The one-step prediction technique appears to provide good admission control in the case of homogeneous traffic loading. Multiple-step prediction would not appear to provide significant improvement.

3.6.2.1 Selection of the Time Step Parameter T

Time T is the basic time step prediction parameter which is used to determine the probabilities of handoff and/or ending the call. The choice of T has a profound impact on the performance of the prediction and of the resulting algorithms. If

T is too large, the Markov assumptions are no longer valid as there exists a non-negligible probability that users are handed off two or more times and thus may be two or more cells away, thereby violating the fundamental assumptions upon which the algorithm is based. We note here that this is not necessarily of concern in the cases we are considering since we assume that the offered traffic load is the same for all cells in the system as is the probability of handoff homogeneous. If T is too small, the state of the system has not significantly changed and the prediction is a poor indicator of the future system state.

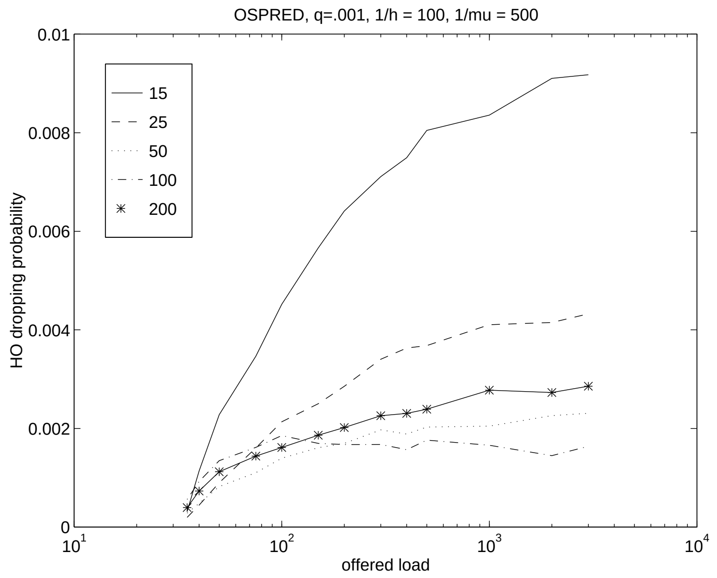


Figure 3-7: OSPRED: $q = .001$ and $15 \leq T \leq 200$ seconds.

Figure 3-7 shows the handoff dropping probability as a function of offered load for the case where $q = .001$, $N = 50$ BBUs, $1/h = 100$ seconds, $1/\mu = 500$ seconds, and $15 \leq T \leq 200$ seconds. For all values of T, the handoff dropping probability reaches an asymptotic maximum value as the load increases. The shape of the curve is similar for all cases. For relatively small values of T, p_d is larger, with the poorest prediction when $T = 15$. Other values of T first provide a lower dropping probability than the $T = 100$ case, but when the system is further overloaded, they exceed the dropping probability achieved by $T = 100$. The minimum is achieved for $T = 1/h = 100$ with the handoff dropping probability increasing when $T = 200$ seconds. By plotting the data as a function of T for different loads, we were able to ascertain that the smallest spread in handoff dropping probability occurs when $T = 100$. This property is evident in Figure 3-7 as the handoff dropping probability profile is flattest for $T = 100$. Additionally, the curve achieves the asymptotic value earliest for $T = 100$ seconds. The case where $T = 100$ is the best predictor of the behavior of the system both in the short and long term which is indicated by the fact that the target prediction is met even where the values of load are close to the system operating point with no deterioration in performance as the system load increases.

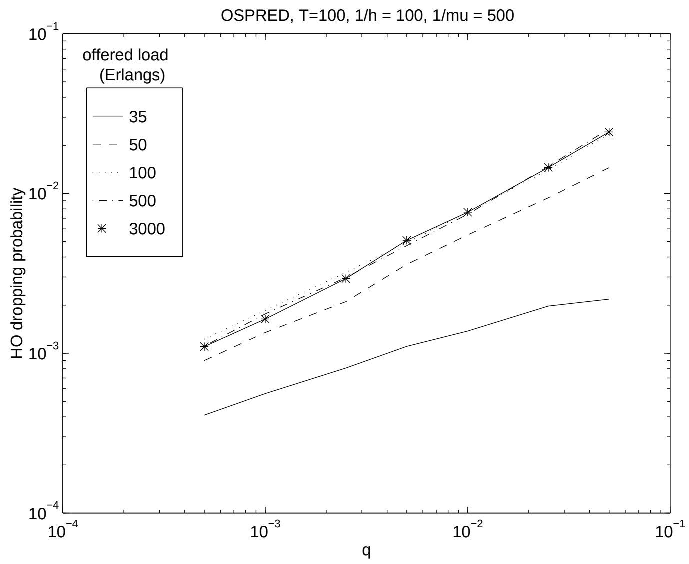
Although we only show a single case, we repeated these experiments for many of the different configurations considered in this chapter and the behavior was similar. From all of these results, it is clear that this is true for cases where the call time is much larger than the handoff time. In cases where the average call time is about the same as the average handoff time or smaller than the average handoff time, it is not as clear. The prediction process might need to take the call process into account in those cases as well in order to achieve good performance. We also surmise that if the average handoff time tends towards infinity, then the same performance would be achieved for all values of T. Further study of the algorithms is needed to completely understand the process. We therefore conclude that the prediction is best in the case where the step is approximately equal to the average handoff time. This supports our previous general assertions and makes intuitive sense. OSPRED (and by extension the other algorithms in this chapter), however, is not that sensitive to the choice of T.

We additionally note that the call blocking probability was completely insensitive to the choice of T. This is to be expected given the fact that we are in the overload region where the offered Erlang load is greater than the channel capacity N. While variations in the value of T have a significant impact on the handoff dropping probability, this is not the case for the call blocking probability.

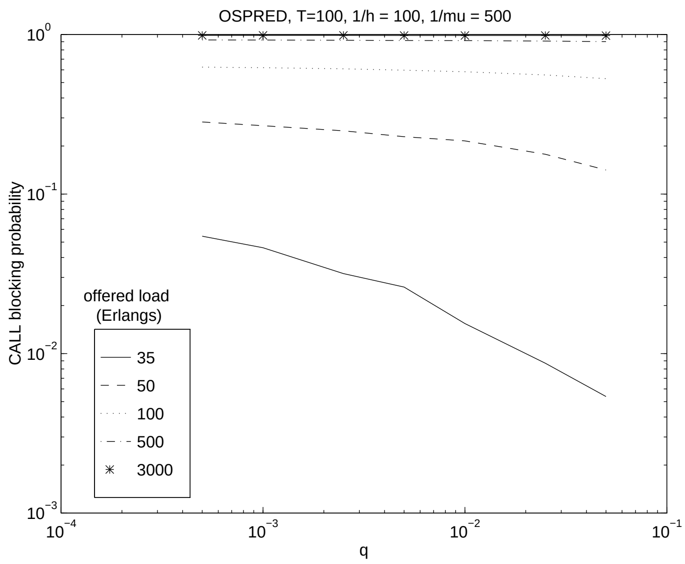
3.6.2.2 Choosing the QoS Parameter q

The second parameter that needs to be chosen in designing OSPRED is the QoS parameter q. Figure 3-8 contains a plot of the handoff dropping probability versus q where q varies between .0005 and .05 and the offered load varies from 35 to 3000 Erlangs on a channel with 50 BBUs. Figure 3-8(b) contains the call blocking probability as a function of q. As noted previously, the offered load of 35 Erlangs represents the approximate nominal load for a system with $q = .01$ and a desired call blocking probability of 1%.

The following observations are made from the data. We note that there is essentially a linear relationship between q and the achieved handoff dropping probability. Thus, the algorithm does indeed perform as was expected. We note from Figure 3-8(a), as pointed out earlier, that the handoff dropping probability remains almost constant over a wide range of offered load once the offered load in Erlangs exceeds the capacity, in this case $N = 50$. At the rated load (or operating point of the system), the relationship is still essentially linear, though the slope is somewhat different. We



(a)



(b)

Figure 3-8: OSPRED: (a) handoff dropping and (b) call blocking probabilities as a function of QoS parameter q.

surmise that since the channel is "emptier," it is more difficult to achieve the rated handoff dropping probability. On closer examination of the call blocking probability curves and their relationship to the handoff dropping probabilities, we suggest the following. In the cases where the the rated load is larger than the capacity of the channel, the variation of q has little impact on the call blocking probabilities (as is to be expected due to the non-linear relationship between the two). In the vicinity of the operating point of the system, the system is much more lightly loaded and variations in q result in corresponding variations in the call blocking probability. As q gets larger, the handoff dropping probability starts to approach the call blocking probability and thus the slopes change there.

The way that the algorithm works is that it trades off the call and handoff dropping probabilities. When the system is fully loaded it is easier to control and predict what will occur.

We conclude that by varying q, we can achieve the desired handoff dropping probability profile. This relationship is one-to-one and essentially linear. Given a desired call dropping probability profile, it is easy to achieve it using the above relationship. All extra bandwidth is then allocated to the new calls and results in greater system throughput.

Since OSPRED is predictive in nature and does not involve the reservation of integral basic bandwidth units (BBUs), any call dropping probability requirement may be met exactly, thus achieving a corresponding high level of throughput. This is justified using the following argument. Assuming that the call dropping probability is held constant, throughput or carried traffic is proportional to the number of users which are admitted into the system. Incremental improvement to the handoff dropping probability (and thus call dropping probability) is achieved through a much greater reduction in users admitted into the system. Thus, using an algorithm which overshoots the handoff dropping requirement automatically reduces system throughput. These results apply both in the case where the offered load is homogeneous over all cells in the network as well as when the offered load is non-homogeneous and individual cells or groups of cells have an average offered load which is different than that of the other cells in the network. By the same token, the algorithm adapts automatically to time-varying fluctuations in offered load. Through the application of this concept, OSPRED provides a handoff dropping probability QoS measure to the different classes of users. This property is also true for the multi-class algorithms which are based on the same principle and are discussed in section 3.6.3.

3.6.2.3 Comparison to Another Prediction Algorithm

The performance of OSPRED is compared with the prediction algorithm proposed by Naghshineh and Schwartz (NPRED) in [58]. The results for the NPRED algorithm are taken from Figure 7 in [58] where the parameter $T = 20$ and a parameter $a = 2.35$.

Figure 3-9 shows the handoff dropping probabilities and call blocking probabilities for the cases discussed above. At low loads (essentially where we generally expect to operate), both algorithms achieve the same performance, both in terms of handoff dropping probabilities and throughput (or call blocking probabilities). However, as the load starts to increase, the OSPRED achieves a higher throughput (lower blocking probability) while the handoff dropping probabilities and call dropping probabilities remain fixed. Note that with NPRED the handoff dropping

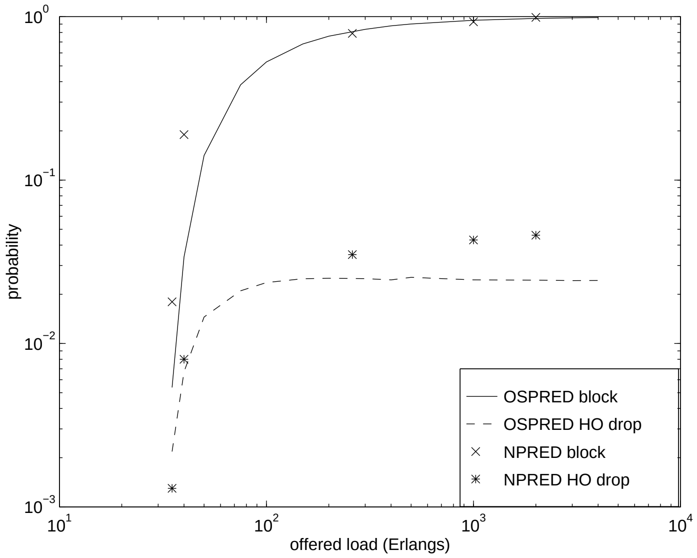


Figure 3-9: OSPRED and NPRED comparison. In both cases, $N = 50$, $1/h = 100$, and $1/\mu = 500$. The OSPRED parameters are $T = 100$ and $q = .05$. NPRED results are from Figure 7 in [58] where $T = 20$ and $a = 2.35$.

probability increases with load while with OSPRED this probability is kept essentially constant for a wide range of loads. (As noted in [15, 58], the standard guard channel reservation algorithm has an even more pronounced increase in handoff dropping probability with increases in offered load).

The OSPRED algorithm requires the input of two parameters, namely q and T, while the NPRED algorithm uses four parameters: T, a, λ, and the average number of users in a cell. The NPRED algorithm only requires knowledge of cell occupancy in 3 cells while OSPRED requires knowledge of cell occupancy in 5 cells. In the case where the offered load may vary significantly from one cell to another, the average cell occupancy may also vary among cells. In that case, the actual knowledge of the occupancy in the 2 outermost cells would lead to better performance of the NPRED algorithm. This would need to be verified by simulation.

3.6.3 Multi-Class Algorithms

In the following sections, we discuss the multi-class extensions of OSPRED: MMOSPRED, IMOSP-CS, and IMOSP-RES. Without loss of generality, we consider the two-traffic-class case. Both classes have the same traffic parameters $1/h = 100$ seconds and $1/\mu = 500$ seconds as above. The channel bandwidth is also unchanged at $N = 50$ BBUs. The narrowband traffic occupies 1 BBU and the wideband traffic 5 BBUs. We chose T for both classes to be equal to $1/h$ as per the results for the single-traffic-class case.

We consider the case where the relative call blocking ratio PB_R is set to one, and the percentage of the offered load (PL) due to the class II wideband traffic is 25%, unless otherwise indicated. We require both classes of traffic to have the same blocking probability to within a difference of 1%. Unless otherwise stated, UP is set to 250 and the blocking threshold, BT, is equal to .01 (or a difference of $\pm 1\%$). Unless otherwise indicated, for MMOSPRED and IMOSP-CS, $q_I = q_{II} = .05$ and $q_I = .02$, and $q_{II} = .0005$ for the IMOSP-RES algorithm. These values were chosen since it is easier to analyze algorithm performance for large values of q_I and q_{II}. However, the parameters may be adjusted to provide any values of asymptotic handoff dropping probabilities.

In discussing and comparing the various algorithms, we use throughput as a performance measure, in addition to handoff dropping and call blocking. We initially focus on those quantities and later discuss throughput as well. Throughput is defined as the percentage of the bandwidth available to the system which, on average, is being utilized. Thus, a call which is admitted into the system but is blocked on handoff contributes to system throughput until it is dropped. We define system throughput for class i to be:

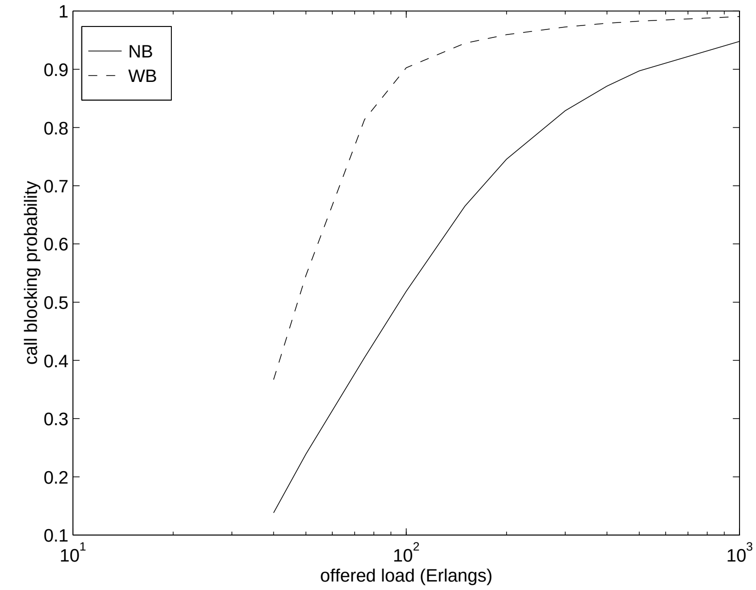
$$\gamma_i = \frac{(A_i + H_i)BW_i/(\mu_i + h_i)}{TT \cdot N \cdot S} \tag{3.25}$$

where A_i is the number of class i new users admitted into the entire system over the simulation time, H_i is the number of class i handoff users admitted into the entire system over the simulation time, BW_i is the number of BBUs occupied by class i, $1/(\mu_i + h_i)$ is the average time a user spends in a cell before either being handed off to an adjacent cell or terminating the call (where $1/\mu_i$ and $1/h_i$ are respectively the average call length and the average time until handoff), TT is the total simulation time, N is the bandwidth available to each cell, and S is the number of cells in the system. The total system throughput is the sum of the throughput due to the different classes in the system.

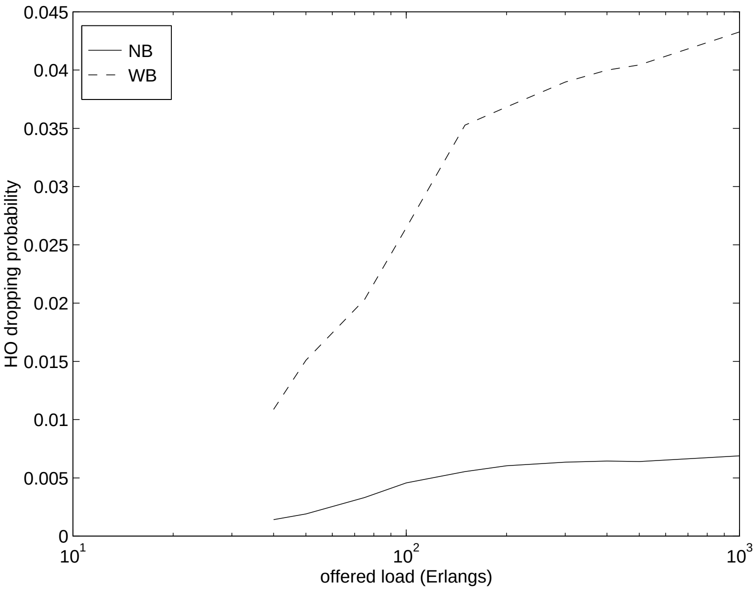
3.6.3.1 MMOSPRED Performance

We consider the case where q was set to .05 for both traffic classes. We first look at results for the basic case considered. They are summarized in Figure 3-10.

We note that handoff and call dropping probabilities exhibit the same properties as in OSPRED. Each traffic class attains an asymptotic dropping probability which is maintained over a very large range of loads. The handoff blocking probabilities also achieve results similar to OSPRED. Results from additional simulations indicate that though the variation of a single parameter q_c impacts the desired class in the expected manner, it also impacts the other traffic class. This is understood since the two traffic classes utilize the same channels and, on call handoff, bandwidth is



(a)



(b)

Figure 3-10: MMOSPRED: $q_I = q_{II} = .05$ (a) call blocking probabilities. (b) handoff dropping probabilities.

completely shared. Therefore, users compete for the free channels without taking into account the capacity bounds computed at the previous call arrival.

We next look at call blocking probabilities. As the load increases, the blocking probability of the wideband traffic is always greater than that of the narrowband traffic. Additionally, when the offered Erlang load is greater than the capacity N of the channel, the wideband blocking probability quickly approaches unity. This is due to the fact that, on the average, there is not enough capacity for a wideband channel to be admitted into the system. As such, the probability that a wideband handoff user will be dropped is usually greater than q_{WB}. This same rational may be applied to other cases where one of the traffic classes has more stringent requirements than another. Thus, even though there is no pre-set or a-priori blocking probability for the different classes, the class with the least stringent QoS criterion (which is a function of the traffic parameters, bandwidth requirements, and parameter q) will essentially hog all the resources. Algorithms such as those proposed in Chapter 2 (see [15] also) could be implemented in conjunction with MMOSPRED to ensure that a minimal bandwidth is available for traffic of a given class. Alternatively, this may be addressed with the call blocking condition imposed for IMOSP-CS and IMOSP-RES and discussed in the following sections.

3.6.3.2 MMOSPRED as Compared to OSPRED

The MMOSPRED algorithm provides dynamic sharing of the channel among different traffic classes by computing the predicted capacity required by each class to satisfy the QoS of that class. The total demand of the system is the sum of the demands of each class. On handoff, however, users are admitted if enough bandwidth is available.

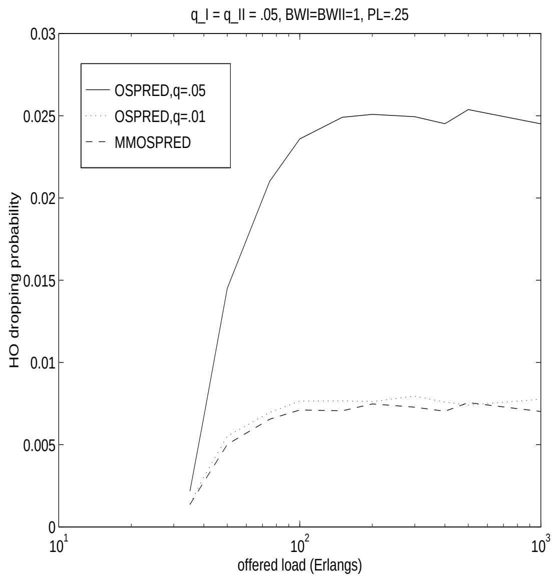
Thus, the admission control decision may be viewed as a dynamic predictive completely partitioned decision since the total requirements of each class are assumed to occupy a channel which is separate from traffic of other classes. The handoff admission procedure, on the other hand, views the channel as a single completely shared entity which allocates bandwidth on a first come first serve (FCFS) basis.

To better understand the impact or the price paid by dividing traffic into separate classes and admitting them using the MMOSPRED algorithm, we conducted the following experiment. We simulated a system where all the traffic of either class occupied 1 BBU and where $1/h = 100$ seconds, $1/\mu = 500$ seconds, and $q = .05$. The only difference between the two traffic types was a notation in the header. Class I traffic generated 75% of the load while class II traffic generated 25% of the load. As was expected, the handoff dropping and call blocking probabilities were the same for both classes.

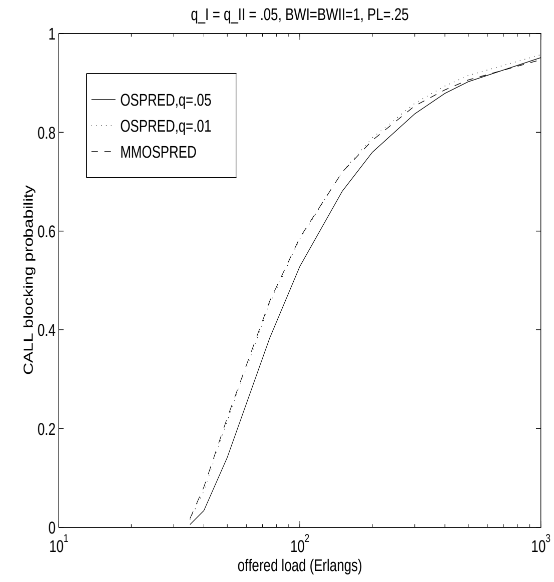
The results were then compared to two different scenarios of the OSPRED algorithm. In the first, $q = .05$ and in the second $q = .01$. Plots of handoff dropping and call blocking are given in Figure 3-11.

These results indicate that the MMOSPRED algorithm with $q = .05$ achieves roughly the same results as does the OSPRED algorithm for $q = .01$. We first note that, due to the completely shared nature of the handoff admission decision, no penalty is paid by the MMOSPRED algorithm due to the fact that all traffic is in essence statistically multiplexed onto the same channel.

However, the increased value of q in the MMOSPRED case which achieves the same QoS as in the OSPRED case indicates that the admission control decision in the



(a)



(b)

Figure 3-11: MMOSPRED and OSPRED performance comparison. For MMOSPRED, $q_I = q_{II} = .05$ and q equals either .01 or .05 for the OSPRED algorithm.

multi-traffic class case is pessimistic in predicting the performance of the algorithm. This is a direct result of the completely partitioned nature of the prediction process. The MMOSPRED prediction process computes the minimum number of channels needed to ensure that the mandated QoS is provided to the users in the system. Each traffic class completes the process independently assuming that the predicted channel demand is the total bandwidth available to that traffic class. However, all the traffic classes are "multiplexed" unto the same channel. This is evident by the completely shared or FCFS nature of the handoff admission control process. This indicates the need for a recalibration of the $q \rightarrow$QoS mapping.

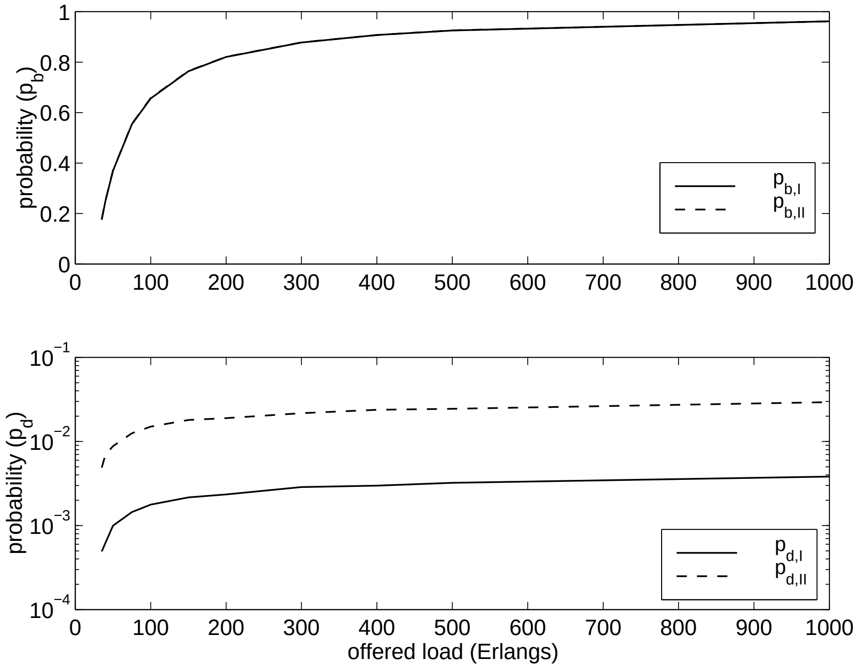
In conclusion, the shared nature of the channel in the MMOSPRED algorithm translates into a high performance multi-media algorithm which performs on the level of completely multiplexed traffic while providing independent QoS ratings as is required by each traffic type with resulting inter-dependence of the different classes because of the reasons previously previously. However, due to the completely partitioned nature of the prediction in the call admission process and the completely shared admission process, different scales are required to relate the QoS requirements to the values of q for each traffic class. We expect that these values are related both to the traffic parameters as well as to the average percentage of the offered load which is provided by each class.

3.6.3.3 IMOSP-CS and IMOSP-RES

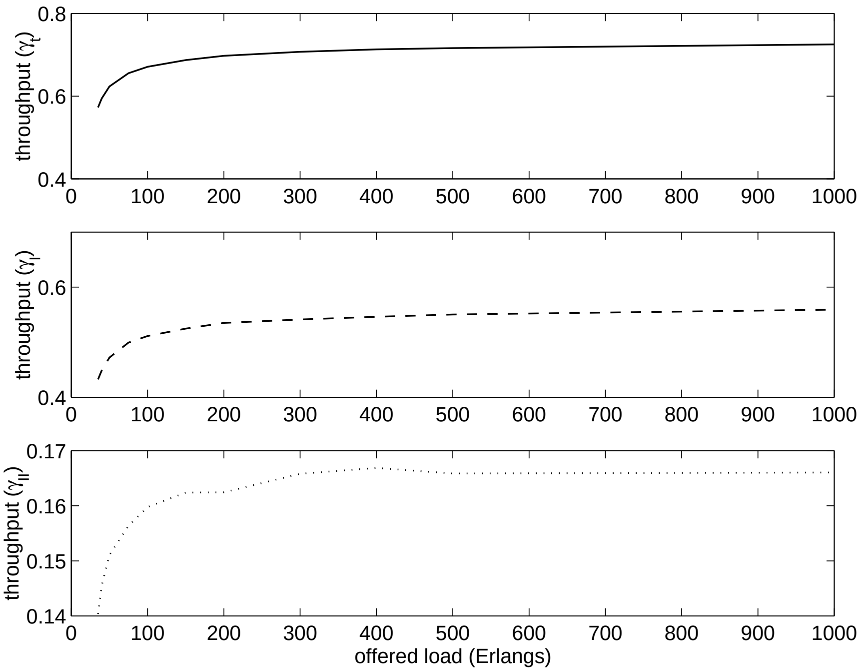
As is the case for the MMOSPRED algorithm, the handoff dropping probabilities in both IMOSP-CS and IMOSP-RES remain constant over a wide range of loads. In addition, with the use of the PB_R parameter of equations (3.19) and (3.21), we control the relationship between the call blocking probabilities. Figures 3-12 and 3-13 give one such example for each algorithm where $q_I = q_{II} = .05$ for IMOSP-CS and IMOSP-RES $q_I = .02$, $q_{II} = .0005$. Other parameters remain the same as previously. Both algorithms perform as expected, with the call blocking probabilities of the two classes being equal to each other and the handoff dropping probabilities maintained at a constant level over the range of overload. Though we note that the operating point for the systems would typically be at the far left of the plot, we primarily study the overload region for the reasons elucidated in section 3.1.

IMOSP-CS is similar to MMOSPRED in that no matter how the individual q_c parameters are assigned, there is a gap between the handoff dropping probabilities such that the wideband handoff dropping probabilities always greater than the narrowband handoff dropping probabilities. Even though the one-step predictions are made independently for each traffic class, handoff users are admitted if sufficient bandwidth is available. In direct contrast, the IMOSP-RES algorithm uses the same one-step prediction algorithm. However, since this is used to set pre-reservation partitions which are used in admitting handoff users into the system, any set of maximum handoff dropping probabilities may be achieved for the different traffic classes. This is illustrated in Figure 3-13 where the maximum handoff dropping probabilities of the two classes are approximately equal to each other. We chose these values to illustrate the flexibility within the algorithm.

Figures 3-14 and 3-15 show the impact of varying the q_I parameter of the class I traffic while maintaining q_{II} at .05 for IMOSP-CS and q_{II} at .0005 for IMOSP-RES. We note that in Figure 3-14, the handoff dropping probabilities of both traffic classes varied with the variation of q_I even though q_{II} was held fixed. Therefore, in satisfying

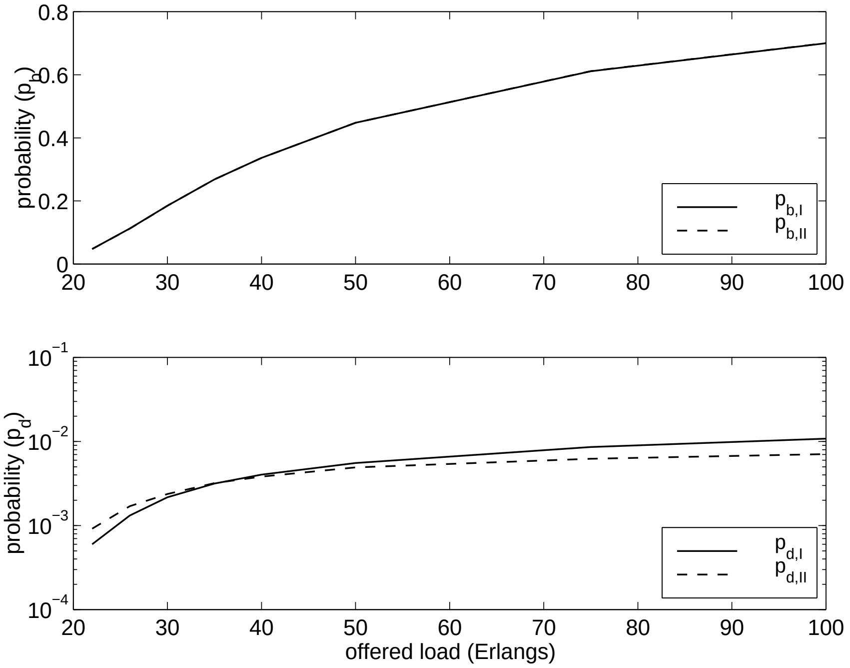


(a)

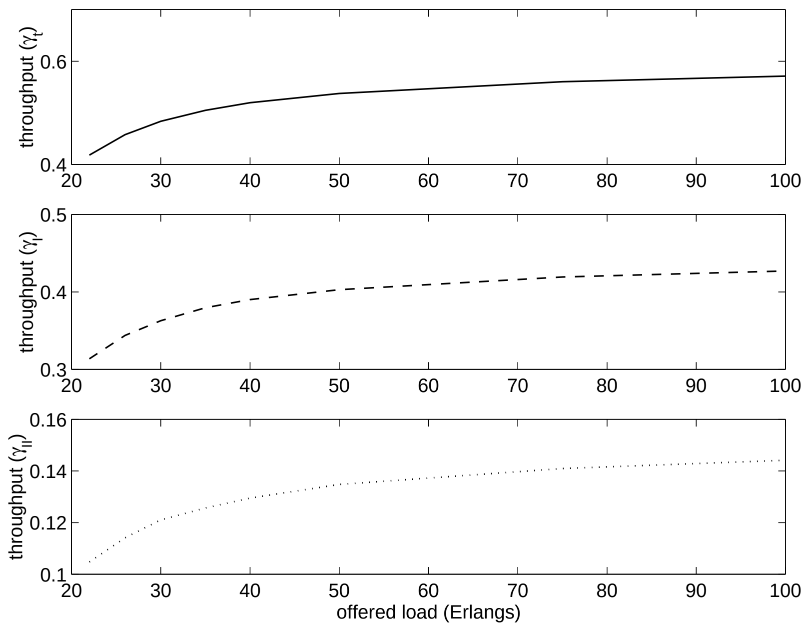


(b)

Figure 3-12: IMOSP-CS, $q_I = q_{II} = .05$. (a) blocking and dropping probabilities and (b) throughput as a function of offered load.

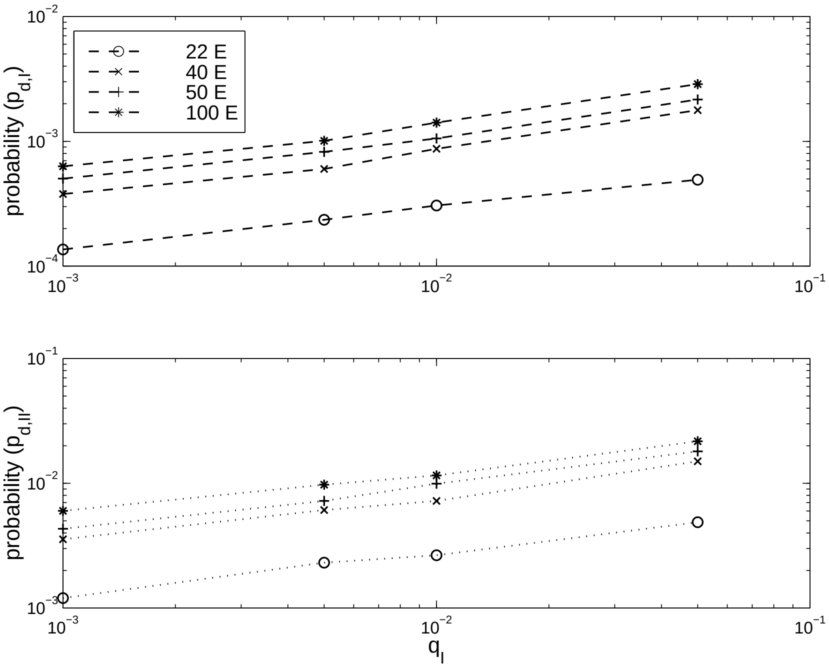


(a)

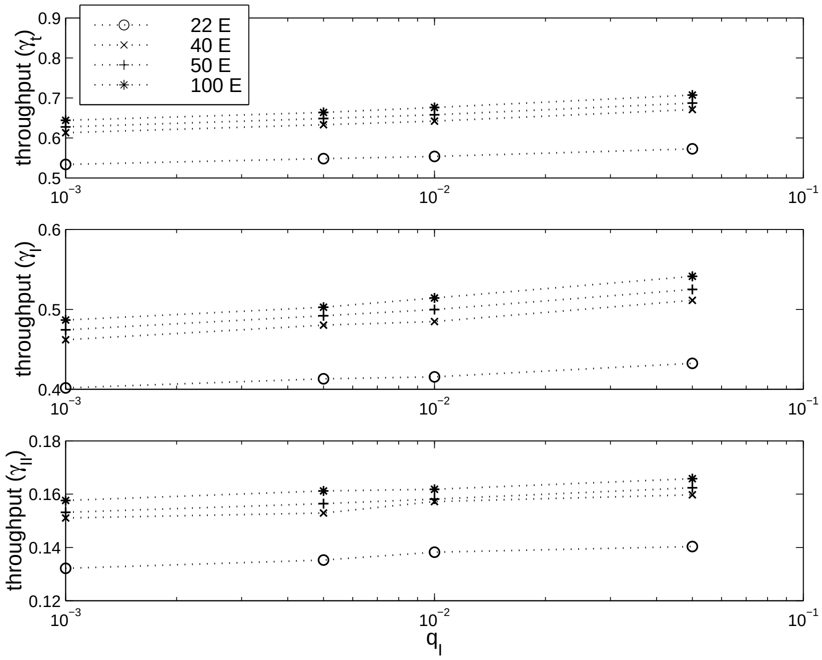


(b)

Figure 3-13: IMOSP-RES, $q_I = .02$ and $q_{II} = .0005$. (a) blocking and dropping probabilities and (b) throughput as a function of offered load.

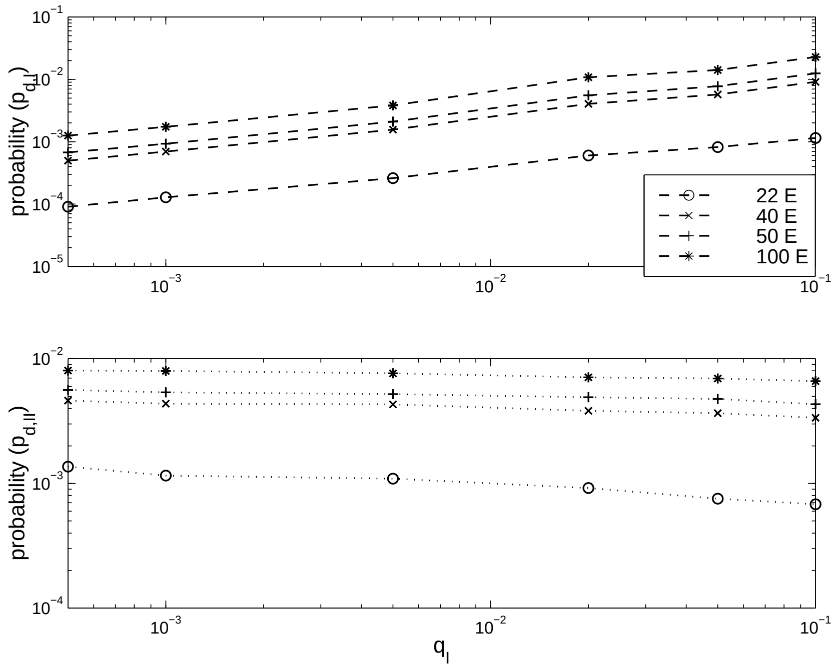


(a)

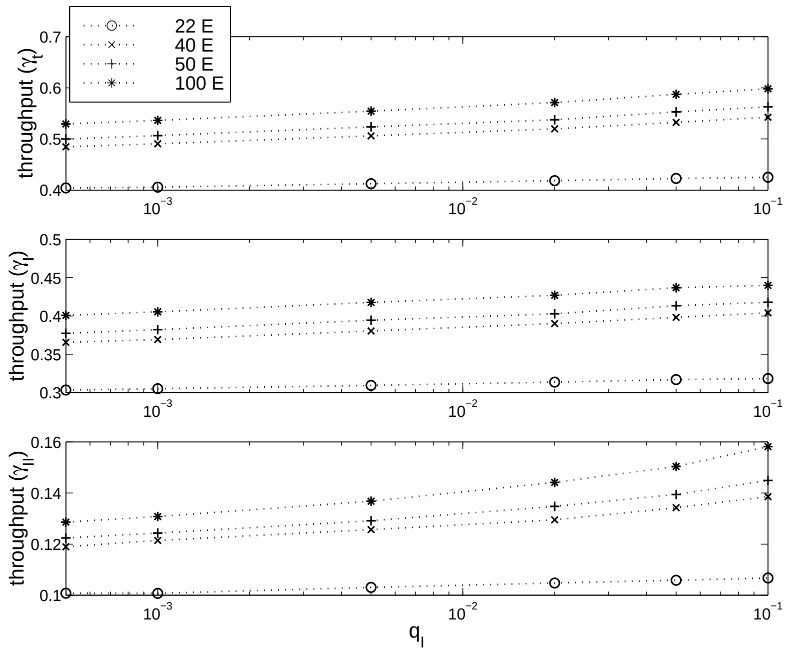


(b)

Figure 3-14: IMOSP-CS: QoS parameter $.001 \le q_I \le .05$, $q_{II} = .05$.



(a)

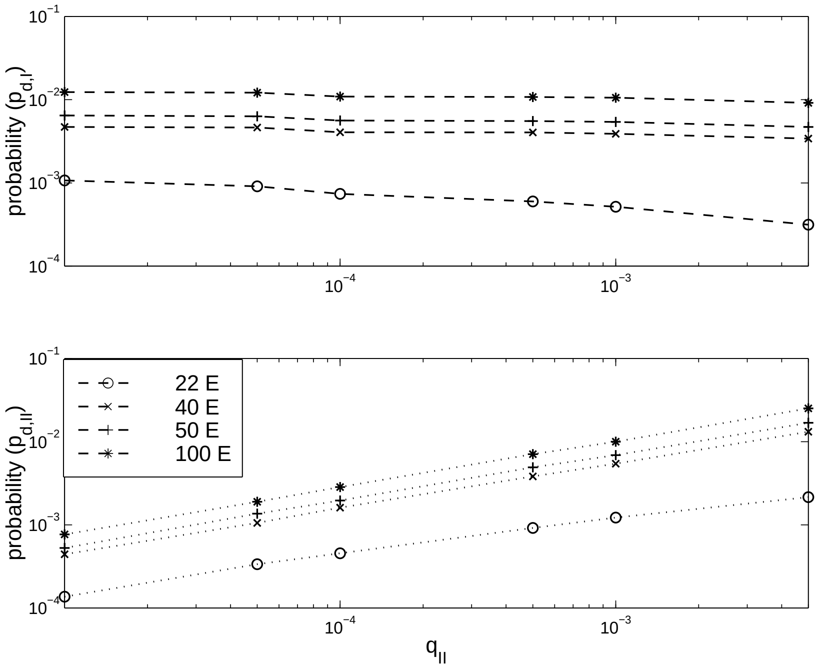


(b)

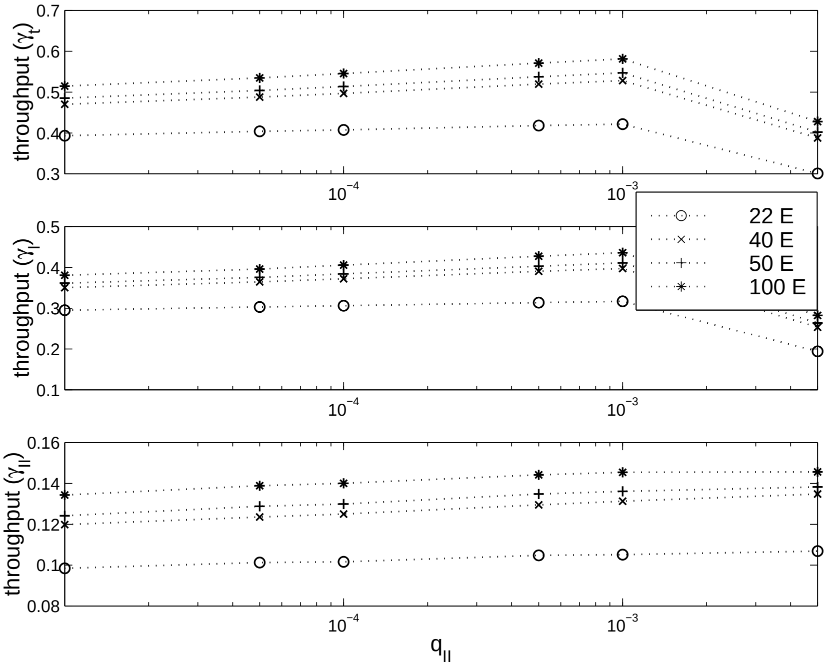
Figure 3-15: IMOSP-RES: QoS parameter $.0005 \le q_I \le .10$, $q_{II} = .0005$.

the QoS constraints for MMOSPRED and IMOSP-CS, we do the following. After choosing a maximum dropping probability for both traffic classes, we choose a q_I, q_{II} pair which satisfies both requirements. Typically, one requirement will be met exactly while the over will be overshot. In IMOSP-RES, on the other hand, while the class I handoff dropping probabilities varied with variation of q_I, the class II handoff dropping probabilities remained fixed as did q_{II}. The same qualitative properties are true for variation of q_{II} and are given by Figure 3-16. However, though the mappings of QoS parameters q_c to handoff dropping probabilities $p_{d,c}$ are one-to-one, they are not linear. Thus, simultaneous variation of q_I and q_{II} does not result in a linear relationship between q_I, q_{II} and $p_{d,I}$, $p_{d,II}$ as is shown in Figure 3-17. These trends corroborate the results discussed above. In sum, variation of the q parameters results in a monotonic one-to-one variation of the handoff dropping probabilities.

The achievement of independent QoS requirements while using IMOSP-RES comes at a cost to the system. This is easily noticed upon examination of Figure 3-18. We see here that both the class I and class II dropping probabilities of IMOSP-RES are greater than IMOSP-CS, and the blocking probabilities of the respective algorithms almost identical, while the throughput of each traffic class as well as the total throughput is greater for IMOSP-CS. This is a result of the multiplexing gain the system benefits from as a result of the complete sharing of the incoming handoff calls of all classes. As noted previously, small variations in handoff dropping probabilities come at a cost to the system. This effect is magnified in IMOSP-RES. We note, though, that we expect that there are cases where greater system throughput would be achieved using IMOSP-RES in cases where one of the QoS handoff dropping requirements is very stringent and the other very lax. However,



(a)



(b)

Figure 3-16: IMOSP-RES: QoS parameter $q_I = .02$, $10^{-5} \le q_{II} \le .005$.

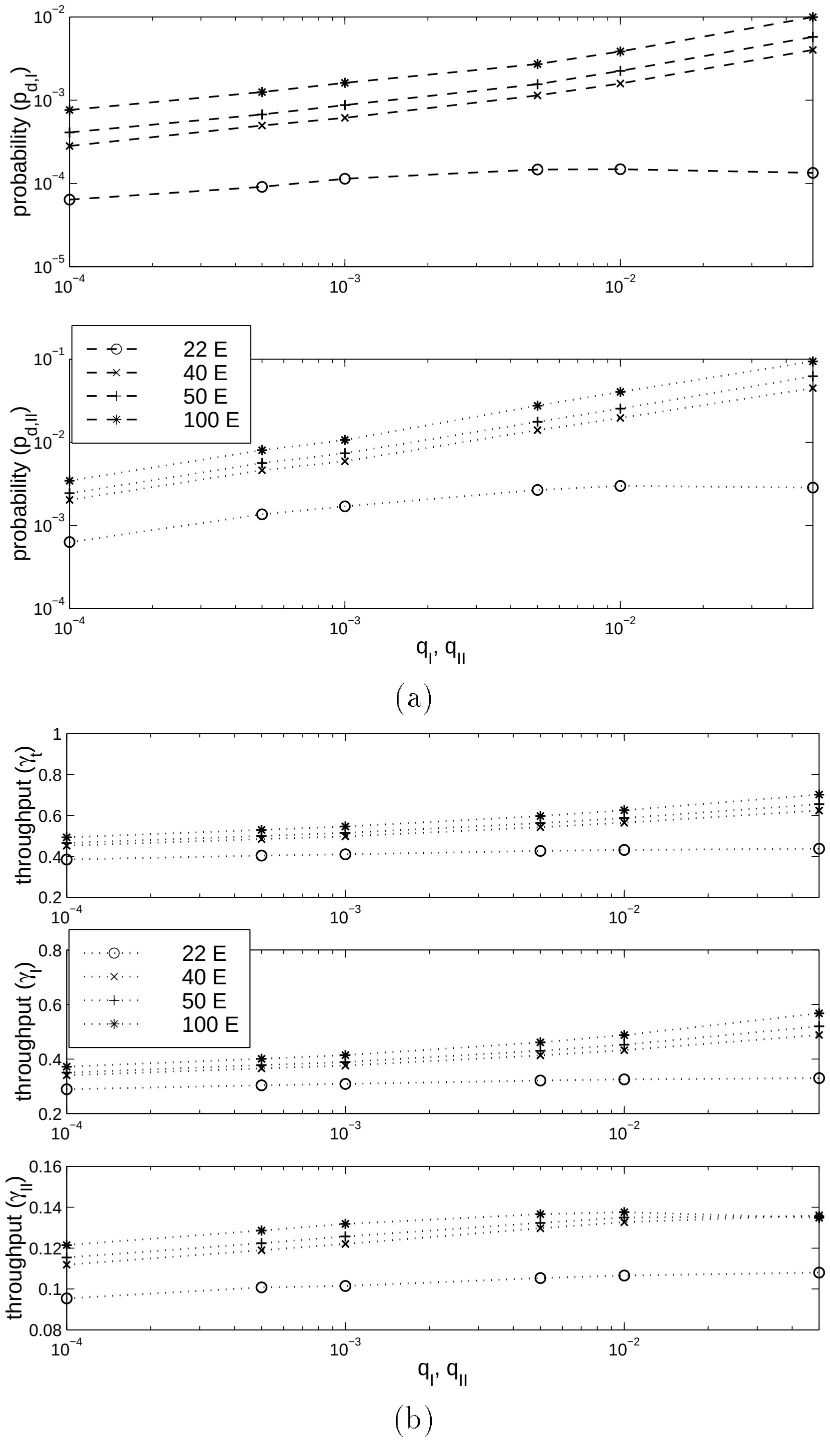
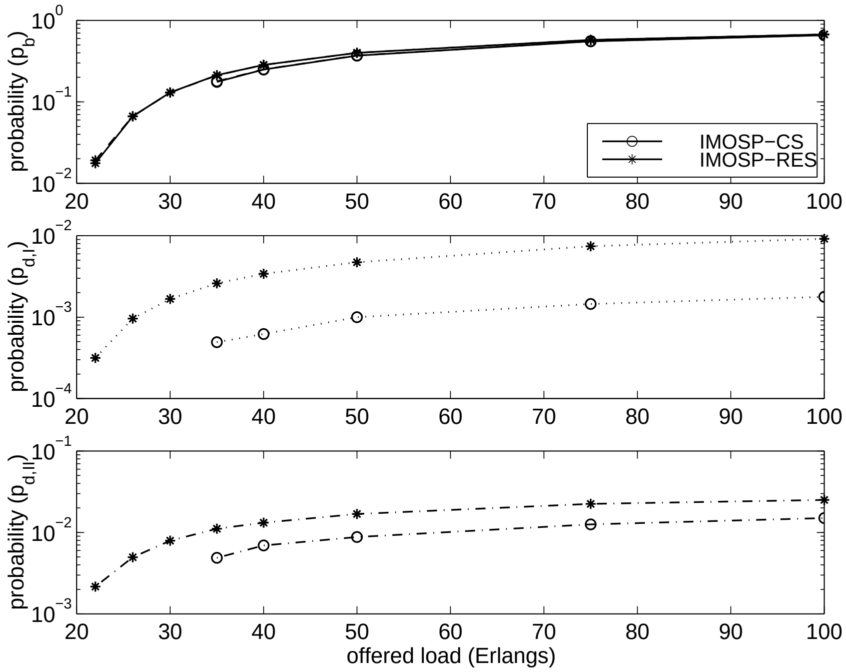
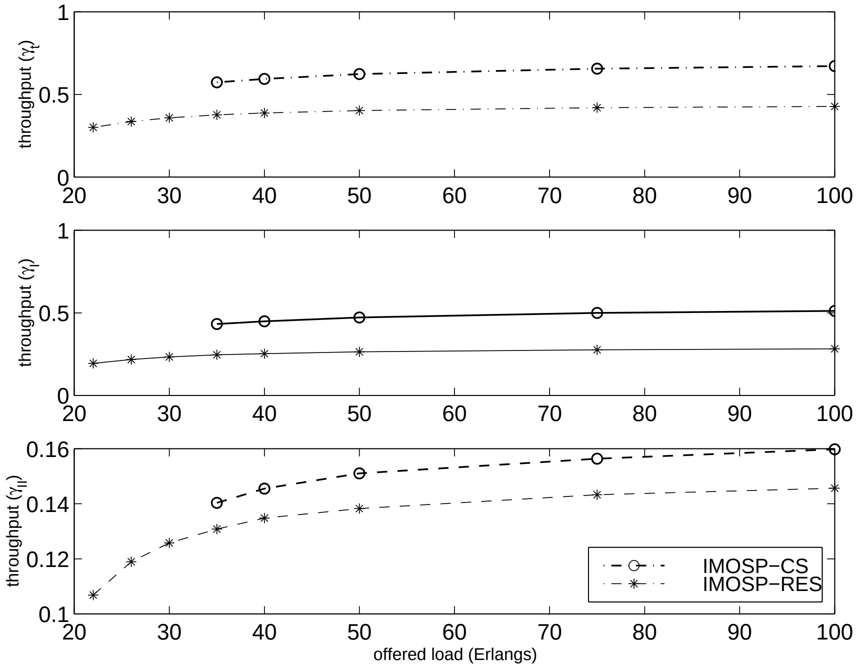


Figure 3-17: IMOSP-RES: QoS parameter $10^{-4} \leq q_I, q_{II} \leq .05$



(a)

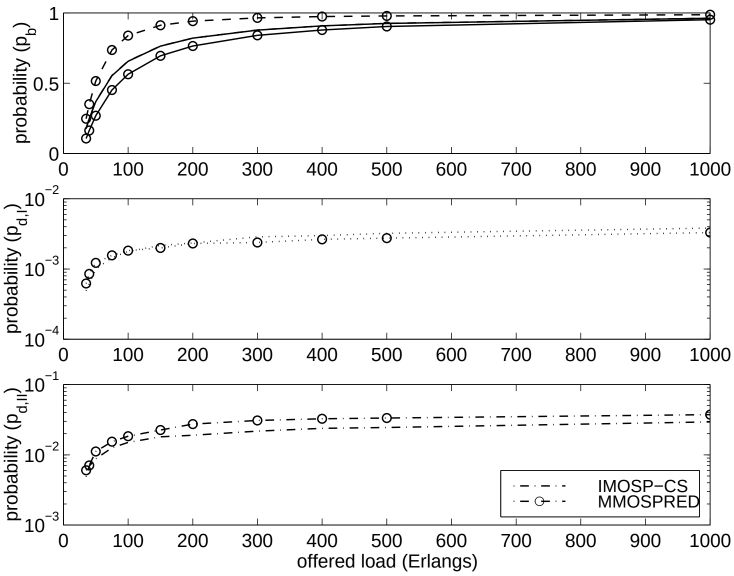


(b)

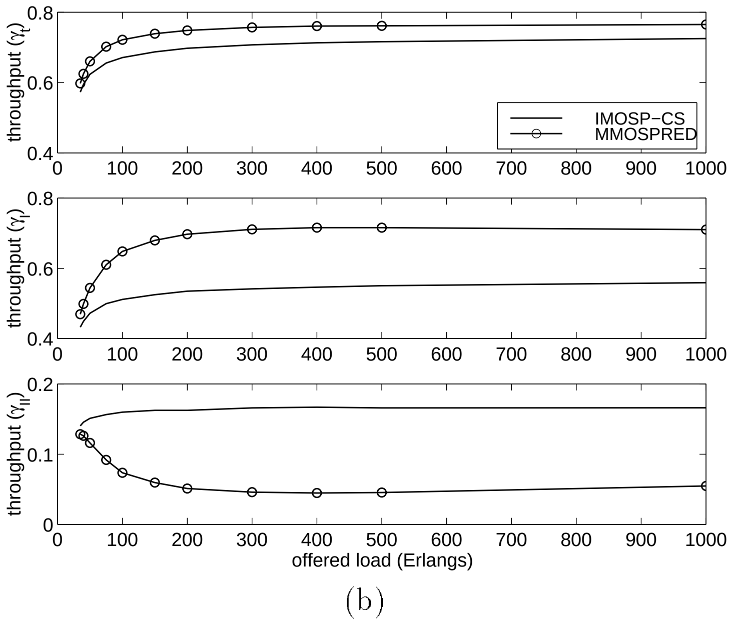
Figure 3-18: Comparison between IMOSP-CS ($q_I = q_{II} = .05$) and IMOSP-RES ($q_I = .02$, $q_{II} = .005$) algorithms. (a) blocking and handoff dropping probabilities, (b) average system throughput in most cases, we expect greater throughput using IMOSP-CS. Hence, IMOSP-CS achieves higher throughput and performance for given constraints at the cost of dependent results for handoff dropping. Thus, the imposition of the independence requirement for handoff dropping probabilities on the system in IMOSP-RES may result in significant throughput loss.

This behavior is similar to that experienced with the imposition of the independence of blocking probability of the different traffic classes. Examination of Figure 3-19 which compares MMOSPRED to IMOSP-CS which differ only in the application of this criterion, bears this out. For MMOSPRED, the wideband traffic throughput decreases as the system goes into overload, while with IMOSP-CS, wideband throughput is independent of traffic load. However, the total throughput of MMOSPRED is greater than that achieved with IMOSP-CS.

Variation of the prediction interval, T, indicated little change in the long-term average blocking and dropping probabilities for both IMOSP-CS and IMOSP-RES algorithms. This indicated a general insensitivity to that parameter as is the case with OSPRED and MMOSPRED. Further experiments with non-homogeneously loaded systems indicate that the choice of T is more important in these cases. We would expect similar phenomena if the length of the calls were taken into account in the prediction process. The systems likewise showed little change with the variation of the UP as is shown in Figure 3-20. These results apply, as expected, to MMOSPRED as well. While the long term average system performance is relatively insensitive to these changes except for small deviations at either extreme, the instantaneous or short term averaging shows more sensitivity to adjustment of this parameter. Additionally, systems with time varying offered cell loads require more

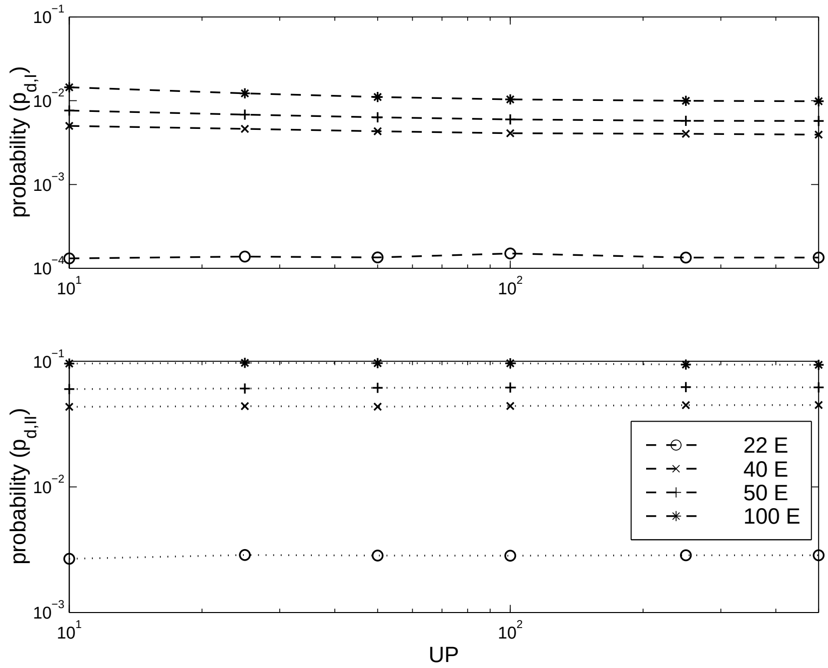


(a)

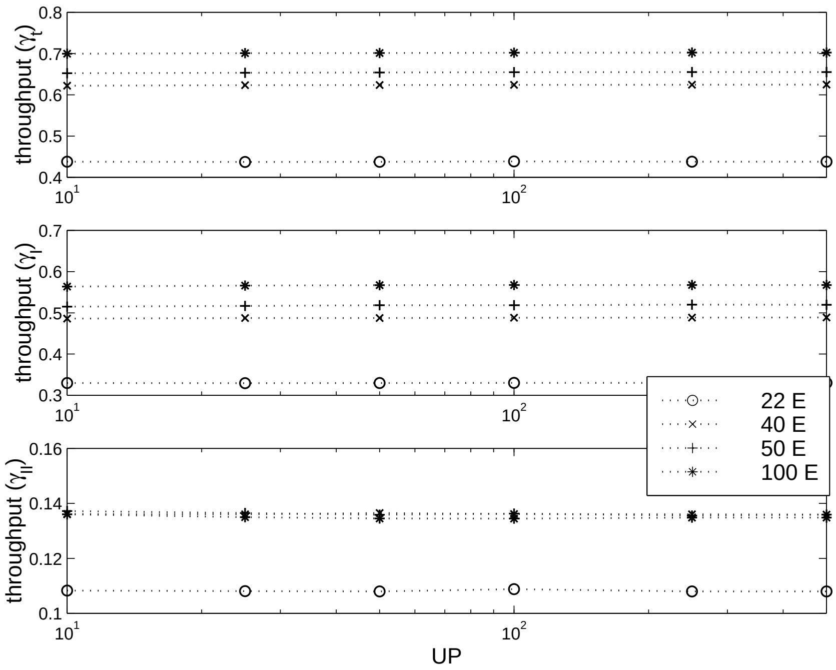


(b)

Figure 3-19: MMOSPRED and IMOSP-CS algorithm comparison, $q_I = q_{II} = .05$. (a) blocking and handoff dropping probabilities, (b) average system throughput



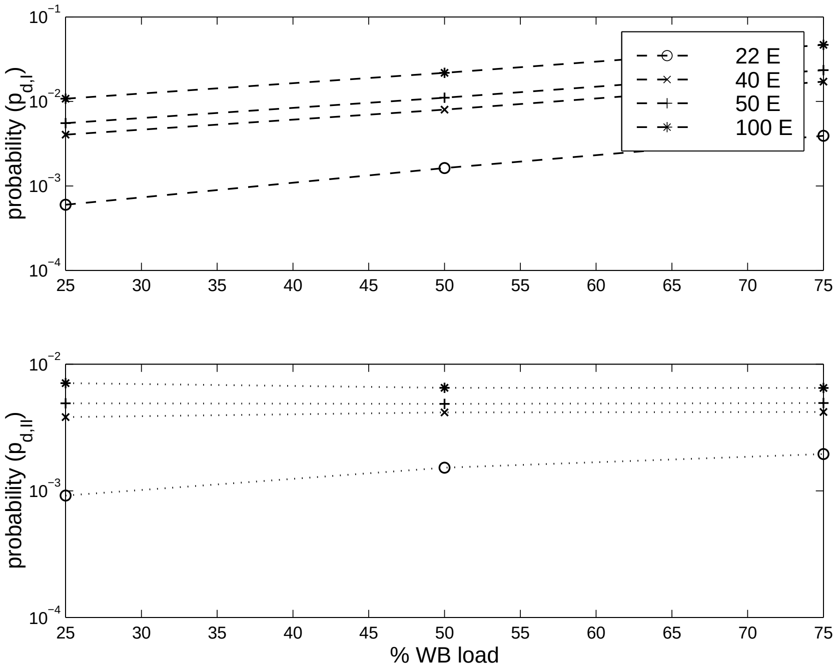
(a)



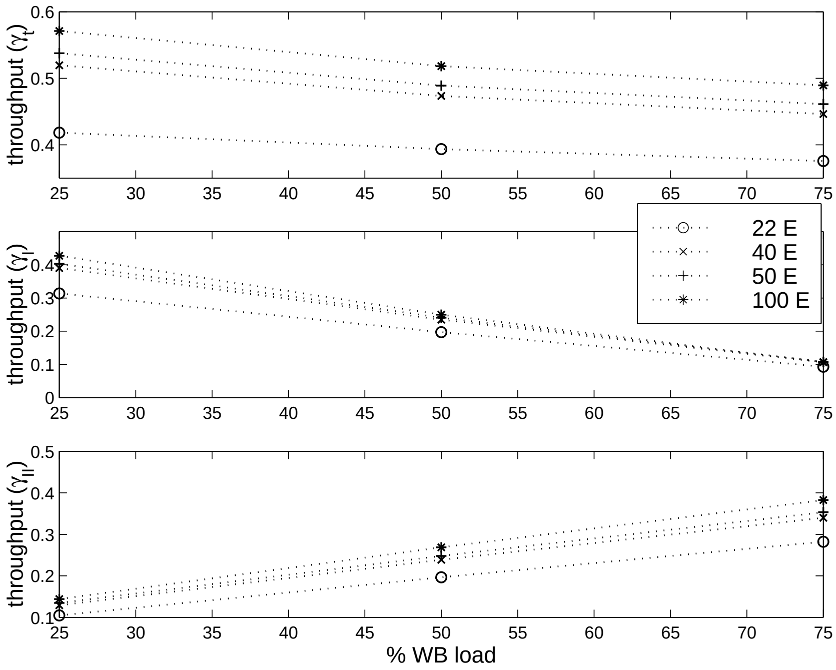
(b)

Figure 3-20: IMOSP-CS: $10 \leq UP \leq 500$, (a) handoff dropping probabilities, (b) average system throughput dynamic and flexible adjustment of the partitions. The choice of UP should be adjusted according to the system parameters such that performance is maximized at the minimum rate thereby minimizing the average number of computations performed per handoff admission.

Another set of experiments varied the ratio of narrowband to wideband traffic from 3 : 1 to 1 : 3 for the values of offered load considered above and as shown in Figure 3-21. In IMOSP-RES, the blocking probability increased a little bit with an increasing percentage of wideband traffic. The difference in required bandwidth between the narrowband and wideband traffic is significant especially given that the total bandwidth of the channel is so small and thus affects the number of users given access to the system. As a result, the total system throughput experiences a corresponding decrease, though as expected, the throughput due to wideband traffic increases while that due to narrowband traffic decreases. The overload handoff dropping probabilities for IMOSP-CS remain essentially constant though the more lightly loaded systems experience increased handoff dropping as the percentage of wideband traffic increases. This is expected since the overload dropping probabilities are at the maximum allowed by the system and therefore remain constant while the more lightly loaded systems are less affected by the dropping probability condition since they fall below the requirement. As the ratio of WB:NB traffic increases, the dropping probability increases since overload is encountered more frequently with larger bandwidth channels. With IMOSP-RES, the dropping probability results differ with the most marked difference occurring with the narrowband probability of dropping. In that case, narrowband dropping probability increases with an increasing percentage of wideband traffic even in the overload case thereby violating



(a)



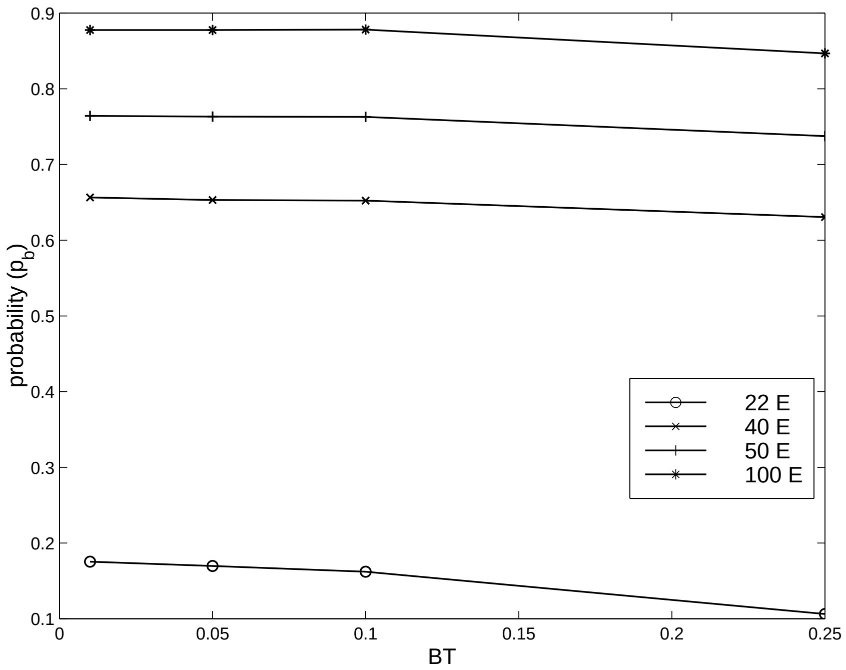
(b)

Figure 3-21: IMOSP-RES: WB:NB ratio varies from 1 : 3 to 3 : 1. (a) handoff dropping probabilities, (b) average system throughput the maximum dropping probability condition. This is directly attributed to the implementation of the partitions which while guaranteeing dropping probabilities for both traffic classes, further degrades performance.

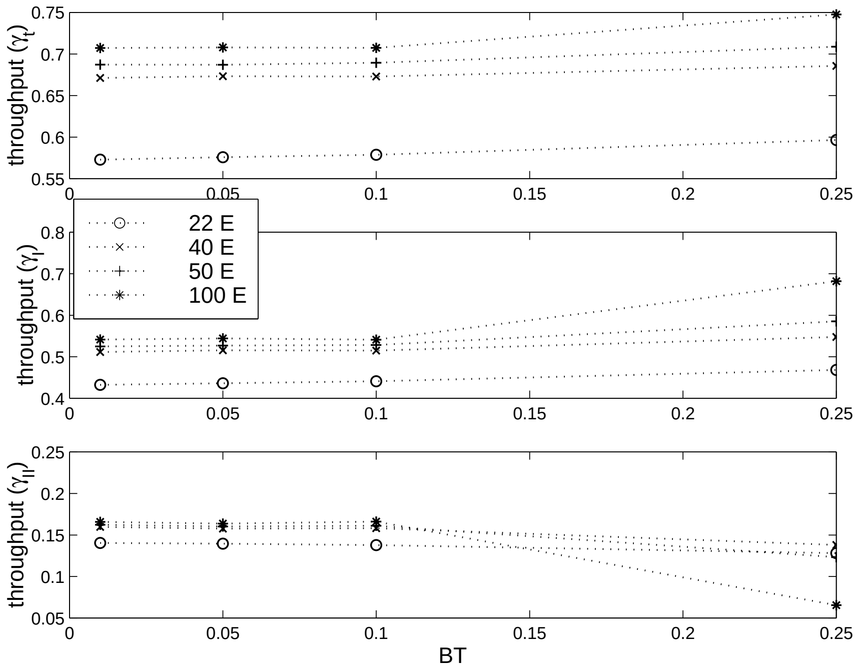
The threshold parameter BT was varied between .01 and .25 as is shown in Figure 3-22. Except for the extreme value of .25, IMOSP-CS and IMOSP-RES performance was relatively insensitive to these changes. The algorithms perform as expected and satisfy the blocking probability QoS requirement.

Experiments were performed to simulate heterogeneous traffic cases where from one to six cells in a ring containing twenty cells were overloaded forming a hotspot region while the rest of the system maintained loads at the upper end of the operation region. The system continued to perform as expected though performance was more sensitive to variation in system parameters. The effect of the increase was noticed only in the hotspot cells themselves and the cells directly adjacent to the hotspot region. Even in these cells, only small changes in dropping probabilities which conformed to the requirements occurred with throughput remaining constant over all cells.

We also varied the average handoff time with IMOSP-RES between 10 and 1000 with results given in Figure 3-23. Results for systems where the average handoff time was varied indicate that one-step prediction is inadequate when users are highly mobile (i.e. the average time until handoff is shortened) and hotspot conditions impact larger areas in the system. We expect this difference to be even more pronounced in a two-dimensional system.

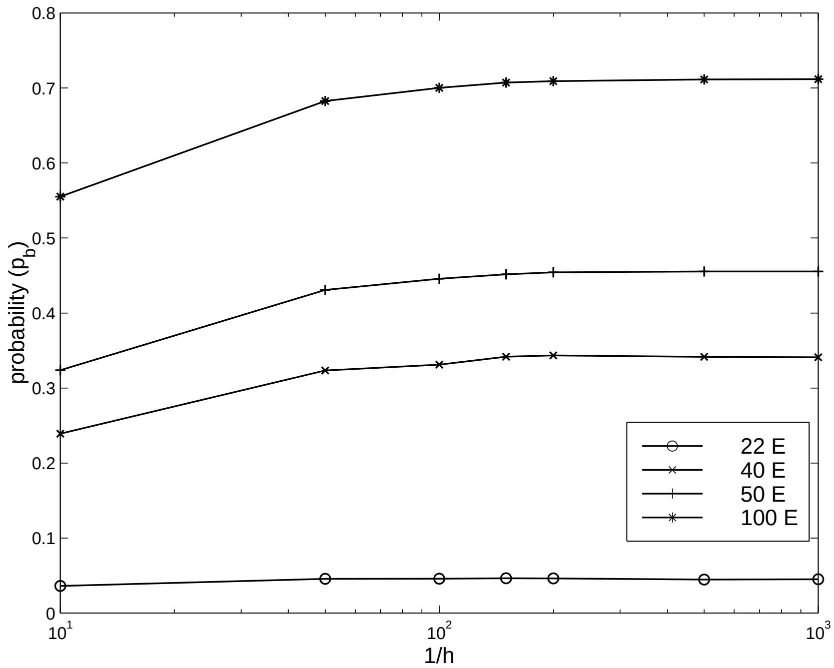


(a)

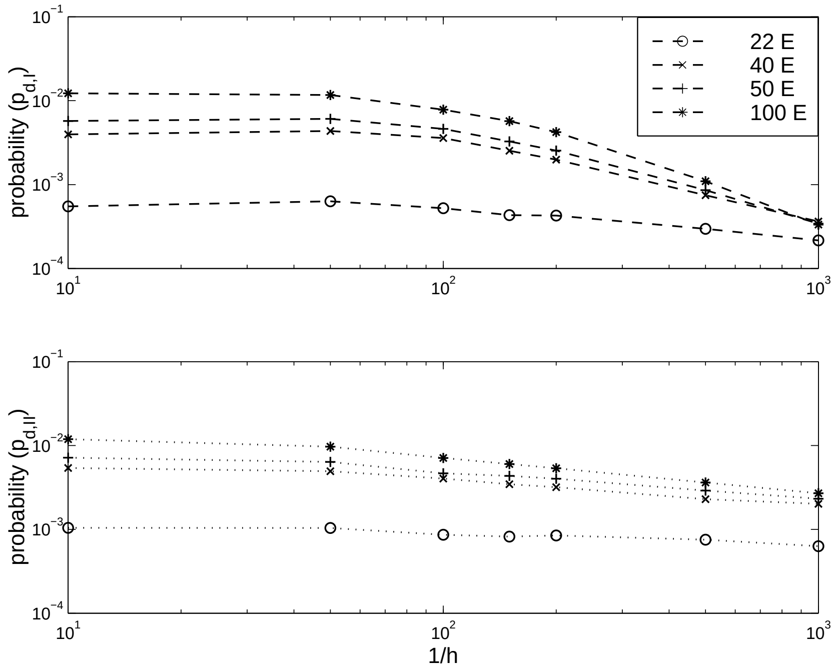


(b)

Figure 3-22: IMOSP-CS: $.01 \leq BT \leq .25$. (a) call blocking probabilities, (b) average system throughput



(a)



(b)

Figure 3-23: IMOSP-RES: $10 \leq 1/h \leq 1000$. (a) call blocking probabilities, (b) handoff dropping probabilities

3.7 Conclusion

The family of prediction algorithms presented in this chapter allocate bandwidth to incoming users in the home cell based on the cell occupancy of the home and neighbor cells. This is done by predicting one time step into the future and estimating the probability that a user in service at the present time will be dropped. The OSPRED and MMOSPRED algorithms for single and multi-class traffic use this as the only QoS condition. The IMOSP-CS and IMOSP-RES algorithms additionally require that a pre-specified call blocking probability be adhered to. The IMOSP-RES algorithm uses adaptive pre-reservation partitions for handoff traffic to admit handoff users into the cell and thus achieves class independance for handoff dropping probabilities. All algorithms aside from IMOSP-RES allocate bandwidth to handoff users by completely sharing the bandwidth among users of all classes. All of the algorithms maintain appropriate throughput for the system even at high overload while insuring that the probability of being dropped on handoff not exceed a pre-defined maximum. The OSPRED algorithm achieves better results than the NPRED algorithm in the overload region. IMOSP-CS and IMOSP-RES additionally insure that a pre-defined class admission condition is attained thus achieving the desired independent blocking probability profile.

Chapter 4

Measurement-Based Reservation for Multi-Class Mobile Admission Control

4.1 Introduction

In Chapter 3, we introduced a family of prediction-based algorithms. These algorithms did not admit new users into the system if the predicted QoS based capacity demand in any of the home and neighbor cells was greater than the bandwidth available in each of those cells. In this chapter, we develop a completely distributed reservation-based algorithm which dynamically adjusts the partitions in the home cell based on class-based statistics collected within the cell. Thus, the prediction condition was replaced by a statistics-based measurement condition.

In the following sections, we define a new multi-media dynamic reservation algorithm (MMDR) for multi-media traffic in wireless networks. Users entering a cell are admitted assuming that after admission into the system a number of guard channels (unique for each call class and call type) are still available. The number of guard channels reserved for each class and type are dynamically adjusted at regular intervals to ensure conformance to the pre-defined QoS profile which is the same as that discussed in Chapter 3.

This reservation algorithm is different than others such as [26] and the static algorithm discussed in Chapter 2 (see also [15]) in that it periodically adjusts the reservation partitions in response to changes in the offered traffic load and that the update algorithm is essentially a two-tier hierarchy. The handoff user (HO) QoS requirement is given in the form of class-specific maximum allowed dropping probabilities. The new user requirement, on the other hand, is relative in nature and involves comparing the blocking probabilities of the different classes to each other. We therefore divide the problem into two parts by separating the setting of partitions into a handoff related set of decisions and into a new user set of decisions. Assuming that we consider K traffic classes, we must consider setting up $2K$ partitions. Taken as a single tier, this is a $2K$-dimensional problem which must consider two different types of QoS criteria. By dividing the problem into handoff and new user sets of decisions, we trade the $2K$-problem for two K-dimensional problems, – each of which contains a single type of QoS criteria.

The MMDR algorithm is a completely distributed measurement-based algorithm. In contradistinction to the prediction algorithms discussed in Chapter 3 which require occupancy information from neighboring cells, the reservation partitions are adjusted based on measurements made in a given cell independent of the activity in adjacent cells. These measurements are used to update functions of the blocking and dropping probabilities in a given cell, giving most weight to the recent past and less to the more distant past. The handoff partitions are adjusted as a result of comparing the dropping probability functions to the QoS handoff dropping requirements specified at the outset. We note that this is equivalent to specifying the QoS criterion in terms of call dropping as was previously discussed in Chapter 3. The new user partitions are adjusted based on the relative comparison of the blocking probability function values using the pre-defined QoS blocking requirements. While MMDR is defined here for two traffic classes, it may be easily extended to three or more traffic classes as well. This is left for further work.

Preliminary results indicate that MMDR operates as expected and achieves the QoS requirements at all loads. As will be shown later, at low loads, all partitions are equal or very close to zero and the algorithm is essentially a complete sharing (CS) algorithm. As the load increases, the partitions increase, eventually achieving a maximum value. As indicated in [84], this behavior results in choosing the smallest partition values for each traffic class which satisfy the requirements at each load. This in turn maximizes the traffic carried by the system.

We additionally compare MMDR to the predictive algorithm IMOSP-RES of the previous chapter and discuss the strengths and weaknesses of each choice. Section 4.2 contains a discussion of the algorithm, section 4.3 a discussion of simulation parameters, section 4.4 some results, and section 4.5 some conclusions.

4.2 Admission Control Algorithm

As first described in the introduction, the multi-media dynamic reservation (MMDR) algorithm utilizes a simple distributed multi-dimensional guard channel implementation in admitting new user and handoff traffic into each cell together with a two-tier hierarchical adjustment method to periodically update the partitions. The update mechanism monitors a measured function of blocking and dropping probabilities in making these changes.

4.2.1 The Requirements

Before commencement of operation, the system defines the following requirements previously defined in Chapter 3. They are: the absolute maximum allowed hand-off (HO) dropping probabilities for each class given by the QoS parameter (q_c, $c = 1, \ldots, K$) and relative new user (NU) blocking probability ratios (PB_R_c, $c = 1, \ldots, K$) for each user class relative to class I (see [3.20]). These requirements may be alternately given as:

$$p_{d,c} \leq q_c \tag{4.1}$$

$$PB\_R_c p_{b,c} = p_{b,1} \tag{4.2}$$

where $p_{d,c}$ is the class c handoff dropping probability and $p_{b,c}$ is the class c call blocking probability. These constraints are thus the same as those invoked previously in the multi-class one-step prediction algorithms IMOSP-CS and IMOSP-RES of Chapter 3. These conditions must independently hold true for all cells in the system at all times. Monitoring of the individual probabilities and a resulting adjustment of partitions is done at update intervals in order to ensure fulfillment of the requirements using partitions of the smallest possible value. The algorithm implementation additionally requires the choice of a handoff update parameter (UP_{HO}), a new user update parameter (UP_{NU}), a blocking threshold parameter (BT), and dropping threshold parameters for each class c (T_c, $c = 1, \ldots, K$). The threshold parameters are used to measure the degree of closeness to the required profile and achievement of those requirements and are analogous to parameters given in Chapter 3.

These parameters will be discussed further in later sections. In the following implementation, we consider the two-traffic class case only. Extensions to three or more traffic classes are left for further work. Without loss of generality, we take $r_2 = PB\_R = 1$. We assume that there are a total of N "channels" or basic bandwidth units (BBUs) available in each cell. The bandwidth of class I traffic is taken as 1 BBU and the bandwidth of class II traffic is BW_{II} BBUs.

We note that the algorithm as is may be used to support three traffic classes where two of the classes require real-time service and the third traffic class best-effort service. This may be sufficient for the large majority of systems currently being planned. A typical example of this would be for voice, video, and data. The results for the two real-time traffic classes would be exactly as shown here, with the best-effort traffic using whatever bandwidth is not used by the other two classes. This class could be implemented as a non-pre-emptive priority traffic class with the real-time traffic pre-empting the best-effort traffic, assuming that it is admitted.

4.2.2 Basic Operation

The algorithm operates in each cell in the system in a completely distributed fashion with reservation partitions being assigned and updates conducted completely independently in each cell. Given K traffic classes, there are a total of $2K + 1$ partitions in each cell. The base partition in cell j, BP_j, separates the new user and hand-off offered traffic from each other. Each traffic class in cell j additionally has new user and handoff user partitions given by $NP_{c,j}$ and $HP_{c,j}$ respectively. Figure 4-1 illustrates the use of the partitions, adjusted in a completely distributed fashion,

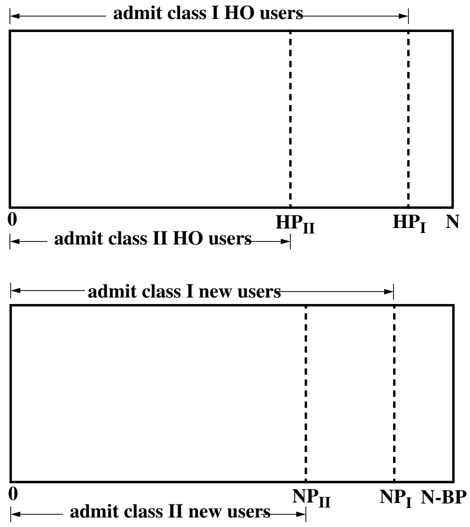


Figure 4-1: MMDR: Partitions for the two-class case. The use of the BP, HP_c, and NP_c partitions used to admit handoff and new users of either class into the system.

which are used to implement the algorithm in the two-class case. The reference to cell j is omitted in this figure for simplicity.

Given the two-traffic class case once again, we assume that there are $s_{I,j}$ class I users in cell j and $s_{II,j}$ class II users there. A class c handoff user is admitted into cell j assuming:

$$(s_{I,j} + BW_{II}s_{II,j}) + BW_c \leq N - HP_{c,j} \tag{4.3}$$

A class c new user is admitted into cell j if:

$$(s_{I,j} + BW_{II}s_{II,j}) + BW_c \leq N - (BP_j + NP_{c,j}) \tag{4.4}$$

We note that this is essentially an extension of the one-dimensional guard channel reservation algorithm first introduced by Hong and Rappaport in [26]. We demonstrate the above operation with the following example. Given a system which allocates $N = 10$ BBUs to each cell, we consider the 0^{th} cell where we assume that the total demand in BBUs including the arriving user in cell 0 is $S = 4$. We first consider new user admission. Let $BP_0 = 5$, $NP_{I,0} = 2$, and $NP_{II,0} = 0$. We look at both class I and class II admission. The class I new user controller sees $4 > 10 - (5 + 2)$ and therefore blocks a new user. The class II new user controller sees $4 < 10 - (5 + 0)$ and therefore admits a new user. We next consider the handoff admission with $HP_{I,0} = 2$ and $HP_{II,0} = 0$ and all other parameters as above. We now see that both class I and class II users are handed off successfully as $4 < 10 - 2$ and $4 < 10 - 0$ respectively.

As noted thus far, the basic operation of MMDR is static as reservation partitions are not adjusted at any point during basic operation. Threshold-driven updates, however, are used to adjust the partitions in order to ensure that the requirements are met at all times while simultaneously allowing the admission of as many users as possible. To this end, the number of new user and handoff admission requests and admissions are tracked on a class basis in each cell. When the number of requests in a given cell exceeds either the new user or handoff update parameter for a particular traffic class, the appropriate update status routine is invoked in that cell.

MMDR is an algorithm which is very easy to implement. Aside from being completely distributed and not requiring any information from any other cells, it is also computationally simple. Given that a comparison is taken as an addition, basic handoff admission requires 4 additions and 1 multiplication per admission and basic new user admission 5 additions and 1 multiplication. In addition, the update status routines require several computations as well. Since they are averaged out over all admission decisions, they add little to the average computational complexity of the algorithm. We comment more generally in sections 4.2.3.1 and 4.2.3.2.

4.2.3 Update Status Routines

We next discuss the update status procedures. In measuring the conformance of the algorithm to the absolute and relative requirements, we utilize dropping probability measurement functions (DPMF) and blocking probability measurement functions (BPMF) for each traffic class in each cell. Each function is a weighted sum of the function from the previous interval added to the appropriate probability from the current interval and is computed independently in each cell. The computation of these quantities is the same as that for BPMF in section 3.4.1.2. We repeat the basic update equations here for the convenience of the reader.

The measurement function in cell j for class c type t at update time $n_{j,c,t}$, $m_{j,c,t}(n_{j,c,t})$, is computed independently for each class, each cell, and each type (new user or handoff). It is given by:

$$m_{j,c,t}(n_{j,c,t}) = (1 - x_{c,t})m_{j,c,t}(n_{j,c,t} - 1) + x_{c,t}\frac{s_{j,c,t}(n_{j,c,t} - 1, n_{j,c,t})}{r_{c,t}} \tag{4.5}$$

where $s_{j,c,t}(n_{j,c,t} - 1, n_{j,c,t})$ is the number of class c calls of type t that were not admitted in the interval $(n_{j,c,t} - 1, n_{j,c,t})$ just concluded, $r_{c,t}$ is the number of class c type t arrivals which is given either by the new user or handoff update parameter as appropriate, and $x_{c,t}$ is equal to $r_{c,t}/M_{c,t}$.

For the handoff calls, the maximum handoff dropping probabilities are limited to the absolute class c dropping probabilities q_c, and

$$M_{c,HO} = \left\lceil \frac{10}{q_c} \right\rceil \tag{4.6}$$

Since each handoff requirement is independent of the other handoff requirements, we compute the value of $M_{c,HO}$ for each class independently. Since the blocking probabilities, on the other hand, are compared to each other, we compute a single value M_{NU}. This value is given by:

$$M_{NU} = \left\lceil \frac{10}{BT} \right\rceil \tag{4.7}$$

where BT is the blocking threshold mentioned above in section 4.2.1.

We note here that $m_{j,c,t}(n_{j,c,t})$ is not the blocking or dropping probability of class c type t during any time interval. It is, instead, a function which is related to that probability.

We parenthetically remark that during the startup phase (before the total number of events exceeds $M_{c,t}$), we do not compute $m_{j,c,t}(n_{j,c,t})$ as above since the total number of events in the interval $(0, n_{j,c,t})$ is smaller than M. Instead, the value of $m_{j,c,t}(n_{j,c,t})$ is given by:

$$m_{j,c,t}(n_{j,c,t}) = \frac{s_{j,c,t}(0, n_{j,c,t})}{h_{j,c,t}(0, n_{j,c,t})} \tag{4.8}$$

where $h_{j,c,t}(0, n_{j,c,t})$ is the number of class c type t admissions in cell j from the startup time until the end of the current interval. While the individual probabilities are not as precise as that which is required to indicate statistical significance at the desired level, equation (4.8) is an adequate approximation in the startup phase.

For the sake of simplicity, we denote the BPMFs by $b_{j,c}(n_{j,c})$ and the DPMFs by $d_{j,c}(n_{j,c})$ where

$$b_{j,c}(\cdot) = m_{j,c,NU}(n_{j,c,NU}) \tag{4.9}$$

$$d_{j,c}(\cdot) = m_{j,c,HO}(n_{j,c,HO}) \tag{4.10}$$

4.2.3.1 New User Partition Updating

The new user partition updating process is the same as that used for IMOSP-CS and IMOSP-RES in Chapter 3. We repeat it here for the ease of the reader.

When the new user update status routine is called, it indicates that there have been UP_{NU} arrivals of one of the traffic classes in that cell in the previous interval. We denote the class prompting the update as UC and the other class by OC. Without loss of generality, let the ratio of blocking probabilities $PB\_R = 1$.

After updating the appropriate BPMF, we compare the value of the adjusted BPMFs of the two classes in cell j to each other. The blocking probability profile defines the desired relationship between the blocking probability function of the traffic classes as:

$$\frac{b_{j,I}}{b_{j,II}} = PB\_R = 1 \tag{4.11}$$

If the absolute value of the difference between the adjusted BPMFs is less than BT (resolution or significance factor) ($|b_{j,I} - PB\_R b_{j,II}| < BT$), nothing is done since the difference between the adjusted BPMFs of the two classes is numerically equivalent as defined by the significance level BT. If it is greater than BT, the new user reservation partitions ($NP_{j,I}$ and $NP_{j,II}$) must be adjusted. The pseudo-code in Figure 4-2 describes the adjustment of the class UC being updated and is compared to the other class OC. The reservation bounds are adaptively adjusted using increments

```
if |b_{j,UC} - b_{j,OC}| > BT
    if b_{j,UC} < b_{j,OC}
        if NP_{j,OC} > 0
            NP_{j,OC} - -
        else if BP_j + NP_{j,UC} < N
            NP_{j,UC} + +
    else            /* b_{j,UC} > b_{j,OC} */
        if # blocks of class UC > 0
            if NP_{j,UC} > 0
                NP_{j,UC} - -
            else if BP_j + NP_{j,OC} < N
                NP_{j,OC} + +
```

Figure 4-2: Pseudo-code for adaptation control of new user reservation bounds

and decrements of a single BBU for all traffic classes, independent of the bandwidth requirements of that class, to reserve more bandwidth for the class with a larger adjusted BPMF which will therefore maintain the desired call blocking probability profile. Additionally, if $b_{j,UC} - b_{j,OC} > BT$, we only increase the priority of class UC if there has been at least one blocked call of class UC in the previous interval. This condition ensures that adjustments are made to the partitions to increase the priority of class UC only when there is an indication that that increase would have potentially lowered the BPMF in the previous interval. This condition also serves to control the extent of the oscillation of partition selection. Some oscillation is inevitable since it takes time for the BPMF to reflect the blocking probability level attained by a particular set of partition levels. The choice of different UP_{NU} values (as well as other algorithm parameters) directly contributes to the degree of oscillation and ultimately the relative stability of the algorithm. Additionally, the more the values oscillate, the greater the impact on the throughput and ultimately the performance of the system. In addition, the reservations are never allowed to exceed the number of BBUs allocated to the cell as this is meaningless and will lead to the entering of a degenerate state where the algorithm breaks down.

Just as noted previously in Chapter 3, computational complexity of the new user partition updating is very low. The BPMF update requires 2 additions and 2 multiplications. The new user reservation partition NP_j adjustment requires between 1 multiplication and 4 additions and 1 multiplication and 7 additions. Figures for the base and handoff partition updating discussed in the next section require the same order of magnitude of calculations. When amortized over the number of admissions between updates, this process adds negligibly to the average complexity per admission decision.

Like the BPMF adjustment discussed in Chapter 3 with regard to IMOSP-CS and IMOSP-RES, the adjustment mechanism discussed here for new user partition updating is simple in nature. We assume that study of this mechanism might indicate a more sophisticated method of adjustment rate which would lead to improved performance in areas relating to faster convergence and better instantaneous call blocking results. We assume that this would apply as well to the base and handoff partitioning discussed next.

4.2.3.2 Base and Handoff Partition Updating

While the use of partitions for updating new user partitions is the same as in Chapter 3 for IMOSP-CS and IMOSP-RES, MMDR differs from those algorithms in that the dropping probability constraint here is also achieved with the use of partitions.

In direct contrast to the adjustment of the new user partitions which involves the comparison of the constituent BPMFs to each other, the adjustment of each of the handoff partitions involves the comparison of each of the DPMFs to the corresponding QoS criterion q_c. To simplify the notation in this discussion, we eliminate all references to a particular cell since the algorithm is completely distributed and all partition adjustments are made on a per cell basis using information local to the cell. For the sake of simplicity, we define UC to be the class being updated and OC to be the other class.

First, the DPMF is computed as above in equations (4.5) and (4.10) for the traffic class UC that triggered the update status routine. The resultant DPMF is compared to the QoS criterion for class UC as given in the following pseudo-code and is assigned a classification value g_{UC}:

if $|d_{UC,c}(\cdot) - q_{UC}| < T_{UC}$

$g_{UC} = EQUAL$ /* *measured value is statistically indistinguishable from requirement* */

else if $d_{UC,c}(\cdot) < q_{UC}$

$g_{UC} = LESS$

else /* $d_{UC,c}(\cdot) > q_{UC}$ */

$g_{UC} = GREATER$

T_{UC} is the class UC threshold value mentioned earlier in section 4.2.1 and functions like the parameter BT does with the call blocking criterion. In other words, if the difference between DPMF and the appropriate dropping QoS criterion is smaller than the corresponding threshold value, the two quantities are considered equal. This threshold value is reflected in the conformance to the standards by a band about the maximum required value q_c. The initial requirement is therefore really equal to $q_c \pm T_c$. Otherwise, we note whether the DPMF is less than or greater than the class UC dropping QoS criterion. The DPMFs and hence g_{OC} value for the other traffic class remain as assigned when the given class last triggered the handoff update status routine.

In section 4.2.1 above, we indicated that the system requirement is that the dropping probability be less than the corresponding q value for each class. We approximate the dropping probability by the measured DPMF and apply the condition. With MMDR, we strive to minimize the partition values while satisfying this requirement, thus ensuring that as much traffic as possible will be admitted to the system. Therefore, when the DPMF is considered *LESS* than the requirement, we lower the appropriate partition values in an attempt to admit more users even though the requirement is already satisfied.

As previously mentioned, g_{UC} takes on three possible values: *EQUAL, LESS,* and *GREATER.* The adjustments made at the update status intervals attempt to adapt the base partition BP and handoff partitions HP_c affected by class UC while taking into account the DPMF values of the other class OC. The following table contains a summary of these adjustments.

g_{UC}	g_{OC}	flag	action
$=$	$<, >, =$	–	do nothing
$<$	$<, =$	–	A
	$>$	–	B
$>$	$>, =$	*true*	do nothing
		false	C
	$<$	*true*	do nothing
		false	D

The following pseudo-code describes the actions A through D in more detail.

A: if $BP_j > 0$

$BP_j --$

else if $HP_{j,OC} > 0$

$HP_{j,OC} --$

else

do nothing

B: if $HP_{j,OC} > 0$

$HP_{j,OC} --$

else

do nothing

C: if $BP_j < N$

BP_j++

else

do nothing

D: if $HP_{j,UC} > 0$

$HP_{j,UC}--$

else if $HP_{j,OC} < N$

$HP_{j,OC}++$

else

do nothing

In the case that g_{UC} is *EQUAL,* the measured performance of class UC is equal to the upper QoS limit defined at system startup. Therefore, the partitions remain at their current values. As such, the partitions need not be increased to attain the requirements. Additionally, any decrease in partition value will likely increase the measured QoS to a value greater than the set limit.

In the second case, g_{UC} is defined as *LESS.* This is an indication that the current partition levels are greater than they need to be in order to attain the defined requirements. The adjustment, however, must take into account g_{OC}. If g_{OC} is either *EQUAL* or *LESS,* we follow the following procedure. If BP_j is greater than zero, we decrease that partition by one. If it is equal to zero, we decrease $HP_{j,OC}$ by one, assuming that it is greater than zero. If they are both equal to zero, no partition adjustment is made since all values are at the minimum possible. If g_{OC} is *GREATER,* we may not decrease BP_j since that would lead to an increase in the DPMF of class OC which is already greater than the maximum desired level as is defined by the QoS requirements. We therefore reduce $HP_{j,OC}$ by one if it is greater than zero and do nothing otherwise. We parenthetically remark that this reduction will additionally result in an increase in the priority of class OC and thus serve to further improve the DPMF of that class.

We finally consider the case where g_{UC} is *GREATER.* In making this adjustment, we define the following *true/false* flag. The flag is *true* if no calls of class UC were blocked in cell j in the class UC update interval currently under consideration and is *false* otherwise. If the flag is *true*, nothing is done since the purpose of any adjustment would be an increase in the priority of class UC so that fewer handoff class UC calls would be dropped in cell j. However, the *true* flag indicates that no handoff class UC calls were dropped in the current interval in cell j. As such, an increase in any particular partition is unlikely to reduce the appropriate DPMF. The DPMF has decreased the maximum amount possible in the previous interval and needs more time to reach the desired DPMF. The following procedure is followed if the flag is *false.* If g_{OC} is either *GREATER* or *EQUAL,* BP_j is increased by one assuming that it will not exceed the number of channels available in the system (N), in which case nothing is done. Assuming that BP_j was incremented by one, we then check that both $BP_j + NP_{j,UC}$ and $BP_j + NP_{j,OC}$ are each less than or equal to N. If either value exceeds N, the appropriate $NP_{j,c}$ value is decremented by one. Finally, in the case that g_{OC} is *LESS* and the flag is *false*, $HP_{j,UC}$ is decremented by one assuming that it is greater than zero. Otherwise, $HP_{j,OC}$ is incremented by one assuming that by increasing it, the value will remain less than or equal to N.

4.3 Simulation Parameters

The simulation parameters discussed here are the same as in section 3.5. They are repeated here for the convenience of the reader.

A ring consisting of ten cells was constructed as in [58, 16]. The probability of a user handing off to any given neighbor is equally likely. Given the mobility parameters of the traffic studied, a ring of size 10 is equivalent to a line of cells. The total channel bandwidth of each cell, N, is given as 50 BBUs. We focus here on the two-traffic (narrowband (NB) and wideband (WB)) case. All narrowband users and users in single class simulations occupy 1 BBU and wideband users occupy 5 BBUs. The narrowband calls may, for example, be voice and the wideband calls low-rate video. The handoff and call times are assumed to be exponential. Both traffic classes were assumed to have the same mobility parameters. Unless otherwise indicated, the average handoff time chosen was 100 seconds and the average call holding time 500 seconds. This is assumed to model an average call in a macrocellular system.

Traffic arrivals are Poisson and are the same for all cells in the network. The offered traffic parameter in units of BBU/second is given by: $\lambda_T = BW_1\lambda_1 + BW_2\lambda_2$. Unless otherwise indicated, 75% of the traffic is due to narrowband traffic in BBU/second requiring 1 BBU and 25% of the traffic to the wideband traffic requiring 5 BBUs. In our case, this gives us: $\lambda_T = \lambda_1 + 5\lambda_2$ which is reduced to $\lambda_1 = 15\lambda_2$. The offered load which is the abcissa of many of the plots to be shown is λ_T/μ where $1/\mu$ is the average holding time.

Without loss of generality, we set the relative new user blocking probability, PB_R, to one giving equal priority to new users of both traffic classes. Likewise, the blocking probability threshold BT was set to 1%, unless indicated. Dropping probability thresholds T_c were typically set to $20 - 25\%$ of the maximum allowed handoff dropping probabilities for each class. Unless otherwise stated, $q_I = q_{II} = .005$, the dropping threshold parameters were taken as $T_I = T_{II} = .001$, and the handoff update parameter as $UP_{HO} = 250$. The q values were arbitrarily chosen for reasons of convenience.

4.4 Results and Analysis

The results in this section are divided into two parts. In the first section, we analyze the performance of MMDR and look at the choice of the algorithm parameters and discuss the impact of the variation of the various parameters on the system. In the latter part of the section, we compare the results to IMOSP-RES described in Chapter 3 and discuss the performance relative to it.

Throughput is used as a measure of performance in the system. It is defined as the percentage of the bandwidth available which, on average, is being utilized. Thus, a call admitted into the system but blocked on handoff contributes to system throughput until it is dropped. This quantity is the same as the one defined in section 3.6 and is repeated here for the convenience of the reader.

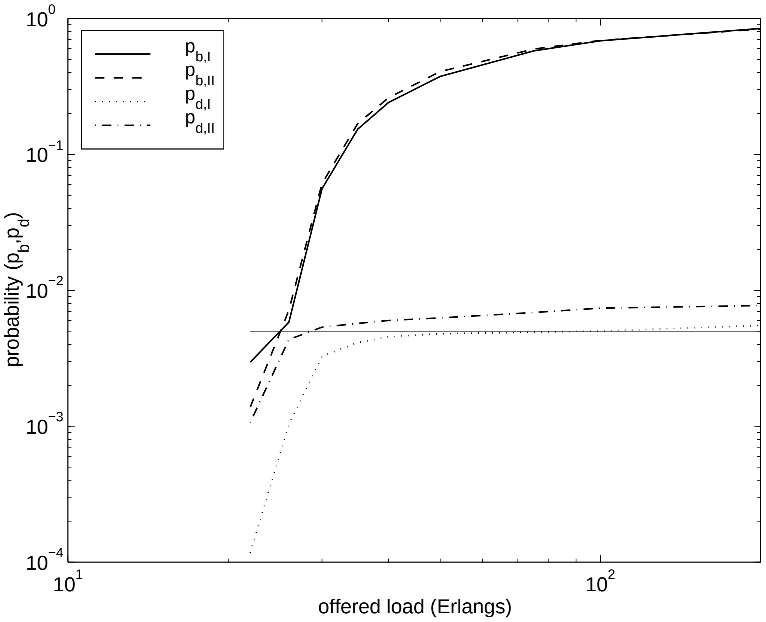
We define system throughput for class i to be:

$$\gamma_i = \frac{(A_i + H_i)BW_i/(\mu_i + h_i)}{TT \cdot N \cdot S} \tag{4.12}$$

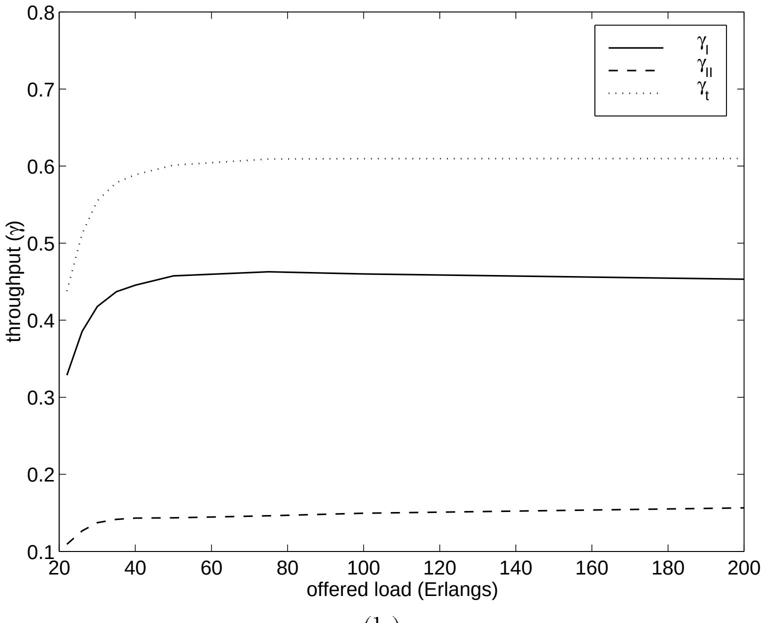
where A_i is the number of class i new users admitted into the entire system over the simulation time, H_i is the number of class i handoff users admitted into the entire system over the simulation time, BW_i is the number of BBUs occupied by class i, $1/(\mu_i + h_i)$ is the average time a user spends in a cell before either being handed off to an adjacent cell or terminating the call (where $1/\mu_i$ and $1/h_i$ are respectively the average call length and the average time until handoff), TT is the total simulation time, N is the bandwidth available to each cell, and S is the number of cells in the system. The total system throughput is the sum of the throughput due to the different classes in the system.

4.4.1 Variation of Load

A typical example of the results achieved with the MMDR algorithm is shown in Figure 4-3 for the case where $q_I = q_{II} = .005$ and other parameters as above. We immediately note that the handoff dropping probability requirements are almost exactly achieved for both traffic classes across the ranges of load shown with almost no deviation when considered in conjunction with the corresponding threshold parameters. Additionally, the new user blocking probability profile is the same for both traffic classes over all ranges of load above the 1% blocking threshold as well. As such, the required QoS profile is met for both traffic classes for new user and handoff users for all ranges of load. These properties indicate that MMDR may be used to ensure that even when the system goes into overload, the basic profile requirements are maintained. Even systems which on average operate at an operating point on the left hand side of the curves will occasionally experience overload in individual cells or across the network. Results from the overload performance region indicate that the network will not experience failure or traffic disruption in the system during



(a)



(b)

Figure 4-3: MMDR: $q_I = q_{II} = .005 \pm .001$. The solid line in (a) indicates the .005 limit set. Relative call blocking is within a 1% threshold for the two classes.

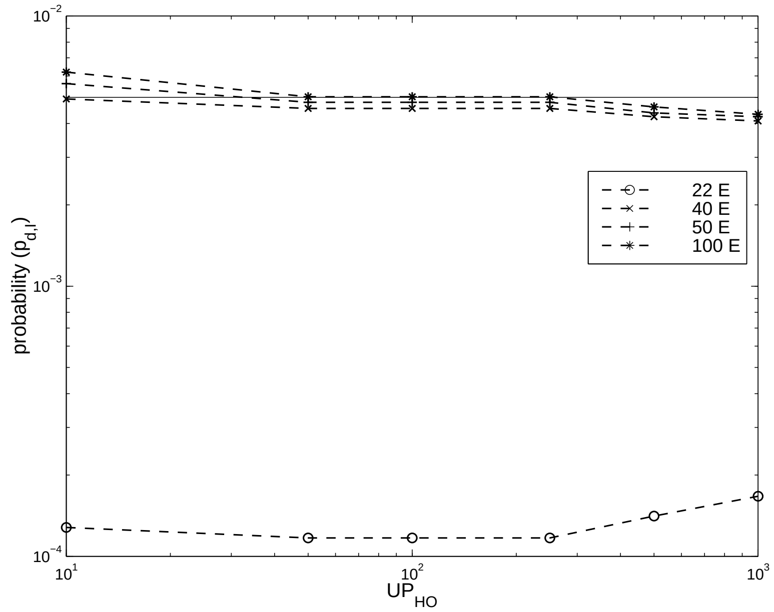
overload.

At the very left hand side of Figure 4-3(a), we note that the dropping probabilities for both classes is less than the requirement and the blocking probabilities are likewise below 1%. At these values, base partition BP and handoff partitions HP_c in each cell go to zero and operate the same way the complete sharing algorithm does. As the load increases, partitions increase automatically and level off at a saturation point tailored to the load, the traffic mix, and the requirements as they are defined by the profile. The adaptive nature of the algorithm further ensures that the partitions are as small as possible at all times. Thus, systems that experience different traffic loads at different times will likewise be accommodated. Because of the dynamic nature of the algorithm, there is no need to overprovision the system during normal operation by imposing large partition values for handoff users in order to accommodate busy period operation. This, in turn, results in more throughput and better system performance overall by raising the operating point of the system and more efficiently utilizing the resources. This adaptivity provides similar results to those considered optimal for a single traffic class in both the extreme high and low load cases [34]. System throughput, as shown in Figure 4-3(b), remains constant over all ranges of overload once a saturation point is reached. This is expected given the handoff dropping and call blocking results. The traffic share allocated to each of the traffic classes remains constant over the range of offered loads ensuring that the different traffic classes have equal access to the system. This property is a direct result of the blocking probability profile requirement.

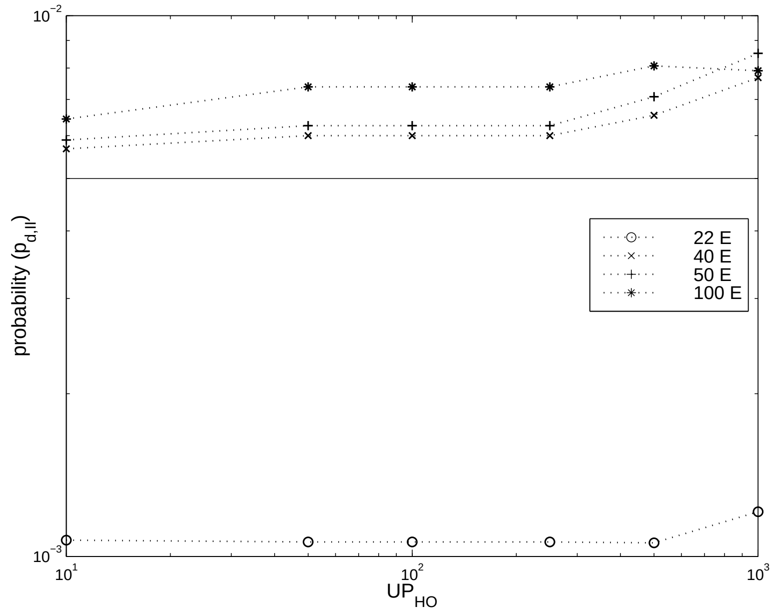
4.4.2 Variation of Handoff Update Parameter UP_{HO}

The UP_{HO} was varied between 10 and 1000 to measure the sensitivity of algorithm performance to the choice of update parameters. Results are shown in Figures 4-4 and 4-5. We noted that the algorithm performed almost identically over the range with small degradations in performance in all respects at both ends of the spectrum. At the low end, this was attributed to updates being performed too quickly and the measurement therefore being inaccurate because the DPMF function introduces some non-linearity into the system. This is especially true for the wideband traffic since on average only 25% of all load is attributed to bandwidth traffic and the $5:1$ bandwidth ratio brings that down to approximately 5% of all arrivals. When $q_{II} = .005$, the denominator of the DPMF is 2000. For $UP_{HO} = 10$, the maximum change in the DPMF is .005 which is very small.

Thus, even when only a small adjustment needs to be made to the partitions to achieve the desired levels, several adjustments will be made to those partitions before it is adequately reflected in the DPMF. This will then result in another needed adjustment in the opposite direction. As a result, there will be oscillations in the partition levels which mean that the instantaneous QoS criterion will not be met though on average it will be achieved. At the other end of the spectrum, partitions are adjusted too slowly and we begin to notice small deviations between the blocking probabilities of the different classes. Assuming that the load and/or relative traffic class composition remain constant, updates should be made at more infrequent intervals so as to minimize the computational complexity burden applied to the system. However, in systems where the load varies dynamically, the update function should be invoked more frequently to follow those changes. The scale of the

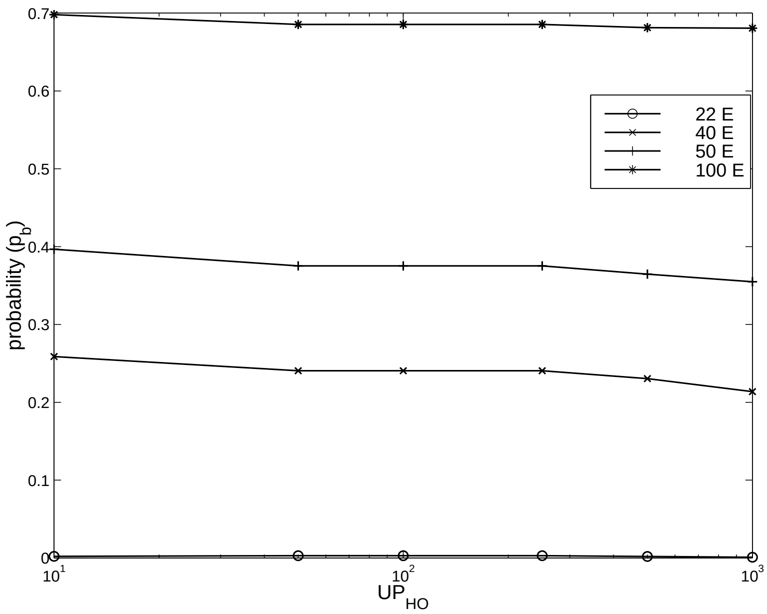


(a)

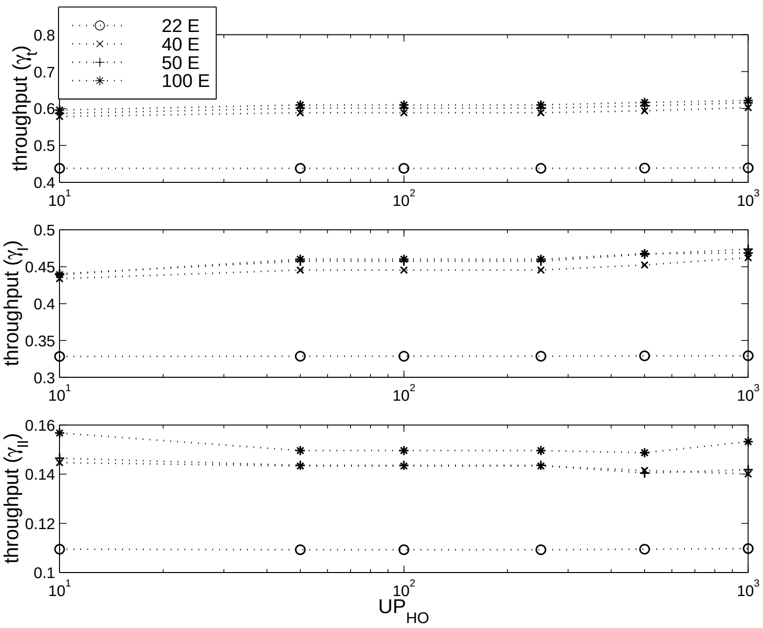


(b)

Figure 4-4: MMDR: $10 \leq UP_{HO} \leq 1000$, handoff dropping probabilities for (a) class I and (b) class II



(a)



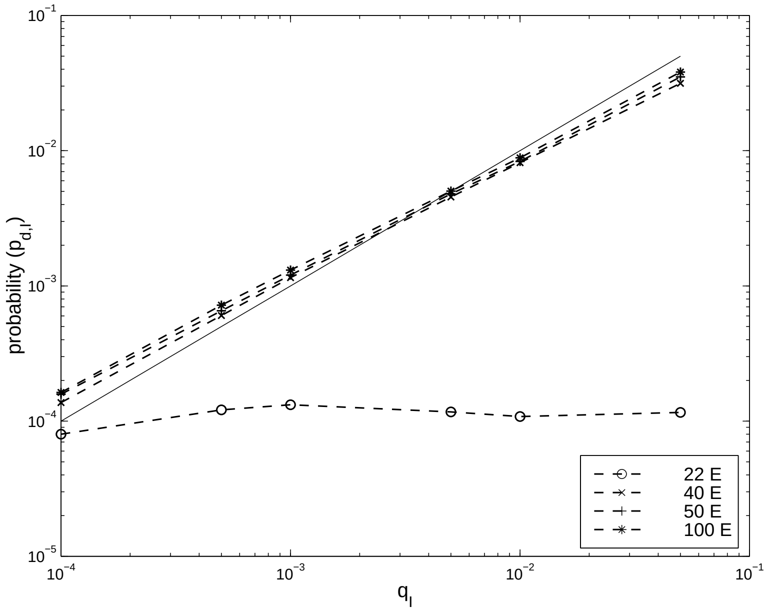
(b)

Figure 4-5: MMDR: $10 \leq UP_{HO} \leq 1000$, (a) call blocking probabilities, (b) normalized average throughput load variation and expected response time should be taken into account when setting the UP_{HO} parameters. The small deviations of the handoff dropping probabilities from the requirements are additionally attributed to the partition update mechanism which has been seen to be non-optimal for similar reasons. Although not considered in this implementation, variations in UP_{HO} on a per class basis and perhaps even dynamically might additionally produce "better" responses. Additionally, varying the step-size of each adjustment might likewise produce both faster convergence and less large scale oscillation. However, this might possibly come at the expense of exact conformance to the requirements. Dynamic variations in the UP_{HO} on a per class basis would presumably address these issues. We finally note that we expect these adjustments to have only second order effects on the system and the manifestations of these results would be subtle.

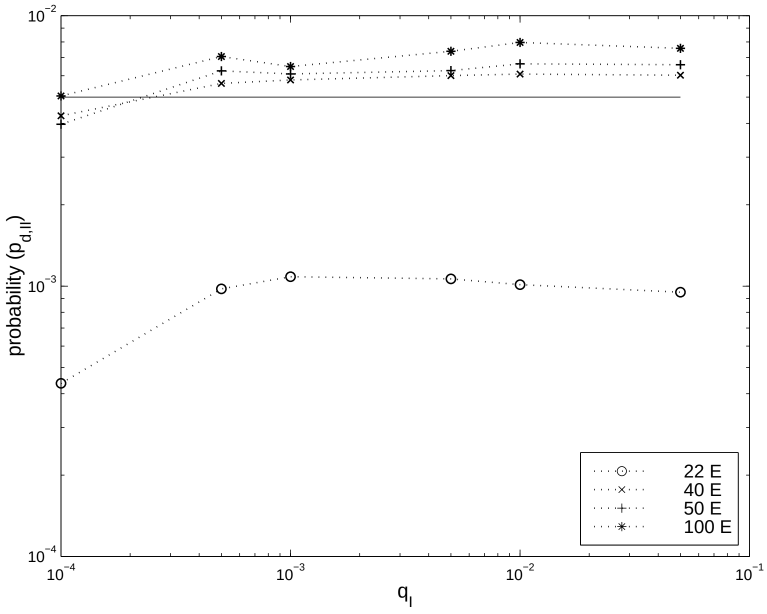
A related point considers the relative ability to measure and follow both "very small" dropping or blocking probabilities and traffic with relatively "small" arrival rates. The length of time needed to adjust to either of these scenarios is relatively long as the time needed to accrue these data points is "large." Though on average we expect that the required profiles will be achieved, the response to dynamic variations or changes in load will be retarded.

4.4.3 Variation of QoS Handoff Dropping Probability Parameters

We next looked at the variation of the class I and II handoff dropping probability parameters as a function of load. We considered three cases. In the first, shown in Figures 4-6 and 4-7, q_I was varied between 10^{-4} and .05 while q_{II} was kept constant at .005. In the second, shown in Figures 4-8 and 4-9, we varied q_{II} between 10^{-4}

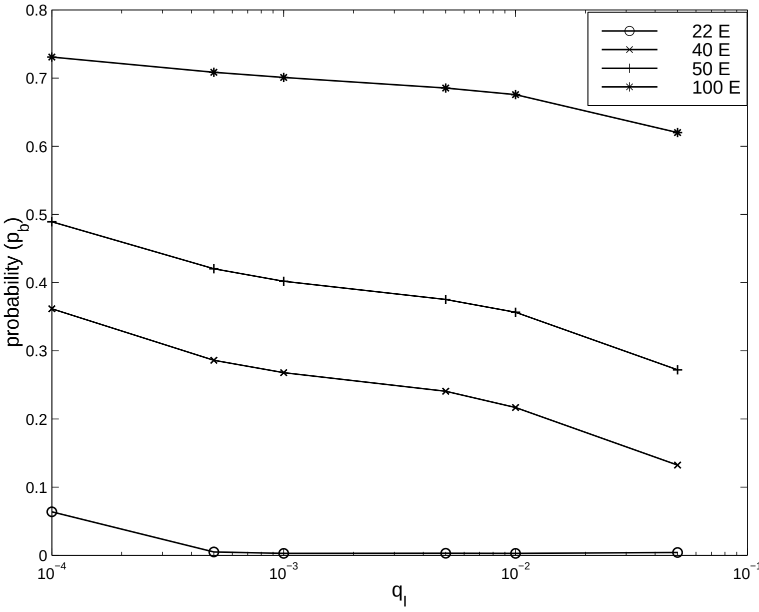


(a)

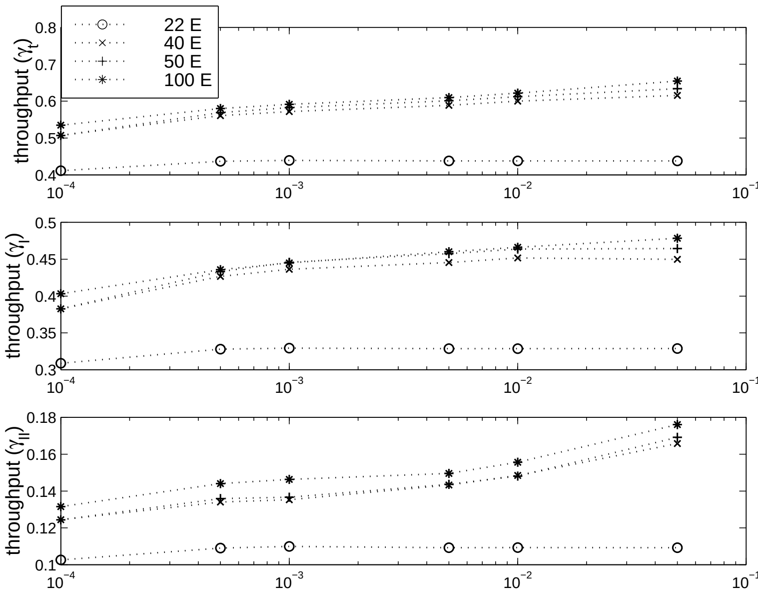


(b)

Figure 4-6: MMDR: $10^{-4} \leq q_I \leq .05$, $q_{II} = .005$. Handoff dropping probabilities for (a) class I and (b) class II

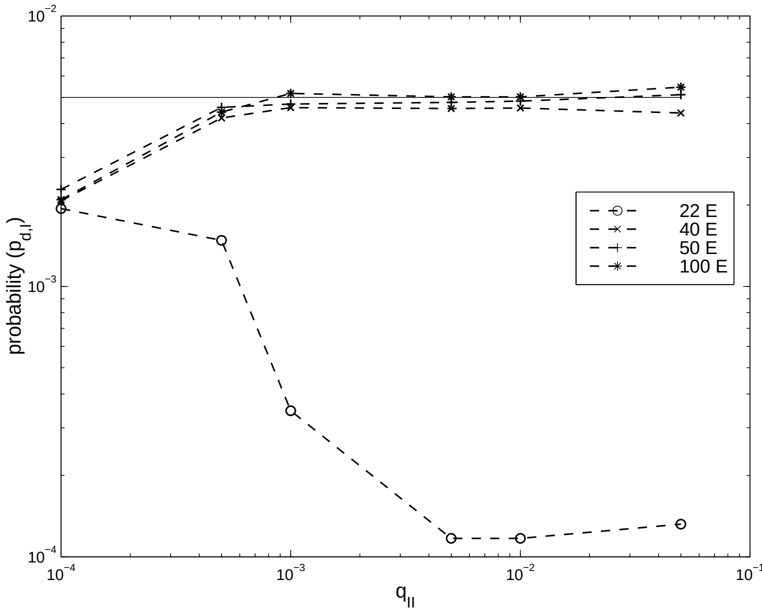


(a)

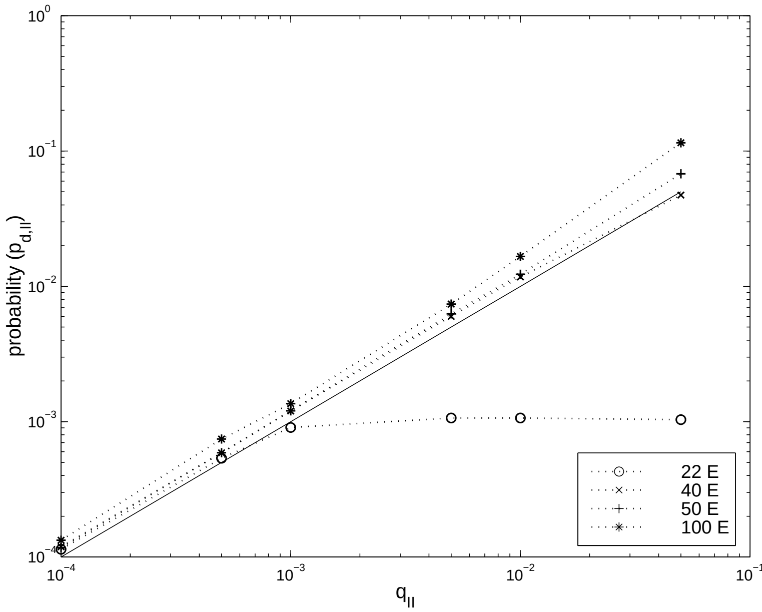


(b)

Figure 4-7: MMDR: $10^{-4} \leq q_I \leq .05$, $q_{II} = .005$ (a) call blocking probabilities, (b) normalized average throughput

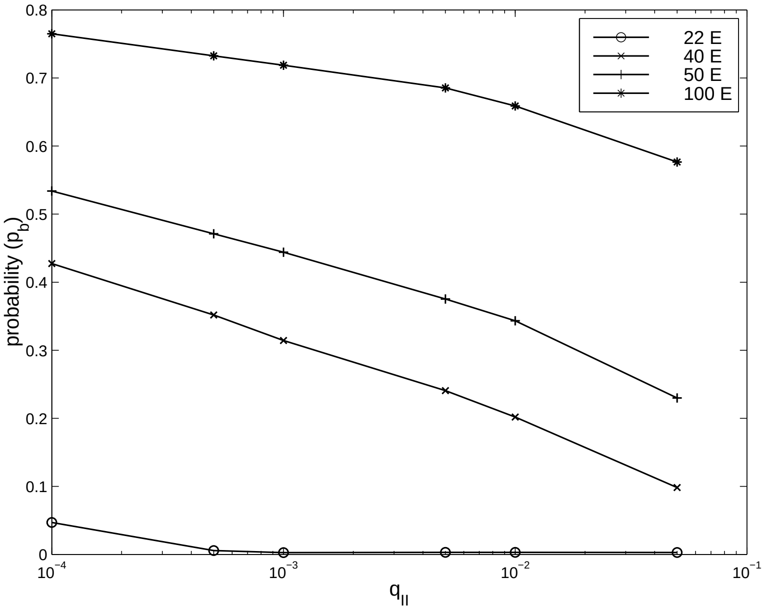


(a)

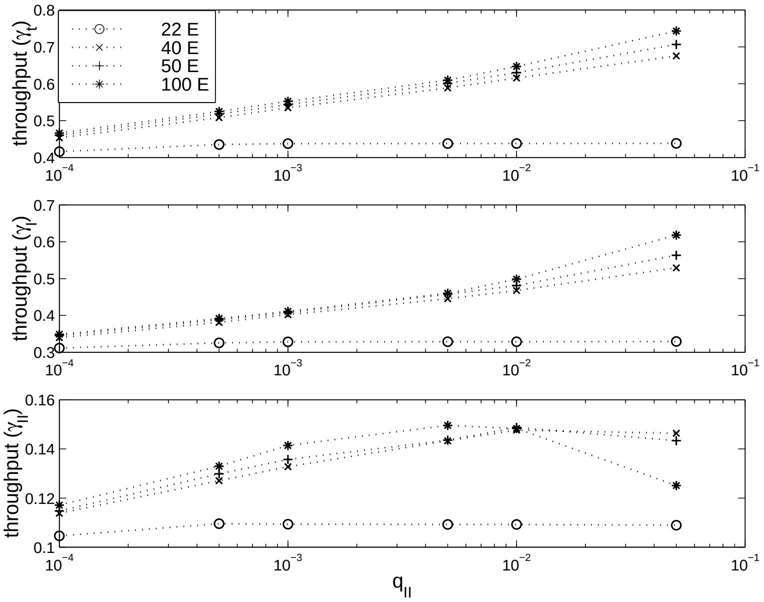


(b)

Figure 4-8: MMDR: $q_I = .005$, $10^{-4} \leq q_{II} \leq .05$. Handoff dropping probabilities for (a) class I and (b) class II



(a)

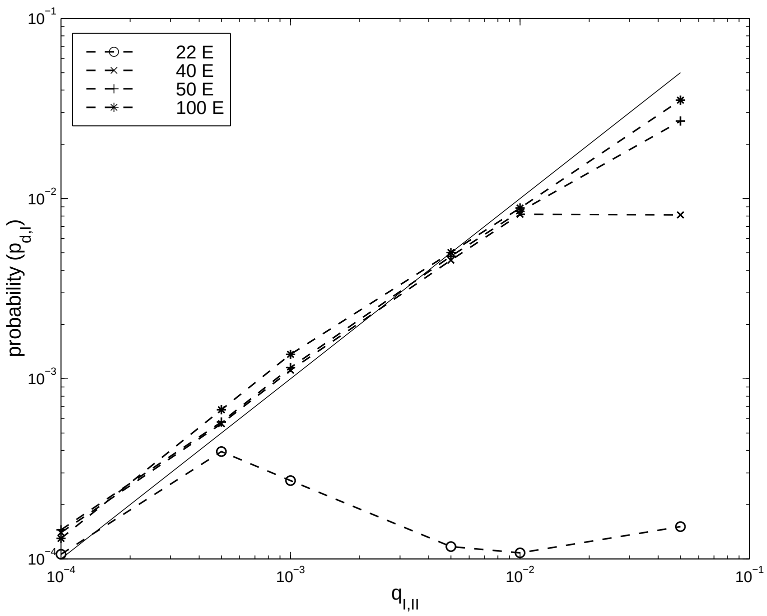


(b)

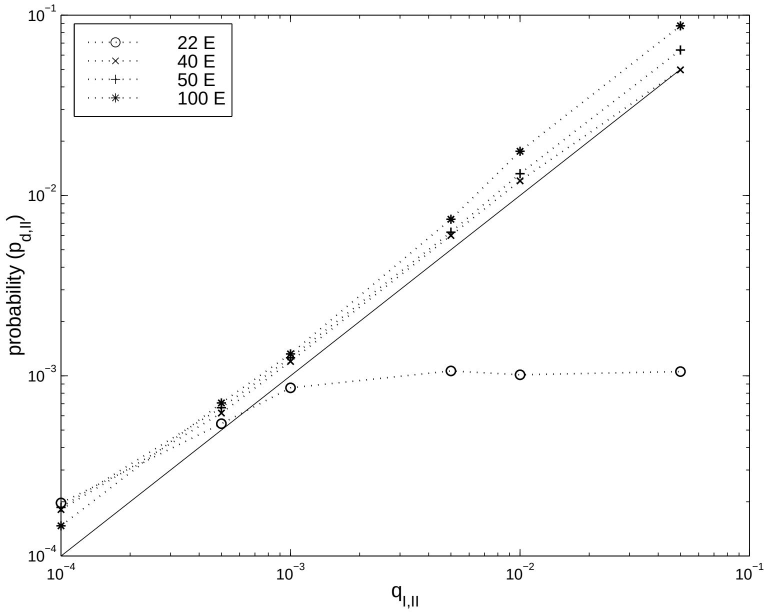
Figure 4-9: MMDR: $q_I = .005$, $10^{-4} \leq q_{II} \leq .05$ (a) call blocking probabilities, (b) normalized average throughput and .05 and kept q_I constant at .005, and in the third, we varied both q_I and q_{II} between 10^{-4} and .05 as shown in Figures 4-10 and 4-11. The solid line in all of the handoff dropping probability plots indicates the targetted q_c value. In reality, the target is a band indicated by $q_c \pm T_c$. This band is not indicated on the plot itself.

In all of the cases, we noted that both QoS criteria were attained. The absolute maximum handoff dropping probabilities were reached to within a small factor of the dropping probability band given by the appropriate q_c plus or minus the given dropping probability threshold T_c in all cases where the load necessitated the implementation of partitions so that the maximum would not be violated. Generally speaking, the upper end of the band was achieved when the q_c was smallest or strictest and the lower end of the band when q_c was largest or least strict. This result is intuitive. Additionally, the measured $p_{d,II}$ showed only small deviations from the band. This was attributed to two factors: the large bandwidth of the class II traffic relative to both the total size of the channel and the class I traffic as well as the relatively small arrival rate of the class II traffic in terms of total arrivals. This was also an intuitive result. We note that this deviation may easily be corrected for by adjusting the q_{II} value an appropriate amount before commencing operation.

At low loads, as discussed previously, the dropping probabilities are below the maximum specified dropping probabilities and an examination of the history of the partition values from simulation data indicates that the system reverts to the complete sharing (CS) algorithm under these loads, thereby achieving maximum throughput. What is a low load is a function of the parameters of the constituent traffic classes, the QoS requirements q_c, and the parameters of the system. In

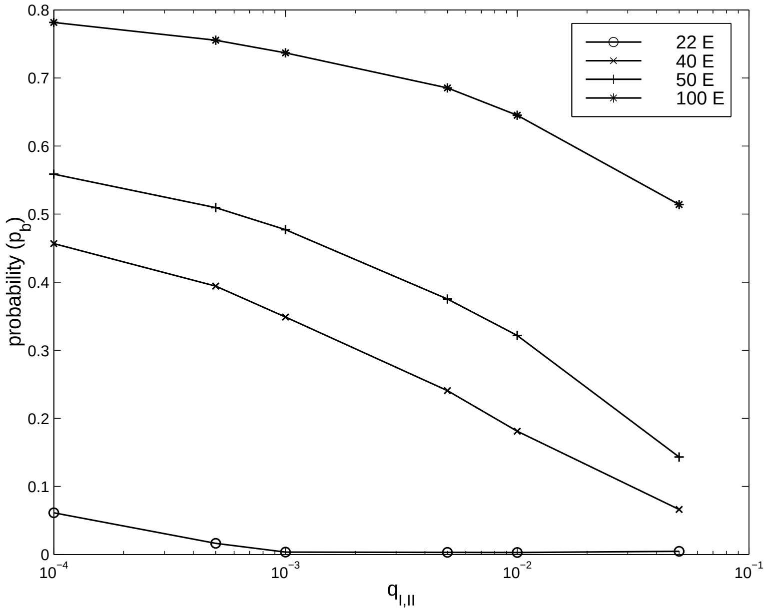


(a)

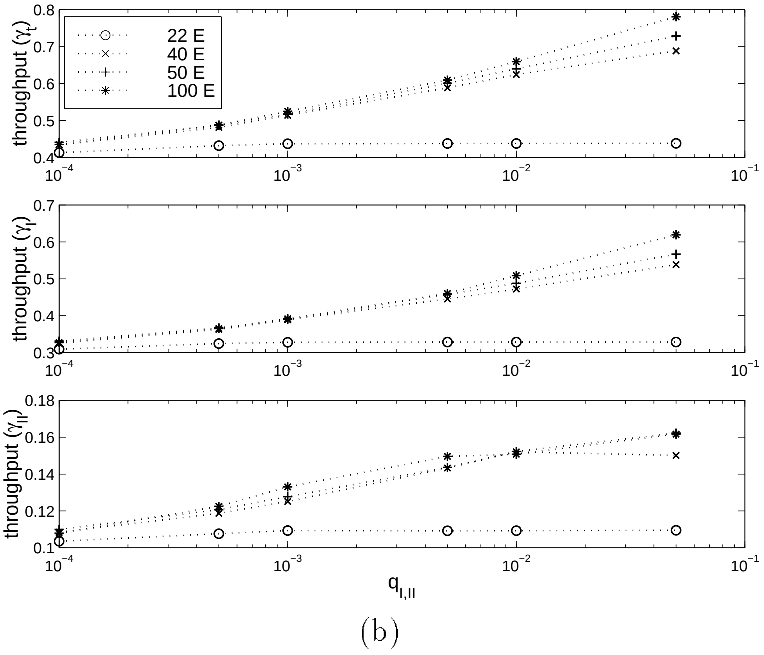


(b)

Figure 4-10: MMDR: $10^{-4} \leq q_I, q_{II} \leq .05$. Handoff dropping probabilities for (a) class I and (b) class II



(a)



(b)

Figure 4-11: MMDR: $10^{-4} \leq q_I, q_{II} \leq .05$ (a) call blocking probabilities, (b) normalized average throughput

Figure 4-10(b), we see that when $q_I = q_{II}$ is below 10^{-3}, the class II traffic is limited by the maximum handoff dropping probability. When $q_I = q_{II} \geq 10^{-3}$, the class II traffic is not limited by the handoff dropping probability requirement when the offered load is $22E$. Figure 4-10(a) is a bit more complex. For heavily loaded systems, once again, we note that the curves are within the band determined by the maximum dropping probability requirement in conjunction with the dropping probability threshold. The $40E$ load curve indicates that when $q_I \leq .01$, the system was saturated and partitions were used to ensure that the bounds were met. When $q_I = .05$, the system was unsaturated and even with minimal partitioning, the maximum $p_{d,I}$ was less than the requirement.

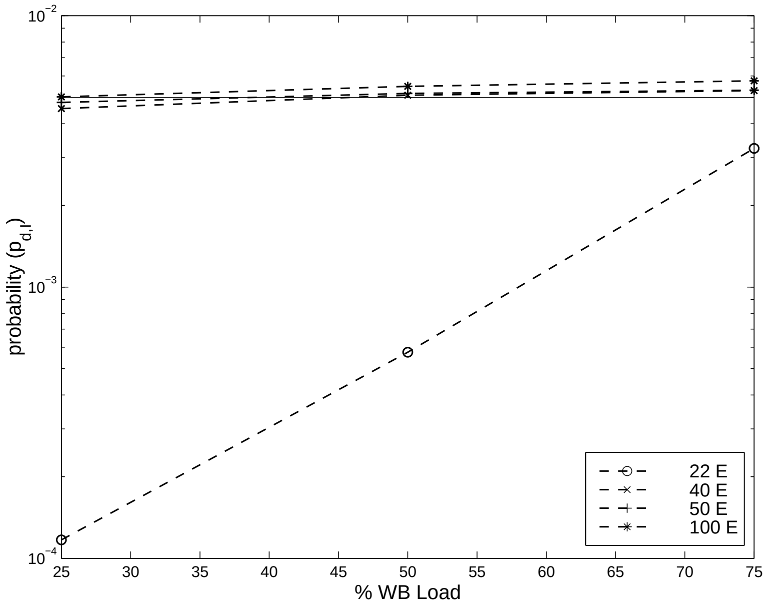
The more interesting case is the behavior of the $22E$ load curve. For the first two points on the curve, the joint requirements were strict enough such that the $p_{d,I}$ was within the band of the maximum allowed by the QoS requirements and partitions were used to ensure that adherence. However, unlike other saturation points, the $p_{d,I}$ first decreases before levelling off. This may be attributed to the interaction between the different classes and partitions. We may divide this curve into three parts. In the first part, both traffic classes require partitions in order to meet the defined QoS requirements. In the second portion of the curve, class I traffic meets the requirements while the class II traffic still requires partitions. As the requirements for the class II traffic decrease, the partition level required to achieve those requirements also decreases which in turn results in more bandwidth available for the other traffic streams (in this case the class I traffic) and a resulting decrease in $p_{d,I}$. Finally, in the last portion of the curve, the partition levels for both traffic classes have been satisfied and the bandwidth is completely shared among the traffic streams. As a result, the further loosening of the QoS constraints does not result in a corresponding decrease in the handoff dropping probabilities. We lastly note that the final point on the curve is statistically equivalent to the two points preceding it and is merely a simulation artifact.

It is important to note that though the QoS profiles provide class independent dropping probabilities, the traffic classes ultimately share the same channel and the loads of the different classes impact one another. Further insight is provided into this performance by the plotted blocking probabilities indicated in Figure 4-11(a).

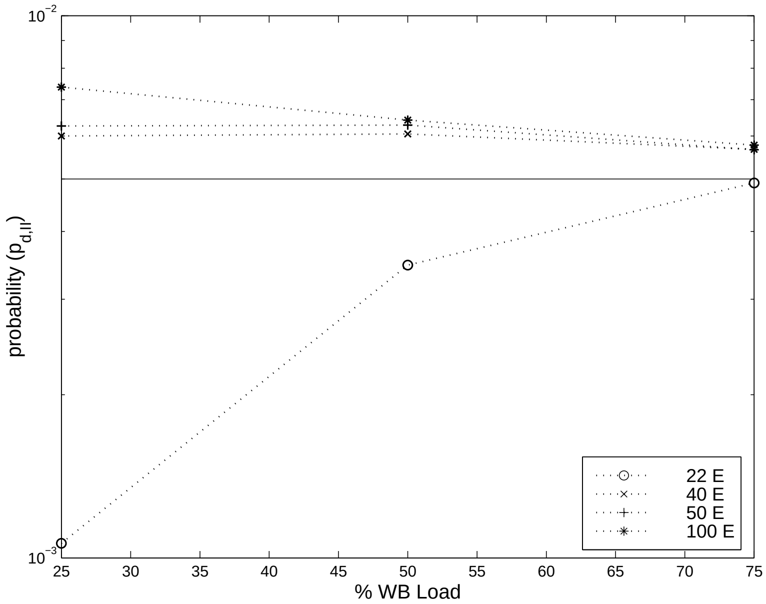
Finally, Figure 4-11(b) corroborates the above data. As noted previously, very lightly loaded systems do not show improvement in performance exhibited by increased throughput after the handoff partitions are not required to enforce the QoS requirements. Even when partitions are required to achieve the targetted handoff dropping probabilities, almost all the traffic is admitted (see Figure 4-11(a)). Further load applied to the system will achieve corresponding improvements in the throughput. However, this will come at the expense of an increase dropping probabilities.

4.4.4 Variation of Traffic Composition

In the next set of experiments, shown in Figures 4-12 and 4-13, we maintained the total load applied to the system constant and measured the performance against a varying mix of class I and class II traffic. In the heavily loaded systems, there was little difference in handoff dropping probabilities of both traffic classes as shown in Figure 4-12. This is indicated by the almost flat slope as the traffic varied between a 25%/75% NB:WB traffic mix ratio and a 75%/25% NB:WB traffic mix ratio.

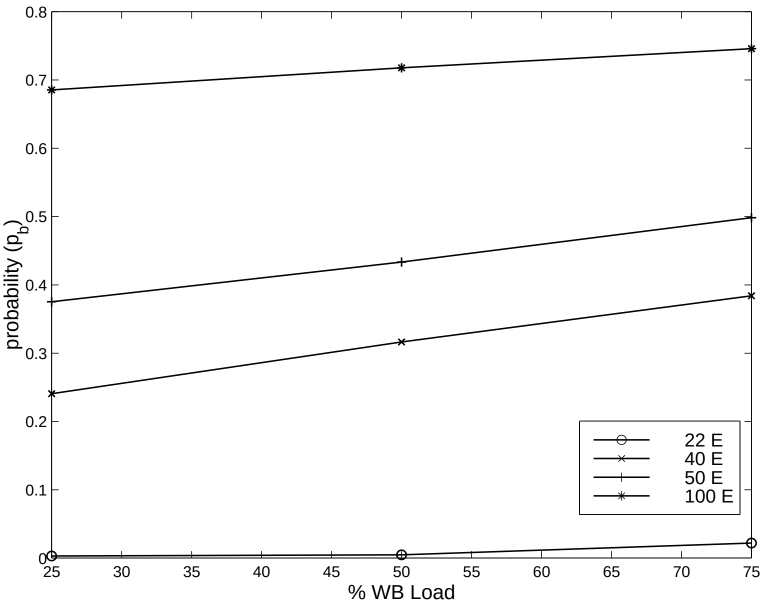


(a)

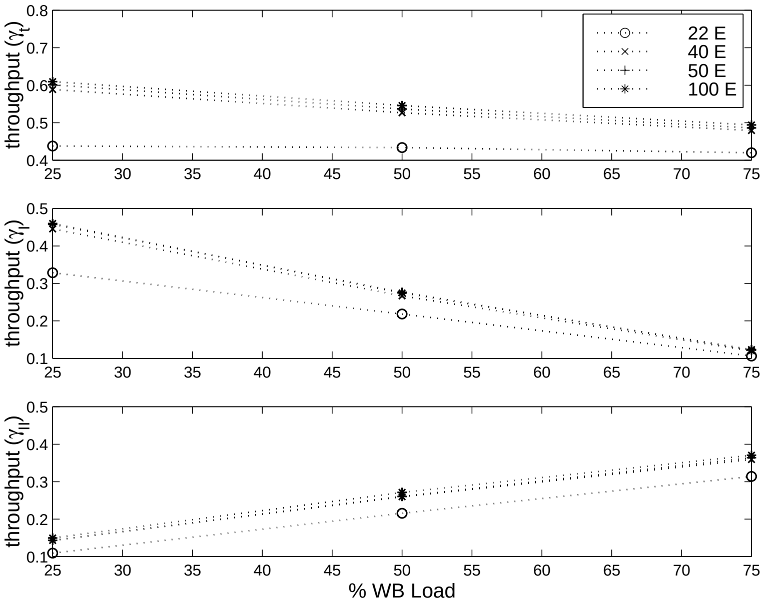


(b)

Figure 4-12: MMDR: WB to NB ratio varies between 1 : 3 and 3 : 1. Handoff dropping probabilities for (a) class I and (b) class II



(a)



(b)

Figure 4-13: MMDR: WB to NB ratio varies between 1 : 3 and 3 : 1 (a) call blocking probabilities, (b) normalized average throughput

Once again, the solid line indicates the targetted handoff dropping probability value q_c. This target is actually a band defined by $q_c \pm T_c$. The band is not reflected on the plots themselves.

When a light load of $22E$ was applied to the system, the results were more interesting. For class I traffic, $p_{d,I}$ seems to exhibit a logarithmic relationship with the change in percentage load. As expected, the dropping probability increases when a greater share of the load consists of class II traffic. This is plausible since the lack of smoothness in the variation of percentage channel occupancy during operation would contribute to increased dropping probability. This is not exhibited in the heavy load cases since the dropping probability is constrained to the maximum allowed value with the use of partitions. The class II traffic case is similar. As previously noted, $p_{d,II}$ exceeds the maximum. However, it is almost always within the allowed band as defined by the blocking threshold, T_c. The light load ($22E$) case does not seem to exhibit the exponential relationship with the variation in traffic mix that the class I traffic did though the general increase in $p_{d,II}$ with increase in percentage of class II traffic is indicated. This would indicate one of the following possibilities. The seeming exponential relationship exhibited by class I was completely accidental. A second possibility indicates that this relationship is applicable to class I traffic, but not class II traffic. A third option would say that it applies to both class I and II traffic assuming that there is no constraint applied by the algorithm to assure that the probability of dropping QoS condition will not be violated (as is the case in the $22E$ case when 75% of the load is due to class II traffic). Further study of simulation data and further simulations would be needed to ascertain which of the above options is correct. The general trend though does apply to both traffic classes and is an intuitive result.

Figure 4-13(a) is a plot of call blocking probability as a function of percentage load. We note that the call blocking probability increased linearly for mid to heavy loads as the percentage load was dominated by class II traffic. The slope of each curve was a function of the load applied to the system with variations ranging from about 2% for a $22E$ load, $12-15\%$ for loads of $40E$ and $50E$, to about 7% at $100E$. These results are intuitive as well and may be attributed to the non-linear nature of these systems.

An examination of the system throughput is given by Figure 4-13(b). The total throughput of the system decreases quasi-linearly for loads of $40E$, $50E$, and $100E$. This is not true for the $22E$ case. We assume that this is a further manifestation of the dropping probability behavior exhibited by the class II traffic. The non-linear nature of the system and transition from completely shared behavior to implementation of partitions would be directly responsible for that. Similarly, the class I throughput for $40E$, $50E$, and $100E$ decreases linearly and at the same slope as is directly attributable to the variation in percentage load. At low loads, it is still linear though the slope is different. This may be attributed to the fact that the blocking probabilities at the left hand side of the graph were below the enforced blocking probability QoS parameter thus allowing the throughput to be composed (between classes) on a new user completely shared basis with assignment of both new user partitions, NP_c, to be zero. The class II throughput is similar though the $22E$ load did exhibit more similar percentage increases to the higher load systems. This is understood in light of the dropping and blocking probability class II plot results discussed above.

4.4.5 Effect of Cell Size Variation

As mobile networks become increasingly ubiquitous, the range of applications and kinds of networks increase as well. In this section, we study the flexibility of MMDR in its performance in networks whose cells range in size from picocell to microcell to macrocell as we vary the average time until handoff $(1/h)$ between 10 and 1000 seconds while keeping the average call time $(1/\mu)$ constant at 500 seconds. These results are shown in Figures 4-14 and 4-15. All other QoS constraints and algorithm parameters are kept as above.

For the most part, we note that dropping probabilities for both traffic classes remain essentially constant. Blocking probability never varies more than a few percentage points with the maximum occuring in the mid-range of $1/h$ values. At the high end, handoff times are long compared to call times. Therefore, there are very few handoffs and little or no partitioning is needed to maintain the appropriate dropping probability QoS. At the low end, many handoffs are admitted per call. The statistical multiplexing of the channels also leads to a gain in the percentage of calls admitted. Further study is needed to better understand this phenomenon. Finally, the maximum system (and class) throughput increases with an increase in the average handoff time. This once again is attributed to the fact that a decrease in handoffs leads to lower partition levels which makes the behavior more closely mimic the complete sharing (CS) algorithm which thus produces higher throughput.

One caveat to comparing behavior across sizes of cells is that as the cell size decreases, the number of handoffs per call increases. Given an absolute bound on the probability of being dropped from service requires a decrease in dropping probability for smaller cells in order to keep that value correct.

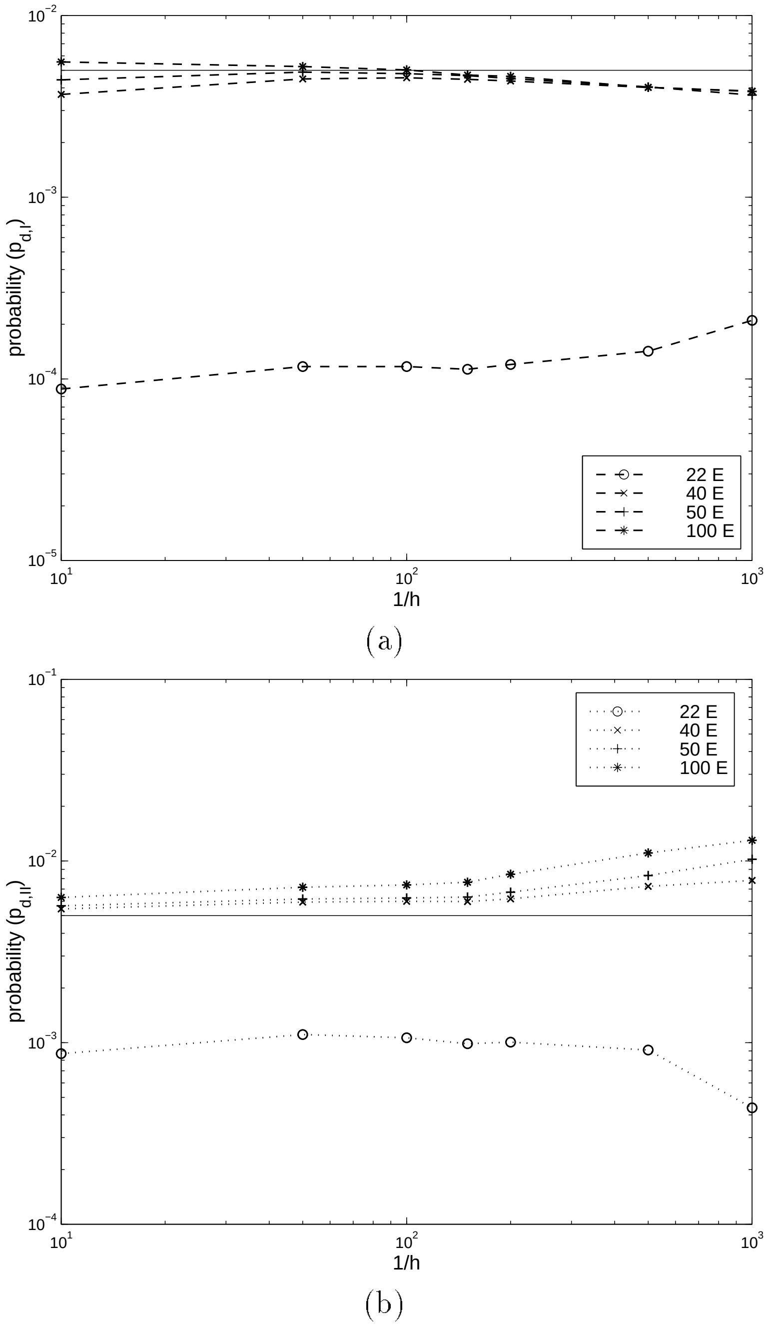
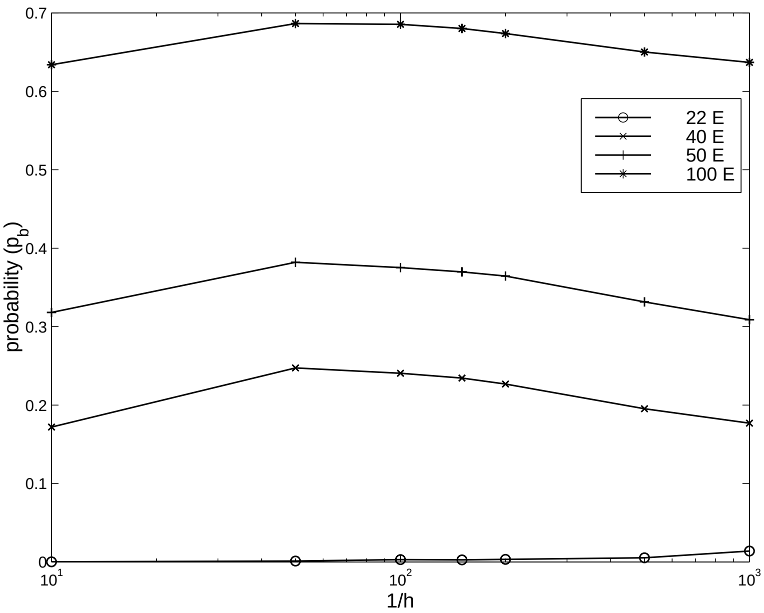
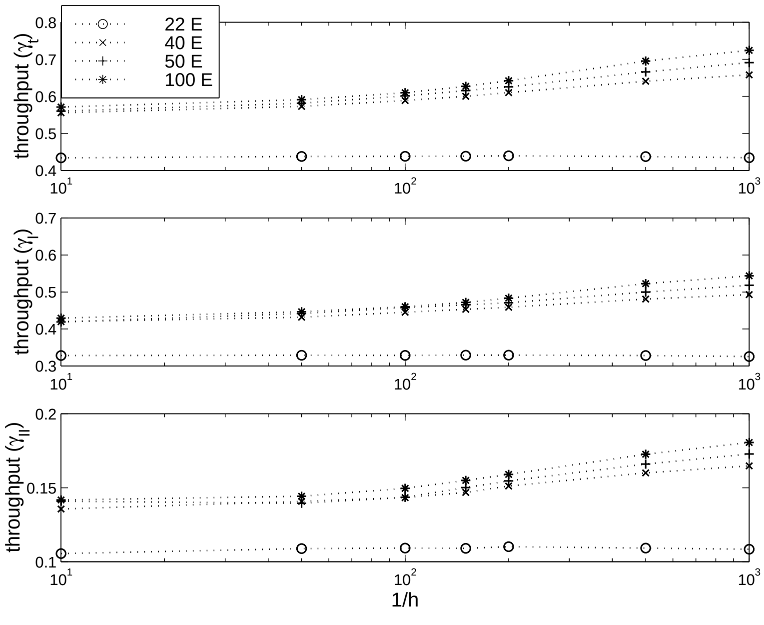


Figure 4-14: MMDR: $10 \leq 1/h \leq 1000$. Handoff dropping probabilities for (a) class I and (b) class II



(a)



(b)

Figure 4-15: MMDR: $10 \leq 1/h \leq 1000$ (a) call blocking probabilities, (b) normalized average throughput

In sum, we note that MMDR performs equally well in systems with cells of any size by automatically adjusting partitions to conform to the required QoS requirements.

4.4.6 Hotspot Scenarios

In the next set of experiments, we looked at hotspots located in a single cell. The hotspot load was set to $50E$, $100E$, and $200E$ successively for each case where the nominal loads were $22E$, $25E$, $35E$, and $50E$ in all of the other cells. Analysis of the results was done for each of the above cases. In both the hotspot and adjacent cells, the dropping probability QoS constraints were met in the same manner as for the homogeneous load. The blocking probabilities for both classes was somewhat elevated in cells adjacent to the hotspot cell as was expected. Likewise, the per cell throughput was elevated in cells closer to the hotspot cell.

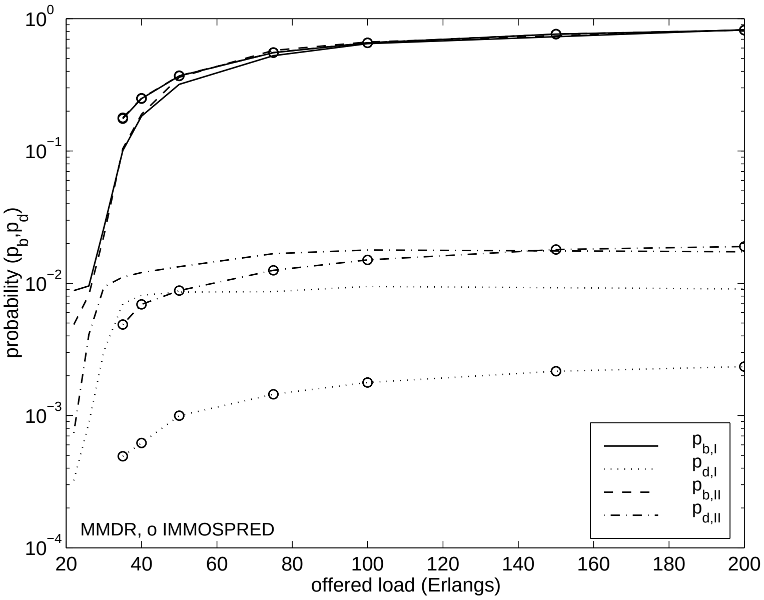
These results shed light on cases where load varies either from cell to cell or from time to time. We note that the completely distributed measurement based reservation system is able to adapt to changing situations and satisfy the QoS demands of the system. This motivates the notion that the only information needed to adequately service the QoS requirements is a measure of the blocking and dropping information in each cell. No information need be exchanged between cells. A potential benefit that may be gained with that information is an increase in the nominal operating load in the system which would directily correspond to an increase in the throughput seen by the system for given QoS requirements.

4.4.7 Comparison to Other Algorithms

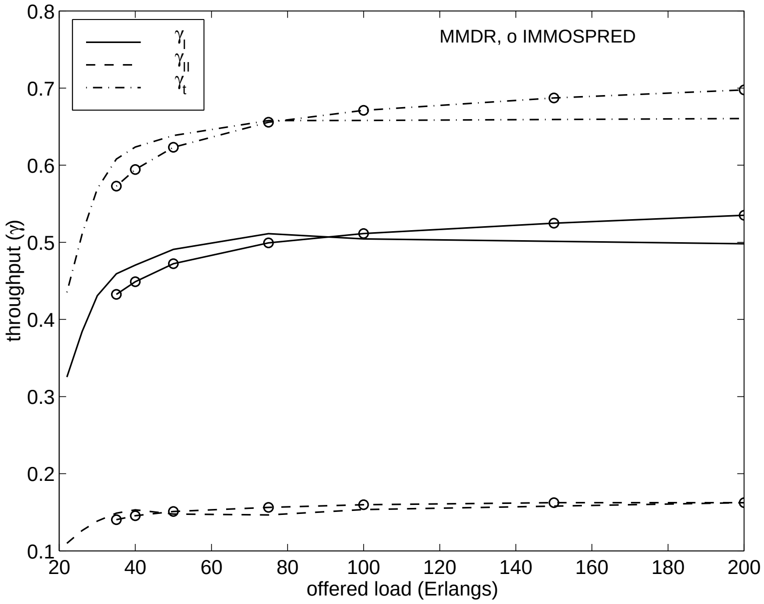
We finally compare MMDR to IMOSP-RES defined and discussed in Chapter 3. IMOSP-RES has the same QoS requirements as MMDR. It allows new users into the system assuming that the predicted handoff dropping probabilities for all users in both the home and adjacent cells will not exceed the QoS requirements given knowledge of the number of class I and class II users in all of the cells. Figure 4-16 contains dropping probability and throughput results where the q_c values are set to .005. We generally note that the performance of the two algorithms is comparable, indicating that it is possible to achieve the required QoS requirements using both prediction and measurement techniques.

Asymptotic overload results indicate that the performance of the two algorithms is very similar though IMOSP-RES achieves slightly higher throughput. At low loads though, MMDR achieves better results both in probability of blocking and in throughput. Since this is where the system operating point is located, the expected total throughput of MMDR is greater than IMOSP-RES. These results are true for traffic which varies slowly and in the case where we only consider long term average results. In systems where the average load varies considerably over short term intervals or where the QoS requirements are very strict and performance is measured on a per call basis, IMOSP-RES may be a better option.

MMDR is thus the algorithm of choice in the general case both for performance reasons and simplicity in implementation since its implementation is completely distributed and does not require cell and class occupancy from neighboring cells.



(a)



(b)

Figure 4-16: Comparison of MMDR to IMOSP-RES from Chapter ??. IMOSP-RES values are indicated by the measurements with circles.

4.5 Conclusion

MMDR is a computationally simple multi-class algorithm which is completely distributed in nature. Reservation partitions are adjusted independently in each cell based on the measured per class blocking and dropping probabilities. This is achieved using a two-tier partition structure which completely separates the handoff and new user QoS requirements into two independent decision mechanisms which allows us to simultaneously achieve the two-different types of QoS requirements. Extensions of MMDR to three or more traffic classes is left for further study. Given the structure of the algorithm, we expect that this may be accomplished directly. However, there is the expectation that performance will be compromised to a degree. Comparison to IMOSP-RES (from Chapter 3) indicates that MMDR is superior in performance when considering performance in terms of long-term averages. Since the partition mechanism is measurement-based, the instantaneous dropping probabilities may achieve inferior performance in some cases.

Chapter 5

Conclusions and Further Work

In this thesis, we have discussed three different approaches which may be used to solve the admission control problem in multi-class wireless networks. We introduced two different sets of QoS measures which define fair bandwidth allocation among different classes of users. Each class has its own requirements both in terms of absolute handoff dropping probabilities and call blocking probabilities. This is fundamentally different than other approaches generally taken. These often consider high and low priority traffic where the QoS of the low priority traffic is completely determined by the load of the high priority traffic in the system. This method is appropriate where the low priority class is used for best effort traffic. Other reservation or prediction algorithms consider the QoS of individual users which enter the system but neglect the relative service available to each of those traffic classes.

The static reservation algorithm introduced in Chapter 2 explored the concepts of pre- and post-reservation and their impact on system performance. We extended the notion of reservation to multiple traffic classes and analyzed the impact that parameter choice has on performance. Additionally, we showed how this algorithm performed relative to the complete sharing (CS) and complete partitioning (CP) algorithms which form the bound on possible algorithm choices in this realm.

We next looked at two different types of dynamic algorithms, predictive and measurement-based. Both of these algorithm types dynamically adjust performance during operation to conform to pre-defined network requirements. These algorithms also would easily adapt to changing standards or requirements both in time and space. While the reservation algorithms require information about the numbers of users in adjacent cells, the measurement algorithm, MMDR, is completely distributed and adapts based only on information collected within the cell.

The family of prediction algorithms we developed in Chapter 3 extended the notion of bandwidth or capacity prediction based on QoS requirements to multiple traffic classes. We explored the impact that variation of parameter choice had on the network. We additionally examined the impact that the addition of reservation partitions for the new and handoff users has on the network. This manifested itself in terms of more precise conformance to QoS standard at a cost to the system.

The last approach we devised is the MMDR algorithm. It uses per class handoff dropping and call blocking statistics collected in the cell to adjust a two-tier hierarchy of partitions to conform to the QoS requirements. We looked at the impact that parameter variation had on performance and compared results to the prediction based algorithms from Chapter 3.

The algorithms as they are described in this thesis assume that the traffic is essentially circuit switched. While in service, users use the total number of BBUs they are allocated upon entry into the system. We used this model to enable us to closely study the impact that parameter variation had on performance irrespective of other unrelated parameters. While we expect that the first order conclusions we reach in this thesis will hold true in packet-based systems, further examination of these systems is warranted with perhaps some small adaptations to take advantage of the mechanics of these systems. One such dimension includes the implementation of queuing and reduced bandwidth availability on handoff or during periods of congestion. This will likely improve QoS performance as described here at a cost in the quality of carried traffic.

In this thesis, we assumed homogeneous mobility. All users were equally likely to travel in any particular direction. It would be instructive to look at performance in networks where users are more likely to move along some pre-defined paths. Additionally, we could adapt the algorithms to these networks and look at how this would improve performance. On a related note, we consider traditional cellular mobile networks. The algorithms developed here could be adapted with very minor changes to other kinds of wireless networks such as ad-hoc wireless networks, mixed wireless-wired networks, and wireless local loop networks.

While it is possible to extend all of the algorithms described in this thesis to provide independent QoS for an arbitrary number of different traffic classes, we maintain that in most cases this will increase algorithm complexity and almost certainly result in a corresponding degradation in performance. This is likely to be true particularly in cases where a reservation type algorithm is implemented or class independent predictions are used and the percentage of traffic attributed to particular classes may be small and the total bandwidth available to the system is also small.

The algorithms as they are, on the other hand, may be used to support three traffic classes where two of the classes require real-time service and the third traffic class best-effort service. This may be sufficient for the large majority of systems currently being planned. A typical example of this would be for voice, video, and data. The results for the two real-time traffic classes would be exactly as shown for the two class cases, with the best-effort traffic using whatever bandwidth is not used by the other two classes. This class could be implemented as a non-pre-emptive priority traffic class with the real-time traffic pre-empting the best-effort traffic, assuming that it is admitted.

Assuming a packet-based system implementation, the best-effort traffic could additionally be queued and each of the calls serviced at a lower rate resulting in little if any degradation to the system. We additionally note that the best-effort traffic class would have a large bandwidth available to it particularly since the operating point of the system would typically be at relative low traffic load in order to ensure low blocking probabilities. When the average load is very low, the probability that there is a large amount of bandwidth available is great and the average delay is low. Additionally, we could provide soft guarantees to that traffic as follows. We measure the average delay to the best-effort traffic and admit best-effort traffic assuming that the required average delay of the users in the system is not violated and that the entering user would, on average, get adequate service. Handoff dropping probabilities could even be supported in the same sense. The major differences would be no blocking probability QoS requirement and the queueing of data when the bandwidth is needed for the real-time traffic.

On a more specific level, we now list several areas to be explored which apply to individual traffic classes. Using the static reservation algorithm, we can generalize the notion of pre- and post-reservation to include overlapping reserved pools such as those in MMDR. Additionally, we can apply and generalize the QoS measure used there to other networks. The prediction algorithm family may be modified in the following ways. Increasing the number of steps in the prediction process would produce more accurate prediction. Call length might also be incorporated into the prediction process. Analysis of the performance impact these changes would have on the system needs to be balanced against the added cost they incur. Additionally, examination of the cost versus performance dimension that the increased knowledge about the system affords could be studied. Finally, with the measurement-based algorithm, extension to a full-fledged multi-class algorithm together with performance analysis of those systems as well as more complex statistical measurement mechanisms is another area to be examined. Additionally, a more complete analysis of the adaptation process and relationship to the size of the requirements to insure good short-term performance in addition to long-term asymptotic performance is warranted. Examination of the partition adjustment increments could be used to maximize results in these dimensions as well.

The algorithms as discussed in this thesis are simulated in a fixed channel allocation (FCA) network where the number of channels or BBUs allocated to each cell is static. The algorithms may be directly implemented or adjusted with minor modifications for application in dynamic channel allocation algorithms. Implementation and performance analysis of the algorithms discussed here is an area for further study.

By the same token, the application of these algorithms and methods may be applied at the cell or packet level as they are additionally resource allocation problems.

Finally, we note that due to the multi-dimensionality of the problem together with the non-linearity introduced by the QoS requirements, capacity and performance analysis of the system and implemented algorithms is largely intractable. However, computation of performance bounds and approximation of individual algorithm performance and the inter-relationship of the different parameters and their impacts would aid in understanding of the systems and their comparisons to theoretically achievable bounds by any of the algorithms. This would aid in understanding how variation of parameters such as bandwidth allocated to a network and changes in cell size would impact the system more generally.

References

[1] A. Acampora and Z. Zhang. A throughput/delay comparison: Narrowband versus broadband wireless LAN's. *IEEE Transactions on Vehicular Technology*, 42(3), August 1993.

[2] A . S. Acampora and M. Naghshineh. An architecture and methodology for mobile-executed handoff in cellular ATM networks. *IEEE Journal on Selected Areas in Communications*, 12, October 1994.

[3] N. Amitay. Distributed switching and control with fast resource assignment/handoff for personal communications systems. *IEEE Journal on Selected Areas in Communications*, 11(6), August 1993.

[4] C. Purzynski ans S. S. Rappaport. Traffic performance analysis for cellular communication systems with mixed platform types and queued handoffs. In *IEEE 43rd Vehicular Technology Conference*, May 1993.

[5] R. Beck and H. Panzer. Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems. In *39th IEEE Vehicular Technology Conference*, 1989.

[6] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1987.

[7] S. K. Biswas and D. Reininger. Bandwidth allocation for VBR video in wireless atm winks. In *MoMuC '96*, pages 1–6, 1996.

[8] S. C. Borst and D. Mitra. Virtual partitioning for robust resource sharing: Computational techniques for heterogeneous traffic. *IEEE Journal on Selected Areas in Communications*, 16(5):668–678, June 1998.

[9] J. E. Wieselthier C. M. Barnhart and A. Ephremides. Admission control policies for multihop wireless networks. *Wireless Networks*, 1:373–387, 1995.

[10] C. Chang and M. Dai. Analysis of packet-switched data in a new basic rate user-network interface of ISDN. *IEEE Transactions on Communications*, 42(12):3129–3136, December 1994.

[11] C. Chao and W. Chen. Connection admission control for mobile multiple-class personal communications networks. *IEEE Journal on Selected Areas in Communications*, 15(8):1618–1626, October 1997.

[12] S. Choi and K. G. Shin. Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks. In *SIGCOMM '98*, pages 155–166, 1998.

[13] R. L. Cruz. Quality of service guarantees in virtual switched networks. *IEEE Journal of Selected Areas in Communications*, 13, August 1995.

[14] D. S. Eom, M. Sugano, M. Murata, and H. Miyahara. Call admission control for QoS provisioning in multimedia wireless ATM networks. *IEICE Transactions on Communications*, E82-B(1):14–23, January 1999.

[15] B. Epstein and M. Schwartz. Reservation strategies for multi-media traffic in a wireless environment. In *1995 IEEE 45th Vehicular Technology Conference*, pages 165–169, July 1995.

[16] B. Epstein and M. Schwartz. QoS-based admission control for independent multi-class traffic in cellular wireless networks. personal communication, 1998.

[17] B. Epstein and M. Schwartz. QoS-based predictive admission control for multi-media traffic. In M. Luise and S. Pupolin, editors, *Broadband Wireless Communications*, pages 213–224. Springer-Verlag, 1998.

[18] D. Everitt and D. Manfield. Performance analysis of cellular communication systems with dynamic channel assignment. *IEEE Journal on Selected Areas in Communications*, 7(9), October 1989.

[19] D. E. Everitt and N. W. Macfadyen. Analysis of multicellular mobile radiotelephone systems with loss. *British Telecom Technology Journal*, 1(2), 83.

[20] Y. Fang and I. Chlamtac. A new mobility model and its application in channel holding time characterization in PCS networks. In *INFOCOM '99*, 1999.

[21] M. Frodigh. Reuse-partitioning combined with traffic adaptive channel assignment for highway microcellular radio systems. In *GLOBECOM*, pages 1414–1418, 1992.

[22] D. Goodman. Cellular packet communications. *IEEE Transactions on Communications*, 38(8), August 1990.

[23] O. Grimlund and B. Gudmundson. Handoff strategies in microcellular systems. In *Forty-first Vehicular Technology Conference*, 1991.

[24] R. Guerin. *Queueing and traffic in cellular radio.* PhD thesis, California Institute of Technology, 1986.

[25] R. Guerin. Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications*, 36:153–163, February 1988.

[26] D. Hong and S. Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, 35(3):77–92, August 1986.

[27] D. Hong and S. S. Rappaport. Priority oriented channel access for cellular systems serving vehicular and portable radio telephones. *IEE Proceedings - I Communications, Speech, and Vision*, 136(5):339–346, October 1989.

[28] R. Howard. *Dynamic Probabilistic Systems, Volume 1: Markov Models.* John Wiley and Sons, 1971.

[29] J. M. Hyman, A. Lazar, and G. Pacifici. Real-time scheduling with quality of service constraints. *IEEE Journal on Selected Areas in Communications*, 9(7):1052–1063, September 1991.

[30] C. I, L. Greenstein, and R. Gitlin. A microcell/macrocell cellular architecture for low- and high-mobility wireless users. In *GLOBECOM*, 1991.

[31] C. I, L. Greenstein, and R. Gitlin. A microcell/macrocell cellular architecture for low- and high-mobility wireless users. *Journal on Selected Areas in Communications*, August 1993.

[32] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A measurement-based admission control algorithm for integrated service services packets networks. In *ACM SIGCOMM*, August 1995.

[33] J. Jeng, Y. Lin, and H. C. Rao. Flexible resource allocation scheme for GSM data services. *IEICE Transactions on Commmunications*, E81-B(10):1797–1802, October 1998.

[34] S. Jordan and A. Khan. A performance bound on dynamic channel allocation in cellular systems: equal load. *IEEE Transactions on Vehicular Technology*, pages 333–344, May 1994.

[35] S. Jordan and P. Varaiya. Control of multiple service, multiple resource communication networks. *IEEE Transactions on Communications*, 42(11):2979–2988, November 1994.

[36] S. Jordan and P. P. Varaiya. Throughput in mutiple service, multiple rescource communication networks. *IEEE Transactions on Communications*, 39:1216–1222, August 1991.

[37] I. Katzela and M. Naghshineh. Channel assignment schemes for cellular mobile telecommunications. *IEEE Personal Communications*, 3(3):10–31, June 1996.

[38] J. Keilson and O. C. Ibe. Cutoff priority scheduling in mobile cellular communication systems. *IEEE Transactions on Communications*, 43(2/3/4):1038–1045, 1995.

[39] F. Kojima, S. Sampei, and N. Morinaga. An intellegent radio resource management scheme for multi-layered cellular systems with different assigned bandwidths. *IEICE Transactions on Communications*, E81-B(12):2444–2453, December 1998.

[40] B. Kraimeche and M. Schwartz. Analysis of traffic access control strategies in integrated service networks. *IEEE Transactions on Communications*, October 1985.

[41] B. Kraimeche and M. Schwartz. Bandwidth allocation strategies in wide-band integrated networks. *IEEE Journal on Selected Areas in Communications*, SAC-4(6), September 1986.

[42] B. Kraimeche and M. Schwartz. A channel access structure for wideband ISDN. *IEEE Journal on Selected Areas in Communications*, October 1987.

[43] M. D. Kulavaratharasah and A. H. Aghvami. Teletraffic performance evaluation of microcellular personal communications networks (PCN's) with prioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, 48(1):137–152, January 1999.

[44] C. Lea and A. Alyatama. Bandwidth quantization and states reduction in the broadband ISDN. *IEEE/ACM Transactions on Networking*, 3(3):352–360, June 1995.

[45] W. Lee. Smaller cells for greater performance. *IEEE Communications Magazine*, November 1991.

[46] K. K. Leung, W. A. Massey, and W. Whitt. Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communications*, 12(8):1353–1364, October 1994.

[47] D. Levine, I. Akyildiz, and M. Naghshineh. A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. *IEEE/ACM Transactions on Networking*, pages 1–12, February 1997.

[48] J. Li, N. Shroff, and E. Chong. A new channel allocation scheme for cellular networks. *Wireless Networks*, July 1997. submitted.

[49] Y. B. Lin, S. Mohan, and A. Noerpel. Queueing priority channel assignment strategies for handoff and initial access for a PCS network. *IEEE Transactions on Vehicular Technology*, 43(3):704–712, 1994.

[50] Y. B. Lin, A. Noerpel, and D. Harasty. A nonblocking channel assignment strategy for handoffs. In *IEEE ICUPC*, September 1994.

[51] D. Macmillan. Delay analysis of a cellular mobile priority queueing network. *IEEE/ACM Transactions on Networking*, 3(3):310–319, June 1995.

[52] B. S. Maglaris and M. Schwartz. Optimal fixed framed multiplexing in integrated line- and packet-switched communications networks. *IEEE Transactions on Information Theory*, II-28(2):263–27, March.

[53] L. Merakos and S. Jangi. Voice packet losses and data integration in reservation random access protocols for wireless access networks. 1992.

[54] J. Misic, S. Chanson, and F. Lai. Admission control for wireless networks with heterogeneous traffic using event based resource estimation. In *IEEE ICCCN 97*, September 1997.

[55] N. Mitrou, G. Lyberopoulos, and A. Panagopoulou. Voice and data integration in the air-interface of a microcellular mobile communication system. *IEEE Transactions on Vehicular Technology*, 42(1), February 1993.

[56] A. Murase, I. Symington, and E. Green. Handover criterion for macro and microcellular systems. In *Forty-first Vehicular Technology Conference*, 1991.

[57] M. Naghshineh. *Distributed control of wireless/mobile networks*. PhD thesis, Columbia University, 1994.

[58] M. Naghshineh and M. Schwartz. Distributed call admission control in mobile/wireless networks. *IEEE Journal on Selected Areas in Communications*, May 1996.

[59] S. Nanda and D. Goodman. Dynamic resource acquisition: Distributed carrier allocation for TDMA cellular systems. In *GLOBECOM*, 1991.

[60] B. Ngo and H. Lee. Queueing analyses of traffic access control stategies with preemptive and nonpreemptive disciplines in wideband integrated networks. *IEEE Journal on Selected Areas in Communications*, 9(7):1093–1109, September 1991.

[61] S. Oh and D. Tcha. Prioritized channel assignment in a cellular radio network. *IEEE Transactions on Communications*, July 1992.

[62] C. Oliveira, J. Kim, and T. Suda. Quality-of-service guarantees in high-speed multimedia wireless networks. In *IEEE ICC '96*, 1996.

[63] C. Olivera, J. B. Kim, and T. Suda. An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks. *IEEE Journal on Selected Areas in Communications*, 16(6):858–874, August 1998.

[64] P. V. Orlik and S. S. Rappaport. A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions. *IEEE Journal on Selected Areas in Communications*, 16(5):788–803, June 1998.

[65] P. Ramsdale and W. Harrold. Techniques for cellular networks incorporating microcells. 1992.

[66] S. S. Rappaport. Modeling the hand-off problem in personal communication networks. In *IEEE Vehicular Technology Conference Proceedings*, pages 517–523, 1991.

[67] S. S. Rappaport. The multiple-call hand-off problem in high-capacity cellular communications systems. *IEEE Transactions on Vehicular Technology*, 40(3):546–557, August 1991.

[68] S. S. Rappaport and C. Purzynsky. Prioritized resource assignment for mobile cellular communications systems with mixed service platform types. *IEEE Transactions on Vehicular Technology*, 45(3):443–458, August 1996.

[69] E. Del Re, R. Fantacci, and G. Giambene. Handover and dynamic channel allocation techniques in mobile cellular networks. *IEEE Transactions on Vehicular Technology*, 44(2):229–237, 1995.

[70] J. Sarnecki, C. Vinodrai, A. Javed, P. O'Kelly, and K. Dick. Microcell design principles. *IEEE Communications Magazine*, 31(4):76–82, April 1993.

[71] M. Schwartz. *Telecommunication Networks: Protocols, Modeling and Analysis.* Addison-Wesley Publishing Company, Reading, Mass., 1987.

[72] M. Schwartz. *Broadband Integrated Networks.* 199.

[73] M. Schwartz. Network management and control issues in multimedia wireless networks. *IEEE Personal Communications*, pages 8–16, June 1995.

[74] N. Srivastava and S. Rappaport. Models for overlapping coverage areas in cellular and micro-cellular communication systems. *GLOBECOM*, 1991.

[75] A. Sutivong and J. Peha. Call admission control algorithms for cellular systems: Proposal and comparison. In *Globecom*, 1997.

[76] S. Tekinay. *Modeling and analysis of cellular networks with highly mobile heterogeneous traffic sources*. PhD thesis, George Mason University, 1994.

[77] S. Tekinay and B. Jabbari. A measurement-based prioritization scheme for handovers in mobile cellular networks. *IEEE Journal on Selected Areas in Communications*, 10(8):1343–1350, October 1992.

[78] L. K. Thong, O. Baldo, and A. H. Aghvami. Performance of distributed call admission control for multimedia high speed wireless mobile ATM networks.

[79] R. Valenzuela. Dynamic resource allocation in line-of-sight microcells. *IEEE Journal on Selected Areas in Communications*, 11(6), August 1993.

[80] R. Vijayan and J. M. Holtzman. Analysis of handoff algorithms using non-stationary signal strength measurements. In *GLOBECOM*, pages 1405–1409, 1992.

[81] W. Wong. Packet reservation multiple access in a metropolitan microcellular radio environment. *IEEE Journal on Selected Areas in Communications*, 11(6), August 1993.

[82] C. H. Yoon and C. K. Un. Performance of personal portable radio telephone systems with and without guard channels. *IEEE Journal of Selected Areas in Communications*, 11(6):911–917, 1993.

[83] I. Yoon and B. G. Lee. A distributed dynamic call admission control that supports mobility of wireless multimedia users.

[84] O. Yu and V. Leung. Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN. *IEEE Journal on Selected Areas in Communications*, 15(7):1208–1225, September 1997.

[85] T. P. Yum and W. Wong. Hot-spot traffic relief in cellular systems. *IEEE Journal on Selected Areas in Communications*, 11(6):934–940, August 1993.

[86] T. S. Yum and K. L. Yeung. Blocking and handoff perfomance analysis of directed retry in cellular mobile systems. *IEEE Transactions on Vehicular Technology*, 44(3):645–650, 1995.

[87] J. Zander and H. Eriksson. Asymptotic bounds on the performance of a class of dynamic channel assignment algorithms. *IEEE Journal on Selected Areas in Communications*, 11(6):926–932, August 1993.