Wireless Personal Communications (2006) 38: 357–376
DOI: 10.1007/s11277-006-9037-6

Admission Control Algorithms for Revenue Optimization with QoS Guarantees in Mobile Wireless Networks

ING-RAY CHEN[1], OKAN YILMAZ[1] and I-LING YEN[2]

[1]*Computer Science Department, Virginia Tech*
E-mail: irchen@vt.edu, oyilmaz@vt.edu
[2]*Department of Computer Science, University of Texas at Dallas*
E-mail: ilyen@utdallas.edu

Abstract. We propose and analyze call admission control algorithms integrated with pricing for revenue optimization with QoS guarantees to serve multiple service classes in mobile wireless networks. Traditional admission control algorithms make acceptance decisions for new and handoff calls to satisfy certain QoS constraints such as the dropping probability of handoff calls and the blocking probability of new calls being lower than a pre-specified threshold. We analyze a class of *partitioning* and *threshold-based* admission control algorithms that make acceptance/rejection decisions not only to satisfy QoS requirements but also to optimize the revenue of the system by taking prices and arrival/departure information of service calls into account. We show that for a "charge-by-time" pricing scheme, there exist optimal resource allocation settings under which the partitioning and threshold-based admission control algorithms would produce the maximum revenue obtainable by the system without sacrificing QoS requirements. Further, we develop a new hybrid admission control algorithm which outperforms both partitioning and threshold-based admission control algorithms over a wide range of input parameters characterizing the operating environment and service workload conditions. Methods for utilizing of the analysis results for real-time admission control for revenue optimization with QoS guarantees are described with numerical data given to demonstrate the applicability.

Keywords: admission control, QoS guarantees, performance analysis, revenue optimization, mobile networks

1. Introduction

Next generation wireless networks will carry real-time multimedia services such as video and audio and non-real-time services such as images and files. Increasing demands from individuals and businesses with different profiles expect a network that can easily adapt to user needs and growing population without compromising the Quality of Service (QoS), while they are traveling away from office or home.

Two of the most important QoS measures in cellular networks are percentages of new and handoff calls blocked due to unavailability of channels. Mobile users in a cellular network establish a connection through their local base station. A base station may support only a limited number of connections (channel assigned) simultaneously due to bandwidth limitations. Handoff occurs when a mobile user with an ongoing connection leaves the current cell and enters into another cell. Thus an ongoing, incoming connection may be dropped during a handoff if there is insufficient bandwidth in the new cell to support it. We can reduce the handoff call drop probability by rejecting new connection requests. Reducing handoff call drop probability could result in an increase in the new call blocking probability. As a result, there is a tradeoff between the handoff and new call blocking probabilities.

Thus far a single class of network traffic, such as voice (real-time), has been studied extensively. The *Guard channel algorithm* [1] assigns a higher priority to handoff calls where a fixed number of channels are reserved for handoff requests. Hong and Rappaport [2] proposed a *cutoff threshold algorithm* with no distinction made initially between new and handoff calls which are treated equally on a FCFS basis for channel allocation until a predetermined channel threshold is reached. When the threshold is reached, new calls are blocked (cutoff), allowing only handoff calls. They also [3] proposed a priority oriented algorithm where queuing of handoff calls is allowed. Guerin [4] demonstrated queuing of both new and handoff calls improves channel utilization while reducing the call blocking probabilities. Most of the above mentioned research methods focus on voice-based cellular systems.

Fang [5] presented a *thinning algorithm* which supports multiple types of services by calculating the call admission probability based on the priority and the current traffic situation. When network approaches congestion, call admissions are throttled based on the priority levels of calls, i.e., lower priority calls are blocked, and hence thinned to allow higher priority calls. Wang, Zeng and Agarwal [6] classified traffic as real-time and non-real time traffic and divided the channels in each cell into three parts, namely, new and handoff real-time calls, new and handoff non-real-time calls. For overflow of real-time and non-real-time service handoff requests from the first two groups of channels, real-time handoff service requests are given a higher priority than non-real-time service requests.

Li, Lin and Chanson [7] proposed a *hybrid cutoff priority algorithm* for multimedia services. Different cutoff thresholds are assigned to individual services. Handoff calls are given a higher threshold while new calls, controlled by a lower threshold, are served only if there are still channels available bounded by the lower threshold. Each service class is characterized by its QoS requirements in terms of the number of channels required. Ye, Hou and Papavassilliou [8] proposed a bandwidth reservation and reconfiguration mechanism to facilitate handoff processes for multiple services. In general these algorithms make acceptance decisions for new and handoff calls to satisfy QoS requirements in order to keep the dropping probability of handoff calls and the blocking probability of new calls lower than pre-specified thresholds.

In this paper, we propose and analyze a class of partitioning and threshold-based admission control algorithms and their hybrids that will make acceptance/rejection decisions not only to satisfy QoS requirements but also to optimize the revenue of the system based on charge-by-time pricing algorithms defined by the service provider. By “partitioning” we mean a number of channels are specifically reserved to serve handoff (or new) calls of a particular service type and calls from other service types would not use the reserved channels. By “threshold-based” we mean handoff (or new) calls of a service type are given a threshold and as long as the threshold is not reached and there are still channels available in the cell, handoff (or new) calls of that service type can be admitted. The threshold-based algorithm derives from [7] augmented with the concepts of revenue optimization consideration and QoS guarantees; it will be used as a baseline algorithm against which performance characteristics are compared. Compared with partitioning algorithms, threshold-based algorithms have an inherent multiplexing property that results in a large shared partition to be opened to accommodate multiple service classes as long as the thresholds are not reached. To take advantage of both partitioning and threshold-based admissions, we develop a hybrid admission control algorithm which outperforms both partitioning and threshold-based admission control algorithms over a wide range of parameter values characterizing operating environment and service workload conditions of a mobile wireless network.

Our work is different from previous ones as we develop partitioning and threshold-based algorithms and their hybrids to target real-time admission control for revenue optimization with QoS guarantees. The goal of revenue optimization with QoS guarantees is achieved by integrating pricing with call admission control to meet QoS requirements. We do not deal with dynamic pricing as proposed in [9] because we believe dynamic pricing (changing the charge rate dynamically) receives little acceptance in user or provider communities. Rather we integrate a static "charge-by-time" pricing algorithm with admission control to maximize the revenue generated subject to the constraints that system imposed QoS requirements are satisfied.

The rest of the paper is organized as follows. Section 2 states the system model and gives assumptions used in characterizing the operational environment of a wireless network. Section 3 develops a methodology for cells in the wireless network to gather input information from individual mobile users and neighbor cells to run the proposed admission control algorithms at runtime. Section 4 describes a class of partitioning and threshold-based admission control algorithms and a hybrid algorithm integrated with pricing for revenue optimization with QoS guarantees for mobile wireless networks. Performance models are developed for assessing and comparing the performance characteristics and these algorithms. Section 5 presents numerical data and provides physical interpretation of the results. It shows that there exists an optimal setting under which the revenue collected is maximized for both partitioning and threshold-based admission control algorithms. Moreover, the hybrid algorithm is demonstrated to outperform both partitioning and threshold-based algorithms. Finally, Section 6 discusses applicability, summarizes the paper, and outlines future research areas.

2. System Model

A cellular network is modeled by a flat architecture in which cells are connected consecutively. In the center of each cell, a base station is used to provide network services to mobile hosts within the cell. We assume there exists a number of distinct *service classes*, S^1, $S^2, \ldots, S^n$, which are characterized by the service type attribute. For example, the service types can be *realtime* and *non-real time*. Further, there are handoff and new calls for each service type with handoff calls having a higher priority than new calls. Each service type, other than requiring a number of bandwidth channels for the intrinsic bandwidth QoS requirement, can possibly impose a system-wide QoS requirement. For example, the handoff call drop probability of a service type being less than 5% could be a QoS requirement. Dropped handoff calls dissatisfy users more than blocked new calls do. Assume that for each service class, say i, a QoS constraint exists on the handoff call blocking probability $B_h^i t$ and the new call blocking probability $B_n^i t$.

From the perspective of a single cell, each service class is characterized by its arrival rate (including new service connections initiated by mobile users in the cell and for handoff service connections from neighbor cells), and departure rate (of leaving the cell). Let λ_n^i denote the arrival rate of *new* calls of service class i and μ_n^i, be the corresponding departure rate. Similarly, let λ_h^i denote the arrival rate of *handoff* calls of service class i, and μ_h^i be the corresponding departure rate. These parameters can be determined by inspecting statistics collected by the base station in the cell and by consulting with base stations of neighbor cells. In Section 3, we describe a method for estimating these parameters. Without loss of generality we assume that a cell has C channels where C can vary depending on the amount of bandwidth available in the cell. When service class *i* enters a handoff area from a neighboring cell, a handoff call request is generated. Each call has its specific QoS bandwidth requirement dictated by its service traffic type attribute. Assume that a service call of service class *i* (regardless of handoff or new) requires k^{i} channels.

From the perspective of a cellular network service provider, each service class also has a "price" associated with it such that the system receives some revenue corresponding to the price associated when the service is rendered. The service provider would like to maximize the total revenue obtained by the system by means of optimal pricing for service classes and performing admission control functions subject to the bandwidth resources available in the system. The system achieves total revenue maximization in a distributed manner by maximizing each individual cell's revenue. That is, each cell makes admission control decisions for new and handoff call requests taking into consideration of the price rate information of these service calls in order to maximize the revenue received from servicing new and handoff calls in the cell.

The total revenue obtained by the system is inherently related to the pricing algorithm employed by the service provider. While many pricing algorithms exist [10], the most prevalent with general public acceptance to date is the "price-rate" scheme by which a user is charged by the amount of time in service. We assume that such a "price-rate" pricing scheme is adopted by the service provider such that a call of service class i has a "charge-rate" of v^{i} per time unit. That is, if a call of service class i is admitted into a cell, and subsequently handed off to the next cell or terminated in the cell, a reward of v^i multiplied with the amount of time the service is rendered in the cell will be "earned" by the system. There is no distinction for handoff vs. new calls in pricing as long as the call is in the same service class. We conjecture that the network service provider will assign proper values v^i to each service class. The performance model developed in the paper will allow the service provider to calculate the revenue earned per unit time under an admission control algorithm by each individual cell such that the revenue obtained by the system is maximized while satisfying some system imposed QoS constraints.

3. Mobility and Service Call Pattern

We propose to use a simple yet efficient learning mechanism to estimate the values of λ_n^i, μ_n^i, λ_h^i and μ_h^i of service class *i* from a cell's perspective. This learning mechanism involves a mobility and service call pattern recognition algorithm being executed on individual mobile devices for scalability reasons. The algorithm summarizes mobility and service call information of each individual mobile user in two data structures, namely, a mobility probability matrix and a service call table.

Specifically, the mobility probability matrix summarizes the probability of the mobile user going from one cell to the next cell and the residence time (distribution) of each cell, given that the mobile user comes from a previous cell. Specifically, the matrix stores P_{BCD} and T_{BCD} where P_{BCD} is the probability of the mobile user entering into cell *D* (the next cell) from cell *C* (the current cell), given that the previous cell is *B*, and T_{BCD} is the average dwell time of the mobile user in cell *C*, given that the previous cell is *B* and the next cell is *D*. For resource management, the information summarized in the mobility probability matrix is provided to the cells when the mobile user migrates into them. A design variation is to consider not only the previous cell but also a path consisting of the previous cell and the second previous cell. In this design variation, the probability matrix will store P_{ABCD} and T_{ABCD} for the probability of the mobile user entering into cell D (the next cell) from cell C (the current cell), given that the previous two cells are A and B in the sequence, and the average dwell time of the mobile user in cell C, given that the next cell is D and the previous two cells are A and B. This design variation trades off storage and processing requirements for information accuracy and improves the accuracy in summarizing the mobility behavior of a mobile user.

Certainly any method used to summarize mobility and service call patterns of a mobile user must be light-weight, given the small processing, storage and communication capabilities of a mobile device nowadays. We propose to adopt a light-weight yet effective self-learning mechanism [11] originally designed for wireless LANs and now applied to cellular networks based on *rewarding* the correct state transition and *penalizing* incorrect state transitions such that the sum of all probabilities is one. Consider the design that takes the previous state, the current state and the next state into account. Also suppose for convenience that a cell has 6 neighbors. If the previous cell is B and the current cell is C, then in the mobility probability matrix we would have entries for P_{BCD1}, P_{BCD2}, P_{BCD3}, P_{BCD4}, P_{BCD5}, and P_{BCD6} for six possible next cells D_1, D_2, D_3, D_4, D_5 and D_6 as neighbors to cell C. Suppose that the mobile user actually goes to cell D_4 from cell C. Then, P_{BCD4} would be rewarded with a probability increment while all others would be penalized with a probability decrement such that the probability sum is still 1.

For a mobile user that exhibits a certain degree of regularity for movements and calls, eventually the mobility probability matrix will concentrate on certain state transition probabilities with values close to 1, while most others will have probabilities close to 0 due to the reward-penalty learning mechanism applied. Thus the mobility probability matrix will summarize the regular paths taken by the mobile user. On the other hand, T_{BCD1}, T_{BCD2}, T_{BCD3}, T_{BCD4}, T_{BCD5}, and T_{BCD6} are updated accordingly depending on the actual path taken by the mobile user. This can be determined by each mobile user easily by keeping track of the average dwell time that the mobile user stays in a particular cell, given the history of the previous cell and the next cell.

The second data structure, a service call table also maintained by individual mobile devices to summarize call patterns, is to be populated as the mobile device goes through a sequence of calls. Specifically, for each cell visited by the mobile device, the table stores four "rate" values related to calls, namely, the arrival rate of a new call made by the mobile device in cell C, denoted by $\Lambda_n(C)$, the departure rate of a new call by the mobile device at cell C, denoted by $\theta_n(C)$, the arrival rate of a handoff call from cell C into its neighbor cells, denoted by $\Lambda_h(C)$, and the departure rate of a handoff call in cell C, denoted by $\theta_h(C)$. These four rate values reflect the frequency at which new calls are initiated and terminated, and also the frequency at which handoff calls arrive from neighbor cells and terminated by the mobile device, thus summarizing the call patterns of a mobile device. The computational procedure for obtaining the values of these rate parameters is also very lightweight, involving only a few variables to be kept in the mobile device's memory which are being updated as calls are made and terminated by the mobile user roaming across cells in the wireless network.

Our approach is to have each cell make admission control decisions to allocate resources to calls based on summarized mobility and service call patterns of those mobile users currently in the cell to intelligently know their call expected arrival and departure rates, as well as of those mobile users in the neighbor cells in order to know their expected handoff call arrival rate. From a cell's perspective (say, cell C), the arrival rate of handoff calls from mobile users in all the neighbor cells (say, B's), denoted by $\lambda_h(C)$, is given by:

$$\lambda_h(C) = \sum_{B \in M} \sum_{\substack{\text{all} \\ \text{users} \\ \text{in}B}} \Lambda_h(B) \times P_{\text{BC}}$$

Here M is the set of neighbor cells of cell C and P_{BC} is the probability that if the mobile user is in cell B it will go to cell C as the next cell, which can be calculated easily by each individual mobile user through a look-up of its mobility probability matrix by summarizing the probabilities of P_{ABC}, i.e.,

$$P_{\text{BC}} = \sum_{\text{all } A} P_{\text{ABC}}$$

Let $\lambda_n(C)$ denote the arrival rate of new calls, $\mu_n(C)$ denote the departure rate of a new call in cell C, and $\mu_h(C)$ denote the departure rate of a handoff call in cell C. Then,

$$\lambda_n(C) = \sum_{\substack{\text{all} \\ \text{users} \\ \text{in}C}} \Lambda_n(C)$$

$$\mu_n(C) = \frac{\sum_{\substack{\text{all} \\ \text{users} \\ \text{in}C}} \theta_n(C)}{\text{number of users}}$$

and

$$\mu_h(C) = \frac{\sum_{\substack{\text{all} \\ \text{users} \\ \text{in}C}} \theta_h(C)}{\text{number of users}}$$

Note that the arrival rate of all new calls is an aggregate measure summing all new call arrival rates by individual users in the cell, while the departure rate per call is an average parameter, averaging over all the mobile users in the cell.

The design of resource management for revenue optimization hinges on the adaptive admission control algorithm executed by individual cells. A cell, say C, will dynamically collect mobility and service call patterns in the form of $\Lambda_n(C)$, $\theta_n(C)$, $\Lambda_h(B)$, where B is a neighbor cell of C, and $\theta_h(C)$ from mobile users of a service class currently in its cell and will communicate with neighbor cells regarding the handoff arrival rate to have knowledge of the expected arrival and departure rates of new calls and handoff calls from these mobile users. Then each cell can intelligently allocate resources to serve new and handoff calls of various service types of different *charge* rates (i.e., a call of service class i has a "charge-rate" of v^i per time unit) to maximize the revenue obtainable, limited by the amount of resources available (bandwidth channels).

4. Admission Control for Revenue Optimization with QoS Guarantees

In this section we develop admission control algorithms integrated with pricing for revenue optimization with QoS guarantees in wireless mobile environments. For ease of presentation, we assume that there are two service types, class 1 (high-priority) and class 2 (low-priority), distinguished primarily by their traffic type, i.e., real-time and non-real-time respectively.

These algorithms can be easily applied to the case in which more than two service classes exist. The traffic input parameters λ_n^1, μ_n^1, λ_h^1 and μ_h^1 for class 1 and λ_n^2, μ_n^2, λ_h^2 and μ_h^2 for class 2 are obtained as described in Section 3. The superscript in the notation denotes the class type; the cell id (X) in the notation is dropped as we now refer to a general cell in the system.

4.1. Partitioning Admission Control

A partitioning call admission control policy divides the total number of channels in a cell into several fixed partitions with each partition specifically reserved to serve a particular service class (real-time vs. non-real-time) and call type (new vs. handoff). Thus for our example system there exist four partitions: high-priority handoff calls, high-priority new calls, low-priority handoff calls, and low-priority new calls, as illustrated in Figure 1.

By "partitioning" here we mean that a fixed number of channels are allocated to a specific service type and a call type and it cannot be used or shared by others. In calculating the expected revenue we assume that we have a priori knowledge of the arrival rate of calls and the service class to which it belongs. This knowledge is essential to justify the admission of a call into a cell in order to maximize the revenue of a cell at any given point of time. The network service provider normally has specified the desired threshold blocking probability for both new and handoff calls for different service classes in order to satisfy QoS constraints.

With the scenario of two service types, the following are the input parameters to a cell: C, λ_h^1, μ_h^1, λ_n^1, μ_n^1, λ_h^2, μ_h^2, λ_n^2, μ_n^2, v^1, v^2, k^1, k^2, $B_h^1 t$, $B_h^2 t$, $B_n^1 t$, and $B_n^2 t$ where k^1 and k^2 are the number of channels used by class 1 and class 2 calls, and $B_h^1 t$, $B_h^2 t$, $B_n^1 t$, and $B_n^2 t$ are threshold blocking probability of high-priority handoff, low-priority handoff, high-priority new, and low-priority new calls, respectively.

Under the partitioning admission control algorithm, the total number of channel C is divided into C_h^1, C_n^1, C_h^2, and C_n^2 channels for high-priority handoff calls, high-priority new calls, low-priority handoff calls, and low-priority new calls, respectively, as shown in Figure 1. These parameters are subject to the constraint:

$$C_h^1, C_n^1, C_h^2, C_n^2 \leq C \tag{1}$$

Let $(n_h^1, n_n^1, n_h^2, n_n^2)$ be the numbers of calls corresponding to the four fixed partitions denoted by $(C_h^1, C_n^1, C_h^2, C_n^2)$. Then $n_h^1\ k^1 = C_h^1$, $n_n^1\ k^1 = C_n^1$, $n_h^2\ k^2 = C_h^2$, and $n_n^2\ k^2 = C_n^2$ subject to the constraint that:

$$C_h^1 + C_n^1 + C_h^2 + C_n^2 = C \tag{2}$$

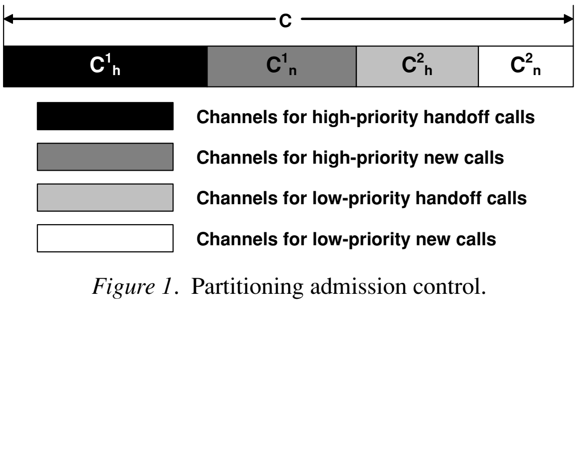
C
C¹h
C¹n
C²h
C²n
Channels for high-priority handoff calls
Channels for high-priority new calls
Channels for low-priority handoff calls
Channels for low-priority new calls

Figure 1. Partitioning admission control.

The QoS constraints to be satisfied are the blocking probability of new and handoff calls for both classes 1 and 2 calls. That is, we would like to partition C channels such that the following QoS constraints are satisfied:

$$B_h^1 < B_h^1 t \tag{3}$$

$$B_n^1 < B_n^1 t \tag{4}$$

$$B_h^2 < B_h^2 t \tag{5}$$

$$B_n^2 < B_n^2 t \tag{6}$$

The revenue that a successfully terminated or handed-off call brings to the cell is calculated by the product of the call's price rate parameter v^i with the duration of the call in the cell. Specifically, suppose that the partitioning algorithm reserves $(C_h^1, C_n^1, C_h^2, C_n^2)$ channels for revenue optimization, resulting in N_h^1, N_n^1, N_h^2, and N_n^2 high-priority handoff calls, high-priority new calls, low-priority handoff calls, and low-priority new calls, respectively, successfully terminated or handed off per unit time in the cell. Then the cell will receive the following revenue *per unit time* due to the deployment of the partitioning admission control algorithm:

$$\left(\frac{N_h^1}{\mu_h^1} + \frac{N_n^1}{\mu_n^1}\right)v^1 + \left(\frac{N_h^2}{\mu_h^2} + \frac{N_n^2}{\mu_n^2}\right)v^2 \tag{7}$$

Thus the optimization problem for the partitioning algorithm is to identify the best partition $(C_h^1, C_n^1, C_h^2, C_n^2)$ that would maximize the cell's revenue subject to the imposed QoS constraints defined by Conditions 3 through 6.

Under the partitioning algorithm, if a new high-priority (i.e., class 1) call arrives at a cell and all the channels allocated to serve high-priority new calls are used up, then the call is rejected. Similar reasoning applies to other service classes, too. No sharing is allowed among multiple partitions that exist. In this case the system behaves as if it is managing four concurrent queues: an $M/M/n_h^1/n_h^1$ queue to serve high-priority handoff calls with arrival rate λ_h^1, service rate μ_h^1, and the number of call slots allocated being n_h^1 (such that $n_h^1 k^1 = C_h^1$), an M/M/n_n^1/n_n^1 queue to serve new high-priority new calls in a cell with arrival rate λ_n^1, service rate μ_n^1, and the number of call slots allocated being n_n^1 (such that $n_n^1 k^1 = C_n^1$), an $M/M/n_h^2/n_h^2$ queue to serve low-priority handoff calls with arrival rate λ_h^2, service rate μ_h^2, and the number of call slots being n_h^2 (such that $n_h^2 k^2 = C_h^2$), and an $M/M/n_n^2/n_n^2$ queue to serve low-priority new calls with arrival rate λ_n^2, service rate μ_n^2, and the number of call slots being n_n^2 (such that $n_n^2 k^2 = C_n^2$).

The call dropping probabilities for handoff calls for various service classes (i.e., B_h^1 and B_h^2) and the blocking probability for new calls for various service classes (i.e., B_n^1 and B_n^2) can be determined easily by calculating the probability of the partition allocated to serve the specific calls being full. We can calculate the revenue generated per unit time by the partition reserved to serve only high-priority handoff calls by associating a reward of $i * v_h^1$ for state i in the M/M/n_h^1/n_h^1 queue. The same way applies to other partitions. Specifically, we can compute the revenue per unit time to the cell by:

$$\mathrm{PR}\big(C, \lambda_h^1, \lambda_n^1, \lambda_h^2, \lambda_n^2\big) = \mathrm{PR}_h^1 + \mathrm{PR}_n^1 + \mathrm{PR}_h^2 + \mathrm{PR}_n^2 \tag{8}$$

where the notation $\mathrm{PR}(C, \lambda_h^1, \lambda_n^1, \lambda_h^2, \lambda_n^2)$ is used to stand for the revenue rate earned by the partitioning algorithm as a function of C, $\lambda_h^1, \lambda_n^1, \lambda_h^2, \lambda_n^2$ (with other parameters not listed), while $\mathrm{PR}_h^1 + \mathrm{PR}_n^1 + \mathrm{PR}_h^2$ and PR_n^2, stand for the revenues generated per unit time due to high-priority handoff calls, high-priority new calls, low-priority handoff calls, and low-priority new calls, respectively, as given by (only PR_h^1 is shown below since expressions for others are similar):

$$\mathrm{PR}_h^1 = \sum_{i=1}^{n_h^1} i v_h^1 \frac{\frac{1}{i!}\left(\frac{\lambda_h^1}{\mu_h^1}\right)^i}{1 + \sum_{j=1}^{n_h^1} \frac{1}{j!}\left(\frac{\lambda_h^1}{\mu_h^1}\right)^j} \tag{9}$$

A partitioning solution is “legitimate” if B_h^1, B_h^2, B_n^1 and B_n^2 obtained satisfy Conditions 3 through 6. A partitioning admission control integrated with pricing for revenue optimization with QoS guarantees aims to find the optimal set $(C_h^1, C_n^1, C_h^2, C_n^2)$ that will yield the maximum revenue obtained among all legitimate solutions.

4.2. Threshold-Based Admission Control

In the threshold-based admission control algorithm, we select a threshold C_T to separate class 1 from class 2 based on the service type, i.e., real-time vs. non-real time. The meaning of the threshold is that when the number of channels used in the cell exceeds C_T then new or handoff calls from service class 2 (low-priority) will not be admitted. Within each service class, we further create thresholds to differentiate handoff from new calls such that C_{hT}^1 is the threshold for class 1 high-priority handoff calls; C_{nT}^1 is the threshold for class 1 high-priority new calls; C_{hT}^2 is the threshold for class 2 low-priority handoff calls; and C_{nT}^2 is the threshold for class 2 low-priority new calls.

Figure 2 illustrates the threshold-based admission control algorithm. Since we give handoff calls a higher priority than new calls, the following additional conditions must also be satisfied:

$$C_{nT}^1 \geq C_T, \; C_{hT}^1 \geq C_T \tag{10}$$

$$C_{nT}^2 \leq C_T, \; C_{hT}^2 \leq C_T \tag{11}$$

A threshold-based admission control integrated with pricing for revenue optimization with QoS guarantees thus aims to find the optimal set $(C_{hT}^1, C_{nT}^1, C_{hT}^2, C_{nT}^2)$ satisfying Conditions 10 and 11 that would yield the highest revenue while satisfying the QoS constraints specified by Conditions 3 through 6.

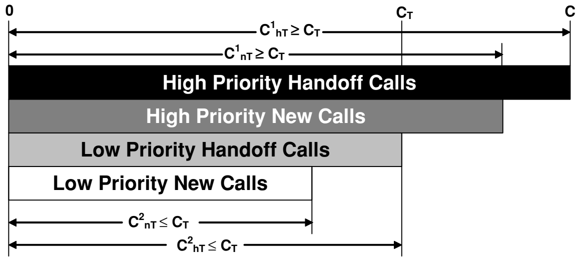

Figure 2. Threshold-based admission control.

We analyze the threshold-based admission control algorithm by using an SPN model. An SPN model is used rather than a Markov model because of the interdependency between thresholds assigned to handoff and new calls of various service classes. The SPN model adopts the idea from [7] and is generically applicable to multiple service classes. Figure 3 shows an SPN model for the threshold-based admission control with two service classes.

The transitions and places shown in Figure 3 are described as follows. For *transitions*, E_n^i models new call arrivals of service class i at rate λ_n^i; E_n^i models handoff call arrivals of service class i at rate λ_h^i; S_n^i models service of new calls of service class *i* with a service rate of $M(\mathrm{UC}_n^i)$ multiplied with μ_n^i where $M(\mathrm{UC}_n^i)$ stands for the number of tokens in place UC_h^i; and S_h^i models service of handoff calls of service class *i* with a service rate of $M(\mathrm{UC}_h^i)$ multiplied with μ_h^i where $M(\mathrm{UC}_h^i)$ stands for the number of tokens in place UC_h^i. For places, UC_n^1 models the execution state of service class 1 new call; UC_h^1 models the execution state of service class 1 handoff calls; UC_n^2 models the execution state of service class 2 new calls; and UC_h^2 models the execution state of service class 2 handoff calls.

A new service request arrival is admitted only if the threshold assigned is not yet reached. Therefore we assign an enabling predicate to guard E_n^1, E_h^1, E_n^2, and E_h^2, with thresholds C_n^1, C_h^1, C_n^2, and C_h^2, respectively. Consequently, the enabling predicate of E_n^1 is $[M(\mathrm{UC}_n^1) + M(\mathrm{UC}_h^1)]\,k^1 + k^1 + [M(\mathrm{UC}_n^2) + M(\mathrm{UC}_h^2)]\,k^2 \leq C_n^1$. The enabling predicate of E_h^1 is $[M(\mathrm{UC}_n^1) + M(\mathrm{UC}_h^1)]\,k^1 + k^1 + [M(\mathrm{UC}_n^2) + M(\mathrm{UC}_h^2)]k^2 \leq C_h^1$. The enabling predicate of E_n^2 is $[M(\mathrm{UC}_n^1) + M(\mathrm{UC}_h^1)]\,k^1 + k^2 + [M(\mathrm{UC}_n^2) + M(\mathrm{UC}_h^2)]k^2 \leq C_n^2$. Finally, the enabling predicate of E_n^2 is $[M(\mathrm{UC}_n^1) + M(\mathrm{UC}_h^1)]\,k^1 + k^2 + [M(\mathrm{UC}_n^2) + M(\mathrm{UC}_h^2)]k^2 \leq C_h^2$.

The blocking probability B_n^1, B_h^1, B_n^2 and B_h^2 are calculated from the SPN model by:

$$B_n^1 = \frac{\left(\lambda_n^1 - \mathrm{rate}\left(E_n^1\right)\right)}{\lambda_n^1}$$

$$B_h^1 = \frac{\left(\lambda_h^1 - \mathrm{rate}\left(E_h^1\right)\right)}{\lambda_h^1}$$

$$B_n^2 = \frac{\left(\lambda_n^2 - \mathrm{rate}\left(E_n^2\right)\right)}{\lambda_n^2}$$

$$B_h^2 = \frac{\left(\lambda_h^2 - \mathrm{rate}\left(E_h^2\right)\right)}{\lambda_h^2}$$

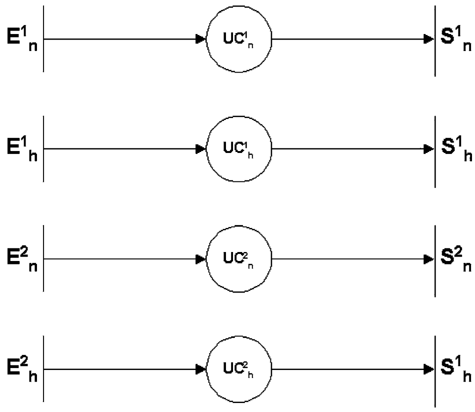

Figure 3. An SPN model for threshold-based admission control with two service classes.

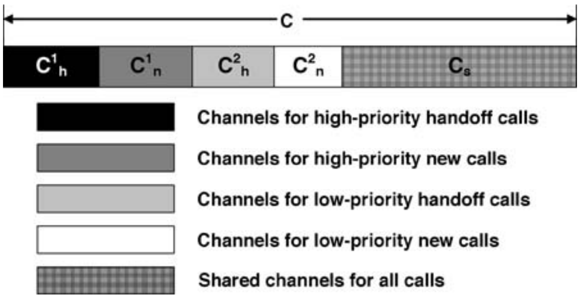
C
C¹h
C¹n
C²h
C²n
Cs
Channels for high-priority handoff calls
Channels for high-priority new calls
Channels for low-priority handoff calls
Channels for low-priority new calls
Shared channels for all calls

Figure 4. Partitioning-threshold hybrid admission control.

where rate (E_c^i) is calculated by finding the expected value of a random variable X defined as $X = \lambda_c^i$ if E_c^i is enabled; 0 otherwise. A "legitimate" solution from a threshold admission control algorithm must generate B_n^1, B_h^1, B_n^2, and B_h^2 to satisfy the QoS constraints specified by Conditions 3 through 6 discussed earlier.

We compute the revenue generated per unit time from the threshold-based admission control algorithm to the cell by:

$$\mathrm{TR}\big(C, \lambda_h^1, \lambda_n^1, \lambda_h^2, \lambda_n^2\big) = \mathrm{TR}_h^1 + \mathrm{TR}_n^1 + \mathrm{TR}_h^2 + \mathrm{TR}_n^2 \tag{12}$$

Here TR_h^1, TR_n^1, TR_h^2, and TR_n^2 stand for the revenues generated per unit time due to high-priority handoff calls, high-priority new calls, low-priority handoff calls, and low-priority new calls, respectively, given by:

$$\mathrm{TR}_h^i = \big(1 - B_h^i\big)\,\lambda_h^i\, v^i/\mu_h^i \tag{13}$$

$$\mathrm{TR}_n^i = \big(1 - B_n^i\big)\,\lambda_n^i\, v^i/\mu_n^i. \tag{14}$$

4.3. Hybrid Partitioning and Threshold-Based Admission Control

We devise a hybrid admission control algorithm to take advantage of both partitioning and threshold-based. The hybrid algorithm also divides the channels into fixed partitions the same way as the partitioning algorithm does. However, to take advantage of multiplexing, a "shared" partition is reserved to allow calls of all service classes/types to compete for its usage in accordance with the threshold algorithm. Figure 1 illustrates the hybrid algorithm. The shared partition is available for use by a service class/type only if the partition reserved for that service class/type is used-up. For example, class 1 handoff calls are allowed to use the channels in the shared partition only if all the channels reserved for class 1 handoff calls in the C_h^1 partition have been used up.

Let n_{hs}^1, n_{ns}^1, n_{hs}^2, n_{ns}^2 be the numbers of high-priority handoff calls, high-priority new calls, low-priority handoff calls, and low-priority new calls, respectively, in the shared partition. Let C_s be the number of channels allocated to the shared partition under the hybrid algorithm. Then, the number of calls of various service classes and types admitted into the shared partition are limited by C_s channels allocated to the shared partition, that is,

$$n_{hs}^1 k^1 + n_{ns}^1 k^1 + n_{hs}^2 k^2 + n_{ns}^2 k^2 \le C_s \tag{15}$$

subject to the constraint that:

$$C_h^1 + C_n^1 + C_h^2 + C_n^2 + C_s = C \tag{16}$$

The QoS constraints specified by (3) through (6) and the revenue earned *per unit time* as specified by Equation (7) remain applicable to the hybrid partitioning algorithm.

Note that the hybrid algorithm encompasses the partitioning algorithm as a special case in which $C_s = 0$ and also the threshold-based algorithm as another special case in which C_h^1, C_n^1, C_h^2, and C_n^2 are all zero. The performance model for the hybrid algorithm is composed of two submodels: one for the partitioning algorithm with the four fixed partitions C_h^1, C_n^1, C_h^2, and C_n^2 and one for the threshold-based algorithm for which $C = C_s$. Since the fixed partitions are modeled as M/M/n/n queues, the arrival rates into the shared partition from high-priority handoff calls (λ_{hs}^1), high-priority new calls (λ_{ns}^1), low-priority handoff calls (λ_{hs}^2), and low-priority new calls (λ_{ns}^2) are simply the spill-over rates from their respective M/M/n/n queues, e.g.,

$$\lambda_{hs}^1 = \lambda_h^1 \frac{\frac{1}{n_h^1!}\left(\frac{\lambda_h^1}{\mu_h^1}\right)^{n_h^1}}{1 + \sum_{j=1}^{n_h^1} \frac{1}{j!}\left(\frac{\lambda_h^1}{\mu_h^1}\right)^j} \tag{17}$$

Here only λ_{hs}^1 is shown since expressions for λ_{ns}^1, λ_{hs}^2, and λ_{ns}^2 are similar.

From the perspective of the shared partition, the arrival rates are thus λ_{hs}^1, λ_{ns}^1, λ_{hs}^2 and λ_{ns}^2 and the total number of channels available is C_s with all other parameters remained the same. Hence we compute the revenue generated per unit time from the hybrid admission control algorithm to the cell by the sum of revenue earned from the fixed partitions plus that from the shared partition:

$$\begin{aligned} \mathrm{HR}\left(C, \lambda_h^1, \lambda_n^1, \lambda_h^2, \lambda_n^2\right) = {} & \mathrm{PR}\left(C - C_s, \lambda_h^1, \lambda_n^1, \lambda_h^2, \lambda_n^2\right) \\ & + \mathrm{TR}\left(C_s, \lambda_{hs}^1, \lambda_{ns}^1, \lambda_{hs}^2, \lambda_{ns}^2\right) \end{aligned} \tag{18}$$

The optimization problem for the hybrid algorithm is to identify the best partition (C_h^1, C_n^1, C_h^2, C_n^2, C_s) that would maximize the cell's revenue subject to the imposed QoS constraints defined by Conditions 3 through 6.

5. Numeric Data and Analysis

In this section we report numerical data obtained from applying Equations 8, 12 and 18 derived for partitioning, threshold-based and hybrid admission control algorithms for revenue optimization with QoS guarantees and compare their performance characteristics with physical interpretations of the results. The charging rate model is based on the popular "charge-by-time" scheme for which a call is charged by time with a fixed rate per time unit. The analysis considers two classes with class 1 (real-time) demanding more resources with higher QoS constraints than class 2 (non-real-time), so class 1 has a higher charging rate per unit time and more stringent thresholds on both the new and handoff call blocking probabilities than class 2.

The input parameters are C, λ_h^1, μ_h^1, λ_n^1, μ_n^1, λ_h^2, μ_h^2, λ_n^2, μ_n^2, v^1, v^2, k^1, k^2, B_h^1t, B_h^2t, B_n^1t, and B_n^2t. We set $C = 80$, $k^1 = 4$ and $k^2 = 1$ for a typical cell in mobile wireless networks to service realtime and non-real-time traffic such that there are 80 channels in the cell with a class 1 call (realtime) consuming 4 channels and a class 2 call (non-real-time) consuming 1 channel. We vary the values of other model parameters, such as the arrival rates of new/handoff calls for different classes, pricing values (v^1 vs v^2), and threshold blocking probabilities (B_h^1t and B_h^2t) to analyze their effects on the maximum revenue obtainable subject to the QoS constraints specified in terms of Conditions 3 through 6 being satisfied.

Table 1 compares the optimal revenue obtained per unit time while the QoS constraints specified are satisfied, under partitioning, threshold-based and hybrid admission control algorithms at optimal settings as a function of the high-priority handoff call arrival rate λ_h^1, with all other parameter values being listed at the bottom of the table. The corresponding optimal (C_h^1, C_n^1, C_h^2, C_n^2) settings under partitioning, optimal (C_{hT}^1, C_{nT}^1, C_{hT}^2, C_{nT}^2) settings under threshold-based and (C_h^1, C_n^1, C_h^2, C_n^2, C_s) settings under hybrid admission control algorithms are also listed in the table to reveal the trend exhibited in resource allocation by these algorithms. One should note that a difference of even 1 unit of revenue per unit time earned by the system as a result of adopting different admission control algorithms could be considered significant because the revenue accumulated over a period of time would be significant.

The data in Table 1 indicate that as λ_h^1 increases, the revenue rate obtainable also increases as long as the QoS constraints can still be satisfied given the amount of resources available ($C = 80$). Nevertheless, as λ_h^1 increases further past a threshold value, all algorithms eventually fail to yield a legitimate solution because the workload is too heavy to satisfy the imposed QoS constraints, as indicated in the table by “None”. One can see that hybrid admission control is the most tolerant algorithm among all in terms of being able to yield a solution under high workload situations, followed by threshold-based and partitioning.

We see that in response to a high arrival rate of λ_h^1, hybrid admission control (in the middle column) tends to increase the size of two partitions, that is, it tends to increase C_h^1 to satisfy the stringent QoS constraint of $B_h^1t = 0.02$ for class 1 handoff calls, and it also tends to increase C_s to exploit the multiplexing power of the shared partition by means of threshold-based admission control to satisfy QoS constraints of all other service calls. The multiplexing power of the shared partition is clearly demonstrated by the fact that hybrid significantly outperforms partitioning in terms of revenue obtainable over a range of λ_h^1 values, while being able to sustain a higher workload of λ_h^1 and provide QoS guarantees. As the arrival rate of class 1 handoff calls increases, on the other hand, threshold-based admission (in the last column) control tends to decrease the threshold values of C_{nT}^1, C_{hT}^2 and C_{nT}^2 while keeping C_{hT}^1 as high as possible to satisfy the stringent QoS constraint of $B_h^1t = 0.02$. We observe that the performance of threshold-based admission control is comparable to hybrid admission control until λ_h^1 becomes high enough, beyond which threshold-based performs significantly worse and eventually fails to yield a legitimate solution compared with hybrid admission control. We attribute the superiority of hybrid admission control over partitioning and threshold-based admission control to the ability to optimally reserve dedicated resources for high-priority classes through fixed partitioning to reduce interference from low-priority classes, and to optimally allocate resources to the shared partition in accordance with threshold-based admission control to exploit the multiplexing power for all classes.

Table 1. Comparing partitioning, threshold-based and hybrid admission control for maximum revenue obtainable while satisfying QoS as a function of λ_h^1.

	Partitioning		Hybrid		Threshold-based	
λ_h^1	$(C_h^1, C_n^1, C_h^2, C_n^2)$	Revenue/Time	$(C_h^1, C_n^1, C_h^2, C_n^2, C_s)$	Revenue/Time	$(C_{hT}^1, C_{nT}^1, C_{hT}^2, C_{nT}^2)$	Revenue/Time
1	(16, 56, 4, 4)	577.391	(8, 32, 0, 0, 36)	580.000	(80, 80, 80, 80)	579.95
1.5	(20, 52, 4, 4)	615.486	(12, 36, 0, 0, 32)	620.000	(80, 80, 80, 80)	619.88
2	(20, 52, 4, 4)	652.304	(12, 32, 0, 0, 36)	659.997	(80, 80, 80, 80)	659.75
2.5	(28, 44, 4, 4)	686.660	(16, 32, 0, 0, 32)	699.986	(80, 80, 80, 80)	699.485
3	(32, 40, 4, 4)	717.032	(16, 32, 0, 0, 32)	739.949	(80, 80, 80, 80)	739.023
3.5	(32, 40, 4, 4)	754.215	(16, 28, 0, 0, 36)	779.842	(80, 80, 76, 76)	778.258
4	None	None	(16, 28, 0, 0, 36)	819.565	(80, 80, 76, 76)	817.058
4.5	None	None	(20, 24, 0, 0, 36)	858.998	(80, 80, 76, 76)	855.266
5	None	None	(20, 24, 0, 0, 36)	897.974	(80, 80, 76, 76)	892.708
5.5	None	None	(20, 24, 0, 0, 36)	936.137	(80, 80, 76, 76)	929.203
6	None	None	(20, 20, 0, 0, 40)	973.303	(80, 80, 76, 76)	964.569
6.5	None	None	(20, 20, 0, 0, 40)	1009.098	(80, 76, 75, 72)	992.917
7	None	None	(24, 20, 0, 0, 36)	1043.262	None	None
7.5	None	None	(24, 20, 0, 0, 36)	1075.786	None	None

$C = 80$, $\mu_h^1 = 1.0$, $\lambda_n^1 = 6.0$, $\mu_n^1 = 1.0$, $\lambda_h^2 = 1.0$, $\mu_h^2 = 1.0$, $\lambda_n^2 = 1.0$, $\mu_n^2 = 1.0$, $v^1 = 80$, $v^2 = 10$, $k^1 = 4$, $k^2 = 1$, $B_h^1 t = 0.02$, $B_h^2 t = 0.04$, $B_n^1 t = 0.05$, $B_n^2 t = 0.1$.

Table 2. Comparing partitioning, threshold-based and hybrid admission control for maximum revenue obtainable while satisfying QoS as a function of λ_h^2 and λ_n^2.

	Partitioning		Hybrid		Threshold-based	
λ_h^2 & λ_n^2	$(C_h^1, C_n^1, C_h^2, C_n^2)$	Revenue/Time	$(C_h^1, C_n^1, C_h^2, C_n^2, C_s)$	Revenue/Time	$(C_{hT}^1, C_{nT}^1, C_{hT}^2, C_{nT}^2)$	Revenue/Time
1	(48, 24, 4, 4)	576.382	(32, 16, 0, 0, 32)	580.000	(80, 80, 80, 80)	579.952
2	(44, 24, 6, 6)	594.268	(28, 12, 0, 0, 40)	599.999	(80, 80, 80, 80)	599.917
3	(44, 20, 8, 8)	610.326	(28, 12, 0, 0, 40)	619.998	(80, 80, 80, 80)	619.855
4	(44, 20, 8, 8)	628.380	(28, 12, 1, 1, 38)	639.993	(80, 80, 80, 80)	639.755
5	(40, 20, 10, 10)	644.636	(28, 12, 2, 2, 36)	659.977	(80, 80, 80, 80)	659.593
6	(40, 20, 11, 9)	660.886	(24, 8, 2, 2, 44)	679.937	(80, 80, 80, 80)	679.338
7	None	None	(24, 8, 2, 2, 44)	699.854	(80, 80, 80, 80)	698.948
8	None	None	(24, 8, 3, 3, 42)	719.675	(80, 80, 80, 80)	718.365
9	None	None	(20, 8, 3, 3, 46)	739.321	(80, 80, 76, 76)	737.525
10	None	None	(20, 8, 3, 3, 46)	758.708	(80, 80, 76, 76)	756.341
11	None	None	(20, 8, 4, 4, 44)	777.650	(80, 80, 76, 76)	774.714
12	None	None	(20, 4, 3, 3, 50)	795.995	(80, 80, 76, 76)	792.533
13	None	None	(16, 4, 3, 3, 54)	813.654	(80, 80, 76, 76)	809.685
14	None	None	(16, 4, 3, 3, 54)	830.339	(80, 80, 76, 76)	826.054
15	None	None	(16, 4, 3, 3, 54)	845.795	(80, 80, 76, 76)	841.533
16	None	None	(16, 4, 6, 6, 48)	859.543	(80, 80, 76, 75)	855.576
17	None	None	(12, 0, 2, 2, 64)	872.773	None	None

$C = 80$, $\lambda_h^1 = 5.0$, $\mu_h^1 = 1.0$, $\lambda_n^1 = 2.0$, $\mu_n^1 = 1.0$, $\mu_h^2 = 1.0$, $\mu_n^2 = 1.0$, $v^1 = 80$, $v^2 = 10$, $k^1 = 4$, $k^2 = 1$, $B_h^1 t = 0.02$, $B_h^2 t = 0.04$, $B_n^1 t = 0.05$, $B_n^2 t = 0.1$.

Next we test the sensitivity of the results with respect to the traffic load of class 2 (low-priority) calls Table 2 shows the revenue rate earned by the system as a function of λ_h^2 and λ_n^2. Here we see that as the arrival rate of the low-priority class increases (shown in the first column), hybrid admission control (shown in the middle column) tends to decrease the number of dedicated channels allocated to high-priority calls, while at the same increasing the number of shared channels to exploit the multiplexing power in the shared partition. The reason is that as the arrival rate of low-priority calls increases the system will gain most of its revenue from low-priority calls. Thus hybrid admission control tempts to allocate as much resources to low-priority calls as possible, to the extent that QoS constraints for both high and low priority calls are satisfied. Since the QoS constraint of high priority handoff calls is stringent (2% drop probability), we see from Table 2 that even when the arrival rate of low-priority class is very high, hybrid admission control still allocates some designated channels to the C_h^1 partition. In all cases we observe that hybrid admission control performs the best among all over a wide range of arrival rate of low-priority calls.

Two other important parameters significantly affect the behavior of admission control for revenue optimization with QoS guarantees. One is the ratio of v^1: v^2; the other is QoS constraints. We present sensitivity analysis of these two parameters below.

Table 3 shows the effect of v^1: v^2 (by varying v^1 while setting v^2 to 10). The results show that hybrid admission control outperforms or is at least as good as partitioning and threshold-based admission control. Moreover, we observe that the difference in revenue earned becomes more significant as the v^1: v^2 ratio increases. This effect is especially pronounced when the system is heavily loaded (shown in the bottom half of Table 3) under which it is necessary to optimally allocate channels to calls of different priority types and service charge rates to maximize the revenue earned by the system while at the same time satisfying the imposed QoS constraints.

Table 4 shows the effect of QoS constraints. We tighten the QoS constraints of handoff calls for both classes 1 and 2 (B_h^1t, and B_h^2t) by a multiplicative factor of 2 successively to see how these admission control algorithms would respond to the change. We observe that under light-load conditions (the first half of Table 4) all three algorithms can reasonably adapt to the QoS change in order to satisfy the QoS constraints. However, partitioning admission control generates relatively lower revenue because without multiplexing power it needs to trade revenue off for QoS satisfaction. When the QoS constraints of handoff calls become extremely tight, both partitioning and threshold-based admission control algorithms fail to provide a legitimate solution, while hybrid admission is still able to provide a legitimate solution due to its ability to exploit the multiplexing power in the shared partition and to reserve dedicated resources for individual service classes.

The adaptability of hybrid admission control with respect to more stringent QoS constraints is especially pronounced under heavy-load situations (shown at the bottom half of Table 4). In response to more stringent QoS constraints on handoff calls under heavy-load conditions, hybrid admission control allocates more channels in the C_h^1 partition and conversely fewer channels in the C_n^1 partition to satisfy the most stringent QoS constraint imposed on class 1 handoff calls. Further, it also allocates more channels in the shared partition to satisfy the stringent QoS requirement of class 2 handoff calls, which through multiplexing also has the benefit of compensating class 1 and class 2 new calls to satisfy their QoS constraints. We conclude that the channel allocation made by the hybrid admission control algorithm represents the best possible way to satisfy varying QoS requirements while maximizing revenue earned by the system.

Table 3. Comparing partitioning, threshold-based and hybrid admission control for maximum revenue obtainable while satisfying QoS as a function of v^1 to v^2 ratio.

	Partitioning		Hybrid		Threshold-based	
$v^1 : v^2$	$(C_h^1, C_n^1, C_h^2, C_n^2)$	Revenue/Time	$(C_h^1, C_n^1, C_h^2, C_n^2, C_s)$	Revenue/Time	$(C_{hT}^1, C_{nT}^1, C_{hT}^2, C_{nT}^2)$	Revenue/Time
			Low class 1 call arrival rates ($\lambda_h^1 = 1.0, \lambda_n^1 = 1.0$)			
1	(20, 16, 22, 22)	219.735	(12, 12, 5, 5, 46)	220.000	(80, 80, 80, 80)	220.000
2	(20, 16, 22, 22)	239.550	(12, 12, 5, 5, 46)	240.000	(80, 80, 80, 80)	240.000
4	(20, 20, 20, 20)	279.381	(16, 12, 5, 5, 42)	280.000	(80, 80, 80, 80)	280.000
8	(20, 20, 20, 20)	359.135	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
16	(20, 20, 20, 20)	518.645	(16, 16, 7, 5, 36)	520.000	(80, 80, 80, 80)	520.000
32	(20, 20, 20, 20)	837.663	(16, 16, 7, 5, 36)	840.000	(80, 80, 76, 76)	840.000
64	(24, 20, 18, 18)	1476.281	(16, 16, 7, 5, 36)	1480.000	(80, 80, 76, 76)	1480.000
128	(24, 24, 16, 16)	2754.232	(16, 16, 7, 5, 36)	2760.000	(80, 80, 76, 76)	2760.000
			High class 1 call arrival rates ($\lambda_h^1 = 3.5, \lambda_n^1 = 4.5$)			
1	None	None	(8, 12, 5, 5, 50)	278.919	(80, 80, 80, 80)	278.280
2	None	None	(12, 16, 4, 4, 44)	358.240	(80, 80, 80, 80)	357.129
4	None	None	(12, 16, 3, 3, 46)	516.987	(80, 80, 80, 80)	514.828
8	None	None	(12, 16, 2, 2, 48)	834.545	(80, 80, 76, 76)	830.611
16	None	None	(12, 16, 1, 1, 50)	1469.720	(80, 80, 72, 72)	1464.843
32	None	None	(12, 16, 0, 0, 52)	2747.443	(80, 80, 72, 69)	2736.794
64	None	None	(12, 16, 0, 0, 52)	5303.173	(80, 80, 71, 66)	5284.153
128	None	None	(12, 16, 0, 0, 52)	10416.435	(80, 80, 71, 66)	10380.503

$C = 80$, $\mu_h^1 = 1.0$, $\mu_n^1 = 1.0$, $\lambda_h^2 = 10.0$, $\mu_h^2 = 1.0$, $\lambda_n^2 = 10.0$, $\mu_n^2 = 1.0$, $v^2 = 10$, $k^1 = 4$, $k^2 = 1$, $B_h^1t = 0.02$, $B_h^2t = 0.04$, $B_n^2t = 0.05$, $B_n^2t = 0.1$.

Table 4. Comparison of partitioning, threshold-based and hybrid admission control for maximum revenue obtainable while satisfying qos as a function of QoS constraints on class 1 and class 2 handoff calls.

	Partitioning		Hybrid		Threshold-based	
$(B_h^1 t, B_h^2 t)$	$(C_h^1, C_n^1, C_h^2, C_n^2)$	Revenue/Time	$(C_h^1, C_n^1, C_h^2, C_n^2, C_s)$	Revenue/Time	$(C_{hT}^1, C_{nT}^1, C_{hT}^2, C_{nT}^2)$	Revenue/Time
			Low class 1 call arrival rates ($\lambda_h^1 = 1.0, \lambda_n^1 = 1.0$)			
$(0.02, 0.04) \times 2^{0}$	(20, 20, 20, 20)	359, 135	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-1}$	(20, 20, 20, 20)	359.135	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-2}$	(20, 20, 20, 20)	359.135	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-3}$	(24, 20, 20, 20)	358.345	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-4}$	(24, 20, 20, 20)	358.345	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-5}$	(24, 20, 21, 19)	358.264	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-6}$	(28, 16, 22, 14)	353.041	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
$(0.02, 0.04) \times 2^{-7}$	(28, 16, 23, 13)	350.311	(16, 12, 5, 5, 42)	360.000	(80, 80, 80, 80)	360.000
...	None	None	...	...	...	...
$(0.02, 0.04) \times 2^{-21}$	None	None	(16, 12, 5, 5, 42)	360.000	(80, 76, 76, 61)	359.991
$(0.02, 0.04) \times 2^{-22}$	None	None	(16, 12, 5, 5, 42)	360.000	(80, 76, 76, 54)	359.904
$(0.02, 0.04) \times 2^{-23}$	None	None	(16, 12, 5, 5, 42)	360.000	(80, 76, 76, 48)	359.409
$(0.02, 0.04) \times 2^{-24}$	None	None	(16, 12, 5, 5, 42)	360.000	(80, 76, 76, 42)	357.231
$(0.02, 0.04) \times 2^{-25}$	None	None	(16, 12, 5, 5, 42)	360.000	None	None
			High class 1 call arrival rates ($\lambda_h^1 = 3.5, \lambda_n^1 = 4.5$)			
$(0.02, 0.04) \times 2^{0}$	None	None	(12, 16, 2, 2, 48)	834.544	(80, 80, 76, 76)	830.610
$(0.02, 0.04) \times 2^{-1}$	None	None	(12, 16, 2, 2, 48)	834.544	(80, 80, 76, 76)	830.610
$(0.02, 0.04) \times 2^{-2}$	None	None	(20, 8, 1, 1, 50)	830.078	(80, 76, 76, 76)	826.208

$C = 80$, $\mu_h^1 = 1.0$, $\mu_n^1 = 1.0$, $\lambda_h^2 = 10.0$, $\mu_h^2 = 1.0$, $\lambda_n^2 = 10.0$, $\mu_n^2 = 1.0$, $v^1 = 80$, $v^2 = 10$, $k^1 = 4$, $k^2 = 1$, $B_n^1 t = 0.05$, $B_n^2 t = 0.1$.

6. Applicability and Summary

In this paper we have proposed and analyzed the design concept for the integration of pricing with admission control algorithms with QoS guarantees in a cellular wireless network. The design concept is based on the idea that an admission control algorithm in deciding which calls to admit should consider not only the QoS constraints imposed by the system, but also the revenue that the admission of such a call will bring to the system. In illustrating our concept, we test a "charge-by-time" pricing scheme being used by the service provider where a user is charged by the amount of time in service. Three admission control algorithms for handling multiple classes of traffic were proposed, namely, partitioning, threshold-based, and hybrid admission control with the intention of maximizing revenue generated by a cell while still satisfying the QoS constraints imposed by the system for distinct service classes. Our analysis results indicated that at optimizing conditions the hybrid admission control algorithm can generate higher revenue with QoS guarantees than the other two admission control algorithms. We attribute the superiority of the hybrid algorithm to the existence of fixed partitions reserved for specific classes/types to avoid interference from other classes/types so as to satisfy the respective QoS requirements, and a shared partition which provides great multiplexing power for sharing the bandwidth among calls of different classes and services. The hybrid algorithm encompasses both the partitioning and threshold-based algorithms as special cases.

To apply the results obtained in the paper, a cell dynamically communicates with mobile users in its cell and neighboring cells to obtain values of arrival and departure rates of new/handoff calls of various service classes periodically and performs a simple table look up at runtime to obtain the optimal $(C_h^1, C_n^1, C_h^2, C_n^2, C_s)$ under hybrid admission control for revenue optimization with QoS guarantees.

Some possible future research directions extending from this work include (a) considering other pricing models and investigating optimal resource allocation settings under which hybrid partitioning threshold-based admission control algorithms can yield the highest revenue with QoS guarantees; (b) considering other revenue collection model, e.g., revenue is collected only on call termination or revenue is lost when a call is terminated prematurely; (c) exploring the relationship between QoS and pricing and determining the optimal pricing for calls of various service classes and types such that the revenue is maximized with QoS guarantees based on anticipated workload conditions and resource availability.

Acknowledgement

This research work was supported by a National Science Foundation IGERT grant #9987586. We like to thank Naresh Verma for his contribution to the initial work of this paper.

References

1. Y. B. Lin and I. Chlamtac, *Wireless and Mobile Network Architecture*, John Wiley and Sons, 2001.
2. D. Hong and S. S. Rappaport, "Priority Oriented Channel Access for Cellular Systems Serving Vehicular and Portable Radio Telephones", *Communications, Speech and Vision, IEE Proceedings I.* Vol. 131, Issue 5, pp. 339–346, Oct. 1989.
3. D. Hong and S. S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Non-Prioritized Handoff Procedures", *IEEE Transactions on Vehicular Technology*, Vol. VT35. No.3, August 1986.

4. R. Guerin, “Queuing-Blocking Systems with Two Arrival Streams and Guarded Channels”, *IEEE Transactions on Communication*, Vol. 36, pp. 153–163, February 1988.
5. Y. Fang, “Thinning Algorithms for Call Admission Control in Wireless Networks”, *IEEE Transactions on Computers*, Vol. 52, No. 5, pp. 685–687, May 2003.
6. J. Wang, Q. Zeng and D.P. Agrawal, “Performance Analysis of a Preemptive and Priority Reservation Handoff Algorithm for Integrated Service-Based Wireless Mobile Networks”, *IEEE Transactions on Mobile Computing*, Vol. 2, No. 1, pp. 65–75, January–March 2003.
7. B. Li, C. Lin, and S.T. Chanson, “Analysis of a Hybrid Cutoff Priority Algorithm for Multiple Classes of Traffic in Multimedia Wireless Networks”, *Wireless Networks*, Vol. 4, pp. 279–290, 1998.
8. J. Ye, J. Hou, and S. Papavassilliou, “A Comprehensive Resource Management for Next Generation Wireless Networks”, *IEEE Transactions on Mobile Computing*, Vol. 1, No. 4, pp. 249–263, October–December 2002.
9. J. Hou, J. Yang and S. Papavassiliou, “Integration of Pricing with Call Admission Control to Meet QoS Requirements in Cellular Networks, ” *IEEE Trans. on Parallel and Distributed Systems*, Vol. 13, No. 9, pp. 898–910, Sept. 2002.
10. N.J. Keon and G. Anandalingam, “Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees, ” *IEEE/ACM Trans. on Networking*, Vol. 11, No. 1, pp. 66–80, Feb. 2003.
11. M. Kyriakakos, N. Frangiadakis, L. Merakos and S. Hadjiefthymiades, “Enhanced path prediction for network resource management in wireless LANs, ” *IEEE Wireless Communications*, Vol. 10, No. 6, pp. 62–69, Dec. 2003.