Integration of Pricing with Call Admission Control for Wireless Networks

Jiongkuan Hou, Jie Yang and Symeon Papavassiliou
New Jersey Center for Wireless Telecommunications
Department of Electrical and Computer Engineering
New Jersey Institute of Technology, Newark, NJ 07102, USA
jxh0350@oak.njit.edu, jxy9918@njit.edu, Papavassiliou@adm.njit.edu

Abstract— **Traditional CAC schemes that mainly focus on the trade off between new call blocking probability and handoff call blocking probability can not solve the problem of congestion in wireless networks. In this paper we investigate the role of pricing as an additional dimension of the call admission control process in order to efficiently and effectively control the use of wireless network resources. First we prove that for a given wireless network there exists a new call arrival rate which can maximize the total utility of users. Based on this result and observation we propose an integrated pricing and call admission control scheme, where the price is adjusted dynamically based on the current network conditions, in order to alleviate the problem of congestion. We compare the performance of our approach with the corresponding results of conventional systems where pricing is not taken into consideration in the call admission control process. These performance results verify the considerable improvement that can be achieved by the integration of pricing in the call admission control process in cellular networks.**

Keywords—**Call Admission Control, Pricing, Wireless Networks**

1 Introduction

THE rising demand for mobile communication services is increasing the importance of efficient use of the limited bandwidth and frequency spectrum. In recent years considerable efforts have focused on the Channel Allocation and Call Admission Control (CAC) problems and many schemes that range from static to dynamic strategies have been proposed in literature [3, 6]. Call Admission Control is a provisioning strategy used to limit the number of call connections into the networks in order to reduce the network congestion and provide the desired Quality of Service (QoS) to users in service.

New call blocking probability and handoff call blocking probability are two important connection level QoS parameters. Handoff calls are commonly given a higher priority since a call being forced to terminate during the service is more annoying than a call being blocked at its start. Various handoff priority-based CAC schemes have been proposed in literature including *Guard Channel Schemes*, *Queuing Priority Schemes* and *Channel Borrowing Schemes*. These research efforts have focused on how to adjust the tradeoff between new call blocking probability and handoff call blocking probability. Within a certain dynamic range of call arrival rate, these schemes can improve the system performance. However, we can observe from the results presented by these research efforts that with the increase of call arrival rate, both the new call blocking probability and the handoff call blocking probability increase. When the call arrival rate is temporarily very high (for example in busy hours), no matter how the parameters are adjusted, these schemes can not guarantee the QoS to users.

The main reason of degradation of QoS stems from the fact that resources in a wireless network, such as timeslots, code and power, are shared by all the users. When one user is admitted into the network, it will cause QoS degradation to other users. In general we can observe that the most serious QoS violation occurs when the system is congested. However, the current CAC schemes can not avoid congestion, because they do not provide incentives for users to use the channel resources effectively. In broadband networks, pricing schemes are widely discussed as means for traffic management and congestion control [1, 4]. Through pricing, the network can send signals to the users, providing incentives that influence their behavior. This provides another dimension for the design of CAC schemes that can be used in wireless networks as well. In this paper we integrate pricing with CAC to address the problem of congestion.

The remaining of this paper is organized as follows. Section 2 provides the description of the proposed integrated pricing and call admission control scheme, while section 3 includes the performance evaluation of our proposed approach and scheme.

2 Integrating Pricing and Call Admission Control

Network users act independently and sometimes "selfishly", without considering the current network traffic conditions. Hence system overload situations are unavoidable. If each user requests the resources that maximize his/her individual level of satisfaction, the total utility of the community will decrease, so that there must be some mechanism to provide incentives for users to behave in ways that improve overall utilization and performance. In commercial networks, this can be most effectively achieved through pricing.

0-7803-7005-8/01/$10.00

Network pricing has recently been embraced by researchers in the multi-service broadband networks [1, 4] not only as an economic issue and element, covering the infrastructure expenses and operational expenses through charging the end users, but also as a resource management issue. The aggregate traffic load on a wireless network is the result of many users' individual decisions about whether and how to use the network, and these decisions are affected by the incentives these users encounter when using the resources of a wireless network. These incentives can take many forms; one of the most important incentives is the monetary incentive [1]–raising unit price that could make some of the users request less resources.

2.1 Optimal New Call Arrival Rate

If we consider the wireless network resources as a public good, the best policy to share this good is the one that can maximize the total user utility [4]. In terms of economics, *utility functions* describe users' level of satisfaction with the perceived Quality of Service [1, 4]; the higher the utility, the more satisfied the users. In general, utility function characterizes how sensitive users are to the changes in QoS. It is sometimes useful to view the utility functions as of money a user is willing to pay for certain QoS. Some utility functions have been suggested in literature in order to model customer behavior and evaluate pricing policies. For example, in [1], Cocchi *et al.* proposed utility function for Email applications to be a decreasing function of both average delay and the percentage of messages not delivered within a delay bound of five minutes. In this paper we define utility function as function of call blocking probabilities, which represent the main QoS metrics in cellular networks

In this section, we prove that there exists a new call arrival rate where the total user utility is maximized and therefore the network resources are optimally utilized. We consider a wireless network that carries out Guard Channel CAC scheme; the arrival process of new calls is assumed to be Poisson and the channel holding time is assumed to have negative exponential distributions. The parameters of the system are given, including the total number of channels, the number of guard channels, the average new call channel holding time and average handoff call channel holding time, so that the performance of the system depends on the new call arrival rate (λ_n) and handoff call arrival rate (λ_h) [3]. Lin *et al.* also proved in [6] that handoff call arrival rate is a function of new call arrival rate and other system parameters. Therefore, in the following we study how the total utility changes with the change of new call arrival rate. Our analysis is based on the following definitions, observations and assumptions.

Definition 1. We define the average number of admitted users (N) as a function of new call arrival rate, i.e. $N = f(\lambda_n)$. $f(\lambda_n)$ is a differentiable and monotonically increasing concave function of λ_n [5]. Therefore:

$$f(\lambda_n) \geq 0; \qquad f'(\lambda_n) > 0; \qquad f''(\lambda_n) < 0 \qquad (1)$$

Definition 2. We define the Quality of Service metric P_b as a weighted sum of new call blocking probability (P_{nb}) and handoff call blocking probability (P_{hb}):

$$P_b = \alpha P_{nb} + \beta P_{hb} \qquad (2)$$

where α and β are constants that denote the penalty associated with rejecting new calls and handoff calls respectively, with $\beta > \alpha$ to reflect the higher cost to block a handoff call. Because both P_{nb} and P_{hb} are monotonically increasing convex functions of λ_n [5], P_b is also a monotonically increasing convex function of λ_n.

$$P_b = g(\lambda_n) \qquad \text{with} \qquad (3)$$

$$g(\lambda_n) > 0; \qquad g'(\lambda_n) > 0; \qquad g''(\lambda_n) > 0 \qquad (4)$$

In the following, we describe the general properties and characteristics of the user utility function. As mentioned before, utility function models network users' preference. We argue that when P_b increases, users will suffer more call blockings and the level of user satisfaction decreases. We also note that when P_b is small, the satisfaction degradation caused by the increase of P_b is not significant; as P_b becomes large, the satisfaction degradation will be substantial. Therefore, throughout this paper we make the following assumption:

Assumption 1. The utility function of a single user (U_s) is a differentiable and monotonically decreasing concave function of the QoS parameter P_b. That is:

$$U_s = h(P_b) \qquad \text{with} \qquad (5)$$

$$h(P_b) > 0; \qquad h'(P_b) < 0; \qquad h''(P_b) < 0 \qquad (6)$$

Note that U_s achieves maximum value at $P_b = 0$, which means that if the blocking probability is 0% the user has the highest level of satisfaction. Moreover, although different applications may have different QoS requirements and therefore different utility functions, without loss of generality we can assume that there exists a P_b^{max} such that $U_s(P_b) = 0$ for all $P_b \geq P_b^{max}$. This means that when call blocking probability is very high, the user satisfaction is zero. In a realistic wireless system P_b^{max} represents the threshold value (maximum) of P_b that can be tolerated so that the Quality of Service is considered acceptable. Based on the above definitions and assumptions we can prove the following theorem:

Theorem 1 For a given wireless network, there exists an optimal new call arrival rate that maximizes the total utility U, where U is defined as:

$$U = N \times U_s = f(\lambda_n) \times h[g(\lambda_n)] \qquad (7)$$

Proof. **(i)**

$$U(\lambda_n = 0) = 0 \qquad (8)$$

This is the case that no user is in the system, i.e. $N = 0$, hence the total utility is zero;

(ii)

$$U(\lambda_n = \lambda_n^{max}) = 0 \qquad (9)$$

where λ_n^{max} is the new call arrival rate that satisfies $g(\lambda_n^{max}) = P_b^{max}$, at which point single user's utility becomes zero. This corresponds to the case that the network is congested because a large number of users are competing for the limited channel resources and as a result no user is satisfied with the QoS.

(iii) From the above definitions and assumption, we have

$$\frac{d^2U}{d\lambda_n^2} < 0 \tag{10}$$

which means that U is a concave function of new call arrival rate λ_n.

From (i) (ii); and by Rolle Theorem we conclude that:

$$\exists\lambda_n^* \quad \cdot \quad \frac{dU}{d\lambda_n}(\lambda_n^*) = 0 \tag{11}$$

Combining this result with equation (10), we can sufficiently ensure a maximum of U at $\lambda_n = \lambda_n^*$. □

It should be noted here that maximum total user utility also means that channel resources are most efficiently utilized. When $\lambda_n < \lambda_n^*$, users can get a better quality than their QoS requirements, but some channel resources are wasted and from the perspective of the service provider, this means less revenue. On the other hand, when $\lambda_n > \lambda_n^*$, a large number of users are blocked when trying to initiate their calls or when trying to handoff to another cell in the middle of a call, which means that the QoS degrades and may become unacceptable. In this case, although on average, more channels are used, because of the increasing handoff failures, it is hard for a user to finish his/her call successfully and as a result the "effective" utilization of channel resources is low. Therefore, we can say that $\lambda_n = \lambda_n^*$ is a point where the number of satisfied users is maximized and channel resources are most efficiently used. When $\lambda_n > \lambda_n^*$, both the total user utility and the QoS decrease with any further increase of λ_n and we may say that the cell enters the congestion state. From the view point of QoS guarantee, it is ideal for a system to operate at the optimal traffic load (λ_n^*) or below.

2.2 System Model

In current wireless networks users are charged with fixed rate or based on the time of the day. The major advantage of these schemes is that the billing and accounting processes are simple. However, the price is independent of the current state of the network. Such systems can not provide enough incentives for users to avoid congestion, and furthermore can not react effectively to the dynamic and sometimes unpredictable variation of the network usage and conditions. This paper proposes a new scheme which integrates the congestion pricing with call admission control to address this problem. Figure 1 provides a schematic representation of the proposed approach and model.

The system is composed of two functional blocks: Pricing block and CAC block. Here we use a guard channel scheme at the CAC block. The pricing block works as follows: when the traffic load is less than the optimal value, $\lambda_n < \lambda_n^*$, a *normal price* is charged to each user. The normal price is the price that is acceptable to every user. When the traffic load increases beyond the optimal value, dynamic *peak hour price* will be charged to users who want to place their calls at this time. It should be noted here that according to our scheme the decision about the peak hour fee is based on the network conditions. This means that the price is continuously and dynamically adjusted according to the changes in the system condition as the system evolves. During the period that dynamic peak hour price is charged to users, if some users are not willing to accept the extra charge, they will choose not to place their calls at this time. These users can make their calls later when the network conditions change and the price decreases. This generates another traffic stream to the pricing block–the retry traffic, whose arrival rate is denoted by λ_r in figure 1.

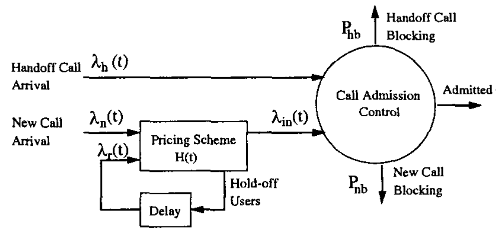

Figure 1. Integration of pricing scheme with call admission control

2.3 Setting Price According to Traffic Load

We define the system function of pricing block ($H(t)$) as the percentage of the incoming users that will accept the price at time t, i. e.

$$(\lambda_n(t) + \lambda_r(t))H(t) = \lambda_{in}(t) \tag{12}$$

where $\lambda_{in}(t)$ is the rate of input traffic to CAC block. The congestion pricing block should be designed in such a way that by adjusting $H(t)$ according to current traffic condition, $\lambda_{in}(t)$ always meets the following requirement:

$$\lambda_{in}(t) \leq \lambda_n^* \tag{13}$$

where λ_n^* is the optimal new call arrival rate we obtained in section 2.1. This requirement guarantees that the cell will not be congested and therefore the quality of service of the callers in service can be guaranteed.

As mentioned before, monetary incentive can influence the way that users use resources and is usually characterized by demand functions. Demand function describes the reaction of users to the change of price. Different demand functions have been proposed in literature. In this paper we use the following demand function [2].

$$D[p(t)] = e^{-(\frac{p(t)}{p_0}-1)^2} \qquad p(t) \geq p_0 \tag{14}$$

where p_0 is the normal price, $p(t)$ is the price charged to users at time t which is the sum of normal price and extra peak hour price (if applicable). $D[p(t)]$ denotes the percentage of users that will accept this price. The control function of $H(t)$ is realized by users' reaction to the current price, therefore we have:

$$H(t) = D[p(t)] \tag{15}$$

Combining equations (12) through (15) we have

$$p(t) = D^{-1}\left(\min(\frac{\lambda_n^*}{\lambda_n(t)+\lambda_r(t)}, 1)\right) \qquad (16)$$

This is the price that should be set at time t in order to obtain the desired QoS.

3 Performance Analysis

In this section we evaluate the performance of the proposed integrated pricing and call admission control in terms of congestion prevention. We observe that our proposed integrated scheme achieves to alleviate the system congestion occurrences and meet the QoS requirements of the users in service, while other conventional CAC schemes fail to do so.

In section 3.1 we describe in detail the basic assumptions about the system under consideration. In section 3.2 we compare the results of conventional guard channel scheme and the proposed integrated scheme.

3.1 Model and Assumptions

The parameters used throughout our performance evaluation are as follows:

(1). Each cell is assigned $C = 40$ channels, and 2 of them are used as guard channels.

(2). Each call requires only one channel for service.

(3). The arrival of new calls initiating in each cell forms a Poisson process with rate $\lambda_n(t)$. The variation of new call arrival rate during a 24-hour period used through out our study is shown in figure 2.

(4). For both new calls and handoff calls, the call duration times are exponentially distributed with mean 240 seconds, and the cell dwell times are also exponentially distributed with mean 120 seconds.

(5). Parameters α and β in equation (2) are set to be $\frac{1}{3}$ and $\frac{2}{3}$ respectively, which means that we treat handoff calls twice more important than new calls.

(6). In the following numerical study, we use the following utility function

$$U_s = \begin{cases} 1 - e^{30(P_b - 0.1)} & \text{when } 0 \le P_b \le 0.01 \\ 0 & \text{when } P_b > 0.01 \end{cases} \qquad (17)$$

The optimal new call arrival rate for this system is $\lambda_n^* = 0.12$ call/sec and at this point the QoS metric is $P_b = 1\%$ Based on the analysis in section 2.1, this is the optimal operation point of the system in the sense that at this point, the total user utility can be maximized given that the QoS requirement (e. g. $P_b \le 0.01$) is met.

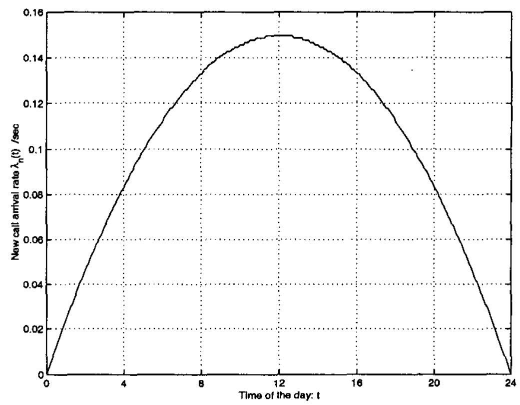

Figure 2. Input new call arrival rate as function of time

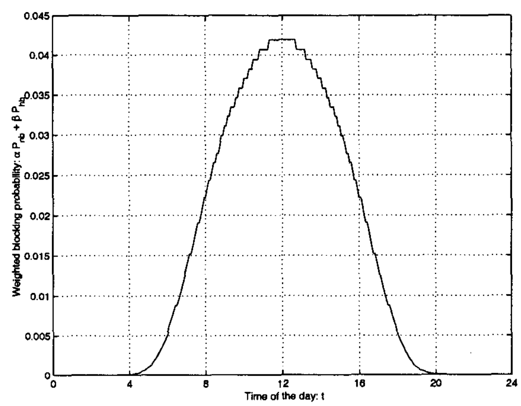

Figure 3. Call blocking probability of conventional system

3.2 Numerical Results and Discussion

Figure 3 shows the results of conventional system that do not use pricing in the call admission control process to control the traffic. From this figure, we can find that when traffic load is heavy (e. g. in noon hours), P_b can be as high as 4.2%. This values is far beyond users' minimum QoS requirement 1%, and therefore we can conclude that cells are seriously congested during this period.

The corresponding results of the proposed scheme are shown in figures 4 through 6. Figure 4 shows how the price is adjusted according to the change of offered traffic load. For the given new call arrival variation, when the offered traffic load into the pricing block is more than the optimal new call arrival rate (6:30AM to 11:00PM in figure 4(a)), the ratio $p(t)/p_0$ becomes more than 1, which means that the pricing mechanism comes into effect and the peak hour prices are charged to users. The heavier the traffic load, the higher the price, so that the less the percentage of users that would like to access the network, as suggested by equation (16). In our scheme, there is no central control mechanism to determine which user can access the channel resources.

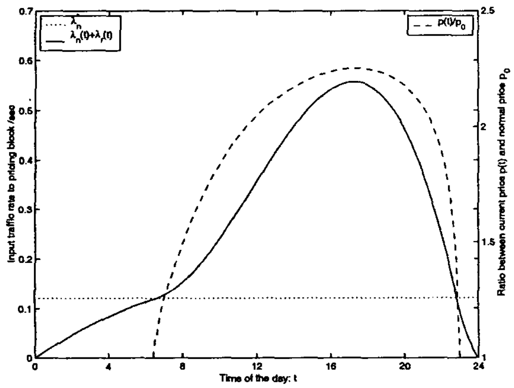

Figure 4. Set price according to traffic condition

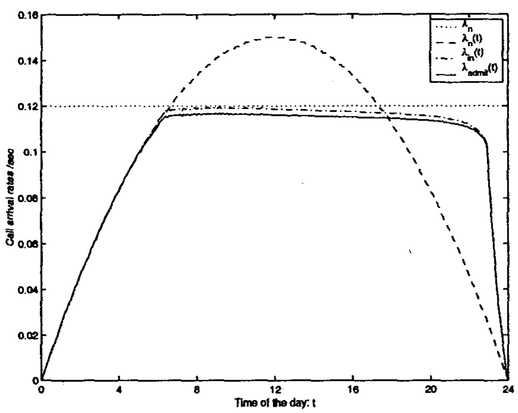

Figure 5. Traffic rates at different points

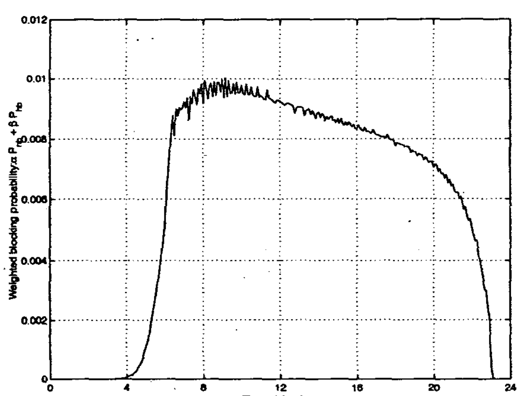

Figure 6. Call blocking probability for proposed scheme

Each base station just sets the price according to current traffic load of the cell, and it is the individual user's decision on whether to accept this price or not that controls the input traffic load to the system at this time. This implicitly implements a distributed user-based prioritization scheme where the priority is set by user's reaction to current price.

Figure 5 shows the traffic rates at different points of the system.From this figure we observe that the inputs to the CAC block are always lower than the optimal rate, i. e. $\lambda_{in}(t) < \lambda_n^*$, which means that the cell is not congested. The reason is that we adjust the price based on the user demand function and current traffic load (equation (16)) so that the price is always set to the appropriate value to guarantee that the traffic rate going through the pricing block is less than the optimal value. This result is justified by figure 6. From this figure we observe that the weighted call blocking probabilities are always lower than 1%, which means the QoS of the users who accept current price can be guaranteed. Comparing figure 5 with figure 2 we observe that our pricing block works like a traffic shaper, which can "move" part of the peak hour (6:30AM to 6:00PM in figure 5 (a)) traffic to relatively idle hours that follow the peak (6:00PM to 11:00PM in figure 5 (a)). The traffic being "moved" is composed of users that will not accept peak hour price.

References

[1] R. Cocchi, S. Shenker, D. Estrin and L. Zhang, "Pricing in Computer Networks: Motivation, Formulation and Example," *IEEE/ACM Transactions on Networking*, Vol. 1, No. 6, December 1993.

[2] P. C. Fishburn and A. M. Odlyzko, "Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet," *ICE'98*, pp. 128-139

[3] J. Hou and Y. Fang, "Mobility-based Channel Reservation Scheme for Wireless Mobile Networks," *Proceedings of IEEE WCNC 2000*, Chicago, September 2000.

[4] H. Ji, J. Y. Hui and E. Karasan, "GoS-Based Pricing and Resource Allocation for Multimedia Broadband Networks," *Proceedings of IEEE INFOCOM 1996*, pp. 1020-1027.

[5] K. R. Krishnan, "The Convexity of Loss Rate in an Erlang Loss System and Sojourn in an Erlang Delay System with Respect to Arrival rate and Service Rate," *IEEE Transactions on Communications*, Vol. 38, No. 9, September 1990.

[6] Y. B. Lin, S. Mohan and A. Noerepel, "Queueing Priority Channel Assignment Strategies for PCS Hand-Off and Initial Access," *IEEE Transactions on Vehicular Technology*, Vol. 43, No. 3, August 1994.