Soft QoS in Call Admission Control for Wireless Personal Communications

CHI WA LEONG and WEIHUA ZHUANG
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario,
Canada N2L 3G1
E-mail: {cleong, wzhuang}@bbcr.uwaterloo.ca

Abstract. In wireless multimedia communication systems, call admission control (CAC) is critical for simultaneously achieving a high resource utilization efficiency and maintaining quality-of-service (QoS) to mobile users. User mobility, heterogeneous nature of multimedia traffic, and limited radio spectrum pose significant challenges to CAC. QoS provisioning to both new calls and handoff calls comes with a cost of low resource utilization. This paper proposes a CAC policy for a wireless communication system supporting integrated voice and data services. In particular, soft QoS (or relaxed target QoS) is incorporated in the CAC policy to make compromises among different objectives. Numerical results are presented to demonstrate that (a) in dealing with the dilemma between QoS satisfaction and high resource utilization, how the resource utilization efficiency can be increased by introducing soft QoS; and (b) in accommodating different types of traffic, how the QoS of low priority traffic can be improved by specifying soft QoS to high priority traffic.

1. Introduction

Future wireless personal communication systems will provide multimedia communication services to mobile users anytime and anywhere [1], with guaranteed quality-of-service (QoS). QoS provisioning in a wireless environment is very challenging, due to limited radio spectrum, user mobility, and the heterogeneous nature of multimedia traffic. Call admission control (CAC) is the first control function imposed on the users and is, therefore, critical for maintaining a certain level of QoS to the users. The decision process of CAC can often be formulated in a concise representation called the CAC policy. The CAC policy consists of steps that the system follows in order to determine whether or not to accept the connection request from a user. Especially, the CAC policy takes call requests as input, and makes the decision based on the current traffic conditions of the system. When a user requests a new connection, the CAC process is to calculate the amount of resources required by (a) the users already in the system, and (b) the pending user. If the sum of the two is not larger than the total system resources, then the user will be admitted to the system; otherwise, the request will be rejected.

With the use of the cellular structure for radio frequency reuse, there are two major types of calls that can arrive at any particular cell: new calls originating from inside and handoff calls coming from outside. From the users' point of view, it is better to be blocked in the beginning rather than dropped in the middle of a connection. As a result, handoff calls should be given a higher priority than new calls. To do so, the usual approach is to reserve some resources exclusively for handoff calls [2–4]. To calculate the right amount of reserved re-

sources requires knowledge of the mobility information such as the users' traveling directions and speeds and the current traffic load in the surroundings. Furthermore, CAC should take into account the behavior of multimedia traffic. Two major classes of multimedia traffic are considered here: constant-rate high priority traffic for real-time applications such as voice and video, and available-rate low priority traffic carrying non real-time information such as data. As the radio spectrum is very limited, the objective of CAC is to achieve a high radio resource utilization and, at the same time, to ensure QoS satisfaction for different applications from different mobile users. Due to the random nature of multimedia traffic and the characteristics of a wireless mobile environment, the available resources in a base station can change dramatically as a mobile user enters or leaves the cell coverage area. As a result, simultaneously providing consistent QoS satisfaction to mobile users with heterogeneous requirements and achieving maximal utilization of system resources is not realistic. Certain compromises between the two objectives have to be made. In order to achieve a good compromise, this paper proposes a CAC policy that is adaptive to system traffic load conditions and to user application characteristics. The emphasis is placed on efficient utilization of system resources while guaranteeing *soft* QoS to mobile users.

At the call level, QoS is represented by a set of QoS parameters such as new call blocking probability and handoff call dropping probability. Normally, QoS requirement is specified by a value for each of the QoS parameters and is referred to as hard QoS requirement. On the other hand, when the QoS requirement is specified by a range (instead of a single value) for each of the QoS parameters, guaranteeing the QoS requirement is referred to as soft QoS provisioning. When the most stringent QoS requirement in soft QoS is the target service quality in the hard QoS, the soft QoS is actually relaxed hard QoS. A concept similar to soft QoS was introduced in [5], where an "adaptive reserved service" framework was proposed to support mobile connections. If mobile users can tolerate a certain degree of fluctuations in the service quality, then the system can exploit the QoS flexibility to improve resource utilization by adaptive resource allocation based on the instantaneous traffic load [6], wireless link quality, user application characteristics, etc. It should be noted that provisioning soft QoS (or relaxed hard QoS) is different from setting hard QoS at a lower level (e.g., the least stringent requirement in the soft QoS). In provisioning hard QoS, there is no adaptive mechanism in resource allocation; for example, when traffic load is low, available resources are not utilized for better QoS and are wasted. Even though it is quite obvious that relaxing QoS constraints can improve system resource utilization, in-depth investigation is necessary to evaluate the tradeoff between service quality and resource utilization.

In this paper, we investigate the effectiveness of using soft QoS in CAC for wireless communications. Our emphasis is placed on the relationship between resource utilization and soft QoS, and the relationship among the QoS parameters of the different traffic classes. In particular, we will study

(a) the relationship between soft QoS and resource utilization for a wireless system with constant-rate traffic. Numerical results are presented to demonstrate how much increase in resource utilization can be obtained by introducing soft QoS;

(b) the effectiveness of introducing soft QoS to a system with both constant-rate traffic and available-rate traffic. Numerical results are presented to demonstrate how much improvement in the QoS of the low priority traffic can be obtained when the QoS of the high priority traffic is relaxed.

The remainder of this paper is organized as follows. Section 2 describes the system model with the constant-rate and available-rate traffic, briefly reviews the limited fractional guard channel policy (LFGCP), presents the CAC policy proposed for the mixed traffic, and analyzes the performance of the CAC policy. Using soft QoS in the CAC is then discussed in Section 3, where how soft QoS can improve resource utilization and how soft QoS specified to constant-rate traffic can improve the QoS of available-rate traffic are investigated. Finally, Section 4 concludes this research.

2. The System Model and CAC Policy

2.1. THE SYSTEM WITH MIXED TRAFFIC

A wireless system with hexagonal radio cells and fixed channel allocations is considered. The system supports both constant-rate traffic and available-rate traffic. All the constant-rate sources require the same transmission bandwidth and all the available-rate sources have the same traffic characteristics. The traffic load is uniformly distributed over the coverage area. Both the system and the traffic are assumed to have attained stationary states. All new and handoff call arrivals to a cell are Poisson, and all service times and cell dwelling times are exponentially distributed and are independent of each other. The handoff rates of the same traffic type at a cell from the neighboring cells can be lumped into one single parameter. After admitted to a cell, a handoff call is treated in the same way as a new call of the same traffic type, i.e., the service rates for both handoff and new calls of the same traffic type are the same. In such a system, the focus can be placed on a single test cell described by the new and handoff call arrival rates, service rates and the total cell capacity. The service time in the test cell is the minimum of call duration and the cell dwelling time. Therefore, the service rate is the sum of the call service rate and the handoff rate of the admitted users to the neighboring cells. It is assumed that: (a) arrivals and services of constant-rate traffic are independent to those of available-rate traffic; and (b) the effects of channel fading are completely mitigated by the physical layer, and the problem of capacity fluctuation due to possible collisions of transmitted signals is eliminated by the data link layer. λ_C, h_C and μ_C denote respectively the new call arrival rate, the handoff call arrival rate and the service rate for constant-rate traffic. The same notations with subscript A are used to represent the corresponding quantities for available-rate traffic. The arrival rates and service rates are represented in calls per unit time. As there is a control/signaling resource overhead associated with each of the available-rate connections, when the available bandwidth for data transmission in each connection is very low, the resources for the connections are not used efficiently due to the relative large overhead. As a result, it is necessary to limit the number of the non real-time data users, even though the available-rate non real-time data sources can adapt to whatever bandwidth available. In practice, the number of base station transceivers for the data services is also limited. Hence, the QoS parameters of interest are new call blocking probability and handoff call dropping probability for both constant-rate and available-rate calls.

Because of the existence of heterogeneous traffic in the system, it is necessary to have a resource allocation scheme that defines the priority for using system resources. The restricted access (RA) scheme [7] is used here. This scheme is preferred over others because low priority traffic will never be deprived of resources indefinitely. The RA scheme is depicted in Figure 1, where C is the total capacity (radio resources) of the test cell. The real-time constant-rate traffic has preemptive priority over the non real-time available-rate traffic, and can occupy up
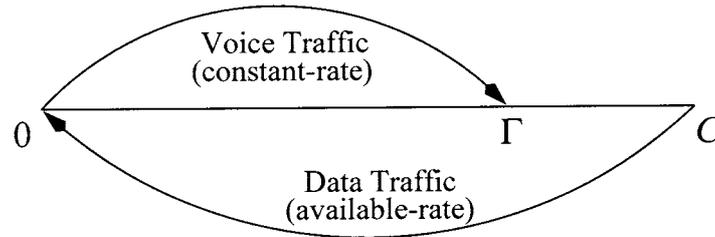
Figure 1. The RA resource allocation scheme for the integrated voice and data services.

to Γ out of the total capacity. The leftover resources in the test cell are equally shared among all the admitted available-rate users. The amount of resources received by each available-rate user is therefore random. Because the amount of resources received by an available-rate user varies, there is a chance that this amount drops to a critical level below which a connection will suffer severe quality degradation. This phenomenon is called overload. Thus, in addition to the call blocking and dropping probabilities, the overload probability is a QoS measure for the admitted available-rate users. The overload probability is different from (even though closely related to) the handoff call dropping probability, due to dynamics of user movements and of resource requirements from the high priority real-time services.

2.2. THE LFGCP POLICY

There are various types of CAC policies proposed in the literature. The one of interest here is the LFGCP [8], because (a) the policy is simple to implement and (b) given certain QoS requirements, the CAC policy parameters can be determined in a systematic way. This policy is originally proposed to work with a single type of traffic, each call occupying one channel. It can be abbreviated as $g_{T,M}^{\beta}$, where M is the number of channels available in a radio cell, $T \ (< M)$ is the number of busy channels over which no new calls are accepted, and β is a constant denoting the probability of accepting a new call when the channel occupancy in the cell is T. Let i denote the current number of occupied channels. The policy can be summarized as: (a) a new call is always accepted if $i \ < \ T$, is accepted with probability β if $i \ = \ T$, and is always rejected if $i \ > \ T$; and (b) a handoff call is always accepted if $i \ < \ M$. With only one type of traffic, let λ and h denote the new and handoff call arrival rates respectively, and let μ denote the service rate. The performance of the LFGCP in terms of new call blocking probability (B_n) and handoff call dropping probability (D_h) can be calculated, by using a Markov chain to model the channel occupancy of the test cell.

Given a certain traffic condition, the three parameters M, T and β of the LFGCP can be determined in a systematic way. In particular, with the objective of maximizing the resource utilization together with constraints on the blocking and dropping probabilities, the three parameters can be obtained as the solution of an optimization problem. As minimizing the value of M is equivalent to maximizing the resource utilization under the constraints, the optimization problem can be formulated as follows.

Given:
λ, h, μ, Q_n (the allowed maximum new call blocking probability) and Q_h (the allowed maximum handoff call dropping probability)

Objective:
Find M, T and β such that M is minimized, $B_n \leq Q_n$ and $D_h \leq Q_h$.

The optimization problem can be solved by the `Min M` algorithm [8] given in Appendix A. Numerical results presented in [9] demonstrate that the CAC policy with parameters determined by the `Min M` algorithm is able to meet the QoS requirements specified by Q_n and Q_h under various traffic conditions.

In the optimization problem formulation, we try to find the minimum M for the QoS satisfaction given the traffic load. However, in a practical wireless system, normally the system resources are fixed and the objective in CAC is to admit a maximal number of users for high resource utilization under the QoS constraints. Nevertheless, the optimization problem formulation and the `Min M` algorithm are still applicable to the CAC under consideration. Given the fixed total resources in the system, the CAC parameter set $\{M, T, \beta\}$ is a function of the traffic load. Due to the dynamics of user mobility, the traffic load changes with time. Indeed, the main motivation for introducing the soft QoS concept is such that the CAC can adapt to the traffic load variation for a better compromise between QoS provisioning and resource utilization efficiency. As a result, the CAC parameter set needs to be redetermined whenever there is a significant change in the traffic load. Given the traffic load and the QoS requirement, the `Min M` algorithm can be used iteratively to determine the CAC parameter set such that M is equal to or very close to (but not larger than) the system total resources.

Even though the LFGCP was original proposed for voice traffic, its main feature of providing higher priority to handoff calls than that to new calls suits well to data traffic. Furthermore, the QoS parameters at the call level for data traffic are the same (except the additional overload probability requirement) as those of voice traffic. As a result, we propose to apply the LFGCP to data traffic.

2.3. THE CAC POLICY FOR THE MIXED TRAFFIC AND PERFORMANCE ANALYSIS

For the system supporting both constant-rate and available-rate traffic using the RA resource allocation scheme, the proposed CAC policy consists of two LFGCPs [9]: one for each traffic type. Let the LFGCPs for constant-rate and available-rate traffic be $g_{T_C, M_C}^{\beta_C}$ and $g_{T_A, M_A}^{\beta_A}$ respectively. However, as data traffic is different from voice traffic in the aspect that the admitted available-rate users are not allocated with a fixed bandwidth, the integer variables T_C, M_C, T_A, and M_A are defined in terms of the number of users, instead of the number of channels. Then the CAC policy for mixed traffic can be represented as follows, where i_C and i_A are respectively the current numbers of constant-rate and available-rate users in the test cell:

```
*********** CAC POLICY FOR MIXED TRAFFIC ***********
```
if (New_Constant-Rate_Call_Request) then
  if ($i_C < T_C$) then
    Admit_Constant-Rate_Call
  else if ($i_C = T_C$) then
    Admit_Constant-Rate_Call_with_Probability_β_C
  else
    Reject_Constant-Rate_Call;
if (Handoff_Constant-Rate_Call_Request) and ($i_C < M_C$) then
  Admit_Constant-Rate_Call

else
  Reject_Constant-Rate_Call;
if (New_Available-Rate_Call_Request) then
  if ($i_A < T_A$) then
    Admit_Available-Rate_Call
  else if ($i_A = T_A$) then
    Admit_Available-Rate_Call_with_Probability_β_A
  else
    Reject_Available-Rate_Call;
if (Handoff_Available-Rate_Call_Request) and ($i_A < M_A$) then
  Admit_Available-Rate_Call
else
  Reject_Available-Rate_Call;
****

Because the two types of traffic share the total capacity C using the RA scheme, the two LFGCPs, $g_{T_C,M_C}^{\beta_C}$ and $g_{T_A,M_A}^{\beta_A}$, are not independent of each other. The correlation is captured in the overload probability given in the following and in the determination of the CAC parameters as described in Section 3.

To analyze the performance of the CAC policy, the following notations are defined:

γ_C – the amount of capacity required by a constant-rate call;

γ_A – the critical amount of capacity for an available-rate call, which is basically the minimum amount of resources required to maintain each available-rate data connection, taking into account the required overhead for control and signaling;

B_{nC} (D_{hC}) – the new call blocking (handoff call dropping) probability for constant-rate calls;

B_{nA} (D_{hA}) – the new call blocking (handoff call dropping) probability for available-rate calls;

Π_{oA} – the overload probability for available-rate calls;

Q_{nC} (Q_{hC}) – allowed maximum new call blocking (handoff call dropping) probability for constant-rate calls;

Q_{nA} (Q_{hA}) – allowed maximum new call blocking (handoff call dropping) probability for available-rate calls;

Q_{oA} – suggested upper bound of the overload probability for available-rate calls.

The traffic model for the constant-rate calls in the test cell can be described by a Markov chain. Using queueing theory [10], it can be derived that the steady state probability of the number of constant-rate users in the cell being j is

$$p_C(j) = \begin{cases} \frac{\rho_C^j}{j!} p_C(0), & 0 \le j \le T_C \\[2mm] \frac{\rho_C^j[\alpha_C+(1-\alpha_C)\beta_C]}{j!} p_C(0), & j = T_C + 1 \\[2mm] \frac{\rho_C^j[\alpha_C+(1-\alpha_C)\beta_C]\alpha_C^{j-T_C-1}}{j!} p_C(0), & T_C + 2 \le j \le M_C \end{cases} \tag{1}$$

where $\rho_C = \frac{\lambda_C + h_C}{\mu_C}$ is the total offered traffic load of constant-rate calls, $\alpha_C = \frac{h_C}{\lambda_C + h_C}$ is the fraction of the total offered constant-rate load which is handoff calls, and $p_C(0)$ is a normalization constant given by

$$p_C(0) = \left\{ \sum_{j=0}^{T_C} \frac{\rho_C^j}{j!} + \frac{\rho_C^{T_C+1}[\alpha_C + (1-\alpha_C)\beta_C]}{(T_C+1)!} + \sum_{j=T_C+2}^{M_C} \frac{\rho_C^j[\alpha_C + (1-\alpha_C)\beta_C]\alpha_C^{j-T_C-1}}{j!} \right\}^{-1}. \quad (2)$$

Consequently, the new call blocking probability is

$$B_{nC} = (1-\beta_C)p_C(T_C) + \sum_{j=T_C+1}^{M_C} p_C(j) \quad (3)$$

and the handoff call dropping probability is

$$D_{hC} = p_C(M_C). \quad (4)$$

Define the resource utilization efficiency, ϕ_c, as the expected fraction of the number of busy channels in the system. It can be calculated by

$$\phi_c = \sum_{j=0}^{M_C} \frac{j \cdot p_C(j)}{M_C}. \quad (5)$$

The same equations can be applied to available-rate traffic by simply replacing the subscript C with A. As to the overload probability (Π_{oA}) for the available-rate calls, by following the approach in [7, 9], it can be derived that

$$\Pi_{oA} = \sum_{\{r \in A_R\} \cap \{r < \gamma_A\}} \Pr[c_A = r] \quad (6)$$

where c_A is the instantaneous amount of resources received by an available-rate user,

$$\Pr[c_A = r] = \frac{\sum_{(i,j) \in A_S} p_C(i) \cdot j \cdot p_A(j)}{\sum_{k=0}^{M_A} k \cdot p_A(k)}.$$

A_R is the set of distinct values in $\left\{ \frac{C - i \cdot \gamma_C}{j} : 0 \le i \le M_C, 1 \le j \le M_A \right\}$, and A_S is $\{(i, j) : 0 \le i \le M_C, 1 \le j \le M_A, \frac{C - i \cdot \gamma_C}{j} = r\}$.

3. Soft QoS in the CAC Policy

From the point of view of the service provider, resource utilization is, besides QoS provisioning, an important factor that cannot be ignored. A highly utilized system that can provide a satisfactory level of QoS to the users is always a desired solution. However, high resource utilization and QoS provisioning are always conflicting goals. As in the case of CAC, resources are set aside for active (or handoff) users in the neighboring cells so that their QoS can be maintained, but unused resources mean low utilization. By slightly relaxing the original target QoS levels, more users can be allowed into the systems, resulting in an increase of the resource utilization efficiency. For example, if Q_n^T is the target allowed maximum new call blocking

probability and Q_h^T is the target allowed maximum handoff call dropping probability, then the soft QoS (i.e., relaxed target QoS) can be expressed as

$$Q_n^S = \eta_n \cdot Q_n^T \quad \text{and} \quad Q_h^S = \eta_h \cdot Q_h^T, \tag{7}$$

where η_n and η_h are constants larger than 1. The values of η_n and η_h depend on the extent of QoS fluctuations that the users can tolerate and on traffic characteristics. If a small increase of the η_n and η_h values results in a significant increase of the resource utilization efficiency, then a small increase in the QoS fluctuation will mean a large reduction in the service cost and the users would be willing to suffer the temporarily QoS degradation when the system is congested. However, there should be upper bound values for η_h and η_n; otherwise the purpose of having soft QoS becomes fallacious. Furthermore, in a multimedia wireless communication system, each multimedia traffic usually has its own QoS requirement. Performing the CAC properly in order to meet all the requirements simultaneously is difficult and sometimes impossible due to the limited resources. By specifying soft QoS levels for some of the traffic types, a compromise among the multiple requirements can be reached. In this section, we present numerical results to demonstrate how the resource utilization efficiency is increased as a function of η_n and η_h and how soft QoS can be introduced to balance different QoS requirements of constant-rate and available-rate traffic. Our focus is placed on the parameter study, including the CAC and QoS parameters and the resource utilization efficiency, as the CAC policy is completely specified by the CAC parameters and the performance of the CAC policy is evaluated in terms of QoS and resource utilization.

3.1. Soft QoS versus Resource Utilization Efficiency

In the following, how the resource utilization efficiency changes with respect to soft QoS will be presented. For simplicity, consider only a single type of traffic (such as constant-rate) in the system with an LFGCP and the corresponding subscript (C) in the mathematical symbols is omitted. In this case, soft QoS is referred to as the relaxed upper bounds on the new call blocking and handoff call dropping probabilities. The investigation is carried out in two steps:

Step 1 *With Hard QoS:* The QoS parameters are specified as $Q_n = Q_n^T (= 0.015)$ and $Q_h = Q_h^T (= 0.001)$, and the traffic conditions are set at $\rho = 25$, $\alpha = \frac{1}{3}$. The three parameters M, T and β of the CAC policy can then be found by using the `Min M` algorithm. The resource utilization efficiency with the hard QoS, denoted by ϕ^T, is found to be 0.6694.

Step 2 *With Soft QoS:* The QoS requirements are relaxed with $Q_n = Q_n^S (= \eta_n \cdot Q_n^T)$ and $Q_h = Q_h^S (= \eta_h \cdot Q_h^T)$. Because of the changes in the QoS parameters, M, T and β need to be recalculated and the resource utilization reevaluated. This is repeated when both η_n and η_h are allowed to vary from 1 to 10 simultaneously, and the increase in the resource utilization efficiency versus η_n and η_h is shown in Figure 2. The increase is represented in percentage of ϕ^T.

In general, the utilization efficiency increases as η_n and/or η_h increases. For the range of values shown, the maximum increase in the efficiency is 17%. To have a closer look of the comparison between hard and soft QoS, Figure 3 shows the efficiency increase as a function of η_n at $\eta_h = 2$. With slightly relaxed QoS such as $\eta_h = 2$ and $\eta_n = 5.5$, an increase of 10% in the efficiency can be achieved. It should be mentioned that the resource utilization with hard
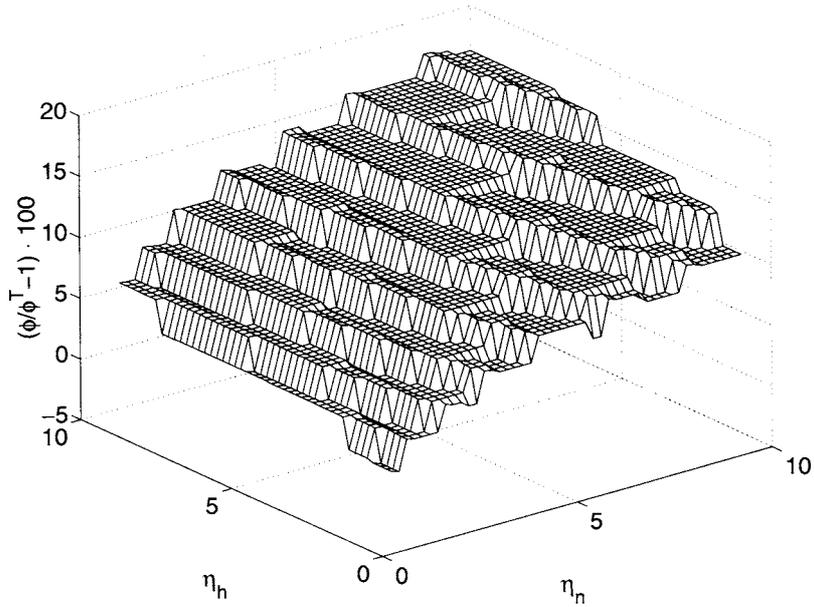
Figure 2. The change in resource utilization in percentage over ϕ^T for various values of η_n and η_h.
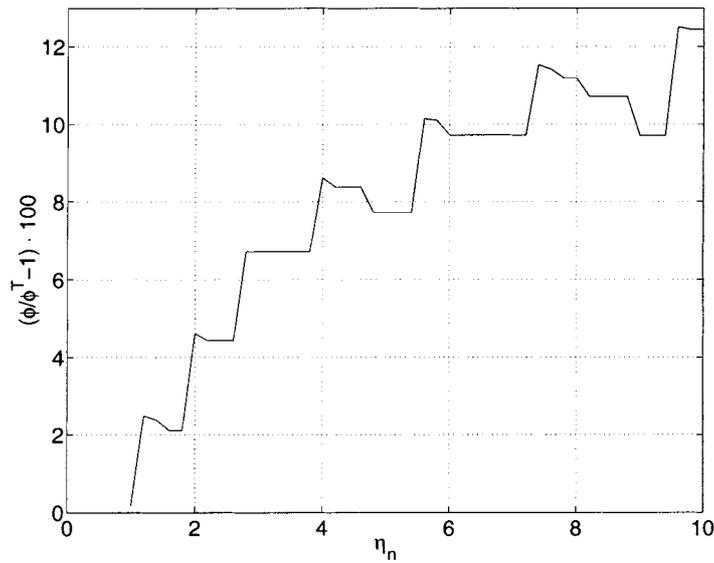

Figure 3. The change in resource utilization in percentage over ϕ^T for $\eta_h = 2$.

QoS has already been maximized by minimizing the value of M in the CAC policy, and the efficiency increase with soft QoS is on top of the maximized value of ϕ^T. In Figures 2 and 3, the change in the utilization efficiency is not a monotonically increasing function of η_n and/or η_h because the channel occupancy is not a continuous variable.

When there are both the constant-rate traffic and the available-rate traffic in the system, it is expected that relaxing the QoS parameters for one traffic or for both traffic will be effective in increasing the resource utilization.

3.2. QoS Compromise between Constant-Rate and Available-Rate Traffic

The effectiveness of maintaining multiple QoS requirements by introducing soft QoS to CAC is investigated in the following. In particular, numerical results will be presented to show how much improvement in the overload probability of available-rate calls can be obtained when both the new call blocking and handoff call dropping probabilities of constant-rate calls are relaxed. That is, soft QoS here is referred to as the relaxed upper bounds on the new call blocking and handoff call dropping probabilities for the constant-rate traffic. In order to do that, two methods of determining the CAC policy parameters will be considered for comparison. The first one takes only target QoS requirements into account, and the second one takes both target and soft QoS requirements into account. First of all, let the symbols with a superscript T represent the values required to maintain the target QoS, while those with a superscript S represent the corresponding values required to maintain the soft QoS. The soft QoS for constant-rate calls is introduced in the way as specified in Equation (7). In the following two algorithms, the function MinM($\cdot$) represents a call to algorithm `Min M`. The function $\Pi_{oA}(g^{\beta_C}_{T_C,M_C}, g^{\beta_A}_{T_A,M_A})$ returns the overload probability when $g^{\beta_C}_{T_C,M_C}$ and $g^{\beta_A}_{T_A,M_A}$ are chosen for the constant-rate and available-rate calls respectively. Both algorithms return a hex-tuple $(M_C, T_C, \beta_C, M_A, T_A, \beta_A)$ which can then be used to specify the CAC policy described in Section 2.3, if the solution exists. The first method is represented by algorithm `Min M_1`:

*************** ALGORITHM `Min M_1` ***************
1. $(M^T_C, T^T_C, \beta^T_C) = \text{MinM}(Q^T_{nC}, Q^T_{hC})$;
2. $(M_A, T_A, \beta_A) = \text{MinM}(Q_{nA}, Q_{hA})$;
3. if $M_C \cdot \gamma_C > C$ OR $\Pi_{oA}\left(g^{\beta^T_C}_{T^T_C,M^T_C}, g^{\beta_A}_{T_A,M_A}\right) > Q_{oA}$

    return ("Error: The QoS requirements can never be
        fulfilled with current resources".)

  else

    return $(M^T_C, T^T_C, \beta^T_C, M_A, T_A, \beta_A)$;
****

The second method is represented by algorithm `Min M_2`:

*************** ALGORITHM `Min M_2` ***************
1. $(M^T_C, T^T_C, \beta^T_C) = \text{MinM}(Q^T_{nC}, Q^T_{hC})$;
2. $(M^S_C, T^S_C, \beta^S_C) = \text{MinM}(Q^S_{nC}, Q^S_{hC})$;
3. if $M^S_C \cdot \gamma_C > C$

    return ("Error: The QoS requirements can never be
        fulfilled with current resources".);

4. $(M_A, T_A, \beta_A) = \text{MinM}(Q_{nA}, Q_{hA})$;
5. if $\Pi_{oA}\left(g^{\beta^T_C}_{T^T_C,M^T_C}, g^{\beta_A}_{T_A,M_A}\right) > Q_{oA}$ OR $M^T_C \cdot \gamma_C > C$

    return$(M^S_C, T^S_C, \beta^S_C, M_A, T_A, \beta_A)$

  else

    return$(M^T_C, T^T_C, \beta^T_C, M_A, T_A, \beta_A)$;
****

The consideration of soft QoS in `Min M_2` is represented in Step **5**, which indicates if the suggested upper bound of overload probability for available-rate calls is violated by the

Table 1. Experimental parameters for the numerical example.

ρ_C	α_C	γ_C	Q_{nC}^T	Q_{hC}^T	C
30	$\frac{1}{3}$	5	10^{-3}	10^{-4}	250

η_n	η_h	γ_A	Q_{nA}	Q_{hA}	Q_{oA}
3.0	2.0	1.0	0.2	0.02	0.003

achievement of the target QoS requirements for constant-rate calls, or if it is impossible to achieve the target QoS requirements with the use of the current available resources, soft QoS requirements for constant-rate calls will be used. Because available-rate calls are of low priority in the system, the suggested upper bound for the overload probability cannot be guaranteed if the constant-rate traffic load is high. With respect to the RA scheme shown in Figure 1, given the total capacity C, the parameter Γ is equal to $M_C^T \cdot \gamma_c$ or $M_C^S \cdot \gamma_C$ depending on which QoS is used. The relation between M_A and $\{\Gamma, C\}$ is not straightforward. Because available-rate traffic is not real-time, the capacity allocated to each available-rate source is equal to the total leftover capacity (after serving constant-rate traffic) divided by the total number of active available-rate sources in the cell. On the average, given Γ and C, a larger M_A value corresponds to a smaller share of bandwidth for each available-rate source. The related QoS is represented only by the overload probability Π_{oA}, as the other two QoS parameters B_{nA} and D_{hA} are guaranteed first by the LFGCP $g_{T_A,M_A}^{\beta_A}$. In the algorithms, an error occurs when no practical value of the CAC parameters can be found to fulfill the QoS requirements for the given fixed total capacity C. In the situation, the CAC parameter set should be redetermined by increasing Q_{nA}^T, the upper bound of the new call blocking probability for the available-rate data users who have the lowest priority among all the users. Similarly, if the overload probability $\Pi_{oA}\left(g_{T_C^T,M_C^T}^{\beta_C^T}, g_{T_A,M_A}^{\beta_A}\right)$ is larger than the required upper bound Q_{oA} even after the QoS for the constant-rate calls is relaxed, the CAC parameter set should be redetermined by increasing Q_{nA}^T. By reducing the number of new available-rate calls, each admitted available-rate user will have a larger share of the bandwidth; hence, the overload probability will be reduced.

To check whether the CAC policies with parameters determined by `Min M_1` and `Min M_2` are capable of maintaining the specified QoS requirements, consider a situation where the traffic parameters for constant-rate calls are fixed while those for available-rate calls, ρ_A and α_A, are allowed to vary from 4 to 40 Erlangs and $1/20$ to $19/20$ respectively. Other parameters are given in Table 1.

Figures 4 and 5 demonstrate the performance of the CAC policy determined by `Min M_2` for constant-rate calls. From the new call blocking probability for constant-rate calls shown in Figure 4, it can be seen that the target QoS ($\log_{10} Q_{nC}^T = -3$) can only be achieved in the light to medium traffic conditions, while the soft QoS ($\log_{10} Q_{nC}^S = -2.52$) is maintained in the heavy traffic conditions. Similar situations can be observed from the handoff call dropping probability for constant-rate calls shown in Figure 5. As the traffic load increases to a certain level, the hard QoS cannot be guaranteed with the given total resources C. At this point, the algorithm `Min M_2` switches to the soft QoS specification. On the other hand, if the soft QoS is not introduced, in the light to medium traffic load, the performance of the CAC policy
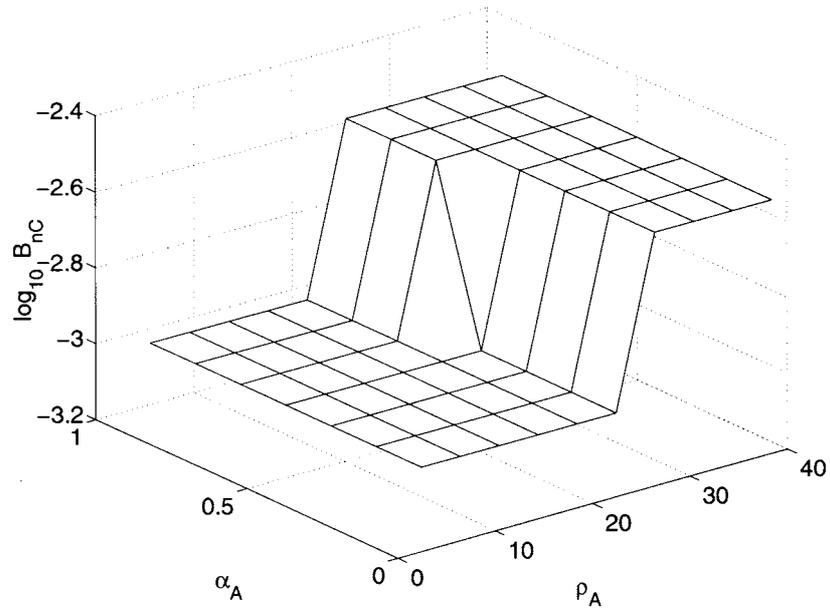
Figure 4. The new call blocking probability for constant-rate calls under CAC determined by `Min M_2`.
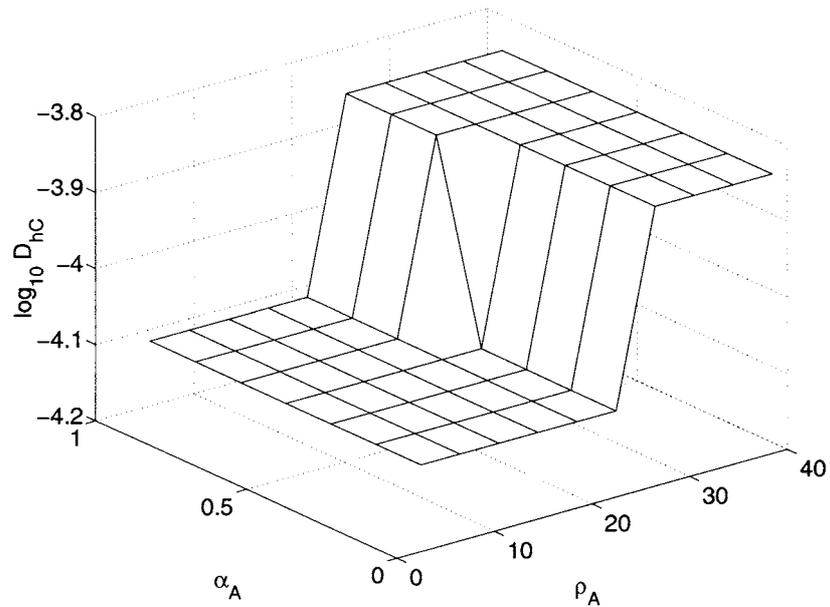

Figure 5. The handoff call dropping probability for constant-rate calls under CAC determined by `Min M_2`.

determined by `Min M_1` will be the same as that shown in Figures 4 and 5; however, when the traffic load falls into the heavy load region, the CAC policy fails as the QoS requirements cannot be satisfied with the given resources.

Figures 6 and 7 show the performance of the CAC determined by `Min M_2` for available-rate calls. From Figure 6, it can be seen that the new call blocking probability increases as the number of new call arrivals to the system increases (small α_A implies more new call arrivals
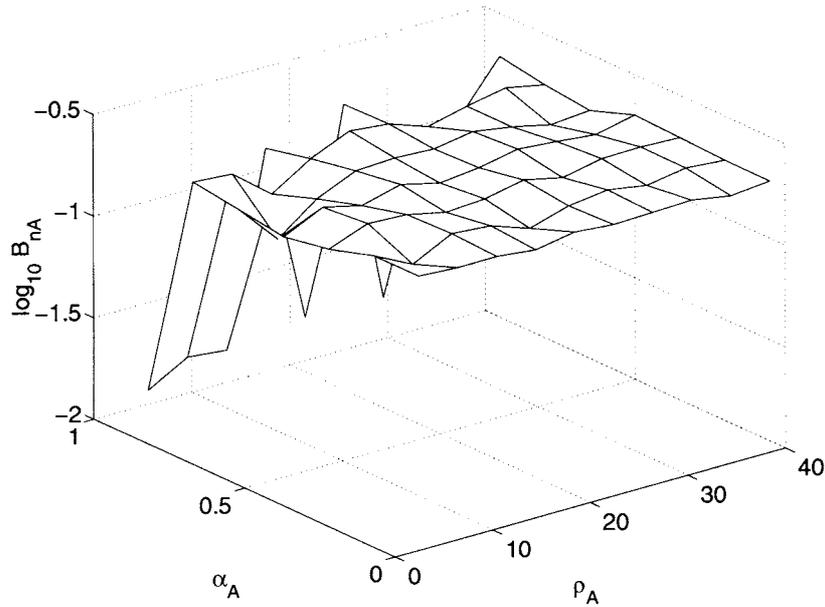
Figure 6. The new call blocking probability for available-rate calls using CAC determined by `Min M_2`.

and vice versa); nevertheless, the QoS requirement ($\log_{10} Q_{nA} = -0.70$) is always maintained for all traffic conditions. The slight fluctuations of the call blocking probability is due to the rounding process as the number of channels is an integer variable. Similar situations can be observed for the handoff call dropping probability from Figure 7. When α_A is small, there are fewer handoff call arrivals to the system and therefore handoff call dropping is less frequent. In addition, the system is able to maintain the QoS requirement ($\log_{10} Q_{hA} = -1.70$) under all traffic conditions considered. Because soft QoS is not specified for available-rate calls, the performance of the CAC policy determined by `Min M_1` is identical to the performance shown in Figures 6 and 7. From Figures 4–7, it is obvious that the CAC policy with soft QoS based on the algorithm `Min M_2` is always able to satisfy the QoS requirements for both new and handoff calls under the traffic conditions considered.

The overload probabilities of the CAC policies determined by `Min M_1` and `Min M_2` are shown in Figures 8 and 9 respectively. As can be seen, the suggested upper bound of the overload probability ($\log_{10} Q_{oA} = -2.52$) will not always be maintained because available-rate calls are of lower priority. When Figures 8 and 9 are compared, it can be noticed that in light to medium traffic conditions when the target QoS for constant-rate calls is maintained, the overload probabilities are identical in both figures. The two figures differ only in the heavy traffic conditions, in which the CAC policy based on algorithm `Min M_2` manage to reduce the overload probability for available-rate calls by relaxing the target QoS for constant-rate calls. To show the difference of the two figures more clearly, Figure 10 plots both overload probabilities together with the suggested upper bound at $\alpha_A = \frac{7}{20}$. The slight fluctuations of the curves are due to the integer parameters of the CAC policies. The maximum improvement is approximately $10^{0.4}$ (≈ 2.5), which means that the overload probability for available-rate calls can be decreased by a factor of 2.5 by introducing soft QoS to constant-rate calls. The difference diminishes slowly when traffic load increases; at $\rho_A = 35$ Erlangs, the CAC using soft QoS can still maintain an overload probability which is $10^{0.15}$ (≈ 1.4) times smaller than
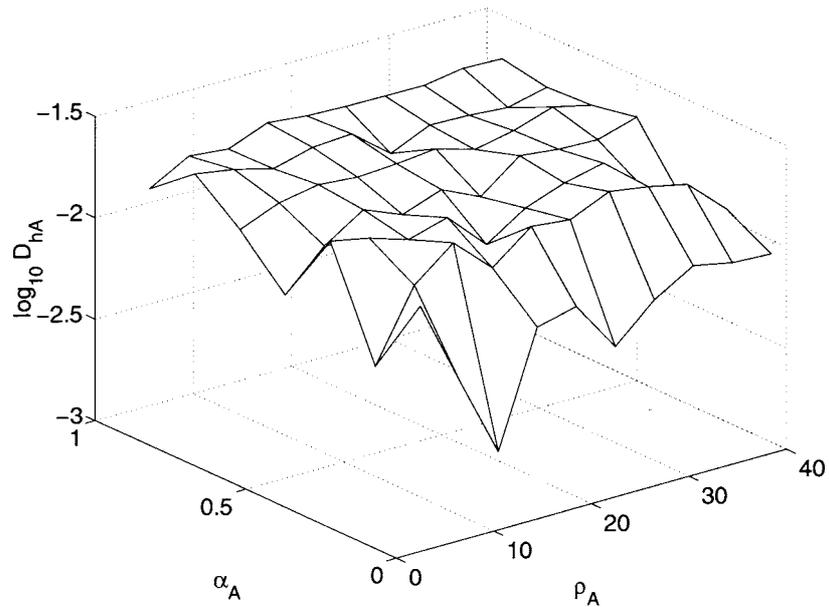
Figure 7. The handoff call dropping probability for available-rate calls using CAC determined by `Min M_2`.
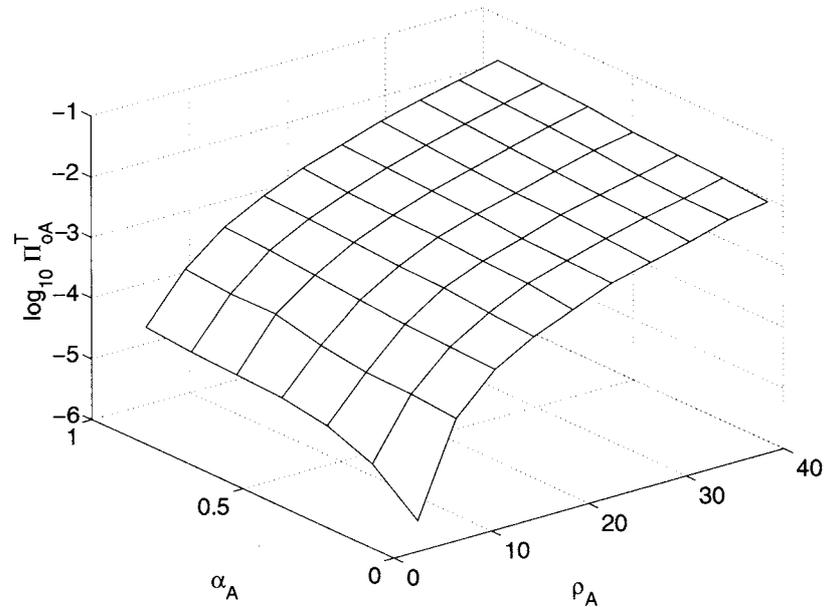

Figure 8. The overload probability for available-rate calls using CAC determined by `Min M_1`.

the overload probability without using soft QoS. In addition, by introducing soft QoS, the maximum traffic load below which a satisfactory overload probability is achieved is extended from 25 Erlangs to 30 Erlangs.

In general, when there is a request for resource allocation from the constant-rate users, the system may release a portion of the resources allocated to the available-rate users and reallocate them after first serving the constant-rate users. To avoid a high probability of the
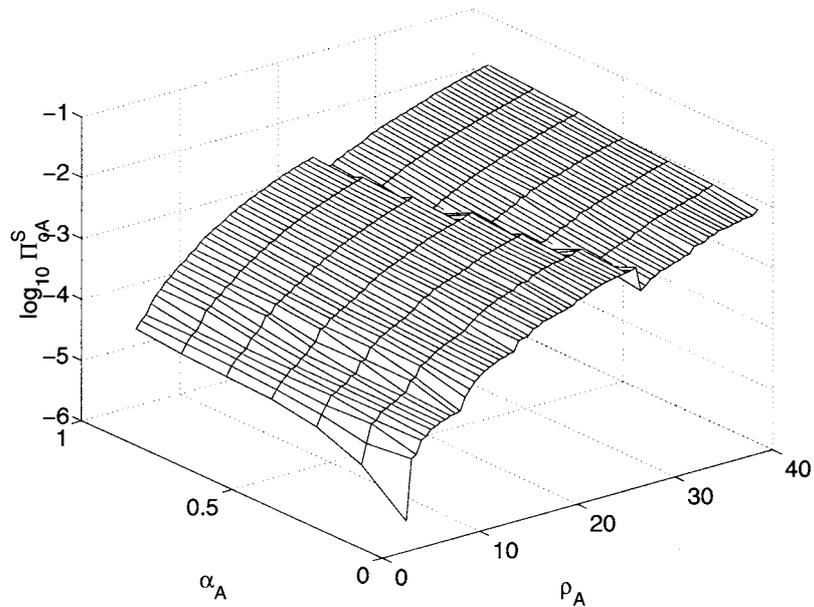
Figure 9. The overload probability for available-rate calls using CAC determined by `Min M_2`.
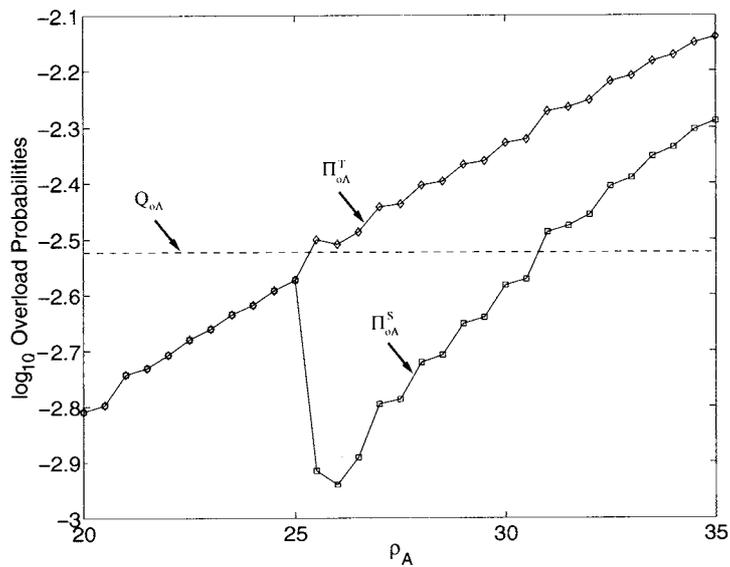

Figure 10. The overload probabilities and the suggested upper bound for comparison.

overload event for the available-rate users, in the above discussion, we limit the number of constant-rate users by increasing their new call blocking and handoff call dropping probabilities. Another way to reduce the overload probability is to limit the active available-rate users by increasing the new call blocking probability and maybe even handoff call dropping probability. In a worst situation, dropping some active available-rate connections can ensure that each of the remaining available-rate users will be allocated with resources not smaller

than the critical amount. The last approach is complex as it involves the fairness issue (as to which connections to drop). Note that dropping admitted users is not a CAC issue.

4. Conclusions

In this paper, we have investigated how soft QoS can be introduced to the CAC policy in wireless systems for increasing the resource utilization efficiency and for making a compromise between QoS requirements of constant-rate and available-rate traffic. For the traffic conditions considered, numerical results demonstrate that (a) by slightly relaxing the target QoS for either the new or handoff calls, a reasonable increase (such as 10%) in the resource utilization efficiency can be obtained, and (b) the overload probability of available-rate traffic can be decreased by a factor of up to 2.5 when soft QoS is specified for constant-rate calls. The idea of soft QoS in CAC will be very useful in the design and operation of wireless systems, in which decisions need to be made on the profitable values of many complicated factors such as resource utilization and multiple QoS requirements. These factors are not independent of each other, and the designer/operator usually faces dilemma when specifying their values. Soft QoS provides flexibility in specifying relationships among these factors.

Appendix: The `Min M` Algorithm [8]

In this algorithm, $B_n(M, T, \beta)$ and $D_h(M, T, \beta)$ are respectively functions that calculate the new call blocking and handoff call dropping probabilities of the system under the control of the LFGCP $g_{T,M}^{\beta}$. The functions $\text{Int}(y)$ and $\text{Frac}(y)$ return respectively the integer part and the fractional part of the number y. The result of the algorithm `min M` is the tri-tuple (M, T, β).

```
***************** ALGORITHM Min M *****************
```

1. $M = 1$;
2. while $B_n(M, M, 0) > Q_n$
      $M = M + 1$;
3. if $D_h(M, M, 0) \leq Q_h$
      return $(M, M, 0)$;
4. $U = M$;
5. $L = 0$;
6. $y = (U + L)/2$;
7. while $B_n(M, \text{Int}(y), \text{Frac}(y)) > Q_n$ XOR $D_h(M, \text{Int}(y), \text{Frac}(y)) > Q_h$ {
      if $B_n(M, \text{Int}(y), \text{Frac}(y)) > Q_n$ {
        $L = y$;
        $y = (U + L)/2$; }
      else if $D_h(M, \text{Int}(y), \text{Frac}(y)) > Q_h$ {
        $U = y$;
        $y = (U + L)/2$; } }
8. if $B_n(M, \text{Int}(y), \text{Frac}(y)) \leq Q_n$ AND $D_h(M, \text{Int}(y), \text{Frac}(y)) \leq Q_h$
   return $(M, \text{Int}(y), \text{Frac}(y))$;
  else {
    $M = M + 1$;
    goto step **3**; }

```
**************************************************
```

Acknowledgements

References

1. V. Li and X. Qiu, "Personal Communication Systems (PCS)", *Proc. IEEE*, Vol. 83, No. 9, pp. 1210–1243, 1995.
2. D. Hong and S.S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures", *IEEE Trans. on Vehicular Tech.*, Vol. 35, 1986.
3. S. Tekinay and B. Jabbari, "Handover and Channel Assignment in Mobile Cellular Networks", *IEEE Commun. Mag.*, Vol. 29, pp. 42–46, 1991.
4. D.A. Levine, I.F. Akyildiz and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept", *IEEE Trans. Networking*, Vol. 5, pp. 1–12, 1997.
5. K. Lee, "Supporting Mobile Multimedia in Integrated Services Networks", *Wireless Networks*, Vol. 2, No. 3, pp. 205–217, 1996.
6. O. Yu and V. Leung, "Adaptive Resource Allocation for Prioritized Call Admission over an ATM-Based Wireless PCN", *IEEE J. Select. Areas Commun.*, Vol. 15, No. 7, pp. 1208–1225, 1997.
7. M. Naghshineh and A. Acampora, "QoS Provisioning in Micro-Cellular Networks Supporting Multiple Classes of Traffic", *Wireless Networks*, Vol. 2, No. 3, pp. 195–203, 1996.
8. R. Ramjee, D. Towsley and R. Nagarajan, "On Optimal Call Admission Control in Cellular Networks", *Wireless Networks*, Vol. 3, No. 1, pp. 29–41, 1997.
9. C.W. Leong and W. Zhuang, "Optimal Call Admission Policy for Wireless Multimedia Networks", in *Proc. IEEE Veh. Technol. Conf. (VTC'99)*, Houston, U.S.A., May 1999.
10. L. Kleinrock, *Queueing Systems, Volume I: Theory*, John Wiley and Sons, 1975.

Chi Wa Leong received the B.Sc. degree in computer engineering and the M.Sc. degree in electrical engineering, both from the University of Manitoba, Canada. He was a research assistant at the University of Waterloo, Canada, where he carried out research on call admission control for mobile networks. He also received the M.A.Sc. degree in electrical engineering from the University of Waterloo.

Weihua Zhuang received the B.Sc. (1982) and M.Sc. (1985) degrees from Dalian Marine University (China) and the Ph.D. degree (1993) from the University of New Brunswick (Canada), all in electrical engineering. Since 1993 she has been a faculty member at the University of Waterloo where she is currently an Associate Professor in the Department of Electrical and Computer Engineering. Her research interests include mobility and resource management for wireless multimedia communications.

Dr. Zhuang is a licensed Professional Engineer in the Province of Ontario, Canada.