Wireless Networks 9, 51–59, 2003

QoS Provisioning in Wireless/Mobile Multimedia Networks Using an Adaptive Framework

TAEKYOUNG KWON * and YANGHEE CHOI
Seoul National University, Seoul, Korea

CHATSCHIK BISDIKIAN and MAHMOUD NAGHSHINEH
IBM Watson Research Center, Hawthorne, NY 10532, USA

Abstract. Recently there is a growing interest in the adaptive multimedia networking where the bandwidth of an ongoing multimedia call can be dynamically adjusted. In the wireless/mobile multimedia networks using the adaptive framework, the existing QoS provisioning focused on the call blocking probability and the forced termination probability should be modified. We, therefore, redefine a QoS parameter – the *cell overload probability* – from the viewpoint of the adaptive multimedia networking. Then, we propose a distributed call admission control (CAC) algorithm that guarantees the upper bound of the *cell overload probability*. Also, a bandwidth adaptation algorithm which seeks to minimize the *cell overload probability* is also presented. Simulation experiments are carried out to verify the performance of the proposed CAC algorithm. Furthermore, the performance of the adaptive wireless/mobile network is compared to that of the existing non-adaptive wireless/mobile networks. As a further step in QoS provisioning, we propose another QoS parameter, the *degradation period ratio*, and discuss analytically how the CAC algorithm guarantees the upper bound of the *degradation period ratio*.

Keywords: QoS provisioning, call admission control, adaptive multimedia, adaptive framework, wireless/mobile multimedia network, bandwidth adaptation

1. Introduction

With the explosive demand on wireless/mobile communications and the emergence of bandwidth-intensive multimedia applications, Quality of Service (QoS) provisioning in wireless/mobile multimedia networks is becoming more and more important. The most significant QoS parameters in the existing wireless/mobile multimedia networks are the call blocking probability and the forced termination probability. It is widely accepted that the forced termination of an ongoing call (e.g., due to handoff failure) is much more unbearable to users than the blocking of a new call, so that there have been numerous schemes to provide satisfiable QoS by exclusively reserving the network resources for handoff calls. The key factor in QoS provisioning is a call admission control (CAC) algorithm that provides acceptable QoS to users and utilizes the system resources (e.g., wireless link bandwidth) efficiently. So far, CAC algorithms in wireless/mobile multimedia networks have been focused on how to block originating calls to satisfy the QoS requirements (e.g., the upper bound of the forced termination probability [10,12,18]).

Recently, in order to overcome the scarce and highly-fluctuating link bandwidth in wireless/mobile networks, leveraging the adaptive multimedia networking is proposed (e.g., [1,13]). In the wireless/mobile multimedia networks using the adaptive framework, the bandwidth of an ongoing call can be dynamically adjusted to adapt to the various communication environments, especially in overloaded situations. With the help of this adaptive framework, the forced termination probability is reduced to a negligible level in moderate traffic load, especially when the minimum possible bandwidth of a call is sufficiently small.

Originally, the concept of the adaptive multimedia networking was introduced in the wired network. In wired broadband networks like ATM, once a call is admitted to the network, a contract between network and application is established. Then, they both try to keep the contract throughout the call's lifetime. In such a paradigm, network congestion can cause fluctuations in the availability of network resources, thereby resulting in severe degradation of multimedia services. To overcome this problem, many adaptive multimedia encoding and/or networking schemes are proposed such as hierarchical encoding [17] and network filters [6] to mitigate the effect of fluctuation in the network resources.

We advocate in this paper that the adaptive multimedia networking paradigm can play an important role to mitigate the highly-varying resource availability in wireless/mobile networks. Compared to wired networks, the fluctuation in resource availability in wireless/mobile networks is much more severe and results from two inherent features of such networks: fading and mobility. The fading in a wireless channel is highly-varying with time and spatial dependencies and interference. We assume that the effect of fading can be mitigated by rich-function transmission/reception wireless subsystems (e.g., [4]). The second reason for the fluctuation in resource availability is mobility (e.g., handoff in the cellular networks), which is inherently unpredictable in public land mobile networks (PLMNs). The adaptive multimedia networking framework in this paper takes into consideration

* Corresponding author.

VTEX() PIPS No:5101561 artty:res (Kluwer BO v.2002/08/19)
n5101561.tex; 12/09/2002; 8:44; p. 1

only mobility. Also, the cellular network environment is the focus of this paper. That is, an ongoing multimedia call in a given cell may change its bandwidth due to a new call arrival, a call completion, or an incoming/outgoing handoff call.

In the existing non-adaptive wireless/mobile networks, an incoming call to a given cell will be forced to be terminated if there is not sufficient wireless bandwidth (e.g., channels) to accommodate the call. However, with the above adaptive multimedia networking paradigm, the bandwidth of the existing calls in the given cell can be reduced to smaller values, hence freeing bandwidth for the incoming call. Also, the bandwidth of the incoming call can be adjusted depending on the situation of the given cell. In conclusion, there is a tradeoff between having adaptive bandwidth and reducing the forced-termination. That is, the problem of forced-termination is moved to bandwidth adaptation which is much more bearable to service users.

In the adaptive multimedia framework, QoS provisioning is performed by two workhorse algorithms: a CAC algorithm and a bandwidth adaptation algorithm (BAA). In this paper, we redefine a QoS parameter – the *cell overload probability* – and propose the CAC algorithm that guarantees the upper bound of the *cell overload probability* in a cell. Also, we devise the BAA that allocates/reallocates bandwidth of ongoing calls in a cell. Here, "bandwidth adaptation" means the bandwidth allocation of incoming calls and/or the reallocation of bandwidth of the existing calls in a given cell. The proposed BAA seeks to minimize the *cell overload probability* at any time.

Recently, some CAC algorithms and/or BAAs have been proposed in wireless/mobile networks using the adaptive framework [2,3,19]. Bharghavan et al. [2] propose an overall framework including CAC and bandwidth adaptation, which seeks to achieve optimal revenue over the whole network. However, the message overhead to ensure optimal bandwidth adaptation is inherently high. Also, it assumes continuous values of bandwidth in the adaptive framework. A more generalized BAA is proposed by Talukdar et al. [19] who investigate the tradeoff between network overhead and optimal bandwidth allocation. The above two approaches aim at optimal bandwidth adaptation over the whole mobile/wireless network. On the other hand, Das and Sen [3] propose an optimal BAA from a cell's perspective. They exploit the tradeoff between the carried traffic and bandwidth degradation.

The rest of this paper is organized as follows. Modeling of the adaptive multimedia networking is presented in section 2. The proposed CAC algorithm is proposed in section 3. Section 4 details the proposed BAA. The results of simulation experiments are shown in section 5. The analysis of the forced termination probability is given in section 6. In section 7, we discuss another QoS parameter: the *degradation period ratio*, the portion of a call's lifetime during which the call is allocated bandwidth less than a predefined target value. Finally, we conclude this paper in section 8.

2. Modeling the adaptive multimedia networking

Recall that in the wireless/mobile multimedia network with the adaptive framework, the bandwidth of a multimedia call can be dynamically adjusted depending on the situation during its lifetime. We assume that the bandwidth of a call takes its discrete value from the set $B = \{b_1, b_2, \ldots, b_n\}$ where $b_i < b_{i+1}$ for $i = 1, \ldots, n-1$. (For the possible values in B, refer to [2,13,19].) We also assume that all calls belong to a single class and all of them take (varying) bandwidth values from the same set B. Here, bandwidth values in B are integers, all of which are assumed to be multiples of a basic bandwidth unit (e.g., a single channel).

As for traffic characterization, we assume a simple model from a cell's perspective. The new call arrivals into a given cell are assumed to form a Poisson process with mean rate λ. Also, the call holding time is assumed to follow an exponential distribution with mean $1/\mu$. The cell residence time, the amount of time during which a call stays in a cell before handoff, is also assumed to follow an exponential distribution with mean $1/h$. Here h is also called as the handoff rate, which means how fast a call will hand off. As an ongoing call moves through the cellular network, the call will be allocated and reallocated various bandwidth values in B during its lifetime, depending on wireless/mobile network situations.

Normally, the adaptive framework aims to allocate as much bandwidth as possible to every call. Alternately said, a call will be allocated bandwidth b_n (the maximum bandwidth) whenever possible. However, network congestion may occur, in which case a cell cannot accommodate all calls with their maximum bandwidth b_n. In this case, one or more calls should be allocated/reallocated lower bandwidth than b_n. To choose which calls and how much bandwidth of the chosen calls to be changed is the role of the BAA.

A BAA that manages the allocation/reallocation of bandwidth of each call is necessary in this adaptive multimedia framework. According to the different QoS objectives, there can be diverse BAAs. Here we adopt a simple objective – to minimize the number of calls whose bandwidth is lower than predefined "target bandwidth". The target bandwidth will be denoted by b_{tar}, and it is assumed to take a predetermined value from the set B. Hereafter a call is referred to be "degraded" if and only if the bandwidth of the call is lower than b_{tar}. Alternately said, a user of a call is assumed to be satisfied if the bandwidth of the call is at least b_{tar}. In our single class network, all calls are assumed to have the same b_{tar}.

Therefore, the proposed BAA in the adaptive multimedia framework seeks to allocate at least the target bandwidth to every call in a cell whenever possible. However, as the traffic load in the wireless/mobile multimedia network becomes higher, there are more "degraded" calls, and thus, users will become dissatisfied. Thus, for the adaptive multimedia framework to be successful, a CAC algorithm that prevents the cellular system from being highly overloaded is crucial. Here a cell is referred to be "overloaded" if there are one or more "degraded" calls in the cell. More precisely, the *cell overload probability*, P_{CO}, is defined as the sum of probabilities for

states where there are one or more "degraded" calls. The proposed CAC algorithm should therefore enforce the *cell overload probability* to be less than a predetermined value, P_{qos}, which will be given as a QoS requirement.

3. The CAC algorithm

Now the problem is how the proposed CAC algorithm can guarantee the upper bound of the *cell overload probability* in a cell. We basically adopt the distributed CAC algorithm proposed in [12], which considers the number of ongoing calls in adjacent cells when a call request arrives in a given cell. We consider one dimensional cellular array such as one used in streets and highways. In addition, we assume that the cellular system uses a fixed channel allocation scheme and that each cell has the same bandwidth capacity (or the same number of channels).

Let us denote n, r, and l as the number of ongoing calls in the given cell (C_n), its right cell (C_r), and its left cell (C_l), respectively. Considering any test mobile in C_n, we assume that the test mobile remains in the same cell with probability p_r, and that it hands off to one of its adjacent cells (either C_r or C_l) with probability $p_h/2$ during time T which we refer to as the *estimation time*. Furthermore, we assume that the probability that a call hands off more than once during T is negligible.

In our model, we approximately estimate the *cell overload probability* of a radio cell after *estimation time* whenever a new call request arrives. Hereafter a cell is referred to be "overloaded" when the number of ongoing calls is greater than a threshold number (N_{th}) which is the maximum number of calls all of which can be allocated at least b_{tar} in a cell. That is, if the number of calls in a cell is greater than N_{th}, there are at least one or more calls whose bandwidth values are less than b_{tar}. Here N_{th} is calculated by $\lfloor C/b_{\text{tar}} \rfloor$ where C is the total bandwidth capacity of a cell in basic bandwidth units.

Accordingly, the cell overload probability, P_{CO}, is expressed by $\sum_{i=N_{\text{th}}+1}^{\infty} P(i)$ where $P(i)$ represents the probability of having i calls in a radio cell. Let us denote P_{qos} as the highest tolerable cell overload probability in wireless/mobile multimedia networks where all calls requires the same P_{qos} throughout their connection.

The CAC algorithm of such system must take two factors into consideration when admitting a new call: (1) by admitting the new call, the desired QoS (in terms of P_{qos}) of existing calls in the system must be maintained; and (2) the system must provide the new call with its desired QoS. To achieve the above objectives, we estimate the state of the given cell after T units of time when the new call arrives. Therefore, a new call is admitted to the given cell at time t_0 if and only if the following condition is satisfied:

At time $t_0 + T$, the overload probability of the given cell estimated by considering handoffs from its adjacent cells and handoffs from the given cell to adjacent cells must be smaller than P_{qos}.

Given the above assumptions, and denoting the number of calls in the given cell as n at time t_0, the probability that i calls (out of n calls) remain in the given cell at time $t_0 + T$ has a binomial distribution given by $B(i; n, p_r)$, and the probability that j calls (out of n calls) hand off from the given cell to the right-hand side of the test cell at time $t_0 + T$ is given by a binomial distribution $B(j; n, p_h/2)$, where $B(i; n, p)$ is defined as

$$B(i; n, p) = \binom{n}{i} p^i (1-p)^{n-i}. \tag{1}$$

Let us assume that by admitting a call to cell C_n at time t_0 there will be n, r, and l calls in cells C_n, C_r, and C_l, respectively. To satisfy the above call admission rule, we should find the probability distribution of the number of calls in C_n at time t_0+T denoted by $P_{t_0+T}(k)$ using a convolution sum of three binomial distributions $B(i_n; n, p_s)$, $B(i_r; r, p_h/2)$, and $B(i_l; l, p_h/2)$, where $n \geqslant i_n \geqslant 0$, $r \geqslant i_r \geqslant 0$, $l \geqslant i_l \geqslant 0$, and $k = i_n + i_r + i_l$. Recall that binomial distribution can be approximated by Gaussian distribution. According to [5], the convolution sum of binomial distributions can also be approximated by Gaussian distribution. Therefore, based on the independence assumption, the number of calls in cell C_n at time $t_0 + T$ has also Gaussian distribution given by

$$P_{t_0+T}(k) \simeq G\left(np_r + (l+r)\frac{p_h}{2}, \sqrt{np_r(1-p_r) + (l+r)\frac{p_h}{2}\left(1 - \frac{p_h}{2}\right)}\right). \tag{2}$$

Hence, the overload probability P_{CO} is given by the tail of the Gaussian distribution which can be calculated by

$$P_{\text{CO}} = \sum_{N_{\text{th}}+1}^{l+n+r} P_{t_0+T}(k) \simeq Q\left(\frac{N_{\text{th}} - (np_r + (l+r)p_h/2)}{\sqrt{np_r(1-p_r) + (l+r)(p_h/2)(1-p_h/2)}}\right). \tag{3}$$

Here $Q(\cdot)$ is the integral over the tail of a Gaussian distribution which can be expressed in terms of the error function [9]. Recall that P_{CO} is the sum of probabilities for states in which there are one or more calls with lower than target bandwidth in a cell. If P_{CO} is expected to be greater than predetermined QoS value (P_{qos}), then the arriving new call is rejected. Thereby the distributed CAC algorithm can enforce P_{CO} to be less than P_{qos}.

4. The BAA

To support the proposed CAC algorithm for QoS requirements, the bandwidth adaptation algorithm (BAA) seeks to minimize the number of calls with lower than *target bandwidth*, b_{tar}. Alternately said, it seeks to maximize the number of calls with equal to or more than b_{tar} at any instant.

There are broadly two cases where the BAA performs the bandwidth adaptation: for reduction and for expansion. The

Table 1
Notation for the BAA.

b_{tar}	target bandwidth
b_{min}	minimum bandwidth (b_1)
b_{max}	maximum bandwidth (b_n)
B_A	available bandwidth in the given cell
B_T	amount of squeezable bandwidth by changing all calls with more than b_{tar} into calls with b_{tar}
B_M	amount of squeezable bandwidth by changing all calls with more than b_{min} into calls with b_{min}

BAA for reduction applies to the case where a new call or an incoming handoff call arrives in the given cell and the cell is overloaded. Depending on the situation, the BAA allocates the suitable bandwidth to the incoming call (the new call or the handoff call) and reallocates the bandwidth of the existing calls, if necessary. The BAA for expansion may expand the bandwidth of calls with lower than target bandwidth to b_{tar} or more when there is an outgoing handoff call or a call completion in the given cell. The notation of the BAA is summarized in table 1.

The description of the BAA for reduction when a call (new or handoff) arrives in the given cell is detailed below. There are six cases in the BAA for reduction. Below, the operation ReduceT(b_{wanted}) squeezes the calls with more than b_{tar} to b_{tar} and adds the squeezed bandwidth to B_A until B_A equals to or exceeds b_{wanted}. Similarly, ReduceM(b_{wanted}) squeezes the calls with more than b_{min} to b_{min} until B_A equals to or exceeds the b_{wanted}.

(1) if ($B_A \geqslant b_{tar}$),
allocate the maximum b_i to the incoming call
($b_i \leqslant B_A$, $b_{tar} \leqslant b_i \leqslant b_{max}$)

(2) else if ($B_A < b_{tar}$ and $B_A + B_T \geqslant b_{tar}$),
ReduceT(b_{tar})
allocate the maximum b_i to the incoming call
($b_i \leqslant B_A$, $b_{tar} \leqslant b_i \leqslant b_{max}$)

(3) else if ($B_A \geqslant b_{min}$ and $B_A + B_T < b_{tar}$),
allocate the maximum b_i to the incoming call
($b_i \leqslant B_A$, $b_{min} \leqslant b_i < b_{tar}$)

(4) else if ($B_A < b_{min}$ and $B_A + B_T \geqslant b_{min}$),
ReduceT(b_{min})
allocate the maximum b_i to the incoming call
($b_i \leqslant B_A$, $b_{min} \leqslant b_i < b_{tar}$)

(5) else if ($B_A < b_{min}$ and $B_A + B_M \geqslant b_{min}$),
ReduceM(b_{min})
allocate the maximum b_i to the incoming call
($b_i \leqslant B_A$, $b_{min} \leqslant b_i < b_{tar}$)

(6) else drop/block the call

When a call leaves the cell, B_A increases. This change in B_A may make it possible for one or more calls to expand their bandwidth. The BAA for expansion is described below, where b_{cur} denotes the currently allocated bandwidth of the corresponding call and b_{req} denotes the required bandwidth to "upgrade" the bandwidth of the corresponding call. Here a call is referred to be "upgraded" if the bandwidth of the call is changed from lower than b_{tar} to b_{tar} or more.

(1) order the degraded calls by decreasing b_{cur}

(2) for each degraded call

(a) $b_{req} = b_{tar} - b_{cur}$

(b) if ($B_A \geqslant b_{req}$),
allocate the maximum b_i to the call
($b_i \leqslant B_A$, $b_{tar} \leqslant b_i \leqslant b_{max}$)

(c) else if ($B_A < b_{req}$ and $B_A + B_T \geqslant b_{req}$),
ReduceT(b_{req})
allocate the maximum possible b_i to the call
($b_i \leqslant B_A$, $b_{tar} \leqslant b_i \leqslant b_{max}$)

(3) repeat step (2) if upgrade is still possible

In the step (1), we order the degraded calls in a given cell according to the decreasing b_{cur}. However, there can be other ordering criteria such as the amount of degraded time, if available.

5. Numerical results

In this section, by simulation experiments, we present how the proposed CAC algorithm can guarantee QoS to users and compare the adaptive multimedia framework with the non-adaptive multimedia framework where the bandwidth of ongoing call is fixed throughout its lifetime. The bandwidth of a call in the non-adaptive multimedia networking paradigm is 10 (bandwidth units) in the simulation experiments. Whereas, there are three different bandwidth values in the set B in the adaptive multimedia framework (see table 2).

In the non-adaptive multimedia framework, P_{qos} represents the upper bound of the forced termination probability. The forced termination probability is the probability that the system (a cell) has no more available bandwidth for incoming calls. If a handoff call arrives in the cell where there are no more channels (more exactly, the available channels are less than 10 in this experiment), the call will be forced to be terminated instead of bandwidth adaptation.

The experimental results here are based on the simulation of a system consisting of 10 cells arranged in one dimension. The cells are configured to form a circle to remove the boundary effect. The probability of a user handing off to any adjacent cell is equally likely. Recall that the call duration is exponentially distributed with mean $1/\mu$, and that the time a call spends in a radio cell prior to handoff to another cell is also

Table 2
Simulation parameters.

b_1	1 ($b_{\min}$)
b_2	9 (b_{tar})
b_3	10 ($b_{\max}$)
$1/\mu$	500 s
$1/h$	100 s
T	20 s

exponentially distributed with mean $1/h$. Then, based on the above assumptions, p_r and p_h can be calculated as follows:

$$p_r = \mathrm{e}^{-(\mu+h)T}, \tag{4}$$

$$p_h = 1 - \mathrm{e}^{-hT}. \tag{5}$$

For the purposes of illustration, we assume an average call duration of 500 s, an average handoff time of 100 s, and an estimation period (T) of 20 s. As a result, p_r is 0.7866 and p_h is 0.1813. Note that $1 - (p_r + p_h)$ is the probability that a call departs from the cellular network in T units of time. The parameters in the simulation experiments are summarized in table 2.

Now, we present the performance of the proposed distributed CAC algorithm. As mentioned earlier, the proposed algorithm enforces P_{CO} to be less than the predetermined value, P_{qos}. Figure 1 shows P_{CO} as Erlang load increases. The P_{CO} is measured by monitoring the state of the cellular network for every second. It represents the portion that the cell is overloaded during the whole measurement period. Here total number of bandwidth units (channels) in each cell (C) is 500. Note that P_{CO} approaches P_{qos} as the offered load increases. Obviously, P_{CO} does not exceed the value of P_{qos}, which proves that the proposed CAC algorithm can guarantee QoS to users.

In comparing the adaptive multimedia framework with the non-adaptive multimedia framework, we adopt the forced termination probability P_{F} and the call blocking probability P_{B} as performance measures. Figures 2 and 3 show the forced termination probability in the non-adaptive multimedia framework when the bandwidth capacity of each cell is 200 and 500 bandwidth units, respectively. Obviously, the forced-termination probability increases as Erlang load and/or P_{qos} increases. In the adaptive multimedia framework, the forced termination probability is 0 due to the simulation configuration that the minimum bandwidth of a call occupies only one bandwidth unit ($b_1 = 1$), which highlights the advantage of the adaptive multimedia framework.

In figure 4, the call blocking probability of the proposed CAC algorithm in both frameworks is shown as the offered load (in Erlangs) increases when C is 200. Note that the call blocking probability of the non-adaptive multimedia framework is lower than that of adaptive multimedia. It is mainly because there are already many calls in each cell in the adaptive multimedia paradigm as there is no forced-termination (no handoff failure). Note that the call blocking probability notably decreases in the non-adaptive multimedia framework as P_{qos} increases. This phenomenon results from the fact that the more cell overload probability is allowed, the more calls

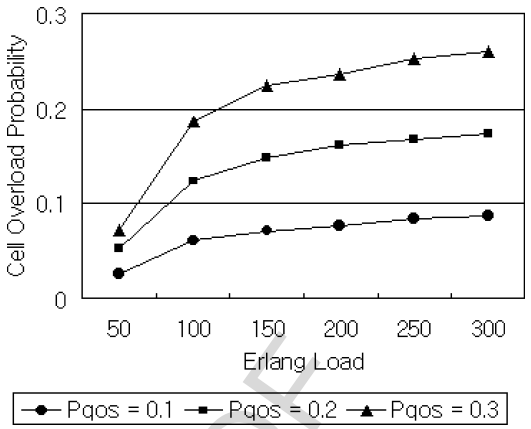

Figure 1. The cell overload probability.

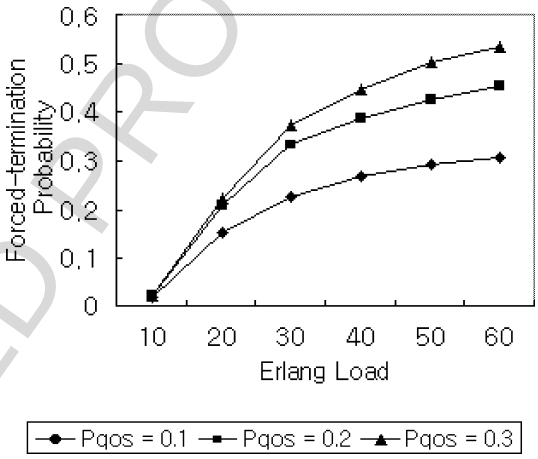

Figure 2. The forced termination probability ($C = 200$).

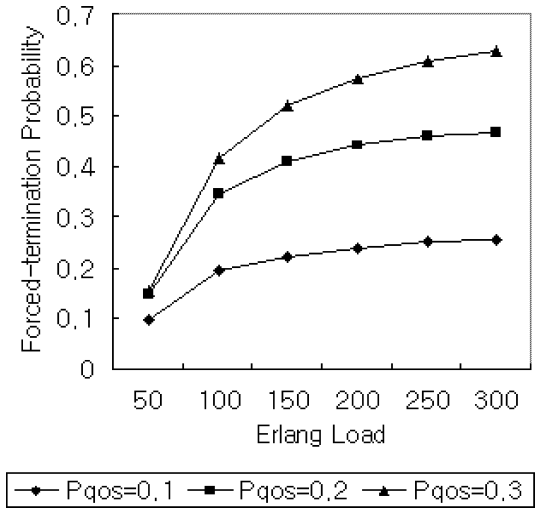

Figure 3. The forced termination probability ($C = 500$).

are accepted by the system and thereby more calls will be forced to be terminated. Again, the forced-terminated calls will make idle channels for newly arriving calls. As a result, the call blocking probability decreases as P_{qos} increases. Whereas, the call blocking probability in the adaptive multimedia framework does not change much despite the vari-

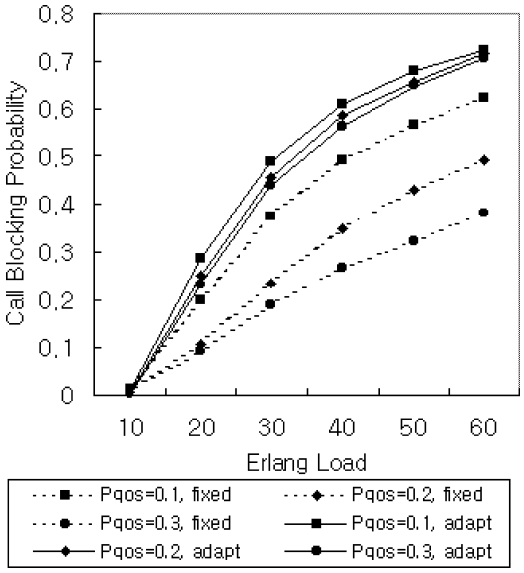

Figure 4. The call blocking probability ($C = 200$).

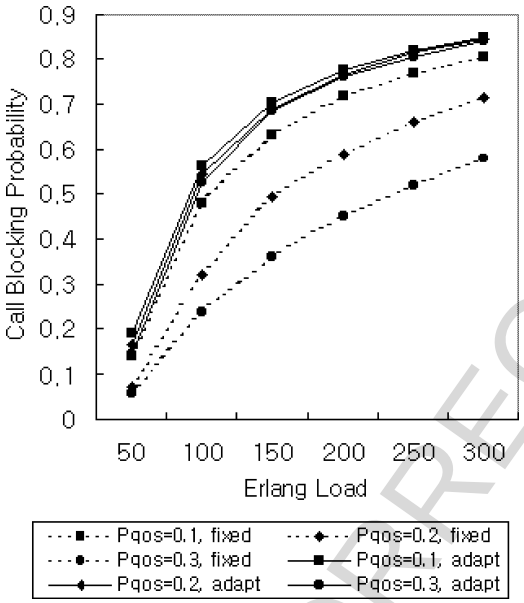

Figure 5. The call blocking probability ($C = 500$).

ation of P_{qos}. This results from the fact that as there is no handoff failure, there is less fluctuation in available channels in each cell. Figure 5 shows similar results when C is 500.

We define another performance measure in this section: "effective utilization". Here "effective utilization" represents the ratio of the bandwidth used by completely serviced calls to the total bandwidth capacity. If a call is forced to be terminated before the completion of service, the bandwidth used by the call is not taken into account. Figure 6 shows that the effective utilization of the adaptive multimedia framework outperforms that of the non-adaptive multimedia framework when there are 200 channels in a cell. Note that effective utilization in the non-adaptive multimedia framework decreases as the offered load increases. Also, effective utilization in the non-adaptive multimedia becomes smaller as P_{qos} increases. This phenomenon can be explained similarly as in figures 4 and 5. Recall that in the non-adaptive multimedia networking paradigm, the more calls are accepted, the more calls are forced to be terminated before the completion of service. Figure 7 shows effective utilization in the case of 500 channels in each cell, which also proves the above reasoning.

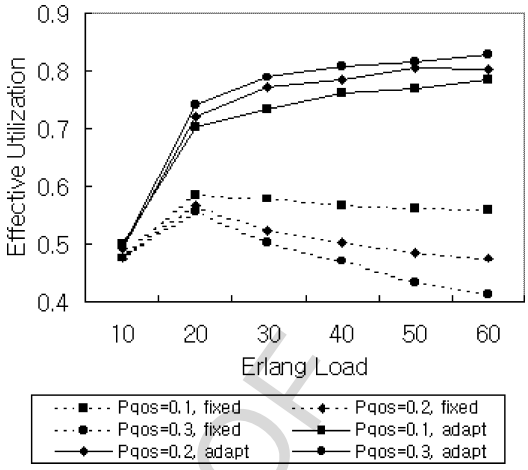

Figure 6. Effective utilization ($C = 200$).

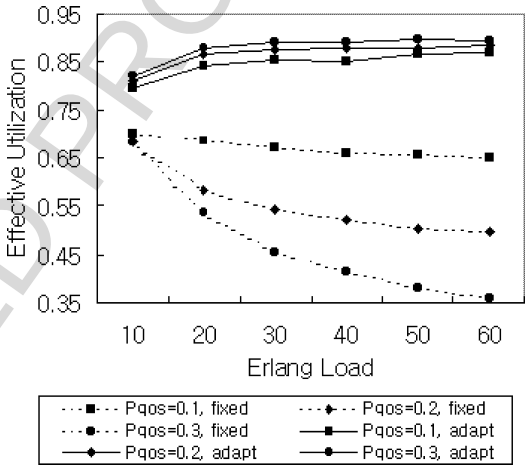

Figure 7. Effective utilization ($C = 500$).

To summarize, although the call blocking probability of the adaptive multimedia framework is greater than that of the non-adaptive multimedia framework, the overall performance of the adaptive multimedia networking is very attractive in that the forced termination probability is negligible and effective utilization increases as the offered traffic load increases.

6. The forced termination probability

In this section, we analyze the forced termination probability in the adaptive multimedia framework. According to [10,15], the forced termination probability is almost directly proportional to the handoff dropping probability. Thus, we con-

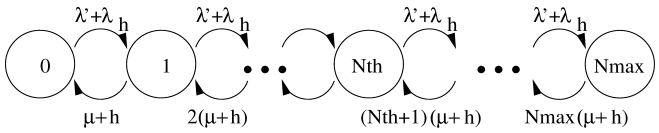

Figure 8. State transition diagram.

sider the handoff dropping probability instead of the forced termination probability. Here the handoff dropping probability means the probability that an attempt of handoff will fail.

As mentioned earlier, the handoff dropping probability in our adaptive framework is negligible in normal cases. However, theoretically it is possible that a handoff fails if a given cell is full of calls with $b_{\min}$ and there are not sufficient channels to allocate $b_{\min}$ to an incoming handoff call. That is, in the adaptive framework, the handoff dropping probability equals the steady state probability of $P(N_{\max})$. Here $N_{\max}$ is the maximum possible number of calls in a cell with bandwidth adaptation. The state-transition diagram in a cell can be described by figure 8.

Here λ_h denotes the handoff call arrival rate into a cell and will be discussed later. Furthermore, λ' denotes the reduced new call arrival rate. It can be approximated by the fixed point approximation [8,14] as follows:

$$\lambda' = \lambda(1 - P_{\mathrm{B}}). \tag{6}$$

Recall that P_{B} is the probability that a newly arriving call will be blocked; it can be calculated by considering the steady state probabilities of three cells (the given cell and two adjacent cells) in one-dimensional cellular network. Let the function $test(n, r, l)$ be the CAC function as described in section 3 which returns 0 when the newly arriving call should be rejected. Then, P_{B} is calculated by

$$P_{\mathrm{B}} = \sum_{n=0}^{N_{\max}} \sum_{r=0}^{N_{\max}} \sum_{l=0}^{N_{\max}} \big(P(n)P(r)P(l)\big),$$
$$\text{where} \quad test(n, r, l) = 0.$$

Accordingly, we can figure out the steady state probabilities of a cell and thereby calculate the handoff dropping probability by

$$P_{\mathrm{HD}} = P(N_{\max}) = \frac{\prod_{i=0}^{N_{\max}-1} (\lambda' + \lambda_h)/(i+1)(\mu + h)}{\sum_{j=0}^{N_{\max}-1} \prod_{i=0}^{j} (\lambda' + \lambda_h)/(i+1)(\mu + h)}. \tag{7}$$

Similar to the fixed point approximation in [8,14], we can solve the above equations by repeated substitutions.

7. Degradation period ratio

With the adaptive framework, we could ignore the forced termination probability in normal traffic load, as it can be made practically zero (see section 5). In this section, to quantify the level of the degradation of a call, we define another QoS parameter: the *degradation period ratio*. The *degradation period ratio* (DPR) is the portion of a call's lifetime during which the allocated bandwidth is lower than the target bandwidth with respect to the whole service time of the call. For example, if a call's DPR is 0.5, the period while the call's allocated bandwidth is less than the target bandwidth is half of the call's lifetime.

Recall that a state is defined by the number of calls in each cell at a given instant. A call may experience a number of states throughout its lifetime. The residence time in a state represents the time interval between every instant of a new call arrival, a call departure, or an incoming/outgoing handoff call. We assume that the time between every state transition follows an exponential distribution with mean rate r. Here r is the state transition rate which reflects how fast the state of the system (a cell) will move to another state.

The state transition rate is a function of the new call arrival rate λ, the handoff call arrrival rate λ_h, the call service rate μ, and the handoff rate h. More exactly, the rate r can be calculated from the effective new call arrival rate, the handoff call arrival rate (see [10]), the call service rate, and the handoff rate as follows:

$$r = \lambda(1 - P_{\mathrm{B}}) + \lambda_h + E[n]\mu + E[n]h. \tag{8}$$

Here, $E[n]$ denotes the average number of calls in each cell. Furthermore, according to [10], λ_h can be expressed by (9) where P_{HD} is the handoff dropping probability:

$$\lambda_h = \lambda(1 - P_{\mathrm{B}}) \frac{h}{\mu + h P_{\mathrm{HD}}}. \tag{9}$$

An accepted call may experience a number of state transitions during its lifetime as (1) a new call is accepted, (2) a call is handed off from adjacent cells to a given cell, (3) calls are terminated, or (4) a call in a given cell is handed off to adjacent cells. Using r, we can calculate the probability distribution of how many states a call will experience throughout its lifetime.

Suppose that a call resides in the ith state during a time period t_i ($i = 1, 2, \ldots$). Then, the time between state change follows the exponential distribution with mean rate r

$$P[t_i \leqslant t] = 1 - \mathrm{e}^{-rt}. \tag{10}$$

Let us denote the call service time by the random variable τ and the number of state transitions experienced by a call by the random variable K. Then, the probability of a k-state call (a call will experience k states during its lifetime) is given as follows.

For $k = 1$,

$$P[K = k] = P[\tau \leqslant t_1] = \int_{t_1=0}^{\infty} \int_{\tau=0}^{t_1} \mu \mathrm{e}^{-\mu\tau} r \mathrm{e}^{-rt_1} \,\mathrm{d}\tau \,\mathrm{d}t_1 = \frac{\mu}{\mu + r}. \tag{11}$$

For $k \geqslant 2$,

$$P[K=k] = P\left[\sum_{i=1}^{k-1} t_i < \tau \leqslant \sum_{i=1}^{k} t_i\right]$$
$$= \int_0^\infty \int_0^\infty \cdots \int_0^\infty \int_{\tau=t_{k-1}}^{t_k} \mu e^{-\mu\tau} r e^{-rt_k} \cdots$$
$$\times r e^{-rt_2} r e^{-rt_1} \, d\tau \, dt_k \ldots dt_2 dt_1$$
$$= \left(\frac{r}{\mu+r}\right)^{k-1} \frac{\mu}{\mu+r}. \quad (12)$$

In general, the probability of a call will experience k states is $(r/(\mu+r))^{k-1}\mu/(\mu+r)$ for $k = 1, 2, \ldots$.

At each state, a call will be allocated/reallocated bandwidth which is lower than the target bandwidth or not. Recall that a call is referred to be "degraded" if the bandwidth of the call is less than the target bandwidth. Let P_D be the "degradation probability" that a call will be allocated bandwidth lower than the target bandwidth in a state. If we assume that the state transition process of a call is a discrete-time Markov process and every state is independent of each other, then, for a call experiencing k states, the number of states with lower than target bandwidth will follow a binomial distribution $B(k, P_D)$. Let X_k denote the random variable of this binomial distribution which represents the number of degraded states of a k-state experiencing call. Then we can calculate the probability distribution of degradation period ratio (DPR) as in (13). Here X is a random variable that represents the expected DPR of a call:

$$P[X \geqslant \alpha] = \sum_{k=1}^{\infty} P[K=k] P\left[X_k \geqslant \lceil \alpha k \rceil\right]. \quad (13)$$

Here, if k is large, the binomial distribution of X_k can be approximated by a Gaussian distribution $G(m_k, \sigma_k{}^2)$. Here $m_k = kP_D$ and $\sigma_k^2 = kP_D(1-P_D)$. As a result, $P[X_k \geqslant \lceil \alpha k \rceil]$ can be approximated by $Q((\alpha k - m_k)/\sigma_k)$. Therefore, (13) can be rewritten as

$$P[X \geqslant \alpha] = \sum_{k=1}^{k=N} P[K=k] P\left[X_k \geqslant \lceil \alpha k \rceil\right]$$
$$+ \sum_{k=N+1}^{k=\infty} P[K=k] Q\left(\frac{\alpha k - m_k}{\sigma_k}\right). \quad (14)$$

Here N is sufficiently large (say, 100). We think that the ultimate QoS to users in the adaptive multimedia framework can be expressed by

$$P[X \geqslant \alpha] \leqslant \beta. \quad (15)$$

Finally, for given QoS parameters α and β, we should find out the maximum P_D that satisfies (15).

8. Conclusion

It is anticipated that multimedia applications with the adaptive networking framework where the bandwidth of an ongoing call can be dynamically adjusted will become widespread in the near future. Although the forced termination probability can be reduced to a negligible level in this adaptive framework, a new kind of QoS provisioning is still required. In this paper, we have defined a QoS parameter, the *cell overload probability*, from the perspective of the adaptive multimedia networking paradigm. The key factor in QoS provisioning is a CAC algorithm which enforces the upper bound of the *cell overload probability*. We have discussed how the proposed CAC algorithm can guarantee QoS in wireless/mobile multimedia networks. Also, we proposed the bandwidth adaptation algorithm to minimize the *cell overload probability* at any time. Numerical results show that the proposed distributed CAC algorithm guarantees the upper bound of the *cell overload probability*. In addition, simulation experiments are conducted to highlight the performance of the adaptive multimedia framework compared to that of the non-adaptive multimedia framework. Even though the functional requirements of the adaptive multimedia networking framework are much more complicated than those of the non-adaptive multimedia framework, we believe that it is worthwhile to deploy the adaptive framework in the wireless/mobile multimedia networks, especially considering the scarcity and fluctuation of wireless link bandwidth. In this study, only a single class of the adaptive multimedia networking has been investigated. As a future work, we will extend the proposed CAC algorithm for the case of multiple classes where there are various adaptive multimedia streams/encodings in integrated services networks.

References

[1] A. Alwan et al., Adaptive mobile multimedia networks, IEEE Communications Magazine 34(4) (April 1996) 34–51.

[2] V. Bharghavan, K. Lee, S. Lu, S. Ha, J. Li and D. Dwyer, The TIMELY adaptive resource management architecture, IEEE Personal Communications Magazine 5(8) (August 1998).

[3] S.K. Das and S.K. Sen, Quality-of-Service degradation strategies in multimedia wireless networks, in: *IEEE Vehicular Technology Conference (VTC'98)*, Ottawa, Canada (May 1998).

[4] J. Gomez, A. Campbell and H. Morikawa, A systems approach to prediction, compensation and adaptation in wireless packet networks, in: *Proc. of ACM/IEEE International Workshop on Wireless and Mobile Multimedia (WoWMoM'98)*, Dallas (October 1998) pp. 92–100.

[5] R. Howard, *Dynamic Probabilistic Systems*, Vol. 1, *Markov Models* (Wiley, 1971).

[6] N. Yeadon, Filters: QoS support mechanisms for multipeer communications, IEEE Journal on Selected Areas in Communications 14(7) (September 1996) 1245–1262.

[7] ITU-T, Recommendation H.263, Video codec for low bitrate communication (1996).

[8] F.P. Kelly, Fixed point models of loss networks, Journal of Australian Mathemathical Society B(31) (1989) 204–218.

[9] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering* (Addison Wesley, 1994).

[10] Y.-b. Lin et al., Queueing priority channel assignment strategies for PCS hand-off and initial access, IEEE Transactions on Vehicular Technology 43(3) (August 1994) 704–712.

[11] S. Lu et al., Adaptive service in mobile computing environments, in: *IWQOS'97*, New York (1997).

[12] M. Naghshineh and M. Schwarz, Distributed call admission control in mobile/wireless networks, in: *IEEE PIMRC'95* (1995) pp. 289–293.
[13] M. Naghshineh and M. Willebeek-LeMair, End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework, IEEE Communications Magazine 35(11) (November 1997) 72–81.
[14] D.L. Pallant, A reduced load approximation for cellular mobile networks including handovers, Australia Telecommunications Research 26(2) (1992) 21–29.
[15] S.S. Rappaport, The multiple-call hand-off problem in high-capacity cellular communications systems, IEEE Transactions on Vehicular Technology 40(3) (August 1991) 546–557.
[16] S. Sen et al., Quality of Service degradation strategies in multimedia wireless networks, in: *IEEE VTC'98*, Ottawa (1998) pp. 1884–1888.
[17] N. Shacham, Multipoint communication by hierarchically encoded data, in: *IEEE INFOCOM'92* (1992) pp. 2107–2114.
[18] A. Sutivong and J.M. Peha, Novel heuristics for call admission control in cellular systems, in: *IEEE ICUPC'97* (1997) pp. 129–133.
[19] A.K. Talukdar, B.R. Badrinath and A. Acharya, Rate adaptation schemes in networks with mobile hosts, in: *ACM/IEEE MobiCom'98* (October 1998).

Taekyoung Kwon is currently a post-doctoral researcher in Computer Science Department at UCLA. He received his Ph.D., M.S., and B.S. degree in computer engineering from Seoul National University in 2001, 1995, 1993, respectively. He was a visiting student at IBM T.J. Watson Research Center in 1998 and was a visiting scholar at University of North Texas in 1999. He was also a part-time lecturer at Hanyang University in 2001, teaching the Internet. He has been working on Call Admission Control in Mobile Cellular Networks, Adaptive Multimedia Networking. His recent research area includes wireless technology convergence and mobility management. He has published over 20 technical papers on wireless/mobile communications and networking.
E-mail: tkkwon@mmlab.snu.ac.kr

Yanghee Choi received B.S. in electronics engineering from Seoul National University, M.S. in electrical engineering from Korea Advanced Institute of Science, and Doctor of Engineering in computer science from École Nationale Supérieure des Telecommunications (ENST) in Paris, in 1975, 1977 and 1984 respectively. Before joining the School of Computer Engineering, Seoul National University in 1991, he has been with Electronics and Telecommunications Research Institute (ETRI) during 1977–1991, where he served as a director of Data Communication Section and Protocol Engineering Center. He was a research student at Centre National d'Étude des Telecommunications (CNET), Issy-les-Moulineaux, during 1981–1984. He was also a Visiting Scientist to IBM T.J. Watson Research Center for the year 1988–1989. He is now leading the Multimedia Communications Laboratory in Seoul National University. He is also a director of Computer Network Research Center in Research Institute of Advanced Computer Technology (RIACT). He was an editor-in-chief of Korea Information Science Society journals. He was a chairman of the Special Interest Group on Information Networking. He has been an associate dean of research affairs at Seoul National University. He is now the President of Open Systems and Internet Association of Korea. His research interest lies in the field of multimedia systems and high-speed networking.
E-mail: yhchoi@mmlab.snu.ac.kr

Chatschik Bisdikian is a Research Staff Member and a manager of the Wireless Services (WISE) Platforms group at IBM's T.J. Watson Research Center, Hawthorne, NY. He received a Ph.D. degree in electrical engineering from the University of Connecticut in 1988 and he has been with IBM ever since. His research interest includes short-range wireless networks, service discovery, spontaneous networking, content delivery, multimedia/broadband communications, and related areas. He has been involved with the development of the Bluetooth protocol specification from its early stages. He is participating in the standardization of the Bluetooth specification within IEEE 802.15 and serves as a vice-chair of the IEEE 802.15.1 task group. He is a senior member of IEEE and has served in the editorial boards of several technical publications. He has authored over 80 technical, peer-reviewed papers, and holds three patents. He is a co-author of *Bluetooth Revealed*, published by Prentice-Hall PTR (2001). He is a 1995 Eta Kappa Nu Outstanding Young EE Award program finalist.
E-mail: bisdik@watson.ibm.com

Mahmoud Naghshineh is a Senior Manager at the IBM T.J. Watson Research Center, Yorktown Heights, NY, where he currently manages the Pervasive Security and Networking Department. He joined IBM in 1988. Since then, he has worked on communication and networking protocols, fast packet-switched/broadband IP and ATM networks, short range Infrared/RF wireless and mobile networking, optical networking, QoS provisioning, call admission, routing, and resource allocation, network security and secure co-processors. He has had several main technical contributions to IBM products in areas of networking technologies and software. He has contributed to IrDA, Bluetooth and IEEE standards. His is currently responsible for IBM research activities in areas of Wireless Internet and Security Systems. He received his doctoral degree from Columbia University, New York, in 1994, Master of Science in electrical engineering from Polytechnic University, and Vordiplom from Technical University of Aachen, Germany. He is a senior member of the IEEE and the Editor-in-Chief of IEEE Personal Communications Magazine. He has served as a technical editorial board member of many wireless and mobile networking/computing journals, as a member of technical program committee, session organizer and chairperson for many IEEE/ACM, NSF and Government conferences and workshops. Currently, he is an adjunct faculty member of the Department of Electrical Engineering at Columbia University teaching a graduate course on wireless and mobile networking. He has published over 60 technical papers and holds a number of IBM awards and patents.
E-mail: mahmoud@watson.ibm.com

VTEX() PIPS No:5101561 artty:res (Kluwer BO v.2002/08/19)
n5101561.tex; 12/09/2002; 8:44; p. 9