A Dynamic Multiple-Threshold Bandwidth Reservation (DMTBR) Scheme for QoS Provisioning in Multimedia Wireless Networks

Xiang Chen, *Student Member, IEEE*, Bin Li, and Yuguang Fang, *Senior Member, IEEE*

***Abstract*—Next-generation wireless networks target to provide quality of service (QoS) for multimedia applications. We study the wireless systems that support two QoS requirements: keeping the handoff dropping probability less than a predefined QoS threshold while maintaining relative priorities for different traffic classes based on blocking probability. To achieve this goal, a dynamic multiple-threshold bandwidth reservation (DMTBR) scheme, which is capable of granting differential priorities to different traffic classes and to new and handoff traffic for each class by dynamically adjusting bandwidth reservation thresholds, is proposed. In this scheme, the thresholds are obtained in two steps. The initial values are estimated based on instantaneous network traffic situation, then the thresholds will be further adapted according to the instantaneous network QoS status. In times of network congestion, a preventive measure is taken to throttle the new connections. Another contribution of this paper is to generalize the concept of relative priority and hence give the network operator more flexibility to adjust admission control policy by taking into account some dynamic factors such as offered load. The extensive simulations are conducted for two purposes. First, we verify the performance of the proposed scheme and show our scheme performs well under various traffic loads. Second, we demonstrate that the DMTBR scheme gains more advantages when taking the offered load into consideration.**

***Index Terms*—Connection admission control, mobile wireless networks, quality of service (QoS).**

I. Introduction

With increasing demands for mobile multimedia services such as audio, video, and data, next-generation wireless networks are expected to provide quality of service (QoS) for such multimedia applications to users on the move. Since the multimedia services have inherently different traffic characteristics, their QoS requirements may differ in terms of bandwidth, delay, and connection dropping probabilities. It is the networks' responsibility to fairly and efficiently allocate network resources among different users to satisfy such differentiated QoS requirements for each type of service.

Manuscript received March 17, 2003; revised November 12, 2003; accepted December 31, 2003. The editor coordinating the review of this paper and approving it for publication is B. Li. This work was supported in part by the U.S. National Science Foundation under Faculty Early Career Development Award ANIR0093241 and under Grant ANI-0220287.

X. Chen and Y. Fang are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: xchen@ecel.ufl.edu; fang@ece.ufl.edu).

B. Li is with the China Motion Telecom Group, Kowloon Bay, Kowloon, Hong Kong, China (e-mail: bin.li@chinamotion.com).

Digital Object Identifier 10.1109/TWC.2004.843053

In wireless cellular networks, a base station (BS) serves mobile users in the cell covered by this BS. When a user engaging a call connection moves from one cell to another, a smooth handoff has to be made to provide uninterrupted service to the previously established connection. If the destination cell does not have enough resources (i.e., wireless channels), the ongoing connection is forced to terminate before normal completion. Since mobile users are more sensitive to the termination of an ongoing connection than the blocking of a new call connection, handoff call connections are usually given higher priority over the new call connection. In this paper, we consider two types of traffic: real-time (rt) traffic and nonreal-time (nrt) traffic. As real-time traffic has a stringent requirement on time delay, it demands higher priority over nonreal-time traffic.

Since radio spectral resource is scarce and valuable in wireless networks, efficient connection admission control (CAC) schemes have to be designed to guarantee QoS requirements. During the past decade, intensive research has been done and various handoff priority-based CAC schemes have been investigated.

In [1]–[4], the well-known guard channel (cutoff priority) scheme and its variations were proposed to give higher priority to handoff connections over new connections by reserving a number of channels called *guard channels* for handoff call connections. All these schemes are static in the sense that the number of guard channels is determined mainly based on *a priori* knowledge of the traffic patterns, thereby being unable to cope with network dynamics. Moreover, only one traffic class, i.e., voice traffic, is considered. There have been a few guard channel-based schemes supporting voice and data traffic in an integrated mobile network [5]–[8]. In [9], Chao and Chen developed a multiple-class call model with user mobility, where various guard channel schemes can be used to satisfy connection-level QoS requirements for different traffic class. Using the stochastic Petri net (SPN) model, Li *et al.* [10] proposed and analyzed a hybrid cutoff priority scheme supporting different QoS requirements for multiclass traffic by adopting multiple thresholds. Similar to the guard channel scheme, these schemes are static as the boundaries and thresholds are always predetermined and lack adaptability.

However, a network with time-varying traffic may require dynamic bandwidth allocation schemes. In [11], Naghshineh and Schwartz developed a distributed call admission control (DCA)

scheme. Epstein and Schwartz extended this scheme to address the QoS requirements of multiclass traffic in wireless cellular networks [12]. By using the predefined call blocking probability profile, their scheme can not only guarantee maximum call dropping probabilities, but also maintain the relative priorities among different classes of traffic. However, since the profiles are predefined, their scheme may not adapt to dynamic traffic makeup. Besides, due to its underlying assumption, the performance may be sensitive to network load, especially in the heterogeneous case. Compared with DCA, a stable dynamic call admission control mechanism (SDCA) [13] is shown to achieve higher channel utilization while satisfying a predetermined bound on the call-dropping probability, since this scheme considers the influences of limited channel capacity and time dependence as well as the influences from the nonneighboring cells. To avoid the frequent information exchanges among cells required by SDCA, Li *et al.* [14] proposed an online local estimation algorithm, which can obtain almost the same performance as that of SDCA without incurring signaling overhead. However, these schemes cannot be directly applied to multiclass traffic scenarios. Ramanathan *et al.* proposed a dynamic resource allocation scheme in [15]. By probabilistically estimating the potential number of handoff connections from neighboring cells and thus reserving the bandwidth needed, the dropping probability of handoff connections is reduced. The drawback of this scheme is that the priority among different traffic classes, either in handoff traffic or new traffic, is not addressed. Oliver *et al.* proposed an adaptive bandwidth reservation scheme in [16]. Although real-time and nonreal-time traffic are differentially treated, the fairness issue is not considered. Hou and Fang used the concept of influence curve to determine the extent of the influence on neighboring cells by each mobile user and reserved the bandwidth in proportion to it [17]. A resource reservation algorithm based on the shadow cluster concept was developed by Levine *et al.* in [18]. Like [17], its performance depends heavily on the detailed knowledge of users' moving pattern.

In this paper, we propose a dynamic multiple-threshold reservation scheme for multimedia mobile wireless systems. The objective of the proposed scheme is twofold. The scheme first provides QoS provisioning by keeping the handoff connection-dropping probability (CDP) below the predefined bound even under a network congestion situation. Second, in a fair manner, it maintains the relative priorities among real-time traffic and nonreal-time traffic in terms of the new connection blocking probability (CBP) according to their traffic profiles and instantaneous traffic situations. To meet the objectives stated above, three bandwidth reservation thresholds are used to grant differential treatments to different traffic classes and to new and handoff traffic belonging to the same traffic class. The thresholds are dynamically adjusted according to the current network traffic situation and QoS status. When the network is under heavy traffic load, to guarantee QoS, we leverage cooperation among neighboring cells. One cell will inform its neighbors to throttle the acceptance of new connection requests, thereby reducing the potential handoff connections to the cell.

In our proposed scheme, a BS needs to communicate with its neighbors to acquire the necessary information for updating these thresholds periodically. Given the amount of information exchanged, however, this will not pose a serious problem on BSs, because they are usually interconnected with high-speed communication links.

In summary, this scheme has the following features.

1) It gives differential priorities to both new and handoff connections with different types of services by keeping multiple bandwidth reservation thresholds, based on which connection admission decision is made.
2) It maintains the relative priorities and fairness among traffic classes by taking the user QoS profile and real traffic conditions into account.
3) It generalizes the concept of relative priorities and fairness among traffic classes. By doing so, we can further reduce the CBPs of some real-time services while not seriously deteriorating the QoS of nonreal-time services.
4) It uses the information in the current cell and in the adjacent cells to periodically update the reservation thresholds, hence is able to respond to the changing network conditions quickly and effectively.

The rest of this paper is organized as follows. Section II describes the traffic model we will use for our study. Our scheme, including the target QoS criteria, is presented in Section III. Then, the threshold adaptation in the scheme is addressed in detail in Section IV. In Section V, the scheme is evaluated through extensive simulations studies. Finally, this paper is concluded in Section VI.

II. Traffic Model

The system under consideration is a wireless multimedia network with a cellular infrastructure, comprising of a number of cells. Each cell is served by a BS, and BSs are interconnected using high-speed communication links. We assume the system uses fixed channel assignment (FCA), which means that each cell has a fixed amount of capacity. Note that no matter which multiple-access technology (frequency-division multiple access, time-domain multiple access, or code-division multiple access) is used, we could interpret system capacity in terms of bandwidth. As in [15], we recognize that connections with different traffic classes may differ in their traffic characteristics, such as constant bit rate, variable bit rate, and peak bit rate and the desired QoS guarantees in terms of delay bound, loss rate, or throughput. For instance, the bandwidth of data service such as web browsing may vary with time. In this paper, we assume a single number, "effective bandwidth" [19], is adequate for guaranteeing the desired QoS for any connection with certain traffic characteristics. Hereafter, whenever we refer to the bandwidth of a connection, we mean its effective bandwidth. We assume each cell has C bandwidth units (BU). There are two classes of incoming traffic: 1) Class I is real-time traffic and 2) Class II is nonreal-time traffic. Typically, Class I traffic includes voice and video service, whereas Class II traffic consists of data services such as email, file transfers, and web browsing. We assume that the arrivals are Poisson processes, with respective arrival rates λ_{rt} and λ_{nrt}. The connection duration times of both connections follow exponential distributions, with means $1/\mu_{\mathrm{rt}}$ and $1/\mu_{\mathrm{nrt}}$, respectively. Furthermore, we assume that

the cell residence time distributions of these two kinds of connections are also exponentially distributed, with means $1/\gamma_{\rm rt}$ and $1/\gamma_{\rm nrt}$, respectively. The number of BUs required by each real-time and nonreal-time connection is $BW_{\rm rt}$ and $BW_{\rm nrt}$, respectively. These assumptions are appropriate and commonly used in the literature.

Following the assumptions, we can easily conclude that for these two types of traffic, the channel holding time (CHT), which is defined as the minimum of connection duration time and cell resident time, is also exponentially distributed, with the mean of $d_{\rm rt} = 1/(\mu_{\rm rt} + \gamma_{\rm rt})$ and $d_{\rm nrt} = 1/(\mu_{\rm nrt} + \gamma_{\rm nrt})$, respectively [20], [21].

III. Proposed Scheme

To meet different QoS requirements for real-time and nonreal-time traffic, we have to take into account their distinct traffic characteristics. Real-time services are sensitive to delay, whereas nonreal-time traffic could tolerate some delay without deteriorating service quality perceived by the mobile users. Moreover, handoff connections should receive higher priority than new connections for the same type. Therefore, it is necessary that they receive differential treatments in terms of access to the network resource. This is the rationale behind our scheme, which completely differentiates connection requests by setting up three reservation thresholds and dynamically adapting their values to network traffic conditions. Next, we describe the two QoS criteria we consider in this paper, followed by a detailed description of our proposed scheme.

A. QoS Criteria

The first criterion is to set an upper bound for the handoff dropping probability, i.e., the probability of a handoff connection being dropped. The criterion is satisfied as long as the following two inequalities are satisfied:

$$\begin{aligned} P_{\rm d,rt} &\le {\rm QoS}_{\rm rt} \\ P_{\rm d,nrt} &\le {\rm QoS}_{\rm nrt} \end{aligned} \tag{1}$$

where $P_{\rm d,rt}({\rm QoS}_{\rm rt})$ and $P_{\rm d,nrt}({\rm QoS}_{\rm nrt})$ are CDPs (the maximum allowable CDPs) for real-time and nonreal-time traffic, respectively.

The second criterion is to maintain the relative priority among different type of traffic in terms of CBPs. When there is no such criterion, the CBP for new connections will not be the same. The traffic classes that require smaller bandwidth will have a lower CBP as compared to those that require larger bandwidth. Obviously, this is unfair to the traffic classes that require larger bandwidth. To address this problem, we assume in the traffic profile for each traffic class, there exists a parameter called the traffic priority weight W, indicating the priority level for the traffic class. This parameter can be set through the negotiation between the user and the network operator and traffic characteristics are taken into account. A smaller weight means a higher priority. To achieve fairness among all traffic classes, the network thus needs to satisfy

$$\frac{P_{\rm b,rt}}{P_{\rm b,nrt}} = \frac{W_{\rm rt}}{W_{\rm nrt}} \tag{2}$$

where $P_{\rm b,rt}(W_{\rm rt})$ and $P_{\rm b,nrt}(W_{\rm nrt})$ are CBPs (the predefined traffic priority weight) for real-time and nonreal-time traffic, respectively.

Normally, there are many factors that affect the CBP. The actual CBP for each traffic class depends on the network capacity, the offered traffic load of each traffic class, the priority of each traffic class, the admission policy adopted to fulfill the QoS criteria related to the handoff traffic, the action the network may take in times of congestion, and so on. While some factors such as the network capacity or the preassigned traffic priority could be static, the offered load and the actions taken to deal with network congestion may be dynamic. In this sense, the criterion in (2) is static since it fails to reflect the real time network situation. Therefore, we generalize the concept of relative priority and propose a more general way to meet the second QoS criterion, i.e., maintain the relative priority among different traffic classes. We use the following equation for this purpose:

$$\frac{P_{\rm b,rt}}{P_{\rm b,nrt}} = \alpha \frac{W_{\rm rt}}{W_{\rm nrt}}. \tag{3}$$

Comparing to (2), we add one factor α on the right-hand side of (2). α can be thought of as a function of some of the dynamic factors mentioned earlier, representing the network's real traffic conditions or some procedures responding to traffic changes or QoS status.

To take into account network traffic load, we allow α to be a function of the offered load, a commonly used measure of traffic load. More precisely, the offered load per cell for each traffic class can be defined as the product of traffic arrival rate, the connection holding time, and the normal bandwidth

$$\begin{aligned} OL_{\rm rt} &= \frac{\lambda_{\rm rt} BW_{\rm rt}}{\mu_{\rm rt}} \\ OL_{\rm nrt} &= \frac{\lambda_{\rm nrt} BW_{\rm nrt}}{\mu_{\rm nrt}}. \end{aligned} \tag{4}$$

Replacing α with the ratio of the offered load of real-time to the offered load of nonreal-time traffic, we obtain

$$\frac{P_{\rm b,rt}}{P_{\rm b,nrt}} = \frac{OL_{\rm rt}}{OL_{\rm nrt}} \times \frac{W_{\rm rt}}{W_{\rm nrt}}. \tag{5}$$

The advantage of this approach will be discussed in Section V. Hereafter, we call the scheme satisfying (1) and (2) DMTBR_A and the scheme satisfying (1) and (5) DMTBR_G.

B. Connection Admission Policy

In the proposed scheme, three reservation thresholds, namely, G_1, G_2, and G_3, are maintained and dynamically adjusted. G_1 is the bandwidth reserved for real-time handoff connections only, and G_2 is the bandwidth reserved for handoff traffic including real-time handoff and nonreal-time handoff traffic. G_3 is the threshold used for admitting the two types of new traffic and is set such that the relative priority is maintained. Its role thus is not fixed, depending on the instantaneous status of the relative priority for the traffic classes. More specifically, in (2) or (5), if the CBP for real-time traffic is too large to meet the requirement, G_3 should be used to favorably accept more of the new real-time connections; if the CBP for real-time traffic is too

```
if (the incoming handoff connection is real-time)
    if (available BUs >= BW_rt)   accept;
    else   reject;
else // the incoming handoff connection is non-real-time
    if (available BUs >= G_1 + BW_nrt)   accept;
    else   reject;

if (switch == true)
    if (the incoming new connection is real-time)
        if (available BUs >= G_2 + BW_rt && rand( ) <= prob_rt)
                accept;
            else reject;
    else // the incoming new connection is non-real-time
        if (available BUs >= G_3 + BW_nrt && rand( ) <= prob_nrt)
                accept;
            else reject;
else // (switch == false)
    if (the incoming new connection is real-time)
        if (available BUs >= G_3 + BW_rt && rand( ) <= prob_rt)
                accept;
            else reject;
    else // the incoming new connection is non-real-time
        if (available BUs >= G_2 + BW_nrt && rand( ) <= prob_nrt)
                accept;
            else reject;
```

Fig. 1. Admission policy.

small, G_3 should be used to favorably accept more of the new nonreal-time connections. By definition, $G_1 < G_2 < G_3 \leq C$. Assume now we know G_1, G_2, and G_3 (more details on adaptation will be given in Section IV), the connection admission policy is given as follows. A handoff real-time connection is always admitted as long as there are $BW_{\rm rt}$ BUs available and rejected otherwise. Upon a handoff nonreal-time arrival, it will be admitted if there are $BW_{\rm nrt} + G_1$ BUs available; otherwise, it is rejected. For a new real-time connection, it is admitted if there are $BW_{\rm rt} + G_2$ or $BW_{\rm rt} + G_3$ BUs available and rejected otherwise. Similarly, for a new nonreal-time connection, it is admitted only if $BW_{\rm nrt} + G_3$ or $BW_{\rm nrt} + G_2$ BUs are available. The pseudo-code for the connection admission control, including the throttling of new connections in times of heavy congestion, is shown in Fig. 1. Note that the parameter *prob* denotes the probability of throttling and the function *rand()* generates a random number uniformly distributed in $[0, 1)$. *switch* can be considered as a Boolean sign, which indicates the role of G_3.

C. Cooperation Among Cells

In the cellular network, traffic in different cells have correlations. We observe that at one time, if one cell is swarmed with a lot of new connection requests due to traffic burstiness and many of them are admitted into the cell, it is likely that some time later, one or several of the neighboring cells will receive heavier handoff traffic from this cell. Hence, it is necessary and more efficient to deal with network congestion in a cooperative manner to prevent this from happening when executing the admission control. We adopt this idea in this scheme to cope with the situation where the network is undergoing heavy traffic load in times of burstiness. The details are as follows.

For a time interval, each cell measures CBPs and CDPs, i.e., $P_{\rm b,rt}$ and $P_{d,rt}$ (or $P_{\rm b,nrt}$ and $P_{\rm d,nrt}$). Each time the cell finds that the measured CDP for a certain traffic class, either real-time or nonreal-time traffic, is equal to or larger than a certain portion of QoS bounds, it will increase the corresponding bandwidth reservation threshold for the traffic class (please refer to Section IV for details). However, in the case of heavy network congestion, even if the total available BUs in the cell are all reserved, the predefined QoS requirement still may not be satisfied. To avoid this situation, we will actively take some preventive actions beforehand. We count the number of times we increase the reservation thresholds for handoff traffic. Once one reservation threshold is consecutively increased for a certain number of times, say three times, the cell is deemed experiencing heavy handoff traffic. In this case, to reduce the potential incoming handoff traffic and keep the CDP below the upper bound, the cell will inform all of its neighbors to further throttle the admissions of new connections of the same traffic class as the handoff traffic class in the current cell. Similar to [4] and [22], our proposed scheme admits the new connection request with a certain probability generated online, which is called the *probability of throttling new connections*. Details on how to generate the probability are given in Section IV.

IV. Threshold Adaptation

The key idea in this scheme is to accurately adjust the values of these three thresholds G_1, G_2, and G_3. First, we introduce the method to calculate G_1. From what is stated in the admission policy, we know G_1 is the bandwidth reservation threshold for real-time handoff connections only; here, we adopt the technique similar to the dynamic resource allocation scheme in [15] to estimate G_1. As we will see clearly later, the techniques we use to estimate G_1 (or G_2', the reserved bandwidth for nonreal-time handoff connections) only serves to provide a good initial value that should be further adapted. Therefore, in general, any method that can provide a good initial value can be used in our scheme. In this sense, our scheme is independent of any specific technique we use to obtain the initial value.

A. Adaptation of G_1

Let time t denote the time instant of the arrival of a new real-time connection request in a typical cell i. Let d be the expectation of the CHT of this new connection in cell i. In the time interval $(t, t+d]$, some of the existing real-time connections may leave cell i either due to completion or due to a handoff to the adjacent cells. We define an outgoing event to be such a departure, which will result in a release of $BW_{\rm rt}$ BUs. Meanwhile, there are some handoff real-time connections coming into cell i from its neighboring cells. We define an incoming event to be a handoff into cell i, which will consume $BW_{\rm rt}$ channels. We want to obtain the net bandwidth changes of cell i in $(t, t+d]$.

Let m denote the number of real-time connections in cell i that will leave cell i, and let n denote the number of handoff real-time connections that will enter cell i in $(t, t+d]$. Therefore, a total of $m+n$ events will occur in cell i during $(t, t+d]$. Let s be a sequence of these $m+n$ events and $S(m, n)$ be the set of all possible sequences that may take place in $(t, t+d]$, then we can obtain the cardinality of $S(m, n)$

$$|S(m,n)| = \frac{(m+n)!}{m!n!}. \tag{6}$$

If we define $X_k(s)$ as the net change in the number of BUs allocated to real-time connections from time t to the end of kth event in s, it is obvious that

$$X_k(s) = \begin{cases} X_{k-1}(s) - BW_{\mathrm{rt}}, & \text{if } k\text{th event is outgoing} \\ X_{k-1}(s) + BW_{\mathrm{rt}}, & \text{if } k\text{th event is incoming.} \end{cases} \tag{7}$$

Here, we assume $X_0(s) = 0$ and $1 \le k \le (m+n)$. Let $Y(s) = \max\{X_k(s) : 0 \le k \le (m+n)\}$, corresponding to the maximum net change in the number of BUs allocated to real-time connections in cell i in $(t, t+d]$. $Y(s)$ could also be thought of as the maximum number of BUs that needs to be reserved to deal with handoff real-time connections that arrive in $(t, t+d]$. Because all these incoming and outgoing events are independent, each possible sequence s occurs with an equal probability, i.e., $1/|S(m,n)|$. We thus set G_1 to the expected value of $Y(s)$

$$G_1 = \sum_{s \in S(m,n)} \frac{Y(s)}{|S(m,n)|}. \tag{8}$$

B. Adaptation of G_2

Again, let time t denote the time instant of the arrival of a new nonreal-time connection request in cell i. Let d' be the expectation of the CHT of a new nonreal-time connection in cell i. (According to our assumptions, the expectations of the CHT for a new connection and a handoff connection are equal.) Before calculating G_2, we calculate the expected value of the maximum number of channels G_2' which needs to be reserved for handoff nonreal-time connections arriving in $(t, t+d']$.

Similarly, let m' denote the number of nonreal-time connections in cell i leaving cell i due to completion or handoff, and let n' denote the number of handoff nonreal-time connections entering cell i in $(t, t+d']$. In a very similar manner, we can obtain G_2'

$$G_2 = \sum_{s' \in S'(m',n')} \frac{Y'(s)}{|S'(m',n')|} \tag{9}$$

where s', $S'(m',n'$, and $Y'(s)$ are the counterparts of s, $Y(s)$, and $S(m,n)$ defined earlier.

By definition, G_2 should be the sum of G_1 and G_2', thus we obtain the estimate of G_2

$$G_2 = G_1 + G_2'. \tag{10}$$

Note that the BS in cell i needs to communicate with its neighbors to acquire the information about how many handoff connections will come into cell i in $(t, t+d]$ or $(t, t+d']$, i.e., to acquire the information to calculate n and n'. According to the traffic model described in Section II, we can easily obtain the parameters involved, i.e., d, d', m, m', n, n', $Y(s)$, and $Y'(s)$. Also note that the calculation order of G_1 and G_2 implies that handoff real-time traffic is granted higher priority than handoff nonreal-time traffic.

C. Adaptation of G_3

Before calculating G_3, we calculate G_3', which could be thought of as the bandwidth that is reserved for new real-time traffic against new nonreal-time traffic or the bandwidth that is reserved for new nonreal-time traffic against new real-time traffic, depending on the instantaneous relative priority status for the traffic classes. For instance, in (2) or (5), if the relative priority between real-time traffic and nonreal-time traffic is violated because real-time traffic is unfairly rejected compared with nonreal-time traffic, G_3' should be the bandwidth reserved for new real-time connections so that more real-time connections can be accepted, and vice versa. The initial value for G_3' could be set as BW_{rt} or BW_{nrt}. Therefore, G_3 can be estimated as

$$\begin{aligned} G_3' &= \begin{cases} BW_{\mathrm{rt}}, & \text{if } G_3' \text{ is for new rt conn.} \\ BW_{\mathrm{nrt}}, & \text{if } G_3' \text{ is for new nrt conn.} \end{cases} \\ G_3 &= G_2 + G_3'. \end{aligned} \tag{11}$$

Notice that in (11), we only set the initial value of G_3'. When the scheme is running, we will adapt the value of G_3' according to the second criterion mentioned earlier.

D. Further Adaptation of Thresholds

It is seen that we use the expected maximum net bandwidth or nominal bandwidth to estimate the reservation thresholds. We may expect deviations from them in a dynamically changing network environment, where the accuracy of the estimation may degrade. Therefore, it is insufficient to depend only on the above three thresholds to fulfill the task of QoS provisioning. To meet the QoS criteria mentioned in Section III, further adaptation of these thresholds is needed. The details are shown in Fig. 2.

In Fig. 2, up_th_1, down_th_1 $(0 < \mathrm{down}\_th\_1 < \mathrm{up}\_th_1 < 1)$ are the threshold factors indicating when the measured CDP is above $\mathrm{up}\_th_1^* \mathrm{QoS}_{\mathrm{rt}}$ or below $\mathrm{down}\_th_1^* \mathrm{QoS}_{\mathrm{rt}}$, the threshold will increase or decrease. Once the threshold is consecutively increased for a certain number of times, denoted by *time_th*, the cell will inform all of its neighbors to do throttling as we described before. $\mathrm{Pow}(\mathrm{up}\_1, rt\_index)$ refers to the *rt_index*th power of $\mathrm{up}_1(>1)$, in which *rt_index* is an integer. The positive or negative value of rt_index means we actually increase or reduce the value of the initial value of G_1, which is obtained as described earlier. It is worth noting that when the measured CDP exceeds $\mathrm{up}\_th_1^* \mathrm{QoS}_{\mathrm{rt}}$, we immediately boost *rt_index* to zero if it was negative in previous step. In this way, this scheme is always able to be responsive enough to fulfill the QoS bound criterion. The portion of how to adapt G_2' is omitted since it is similar to that of adapting G_1.

To guarantee the second QoS criterion, G_2 and G_3 are used to make (2) or (5) hold. As we can see in the pseudo-code regarding how to adapt G_3, there are three parameters, namely, *switch*, *percentage*, and *adj_index*. *switch* is defined as before. *Percentage* refers to the deviation error the scheme may tolerate, i.e., the second criterion is still considered to be met. For instance, if *percentage* is set to be 0.1, this means that as long as the ratio of the right-hand side and the left-hand side of (2) or (5) is within the range $[0.9, 1.1]$, the equations hold and the QoS criterion is met. The role of *adj_index* is very similar to *rt_index*. However, in the adaptation here, we change $\mathrm{up}_3(>1)$, according the value of *adj_index* in a way that, the larger the absolute value of *adj_index*, the faster the adaptation speed.

```
/* Assuming the initial values of G1, G2', G3' are obtained. */
time1 = 0, time2 = 0;
rt_index = 0, nrt_index = 0;
switch = TRUE;
if (P_d_rt >= up_th1 * QoS_rt) {
    if (rt_index < 0)    rt_index = 0;
    else    rt_index++;
    G1 = G1*pow(up1, rt_index);
    time1++;
    if ((time1 % time_th) = 0)
        {asking neighboring cells to throttle; time1 = 0;}
    }
else if (P_d_rt < down_th1 * QoS_rt) {
    rt_index--;
    G1 = G1 * pow(up1, rt_index);
    time2++;
    if ((time2 % time_th) = 0)
        {asking neighbors to de-throttle; time2 = 0; }
    }

G2 = G1 + G2'; /* Adaptation of G2' is omitted. */

if (switch = = TRUE){
    if (P_h_rt / P_h_nrt >= W_rt / W_nrt * [OL_rt / OL_nrt] * (1+percentage))
        adj_index++;
    else if (P_h_rt / P_h_nrt <= W_rt / W_nrt * [OL_rt / OL_nrt] * (1-percentage))
        adj_index--;
    }
else{
    if (P_h_rt / P_h_nrt >= W_rt / W_nrt * [OL_rt / OL_nrt] * (1+percentage))
        adj_index--;
    else if (P_h_rt / P_h_nrt <= W_rt / W_nrt * [OL_rt / OL_nrt] * (1-percentage))
        adj_index++;
    }
G3' = G3' * pow(up3, adj_index);
if (adj_index < adj_index_th) {
            adj_index = 0;
            switch = ! switch;
            }
G3 = G2 + G3';
```

Fig. 2. Reservation thresholds adaptation.

This ensures the adaptation of G_3 can promptly respond to the change of the incoming traffic and/or QoS status. Finally, when *adj_index* is less than a threshold, *adj_index_th*, which means G_3' is nearly zero, the scheme will reverse the parameter *switch*, letting G_3 be reserved for the other traffic class instead of the current traffic class it is in.

E. Probability of Throttling

Each cell keeps a $J \times K$ nonnegative integer array A for each traffic class for its neighbors. J is the number of traffic classes and K is the number of neighboring cells. Since we consider real-time and nonreal-time traffic in this paper, J is equal to 2. If the cell's ith $(i = 0, 1, \ldots, K-1)$ neighbor sends a message to the cell to throttle or dethrottle a real-time traffic class, then $A[0][i]$ is incremented or decremented by one. It is similar for nonreal-time traffic. When there is an incoming new connection request, the cell will use the following equation to generate the probability of throttling new connections:

$$\begin{aligned} \mathrm{prob}_{\mathrm{rt}} &= b^{\max(A[0][i])} \\ \mathrm{prob}_{\mathrm{nrt}} &= b^{\max(A[1][i])} \end{aligned} \tag{12}$$

where b is a real number less than and close to 1, say 0.9. We can see $\mathrm{prob}_{\mathrm{rt}}$ (or $\mathrm{prob}_{\mathrm{nrt}}$) is equal to one if $\max(A[0][i])$ (or $\max(A[1][i])$) is zero. This means we do not need to throttle the new real-time (or nonreal-time) connections.

F. Estimation of Arrival Rate λ_{rt} and λ_{nrt}

In order to make scheme DMTBR_G work properly according to the current network traffic situation, we need to provide real-time information about the incoming traffic. Thus, we need to estimate the current arrival rate of real-time connection (nonreal-time connection) $\lambda_{\mathrm{rt}}(\lambda_{\mathrm{nrt}})$. Assume that we measure the arrival rate at a fixed period p, and we denote the measured arrival rate of real-time connections and nonreal-time connections at the nth $(n = 1, 2, \ldots,)$ measurement period as $M_{\mathrm{rt}}(n)$ and $M_{\mathrm{nrt}}(n)$, then we can estimate the arrival rate using a low-pass filter

$$\begin{aligned} \lambda_{\mathrm{rt}}(n+1) &= \alpha\lambda_{\mathrm{rt}}(n) + (1-\alpha)M_{\mathrm{rt}}(n) \\ \lambda_{\mathrm{nrt}}(n+1) &= \alpha\lambda_{\mathrm{nrt}}(n) + (1-\alpha)M_{\mathrm{nrt}}(n) \end{aligned} \tag{13}$$

where $M_{\mathrm{rt}}(n)$ and $M_{\mathrm{nrt}}(n)$ can be obtained by

$$\begin{aligned} M_{\mathrm{rt}}(n) &= \frac{\#\text{ of new rt arrivals in } n\text{th period}}{p} \\ M_{\mathrm{nrt}}(n) &= \frac{\#\text{ of new nrt arrivals in } n\text{th period}}{p} \end{aligned} \tag{14}$$

and α is a weighting factor, usually $0.5 < \alpha < 1$. We can see that more weight is given to the arrival rates recently observed.

G. Updating Frequency

In responding to the network traffic condition as promptly as possible, it would be ideal that each time a new or handoff connection request arrives in cell i, all three reservation thresholds are updated so that an admission decision could be made for the request. However, the BS may update the thresholds each time it has received N connection requests, considering that each update may incur some communication and computation overheads. N could be chosen to provide the tradeoff between system performance and overheads. Its impact on the performance will be investigated in Section V.

V. Performance Evaluation

In this section, we present the performance results for our proposed scheme based on extensive simulations. Our simulation is carried out using OPNET Modeler. The simulation model is a wrap-around model as shown in [23], which comprises 37 cells. Each cell, represented by a hexagon, has six neighbors so that handoff departure from the edge cells will not be ignored.

We consider the following simulation parameters. The total number of BUs in each cell is 50. The number of BUs each real-time or nonreal-time connection will need is $BW_{\mathrm{rt}} = 1$ or $BW_{\mathrm{nrt}} = 4$. The real-time connection may be voice calls and the nonreal-time connection may represent file transfer or web browsing. For real-time traffic, the mean duration $1/\mu_{\mathrm{rt}} = 300$ s and the mean cell residence time $1/\gamma_{\mathrm{rt}} = 150$ s. For nonreal-time traffic, $1/\mu_{\mathrm{nrt}} = 1500$ s and $1/\gamma_{\mathrm{nrt}} = 750$ s. On average, each connection will handoff once during its lifetime.

Whenever there is a handoff request, it will randomly choose a destination from the six neighboring cells. We assume that 25% of traffic is real-time traffic, and 75% of the traffic is nonreal-time traffic. This is consistent with the fact that data services will dominate network traffic in the near future. New connections, including real-time and nonreal-time connections, arrive according to a Poisson process. According to the assumption, 86.96% of the new connection arrivals are real-time connections, and the rest are nonreal-time connections. For both DMTBR_A and DMTBR_G, $\mathrm{QoS}_{\mathrm{rt}} = 0.01$ and $\mathrm{QoS}_{\mathrm{nrt}} = 0.05$. The ratio $W_{\mathrm{rt}}/W_{\mathrm{nrt}}$ is equal to one, with *percentage* set to 0.1. The threshold factors down_th_1 and down_th_2 are set to 0.5; up_th_1 and up_th_2 are set to 0.8. $\mathrm{time}\_th = 3$ and $\mathrm{adj\_index}\_th = -30$. The base up_1 and up_2 equal 1.1. The default value of up_3 is 1.1. However, it changes to 1.15 or 1.2 when $|\mathrm{adj\_index}\_th|$ is larger than 10 or 20. The update frequency is set to 20 unless otherwise specified. The simulation time is 3 h.

Fig. 3(a) and (b) shows in a homogeneous environment the new connection blocking probabilities and handoff dropping probabilities for both traffic classes, as a function of average new connection arrival rate for both DMTBR_A and DMTBR_G. Through calculation, we know that arrival rate 0.1 connection/s corresponds to about 110 Erlangs, which is 220% of the full load. In Fig. 3(a), as expected, we can see that for DMTBR_A, the blocking probabilities for the two classes are almost equal to each other. For DMTBR_G, since the ratio of offered load of each traffic class is taken into consideration, which is equal to 1:3, the ratio of blocking probabilities for the two traffic classes is also about 1:3. Through direct calculation, we find out that on average, the CBP for real-time traffic is reduced 58.23% in DMTBR_G compared to that in DMTBR_A, whereas the CBP for nonreal-time traffic is only increased 9.89% in DMTBR_G compared to that in DMTBR_A. In Fig. 3(b), we see both schemes successfully keep the handoff dropping probabilities of both traffic classes under the predefined QoS bounds as expected, even when the network is experiencing heavy traffic. Also, there is not big difference in these two schemes in terms of handoff-dropping probabilities.

In addition to CBPs and CDPs, we check the performance of these two schemes in terms of traffic throughput. The traffic throughput for real-time (or nonreal-time) traffic is defined as follows:

$$\mathrm{TP}_{\mathrm{rt(nrt)}} = \frac{BW_{\mathrm{rt(nrt)}} \sum \text{CHT of rt (nrt) conn.}}{C^{*}\mathrm{CELL\_NUM}^{*}ST} \tag{15}$$

where C, as mentioned before, is the total number of BUs available in each cell, $\mathrm{CELL_NUM}$ is the total number of cells in the entire network and *ST* is the total simulation time. Rather than using the average time spent by each connection in a cell as [12] did, we count the actual time spent by each connection. Obviously, this will give us a more accurate result. The entire network throughput is the sum of the throughput for each traffic class in the network.

The traffic throughput of each traffic class for both schemes is shown in Fig. 3(c). In DMTBR_A, the traffic throughput is approximately increasing with the connection arrival rate. For DMTBR_G, when traffic arrival rate is low, it performs just like

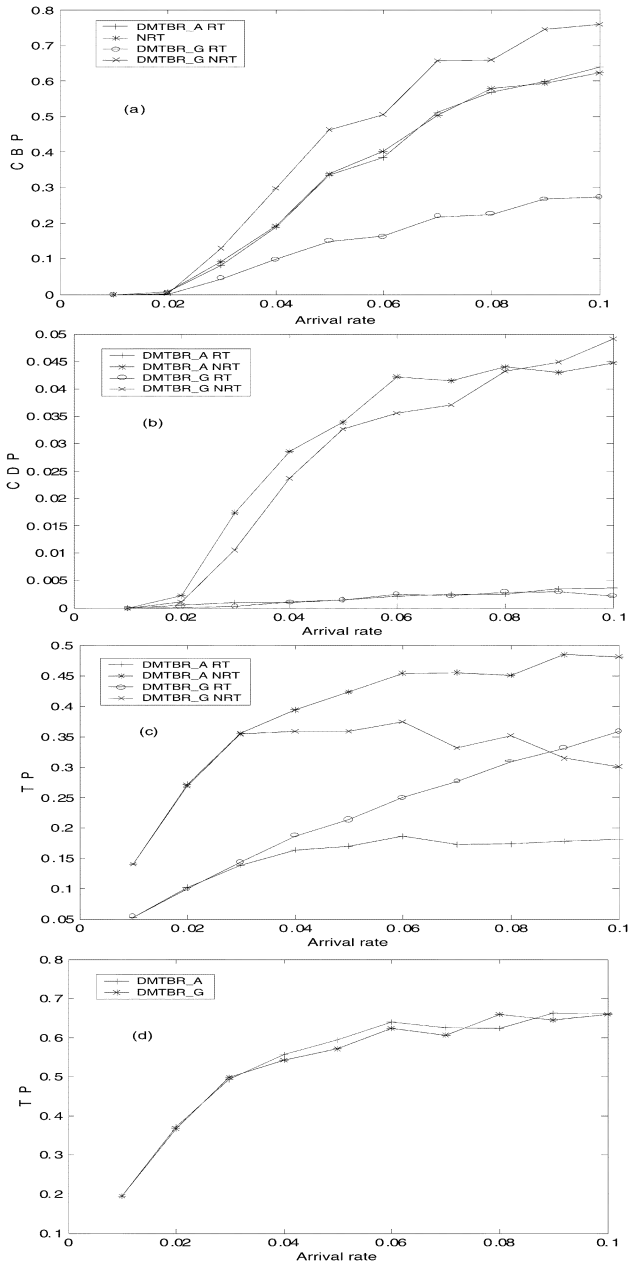


Fig. 3. Performance in homogeneous environment: (a) CBP, (b) CDP, (c) traffic throughput, and (d) system throughput.

DMTBR_A. However, when the traffic arrival rate is getting higher, or when the network is overloaded, it behaves a little differently. The real-time traffic throughput is still increasing with the arrival rate, whereas nonreal-time traffic throughput remains in a certain level and lowers a bit at the end. This is because in DMTBR_G the lower CDP for real-time traffic is achieved at the cost of the higher CBP for nonreal-time traffic; accordingly, we observe the decrease in the throughput of nonreal-time traffic and the increase in the throughput of real-time traffic. Nevertheless, it can be observed that both schemes successfully achieve a stable throughput even under heavy traffic situations. From a system's point of view, these two schemes differ very little in terms of network throughput, which can be observed in Fig. 3(d). The network throughput keeps increasing as the offered load increases, showing very little difference.

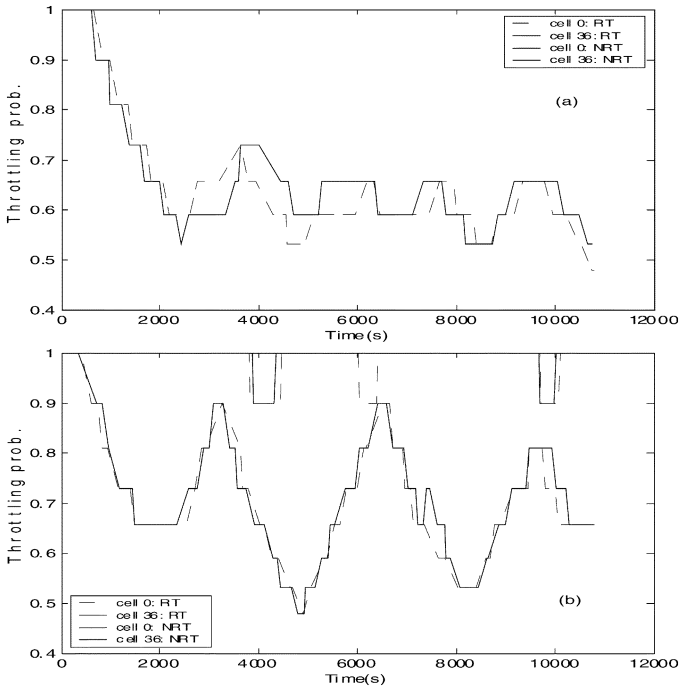


Fig. 4. Throttling probability: (a) DMTBR_A and (b) DMTBR_G.

Combining the observation in Fig. 3, we discover the advantage of DMTBR_G over DMTBR_A which is gained by generalizing the concept of relative priority. Without compromising network throughput, DMTBR_G significantly improves the user satisfaction for new real-time traffic (in terms of the CBP) while only slightly affecting the user satisfaction for nonreal-time new traffic. Meanwhile, the user satisfaction for handoff traffic (in terms of the CDP) is well maintained for both schemes.

Next, we consider the detailed throttling operations in each cell. Fig. 4(a) and (b) shows, in each scheme, the throttling probabilities for both types of traffic, starting from the beginning of a simulation run (i.e., $t = 0$) for arrival rate $= 0.1$ in cell 0 and cell 36. In the simulation model, cell 0 is in the center and cell 36 is located at the edge. In both figures, as time passes, the throttling probabilities for real-time traffic are almost one, which means the neighboring cells of cell 0 or 36 rarely throttle the acceptance of new real-time connections. This is consistent with Fig. 3(b), where the CDP for real-time traffic is well below the predefined QoS bounds, indicating there is no need to reduce the new connection admission for fulfilling the first QoS criterion. For nonreal-time traffic, as time passes, the throttling probabilities first drop then fluctuate around a certain value after the network is in a stable state. Thus, we know that the cells keep the first QoS criterion for nonreal-time traffic with the help of cooperative neighbors, which reduce the admission probability for new connections due to nonreal-time traffic when necessary.

Fig. 5 shows how varying QoS requirements affect the performance of the scheme. Since we assume that a large portion of traffic is due to nonreal-time traffic, we present the results obtained for DMTBR_A with the variation of $\mathrm{QoS}_{\mathrm{nrt}}$ for the purpose of illustration. As observed, the CBP decreases as the QoS bound $\mathrm{QoS}_{\mathrm{nrt}}$ is relaxed. The reason is as follows. A loose QoS bound means less bandwidth needed to be reserved for the

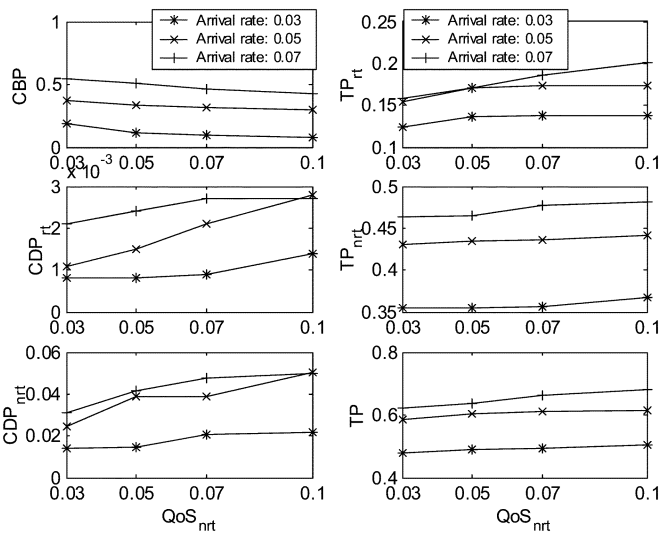


Fig. 5. Performance versus QoS.

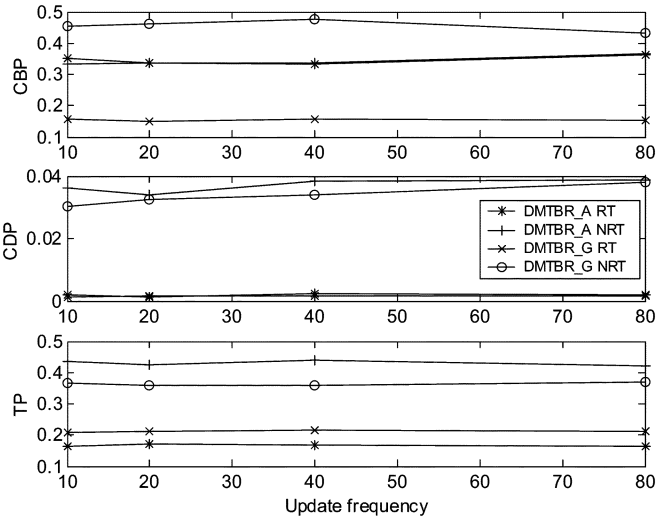


Fig. 6. Performance versus update frequency.

handoff traffic. As a result, the more available bandwidth can be used to accept more new connections. Since a fixed ratio of the CBPs for both traffic classes is maintained, they together will increase at the same pace. CDPs for both traffic classes are getting larger. For the CDP of nonreal-time traffic, it is easy to understand since we loosen the QoS bound. For the CDP of real-time traffic, since more new connections are accepted, it is likely that its corresponding CDP will increase, given that the entire network bandwidth is fixed. The throughputs for both traffic classes increase because the effect of the relaxed $\mathrm{QoS}_{\mathrm{nrt}}$ is magnified and then reflected on the reduction of CBPs, which result in a larger throughput. Accordingly, the network throughput is also increasing.

We also investigate how dependent the system performance is on the value of update frequency N. Larger N means a slower update rate. Fig. 6 shows that all the performance metrics, including the CBP, the CDP, and throughput, are not very sensitive to the value of N, even when N is equal to 80. Therefore, the system can work well without incurring too frequent updates, hence communication and calculation overheads among BSs.

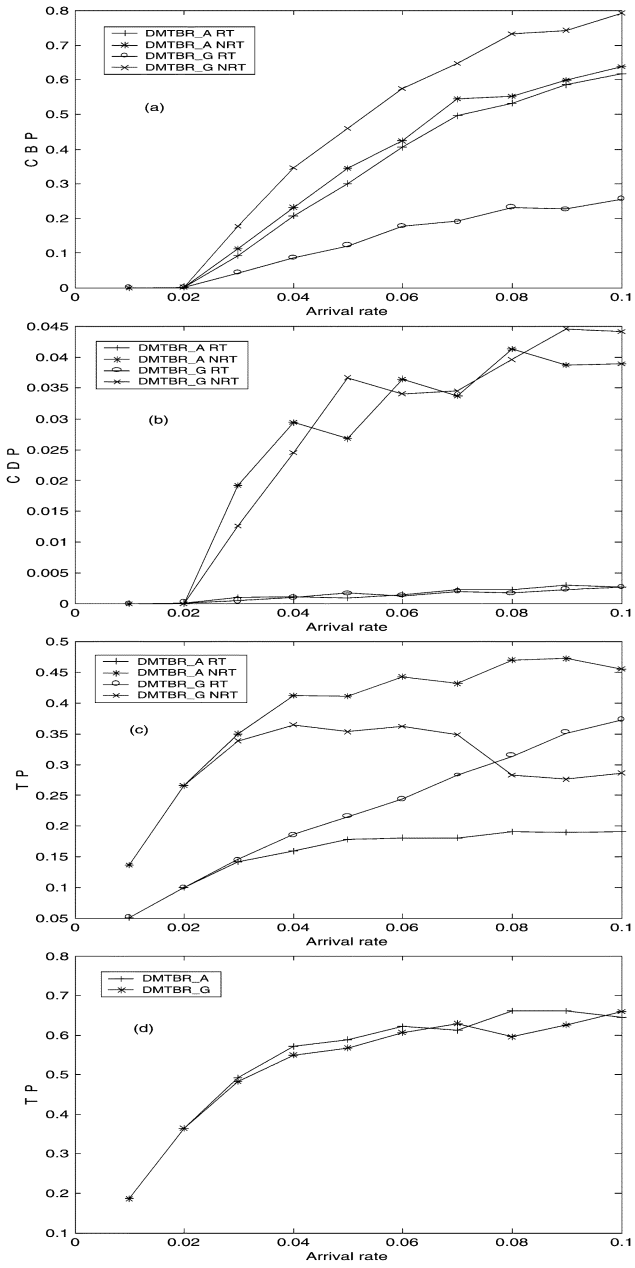


Fig. 7. Performance in heterogeneous environment: (a) CBP, (b) CDP, (c) traffic throughput, and (d) system throughput.

As a consequence, this scheme could be implemented and used in current systems.

Finally, we measure the performance of these two schemes in a heterogeneous environment. Although the network traffic composition is fairly stable in a long run, it may fluctuate sometimes. In the simulation, we model a system where the portion of the nonreal-time traffic in the entire traffic is changing from 70%, to 75%, then to 80%, while all the other traffic parameters are the same. For each specific value of the traffic portion, the network stays for one-third of the total simulation time. The results are shown in Fig. 7. The results show that the schemes still behave the same way, which means our scheme could work well in a heterogeneous case, too. This is expected since the scheme is, in essence, an adaptive one.

VI. Conclusion

In this paper, a DMTBR scheme is proposed to guarantee QoS provisioning in wireless multimedia networks. According to network traffic situations and QoS status, three bandwidth thresholds are dynamically adapted. In addition, when the network is under heavy traffic load, cooperation among neighboring cells is exploited. As a result, this scheme is able to provide a QoS guarantee while efficiently utilizing network resources under various traffic loads in both homogeneous and heterogeneous environments. Also, we generalize the concept of relative priority and show the acquired performance gain. With these desirable features, our proposed scheme is likely to be useful in future wireless systems.

References

[1] 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. Veh. Technol.*, vol. VT-35, no. 8, pp. 77–92, Aug. 1986.

[2] R. Guerin, "Queueing-blocking system with two arrival streams and guard channels," *IEEE Trans. Commun.*, vol. 36, no. 2, pp. 153–163, Feb. 1988.

[3] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," *IEEE/ACM Trans. Networking*, vol. 2, no. 4, pp. 166–175, Apr. 1994.

[4] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *ACM Wireless Networks*, vol. 3, pp. 29–41, 1997.

[5] Haung, Y.-B. Lin, and J. M. Ho, "Performance analysis for voice/data integration on a finite-buffer mobile system," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, Feb. 2000.

[6] L. Yin, B. Li, Z. Zhang, and Y.-B. Lin, "Performance analysis of a dual-threshold reservation (DTR) scheme for voice/data integrated mobile wireless networks," in *Proc. IEEE WCNC'02*, Sep. 2002.

[7] B. Li, L. Li, B. Li, and X. Cao, "On handoff performance for an integrated voice/data cellular system," *ACM Wireless Networks*, vol. 9, pp. 393–402, 2003.

[8] T.-C. Chau, K. Y. M. Wong, and B. Li, "Optimizing call admission control with QoS guarantee in a voice/data integrated cellular network using simulated annealing," in *Proc. IEEE Globecom'02*, Taipei, Taiwan, Nov. 17–21, 2002.

[9] C. Chao and W. Chen, "Connection admission control for mobile multiple-class personal communications networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 10, pp. 1618–1626, Oct. 1997.

[10] B. Li, C. Lin, and S. T. Chanson, "Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks," *ACM Wireless Networks*, vol. 4, pp. 279–290, 1998.

[11] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 711–717, May 1996.

[12] B. M. Epstein and M. Schwartz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 523–534, Mar. 2000.

[13] S. Wu, K. Y. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 4, pp. 257–271, Apr. 2002.

[14] B. Li, L. Yin, K. Y. Wong, and S. Wu, "An efficient and adaptive bandwidth allocation scheme for mobile wireless networks using an on-line local estimation technique," *ACM Wireless Networks*, vol. 7, pp. 107–116, 2001.

[15] P. Ramanathan, K. M. Sivalingam, P. Agrawal, and S. Kishore, "Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, no. 7, pp. 1270–1283, Jul. 1999.

[16] C. Oliver, J. B. Kim, and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 858–874, Aug. 1998.

[17] J. Hou and Y. Fang, "Mobility-based call admission control schemes for wireless mobile networks," *Wireless Commun. Mobile Computing*, vol. 1, pp. 269–282, Jul.–Sep. 2001.

[18] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using shadow cluster concept," *IEEE/ACM Trans. Networking*, vol. 5, no. 2, pp. 1–12, Feb. 1997.

[19] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high-speed networks," *IEEE/ACM Trans. Networking*, vol. 2, no. 4, pp. 166–175, Apr. 1994.

[20] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling for PCS networks," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1062–1072, Jul. 1999.

[21] Y. Fang, I. Chlamtac, and Y. B. Lin, "Channel occupancy times and handoff rate for mobile computing and PCS networks," *IEEE Trans. Comput.*, vol. 47, pp. 679–692, Jun. 1998.

[22] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 3, pp. 371–382, Mar. 2002.

[23] H. Zeng, Y. Fang, and I. Chlamtac, "Call blocking performance study for PCS networks under more realistic mobility assumptions," *Telecommun. Syst.*, vol. 19, pp. 125–146, Feb. 2002.

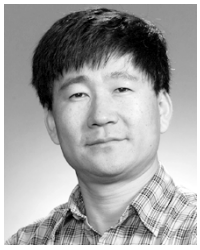
Xiang Chen (S'03) received the B.E. and M.E. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1997 and 2000, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Florida, Gainsville.

He worked as a member of technical staff for Bell Laboratories, Beijing, China. His research interests include resource allocation, call admission control, and quality of service in wireless networks, including cellular networks, wireless LAN, and mobile *ad hoc* networks.

Mr. Chen is a member of Tau Beta Pi.

Bin Li received the B.Eng. degree in automatic control from the Huazhong University of Science and Technology, Wuhan, China, in 1991, and received the M.Phil. and Ph.D. degrees in the electrical and electronic engineering from the Hong Kong University of Science and Technology, Hong Kong, China, in 1996 and 2003, respectively.

Between 1991 and 1994, he worked in the China Telecom Guangdong Branch, where he helped in designing, building, and maintaining the fixed line and mobile network. Since July 1997, he has been with the China Motion Telecom Group, Kowloon, Hong Kong, where he is now the Executive Director and Chief Operating Officer. His current research interests include the multimedia communications in the Internet, traffic engineering in wireless cellular networks, and data management in mobile systems. He has published 20 papers in IEEE conference proceedings and journals.



Yuguang Fang (S'96–M'97–SM'99) received the Ph.D. degree in systems and control engineering from Case Western Reserve University, Cleveland, OH, in 1994, and the Ph.D. degree in electrical engineering from Boston University, Boston, MA, in May 1997.

From September 1989 to December 1993, he was a Teaching/Research Assistant in the Department of Systems, Control and Industrial Engineering at Case Western Reserve University, and he held a Research Associate position from January 1994 to May 1994. He held a post-doctoral position in the Department of Electrical and Computer Engineering, Boston University, from June 1994 to August 1995. From September 1995 to May 1997, he was a Research Assistant in the Department of Electrical and Computer Engineering, Boston University. From June 1997 to July 1998, he was a Visiting Assistant Professor in the Department of Electrical Engineering at the University of Texas, Dallas. From July 1998 to May 2000, he was an Assistant Professor in the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark. In May 2000, he joined the Department of Electrical and Computer Engineering at the University of Florida, Gainesville, where he has been an Associate Professor since 2003. His research interests include wireless networks, mobile computing, mobile communications, automatic control, and neural networks. He has published over 100 papers in refereed professional journals and conferences.

Dr. Fang received the National Science Foundation Faculty Early Career Award in 2001 and the Office of Naval Research Young Investigator Award in 2002. He is a member of the ACM. He is an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING, a Technical Editor for *IEEE Wireless Communications Magazine*, an Editor for *ACM Wireless Networks*, and an Area Editor for *ACM Mobile Computing and Communications Review*. He was an Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS from May 1999 to December 2001, an Editor for *Wiley International Journal on Wireless Communications and Mobile Computing* from April 2000 to January 2004, and the Feature Editor for Scanning the Literature in IEEE PERSONAL COMMUNICATIONS (now the IEEE WIRELESS COMMUNICATIONS) from April 2000 to April 2003. He has also actively involved with many professional conferences. He was the Program Co-Chair for the Global Internet and Next Generation Networks Symposium in IEEE Globecom'2004 and was the Program Vice Chair for 2000 IEEE Wireless Communications and Networking Conference (WCNC'2000), where he received the IEEE Appreciation Award for the service to this conference. He has been serving on many Technical Program Committees such as IEEE INFOCOM (1998, 2000, 2003, 2004, 2005), IEEE ICC (2004), IEEE Globecom (2002–2004), IEEE WCNC (1999, 2000, 2002, and 2004), and ACM MobiCom (2001). He served as the Committee Co-Chair for the Student Travel Award for 2002 ACM MobiCom. He was the Vice-Chair for the IEEE Gainesville Section in 2002 and 2003 and is the Chair in 2004.