Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service

Chun-Ting Chou and Kang G. Shin, *Fellow*, *IEEE*

Abstract—A wireless/mobile network supporting multilevel Quality of Service (QoS) is considered. In such a network, users or applications can tolerate a certain degree of QoS degradation. Bandwidth allocation to users can, therefore, be adjusted dynamically according to the underlying network condition so as to increase bandwidth utilization and service provider's revenue. However, arbitrary QoS degradation may be unsatisfactory or unacceptable to the users, hence resulting in their subsequent defection. Instead of only focusing on bandwidth utilization or blocking/dropping probability, two new user-perceived QoS metrics, *degradation ratio* and *upgrade/degrade frequency*, are proposed. A Markov model is then provided to derive these QoS metrics. Using this model, we evaluate the effects of adaptive bandwidth allocation on user-perceived QoS and show the existence of trade offs between system performance and user-perceived QoS. We also show how to exploit adaptive bandwidth allocation to increase system utilization (for the system administrator) with controlled QoS degradation (for the users). By considering various mobility patterns, the simulation results are shown to match our analytical results, demonstrating the applicability of our analytical model to more general cases.

Index Terms—Wireless/mobile networks, quality of service (QoS), adaptive resource allocation.

1 Introduction

Many real-time applications can use different encoding schemes according to their desired quality and generate traffic with different bandwidth requirements. For example, generic video telephony may require more than 40 Kbps, but low-motion video telephony requiring about 25 Kbps may be acceptable [1]. From the standpoint of a system administrator, this property provides an alternative for resource planning, especially for bandwidth allocation in wireless networks. In wireless networks where the bandwidth is a scarce resource, the system may need to block incoming users if all of the bandwidth has been used up to provide the highest QoS to existing users. However, if these users can be degraded to a lower QoS level, it is possible to reduce the blocking probability without degrading the QoS of existing users to an "unacceptable" level.

Various approaches and algorithms adopting this idea have been proposed. A graceful degradation mechanism is proposed in [2] to increase bandwidth utilization by adaptively adjusting bandwidth allocation according to user-specified loss profiles. Thus, a system could free some bandwidth for new users by lowering the QoS levels of existing users. Sen et al. [1] proposed an optimal degradation strategy by maximizing a revenue function and Sherif et al. [3] proposed an adaptive resource allocation algorithm to maximize bandwidth utilization and attempted to achieve fairness with a generic algorithm.

In these papers, system performance, in terms of bandwidth utilization or service provider's revenue, can be improved significantly by allowing QoS degradation. However, the impact of quality degradation on individual users, which is crucial to QoS provisioning, was overlooked. For example, even though the users can tolerate some quality degradation, it is still desirable to provide them higher QoS when more resources become available. Thus, some performance metrics which reflect the average quality level that a user receives should be considered. Kwon et al. [4] derived a *degradation period ratio* to represent the time a user receives degraded quality. However, their formula hinges on the assumption that the *mean degradation time* and *degradation states* are independent variables. We show that these two variables are dependent and derive a new degradation ratio. In addition to this degradation ratio, we argue that another new performance metric, the frequency of switching between different quality levels, should also be taken into account because users may feel more disturbed by frequent switches between different quality levels than by poor and steady quality. It is shown numerically that *degradation ratio* does not suffice to reflect the QoS guarantees given to individual users. Frequent switching of QoS level may be even worse than a large degradation ratio [5]. So, we also derive a formula for the frequency of changing the QoS level and show the trade offs between this metric and other performance metrics, such as system utilization and fairness among users.

The problem of providing adaptive QoS in a wireless/mobile network would be similar to that in its wired counterpart if we do not consider user mobility. In a wireless/mobile network, a user may move across different cells or administration domains. Thus, we have to consider the user-perceived QoS not only during his stay in a single cell, but in all cells he may traverse throughout the connection lifetime. Moreover, the potential dropping due

• *The authors are with the Real-Time Computing Laboratory, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109-2122. E-mail:{choujt, kgshin}@umich.edu.*

Manuscript received 23 Jan. 2003; revised 7 July 2003; accepted 12 Aug. 2003. For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number 10-012003.

to such cell crossings (i.e., handoffs) has to be taken into account. The *forced-termination* (or *dropping*) *probability* is a widely used metric to represent the compromise of QoS due to user mobility. This probability should be made as small as possible because admitting a user and then terminating his session before its completion would make the user even unhappier. In order to reduce this probability, many admission control algorithms give handoff users priority over new users. Lin et al. [6] proposed an analytical model for a so-called *Guard Channel* system where a portion of bandwidth is reserved for handoff users, while, in [7], [8], handoff users will more likely be accepted once the system load exceeds some predefined threshold. Some other admission algorithms treat new and handoff users equally, but estimate the traffic loads of adjacent cells [9] or the handoff rates from the adjacent cells [10] such that the potential overload (and, consequently, higher forced-termination probabilities) can be prevented in advance. In contrast to these proposals which use fixed bandwidth allocation, we will show that adaptive bandwidth allocation can be also used to further reduce the forced-termination probability.

In this paper, we exploit the adaptive bandwidth allocation for QoS provisioning in wireless/mobile networks. An analytical model for a wireless/mobile network with multilevel degradable QoS is provided. This model includes two very important QoS metrics—degradation ratio and upgrade/degrade frequency—both of which are necessary for QoS provisioning. Moreover, our analytical model includes user mobility to enable the study of its impact on user-perceived QoS. Our work not only provides an analytical framework for predictive or adaptive bandwidth allocation algorithms [11], [12], but also helps decide the operation region based on some desired criteria. It should be noted that our scheme can be applied to various wireless architectures. For a code division multiple access (CDMA) system, the multicode CDMA [13] can be used for service degrade/upgrade; for a time division multiple access (TDMA) system (e.g., Bluetooth), service degrade/upgrade can be achieved by an adequate assignment of time slots (i.e., polling policy) [14].

The rest of this paper is organized as follows: In Section 2, the system model and the assumptions used in this paper are introduced. Section 3 provides an analytical model for the system under consideration and the QoS metrics mentioned above are derived. The numerical results based on the analytical model are presented in Section 4, while Section 5 discusses the simulation results. Finally, conclusions are drawn and directions of our future work are discussed in Section 6.

2 System Model and Assumptions

We consider a wireless network in which the base station takes charge of both admission control and bandwidth allocation for mobile users in its cell. While residing in the cell of a base station, a mobile user communicates with others via that base station. A "wireless network" can be a conventional cellular phone network or an office building with interconnected IEEE 802.11 wireless LANs. In such a network, a mobile user could either be successfully handed off to a new base station or simply dropped when it is about to leave the present cell. As mentioned in the introduction, we give handoff users priority over new users since dropping a handoff user is usually less desirable and less tolerable than blocking new users. This is achieved by restricting a new user into the cell once the total number of users or the total occupied bandwidth exceeds a prespecified threshold, N_{thresh}. Handed-off and newly initiated users are treated equally once they are admitted into a cell.

2.1 QoS Metrics

We are primarily interested in quality-degradable connections as long as the resultant quality is within the user-specified QoS profile. The only QoS requirement we discuss here is the bandwidth. For example, it can be a video-streaming application with multiple transmission rates depending on the encoding schemes and resolution. We assume that there are K different quality levels. The bandwidth requirement of the ith quality level is denoted as W_i and $W_{max} = W_1 > W_i > W_K = W_{min}$. With such a degree of freedom, a base station may try to degrade the quality levels of some existing users in order to admit more users so as to improve the overall system performance. For example, we may be able to achieve high bandwidth utilization and maintain a small blocking and/or forced-termination probability.

In a system with degradable QoS, a user may receive different levels of QoS during the entire duration of his connection, depending on the loads of cells he traverses. Even if a user receives the highest level of QoS when he is admitted to a cell, the QoS may still be degraded when some other base stations on his "path" decide to degrade his QoS in order to accept more users. From the users' perspectives, this may raise two important questions: 1) How long does his connection stay at each individual QoS level? 2) How often does the received QoS switch between these levels? Even though these two questions are interrelated, the first question does not necessarily imply the second, or vice versa. Therefore, two performance metrics associated with these questions, degradation ratio and upgrade/degrade frequency, are proposed as follows:

- *Degradation ratio* (DR). The fraction of time a user receives degraded QoS. Since we consider a multilevel QoS system, DR is defined as

$$\mathrm{DR} = \frac{\sum_i \frac{(W_{max} - W_i)}{W_{max}} \cdot T_i}{\sum_i T_i}, \tag{1}$$

  if a user receives level-i QoS for T_i seconds.
- *Upgrade/degrade frequency* (UDF). The frequency of changing the QoS level an admitted user receives.

These two metrics, along with the probability of blocking new users and the probability of dropping handoff users, will be the key performance metrics that we will consider throughout this paper.

2.2 Traffic Models

We assume that the arrivals of new users into a cell is a Poisson process with a rate λ_0. The Poisson process works well in modeling call arrivals in a public telephone network.

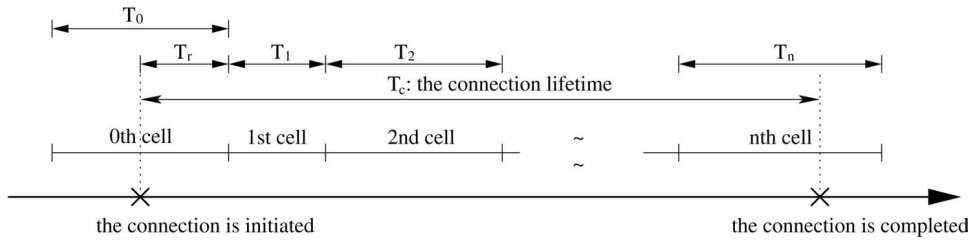

Fig. 1. A connection with n handoffs.

Even though recent studies found that the "packet" arrival process in the Internet switches/routers is not Poisson, the network traces also show that the user-generated connection requests, such as Telnet or FTP connection requests, can still be modeled as a Poisson process [16]. Since we mainly focus on the admission control and bandwidth allocation, which are the connection-level (as opposed to packet-level) resource management, the Poisson process is still a good approximation for our purpose. For mathematical tractability, we also assume the lifetime of each connection to be exponentially distributed with mean $\frac{1}{\mu_0}$. Note that the exponential distribution has been used to accurately model the intervals of talk spurt and silence in a phone call [17]. Thus, this assumption captures the reality of some real-time applications.

To evaluate the effects of user mobility on system performance, we use the cell-sojourn time—the time a user stays in a cell—to account for his mobility. As far as connection-level resource management is concerned, the cell-sojourn time and connection lifetime together determine the duration for which a user will occupy the bandwidth in a cell. Thus, it is more important to model this sojourn time than modeling the user's actual movement. In view of the fact that the cell-sojourn time depends on many time-varying factors, such as the user's speed (which, in turn, depends on his geographical location), the moving direction (i.e., whether or not he is heading toward a congested spot), and the cell size (or, more precisely, the portion of a cell the user traverses), we assume that the distribution of sojourn time in each cell is independent and identical. Furthermore, this cell-sojourn time is assumed to be exponentially distributed with mean $\frac{1}{\eta}$ for the purpose of mathematical derivation as in [7], [8], [15]. However, we will show via simulation that the formulas for QoS metrics derived under this model still match the simulation results well even when different distributions of cell-sojourn time are used.

The relation between connection lifetime, cell-sojourn time, and the number of cells that a mobile user will traverse is shown in Fig. 1. Since the connection lifetime is mainly decided by the communication contents, such as the length of conversation during a phone call or the size of a transferred file,[1] while the cell-sojourn time is decided by the aforementioned factors, we assume that these two random variables are independent. Thus, the probability that a mobile user will experience handoffs H times can be calculated as

$$\begin{aligned} P(H=n) =& P(T_r + T_1 + T_2 + \cdots + T_{n-1}) \\ & < T_c < P(T_r + T_1 + T2 + \cdots + T_n), \end{aligned} \tag{2}$$

where T_r is the remaining cell-sojourn time in the cell where a user's connection is initiated, T_i is the cell-sojourn time in the ith cell, and T_c is the connection lifetime.

Furthermore, if we consider the potential forced termination during a handoff (i.e., a handoff drop), the handoff rate can be derived as in [6]:

$$\lambda_h = \frac{\eta(1-p_b)}{\mu_0 + \eta p_f}\lambda_0, \tag{3}$$

where p_f is the probability of terminating handoff users and p_b the probability of blocking new users. The handoff rate in (3) is a function of p_f which itself is also a function of λ_h, but it can still be solved recursively as suggested in [6]. The channel-holding time of an admitted user in a cell—the time he occupies some bandwidth in that cell—can be computed by taking the minimum of the remaining connection lifetime and cell-sojourn time. Since we assume that both connection lifetime and cell-sojourn time are exponentially distributed, the distribution of channel-holding time can be derived as

$$f_{c0} = (\mu_0 + \eta)e^{-(\mu_0+\eta)t}. \tag{4}$$

Under the proposed degradation scheme, both blocking and forced-termination probabilities can be improved. However, some users may receive severely degraded QoS. In the following section, we investigate the trade offs among the QoS metrics, especially between the blocking probability and the other three QoS metrics.

3 Analysis

We first consider the general case in which there are K different QoS levels and derive the formulas for both DR and UDF. To further demonstrate our model, we will present a simple example and illustrate these derivations. The notations used in this section are listed in Table 1.

Since there are K different QoS levels, we can define the system state, $\bar{\mathbf{n}}$, as

$$\bar{\mathbf{n}} = (n_1, n_2, \ldots, n_K), \tag{5}$$

where n_i is the number of users in the ith QoS level in a cell. Such a system can be modeled as a Markov chain and the transition probabilities can be obtained accordingly. In our model, the transition probabilities depend on the admission

1. We do not consider the change of transmission rate resulting from user mobility.

TABLE 1
The List of Symbols/Notations

N_{thresh}	Restriction threshold for new users
K	Number of QoS levels
W_i	Required bandwidth at QoS level i
n_i	Number of QoS level-i users in a cell
C	Cell capacity
λ_0	Arrival rate of new users
λ_h	Arrival rate of handoff users
$\frac{1}{\mu_0}$	Average connection lifetime
$\frac{1}{\eta}$	Average cell sojourn time

control (i.e., the value of N_{thresh}) and the upgrade/degrade policy. A possible admission and bandwidth allocation algorithm is presented in Fig. 2, where W_a is the currently unused bandwidth and C is the cell capacity. When there is a shortage of bandwidth, allocating only W_{min} to an incoming user minimizes the need to degrade the QoS levels of existing users and, hence, results in smaller DR and UDF. On the other hand, fairness is an important issue when we consider bandwidth reallocation in a system supporting multilevel QoS. We may evenly degrade the QoS of existing users to accommodate a new user or degrade as few users as possible so as to minimize the change of current bandwidth constellation. Therefore, one can make a trade off between fairness and UDF. When a fair degradation algorithm is used, the probability that a user's QoS will be degraded increases (and so does the value of UDF), while using an unfair algorithm as shown in lines 06-11 of Fig. 2 ensures a lower value of UDF. This trade off needs to be made when the bandwidth degradation algorithm is chosen and will be investigated more thoroughly in Section 5. The corresponding upgrade algorithm is shown in Fig. 3 when a QoS level-i user leaves the cell such that an amount of bandwidth, $W_r = W_i$, becomes available in the cell. Here, a fair upgrade algorithm is used to ensure the fairness among the existing users.

3.1 Stationary Distribution of the Number of Connections in a Cell

In order to obtain the stationary distribution of the system state upon arrival of a user's connection request or upon departure of an existing user, we first need to obtain the transition probabilities. Given that the system is in state $\bar{\mathbf{n}} = (n_1, n_2, \ldots, n_K)$ and $\sum_i n_i < N_{thresh}$, if a user arrives at the cell before the departure of any existing user,

$$P_{\bar{\mathbf{n}},\bar{\mathbf{n}}'} = \frac{\lambda_0 + \lambda_h}{\sum n_i\mu + \lambda_0 + \lambda_h}, \tag{6}$$

where $\bar{\mathbf{n}}'$ is determined by lines 06-11 of Fig. 2. If a level-i user leaves the cell,

$$P_{\bar{\mathbf{n}},\bar{\mathbf{n}}'} = \frac{n_i\mu}{\sum n_i\mu + \lambda_0 + \lambda_h}, \tag{7}$$

where $\bar{\mathbf{n}}'$ is determined by the algorithm in Fig. 3. If $\sum_i n_i \geq N_{thresh}$, the transition probabilities can be obtained as (6) and (7) by replacing $\lambda_0 + \lambda_h$ with λ_h. The stationary state distribution, π, can be obtained by solving the equation

$$\pi P = \pi. \tag{8}$$

```
01.  if (the connection request is from a handoff user,
         or from a new user when Σ_{i=1}^{K} n_i < N_thresh) {
02.    if (W_a ≥ W_min)
03.      W_allocated = min(W_max, W_a);
04.    else if (W_a ≤ W_min & (C − Σ_i^K n_i * W_min) ≥ W_min)
05.      { W_allocated = 0;
06.        for (i = 1, i < K, i++)
07.          while (W_allocated < W_min & n_i > 0) {
08.          Randomly degrade one of the n_i connections by
             an amount of W_d = min(W_min, W_i − W_min);
09.          n_i = n_i − 1;
10.          n_j = n_j + 1, where j is such that W_j = W_i − W_d;
11.          W_allocated = W_allocated + W_d; }
12.          n_k = n_k + 1, where W_k = W_allocated; }
13.    else
14.      Reject the connection request; }
15.  else
16.    Reject the connection request;
```

Fig. 2. A pseudocode of the bandwidth degradation algorithm.

Fig. 4 shows the resulting Markov chain for a simple case of $K = 2$, $W_1 = 2$, and $W_2 = 1$. If new users are not differentiated from handoff users (i.e., $N_{thresh} = C$), the stationary distribution of the number of users in a cell can be obtained by Erlang's formula with the arrival rate λ_i set to $\lambda_0 + \lambda_h$ (the arrival rate of newly initiated connections plus that of the handoff connections) and service rate μ_i set to $i \cdot (\mu_0 + \eta)$. If $N_{thresh} < C$, the stationary distribution can still be obtained by solving the local balance equations of the Markov chain in Fig. 4. The stationary state distribution is similar to Erlang's formula except that we now have state-dependent arrival rates,

$$\pi_{n_1,n_2} = \frac{1}{\sum_{i=0}^{C} \frac{\prod_{k=0}^{i-1}\lambda_k}{\mu^i i!}} \times \frac{\prod_{k=0}^{n_1+n_2-1}\lambda_k}{\mu^{n_1+n_2}(n_1+n_2)!}, \tag{9}$$

where $\lambda_k = \lambda_0 + \lambda_h$ if $k < N_{thresh}$ and $\lambda_k = \lambda_h$ otherwise. In either case, the blocking probability p_b is $\sum_{i+j=N_{thresh}}^{C} \pi_{i,j}$, and the forced-termination probability p_f is $\pi_{0,C}$, which can be obtained from (9).

Thanks to the assumptions of homogeneous cells, Poisson arrival process, and the resultant exponential channel-holding time, the statistics for all cells are identical and independent, so the analysis of only one cell is statistically sufficient. Moreover, this stationary distribution is also the probability distribution of the number of connections observed at any arbitrary time instant.

```
01.  n_i = n_i − 1;
02.  for (k = K, k > 1, k − −)
03.      while (W_r > 0 & n_i > 0) {
04.          Randomly upgrade one of the n_k connections
             by one level of quality;
05.          n_k = n_k − 1;
06.          n_{k−1} = n_{k−1} + 1;
07.          W_r = W_r − (W_{k−1} − W_k);}
```

Fig. 3. A pseudocode of the bandwidth upgrade algorithm.

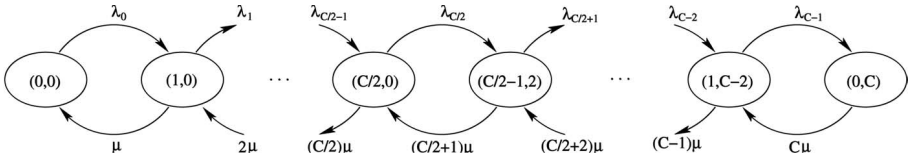

Fig. 4. State transitions of the number of users in one cell.

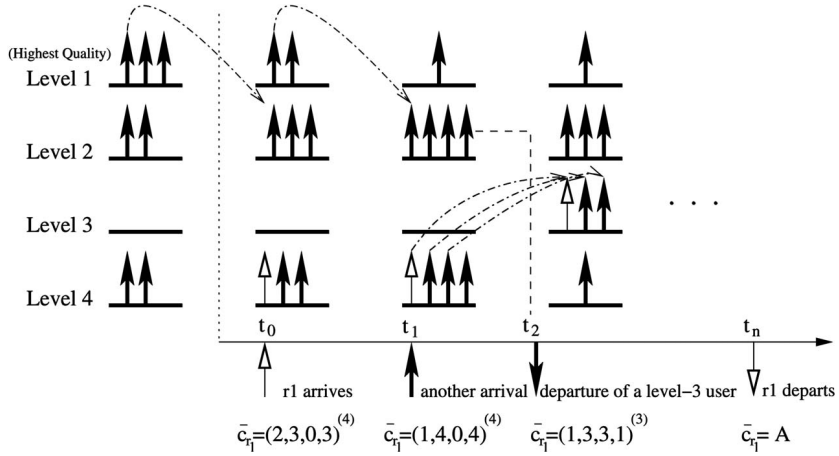

Fig. 5. Transitions of connection states: $K = 4$.

3.2 QoS Metrics

As mentioned in the previous section, the QoS assigned to an admitted user may vary during his connection lifetime. From the perspective of an admitted user, giving only the system state $\bar{\mathbf{n}} = (n_1, n_2, \ldots, n_k)$ does not tell the QoS level it is receiving. In order to derive the DR and UDF of an admitted user, we create a new state called "connection state" to correctly reflect the evolution of the QoS level an admitted user is receiving. The connection state is defined as

$$\bar{c} = \bar{\mathbf{n}}^{(i)}, \tag{10}$$

where i is the quality index which represents the current QoS level of an admitted user. For example, let us consider a system with $K = 4$, $W_i = 5 - i$ for $i = 1$ to K, and $C = 20$. We focus on a specific user, r_1, and trace the changes of its received QoS during its stay in a cell. If it arrives when the cell is in system state $(3, 2, 0, 2)$, r_1's initial connection state will be $\bar{\mathbf{c}}_{r_1} = (2, 3, 0, 3)^{(4)}$ simply because one of the level-1 users is degraded to level 2 and r_1 receives level-4 (minimum) QoS, according to the algorithms introduced before. If another handoff user joins the cell at $t = t_1$, r_1's connection state will change to $\bar{\mathbf{c}}_{r_1} = (1, 4, 0, 4)^{(4)}$ since one of the level-1 users is degraded, but r_1 still receives level-4 QoS. If a level-3 user leaves the cell at $t = t_2$ and r_1 is chosen to be upgraded (with probability $\frac{3}{4}$) with the other two level-4 users, $\bar{\mathbf{c}}_{r_1} = (1, 3, 3, 1)^{(3)}$. These transitions are illustrated in Fig. 5 and we can again model r_1's received QoS levels as an embedded Markov chain $\{Y_{t_n}\}$. In the above example, $Y_{t_0} = (2, 3, 0, 3)^{(4)}$, $Y_{t_1} = (1, 4, 0, 4)^{(4)}$, and $Y_{t_2} = (1, 3, 3, 1)^{(3)}$, where t_i is the occurrence time of the ith event (either an arrival of a connection request or a departure of any existing user). Finally, if r_1 leaves the cell at t_n, we let $Y_{t_n} = A$. Here, A is a completion (absorption) state because, once it enters A, it will stay there forever. For simplicity of notation, we use $\bar{\mathbf{c}}$ as the connection state of any arbitrary user. The transition probability $P_{\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2}$ can be obtained based on the algorithms introduced in the previous section and the detailed derivation will be presented later for the case of $K = 2$.

3.3 Degradation Ratio

We now derive the DR of an admitted user, based on the embedded Markov chain described above. First, we need to derive $N_{\bar{c}_j}$, the number of visits to state $\bar{c}_j$ before entering the completion state A, given that the initial state is $\bar{c}_i$:

$$E_{\bar{c}_i}(N_{\bar{c}_j}) = E_{\bar{c}_i}\left[\sum_{n=0}^{\infty} 1_{\{Y_{t_n} = \bar{c}_j\}}\right] = \sum_{n=0}^{\infty} P_{\bar{c}_i \bar{c}_j}(n), \tag{11}$$

where Y_{t_n} is the state after the nth transition and $P_{\bar{c}_i \bar{c}_j}(n)$ is the n-step transition probability from state $\bar{c}_i$ to state $\bar{c}_j$. $\sum_{n=0}^{\infty} P_{\bar{c}_i \bar{c}_j}(n)$ is also the (i, j)th element of potential matrix G, which can be obtained by the following equation

$$G = \sum_{n=0}^{\infty} P^n, \tag{12}$$

where P is the transition matrix of the embedded Markov chain. We can rewrite P as

$$P = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{T_A} & \mathbf{T_T} \end{bmatrix}, \tag{13}$$

where $P_{AA} = 1$ and $P_{A\bar{c}_i} = 0$ since A is an absorption state. T_T is the restriction of P to the transient set, or the transition probabilities between transient states, while T_A represents the transition probabilities between transient states and A. Since we only consider the number of visits to the transient states before entering the completion state A, the potential matrix can be further rewritten as

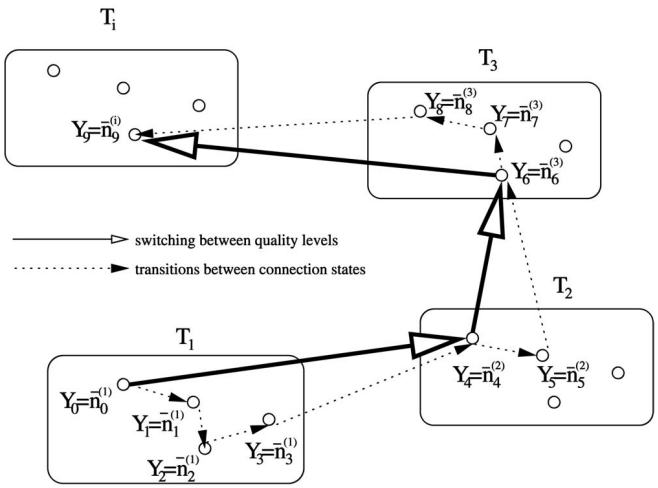

Fig. 6. Transitions between different QoS levels.

$$G = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{F} & \mathbf{S} \end{bmatrix}, \tag{14}$$

where $S = \sum_{n=0}^{\infty} T_T^n$. Thus, $E_{\bar{c}_i}(N_{\bar{c}_j})$ is just the (i,j)th element of matrix S. By matrix manipulation, S can be computed by the following equation [18],

$$\mathbf{S} = (\mathbf{I} - \mathbf{T_T})^{-1}. \tag{15}$$

Before deriving the formula for DR, we first need to calculate DR given that the initial state is $\bar{\mathbf{c}}_i$,

$$\mathrm{DR}_{\bar{c}_i} = \mu \sum_{k=1}^{K} \frac{W_{max} - W_k}{W_{max}} \sum_{\{\bar{c}=\bar{n}^{(k)}:\, k>0\}} \frac{E_{\bar{c}_i}(N_{\bar{c}})}{\lambda + \sum n_j \mu}, \tag{16}$$

where $\lambda = \lambda_0 + \lambda_h$ if $\sum n_i < N_{thresh}$ and $\lambda = \lambda_h$ otherwise. The second summation of (16) is the average time a user spends in the kth QoS level and is calculated by taking the product of the average number of visits to that QoS level and the average duration of each visit. The inverse of μ is the average total time a user spends in a cell, namely, the term $\sum_i T_i$ in (1). Finally, DR can be obtained by (16) as

$$\mathrm{DR} = \sum_{\bar{n}} \pi_{\bar{n}} \cdot P(\bar{c}|\bar{n}) \cdot \mathrm{DR}_{\bar{c}}, \tag{17}$$

where $\pi_{\bar{n}}$ is the stationary distribution of the system state and can be obtained by (8). The conditional probability, $P(\bar{c}|\bar{n})$, is decided by the admission control and degradation policy. For example, we have $P(\bar{c} = (2,3,0,3)^{(4)}|\bar{n} = (3,2,0,2)) = 1$ in the previous example.

3.4 Upgrade/Degrade Frequency

Let us consider how to derive UDF—the average number of switches per unit time between different QoS levels. Since only the transition between a user's connection states with different quality indexes will be counted as a QoS-level switch, we have to group the connection states with the same quality index into a "super state." Let $\mathbf{T_i}$ be such a super state (or set): $\{\bar{c} : \bar{n}^{(i)}\ \forall \bar{n} \text{ and } n_i > 0\}$ for $i = 1$ to $i = K$. For example, as shown in Fig. 6, even though there are three transitions at $t = 0$, 1 and 2, there is no quality switch because the user r_1 keeps receiving level-1 QoS. The only difference between $t = 0$ to $t = 2$, from r_1's perspective, is the number of users in other QoS levels. Until $t = 3$, r_1 is chosen to be degraded (i.e., the fourth transition between $Y_3 = \bar{n}_3^{(1)} \rightarrow Y_4 = \bar{n}_4^{(2)}$) and so is $Y_5 \rightarrow Y_6$ and $Y_8 \rightarrow Y_9$. Let $\tilde{Y}_n = Y_n$ be the new process which samples the original Y_n at the times of QoS-level switch, we have $\tilde{Y}_0 = Y_0$, $\tilde{Y}_1 = Y_4$, $\tilde{Y}_2 = Y_6$, and so on. Thus, $\{\tilde{Y}_n\}$ can also be modeled as a discrete Markov chain with the transient probability matrix $\tilde{\mathbf{P}}$ obtained as follows:

- If $\bar{c}_i \in \mathbf{T_h}$, then $\tilde{p}_{\bar{c}_i\bar{c}_j} = 0$ for $\bar{c}_j \in \mathbf{T_h}$.
- For $\bar{c}_i \in \mathbf{T_h}$ and $\bar{c}_j \in \cup_{k=1,k\neq h}^{K} \mathbf{T_k}$, $\tilde{p}_{\bar{c}_i\bar{c}_j}$ is the probability of being absorbed in the states, $\cup_{k=1,k\neq h}^{K} \mathbf{T_k}$ of the Markov chain with transition matrix $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}} = \begin{matrix} \\ \cup_{k=1,k\neq h}^{K} \mathbf{T_k} \\ \mathbf{T_h} \end{matrix} \begin{matrix} \cup_{k=1,k\neq h}^{K} \mathbf{T_k} & \mathbf{T_h} \\ \end{matrix} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{B_h} & \mathbf{Q_h} \end{pmatrix}, \tag{18}$$

  where $\mathbf{B_h}$ is the transition probability matrix of the set $\mathbf{T_h}$ to all other states, and $\mathbf{Q_h}$ is the restriction of $\hat{\mathbf{P}}$ to the set $\mathbf{T_h}$. Then, $\tilde{p}_{\bar{c}_i\bar{c}_j} = ((I - \mathbf{Q_h})^{-1}\mathbf{B_h})_{\mathbf{\bar{c}_i\bar{c}_j}}$ as we derived (15).

Having $\tilde{\mathbf{P}}$ this way, the average time to the absorption state A is then the number of switches between T_is. If we rewrite $\tilde{\mathbf{P}}$ as

$$\tilde{P} = \begin{matrix} A \\ \cup_{k=1}^{K} \mathbf{T_k} \end{matrix} \begin{matrix} A & \cup_{k=1}^{K} \mathbf{T_k} \\ \end{matrix} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{T_A} & \tilde{\mathbf{Q}} \end{pmatrix}, \tag{19}$$

then the average number of QoS-level switches before a connection is completed or handed off, given an initial state $\bar{c}$, is

$$E[N_d]_{\bar{c}} = (1 - \tilde{\mathbf{Q}})^{-1}\mathbf{1} - 1. \tag{20}$$

Finally, we can obtain UDF as we derived DR in (17)

$$\mathrm{UDF} = \mu \sum_{\bar{n}} \pi_{\bar{n}} \cdot P(\bar{c}|\bar{n}) \cdot E[N_d]_{\bar{c}}. \tag{21}$$

3.5 A Simple Case: $K = 2$

With the analytical model in the previous section, let us consider a simple case with $K = 2$, $W_1 = 2$, and $W_2 = 1$ (e.g., a video telephony with standard quality (= 40 Kbps) and low-motion quality (= 20 Kbps)). The resulting embedded Markov chain for the connection state of an admitted user, r_1, is shown in Fig. 7 and the transition probabilities can be derived as follows: Since there are only two QoS levels, we will denote r_1's connection states $\bar{c} = (n1, n2)^{(1)}$ as f_{n1+n2} ("f" as full quality), and $\bar{c} = (n1, n2)^{(2)}$ as d_{n1+n2} ("d" as degraded quality). Now, let r_1 be in any given state. Then, three different events may occur: the arrival of a new user, the departure of r_1, or the departure of any other existing users. We need to differentiate several situations in order to calculate the transition probabilities as follows:

- For state f_i, $1 \leq i \leq \frac{C}{2} - 1$, all existing users receive full quality. Three transition probabilities of this state are $P_{f_i,f_{i+1}} = \frac{\lambda_i}{\lambda_i + i\mu}$ when another user joins the cell, $P_{f_i,A} = \frac{\mu}{\lambda_i + i\mu}$ when r_1 leaves the cell before any other event, and $P_{f_i,f_{i-1}} = \frac{(i-1)\mu}{\lambda_i + i\mu}$ when any other existing user leaves the cell earlier than r_1.

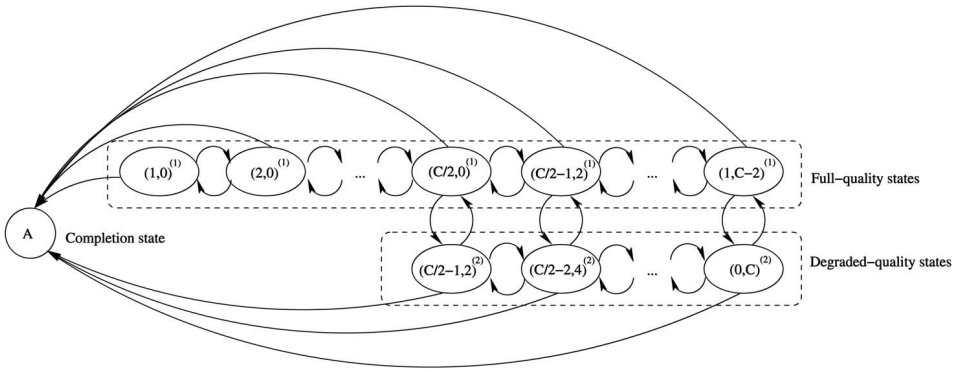

Fig. 7. Transitions of r_1's connection states in a cell.

- For state f_i, $\frac{C}{2} \le i \le C-1$, the arrival of a new user may result in two different transitions. One is that r_1 is degraded such that its connection state transits to the degraded state d_{i+1}. The other is that r_1 is not degraded such that its connection state transits to f_{i+1}. The associated transition probabilities are $P_{f_i,d_{i+1}} = \frac{\lambda_i}{(C-i)(\lambda_i+i\mu)}$ and $P_{f_i,f_{i+1}} = \frac{(C-i-1)\lambda_i}{(C-i)(\lambda_i+i\mu)}$, respectively. The other transition probabilities are $P_{f_i,A} = \frac{\mu}{(\lambda_i+i\mu)}$ and $P_{f_i,f_{i-1}} = \frac{(i-1)\mu}{(\lambda_i+i\mu)}$, which represent the probabilities that r_1 and any other user leaves the cell, respectively.
- For state $d_i, \frac{C}{2}+1 \le i \le C$, the departure of any other user may result in two different transitions. One is that r_1 is upgraded because of the others' departure(s) such that its connection state transits to f_{i-1}. The other is that r_1 continues receiving degraded quality and, thus, its connection state transits to d_{i-1}. The associated transition probabilities are $P_{d_i,f_{i-1}} = \frac{\frac{C}{2}}{(i-\frac{C}{2})\frac{\mu}{\lambda_i+i\mu}}$ and
$$P_{d_i,d_{i-1}} = i\left(1-\frac{1}{i-\frac{C}{2}}\right)\frac{\mu}{\lambda_i+i\mu}.$$
The other transition probabilities are $P_{d_i,d_{i+1}} = \frac{\lambda_i}{\lambda_i+i\mu}$ when another user joins the cell and $P_{d_i,A} = \frac{\mu}{\lambda_i+i\mu}$ when r_1 ends its stay in this cell.
- Note that $\lambda_N = 0$.

DR_i can be obtained as in (16), but here we slightly change it to

$$\mathrm{DR}_{\vec{c}} = \sum_{d_j \in \{degraded\ quality\}} \mu E_{\vec{i}}(N_{d_j}) T_{sojourn,d_j} \tag{22}$$

such that DR will be the fraction of time the user spends in the degrade QoS level. The mean sojourn time in state d_j, $T_{sojourn,d_j}$, is $\frac{1}{\lambda_j+j*\mu}$ according to (4). Finally, DR can be computed according to (17)

$$\mathrm{DR} = \sum_{i=0}^{\frac{C}{2}-1} \pi_{i,0}\mathrm{DR}_{f_{i+1}} + \sum_{i=\frac{C}{2}}^{C-1} \pi_{C-i,2i-C}\mathrm{DR}_{d_{i+1}}, \tag{23}$$

where π_{n_1,n_2} is given in (9).

Since there are only two types of switching (i.e., QoS degradation: $f_i \to d_{i+1}$ or QoS upgrade: $d_i \to f_{i-1}$), we can use the first-step analysis to derive UDF, and the following system of linear equations can be obtained:

$$E(D_{f_i}) = \sum_{j,j\neq i} P_{f_i,f_j} E(D_{f_j}) + \sum_j P_{f_i,d_j}(E(D_{d_j})+1), \tag{24}$$

$$E(D_{d_i}) = \sum_j P_{d_i,f_j}(E(D_{f_j})+1) + \sum_{j,j\neq i} P_{d_i,d_j} E(D_{d_j}), \tag{25}$$

where $E(D_{f_i})$ ($E(D_{f_i})$) is the average number of quality switches r_1 will perceive given that its initial connection state is f_i (d_i). The solution to this system of linear equations can be computed as

$$\mathbf{E(D)} = (\mathbf{I} - \mathbf{T_T})^{-1}\mathbf{C}, \tag{26}$$

where $\mathbf{C}$ is the column vector with the ith element equal to $P_{f_i,d_{i+1}}$ for $1 \le i \le C-1$ or $P_{d_i,f_{i-1}}$ for $\frac{C}{2}+1 \le i \le C$. By using (15), the vector $\mathbf{E(D)}$ can be rewritten as

$$\mathbf{E(D)} = \mathbf{SC}. \tag{27}$$

UDF can then be obtained as:

$$\mathrm{UDF} = \sum_{i=0}^{\frac{C}{2}-1} \mu\pi_{i,0} E(D_{f_{i+1}}) + \sum_{i=\frac{C}{2}}^{C-1} \mu\pi_{C-i,2i-C} E(D_{d_{i+1}}). \tag{28}$$

Note that the DR and UDF derived so far are the QoS metrics a handoff user may experience in each cell. The values of these QoS metrics for a user in the cell where his connection was initiated are different, but similar formulas can still be derived by considering the restriction threshold

$$\mathrm{DR}_I = \sum_{i=0}^{min(N_{thresh},\frac{C}{2})-1} \pi_{i,0} DR_{f_{i+1}} + \sum_{i=min(N_{thresh},\frac{C}{2})}^{N_{thresh}-1} \mu\pi_{C-i,2i-C} DR_{d_{i+1}}$$

$$\mathrm{UDF}_I = \sum_{i=0}^{min(N_{thresh},\frac{C}{2})-1} \mu\pi_{i,0} E(D_{f_{i+1}}) + \sum_{i=min(N_{thresh},\frac{C}{2})}^{N_{thresh}-1} \mu\pi_{C-i,2i-C} E(D_{d_{i+1}}),$$

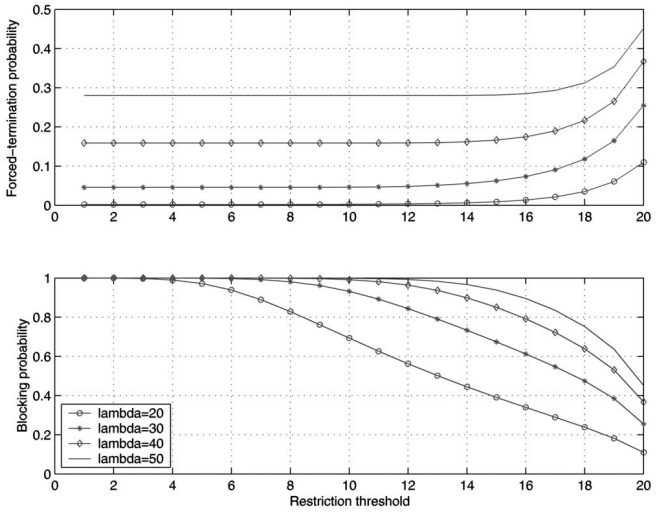

Fig. 8. P_b and P_f versus arrival rate: no quality degradation.

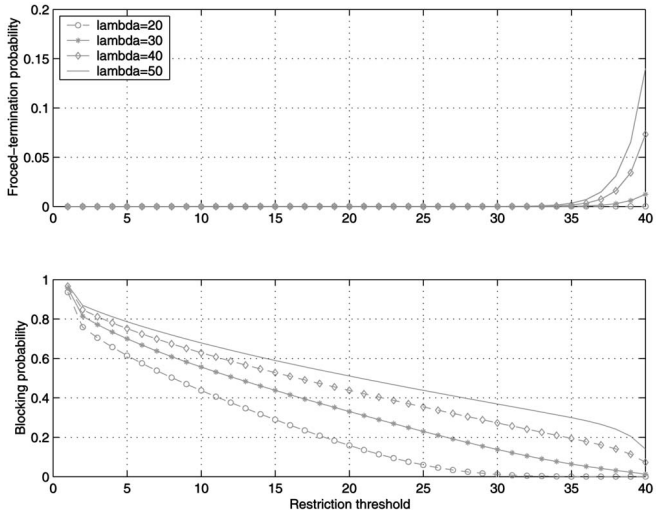

Fig. 9. P_b and P_f versus arrival rate with quality degradation.

where DR_I and UDF_I will be the QoS metrics for a user in the cell where his connection was initiated (i.e., the 0th cell in Fig. 1).

4 Numerical Results

We use the traffic models in Section 2, where the arrival process of new users is assumed to be Poisson and the connection lifetime and cell-sojourn times are exponentially distributed. The formulas for the resulting handoff rate and channel-holding time can be found in (3) and (4). For illustrative purposes, we first consider the case with $K = 2$ and assume that each full-quality connection requires twice the bandwidth of a degraded-quality connection. The impact of arrival rates, connection lifetime, and user mobility on the QoS metrics are discussed. Then, we consider the case of $K = 3$ and show how the bandwidth allocation algorithm will affect the QoS metrics, especially the trade off between UDF and user fairness.

4.1 $K = 2$: Full and Degraded Quality

We assume that each cell can accommodate up to 40 degraded-quality connections in the following discussion. Four QoS metrics—blocking probability of new users (P_b), forced-termination probability of handoff users (P_f), degradation ratio (DR), and upgrade/degrade frequency (UDF)—are evaluated. Since the users' arrival rate, connection lifetime, and mobility ($\propto \frac{1}{\eta}$) could significantly affect these metrics, three sets of analysis are performed to investigate their impact under various restriction thresholds. The restriction threshold ranges from 1 to 40 in each numerical analysis. If the restriction threshold is 1, the traffic restriction is applied at state $(1, 0)$ and higher states of the Markov chain shown in Fig. 4 and at most one new user could be admitted into the system (e.g., most users in a cell are handoff users from adjacent cells). On the other hand, if the restriction threshold is 40, no bandwidth is reserved for handoff users and, thus, there is no distinction between new and handoff users. Selection of the restriction threshold under different traffic loads is also discussed at the end of this section.

4.1.1 QoS Metrics versus Arrival Rate of Connection Requests

Four different arrival rates—20, 30, 40, 50 users per unit time are considered to represent different cell loads. We first consider the case of no QoS degradation for comparison with our scheme. The resulting P_b and P_f are plotted in Fig. 8. The figure shows how the handoff users can be given higher priority by controlling the restriction threshold. A lower restriction threshold prevents a new user from joining a cell, even when there are only few users in that cell, and thus results in a higher P_b. On the other hand, a handoff user will have a better chance of being successfully handed off to a cell since most of the new users have been blocked.

Even though the restriction threshold provides differentiated treatment for new and handoff users, both P_b and P_f are still high. For example, for $\lambda_0 = 30$ and $N_{thresh} = 16$, P_b is as high as 0.6 and P_f is about 0.1. Fig. 9 plots P_b and P_f when QoS degradation is applied. It should be noted that the maximum number of admissible users in a cell now is 40, as compared to 20 in the case of no QoS degradation, since each degraded-quality user requires only one unit of bandwidth. Therefore, the maximum restriction threshold we can choose is 40. With the help of QoS degradation, P_f is negligible and P_b is less than 0.1 for $\lambda_0 = 30$ when a high restriction threshold is used. Even in the case of heavy loads ($\lambda = 50$), P_b and P_f are only 0.18, as compared to 0.45 if using the restriction mechanism only. This result shows that QoS degradation is an effective way of reducing both P_b and P_f. Together with a proper choice of the restriction threshold, we are able to maintain much lower P_b and P_f while still giving handoff users priority over new users.

Fig. 10, however, shows the fact that the improvement on P_f and P_b (by using the QoS degradation scheme) could cause individual users' severe QoS degradation. Both DR and UDF increase with the cell's load which can be increased either by increasing the user arrival rate or the restriction threshold. For example, DR could be as high as 0.8 when $\lambda = 50$ and $N_{thresh} = 35$, mainly because the cell keeps admitting users to the extent that most of them have to receive degraded QoS. UDF increases even more quickly

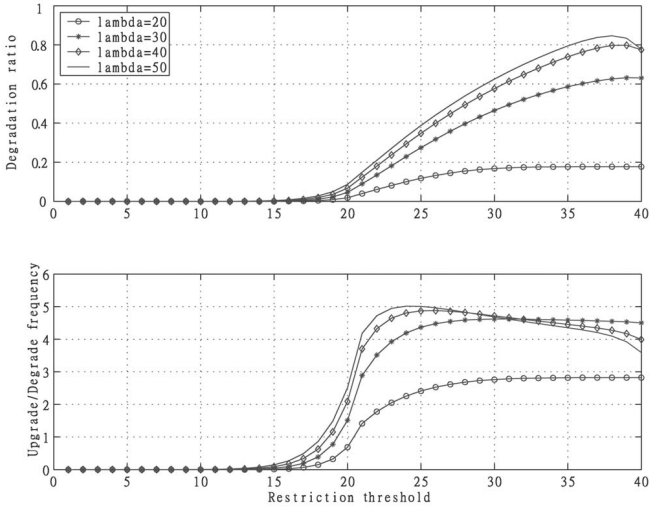

Fig. 10. DR and UDF versus arrival rate.

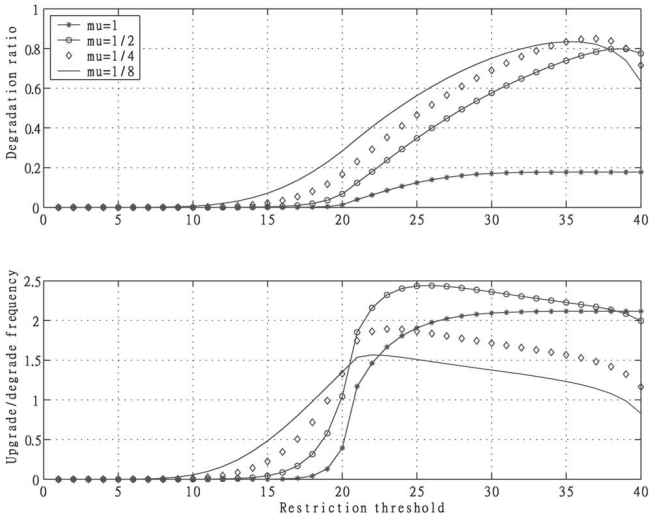

Fig. 11. DR and UDF versus connection lifetime.

than DR as we increase the restriction threshold. For example, UDF can be as high as five in the case of moderate load even though a cell reserves 40 percent of bandwidth for handoff users. One may find that there is a slight drop in the UDF value in the case of high loads and high restriction threshold. This is because of a sharp increase of P_f (in Fig. 9), which, in turn, reduces the handoff rate. In summary, the results show that, even though P_b and P_f can be improved significantly (and, thus, the bandwidth utilization) by QoS degradation, each individual user may receive very poor QoS. Therefore, one must control both DR and UDF when applying this degradation mechanism in order to maintain a balance between the "benefits" of the service provider and users. We will discuss more about this issue in the following sections.

4.1.2 QoS Metrics versus Connection Lifetime

In this section, we vary the user's average connection lifetime: $\frac{1}{\mu} = 8$, 4, 2, and 1 units of time, with the new users' arrival rate fixed at 20 users per unit time. The plots for P_b and P_f are similar to Fig. 9 and, therefore, are omitted here. DR and UDF under these connection lifetimes are plotted in Fig. 11. Unlike the previous case in which both DR and UDF increase with the cell's load, DR and UDF here react to the changes differently. DR still increases with the load, which is now caused by the increase of the connection lifetime or restriction threshold. However, it is much more complicated to interpret the result of UDF. In the case of higher restriction threshold (e.g., 35), the UDF for $\mu = \frac{1}{8}$ is a half of that for $\mu = \frac{1}{2}$. That is, UDF decreases with the increase of connection lifetime. This is because, when the cell is heavily loaded, a longer connection lifetime helps reduce the departure rate of the existing users. Thus, fewer adjustments of bandwidth constellation are needed, resulting in a smaller UDF. In the case of a low restriction threshold, many new users are blocked. Together with a shorter user connection lifetime, the total traffic load is much smaller[2] such that most users would not interfere with one another. This, in turn, results in a smaller UDF, explaining the crossover of UDF under different μs when the threshold changes.

The complexity of UDF's dependency on a cell's load makes it much more difficult to control the QoS metrics. If the cell's load increase is due to a higher user arrival rate, reducing the restriction threshold will be necessary to lower UDF. However, if it is due to a longer user connection lifetime, reducing the threshold only increases the blocking probability without decreasing UDF much. Thus, monitoring only the cell load will not be enough. Instead, the admission and bandwidth allocation policy has to adapt to each of these affecting factors (i.e., arrival rate or connection lifetime) in order to achieve better performance.

4.1.3 QoS Metrics versus Mobility

We now vary the average cell-sojourn time— $\frac{1}{\eta} = 0.5$, 1, 2, and 4 units of time—to study the effects of user mobility on the QoS metrics. In all cases, the plots for P_b and P_f are still similar to Fig. 9 when we change the restriction threshold, but they are much less sensitive to the changes of cell-sojourn time. Even though higher user mobility

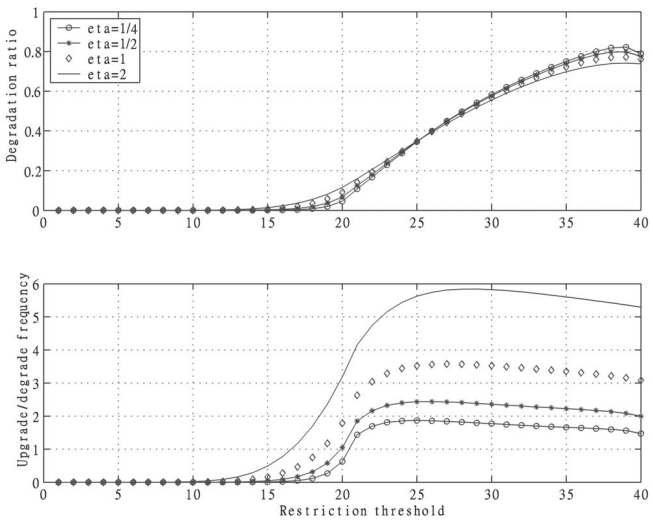

Fig. 12. DR and UDF versus mobility.

2. Recall that λ is fixed here.

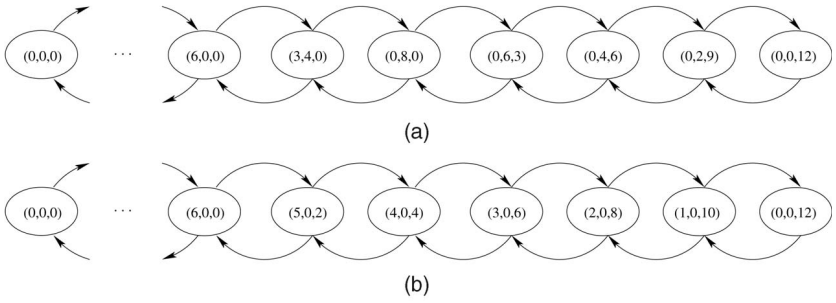

Fig. 13. State-transition diagram. (a) Complete-fair algorithm. (b) UDF-minimizing algorithm.

could contribute to more user's cell crossings and, thus, higher handoff rates, it also reduces the user's channel-holding time in each cell. Thus, the actual cell load does not change much when we change the user mobility (or, more precisely, the cell-sojourn time) and, thus, P_b and P_f do not change much either. This insensitivity to the changes of user mobility can also be found in DR as shown in Fig. 12. Since the load of a cell remains unchanged, the amount of time a user stays at each QoS level will be statistically the same, thus keeping DR unchanged when we vary η. However, UDF can be much larger in the case of higher mobility than in the case of lower mobility, despite the fact that changing user mobility hardly affects the cell's load. For instance, $\mathrm{UDF} \approx 6$ when $\eta = 2$, but $\mathrm{UDF} \approx 2$ when $\eta = \frac{1}{4}$, both in the case of $\mathrm{threshold} = 27$. These findings confirm the need for considering both DR and UDF for adaptive QoS provisioning. In the case of higher mobility, UDF is the dominant factor in individual connections' QoS.

Based on the results in the previous two sections and those in this section, we can conclude that DR depends almost only on the cell's load, irrespective of the factor that contributes to it. This makes it much easier to maintain a preferred DR since we only need to control the cell's load. On the other hand, the impact of arrival rate, connection lifetime, and user mobility on UDF are all different from each other. Depending on which factor is in effect, we may need a different approach to maintaining UDF. That is, the real challenge for maintaining the user-perceived QoS lies in the control of UDF, instead of the commonly believed DR.

4.2 $K = 3$: Fairness versus UDF

As we mentioned in Section 3, the bandwidth reallocation algorithm may affect not only the DR/UDF but also fairness among the existing users. By "fairness" we mean that the service provider should allocate the bandwidth to all existing users in an egalitarian way. Therefore, if the user is admitted into a cell, he should receive a QoS level as close as possible to that of the existing users. On the other hand, if the existing users' QoS needs to be degraded, we should choose users in an ascending order of their current QoS levels and evenly degrade their QoS to ensure fairness among the users. The same policy should be applied in upgrading users' QoS levels except in a reverse order. Fig. 13a shows the system state transitions under such a fair reallocation algorithm in the case of $C = 24$, $K = 3$ with $W_1 = 4$, $W_2 = 3$, and $W_3 = 2$. For example, when a user arrives at state $(6, 0, 0)$, in order to allocate as much bandwidth as possible to the new user, three level-1 users are degraded by one QoS level. The resulting state is then $(3, 4, 0)$. Obviously, the fairness is achieved at the expense of more QoS-level switches for the existing users. At the other end of the spectrum, we may allocate a minimum amount of bandwidth to an incoming user by degrading as few existing users as possible. When an existing user leaves the cell, we may reallocate the freed bandwidth with a minimum adjustment of the current bandwidth constellation. The state transitions of such an "unfair" algorithm are shown in Fig. 13b. This time, if a user arrives when the cell is in state $(6, 0, 0)$, only one existing user is degraded by two QoS levels with the freed bandwidth reallocated to the new user. The resulting state will be $(5, 0, 2)$. Since this unfair algorithm only requires a minimum adjustment of the current bandwidth allocation, a minimum UDF can be achieved. Thus, we will call it a *UDF-minimizing* algorithm in the following discussion.

Fig. 14 plots DR under the completely-fair and UDF-minimizing algorithms. The values of DR under these two algorithms are the same for all restriction thresholds. This is because, when the cell is fully utilized, the total amount of degradation is independent of the bandwidth allocation algorithm given that the total number of users in the cell is the same. For example, the total amount of degradation in state $(3, 4, 0)$ of Fig. 13a is $4 = 4 \cdot 1$ because four users receive level-2 quality. It can also be calculated by

$$\sum_{i=1}^{K} n_i \cdot W_1 - C \tag{29}$$

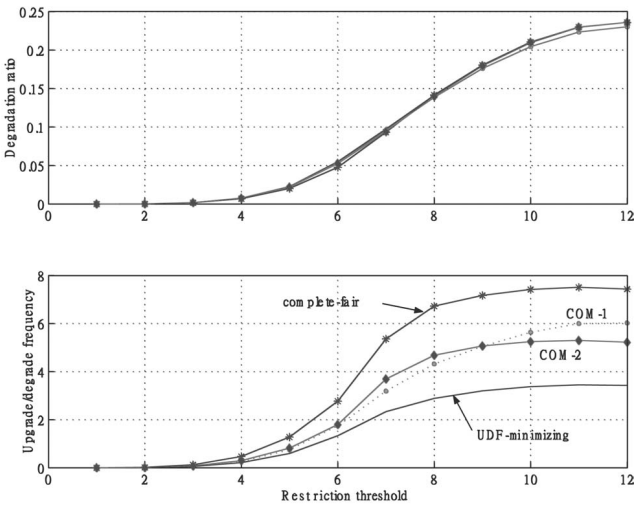

Fig. 14. Fairness versus UDF.

for $(i = K,\ i > 0,\ i - -)$
  while $(W_{allocated} < W_{min}\ \&\ n_i > 0)$ {
    Randomly degrade one of the n_i users by 1 quality level
    $n_i = n_i - 1$;
    $n_{i-1} = n_{i-1} + 1$.
    $W_{allocated} = W_{allocated} + (W_{i-1} - W_i)$; }}

(a)

for $(i = 1,\ i < K,\ i + +)$;
  while $(W_r > 0\ \&\ n_i > 0)$ {
    Randomly upgrade one of the n_i users
    by $\min(W_r, W_{max} - W_i)$ units of bandwidth.
    $n_i = n_i - 1$.
    $n_j = n_j + 1$, where j is such that
    $W_j = \min(W_r, W_{max} - W_i)$.
    $W_r = \max(0, W_r - W_{max} + W_i.$ }

(b)

Fig. 15. Bandwidth reallocation algorithm: Com-2.

because the cell is now fully utilized. If we use the UDF-minimizing algorithm, the total amount of degradation in state $(2, 0, 5)$ of Fig. 13b is also $4 = 2 \cdot 2$, simply because two users receive level-3 quality. Therefore, the average quality degradation for each users will be the same (i.e., $\frac{4}{7}$), regardless of the algorithm used. However, the impact of the reallocation algorithm on the UDF is significant. As also shown in Fig. 14, the values of UDF under the completely-fair algorithm are almost twice those under the UDF-minimizing algorithm. Even though UDF can be minimized as a result of the minimal adjustment of resource allocation, it is extremely unfair in the sense that some users are severely degraded while others receive full quality (e.g., in state (5,0,2), (4,0,4), etc., in Fig. 13b). Between these two extremes there are algorithms with different combinations of fair and unfair bandwidth reallocation. "COM-1" is our proposed bandwidth allocation policy which applies unfair degradation and fair upgrade, while "COM-2" enforces fair degradation and unfair upgrade as shown in Fig. 15. With the help of this combination, a fairer algorithm with a smaller UDF can be derived as shown in Fig. 14.

4.3 Adaptive Admission Control and Bandwidth Allocation

In our proposed scheme, there are two mechanisms, restriction threshold and bandwidth reallocation algorithm, that we can use to control the QoS. Clearly, we must adapt these two mechanisms to different traffic characteristics or the QoS metrics of interest in order to achieve the best performance.

4.3.1 Adaptive Restriction Threshold

From the previous discussion, we know that there exists a trade off between the blocking probability of new users and the other QoS metrics as we change the restriction threshold. So, there does not exist an optimal value of restriction threshold that optimizes *all* four QoS metrics. Since the forced-termination probability increases sharply only when the restriction threshold is close to the cell capacity, we may let $N_{thresh} \approx C$ if the blocking and forced-termination probabilities are the only metrics of interest. For example, the restriction threshold can be 37 or 38 as shown in Fig. 9. However, DR usually attains its maximal value when the restriction threshold is large, meaning that users' received QoS is severely degraded. Thus, we want to choose the threshold to be about a half of cell capacity (e.g., ≈ 25) and then DR can be significantly improved (from 0.8 to 0.4) with only a slight increase of P_b (by 0.1). It should be noted that P_f is negligible and UDF almost remains unchanged between $N_{thresh} = \frac{C}{2}$ and C. This means that a user could have a 50 percent better DR with the same P_f and UDF at the expense of 10 percent more chance of blocking. The same conclusion can be drawn from the results in Fig. 12 if we set the threshold close to one half of the cell capacity instead of setting to higher values. Both DR and UDF decrease significantly—DR decreases from 0.6 to 0.1 in all cases, while UDF decreases from 6 to 3 in case of high mobility and from 2 to 0.8 in case of low mobility—with an increase of P_b by 0.2 in the worst case.

4.3.2 Adaptive Bandwidth Allocation

In addition to adjusting the restriction threshold, we may also use different bandwidth allocation policies to achieve the desired QoS. Since DR only depends on the cell load, as we discussed earlier, changing the bandwidth allocation policy offers the advantage of improving the other QoS metrics, especially UDF, without degrading users' DR. For example, if the user's mobility is high such that UDF is unacceptable, we may use the UDF-minimizing algorithm to get a lower UDF. As shown in Fig. 14, UDF can be reduced by 40 percent as compared to COM-1 without changing P_b, P_f, and DR.

5 Simulation

In the previous analysis, we assumed that the cell-sojourn time is exponentially distributed mainly for mathematical tractability. In order to verify the applicability of our model when this assumption does not hold, we built a simple event-driven simulator written in C++. We consider a wireless network of 30 cells, as shown in Fig. 16. Each cell generates new connection requests according to the Poisson arrival process. Upon receiving a user's request for setting up a connection, each cell will

1. perform the admission control,
2. if the user can be admitted, decide and schedule the departure time according to the distributions of his connection lifetime and cell-sojourn time,
3. if the answer to item 2 is yes, randomly choose a target cell from the neighbor list of the current cell if he needs to be handed off, and
4. perform the bandwidth reallocation algorithm.

A complete flow-chart of our simulator is given in Fig. 17.

For the purpose of comparison, we assume that each cell can accommodate 40 degraded-quality connections. Moreover, the statistics of boundary cells (i.e., cells 7-11, 17-21, 24-29 in Fig. 16) are not taken into account in comparison with the numerical analysis of the previous section. Both heavy-load (40 new users per unit time) and light-load (20 users per unit time) cases are considered. Three

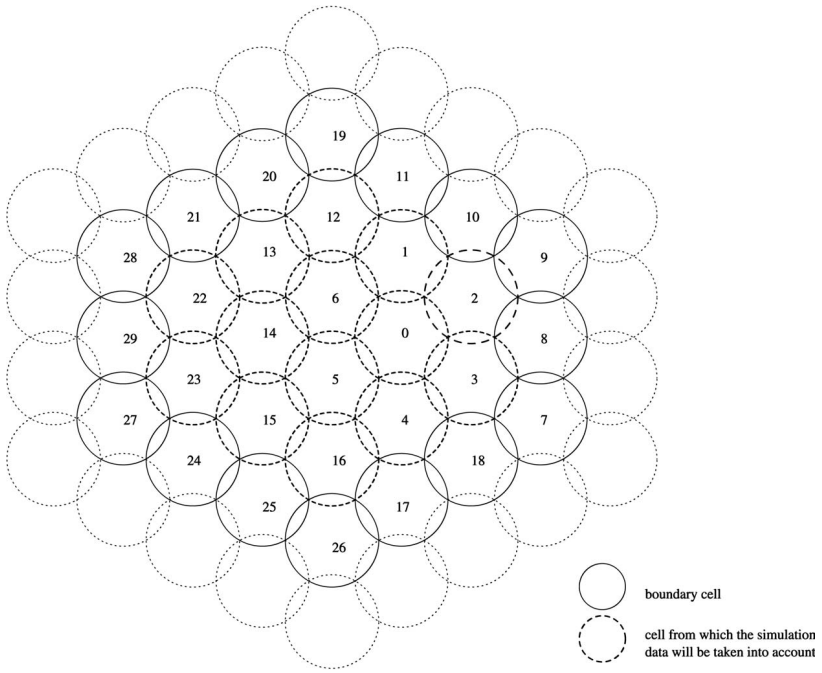

Fig. 16. The cellular network used in the simulation.

distributions of the cell-sojourn time—exponential, uniform, and normal distributions—are used with mean of 1 unit of time and variance of 1 (except for the case of uniform distribution). All other parameters are kept the same as in the numerical analysis. The simulation results are plotted in Fig. 18. Both DR and UDF are plotted with the numerical results of the previous section (solid lines). In both cases, most of the simulation results are close to our numerical results. The largest error of DR is about 15 percent when the arrival rate is 40 and the restriction threshold is 25, while the largest error of UDF is 18 percent when the arrival rate is 40 and the restriction threshold is 20. A reason for such deviation could be that the number of cells is not infinite and, thus, the effect of boundary cells may introduce the error. However, even when the distribution of cell-sojourn time is uniformly or normally distributed, the results are, in general, consistent with our analytical model. As mentioned in Section 2, we model user mobility with the cell-sojourn time because it is the main factor in deciding each user's channel-holding time during his stay in that cell. Moreover, the cell-sojourn time also decides the frequency of a user's handoff before he completes his connection. Thus, as long as the average cell-sojourn times are the same, the average channel-holding time or handoff frequency should not differ much, even though the exact distributions are different. Since DR and UDF mainly depend on the channel-holding time and handoff frequency, DR or UDF will be similar even if different distributions of cell-sojourn time are used. This insensitivity to the distribution of cell-sojourn time implies that our

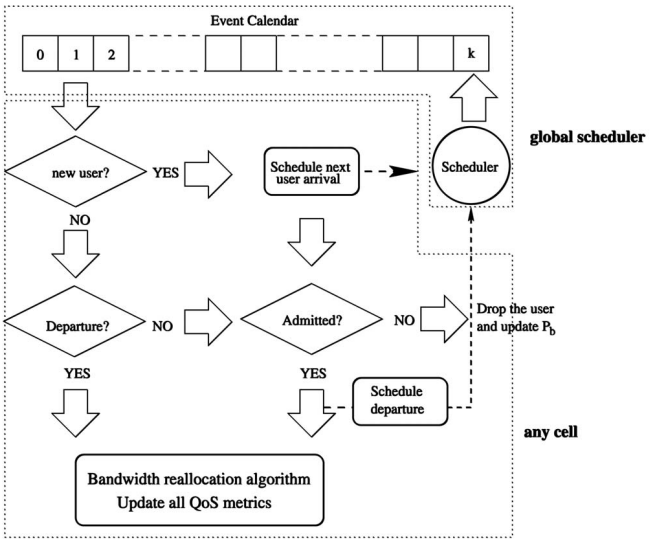

Fig. 17. The flow-chart of event-driven simulator.

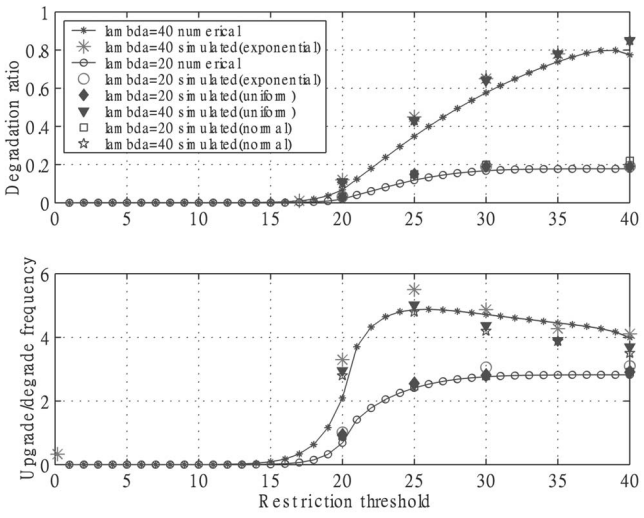

Fig. 18. DR and UDF under different mobility models.

analytical model can still be used to predict and control user-perceived QoS under different user mobility models.

6 Conclusions

In this paper, we derived an analytical model for a wireless network which uses adaptive bandwidth allocation to provide users multilevel QoS. Four performance metrics—blocking probability, forced-termination probability, degradation ratio, and upgrade/degrade frequency—are derived. Using numerical analysis, we investigated their dependency on user arrival rate, connection lifetime, and user mobility, and showed the trade off among these metrics. Moreover, we showed the importance of upgrade/degrade frequency to QoS provisioning, especially because of its strong dependency on user mobility and trade off with user fairness. Fair admission and bandwidth allocation algorithms are then provided such that low DR and UDF can be achieved with only a slight increase in the blocking probability of new users. Our simulation results indicate the applicability of our proposed model even if different mobility models are used. In summary, we have provided an analytical framework for predictive or adaptive bandwidth allocation algorithms. It can also determine an appropriate restriction threshold and a bandwidth allocation algorithm, according to the traffic characteristics or desired QoS criteria.

Acknowledgments

The work reported in this paper is supported in part by the US Air Force Office of Scientific Research under Grant No. F49620-00-1-0327. A subset of materials in this paper was presented at IEEE INFOCOM 2002.

References

[1] S. Sen, J. Jawanda, K. Basu, and S. Das, "Quality-of Service Degradation Strategies in Multimedia Wireless Network," *Proc. IEEE Vehicular Techonology Conf.,* vol. 3, pp. 1884-1888, May 1998.
[2] S. Singh, "Quality of Service Guarantees in Mobile Computing," *Computer Comm.,* no. 19, pp. 359-371, 1996.
[3] M.R. Sherif, I.W. Habib, M.N. Nagshineh, and P.K. Kermani, "Adaptive Allocation of Resources and Call Admission Control for Wireless ATM Using Generic Algorithm," *IEEE J. Selected Areas in Comm.,* vol. 18, no. 2, pp. 268-282, Feb. 2000.
[4] T. Kwon, Y. Choi, C. Bisdikian, and M. Naghshineh, "Call Admission Control for Adaptive Multimedia in Wireless/Mobile Network," *Proc. First ACM Int'l Workshop Wireless Mobile Multimedia,* pp. 111-116, Oct. 1998.
[5] S. Choi and K.G. Shin, "Location/Mobility-Dependent Bandwidth Adaptation in QoS-Sensitive Cellular Networks," *Proc. IEEE Vehicular Technology Conf.,* vol. 3, pp. 1593-1597, 2001.
[6] Y.B Lin, S. Mohan, and A. Noerpel, "Queueing Priority Channel Assignment Strategy for PCS Handoff and Initial Access," *IEEE Trans. Vehicular Technology,* vol. 43, no. 3, pp. 704-712, Aug. 1994.
[7] R. Ramjee, R. Nagarajan, and D. Towsley, "On Optimal Call Admission Control in Cellular Networks," *Proc. IEEE INFOCOM '96,* vol. 1, pp. 43-50, 1996.
[8] K. Mitchell and K. Sohraby, "An Analysis of the Effects of Mobility on Bandwidth Allocation Strategies in Multi-Class Cellular Wireless Networks," *Proc. IEEE INFOCOM '01,* vol. 2, pp. 1075-1084, 2001.
[9] A. Sutoving and J.M. Peha, "Novel Heuristic for Call Admission Control in Cellular Systems," *Proc. IEEE Int'l Conf. Universal Personal Comm.,* vol. 1, pp. 129-133, 1997.
[10] M. Naghshineh and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 14, no. 3, pp. 289-293, May 1994.
[11] S. Choi and K.G. Shin, "Predictive and Adaptive Reservation for Handoffs in QoS-Sensitive Cellular Networks," *Proc. ACM SIGCOMM '98,* pp. 155-166, 1998.
[12] A. Aljadhai and T. Znati, "A Framework for Call Admission Control and QoS Support in Wireless Environments," *Proc. IEEE INFOCOM '99,* vol. 3, pp. 1019-1026, 1999.
[13] Z. Liu, M.J. Karol, M.E. Zarki, and K.Y. Eng, "Channel Access and Interference Issues in Multi-Code DS-CDMA Wireless Packet (ATM) Networks," *Wireless Networks,* vol. 2, no. 3, pp. 173-193, 1996.
[14] J.C. Haartsen, "The Bluetooth Radio System," *IEEE Personal Comm.,* vol. 7, no. 1, pp. 28-36, Feb. 2000.
[15] C. Chao and W. Chen, "Connection Admission Control for Mobile Multiple-Class Personal Communications Networks," *IEEE J. Selected Areas in Comm.,* vol. 15, no. 8, pp. 1618-1626, 1997.
[16] V. Paxson and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling," *IEEE Trans. Networking,* vol. 3, no. 3, pp. 226-244, 1995.
[17] S.N. Subramanian and T. Le-Ngoc, "Traffic Modeling in a Multi-Media Environment," *Proc. IEEE CCECE/CCGEI '95,* pp. 838-841, 1995.
[18] P. Bremaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues.* New York: Springer, 1999.

Chun-Ting Chou received the BS and the MS degrees, both from the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, in 1995 and 1997, respectively. He is currently a PhD candidate at The University of Michigan, Ann Arbor. Since 2000, he has been a research assistant in the Real-Time Computing Laboratory in the Department of Electrical Engineering and Computer Science, The University of Michigan. His research interests include quality of service support in mobile networks, mobility management, TCP performance in mobile networks, MAC-layer resource allocation and scheduling, and spectrum agile communication systems.

Kang G. Shin received the BS degree in electronics engineering from Seoul National University, Seoul, Korea, in 1970, and both the MS and PhD degrees in electrical engineering from Cornell University, Ithaca, New York, in 1976 and 1978, respectively. From 1978 to 1982, he was on the faculty of Rensselaer Polytechnic Institute, Troy, New York. He has held visiting positions at the US Airforce Flight Dynamics Laboratory, AT&T Bell Laboratories, the Computer Science Division within the Department of Electrical Engineering and Computer Science at UC Berkeley, and International Computer Science Institute, Berkeley, California, IBM T. J. Watson Research Center, Software Engineering Institute at Carnegie Mellon University, and HP Research Laboratories. He also chaired the Computer Science and Engineering Division, EECS Department, The University of Michigan for three years beginning in January 1991. He is the Kev and Nancy O'Connor Professor of Computer Science and founding director of the Real-Time Computing Laboratory in the Department of Electrical Engineering and Computer Science. His current research focuses on QoS-sensitive networking and computing as well as on embedded real-time OS and middleware and applications, all with an emphasis on timeliness and dependability. He has supervised the completion of 46 PhD theses and authored/coauthored more then 500 technical papers and numerous book chapters in the areas of distributed real-time computing and control, computer networking, fault-tolerant computing, and intelligent manufacturing. He has coauthored (jointly with C.M. Krishna) a textbook, *Real-Time Systems* (McGraw Hill, 1997). He has received a number of best paper awards and has also coauthored papers with his students which received the Best Student Paper Awards. He has also received several institutional awards, including the Distinguished Faculty Achievement Award in 2001 from The University of Michigan and a Distinguished Alumni Award of the College of Engineering, Seoul National University in 2002. He is a fellow of IEEE and ACM, and a member of the Korean Academy of Engineering.