Analysis of Combined Adaptive Bandwidth Allocation and Admission Control in Wireless Networks

Chun-Ting Chou and Kang G. Shin

Real-Time Computing Laboratory
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109-2122, U.S.A.
{choujt, kgshin}@umich.edu

Abstract— **An analytical model is developed for cellular networks with a combined adaptive bandwidth allocation and traffic-restriction mechanism. Instead of focusing only on the bandwidth utilization and forced-termination probability, we derive two important Quality-of-Service (QoS) metrics,** *degradation ratio* **and** *upgrade/degrade frequency***. We show numerically that these two metrics must be taken into account in order to support the QoS specified by each client. The effects of system loads and clients' mobility on system performance are also investigated. Even under the various distributions of mobility, the simulation results are shown to match our analytical results, implying the applicability of our analytical model to more general cases.**

I. Introduction

With the proliferation of wireless personal devices such as laptops, PDAs, and mobile phones, the demand for wireless communications has grown exponentially over the last decade and is expected even more in the future. More and more multimedia data are being transmitted via wireless media, and such applications require diverse QoS. Due to the intrinsic scarcity of wireless bandwidth, it is challenging to provide diverse QoS while achieving high bandwidth utilization. For example, a system may allocate higher bandwidth for multimedia applications to satisfy their QoS at the expense of rejecting new calls that require less bandwidth. In order to enhance bandwidth utilization while satisfying the QoS of existing connections, numerous approaches have been proposed. A graceful degradation mechanism is proposed by Singh [1] to increase bandwidth utilization by adaptively adjusting bandwidth allocation according to the user-specified loss profiles. For most multimedia applications (e.g., voice, video telephony or video conferencing), service can be degraded in case of congestion as long as it is still within the pre-specified tolerable range. Take video telephony as an example: generic video telephony may require over 40 Kbps but low-motion video telephony requiring about 25 Kbps is acceptable [2]. Thus, a system could free some channels for new calls by lowering the QoS levels of ongoing calls. Sen *et al.* [2] proposed an optimal degradation strategy by maximizing a revenue function. Sherif *et al.* [3] proposed an adaptive resource allocation algorithm to maximize bandwidth utilization and tried 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 graceful QoS degradation. However, they did not provide any analysis for service degradation of individual calls, which is crucial to QoS provision. Kwon *et al.* [4] derived a *degradation period ratio* under the assumption that the degradation probability and mean degradation time are kept intact in all degradation states. However, we show that these metrics are dependent on the degradation state in which a given call resides, and hence, derive a new degradation ratio. Moreover, it is shown numerically that the degradation ratio does not suffice to reflect the QoS guarantees given to individual calls. Frequently switching among the different degradation levels may be even worse than a large degradation ratio [5]. So, we also derive a formula for switching QoS levels.

Another important issue in wireless communication is the forced-termination (or call dropping) probability. In case of shortage of bandwidth, hand-off calls may be dropped, thus compromising their QoS. In order to prevent ongoing calls from potential dropping/termination, Lin *et al.* [6] gave priority to hand-off calls over new calls, such that the forced-termination probability is improved without seriously degrading the blocking probability of new calls. Naghshineh *et al.* [7] proposed a distributed call admission control scheme by estimating the possible number of hand-off calls from adjacent cells. Various reservation-based admission control schemes (or so called *Guard Channels*) have also been proposed to reduce the probability of terminating ongoing or hand-off calls [8], [9]. Some optimal solutions subject to different constraints have also been proposed in [10], [11]. Slightly different from the reservation-based call admission control (CAC), once the system load exceeds a predefined threshold, we restrict the traffic of newly-initiated calls so as not to drop hand-off calls.

In this paper, we derive an analytical model for the combined graceful degradation and traffic restriction mechanism. This model is based on four QoS metrics: blocking probability, forced-termination probability, degradation ratio, and upgrade/degrade frequency. This study provides an analytical framework for predictive or adaptive bandwidth allocation algorithms [12], [13], and helps decide the operation region based on some desired criteria.

Our scheme can be built on various wireless architectures.

The work reported in this paper is supported in part by the US AFOSR under Grant No. F49620-00-1-0327

For a DS-CDMA system, the multi-code CDMA [14] can be used for service degrade/upgrade; for a FH-TDMA system (e.g., Bluetooth), service degrade/upgrade can be achieved by adequate assignment of time slots (i.e., polling policy) [15]. Resource allocation that considers channel deficits in the wireless media, is also related to our scheme, but it is beyond the scope of this paper. Interested readers may refer to [16], [17] for time-slot assignment and [12] for CDMA systems.

The rest of this paper is organized as follows. In Section II, the system environment and the assumptions used in this paper are introduced. Section III provides an analytical model for the proposed scheme, and the QoS metrics mentioned above are derived. The numerical analysis results based on the analytical model are presented in Section IV, while Section V discusses the simulation results. Finally, conclusions are drawn and direction of our future work is discussed in Section VI.

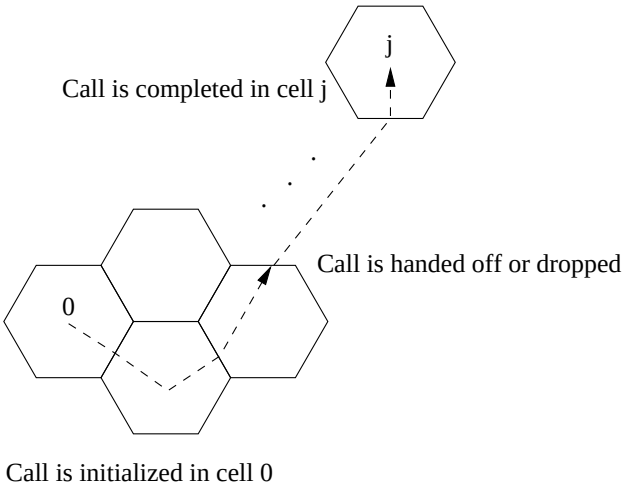

Fig. 1. A wireless cellular network

II. System Description And Assumptions

We consider a cellular network (Figure 1), in which a mobile communicates with others via a base station while residing in the cell of that base station. When a mobile leaves a cell, it could be either successfully handed off, or dropped in case of shortage of channels in the new cell. Since dropping hand-off calls is usually less desirable and less tolerable than blocking newly-initiated calls, hand-off calls are given priority over new calls. This is achieved by restricting new incoming calls into the system once the system load exceeds a certain threshold. Obviously, this threshold is a design parameter, and one of the objectives in this paper is to determine the proper value of this threshold. Moreover, we assume that each call could receive degraded service as long as this degraded service is within the user-specified QoS profile. Therefore, once the total required channels exceed the cell capacity (or the total available channels in that cell), the system may try to degrade the QoS of some existing calls in order to admit more (both new and hand-off) calls, hence reducing the blocking or forced-termination probability.

In this paper, we assume that the call arrival process is Poisson with the new call arrival rate λ_0, and the call-holding time is exponentially distributed with mean $\frac{1}{\mu_0}$. To evaluate the effects of user mobility on system performance, the call sojourn time, which is the time a call spends in a cell, is also taken into account and is assumed to be exponentially-distributed with mean $\frac{1}{\eta}$ as in [10], [11], [18] for mathematical tractibility. However, we will show by simulation in Section V that the formulations for QoS metrics derived under this model are still valid even using different mobility distributions.

Under these assumptions, the hand-off rate can be derived as in [6]:

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

where p_f is the forced-termination probability of hand-off calls and p_b is the blocking probability of new calls. The channel-occupancy time of an admitted call in a cell is the minimum of the remaining call-holding time and call sojourn time. Since we assume that both call-holding time and call sojourn time are exponentially-distributed, the distribution of channel occupancy time is

$$f_{c0} = (\mu_0 + \eta)e^{-(\mu_0+\eta)t}. \tag{2}$$

Under this degradation scheme, both call blocking and forced-termination probabilities are improved. However, some calls may receive severely degraded service. In the following section, we investigate the tradeoff among the QoS metrics, especially between the call blocking probability and the other three QoS metrics.

III. System Analysis

In the analytical model described below, we assume that each call receives either full or degraded service, depending on the system load at the time of its arrival. To simplify the analysis without loss of generality, we also assume that the number of channels required by full service is twice the number of channels required by degraded service (degraded service only requires one unit of channel). If a call can be admitted but there are not enough idle channels for full service, one of the existing full-service calls is randomly chosen to be degraded and the released channel is allocated to the new call. On the other hand, the released channels of a departing call are randomly reallocated to the ongoing calls that receive the degraded service. A generalization for a multi-level degradable service is also given at the end of this section.

A. Stationary distribution of the number of calls in a cell

The number of calls in a cell equipped with N channels can be simply modeled as a one-dimensional Markov chain $X_t = (m, n)$ as shown in Figure 2, where m is the number of calls receiving full service and n is the number of degraded calls. However, due to different admission policies (i.e., restriction thresholds), the effective call arrival rate which results in state transitions may vary. If new calls are not differentiated from hand-off calls (i.e., hand-off calls from adjacent cells are regarded as newly-initiated calls in this cell), the stationary distribution of the number of calls in a cell can be obtained by Erlang's formula by setting the arrival rate λ_i to $\lambda_0 + \lambda_h$

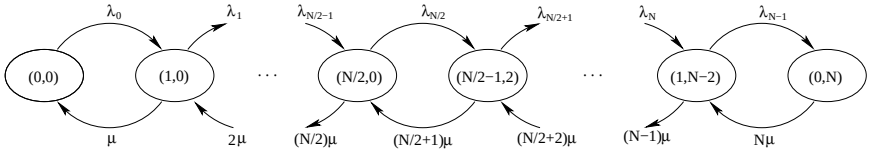

Fig. 2. State transitions of the number of calls in one cell

(the new call arrival rate plus hand-off rate) and service rate μ_i to $i \cdot (\mu_0 + \eta)$ (as suggested in Section II). If the traffic restriction is applied to the new calls and the restriction threshold is set at state $(m', n' = 0)$ (which may result in a higher blocking probability and under-utilization of resources), then $\lambda_i = \lambda_0 + \lambda_h$ for $i < m'$ and $\lambda_i = \lambda_h$ for other states. If the restriction threshold is set at state $(m', n' \neq 0)$ (which may result in a higher forced-termination probability and severely degraded service), then $\lambda_i = \lambda_0 + \lambda_h$ for $i < m' + n'$ and $\lambda_i = \lambda_h$ for $i \geq m' + n'$. In these cases, the stationary distribution can still be obtained as a general Erlang's formula with variable arrival rates. The stationary distribution is given as:

$$\pi_{m,n} = \frac{1}{\sum_{i=0}^{N} \frac{\prod_{k=0}^{i-1} \lambda_k}{\mu^i i!}} \times \frac{\prod_{k=0}^{m+n-1} \lambda_k}{\mu^{m+n}(m+n)!}, \qquad (3)$$

where $\mu = \mu_0 + \eta$. In either case, the blocking probability p_b is $\sum_{i+j=m'+n'}^{N} \pi_{i,j}$, and the forced-termination probability p_f is $\pi_{0,N}$, which can be obtained from Eq. (3).

Thanks to the assumptions of homogeneous cells, Poisson arrival process and exponential channel occupancy 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 calls observed at the time of each call's arrival.

B. QoS metrics

In a system with degradable service, a call may receive full or degraded service, depending on the system load at the time of its arrival (this probability is given in the previous subsection). Even if a call receives full service upon its arrival, it can be degraded when the system tries to accept more calls. From the users' perspectives, this may raise two important questions: (1) how long does a call receive full service or degraded service?, and (2) how often does the QoS level switch? Even though these two questions may be inter-related, the first question does not necessarily imply the second, or vice versa. Therefore, the QoS metrics associated with these two questions, degradation ratio and upgrade/degrade frequency, are defined as follows.

- *Degradation ratio* (DR): the ratio of the time a call receives degraded service to the total channel occupancy time in each cell.
- *Upgrade/degrade frequency* (UDF): the frequency of switching between full and degraded service by an admitted call.

In order to analytically derive these two QoS metrics, we build a discrete-time Markov chain $Y_t = (m, n)$ that models the evolution of any arbitrary call (C) which is admitted to a cell, where m is the number of full-service calls and n is the number of degraded-service calls observed at the time of a new call arrival or the time of a call departure when call C is in the system. However, the call C may receive full service or degraded service, so we should distinguish these situations as follows.

1) If call C arrives and the system is fully-occupied (e.g., $X_t = (\frac{N}{2} - i, 2i), i = 0, 1, \ldots, \frac{N}{2} - 1$), it will receive degraded service and we denote $Y_t = d_i(\frac{N}{2} - i - 1, 2i + 2)$ ('d' for 'degraded service'). On the other hand, if call C receives degraded service (e.g., $Y_t = d_i(\frac{N}{2} - i, 2i), i = 1, \ldots, \frac{N}{2}$) but it is upgraded due to the departure of any other existing calls, then $Y_t = f_i(\frac{N}{2} - i + 1, 2i - 2)$ ('f' for 'full service'). Since call C always receives full-service when $X_t = (j, 0), j = 0, 1, \ldots, \frac{N}{2} - 1$, we denote $Y_t = f_j(j, 0), j = 1, 2, \ldots, \frac{N}{2}$ for these cases.
2) Completion state (A): either the hand-off or completion of servicing call C will lead to the completion state. Once call C enters this state, it returns the allocated channels back to the system (e.g., it is leaving the cell). Obviously, this is an absorption state.

The resulting embedded Markov chain is shown in Figure 3 with the transition probabilities described below.

Consider the admitted call C in any state. Three different events may occur: arrival of a new call, departure of call C or departure of any other existing calls. We need to differentiate several situations in order to calculate the transition probabilities as follows.

- For state f_i $(m = i, n = 0)$, all existing calls receive full service. Three transition probabilities in these states are $P_{f_i,f_{i+1}} = \frac{\lambda_i}{\lambda_i + i\mu}$, $P_{f_i,A} = \frac{\mu}{\lambda_i + i\mu}$ and $P_{f_i,f_{i-1}} = \frac{(i-1)\mu}{\lambda_i + i\mu}$ for $1 \leq i \leq \frac{N}{2} - 1$.
- For state f_i $(m = N - i, n = 2i - N)$ where $\frac{N}{2} \leq i \leq N - 1$, an arrival of a new call may result in two different transitions. One is that call C is degraded such that the state transits to degraded state $d_{i-\frac{N}{2}+1}$. The other is that C is not degraded so that the state transits to f_{i+1}. The associated transition probabilities are $P_{f_i,d_{i-\frac{N}{2}+1}} = \frac{\lambda_i}{(N-i)(\lambda_i + i\mu)}$ and $P_{f_i,f_{i+1}} = \frac{(N-i-1)\lambda_i}{(N-i)(\lambda_i + i\mu)}$, respectively. The other transition probabilities are $P_{f_i,A} = \frac{\mu}{(\lambda_i + i\mu)}$ and $P_{f_i,f_{i-1}} = \frac{(i-1)\mu}{(\lambda_i + i\mu)}$.
- For state d_i $(m = N' - i, n = 2i)$ where $1 \leq i \leq N' = \frac{N}{2}$, the departure of any other calls may result in two different transitions. One is that C is upgraded because of the others' departure such that the state transits to $f_{i+N'-1}$. The other is that C continues receiving degraded service and the state transits to d_{i-1}. The associated transition probabilities are $P_{d_i,f_{i+N'-1}} = \frac{N'}{i} \frac{\mu}{\lambda_{i+N'} + (N'+i)\mu}$ and $P_{d_i,d_{i-1}} = (1 - \frac{1}{i})(N' + i) \frac{\mu}{\lambda_{i+N'} + (N'+i)\mu}$. The other transition probabilities are $P_{d_i,d_{i+1}} = \frac{\lambda_{i+N'}}{\lambda_{i+N'} + (N'+i)\mu}$ and

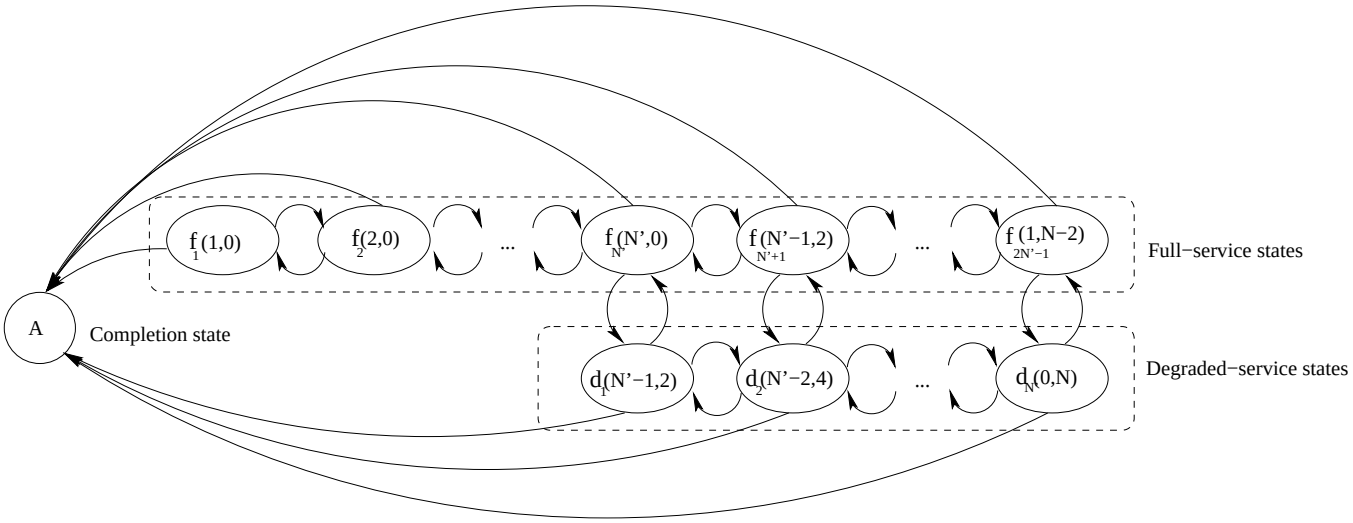

Fig. 3. State transitions of a call admitted into any cell

$P_{d_i,A} = \frac{\mu}{[\lambda_{i+N'}+(N'+i)\mu]}$.

- Note that $\lambda_N = 0$.

C. Degradation ratio

We now derive the DR based on the modified Markov chain shown in Figure 3. First, we need to derive N_j, the number of visits to state j before entering the completion state A, given that the initial state is i:

$$E_i(N_j) = E_i[\sum_{n=0}^{\infty} 1_{\{Y_n=j\}}] = \sum_{n=0}^{\infty} P_{ij}(n), \tag{4}$$

where Y_n is the state after the n-th transition and $P_{ij}(n)$ is the n-step transition probability from state i to state j. The $\sum_{n=0}^{\infty} P_{ij}(n)$ is also the (i,j)-th element of potential matrix G, which can be obtained by the following equation:

$$G = \sum_{n=0}^{\infty} P^n. \tag{5}$$

P is the transition matrix of the modified Markov chain shown in Figure 3 and can be written as

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

where T_T is the restriction of P to the transient set $T = \{f_1, f_2, \ldots, f_{N-1}, d_1, d_2, \ldots, d_{N'}\}$ and how to find these elements is treated in the previous section. Since we only consider the number of visits to the transient states before entering the completion state A, the potential matrix can be rewritten as

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

where $S = \sum_{n=0}^{\infty} T_T^n$ and $E_i(N_j)$ is just the (i,j)-th element of matrix S. By matrix manipulation, S can be computed by the following equation [19],

$$\mathbf{S} = (\mathbf{I} - \mathbf{T_T})^{-1}. \tag{6}$$

Next, we define the expected degradation time given that the initial state is i

$$T_{d,i} = \sum_{d_j \in \{degraded\ class\}} E_i(N_{d_j}) T_{sojourn,d_j}. \tag{7}$$

$T_{sojourn,d_j} = \frac{1}{\lambda_{j+N'}+(j+N')*\mu}$ is the mean sojourn time in state d_j for $1 \le j \le N'$. Then, the degradation ratio can be computed as

$$\text{DR} = \sum_{i=0}^{N'-1} \mu\pi_{i,0} T_{d,i+1} + \sum_{i=N'}^{N-1} \mu\pi_{N-i,2i-N} T_{d,i-N'+1}, \tag{8}$$

where $\pi_{m,n}$ is given in Eq. (3).

D. Upgrade/degrade frequency

Let's consider how to derive UDF. As shown in Figure 3, there are two levels of service a call may receive: full or degraded service. The QoS metric of interest is the average number of times the QoS level changes per unit time between these two service levels:

$$UDF = \frac{N_{full \to degraded} + N_{degraded \to full}}{mean\ occupancy\ time\ in\ a\ cell}.$$

We use the first-step analysis to compute this metric as follows. Let D_i be the number of switches between two service levels, C_f and C_d, given that the initial state is i. (Note that transition $C_f \to C_d$ is service degradation and transition $C_d \to C_f$ is service upgrade.) By the first-step analysis, the following system

of linear equations can be obtained:

$$\begin{cases} E(D_{f_1}) &= P_{f_1,f_2}E(D_{f_2}) \\ E(D_{f_i}) &= P_{f_i,f_{i-1}}E(D_{f_{i-1}}) + P_{f_i,f_{i+1}}E(D_{f_{i+1}}) \\ & \text{for } i\text{=2,3,}\ldots\text{,}N'-1 \\ E(D_{f_i}) &= P_{f_i,f_{i-1}}E(D_{f_{i-1}}) + P_{f_i,f_{i+1}}E(D_{f_{i+1}} \\ & +P_{f_i,d_{i-N'+1}}[1+E(D_{d_{i-N'+1}})] \\ & \text{for } i\text{=}N'\text{,....,}N-1 \\ E(D_{d_1}) &= P_{d_1,f_{N'}}[1+E(D_{f_{N'}})] + P_{d1,d2}E(D_{d_2}) \\ E(D_{d_i}) &= P_{d_i,f_{i+N'-1}}[1+E(D_{f_{i+N'-1}})] \\ & +P_{d_i,d_{i-1}}E(D_{d_{i-1}}) + P_{d_i,d_{i+1}}E(D_{d_{i+1}}) \\ & \text{for } i\text{=2,...}N'-1 \\ E(D_{d_{N'}}) &= P_{d_{N'},f_{2N'-1}}[1+E(D_{f_{2N'-1}})] \\ & +P_{d_{N'},d_{N'-1}}E(D_{d_{N'-1}}) \end{cases}$$

The solution to this system of linear equations can be computed as

$$\mathbf{E} = (\mathbf{I} - \mathbf{T_T})^{-1}\mathbf{C}, \tag{9}$$

where $\mathbf{C}$ is the column vector with the i-th element equal to $P_{f_i,d_{i-N'+1}}$ for $1 \le i \le N-1$ or $P_{d_{i-N},f_{i-N'-1}}$ for $N+1 \le i \le \frac{3}{2}N$. By using Eq. (6), the matrix $\mathbf{E}$ can be rewritten as

$$\mathbf{E} = \mathbf{SC}. \tag{10}$$

UDF can then be obtained as:

$$\text{UDF} = \sum_{i=0}^{N'-1} \mu\pi_{i,0}E(D_{f_{i+1}}) + \sum_{i=N'}^{N-1} \mu\pi_{N-i,2i-N}E(D_{d_{i-N'+1}}). \tag{11}$$

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

$$DR_I = \sum_{i=0}^{min(j,N'-1)} \mu\pi_{i,0}T_{d,i+1} + \sum_{i=min(j,N')}^{j-1} \mu\pi_{N-i,2i-N}T_{d,i-N'+1}$$

$$UDF_I = \sum_{i=0}^{min(j,N'-1)} \mu\pi_{i,0}E(D_{f_{i+1}}) + \sum_{i=min(j,N')}^{j-1} \mu\pi_{N-i,2i-N}E(D_{d_{i-N'+1}}),$$

where j is the restriction threshold, and DR_I and UDF_I are the QoS metrics for a call in the cell where the call was initiated.

E. Generalization for multi-level degradable service

In this subsection, we consider multi-level service with minimum bandwidth requirement, $W_1 = W_{min}$ channels, and maximum bandwidth requirement, $W_K = W_{max}$ channels (full-service). Any amount of bandwidth allocation between them is deemed acceptable. The state should be modified as $Y_t = L^{(i)}(n_1, n_2, \ldots, n_K)$, where n_k is the number of the level-k calls when call C is receiving the level-i service. The single-step transition matrix $\mathbf{P}$ (or more precisely, $\mathbf{T_T}$, the restriction of $\mathbf{P}$ to the transient set) can be obtained as in the previous subsection according to degrade/upgrade algorithm described below. The equations for the QoS metrics in the previous subsections can be directly applied once $\mathbf{T_T}$ is calculated.

Let W_a be the number of available channels, N_i be the number of calls with i units of channels, where $W_{min} \le i \le W_{max}$ and N_T be the total number of existing calls in the system at the time of a call's arrival. The degradation algorithm is presented in Figure 4.

```
01. if (W_a ≥ W_min)
02.    W_allocated = min(W_max, W_a)
03.  elseif (W_a ≤ W_min & (N − N_T * W_min) ≥ W_min){
04.    W_allocated = 0.
05.    for (i = W_max, i > W_min, i − −)
05.       while (W_allocated < W_min & N_i > 0) {
06.          Randomly degrade one of the N_i calls
             by one unit of channel.
07.          N_i = N_i − 1;
08.          N_{i−1} = N_{i−1} + 1;
09.          W_allocated = W_allocated + 1; }}
10. else
11.    Reject the call request.
```

Fig. 4. A pseudo-code of the bandwidth degradation algorithm

Allocating only W_{min} units of channels when there is a shortage of bandwidth, minimizes the need to switch the QoS levels of the existing calls, and hence, a smaller UDF can be achieved without compromising the DR. Fairness is also considered by randomly choosing the calls to be degraded. The corresponding upgrade algorithm is described as follows. Let W_r be the channels that the departing call (either handed off to an adjacent cell, or completion of service) returns to the cell. The released channels are randomly reallocated to the ongoing calls as shown in Figure 5.

```
01. for (i = W_min, i < W_max, i + +);
02.    while (W_r > 0 & N_i > 0) {
03.       Randomly upgrade one of the N_i calls
          by one unit of channel.
04.       N_i = N_i − 1.
05.       N_{i+1} = N_{i+1} + 1.
06.       W_r = W_r − 1.
07.    }
```

Fig. 5. A pseudo-code of the bandwidth upgrade algorithm

Since the system supports the multi-level service, instead of using DR as one of the QoS metrics, a weighted DR should be used, as there are $(W_{max} - W_{min} + 1)$ QoS levels. The resulting DR can still be obtained by Eq. (8) with the weighted

degradation time

$$T'_{d,i} = \sum_{d_j} \frac{W_{max} - W_{d_j}}{W_{max}} E_i(N_{d_j}) T_{sojourn,d_j}, \qquad (12)$$

where d_j is a degraded QoS level and W_j is the allocated bandwidth (in units of channels) in that degraded QoS level.

IV. Numerical Results

We consider a cellular network, in which each cell has 40 units of channels. The arrival process of new calls is assumed to be Poisson, and the call-holding and call-sojourn times are exponentially-distributed. The formula for the resulting hand-off rate and channel-occupancy time can be found in Eqs. (1) and (2). For illustrative purposes, it is assumed that each full service requires 2 units of channels and each degraded service requires only 1 unit of channel. As we pointed out in the previous section, this model can be applied to any degradable service with bandwidth allocation between W_{min} and W_{max}.

Four QoS metrics — blocking probability of new calls (P_b), forced-termination probability of hand-off calls (P_f), degradation ratio (DR) and upgrade/degrade frequency (UDF) — are evaluated. Since the call-arrival rate, call-holding time, and mobility ($= \frac{1}{\eta}$) of each call could significantly affect these metrics, three sets of numerical results are shown for these factors under various settings of the restriction threshold. The restriction threshold (defined as $m' + n'$ in Section III) 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 as shown in Figure 2, and at most one newly-initiated call could be admitted into the system (e.g., most calls in cells are hand-off calls from the adjacent cells). On the other hand, if the restriction threshold is 40, no channel is reserved for hand-off calls, and there is no distinction between new and hand-off calls. Selection of the restriction threshold under different traffic loads is also discussed at the end of this section.

A. QoS metrics vs. call arrival rate

Figure 6 plots P_b and P_f under four call-arrival rates: $\lambda =$ 20, 30, 40, 50 calls per unit time. The tradeoff between P_b and P_f is obvious under different restriction thresholds. In the case of light traffic ($\lambda = 20$) with a high restriction threshold, P_b and P_f are negligible. Even in the case of heavy loads ($\lambda = 50$), both P_b and P_f are still only 0.13 and 0.18, respectively (compared to 0.45 without any degradation and traffic restriction).

Figure 7, however, shows that the decrease of P_f and P_b by the degradation scheme results in severe service degradation of individual calls. DR increases with the restriction threshold under different loads and is higher than 0.8 in the case of high loads and high restriction thresholds. UDF increases more quickly than DR as the restriction threshold increases. Even when the system reserves 40% of channels for hand-off calls, UDF is still as high as 5 in the case of moderate traffic load. A drop in UDF can also be observed in case of high loads and high restrictions, because there is a sharp increase of P_f, and consequently the hand-off rate may significantly decrease.

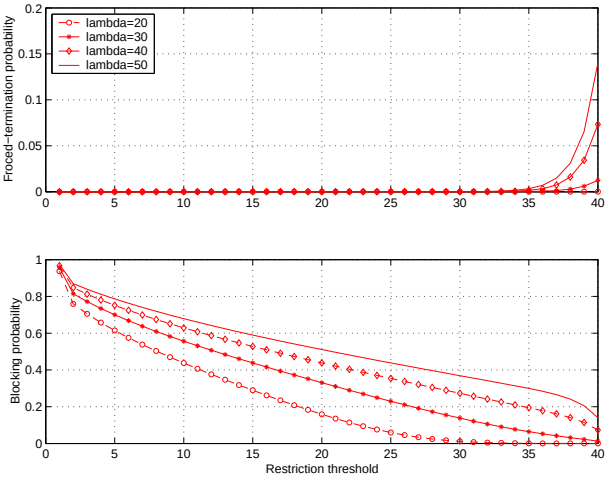

Fig. 6. P_b and P_f vs. call-arrival rate

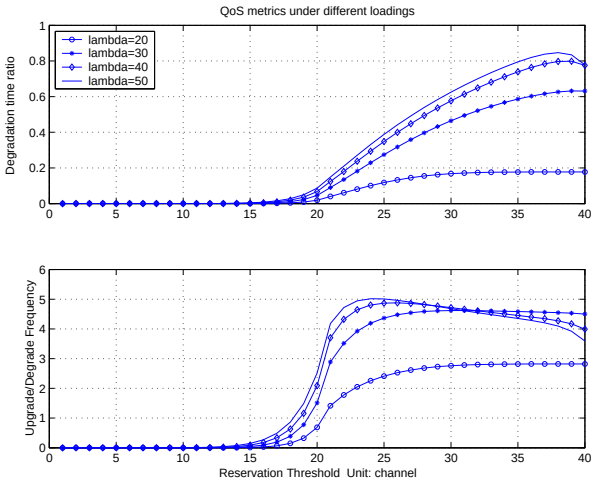

Fig. 7. DR and UDF vs. call-arrival rate

B. QoS metrics vs. call-holding time

Figure 8 shows P_b and P_f under four different call-holding times: $\frac{1}{\mu} = 8$, 4, 2, and 1 unit of time. In this case, the call-arrival rate is 20 calls/unit of time. P_b is much more sensitive to call-holding time than P_f. When the restriction threshold is high (e.g., 35), the blocking probability is still large (e.g., 0.5 in case of $\mu = 0.25$). But we still could simultaneously achieve low probabilities with the help of service degradation, even in the case of a larger call-holding time.

DR and UDF under the four call-holding times are plotted in Figure 9. In the case of a larger call-holding time, both QoS metrics show a drop when the threshold is high, because of the sharp increase in the forced-termination probability as shown in Figure 8. However, unlike DR, UDF tends to decrease with the increase of call-holding time. In the case of a higher restriction threshold (e.g., 35), the UDF value when $\mu = \frac{1}{8}$ is half of that when $\mu = \frac{1}{2}$. However, the UDF is not only dependent on μ but also on the threshold as shown in Figure 9. When the threshold is high and the call-holding time is longer, the service switching due to the departures of other calls is lessened and thus, the UDF decreases with the increase of call-holding time. However, when the threshold is low (more new calls are

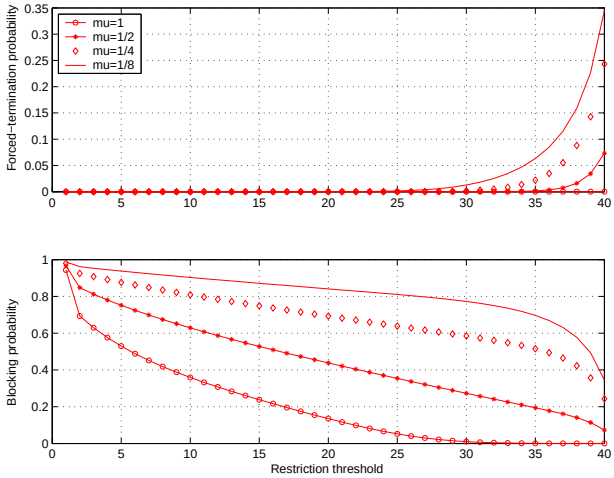

Fig. 8. P_b and P_f vs. call-holding time

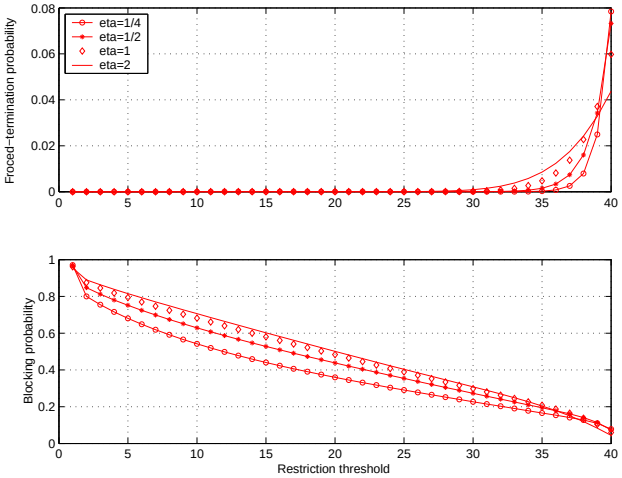

Fig. 10. Pb and Pf vs. mobility

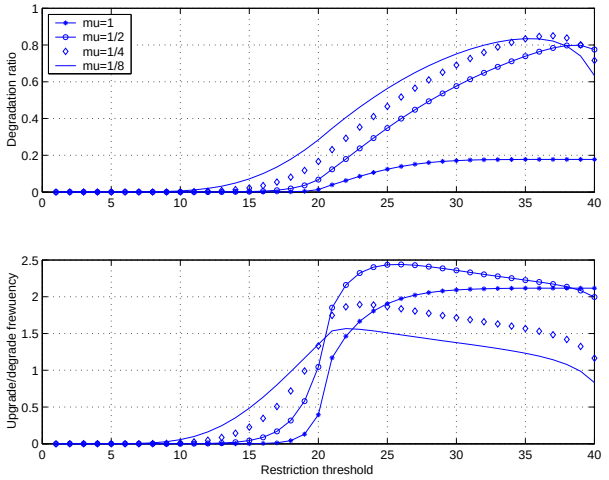

Fig. 9. DR and UDF vs. call-holding time

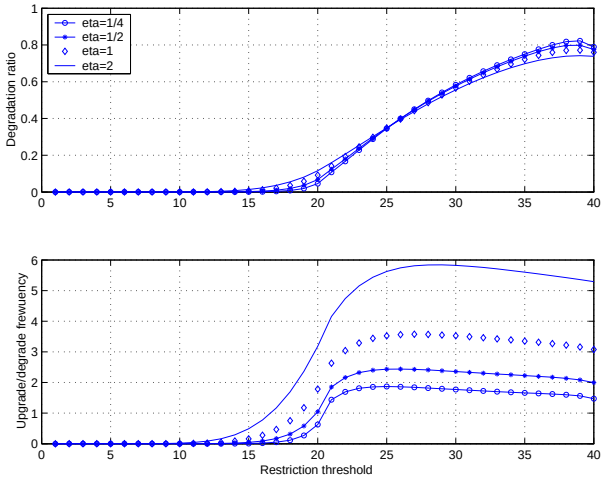

Fig. 11. DR and UDF vs. mobility

blocked) and the call-holding time is shorter, the total traffic load is smaller (note that λ is fixed in this subsection), and thus, most calls would not interfere with one another, which results in a smaller UDF. This explains the crossover of UDF under different μ's when the threshold increases. These different dependencies on call-holding time also justify the need for considering both metrics.

C. QoS metrics vs. mobility

Figure 10 shows P_b and P_f under four different call-sojourn times: $\frac{1}{\eta} = 0.5$, 1, 2, and 4 units of time. In all cases, P_b and P_f only slightly increase with mobility. Even in case of higher mobility, both P_b and P_f can be as low as 0.1 or less with the help of a high restriction threshold and service degradation.

DR and UDF are plotted in Figure 11, and these two metrics exhibit inverse dependence on mobility. DR remains almost the same under the different cases of mobility. However, UDF can be three times larger in the case of higher mobility than in the case of lower mobility (e.g., UDF≈ 6 when $\eta = 2$, but UDF≈ 2 when $\eta = \frac{1}{4}$, in the case of threshold=27). The reason for this is that high mobility results in frequent switches between different QoS levels, but the amount of time a call resides in each level is statistically the same. Therefore, we should consider both DR and UDF for QoS provision. In the case of higher mobility, UDF is the dominant factor of QoS for individual calls.

D. System operation region

There is an obvious tradeoff between the blocking probability of new calls and the other QoS metrics under the proposed degradation and restriction scheme. Therefore, there does not exist an absolutely optimal operation point in terms of all of the four parameters. Since the forced-termination probability rises sharply only when the restriction threshold is close to the system capacity, the possible choice of restriction threshold should be between $\frac{N}{2}$ and N. If we only consider the blocking probability and forced-termination probability, the optimal operation region should be very close to system capacity (e.g., the threshold is 37 or 38 as shown in Figure 8). However, DR has a maximal value (≈ 0.8 in Figure 9), meaning that calls are severely degraded. If we choose the threshold ≈ 25, DR can be significantly improved (from 0.8 to 0.4) with only a slight increase of P_f by 0.12 (P_b is negligible and UDF is almost the same). This means that admitted calls could receive much better service at the expense of blocking only 12% more calls. The same

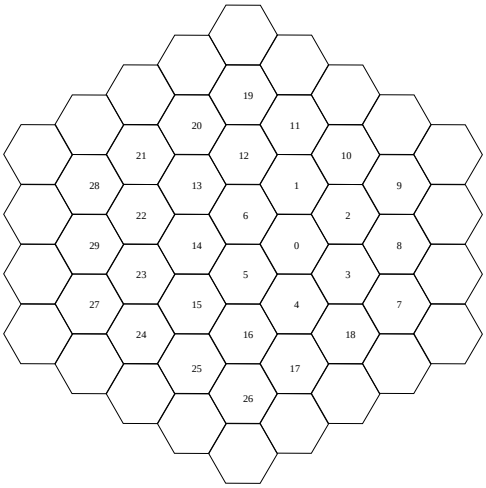

Fig. 12. The cellular network used in simulation

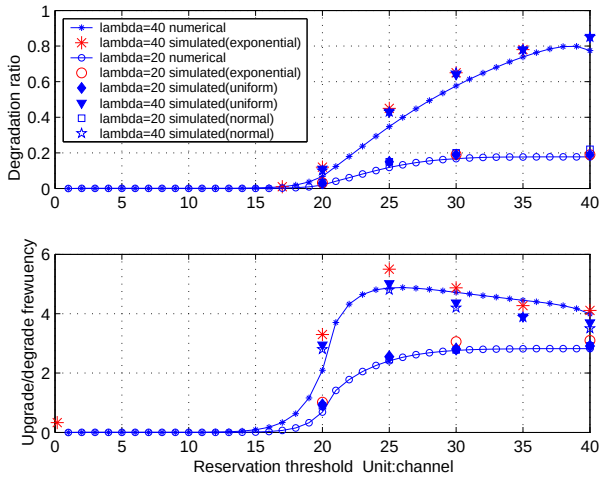

Fig. 13. DR and UDF under different mobility models

conclusion can be drawn from the results in Figures 10 and 11. Both DR and UDF decrease significantly (DR decreases from 0.6 to 0.1 in all cases, and 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 less than 0.2 in most cases, if we set the threshold close to one half of the system capacity, instead of setting to the higher values. We show that if only P_b and P_f are considered, even though we can simultaneous achieve low P_b and P_f, each call endures severely degraded service and frequent switching of service levels. By considering both DR and UDF, each call can receive much better QoS (much smaller DR and much less service switchings) without sacrificing P_f much.

As the numerical results shown in the previous subsection, the choice of operation point may also vary under different traffic loads and mobility. For example, if customers have longer call-holding times, the operation point may be chosen to be close to the system capacity. On the other hand, if the mobility of customers is high, the operation point may be chosen to be close to one half of the system capacity such that UDF is acceptable, as suggested in the set of the third numerical results.

V. Simulation

A cellular network of 30 cells is used in our simulation. As shown in Figure 12, the statistics of boundary cells (e.g., cells 7, 8, 9, 20) are not taken into account in the comparison with the numerical analysis in the previous section. The call-arrival process is still Poisson, call-holding time is exponentially-distributed but the assumption of exponentially-distributed call sojourn times is relaxed since the stochastic model for mobility may still be arguable. For comparative purposes, we assume that each cell has 40 units of channels. Both heavy-load (40 calls per unit of time) and light-load (20 calls per unit of time) cases are considered. Three distributions of the call-sojourn time — exponential, uniform, and normal distributions — are considered with mean of 1 unit of time and variance of 1 (except for the case of uniform distribution).

The simulation results are plotted in Figure 13. Both DR and UDF are plotted with the numerical results in the previous section (solid lines). In both cases, most of the simulation results are close to the numerical results (the largest error of DR is about 15% when the arrival rate is 40 and the threshold is 25, and the largest error of UDF is 18% when the arrival rate is 40 and the threshold is 20). A reason for this is that the number of cells is not infinite, and thus, the effect of the boundary cells introduces the error. However, it is surprising to see the phenomenon that, even the distribution of call-sojourn time is uniformly- or normally- distributed, the results are still consistent with the proposed analytical model. We conjecture that the assumption of independent call-sojourn times in each cell may possibly contribute to this result. Moreover, the insensitivity of P_b, P_f and DR to different mobility values (as shown in Figures 10 and 11) could also explain the independence of performance metrics (except UDF) from mobility distributions. This insensitivity to the distribution of mobility implies the applicability of our model to more general cases.

VI. Conclusions

In this paper, we derived an analytical model for wireless networks with adaptive bandwidth allocation and traffic-restriction CAC. Four QoS metrics — blocking probability, forced-termination probability, degradation ratio, and upgrade/degrade frequency — are derived, and these formulas can be directly applied to the case of multi-level QoS degradation. Moreover, this study provides the analytical framework for predictive or adaptive bandwidth allocation algorithms and helps decide the operation point under different traffic conditions. Using numerical analysis, we show the effects of call-arrival rate, call-holding time, and mobility of users on these QoS metrics and the importance of upgrade/degrade frequency to QoS provision, especially with consideration of mobility. Our simulation results indicate the applicability of our proposed model to the general cases with different mobility models. With this model, more complicated adaptive bandwidth allocation schemes can be analyzed, and their impacts on QoS can also be evaluated, which are matters of our future work.

References

[1] S. Singh, “Quality of Service guarantees in Mobile Computing”, *Computer Communications*, no.19, 1996, pp. 359-371.

[2] S. Sen, J. Jawanda, K. Basu, and S. Das,"Quality-of Service Degradation Strategies in Multimedia Wireless Network", *IEEE Vehicular Techonology Conference*, Vol.3, May 1998, pp. 1884-1888.

[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 Journal of Selected Areas in Communications*, Vol.18, no.2, Feb. 2000, pp. 268-282.

[4] T. Kwon, Y. Choi, C. Bisdikian, and M. Naghshineh, "Call Admission Control for Adaptive Multimedia in Wireless/Mobile Network", *Proceedings of first ACM international workshop on Wireless mobile multimedia*, Oct. 1998, pp. 111-116.

[5] S. Choi and K. G. Shin, "Location/Mobility-Dependent Bandwidth Adaptation in QoS-Sensitive Cellular Networks", *IEEE Vehicular Technology Conference*, 2001 (in press).

[6] Y.B Lin, S. Mohan, and A. Noerpel, "Queueing Priority Channel Assignment Strategy for PCS Hand-off and Initial Access", *IEEE Transaction on Vehicular Technology*, Vol.43, no.3, Aug. 1994, pp. 704-712.

[7] M. Naghshineh, and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks", *IEEE Journal of Selected Areas in Communications*, Vol.14, no.3, May 1994, pp. 289-293.

[8] W. Lee, and B. Sabata, "Admission Control and QoS Negotiations for Soft-Real Time Applications", *IEEE International Conference on Multimedia Computing and Systems*, Vol.1, 1999, pp. 147-152.

[9] A. Sutoving, and J.M. Peha, "Novel Heuristic for Call Admission Control in Cellular Systems", *IEEE International Conference on Universal Personal Communications*, Vol.1, 1997, pp. 129-133.

[10] R. Ramjee, R. Nagarajan, and D. Towsley, "On Optimal Call Admission Control in Cellular Networks", *IEEE INFOCOM '96*, Vol.1, pp. 43-50.

[11] K. Mitchell, and K. Sohraby, "An Analysis of the Effects of Mobility on Bandwidth Allocation Strategies in Multi-Class Cellular Wireless Networks", *IEEE INFOCOM '01*, Vol.2, pp. 1075-1084.

[12] S. Choi, and K. G. Shin, "Predictive and Adaptive Reservation for Handoffs in QoS-Sensitive Cellular Networks", *Proceedings of ACM SIGCOMM '98*, pp. 155-166.

[13] A. Aljadhai, and T. Znati, "A Framework for Call Admission Control and QoS Support in Wireless Environments", *IEEE INFOCOM '99*, Vol.3, pp. 1019-1026.

[14] 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, 1996.

[15] J.C. Haartsen,"The Bluetooth Radio System", *IEEE Personal Communications*, Vol.7, no.1, Feb. 2000, pp. 28-36.

[16] C. Fragouli, V. Sivaraman, and M.B. Srivastava,"Controlled multimedia wireless link sharing via enhanced class-based queuing with channel-state-dependent packet scheduling", *IEEE INFOCOM '98*, Vol.2, pp. 572-580.

[17] D.A. Eckhardt, and P. Steenkiste,"Effort-limited Fair(ELF) Scheduling for Wireless Networks", *IEEE INFOCOM '00*, Vol.3, pp. 1097-1106.

[18] C. Chao, and W. Chen, "Connection Admission Control for Mobile Multiple-Class Personal Communications Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 15, no.8, 1997, pp. 1618-1626.

[19] P. Bremaud, "Markov chains : Gibbs fields, Monte Carlo simulation and queues", Springer, New York, 1999.