Wireless Networks 7, 107–116, 2001

An Efficient and Adaptive Bandwidth Allocation Scheme for Mobile Wireless Networks Using an On-Line Local Estimation Technique *

BO LI ** and LI YIN
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

K.Y. MICHAEL WONG
Department of Physics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

SI WU
Laboratory for Information Synthesis, RIKEN Brain Science Institute, Hirosawa 2-1, Wako-shi, Saitama 351-01, Japan

Abstract. The next generation of mobile wireless networks has to provide the quality-of-service (QoS) for a variety of applications. One of the key generic QoS parameters is the call dropping probability, which has to be maintained at a predefined level independent of the traffic condition. In the presence of bursty data and the emerging multimedia traffic, an adaptive and dynamic bandwidth allocation is essential in ensuring this QoS. The paradox, however, is that all existing dynamic bandwidth allocation schemes require the prior knowledge of all traffic parameters or/and user mobility parameters. In addition, most proposals require extensive status information exchange among cells in order to dynamically readjust the control parameters, thus making them difficult to be used in actual deployment.

In this paper, we introduce a novel adaptive bandwidth allocation scheme which estimates dynamically the changing traffic parameters through *local on-line estimation*. Such estimations are restricted to each individual cell, thus completely eliminating the signaling overhead for information exchange among cells. Furthermore, we propose the use of a *probabilistic control policy*, which achieves a high channel utilization, and leads to an effective and stable control. Through simulations, we show that our proposed adaptive bandwidth allocation scheme can guarantee the predetermined call dropping probability under changing traffic conditions while at the same time achieving a high channel utilization.

Keywords: bandwidth allocation, call admission control

1. Introduction

We have recently witnessed a proliferation and rapid deployment of the wireless cellular communication services. One of the major challenges is to effectively utilize the prime scarce resource (i.e., radio channels) in the emerging micro-cell and pico-cell environment while at the same time guaranteeing the QoS of the on-going calls [18,19]. There are two generic and critical QoS parameters in mobile wireless networks, namely the call (handoff) dropping probability and the new call blocking probability. Dropping a call in progress is generally considered to be more severe, and needs to be kept under control. An efficient bandwidth allocation scheme has to ensure that the call dropping probability is maintained at a predefined level while at the same time minimizing the new call blocking probability (or maximizing the channel utilization).

* A preliminary version of this work was presented at PIMRC'99 [10]. This work was supported in part by a grant from Hong Kong Telecom Institute of Information Technology under contract HKTIIT97/98.EG01.
** In addition, the research of Bo Li was also supported by grants from Research Grant Council (RGC) under contracts HKUST6071/97E, HKUST6157/98E and CRC 98/01.EG03.

The *trunk reservation scheme* (also called the *guarded channel scheme*) has been extensively studied in the traditional voice-centric cellular networks [4,5]. The basic idea is to reserve a fixed number of channels in each cell exclusively for handoffs. This is shown to be able to decrease the dropping probability for the admitted calls. Moreover, the schemes proposed in [1,9,13] allow the queueing of the handoff requests when there is no (reserved) channel available, which can further reduce the dropping probability at the expense of higher new call blocking. Ramjee et al. proved that such a scheme is optimal for a linear objective function of call dropping and new call blocking probabilities [16]. In addition, they proposed a *fractional guarded channel policy* that is optimal for minimizing the call blocking probability subject to a hard constraint on the call dropping probability. In other words, for a given set of parameters including traffic parameters and mobility characteristics, the fixed bandwidth allocation scheme based on guarded channel assignment or its variants can yield an optimal solution. All such schemes, however, by reserving a fixed number of channels, cannot adapt to changes in the network conditions due to its static nature. This is clearly not suitable in the presence of bursty data and the emerg-

ing multimedia traffic. Therefore, an adaptive and dynamic bandwidth allocation is essential.

A call admission control scheme was proposed in [20]. This scheme is adaptive to changing traffic by evaluating the network conditions before each new call can be established. However, this scheme cannot guarantee the more stringent QoS parameter as the bandwidth is only reserved at the cell where the new call is initiated, and thus subsequent handoff calls have higher risks of being dropped. *The shadow cluster concept* proposed by Levine et al. allows predictive bandwidth allocation [8], wherein upon each call set up request, the mobile needs to provide the bandwidth requirements and accurate mobility parameters (position and movement). Such information is passed to the base stations of the cell that the mobile resides as well as the neighboring cells, all of which reserve the bandwidth in advance accordingly. This scheme is shown to be able to guarantee the QoS. The major drawback is the requirement of the detailed trajectory information and the signaling involved for each call setup.

Another adaptive bandwidth reservation scheme was proposed by Oliveira et al. [14], in which the bandwidth is allocated for a new call in the cell where the call request originates, and in addition the bandwidth is reserved in all neighboring cells. When a handoff occurs, similar bandwidth allocation and reservation are carried out, and reserved bandwidths in some original neighboring cells are released. The novelty of the scheme is that the amount of bandwidth reserved can be dynamically adjusted, reflecting the actual traffic conditions in the network. A similar approach was proposed in the adaptive admission control scheme by Mišić et al. [11]. Resource estimations are triggered by the events of call handoff, origination and termination. Bandwidths reserved for possible handoffs are estimated using the "spatial activity factors", providing an approximate control of the QoS. However, the computational and signaling complexities of these approaches are still heavy due to the updates required for each call event.

We recently proposed a dynamic call admission scheme [21] based on a periodical control, similar to the one proposed by Naghshineh and Schwartz [12]. The SDCA can *precisely* guarantee the target call dropping probability while at the same time maximizing the channel utilization. The precision is obtained by taking into account the time dependence of the call dropping probability and the effect of the non-neighboring cells. In addition, it also greatly improves over the Gaussian approximation commonly used [7,12]. However, this also requires periodic status information exchange among different cells.

The paradox, however, is that all proposed dynamic bandwidth allocation schemes require the prior knowledge of the traffic parameters or/and user mobility parameters. Under such conditions, the fixed bandwidth assignment based on the guarded channel scheme can yield optimal solutions for steady state [16], but cannot adapt to changing traffic conditions. In addition, most of the proposed schemes require the status information exchange among different cells in order to dynamically readjust the control parameters, thus making them difficult to be used in actual deployment. *Our major motivation* in this paper is to design an adaptive and dynamic bandwidth allocation scheme that can overcome *these two major deficiencies*. The main features of the proposed algorithm are:

- The estimation is done *on-line* and *periodically*, hence, it can effectively adapt to the changing traffic. This is particularly suitable for the emerging multimedia type of traffic in that the statistical behavior of the traffic is either not available or is difficult to obtain.
- The bandwidth allocation is implemented by a *probabilistic* mechanism, which can reserve the bandwidth in an efficient statistical multiplexing manner. This eliminates the need to reserve bandwidth explicitly for each call set up. In addition, this can spread new arrivals evenly over a control period, thus leading to more effective and stable control.

The rest of the paper is organized as follows. We describe the bandwidth allocation algorithm and on-line estimation algorithm in section 2. In section 3, we study the performance of our proposed bandwidth allocation scheme through simulations, and further investigate the impact on its performance under a variety of changing traffic conditions. We present the conclusion in section 4.

2. The adaptive bandwidth allocation algorithm

We consider a cellular network consisting of close packed hexagonal cells and using a fixed channel allocation scheme. Each cell has a capacity of N channels. New calls arrive in cell i at a rate of λ_i. Connected calls terminate at a rate of μ (i.e., $1/\mu$ is the average call duration time). In addition, calls hand off from cell k to a neighboring cell i at a rate of h_{ik} per call. Let $h_k \equiv \sum_i h_{ik}$ be the handoff rate per call out of cell k.

The bandwidth allocation algorithm is executed in a distributed and periodic fashion. Each cell executes the identical algorithm based on local estimations. The length of the control period is T. At the beginning of a control period, the bandwidth allocation algorithm determines the amount of bandwidth reserved in the next control period for the particular cell by taking into consideration the network traffic conditions. Notice that such bandwidth reservation is done for *all potential handoffs* in a control period, thus eliminating the need for reserving bandwidth for each call required in other schemes [8,14]. More importantly, this enables a more efficient statistical multiplexing, leading to a more effective use of the bandwidth. Before we present the bandwidth allocation algorithm, we first summarize its key features below:

(1) The QoS requirement that the algorithm provides is the dropping probability (P_{QoS}). We calculate an expression of the P_{QoS} for the *call acceptance ratio* a_i, which is defined as the fraction of new calls to be admitted into cell i in the coming control period. Instead of using it

to determine an admission threshold (i.e., the number of new calls that are allowed) as in a guard channel policy, we stochastically accept each new call with probability a_i, which can spread the new calls uniformly over the period. This avoids a sudden overload of the network at the beginning of the control period during congestion, leading to a more effective and stable control.

(2) We derive the call dropping probability as a *time-dependent function* $D_i(t)$ in a cell while taking into account its *finite capacity*. It greatly improves over the Gaussian approximation commonly used [7,12]. We then compute the average dropping probability $\widetilde{D}_i$ as in equation (17) over a control period, taking into account its time dependence. This increases the precision over a single-value approximation within the control period.

The major computational complexity of the control algorithm is to obtain the acceptance ratio by solving the nonlinear equation (18) for the average dropping probability on-line. However, since the control is stochastic, a coarse-grain integration of the average dropping probability is already sufficient.

To derive the control algorithm, the key is to obtain the acceptance ratio a_i for cell i via equation (18) according to the following steps: (a) the arrival rate, the handoff rate and the average channel occupancy in the local cell are estimated on-line in section 2.1; (b) the time-dependent survival probability of a call in a cell is computed with the estimated handoff rate in section 2.2 to be used in the following step; (c) these parameters are used to compute the evolution of the mean $\langle n_i(t)\rangle$ and variance $\sigma_i(t)^2$ of the time-dependent occupancy distribution in each cell in section 2.3; (d) in turn, these enable us to find the evolution of the dropping probability in section 2.4, prescribed by the off-line solution to a diffusion equation applicable to sufficiently large systems.

2.1. The on-line estimation algorithm

The on-line estimation algorithm is developed to reduce the signaling load required in most dynamic call admission algorithms [7,8,11,12,21]. For example, in [21], cell i has to obtain status information from all cells that have potential handoffs at the beginning of each control period. For all neighboring cells k, the signaled parameters include the channel occupancy n_{k0}, the traffic arrival λ_k and the acceptance ratio a_k in previous control period. In addition, it has to pass exactly the same set of parameters of its own (n_{i0}, λ_i and a_i) to all those cells. This requires a significant amount of signaling. One potential solution is to enlarge the control period T so as to reduce the signaling frequency. However, this can result in the inaccuracy in computing the channel occupancy distribution, since statistical uncertainty grows with time. More seriously, if the traffic condition changes in any cell, the control algorithm will not be able to adapt its control parameters in time. These two factors hinder the deployment of such algorithms in actual implementation.

To overcome this limitation, an on-line estimation algorithm is implemented by restricting the use of information to those only from the local cell i, while the status of the neighboring cells are derived by estimation rather than actual signaling. An exponential smoothing technique from time series analysis is adopted to compute the expected arrival rate from the observed values. Noticeably, a similar technique was used in TCP adaptive retransmission to estimate the round-trip time (RTT) [6].

Concretely, let $\lambda_i^{(o)}(j)$ be the observed arrival rate in cell i for the jth control period. This value is needed and is available at the beginning of the $(j+1)$th control period (i.e., the end of jth control period). Let $\lambda_i^{(e)}(j)$ be the estimated arrivals for the jth control period (at the beginning of the jth control period). Using exponential smoothing, we have

$$\lambda_i^{(e)}(j+1) = \alpha_1 \lambda_i^{(e)}(j) + (1-\alpha_1)\lambda_i^{(o)}(j). \qquad (1)$$

Under the uniform handoff rate case, the handoff rate h can also be obtained similarly:

$$h_i^{(e)}(j+1) = \alpha_2 h_i^{(e)}(j) + (1-\alpha_2)h_i^{(o)}(j). \qquad (2)$$

For the observed channel occupancy $n_{i0}(j+1)$ at the beginning of the $(j+1)$th control period, we note that it consists of two components. First, the *controllable* component consists of the channels occupied by calls admitted to cell i during the jth control period. It is directly controlled by the admission actions of the local cell, and is approximated by $a_i(j)\lambda_i^{(o)}(j)T$. Secondly, the *background* component consists of the channels occupied by all the ongoing calls which were admitted in the previous control periods, and handoff calls which entered the cell during the jth period or beforehand. The background occupancies cannot be controlled directly by the actions of the local cell, but it is expected to exhibit some long term statistical behavior given the traffic does not change too rapidly. This will be further elaborated using examples in section 3. Letting $N^{(o)}(j)$ be the *observed background channel occupancy* at the end of the jth control period, we have

$$N^{(o)}(j) = n_{i0}(j+1) - s_i(T)a_i(j)\lambda_i^{(o)}(j)T, \qquad (3)$$

where $s_i(t)$ is the survival probability of a call in cell i averaged over a time interval t, the expression with $t = T$ being the average over the entire jth control period. $s_i(t)$ is computed in the following subsection.

The *estimated* background channel occupancy for the $(j+1)$th control period is given by

$$N^{(e)}(j+1) = \alpha_3 N^{(e)}(j) + (1-\alpha_3)N^{(o)}(j). \qquad (4)$$

Notice that the coefficients α_i $(i = 1, 2, 3)$ used in equations (1), (2) and (4) have to be properly selected to *smooth* all the estimated values. In general, a small value of α_i (thus, a large value of $1-\alpha_i$) can keep track of the changes more accurately, but is perhaps too heavily influenced by temporary fluctuations. On the other hand, a large value of α_i is more stable but could be too slow in adapting to real traffic changes. In our experiment, we find the setting of α_1 and α_2

between 0.6 and 0.7, and α_3 between 0.8 and 0.9 are adequate for the estimation.

2.2. The survival probability for uniform handoff rate

In this subsection, we present the derivation for the survival probability for uniform handoff rate ($h_{ik} = h/6$). Specifically, we use the estimated value of $h = h^{(e)}$ obtained at the beginning of each control period.

Consider the single-call transition probability $f_{ik}(t)$ that an ongoing call in cell k at the beginning of the control period ($t = 0$) is located in cell i at time t. In particular, $f_{ii}(t)$ is the survival probability of a call in cell i at time t. For an effective control enforcing dropping probabilities of the order 10^{-3}–10^{-2}, we assume that essentially all calls hand off successfully, resulting in the evolution equation

$$\frac{\mathrm{d}f_{ik}(t)}{\mathrm{d}t} = -\sum_j J_{ij} f_{jk}(t) \quad \text{and} \quad f_{ik}(0) = \delta_{ik}, \tag{5}$$

where J_{ik} is the transition matrix given by $J_{ii} = h_i + \mu$ and $J_{ik} = -h_{ik}$ for $i \neq k$. The solution to (5) is

$$f_{ik}(t) = \big[\exp(-Jt)\big]_{ik}. \tag{6}$$

The computational complexity of this matrix operation can be reduced by considering the off-diagonal terms as perturbations to the diagonal part of J. Each term in the resultant perturbation series of $f_{ik}(t)$ corresponds to the contribution of a path connecting k and i by cell hopping. While the perturbation technique is applicable to non-uniform handoff rates in general [21], results for the case of uniform rates is particularly illustrating. Notice that the matrix J can be written as

$$J = J_0 + J_1, \tag{7}$$

where

$$\begin{aligned} (J_0)_{ik} &= (h_k + \mu)\delta_{ik}, \\ (J_1)_{ik} &= \begin{cases} -h_{ik}, & k, i = \text{nearest neighbors}, \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{8}$$

For homogeneous handoff rates, $h_k = h$ and $h_{ik} = h/6$. Considering J_1 as the perturbation, we have

$$\big[\exp(-Jt)\big]_{ik} = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!}\big[J_0^r + \big(J_0^{r-1}J_1 + \cdots + J_1 J_0^{r-1}\big) + \cdots\big]_{ik}. \tag{9}$$

The zeroth-order term consists of those terms in (9) which contain no J_1. Hence, the zeroth-order contribution goes only to $i = k$, with

$$q_0(t) = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!}\big(J_0^r\big)_{ii} = \exp\big[-(h+\mu)t\big]. \tag{10}$$

It corresponds to the case that no handoff events take place between time 0 and t.

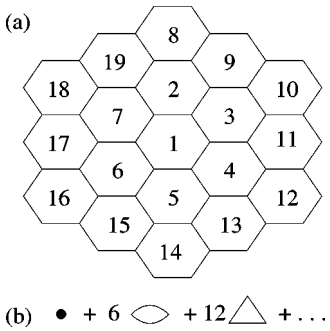

Figure 1. (a) A 19-cell hexagonal cellular network with wrap-around connection. (b) Topology of paths connecting $k = i$.

The first-order terms consist of one and only one J_1. Since the elements of J_1 are nonzero only for neighboring cells, the first-order contributions go only to neighboring cells i and k, with

$$\begin{aligned} q_1(t) &= \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} \\ &\quad \times \left[(h+\mu)^{r-1}\frac{h}{6} + \cdots + \frac{h}{6}(h+\mu)^{r-1}\right] \\ &= \frac{ht}{6}\exp\big[-(h+\mu)t\big]. \end{aligned} \tag{11}$$

It corresponds to the event that a call is handed off from cell k to i.

Higher order contributions can be evaluated similarly. For a path with n hops along the path from k to i, we obtain

$$q_n(t) = \frac{1}{n!}\left(\frac{ht}{6}\right)^n \exp\big[-(h+\mu)t\big]. \tag{12}$$

This equation can be intuitively interpreted by counting the number of handoff events along the path. Since all hopping and termination events are assumed to take place independently at the same rate, the occurrence of n such events in time t obeys a Poisson distribution with mean $(h + \mu)t$. A path with n hops requires each of the n events to be a handoff to a specified neighboring cell, excluding the other five neighbors and the termination event. Hence, the probability $q_n(t)$ is given by $[h/6(h+\mu)]^n p_n$, where the Poisson distribution $p_n = [(h+\mu)t]^n \exp[-(h+\mu)t]/n!$, resulting in equation (12).

Hence, $f_{ik}(t)$ is obtained by summing over all possible paths between k and i. For the cellular network in figure 1(a), figure 1(b) shows the example of $k = i$, in which each diagram represents the topology of a path connecting i to itself, with vertices and edges representing cells and paths, respectively. Hence, there are one path of 0 hops, no path of 1 hop, six paths of 2 hops and twelve paths of 3 hops, leading to

$$f_{ii}(t) = q_0(t) + 6q_2(t) + 12q_3(t) + \cdots. \tag{13}$$

Since ht is the average number of hops in time t, the resultant perturbation series is rapidly converging for ht up to

O(1). For a handoff rate h as high as 0.05 s^{-1} and $t = 20$ s, $\mu = 0.005$ s^{-1}, the computed values for $f_{ii}(t)$ are lower than the true values by 1% up to 2 hops, and 0.3% up to 3 hops.

The survival probability averaged over time t is given by

$$s_i(t) = \frac{1}{t}\int_0^t \mathrm{d}t'\, f_{ii}(t-t'), \tag{14}$$

whose closed form can be easily obtained using integration by parts.

2.3. The mean and variance of the occupancy distribution

The channel occupancy distribution at a cell $p_{n_i}(t)$ is determined by the superposition of the ongoing calls and the new calls admitted to the network. Based on the estimation obtained in section 2.1, the mean of the occupancy distribution $p_{n_i}(t)$ in cell i at time t of the $(j+1)$th control period consists of both the background and controllable components, given by

$$\langle n_i(t)\rangle = N^{(\mathrm{e})}(j+1) + s_i(t)a_i(j+1)\lambda_i^{(\mathrm{e})}(j+1)t, \tag{15}$$

where the background component $N^{(\mathrm{e})}(j+1)$ is obtained in equation (4), $\lambda_i^{(\mathrm{e})}(j+1)$ is the estimated new call arrival in the next $(j+1)$th control period (equation (1)), and $a_i(j+1)$ is the acceptance ratio for the next control period *that needs to be computed*.

Similarly, we can obtain the estimation for the *variance* $\sigma_i(t)^2$ of the channel occupancy distribution at time t. As it turns out that the channel occupancy distribution can be approximated by a Poisson one, we take the variance to be the same as the mean $\langle n_i(t)\rangle$.

The Poisson nature of the channel occupancy is justified for the new calls within the same control period, either directly arriving at the local cell, or entering after first arriving at a neighboring cell. It is also a good approximation for handoff calls entering the local cell, provided that they are only a small fraction of the active calls in the originating cells. However, the Poisson assumption is only a rough approximation for the surviving calls in the local cell. Indeed, the number of calls surviving from the previous control periods obey a binomial distribution with variance $f_{ii}(t)[1-f_{ii}(t)]n_{i0}$, rather than a Poisson distribution with a slightly larger variance of $f_{ii}(t)n_{i0}$. As a result, this discrepancy leads to an underestimation of the variance, and hence, a slight overprovision of the bandwidth. Nevertheless, as justified by the results in section 3, the approximation is adopted because of the reduced computational complexity at the expense of only a slight and acceptable overprovision.

2.4. The call dropping probability

Now we introduce the time-dependent call dropping probability $D_i(t)$ for cell i. The dropping probability can be expressed in terms of the quantities n_{i0} (channel status at the beginning of a control period), $\langle n_i(t)\rangle$ (the *mean* of the channel occupancy distribution) and $\sigma_i(t)^2$ (the *variance* of the channel occupancy distribution). It is obtained by solving the diffusion equation describing the evolution of the channel occupancy distribution while taking into account the *finite capacity* of each cell. The derivation is outlined in appendix, with the result

$$D_i(t) = 2\frac{\exp[-\xi_i(t)^2/2]}{\sqrt{2\pi\sigma_i(t)^2}} + 2\frac{[\langle n_i(t)\rangle - n_{i0}]}{\sigma_i(t)^2}\,H\big(\xi_i(t)\big), \tag{16}$$

where $\xi_i(t) \equiv (N - \langle n_i(t)\rangle)/\sigma_i(t)$ is the normalized vacancy in cell i at time t, with N being the capacity of cell i, and $H(x)$ is related to the *complementary error function* via $H(x) = \mathrm{erfc}(x/\sqrt{2})/2$ [15]. The $\langle n_i(t)\rangle$ and $\sigma_i(t)^2$ are the *mean* and *variance* of the channel occupancy distribution, which are given in section 2.3.

The average dropping probability over a control period is obtained by

$$\widetilde{D}_i = \frac{1}{T}\int_0^T \mathrm{d}t\, D_i(t). \tag{17}$$

For an on-line periodic control, the complexity of the integration could be very high. However, since our control is based on a probabilistic model, the precision for integration needs not be high. We found that it is sufficient to use a 7-point Simpson rule [15]. The acceptance ratio a_i can then be easily obtained by solving numerically

$$\widetilde{D}_i = P_{\mathrm{QoS}}. \tag{18}$$

At low traffic, it may happen that $\widetilde{D}_i < P_{\mathrm{QoS}}$ even for $a_i = 1$. Then a_i is set to 1. Similarly, at high traffic, a_i is set to 0 if $\widetilde{D}_i > P_{\mathrm{QoS}}$ even for $a_i = 0$.

3. Results

Simulations were performed on a hexagonal cluster of 19 cells given in figure 1. To alleviate *finite size effects*, we implement periodic connections on the 3 pairs of opposite sides of the cluster (wrap-around). The parameters used in the simulation are: $N = 100$, $\mu = 0.005$ s^{-1}, $h_i = h = 0.01$ s^{-1}, $T = 20$ s, and $P_{\mathrm{QoS}} = 0.01$. Under such a setting, a connection lasts on average 200 s and the mobile hands off twice during its life time. The coefficients α_1 and α_2 used are 0.6, unless specified otherwise. Except for the figures 13 and 14, the handoff rate is assumed to be given.

We first compare the result with that of SDCA proposed in [21]. Figure 2 shows that both schemes can guarantee the target call dropping probability ($P_{\mathrm{QoS}} = \widetilde{D}_i = 0.01$), but the scheme based on local parameter estimation has a slight overprovision of the bandwidth, thus yielding a call dropping probability slightly lower than the target. This is caused by the conservative estimation in the control algorithm, partly due to the Poisson approximation for the *variance*. Notice that this is also evident from the utilization curve shown in figure 3.

We next present the results when the traffic condition changes, by considering the scenario that the traffic input

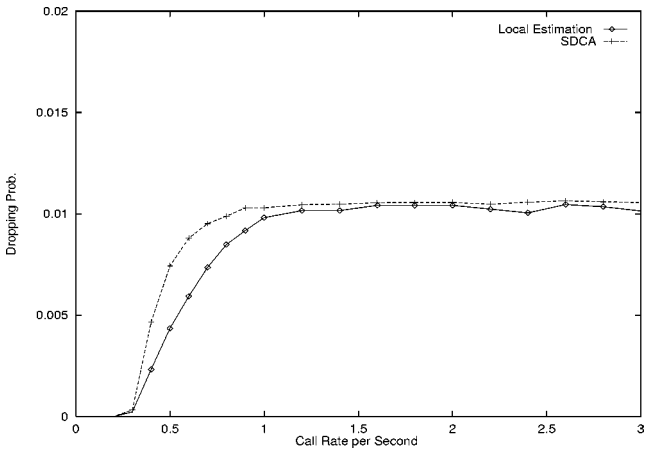

Figure 2. The comparison of call dropping probability with the SDCA.

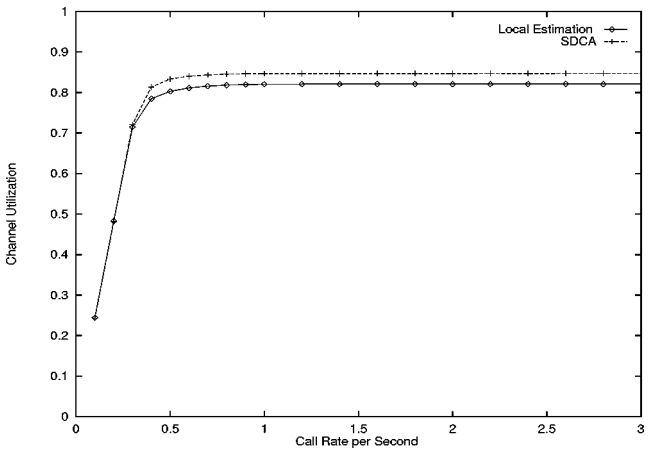

Figure 3. The comparison of channel utilization with the SDCA.

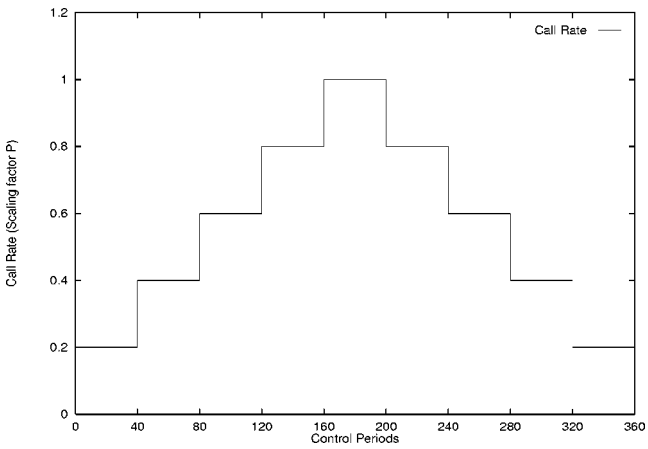

Figure 4. The evolution of the traffic input scaled by the factor P.

changes periodically, as is best reflected from daily telephone operations. Specifically, the traffic input evolves as the staircase function shown in figure 4, in which each step of the staircase is 40 control periods or 800 s. The parameter P is the scaling factor, and the traffic changes in steps of $0.2P$.

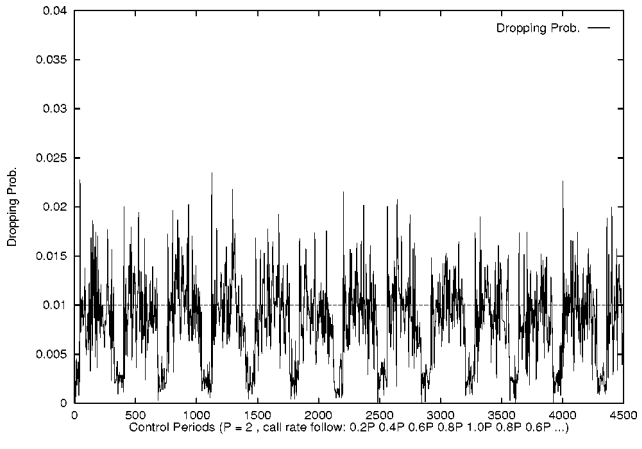

Figure 5. The dynamic call dropping probability for each control period at $P = 2\ \mathrm{s}^{-1}$.

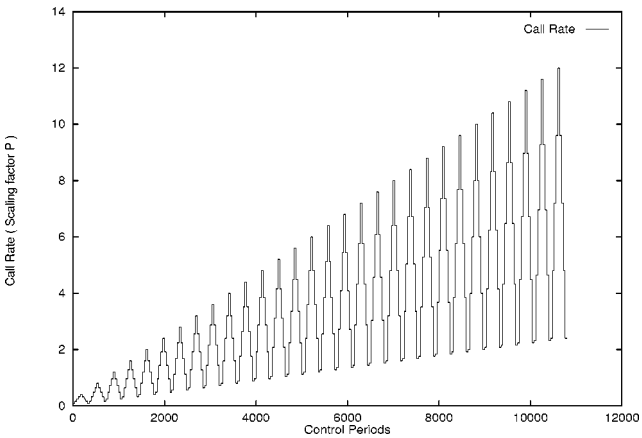

Figure 6. The traffic input with scaling factor P increased from 0.4 to $12\ \mathrm{s}^{-1}$.

Figure 5 demonstrates the dynamic behavior of the call dropping probability for each control period when the network is under saturation loading $P = 2\ \mathrm{s}^{-1}$. In this case, the long term call dropping probability is maintained at 0.0085 (below the target dropping probability 0.01). The utilization obtained is 0.81. The new call blocking probability is 0.67 caused by the overloading condition. One observes from figure 5 that the temporal behavior of the call dropping probability is also periodic, well matching the periodic changes of the traffic input. In addition, careful studies show that the target call dropping probability is violated for the first few control periods when the traffic λ_i changes, due to inaccurate estimations of the traffic. Once the estimation becomes stable, the target dropping probability is guaranteed for subsequent control periods with the same traffic.

We next investigate the performance under fluctuating traffic conditions. We consider a similar traffic pattern but with increasing overall traffic intensity given in figure 6. This is the same as previous one shown in figure 5 except that the scaling parameter P increases after every long cycle of 360 control periods. Figure 7 demonstrates that the cumulative average of the call dropping probability is still well maintained below the target. The call dropping prob-

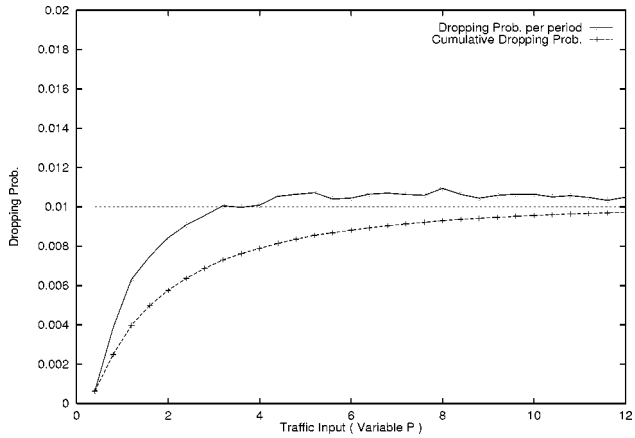

Figure 7. The call dropping probability versus changing P.

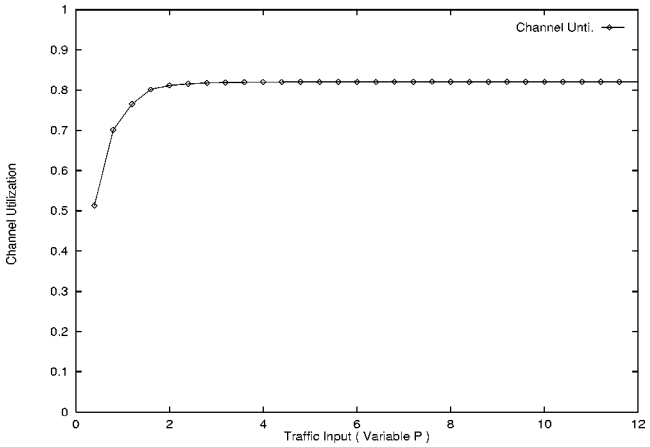

Figure 8. The channel utilization versus changing P.

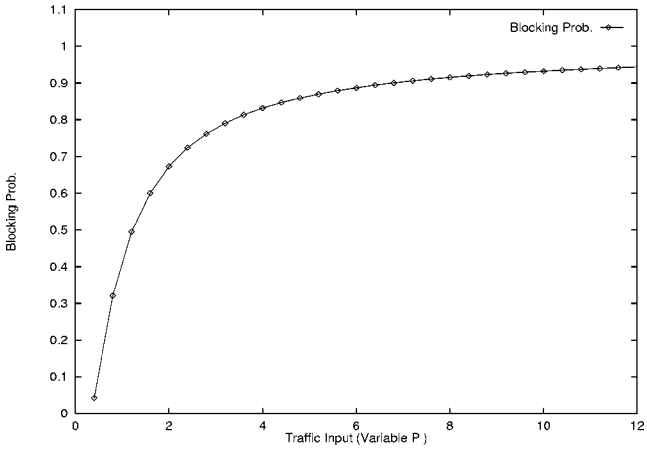

Figure 9. The new call blocking probability versus changing P.

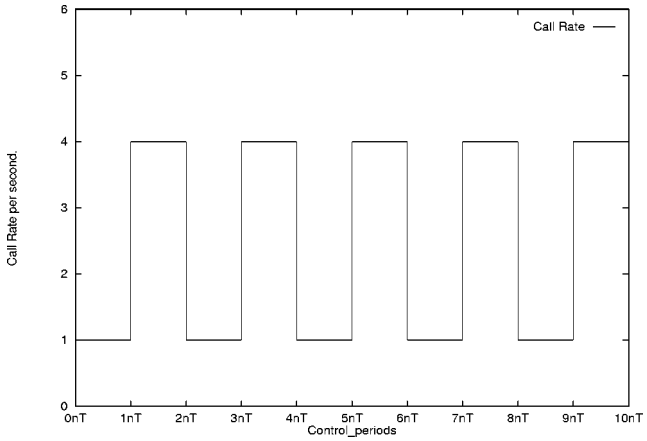

Figure 10. The traffic input with periodic change.

ability for each cycle (i.e., for a fixed P) is also presented in figure 7. Notice that the cumulative average is considerably lower than the individual measurements, but the two converge when the scaling factor P becomes significantly large, as expected. The corresponding utilization and new call blocking probability are plotted in figures 8 and 9, respectively. The high blocking probability for new calls is a consequence of the overloaded traffic, which is the range of interest. This clearly demonstrates the robustness and stability of our bandwidth allocation mechanism. Under such heavy loading, the channel utilization is maintained at over 80%. This also illustrates the fact that on average about 20% of bandwidth is reserved for handoff in order to maintain the target call dropping probability.

We next consider more volatile traffic conditions, and study the impact on the call dropping probability of different estimation coefficients α_i (i.e., α_1 and α_2, since the handoff rate is assumed to be given). While larger α_i's yield better performance at the steady state, smaller α_i's are necessary to cope with volatility. The input traffic follows a periodic change given in figure 10, with n being the number of control periods. Note that under the given system parameters, the system saturates when the average traffic input is $0.5\ \mathrm{s}^{-1}$, therefore, under both low input ($\lambda = 1\ \mathrm{s}^{-1}$) and high input

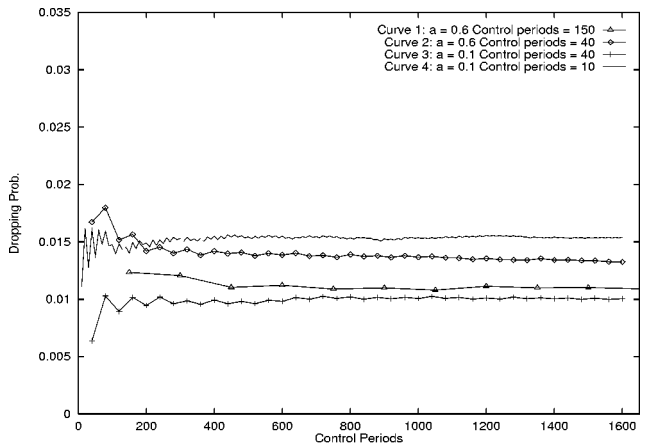

Figure 11. The call dropping probability versus adjusted α values.

($\lambda = 4\ \mathrm{s}^{-1}$), the system is under saturation. Figure 11 describes the behavior of the call dropping probability under a variety of traffic input and different values of α_i. Specifically, curve 1 shows the case of $n = 150$ (i.e., the traffic is changed every 150 control periods) and $\alpha_i = 0.6$. In this case the target call dropping probability is well maintained around $P_{\mathrm{QoS}} = 0.01$. Curve 2 presents a similar scenario

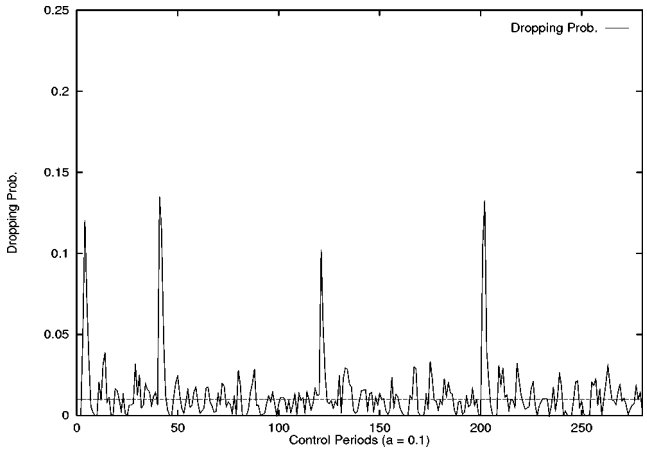

Figure 12. The dynamic call dropping probability for periodic traffic input.

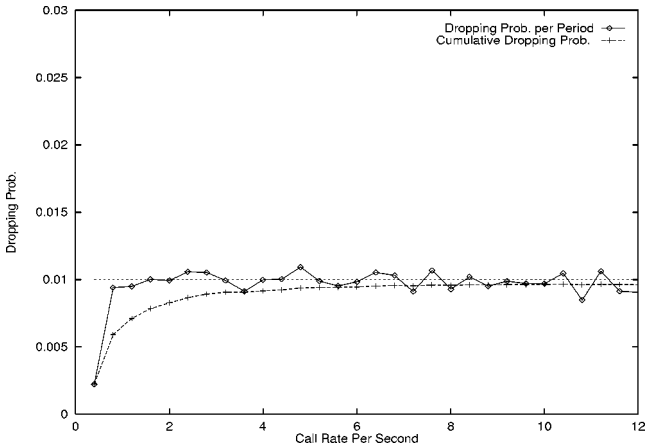

Figure 13. The call dropping probability based on local estimation of the handoff rates.

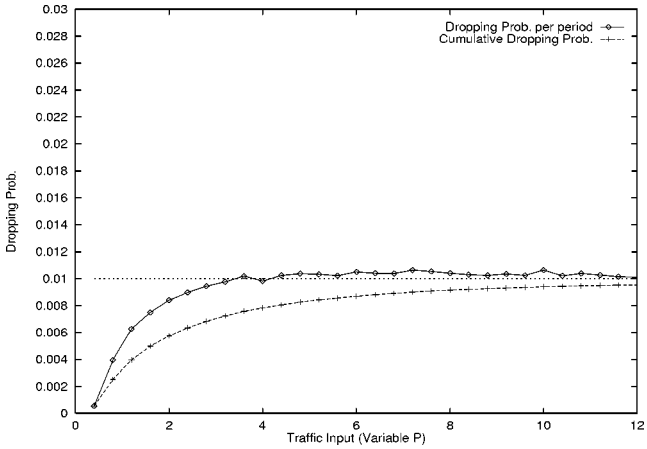

Figure 14. The call dropping probability with estimated handoff rates.

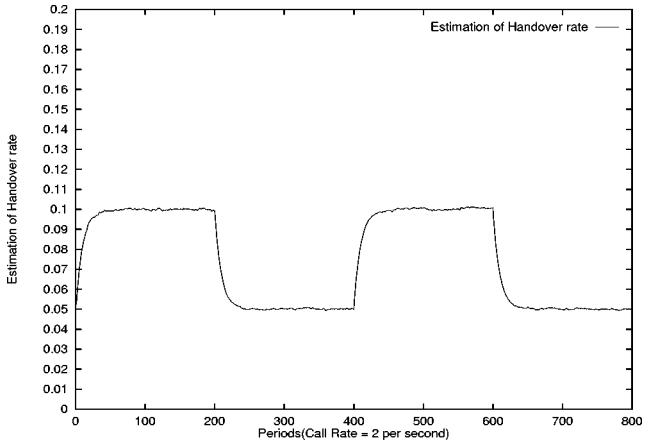

Figure 15. The estimated handoff rate versus time.

with the same α_i, but the traffic is changing more frequently, i.e., the traffic is changed every $n = 40$ control periods. The result from curve 2 shows that the target dropping probability cannot be satisfied. The major reason is that the chosen value of $\alpha_i = 0.6$ is not adequate for keeping track of such frequent traffic changes. Adjusting the value of α_i can clearly improve the performance guarantee. The result is illustrated in curve 3 of figure 11 with α_i set to 0.1, and the target dropping probability is indeed guaranteed. However, the target cannot be met under more frequent traffic updates such as curve 4 in the same figure, in which the traffic is updated every $n = 10$ control periods.

This can also be better observed from the dynamical behavior shown in figure 12. The call dropping probability obtained reflects the periodic change of the input traffic every $n = 40$ control periods. More importantly, at the beginning of the traffic change from low input ($\lambda = 1\ s^{-1}$) to high input ($\lambda = 4\ s^{-1}$), i.e., every 80 control periods, the instantaneous call dropping probability is increased significantly, as much as about 10–15 times the target call dropping probability ($P_{QoS} = 0.01$) in figure 12. This is caused by an excessive underestimation of the input traffic during the initial control periods when traffic increases. Such an impact can be leveraged over longer control periods, such as $n = 150$ when $\alpha_i = 0.6$ and $n = 40$ when $\alpha_i = 0.1$, but clearly cannot be compensated under $n = 10$ with any α_i setting since the call dropping probability in the single control period immediately after the hike in call rate can account for more than 10 periods' target call dropping probability. Therefore, the target call dropping probability cannot be guaranteed for $n = 10$ shown in curve 4 in figure 11.

Finally, we are interested in the system performance when the handoff rates are estimated. First we adjust the traffic input and use the estimation algorithm to trace the handoff changes accordingly. Figure 13 essentially recaptures the call dropping probability shown in figure 2. The only difference is that in figure 13 the handoff rate is on-line periodically estimated as in equation (2), thus the survival probability $f_{ii}(t)$ is computed according to equation (13). It shows that the cumulative (target) call dropping probability can be guaranteed. Figure 14 presents the results for the traffic input given in figure 6, and is similar to those presented in figure 7. Next we assume that the user mobility pattern is periodically changed, specifically, the user handoff rate h is oscillated between 0.05 and 0.1 every 200 control periods. Figure 15 illustrates the handoff rate obtained by the estimation algorithm, which accurately reflects the real changes.

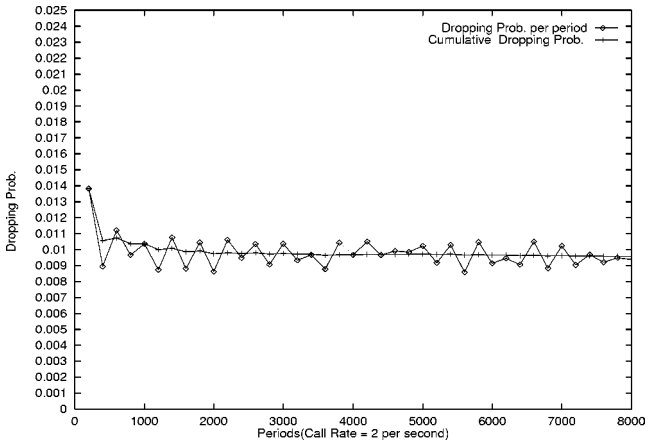

Figure 16. The call dropping probability for periodic handoff rate.

The corresponding call dropping probability is on target and plotted in figure 16.

4. Conclusion

In this paper, we introduce a novel adaptive bandwidth allocation scheme for mobile wireless networks based on local on-line parameter estimations. The novelties of the proposed scheme are: (1) both estimation and bandwidth allocation are carried out periodically, thus, can effectively adapt to the changing traffic conditions; (2) the estimation is restricted to the local cell, thus, eliminating the signaling overhead often required by all existing bandwidth allocation schemes; (3) the allocation algorithm is based on a stochastic control, which results in an efficient use of the bandwidth and leads to an effective and stable control. The results demonstrate that the proposed adaptive bandwidth allocation can guarantee the predefined bound on the call dropping probability under changing traffic conditions, while at the same time achieving high bandwidth utilization.

One limitation is that the control algorithm still relies on the assumptions that the arrival obeys a Poisson process and call durations follow an exponential distribution, other distributions such as Pareto distribution for data traffic [2] and hyper-Erlang distribution [3] will be considered in the future work. In addition, we are investigating other QoS parameters guarantee such as end-to-end delay and considering multiple types of traffic [9,14,17].

Appendix. The time-dependent call dropping probability

For a given cell, let $p_n(t)$ be the probability of having n occupied channels at time t. Assume that the total arrival and departure rates for calls in cell i are Λ and M, respectively. The evolution equation for $p_n(t)$ is then given by

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = \Lambda p_{n-1}(t) - (\Lambda + \mathrm{M})p_n(t) + \mathrm{M}p_{n+1}(t), \quad n < N, \tag{A.1}$$

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = \Lambda p_{n-1}(t) - \mathrm{M}p_n(t), \quad n = N. \tag{A.2}$$

In the limit of large N, the evolution equation (A.1) reduces to a diffusion equation for the continuous distribution $P(x,t)$, where $x \equiv n/N$:

$$\frac{\partial P(x,t)}{\partial t} = -v\,\frac{\partial P(x,t)}{\partial x} + D\frac{\partial^2 P(x,t)}{\partial x^2}, \tag{A.3}$$

where $v \equiv (\Lambda - \mathrm{M})/N$ is the drift velocity, and $D \equiv (\Lambda + \mathrm{M})/2N^2$ is the diffusion coefficient, in analogy with particle diffusion. The boundary condition at $x = 1$ can be obtained from equation (A.2), yielding

$$vP(x,t) = D\,\frac{\partial P(x,t)}{\partial x} \quad \text{at } x = 1. \tag{A.4}$$

The other boundary condition is $P(x,t) = 0$ at $x = -\infty$. The initial condition is $P(x,t) = \delta(x - x_0)$ at $t = 0$, where $x_0 = n_{i0}/N$.

The diffusion equation is solved by Laplace transform. At $x = 1$, the solution is

$$P(1,t) = 2\frac{\exp[-(1 - x_0 - vt)^2/(4Dt)]}{\sqrt{4\pi Dt}} + \frac{v}{D}\,H\left(\frac{1 - x_0 - vt}{\sqrt{2Dt}}\right). \tag{A.5}$$

The dropping probability is given by $D(t) = p_N(t) = P(1,t)/N$, which reduces to equation (16).

References

[1] C.-J. Chang, T.-T. Su and Y.-Y. Chiang, Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping, IEEE/ACM Transactions on Networking 2(2) (April 1994).

[2] M. Cheng and L.-F. Chang, Wireless dynamic channel assignment performance under packet data traffic, IEEE Journal on Selected Areas in Communications 17(7) (July 1999).

[3] F. Fang and I. Chlamtac, Teletraffic analysis and mobility modeling of PCS networks, IEEE Transactions on Communications 47(7) (July 1999).

[4] R. Guerin, Queueing-blocking systems with two arrival streams and guarded channels, Transactions on Communications 36(2) (February 1988).

[5] D. Hong and S.S. Rappaport, Priority oriented channel access for cellular systems serving vehicular and portable radio telephones, IEE Proceedings 136(5) (October 1989).

[6] V. Jackson, Congestion avoidance and control, in: *Proceedings of the ACM SIGCOMM* (August 1988) pp. 314–329.

[7] S.M. Jiang, D.H.K. Tsang and B. Li, Subscriber-assisted handoff support in multimedia PCS, ACM Mobile Computing and Communication Review 1(3) (September 1997).

[8] D.A. Levine, I.F. Akyildiz and M. Naghshineh, A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept, IEEE/ACM Transactions on Networking 5(1) (February 1997).

[9] B. Li, C. Lin and S. Chanson, Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks, Wireless Networks 4(4) (August 1998).

[10] B. Li, L. Yin, K.Y.M. Wong and S. Wu, An efficient and adaptive bandwidth allocation scheme for mobile wireless networks based on on-line local parameter estimations, in: *The 10th IEEE International*

Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC'99), Osaka, Japan (September 1999).

[11] J. Mišić, S.T. Chanson and F.S. Lai, Admission control for wireless multimedia networks with hard call level Quality of Service bounds, Computer Networks and ISDN Systems 31(1–2) (January 1999).

[12] M. Naghshineh and M. Schwartz, Distributed call admission control in mobile/wireless networks, IEEE Journal on Selected Areas in Communications 14(4) (May 1996).

[13] S. Oh and D. Tcha, Prioritized channel assignment in a cellular radio network, IEEE Transactions on Communications 40(7) (July 1992).

[14] C. Oliveira, J.B. Kim and T. Suda, An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks, IEEE Journal on Selected Areas in Communications 16(6) (August 1998).

[15] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1990).

[16] R. Ramjee, R. Nagarajan and D. Towsley, On optimal call admission control in cellular networks, in: *IEEE INFOCOM'96*, San Francisco (March 1996).

[17] P. Ramanathan, K.M. Sivalingam, P. Agrawal and S. Kishore, Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks, IEEE Journal on Selected Areas in Communications 17(7) (July 1999).

[18] T.S. Rappaport, *Wireless Communications: Principles and Practice* (Prentice-Hall, 1996).

[19] M. Schwartz, Network management and control issues in multimedia wireless networks, IEEE Personal Communications (June 1995).

[20] J. Tajima and K. Imamura, A strategy for flexible channel assignment in mobile communication systems, IEEE Transactions on Vehicular Technology 37(5) (May 1988).

[21] S. Wu, K.Y.M. Wong and B. Li, A new, distributed dynamic call admission policy for mobile wireless networks with QoS guarantee, in: *Ninth International IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'98)*, Boston, MA (September 1998); also to appear in IEEE/ACM Transactions on Networking.

Bo Li received the B.S. (cum laude) and M.S. degrees in the Computer Science from Tsinghua University (Beijing) in 1987 and 1989, respectively, and the Ph.D. degree in computer engineering from University of Massachusetts at Amherst in 1993. Between 1994 and 1996, he worked on high performance routers and ATM switches in IBM Networking System Division, Research Triangle Park, North Carolina. He joined the faculty in the Computer Science Department of the Hong Kong University of Science and Technology in January 1996. He has been on editorial board for ACM Mobile Computing and Communications Review (MC2R) and Journal of Communications and Networks (JCN), he is serving as an Editor for Wireless Networks (WINE) and IEEE Journal on Selected Areas in Communications–Wireless Communications series. He has co-guest edited special issues for IEEE Communications Magazine, IEEE Journal on Selected Areas in Communications, SPIE/Baltzer Optical Networks Magazine and ACM Sigmetrics Performance Evaluation Review. He has been involved in organizing many conferences such as IEEE Infocom, ICDCS and ICC. He will be the international vice-chair for INFOCOM'2001. His current research interests are: wireless mobile networking supporting multimedia, voice and video (MPEG-2 and MPEG-4) transmission over the Internet, all optical networks using WDM.

E-mail: bli@cs.ust.hk

Li Yin received her B.S. degree in computer science from Tsinghua University, Beijing, China, in 1998. Since September 1998, she has been a PhD student in the Computer Science Department, Hong Kong University of Science and Technology. Her current research focus is on mobile wireless networks, specifically, resource management and call admission control.

E-mail: yinli@cs.ust.hk

K.Y. Michael Wong received the B.S. degree in physics from the University of Hong Kong in 1978, the M.S. degree in physics in 1982, and the Ph.D. degree in physics in 1986, both from the University of California, Los Angeles. Subsequently he became a Postdoctoral Research Associate in Imperial College, London, and the University of Oxford, UK. In 1992 he became a Faculty Member in the Hong Kong University of Science and Technology, and is now an Associate Professor in Physics. His research interests include learning theory and neural computation, complex optimization, stochastic processes, and applications of learning theory in telecommunications.

E-mail: phkywong@usthk.ust.hk

Si Wu received from the Beijing Normal University, Beijing, China, the B.S. degree in physics in 1990, the M.S. degree in general relativity in 1992, and the Ph.D. degree in statistical physics in 1995. In September 1995, he joined the Physics Department of the Hong Kong University of Science and Technology, where he worked as a Postdoc. He is currently a Staff Scientist in the RIKEN Brain Science Institute, Japan. His research interests include machine learning, neural network, information geometry, computational neuroscience, and the application of neural networks for telecommunication control.

E-mail: phwusi@islab.brain.riken.go.jp