A Dynamic Call Admission Policy With Precision QoS Guarantee Using Stochastic Control for Mobile Wireless Networks

Si Wu, K. Y. Michael Wong, and Bo Li, *Senior Member, IEEE*

***Abstract*—Call admission control is one of the key elements in ensuring the quality of serivce in mobile wireless networks. The traditional trunk reservation policy and its numerous variants give preferential treatment to the handoff calls over new arrivals by reserving a number of radio channels exclusively for handoffs. Such schemes, however, cannot adapt to changes in traffic pattern due to the static nature. This paper introduces a novel stable dynamic call admission control mechanism (SDCA), which can maximize the radio channel utilization subject to a predetermined bound on the call dropping probability. The novelties of the proposed mechanism are: 1) it is adaptive to wide range of system parameters and traffic conditions due to its dynamic nature; 2) the control is stable under overloading traffic conditions, thus can effectively deal with sudden traffic surges; 3) the admission policy is stochastic, thus spreading new arrivals evenly over a control period, and resulting in more effective and accurate control; and 4) the model takes into account the effects of limited channel capacity and time dependence on the call dropping probability, and the influences from nearest and next-nearest neighboring cells, which greatly improve the control precision. In addition, we introduce local control algorithms based on strictly local estimations of the needed traffic parameters, without requiring the status information exchange among different cells, which makes it very appealing in actual implementation. Most of the computational complexities lie in off-line precalculations, except for the nonlinear equation of the acceptance ratio, in which a coarse-grain numerical integration is shown to be sufficient for stochastic control. Extensive simulation results show that our scheme steadily satisfies the hard constraint on call dropping probability while maintaining a high channel throughput.**

***Index Terms*—Call admission control, mobile wireless networks, QoS guarantee.**

I. INTRODUCTION

THERE HAS been a rapid development in wireless cellular communications, in which the quality-of-service (QoS) guarantee remains one of the most challenging issues [2], [23]. One of the key elements in providing QoS guarantees is an effective call admission control (CAC) policy, which not only has to ensure that the network meets the QoS of the newly arriving calls if accepted, but also guarantees that the QoS of the existing calls does not deteriorate.

This paper deals with admission control related to the radio channel assignment. When a mobile moves across cells during its lifetime, dropping is primarily caused by the unavailability of the channels in the new cell. Dropping a call in progress is generally considered to have more negative impact from users' perception than rejecting (blocking) a newly requested call. Therefore, one of the key design goals is to minimize the call dropping probability, which is precisely the objectives of most existing proposals on call admission control. This, however, usually comes at the expense of potentially poor channel utilization by admitting fewer new calls. Given that radio channels are considered to be the primary scarce resource in mobile wireless networks, the *main challenge* in the design of an efficient admission control scheme is to balance these two conflicting requirements. Hence the major performance parameters of interest in this paper are the *call dropping probability*, *channel utilization*, and *new call blocking probability.*

There are a number of unique aspects in the next generation of multimedia wireless networks that the design of an effective admission control scheme needs to take into account.

- Smaller cells will be employed (microcells or picocells), thus the number of handoffs during a call's lifetime is likely to be increased; additionally, there is an increased influence from *neighboring cells* and even *next-neighboring cells* [22].
- Possibly different QoS requirements for different calls, and potentially more stringent QoS requirements of individual calls mandate a highly precise resource allocation [15].
- Diversified traffic load requires that admission control has to be *adaptive* to the changing traffic pattern. Therefore, a dynamic approach is preferred.

This paper introduces a novel *stable dynamic call admission control mechanism* (SDCA) that aims to maximize the radio channel utilization while satisfying a predetermined bound on the call dropping probability. The novelties of the proposed mechanism are as follows.

1) It is dynamically adaptive to a wide range of system parameters and traffic condition due to its dynamic nature.
2) The model takes into account the effects of limited channel capacity and time dependence on the call drop-

Manuscript received February 18, 1999; revised February 21, 2000 and December 8, 2000; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Daigle. This work was supported in part by a grant from Hong Kong Telecom Institute of Information Technology under Contract HK-TIIT97/98.EG01, and by grants from the Research Grant Council (RGC) under Contracts HKUST6157/98E and HKUST6163/00E.

S. Wu is with the Department of Computer Science, Sheffield University, Sheffield S1 4DP, U.K. (e-mail: s.wu@dcs.shef.ac.uk).

K. Y. Wong is with the Department of Physics, Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail: phkywong@ust.hk).

B. Li is with the Department of Computer Science, Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail: BLI@cs.ust.hk).

Publisher Item Identifier S 1063-6692(02)03108-4.

ping probability, and the influences from nearest and next-nearest neighboring cells, which greatly improve the control precision.

3) The admission policy is probabilistic, thus spreading new arrivals evenly over a control period, leading to a more effective and stable control.

We compare our method with a static control, and a recently proposed approximate method for dynamic control. Extensive simulation results show that our scheme outperforms the others by steadily satisfying the hard constraint on call dropping probability while maintaining a high channel throughput.

The paper is organized as follows. We review the relevant work in Section II. In Section III, we describe the control algorithm. In Section IV, we study and compare the performance of our proposed call admission scheme through extensive simulations, and further investigate the impact on its performance from a variety of parameters. To avoid signaling overhead between cells, we introduce localized versions of the admission control in Section V. We present the conclusion in Section VI. Mathematical details for deriving the algorithm are included in Appendices A–C.

II. Existing Call Admission Strategies

The rationale behind the traditional *guard channel scheme* (or *trunk reservation policy*) is to give preferential treatment to the handoff calls, which reserves a fixed number of channels exclusively for handoffs [5], [19]. This scheme was shown by Ramjee *et al.* [21] to be optimal for the linear objective functions of the dropping and blocking probabilities defined above. However, it has a number of deficiencies, in particular, the guard channel scheme cannot satisfy the hard constraints on the call dropping probability often required by multimedia applications. The *fractional guard channel policy* proposed in [21] is shown to be optimal for minimizing the call blocking probability subject to a hard constraint on the call dropping probability. In addition, there have been numerous extensions based on the guard channel scheme. Epstein and Schwartz [4] considered a mixed traffic type with narrow-band and wide-band calls. Another proposal by Acampora and Naghshineh [1] suggests to cluster a group of neighboring cells and allocates a portion of the channels from those cells for handoffs. Li *et al.* [15] extended the guard channel scheme to handle multiple streams of traffic, each having potentially different QoS requirements, thus requiring potentially different channel thresholds. All such policies, however, are *static* in that they do not adapt to changes in the traffic pattern.

A number of recent proposals have made fine attempts to implement dynamic control in the above schemes. The proposed schemes make the admission decision in a distributed manner relying on the status information exchange between adjacent cells, taking into consideration the active calls in the cell where a new call arrives, as well as its neighboring cells to which the call is likely to be handed off.

The *shadow cluster mechanism* by Levine *et al.* [14] is based on the observation that "every mobile terminal with an active wireless connection exerts an influence upon the cells (and their base stations) in the vicinity of its current location and along its direction of travel." The coverage of a shadow cluster for a given active mobile mainly consists of the cell where the mobile is currently present (i.e., the center of the shadow cluster) and all its adjacent cells along the direction of travel. This area changes when the mobile call is handed off to other cells, thus a *tentative shadow cluster* needs to be implemented for every new call as well as every handoff call. Simulations show that the shadow cluster mechanism is able to reduce the percentage of dropped calls in a controlled fashion. The efficiency of this scheme depends on the accuracy of prediction of the future mobile movement, which makes it most suitable for a strong directional environment such as the highway.

On the other hand, the distributed call admission (DCA) scheme by Naghshineh and Schwartz [18] does not need the status information exchange upon each call arrival (new call and handoff). Rather, it only requires the exchange of such information periodically [18]. The admission control algorithm calculates the maximum number of calls that can be admitted to a given cell without violating the QoS of the existing calls in the cell as well as calls in its adjacent cells. One of the main features of the DCA is its simplicity in that the admission decision can be made in real time and does not require much computational effort.

The DCA cannot always guarantee the target call dropping probability, which can be observed from the limited results in the original paper [18] (e.g., Fig. 7) and our own reproduced results shown later. This is due to a number of simplifying approximations in the control mechanism used in the DCA, which potentially can lead to imprecise control decisions. Specifically:

1) It approximates the dropping probability by the tail of a Gaussian distribution, which is applicable only for cells with *infinite capacity.*
2) It neglects the time dependence of the dropping probability, and the estimate *at the end* of the control period is assumed to be the average.
3) It approximates that all the admitted new calls are in progress *at the beginning* of the control period.
4) It neglects the probability that a call can hand off more than once. As a result, the performance of DCA becomes very sensitive to network load. Generally speaking, it yields an excessively low dropping probability at intermediate traffic, which grows rapidly with increasing traffic load.

III. The Control Algorithm

Our main motivations are twofold: 1) an effective call admission control mechanism has to always guarantee the QoS (call dropping probability) under a variety of system configuration and traffic settings; and 2) the scarce radio channel has to be efficiently utilized.

We consider a cellular network consisting of closely packed hexagonal cells, using a fixed channel allocation scheme. Each cell has a capacity of N channels. Since we are dealing with the circuit-switched voice traffic in this paper, following the convention, we assume that new calls arrive according to a Poisson distribution with the rate of λ_i in cell i, call duration time or call holding time is exponentially distributed with the average call duration time $1/\mu$ (i.e., connected calls terminate at a rate of μ). Channel holding time, however, does not necessarily obey the exponential assumption [6], [7] as there exist certain conditions to be held (a necessary and sufficient condition is given in [5]). Jedrzycki ad Leung [11] showed that a lognormal distribution is a more accurate model for channel holding time through field data, and a similar conclusion was drawn in [3]. In this paper, for mathematical derivation, we assume that channel holding time is exponentially distributed with h_{ik} being the rate of handoff from cell k to a neighboring cell i. It turns out that the control algorithm is rather insensitive to this assumption, since we adopt a periodic control in which the length of the control periods is set to less than the dwell time of a call in a cell. This will be further discussed in Section IV using examples.

The objective of our call admission scheme is to maximize the channel utilization (minimize the new call blocking probability) subject to a hard constraint that the handoff dropping probability should be maintained below a predefined threshold P_{QoS} required by a QoS guarantee. In a dynamic and distributed control scheme, this is implemented by periodically exchanging status information between neighboring, and even next-neighboring cells if necessary. Each cell updates its control action at the beginning of the control period of duration T. The exchanged information includes the channel occupancies and the new call arrival rates. In SDCA, the control action is to determine for the next control period the fraction of new calls to be admitted. We summarize the key features of the algorithm below.

1) We estimate *the time-dependent dropping probability* in a cell, taking into account its *finite capacity.* The derivation is based on the solution to the evolution equation of the occupancy distribution. It greatly improves over the Gaussian approximation.
2) We compute the average dropping probability over a control period, taking into account its time dependence. This increases the precision over a single-value approximation within the control period.
3) To alleviate the effects of multiple handoffs over a control period, we base our estimation of the dropping probability on the call transition probabilities between nearest as well as second and third nearest neighboring cells. We show that this has significant impact under certain parameters. While the exact computation of the transition probabilities involves the exponentiation of a matrix whose dimension is the number of cells in the network, we introduce a *local approximation* which reduces the computational complexity, yet yields an excellent precision.
4) The QoS requirement on the dropping probability yields an expression for the *acceptance ratio* a_i, which is the maximum fraction of new calls to be admitted into cell i in the coming control period. Instead of using it to determine the admission threshold as in a guard channel policy, we spread the new calls uniformly over the period, by stochastically accepting each new call with probability a_i. This avoids a sudden overload of the network at the beginning of the control period during congestion, leading to more effective and stable control.

The rest of the section describes the details of the computation. The key is to derive the acceptance ratio a_i for cell i periodically, which is obtained via (21) according to three steps.

1) We compute the intercellular transition probabilities using a local approximation, and hence the mean and variance of the time-dependent occupancy distribution in each cell, summarized in (8) and (9).
2) We derive a diffusion equation whose solution describes the evolution of the occupancy distribution.
3) We introduce a mean-rate approximation which enables us to obtain the dropping probability in (9) by combining the results of 1) and 2).

The major computational complexity of the control algorithm is to obtain the acceptance ratio by solving a nonlinear equation (21) 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. On the other hand, the call transition probabilities for constant handoff rates can be precalculated off-line, either by exact matrix computation or local approximation. For evolving handoff rates, they can also be easily computed on-line using the local approximation.

A. The Local Approximation

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. For an effective control enforcing dropping probabilities of the order 10^{-4} to 10^{-2}, we assume that essentially all calls hand off successfully, resulting in the evolution equation

$$\frac{df_{ik}(t)}{dt} = -\sum_j J_{ij} f_{jk}(t) \quad \text{and} \quad f_{ik}(0) = \delta_{ik} \tag{1}$$

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 (1) is

$$f_{ik}(t) = [\exp(-Jt)]_{ik}. \tag{2}$$

The matrix elements could be obtained by using $[\exp(-Jt)]_{ik} = \sum_\alpha U_{i\alpha} e^{-\lambda_\alpha t} U^{-1}_{\alpha j}$, where λ_α is an eigenvalue of J, and $U_{i\alpha}$ is the ith element of the corresponding eigenvector. However, 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. For illustration, we consider in Appendix A the case of homogeneous handoff rates, i.e., $h_{ik} = h/6$ for all pairs of nearest neighbors i and k. In this case, the contribution of a path takes the simple form

$$q_n(t) = \frac{1}{n!}\left(\frac{ht}{6}\right)^n \exp[-(h+\mu)t] \tag{3}$$

where n is the number of hops along the path from k to i. Hence $f_{ik}(t)$ is obtained by summing over all possible paths between k and i. For the cellular network in Fig. 1(a), Fig. 1(b) shows the example of $k = i$, in which each diagram represents the

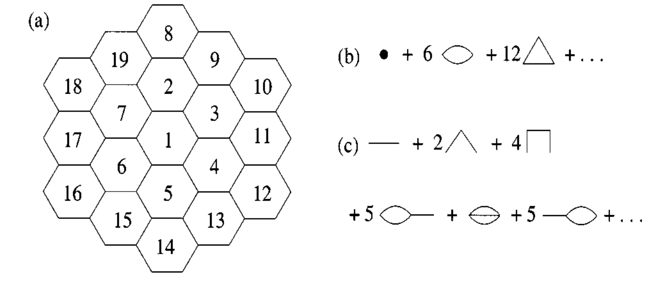

Fig. 1. (a) A 19-cell cellular network. (b) Topology of paths connecting $k = i$. (c) Topology of paths connecting k, $i =$ nearest neighbors. In (b) and (c), the vertices and edges of each graph represent, respectively, the cells and hops of the corresponding path.

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

$$f_{ii}(t) = q_0(t) + 6q_2(t) + 12q_3(t) + \cdots. \tag{4}$$

Similarly, Fig. 1(c) shows the paths connecting k, $i =$ nearest neighbors, giving

$$f_{ik}(t) = q_1(t) + 2q_2(t) + 15q_3(t) + \cdots. \tag{5}$$

Likewise

$$\begin{aligned} f_{ik}(t) =& 2q_2(t) + 6q_3(t) + \cdots, \\ & \text{for } k,\ i = \text{2nd nearest neighbors} \\ f_{ik}(t) =& q_2(t) + 6q_3(t) + \cdots, \\ & \text{for } k,\ i = \text{3rd nearest neighbors}. \end{aligned} \tag{6}$$

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 transition probabilities enable us to estimate the distribution of ongoing calls in cell i at time t. At $t = 0$ there are n_{k0} calls in cell k initially. Since all events (call arrivals, handoffs, and terminations) are stochastic, the number of ongoing calls in cell i at time t is then a superposition of binomial variables, with resultant mean $\sum_k f_{ik}(t)n_{k0}$ and variance $\sum_k f_{ik}(t)[1 - f_{ik}(t)]n_{k0}$.

For the new calls arriving in cell k at time t', the probability of finding them in cell i at time t is $f_{ik}(t - t')$. Since they are evenly distributed over time, the number of new calls in cell i at time t obeys a Poisson distribution with mean $\sum_k g_{ik}(t)a_k\lambda_k$, where $g_{ik}(t)$ is the integrated transition probability given by

$$g_{ik}(t) = \int_0^t dt'\, f_{ik}(t - t'). \tag{7}$$

Knowing the algebraic expressions of $f_{ik}(t)$, it is straightforward to obtain close forms for $g_{ik}(t)$. Thus the mean of the occupancy distribution $p_{n_i}(t)$ in cell i at time t is given by

$$\langle n_i(t)\rangle = \sum_k f_{ik}(t)n_{k0} + \sum_k g_{ik}(t)a_k\lambda_k \tag{8}$$

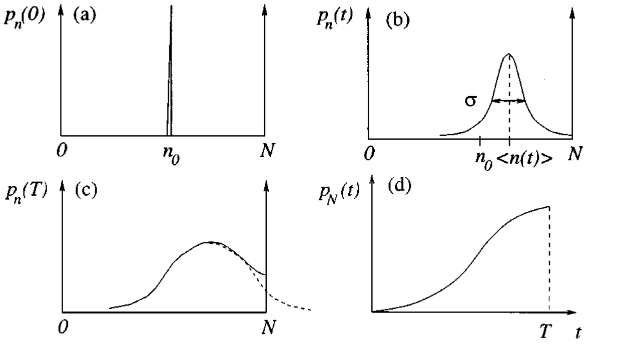

Fig. 2. Schematic diagram of the channel occupancy distribution at (a) $t = 0$, and (b) a subsequent time $t > 0$. (c) The finite capacity N causes a deviation from the Gaussian tail. (d) The evolution of the dropping probability $p_N(t)$.

and the variance is

$$\sigma_i(t)^2 = \sum_k f_{ik}(t)[1 - f_{ik}(t)]n_{k0} + \sum_k g_{ik}(t)a_k\lambda_k. \tag{9}$$

Equations (8) and (9) include the contributions from all calls that can be possibly handoff to cell i from other cells (i.e., $f_{ik}(t)n_{k0}$ and $i \neq k$) during a control period, and cell i's calls that stay at the same cell ($f_{ii}(t)n_{i0}$), new calls admitted to a cell k that are handoff to cell i during the control period ($g_{ik}(t)a_k\lambda_k$ and $i \neq k$), and new calls admitted in cell i ($g_{ii}(t)a_i\lambda_i$). The equations are applicable to long control periods, and the label k can include cells up to arbitrary distance. In practice, we truncate the summation beyond third nearest neighboring cells, since long-range cell hopping can be neglected in a single control period.

The transition probabilities for inhomogeneous handoff rates are derived in Appendix B.

B. The Diffusion Equation

As shown in Fig. 2(a), the initial channel occupancy distribution at the beginning of a control period is a *delta function* given by $p_{n_i}(0) = \delta_{n_i,n_{i0}}$. At a subsequent time t within the control period, the distribution broadens because of the stochastic nature of the events of call handoffs, arrivals, and departures, as schematically shown in Fig. 2(b). For large system sizes, the distribution evolves into a Gaussian distribution with mean $\langle n_i(t)\rangle$ and variance $\sigma_i(t)^2$.

However, the limited capacity of cells modifies the occupancy distribution to non-Gaussian, especially for nearly full occupancy, which is the region of interest in estimating the dropping probability [Fig. 2(c)]. The modified distribution is derived by considering the evolution equation for $p_{n_i}(t)$:

$$\frac{dp_n(t)}{dt} = \Lambda p_{n-1}(t) - (\Lambda + \mathrm{M})p_n(t) + \mathrm{M}p_{n+1}(t), \quad n < N \tag{10}$$

$$\frac{dp_n(t)}{dt} = \Lambda p_{n-1}(t) - \mathrm{M}p_n(t), \quad n = N \tag{11}$$

where Λ and M are the total arrival and departure rates for calls in cell i. Their dependence on n and t is assumed to be negligible. (Unless explicitly specified, the subscript i is omitted hereafter.) In the limit of large N, the evolution equation (10) reduces to a diffusion equation for the continuous distribution $P(x,\,t)$, where $x \equiv n/N$:

$$\frac{\partial P(x,\,t)}{\partial t} = -v\,\frac{\partial P(x,\,t)}{\partial x} + D\,\frac{\partial^2 P(x,\,t)}{\partial x^2} \qquad (12)$$

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 [13].

To consider the boundary conditions in the limit of large N, we first estimate the scaling of various terms. In typical network problems, $\Lambda \sim \mathrm{M} \sim \mathrm{O}(N)$. For an effective admission control, the arrival rate Λ should be adjusted so that $\Lambda - \mathrm{M}$ vanishes to the leading order $\mathrm{O}(N)$, and only statistical fluctuations should contribute, hence $\Lambda - \mathrm{M} \sim \mathrm{O}(\sqrt{N})$. This yields $v \sim \mathrm{O}(1/\sqrt{N})$ and $D \sim \mathrm{O}(1/N)$, and (12) implies that $P(x,\,t)$ should vary significantly over a range of $x \sim \mathrm{O}(1/\sqrt{N})$ and a range of $t \sim \mathrm{O}(1)$.

Applying the scaling argument to (11), the time derivative of $p_n(t)$ becomes negligible, and we arrive at the boundary condition on the right-hand side

$$vP(x,\,t) = D\,\frac{\partial P(x,\,t)}{\partial x} \quad \text{at } x = 1. \qquad (13)$$

Since $P(x,\,t)$ varies significantly in the range $x \sim 1 - \mathrm{O}(1/\sqrt{N})$, we assume that the boundary condition on the left hand side is $P(x,\,t) = 0$ at $x = -\infty$.

The initial condition is

$$P(x,\,t) = \delta(x - x_0) \quad \text{at } t = 0 \qquad (14)$$

where $x_0 = n_{i0}/N$.

This equation is solved in Appendix C. In particular, the solution at the boundary $x = 1$ is

$$P(1,\,t) = 2\,\frac{\exp\left[-\frac{(1-x_0-vt)^2}{4Dt}\right]}{\sqrt{4\pi Dt}} + \frac{v}{D} H\left(\frac{1 - x_0 - vt}{\sqrt{2Dt}}\right) \qquad (15)$$

where

$$H(x) = \int_x^{\infty} du\,\frac{\exp(-u^2/2)}{\sqrt{2\pi}} \qquad (16)$$

which is related to the *complementary error function* via $H(x) = \mathrm{erfc}(x/\sqrt{2})/2$ [20]. The dropping probability is given by $D(t) = p_N(t) = P(1,\,t)/N$.

C. The Mean-Rate Approximation

By (15), the dropping probability is determined by the drift velocity v and the diffusion coefficient D. They determine, respectively, the evolution of the peak and width of the occupancy distribution. To estimate these parameters, we focus on the distribution for occupancies far away from the boundary $x = 1$. For an effective control, this is the region with dominant contribution to the distribution, and hence should give a reliable estimate on the peak and width. In this case, the boundary condition (13) is replaced by $P(x,\,t) = 0$ at $x = \infty$, leading to

$$P(x,\,t) = \frac{\exp\left[-\frac{(x-x_0-vt)^2}{4Dt}\right]}{\sqrt{4\pi Dt}} \qquad (17)$$

which is a Gaussian distribution with mean $x_0 + vt$ and variance $2Dt$. This solution also applies to cells with finite capacity except for the region near the boundary.

A comparison of (17) with (8) and(9) allows us to identify for cell i

$$v = \frac{\langle n_i(t)\rangle - n_{i0}}{Nt}, \qquad D = \frac{\sigma_i(t)^2}{2N^2 t}. \qquad (18)$$

This amounts to a mean-rate approximation, which assumes that the occupancy distribution at time t is the consequence of a constant drift velocity and diffusion coefficient from time 0 up to t. These constants are assigned so that they yield the correct mean and variance of the occupancy distribution at that particular instant. They are updated for every instant the dropping probability is computed. However, since they are only mildly dependent on time, the approximation is very satisfactory.

Hence the dropping probability $D_i(t)$ for cell i can be expressed in terms of the quantities n_{i0}, $\langle n_i(t)\rangle$ and $\sigma_i(t)^2$:

$$D_i(t) = 2\,\frac{\exp[-\xi_i(t)^2/2]}{\sqrt{2\pi\sigma_i(t)^2}} + 2\,\frac{[\langle n_i(t)\rangle - n_{i0}]}{\sigma_i(t)^2}\,H(\xi_i(t)) \quad (19)$$

where $\xi_i(t) \equiv (N - \langle n_i(t)\rangle)/\sigma_i(t)$ is the normalized vacancy in cell i at time t. The average dropping probability over a control period is obtained by

$$\tilde{D}_i = \frac{1}{T}\int_0^T dt\,D_i(t). \qquad (20)$$

For stochastic control, the precision for integration does not need to be high. We found that it is sufficient to use a 7-point Simpson rule [20]. The acceptance ratio a_i can be obtained by solving numerically

$$\tilde{D}_i = P_{QoS}. \qquad (21)$$

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

IV. Results

At the beginning of the control period, cells exchange their status information. To avoid excessive status information exchange, the exchange is limited to first, second, and third nearest neighboring cells, and the computation of the average dropping probability is truncated accordingly. The information transmitted by cell i includes its cell occupancy n_{i0} at that instant and the number of admitted new calls α_i in the previous period. The average new call arrival rate is computed by the moving average $(1-\epsilon)a_i\lambda_i + \epsilon\alpha_i/T \rightarrow a_i\lambda_i$. The transition probabilities are computed in the local approximation for paths up to 3 hops. These parameters are then used to compute the admission ratio in cell i, assuming that the admission ratios in other cells take

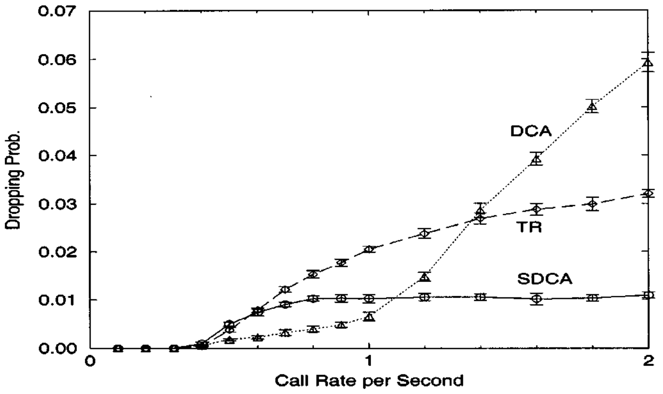

Fig. 3. Dropping probabilities for uniform traffic.

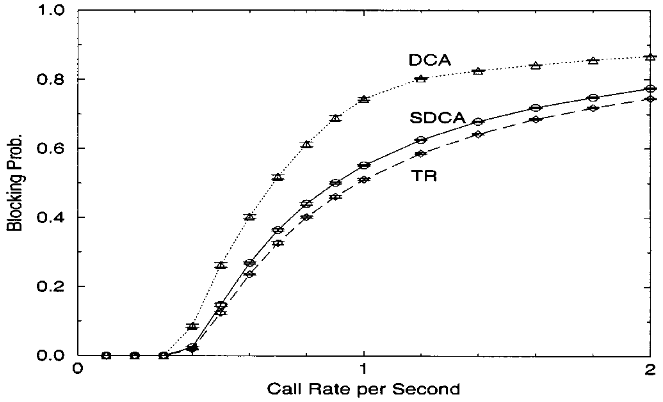

Fig. 4. New call blocking probabilities for uniform traffic.

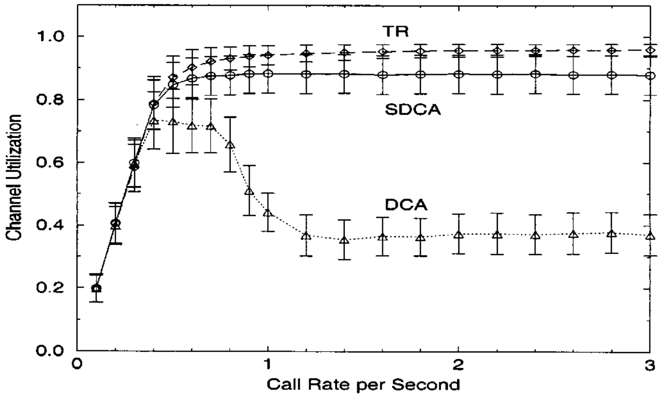

Fig. 5. Channel utilizations for uniform traffic.

their average values. This is done by substituting (8), (9), (19), and (20) into (21), and solving for a_i numerically. We use the bisection method to find the solution.

Simulations were performed on a hexagonal cluster of 19 cells. To alleviate finite size effects, we implement periodic connections on the three pairs of opposite sides of the cluster (wrap-around). The parameters used in Figs. 3–5 are $N = 100$, $\mu = 0.005$ s^{-1}, $h_i = h = 0.01$ s^{-1}, $T = 20$ s, and $P_{QoS} = 0.01$, and the results for SDCA are used as benchmarks for comparing the effects of various parameters in subsequent figures. In Figs. 3–5, the dropping probability and new call blocking probability are compared for SDCA, DCA, and the trunk reservation scheme (TR). Since TR is designed for a static traffic pattern, the dropping probability increases rapidly with the network load when the guard channels are few, but remains too low when the guard channels are many. Here, we choose the number of guard channels to be 5 for comparable performance when the network starts to get overloaded. In each case, simulations were done by averaging over 10 samples, each for 10^4 s of traffic [except in Case 11), where the simulation time is 10^6 s]. Below, we present the results for a number of cases.

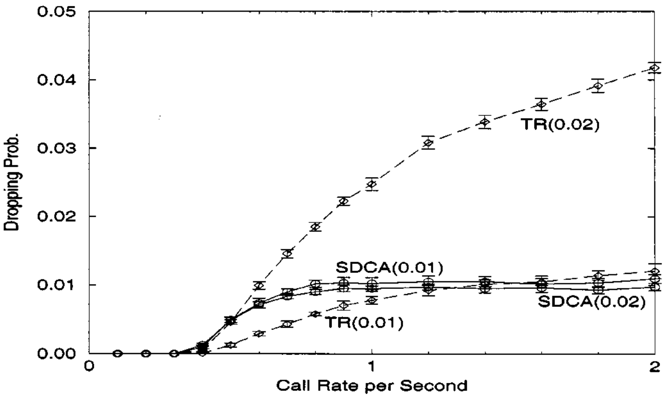

Fig. 6. Dropping probabilities for SDCA and TR when the handoff rate changes from 0.01 to 0.02 s^{-1}.

1) *Uniform traffic:* Fig. 3 shows that the proposed method maintains an almost constant dropping probability for a large range of call rates. Results for DCA are qualitatively the same as in [18], except that here the dropping probabilities are presented in linear rather than logarithmic scale. We remark that in most previous work, QoS guarantees were at most achieved up to the same order of magnitude, i.e., the *logarithm* of the dropping probabilities agrees with the QoS guarantee, whereas here we have enforced the QoS guarantee with a much higher precision. This precision and stability makes it a suitable method for dynamic control. In contrast, neither TR and DCA can enforce the QoS guarantee for the dropping probability. For TR, it grows with the call arrival rate. For DCA, it is far below the QoS requirement for call arrival rates up to 1 per second, and Fig. 4 shows that this is accompanied by a substantial sacrifice in the new call blocking probability. For higher call rates, the dropping probability for DCA grows rapidly. Fig. 5 shows that the channel utilization of SDCA is comparable to TR, and significantly higher than DCA.
2) *Changes in handoff rates:* Fig. 6 shows the performance of SDCA when the handoff rate h changes from $h = 0.01$ s^{-1} to 0.02 s^{-1}. In both cases, SDCA maintains the same level of dropping probability. For comparison, we also show the performance of TR with six guard channels, which is comparable to SDCA at $h = 0.01$ s^{-1}, but yields a significantly higher dropping probability at 0.02 s^{-1}.

This experiment illustrates that SDCA is applicable to a range of handoff rates. Hence if on-line statistics of the handoff rates is available, SDCA can adapt to their

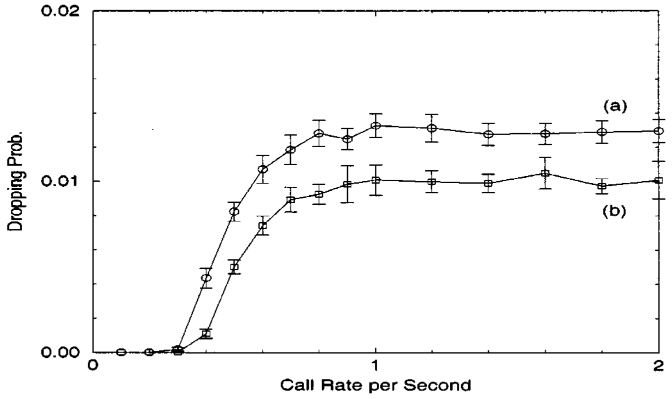

Fig. 7. The dropping probabilities for (a) nonuniform traffic, and (b) inhomogeneous handoff rates.

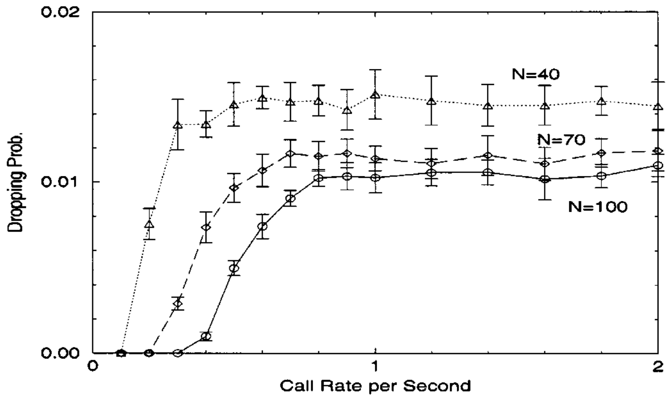

Fig. 8. Dropping probabilities for cell sizes N =100, 70, 40.

changes. Since on-line computation of transition probabilities is greatly simplified in the local approximation, the computational complexity is acceptable.

3) *Non-uniform traffic:* We also simulate a situation for nonuniform traffic, in which the central cell has 2 times the normal call rate and its nearest neighbors have 1.5 times the normal call rate and $h = 0.05\ \mathrm{s}^{-1}$. Fig. 7(a) shows that the proposed method continues to maintain a stable level of dropping probability.

4) *Inhomogeneous handoff rates:* We consider the situation in which the handoff pattern is inhomogeneous but radially symmetric. Referring to the network in Fig. 1(a), cells are arranged in three consecutive rings. For the innermost ring, $h_{1j} = h_{j1} = 0.02/6\ \mathrm{s}^{-1}$. For outer rings, $h_{ij} = h_{ji}$ and the inward, sideways and outward handoff rates have the ratio 3 : 2 : 1. The transition probabilities are computed as in Appendix B. Fig. 7(b) shows that SDCA can cope with the situation very well.

5) *Effects of cell capacity:* Although the SDCA algorithm is derived in the limit of large N, Fig. 8 shows that it is still very effective for N down to 40.

6) *Effects of control period:* To reduce signaling and computational load, it is desirable to increase the control period. Fig. 9 shows that SDCA has a steady performance when the control period is lengthened up to 80 s, roughly comparable to the dwell time of a call in a cell. However, in order to remain adaptive to sudden changes in network situations, the control period should not be lengthened indefinitely.

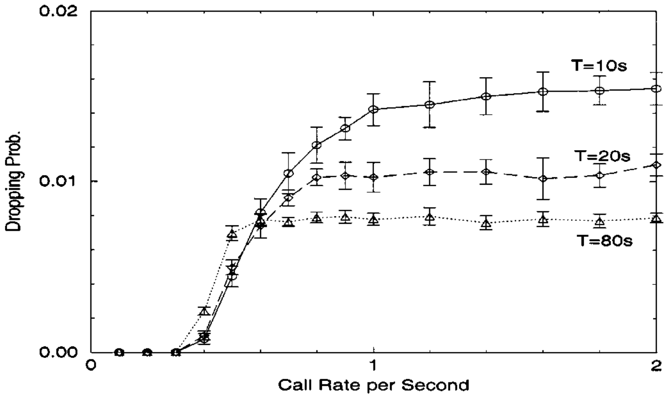

Fig. 9. Dropping probabilities for control periods T =10, 20, 80 s.

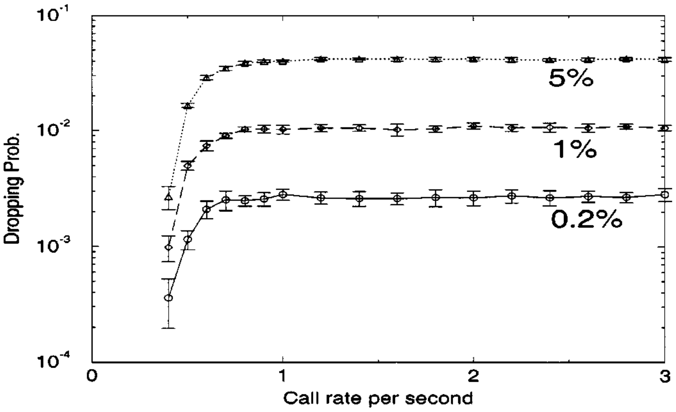

Fig. 10. Dropping probabilities for different QoS requirements.

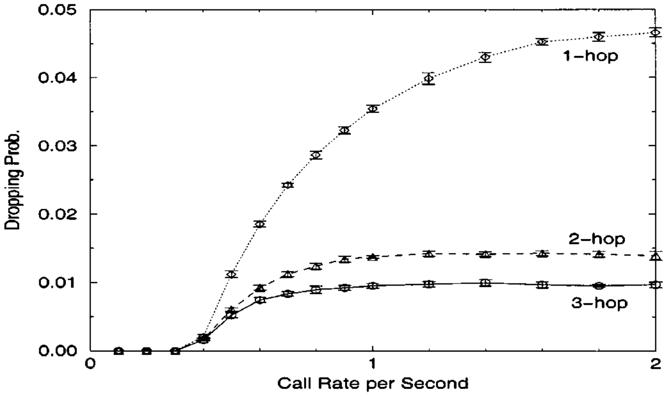

Fig. 11. Dropping probabilities for including up 1, 2, and 3 hops in the local approximation.

7) *Effects of QoS requirement:* As shown in Fig. 10, SDCA can adjust the dropping probability to suit the changes in the QoS requirement, both when P_{QoS} is lowered to 0.2%, or raised to 5%.

8) *Effects of lower order local approximation:* To demonstrate the relevance of the higher order terms used in computing the transition probabilities, we compare in Fig. 11 the results of including up to 1, 2, and 3 hops in the local approximation. Here, we consider $h = 0.05\ \mathrm{s}^{-1}$ and $T = 20$ s so that an average of one

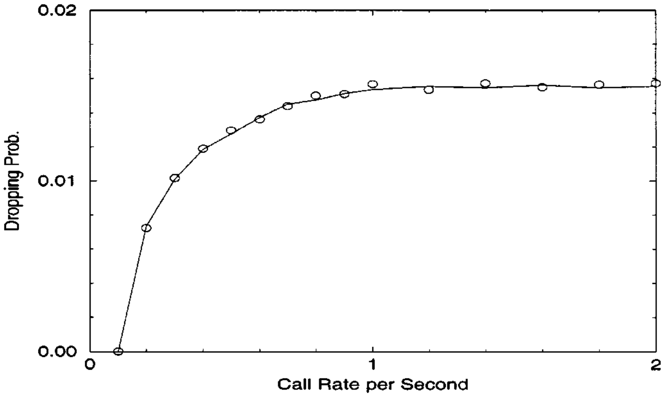

Fig. 12. Dropping probabilities for the mean-field algorithm (symbols) and the measurement-based algorithm (line) in a network with nonuniform traffic.

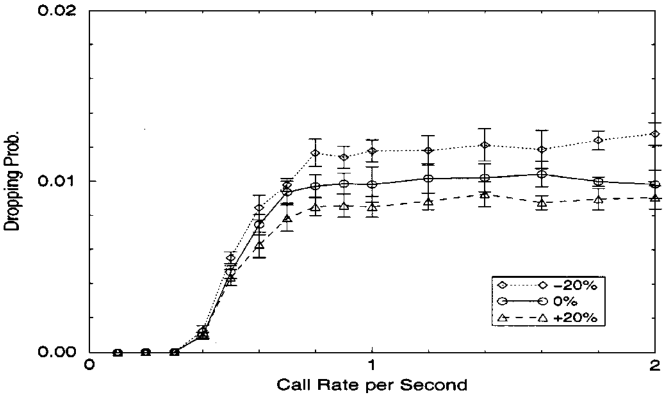

Fig. 13. Dropping probabilities for erroneous measurements of the handoff rate.

handoff event per call takes place in a control period. Since the dropping probabilities for 1 and 2 hops both exceed the requirement of $P_{\mathrm{QoS}} = 0.01$, we conclude that multiple hops are essential to a precise control.

9) *Effects of reducing the distance of information exchange:* The significance of multiple hops is relevant to the issue of reducing the signaling load in the network. We have based the control function on periodic exchange of status information among cells up to third nearest neighbors. Restricting the distance of exchange to nearest neighboring cells will reduce the signaling load, but merely neglecting handoffs from cells beyond will inevitably sacrifice the control precision.

Fig. 12 illustrates a mean-field algorithm which maintains the control precision while restricting the distance of exchange. Here the transition probabilities are still computed up to 3 hops, with the status of the nearest neighboring cells being *measured,* whereas those of the second and third nearest neighbors being *assumed* to take the average values of the nearest neighbors. Using the network with nonuniform traffic in 4), the control performance essentially matches its measurement-based counterpart.

The effectiveness of the mean-field algorithm can be attributed to the fact that fluctuations of the status of the twelve second and third nearest neighbors cancel out, thus improving the precision of estimating their average. Even for networks with nonuniform traffic, the admission control maintains their occupancy and controlled arrival rate at about the same level.

10) *Robustness against measurement errors:* To study the robustness against measurement errors in the handoff rate, we simulate a network whose traffic is characterized by $h = 0.01\ \mathrm{s}^{-1}$, while the control is based on an inaccurate estimation h_{est}. Fig. 13 shows that when h_{est} differs

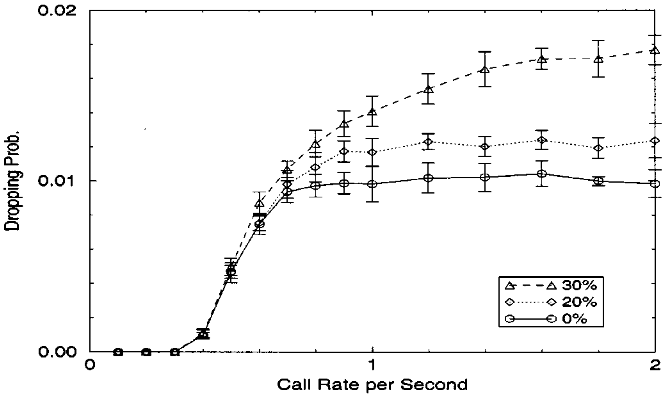

Fig. 14. Dropping probabilities for erroneous measurements of the call arrival rate.

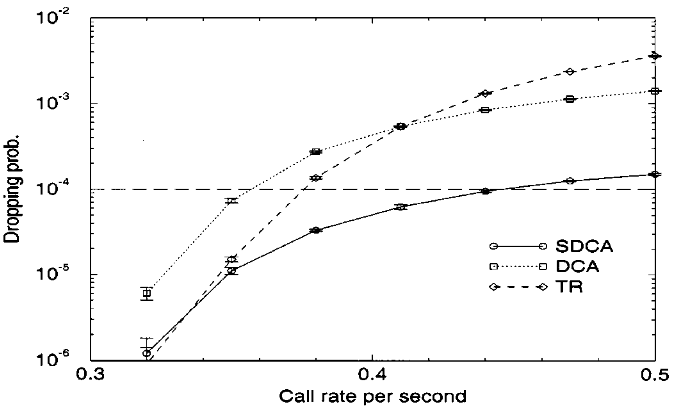

Fig. 15. Dropping probabilities for $P_{\mathrm{QoS}} = 10^{-4}$ when the network starts to saturate. In this and Fig. 16, there are six guard channels in TR.

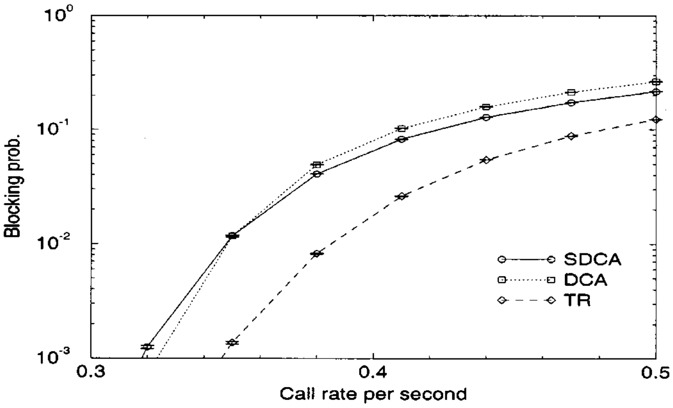

Fig. 16. New call blocking probabilities for $P_{\mathrm{QoS}} = 10^{-4}$ when the network starts to saturate.

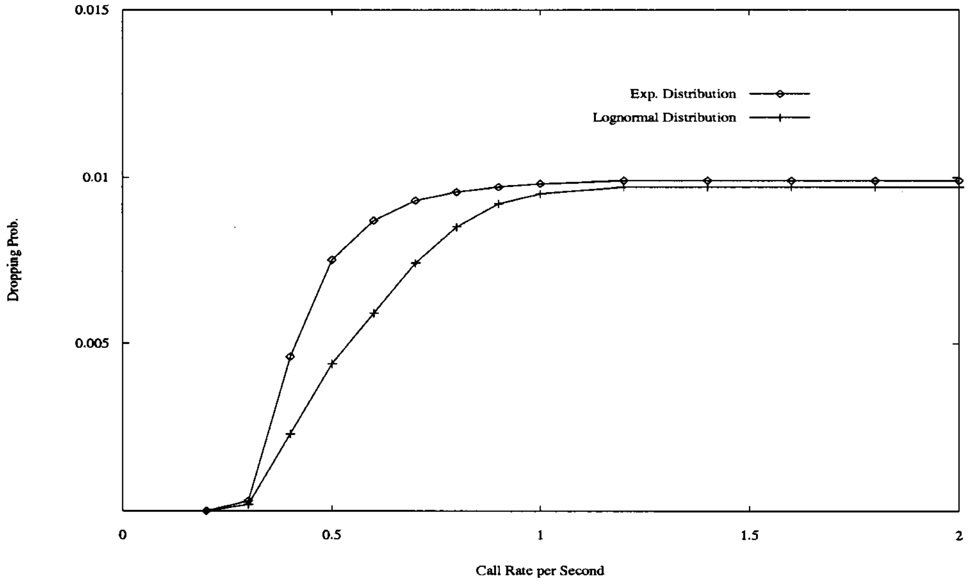

Fig. 17. Handoff dropping probabilities under exponential channel holding time versus under lognormal distribution.

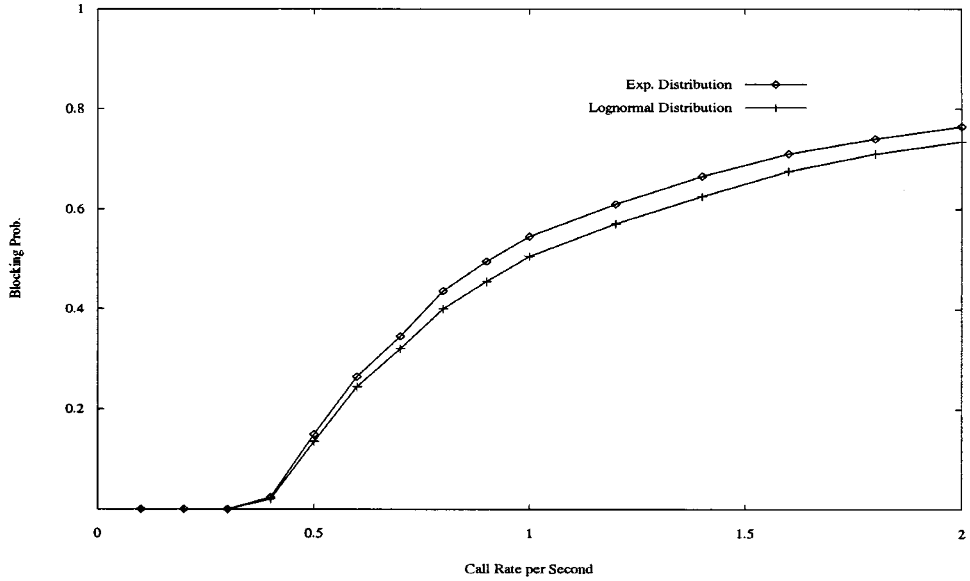

Fig. 18. New call blocking dropping probabilities under exponential channel holding time versus under lognormal distribution.

from h by up to 20%, the dropping probability is still acceptable. For larger differences, the error becomes unacceptable.

Similarly, we study the robustness against measurement errors in the call arrival rate, by simulating a network whose traffic is characterized by λ, while the control is based on an erroneous estimation of $\lambda_{\text{est}} = \lambda(1 + a\epsilon)$ at each control period, ϵ being a Gaussian variable with zero mean and unit variance. Fig. 14 shows that for a up to 20%, the performance is still acceptable.

11) *Normal load:* While the emphasis so far is on the stability of the control under heavily loaded conditions, it is interesting to consider the situation under normal load (and usually with stricter QoS requirements). In Fig. 15, we use $P_{\text{QoS}} = 10^{-4}$ and consider the range of call rates which just start to saturate the network. Again, we see that SDCA satisfies the QoS requirement up to the call rate of 0.44 s^{-1}, much higher than those of DCA and TR. At the same time, Fig. 16 shows that a reasonable level of blocking probability is maintained.

12) *Non-exponential channel holding time:* As stated earlier, one of the key assumptions in our model is the exponential channel holding time, which is necessary for the derivation of the corresponding evolution equations [(10) and (11)]. We now compare the results with those obtained under more realistic assumptions, specifically, the lognormal distribution described in [11]. The same mean value (i.e., the average channel holding time) and the variance are used in both distributions. Observed from Figs. 17 and 18, the exponential channel holding time yields accurate control. This shows that the control algorithm is rather insensitive to this assumption, mainly

because we adopt a periodic control in which the length of the control period is set to be less than the dwell time of a call, and effectively, the exponential distribution is a good approximation in the time interval truncated by the control period.

V. Local Estimation Algorithms

One practical limitation of this algorithm is that it requires the periodical status information changes (signaling overhead) among neighboring cells for the precision of the control, as observed from Fig. 11. This of course depends on the relative magnitude of mobility and the length of the control period. As discussed in Section III-A on the convergence of the local approximation, and confirmed in the result of Fig. 11, such status information exchange is indispensable, unless the probability that a call can handoff more than once in a control period is negligible. Though reducing the length of the control period may alleviate this problem, the frequency of the signaling is increased at the same time.

To overcome this limitation, we propose local estimation algorithms in which the information used by a cell is restricted to those available locally, while the status of the neighboring cells is derived by estimation rather than actual signaling. An exponential smoothing technique from time series analysis is adopted to compute the expected values from the periodically observed values. Such a technique was used in TCP adaptive retransmission to estimate the round-trip time (RTT) [10]. The detailed algorithm based on local control can be found in [16].

Specifically, the mean channel occupancy $\langle n_i(t)\rangle$ and variance $\sigma_i(t)^2$ at time t can be estimated from local information, although their explicit evaluations in (8) and (9) require information from both local and neighboring cells. The key is to note that at the end of a control period, the channel occupancy becomes the initial occupancy of the *next* control period, which is the local information readily available without extra measurements. One can then subtract from it the estimated number of new and ongoing calls which originate from the local cell and survive at the end of the period. The difference yields the number of background calls $n_b(T)$ which originate from neighboring cells. From (8)

$$\langle n_i(T)\rangle = f_{ii}(T)n_{i0} + g_{ii}(T)a_i\lambda_i + n_b(T). \qquad (22)$$

In the last term, $n_b(T)$ is expected to exhibit some long-term statistical behavior provided that the traffic does not change rapidly. Hence its estimated value can be updated at the end of each control period by exponential smoothing via

$$\langle n_b(T)\rangle \leftarrow (1-\epsilon)\langle n_b(T)\rangle + \epsilon[n_i(T) - f_{ii}(T)n_{i0} - g_{ii}(T)a_i\lambda_i]. \qquad (23)$$

The coefficients ϵ used in (23) need to be properly selected to *smooth* the estimated value. In general, a large value of ϵ can keep track of the changes more accurately, but can be too heavily influenced by temporary fluctuations. On the other hand, a small value of ϵ is more stable, but could be too slow in adapting to real traffic changes.

Next, we need to estimate the number of background calls $n_b(t)$ at a general time t during a control period. We expect that this number would increase linearly with time at the beginning of a control period. However, it levels off subsequently because the background calls themselves can handoff or depart. Neglecting this leveling-off effect would inevitably lead to erroneous control.

Without the intercellular information, it is easier to estimate the number of local calls and infer back the number of background calls. This estimation is based on the typical condition of a cell after averaging over different control periods, in which case we can assume:

1) the average occupancy in a cell is $\langle n_{i0}\rangle$, independent of time t;
2) the average number of ongoing local calls at time t is $\langle n_{i0}\rangle f_{ii}(t)$;
3) the average number of new local calls from time 0 to t is $\langle a_i\lambda_i\rangle g_{ii}(t)$;
4) the average number of handout and hand-in calls balance each other, so that the average number of admitted new calls is balanced by the average number of departures only, yielding $\langle a_i\lambda_i\rangle = \langle n_{i0}\rangle\mu$.

Hence the average number of background calls at time t is given by

$$\langle n_b(t)\rangle = [1 - f_{ii}(t) - \mu g_{ii}(t)]\langle n_{i0}\rangle. \qquad (24)$$

Eliminating the average occupancy $\langle n_{i0}\rangle$, $\langle n_b(t)\rangle$ is now related to the measured quantity $\langle n_b(T)\rangle$ via

$$\langle n_b(t)\rangle = \frac{1 - f_{ii}(t) - \mu g_{ii}(t)}{1 - f_{ii}(T) - \mu g_{ii}(T)}\langle n_b(T)\rangle. \qquad (25)$$

Indeed, this result is confirmed by more elaborate analyses. Assuming that all neighboring cells have identically an average initial occupancy of $\langle n_{k0}\rangle$ and an average call arrival rate of $\langle a_k\lambda_k\rangle = \langle n_{k0}\rangle\mu$, we estimate from (8) that

$$\langle n_b(t)\rangle = \left[\sum_{k\neq i} f_{ik}(t)\right]\langle n_{k0}\rangle + \left[\sum_{k\neq i} g_{ik}(t)\right]\langle a_k\lambda_k\rangle. \qquad (26)$$

The sum of the transition probabilities in the first term is the probability that a call survives in *any* neighboring cell, provided that it is still ongoing. Hence we have

$$\begin{aligned}\sum_{k\neq i} f_{ik}(t) &= e^{-\mu t} - f_{ii}(t) \quad \text{and}\\ \sum_{k\neq i} g_{ik}(t) &= \frac{1}{\mu}(1 - e^{-\mu t}) - g_{ii}(t)\end{aligned} \qquad (27)$$

yielding (24) and hence (25).

Thus the mean occupancy at a general time t, to be used in computing the evolving dropping probability, is estimated by

$$\langle n_i(t)\rangle = f_{ii}(t)n_{i0} + g_{ii}(t)a_i\lambda_i + \frac{1 - f_{ii}(t) - \mu g_{ii}(t)}{1 - f_{ii}(T) - \mu g_{ii}(T)}\langle n_b(T)\rangle. \qquad (28)$$

To compute the variance $\sigma_i(t)^2$ of the channel occupancy distribution, we note that in (9), the background calls of cell i

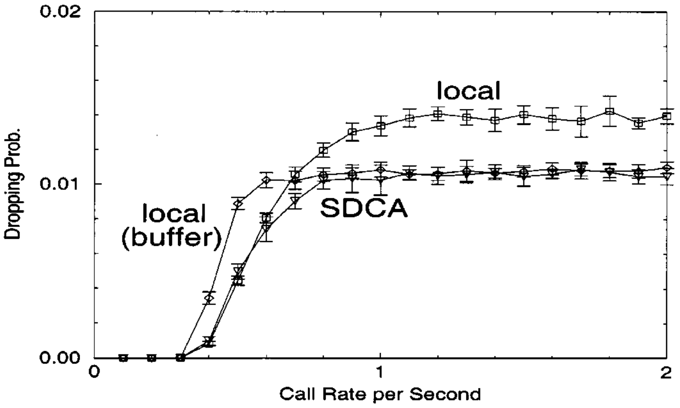

Fig. 19. Handoff dropping probabilities for the local estimation algorithm, with and without the QoS buffer, compared with that of SDCA.

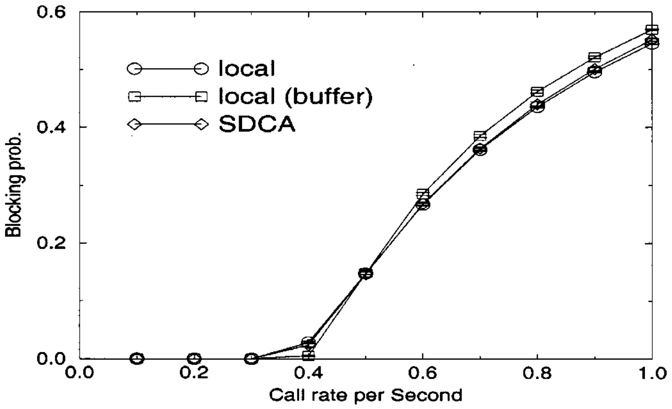

Fig. 20. New call blocking probabilities for the local estimation algorithm, with and without the QoS buffer, compared with that of SDCA.

consist of both the new and ongoing calls originating from the neighboring cells k. The numbers of new calls are Poisson distributed, but the numbers of ongoing calls are binomially distributed with means and variances of $f_{ik}(t)n_{k0}$ and $f_{ik}[1 - f_{ik}(t)]n_{k0}$, respectively. However, for $k \neq i$, $f_{ik}(t)$ remains small within a control period, and both the mean and variance are approximately $f_{ik}(t)n_{k0}$. This implies that the number of background calls is approximately Poisson distributed. Its variance is thus identical to $\langle n_b(t)\rangle$ given in (25). Similar to (28), the variance at a general time t is estimated by

$$\sigma_i(t)^2 = f_{ii}(t)[1 - f_{ii}(t)]n_{i0} + g_{ii}(t)a_i\lambda_i + \frac{1 - f_{ii}(t) - \mu g_{ii}(t)}{1 - f_{ii}(T) - \mu g_{ii}(T)} \langle n_b(T)\rangle. \quad (29)$$

Summarizing, the local estimation algorithm is given by assigning the acceptance ratios a_i according to the solution of (21), where $\tilde{D}_i$ is time averaged by (20) for $D_i(t)$ given in (19), in which the parameters $\langle n_i(t)\rangle$ and $\sigma_i(t)$ are given by (28) and (29), respectively, and $\langle n_b(T)\rangle$ is estimated by exponential smoothing via (23). All information is derived locally, and the only transition probability used is the local survival probability in (4) and its integrated value in (7) for $k = i$.

As shown in Fig. 19, the local estimation algorithm results in stable control, with similar accuracy as the version of SDCA with intercellular communications. To further improve the local estimation algorithm, we introduce the concept of a QoS buffer. When a cell is overloaded, its local dropping probability may be higher than the QoS even when the acceptance ratio is set to 0. In this case, it may be advantageous to reduce the target QoS in the following control periods, so that the QoS averaged over control periods can be still be maintained. Similarly, when the load of a cell is light, it may be advantageous to set a higher target QoS in the following control periods. This mechanism can be implemented by introducing a QoS buffer β_i for cell i. At the beginning of each control period, the acceptance ratio is obtained by solving, instead of (21)

$$\tilde{D}_i = \max(P_{QoS} + \beta_i, 0) \quad (30)$$

followed by an update of the QoS buffer using

$$\beta_i \leftarrow \beta_i + (P_{\mathrm{QoS}} - \tilde{D}_i). \quad (31)$$

As shown in Fig. 19, the local estimation algorithm with a QoS buffer performs comparably with SDCA. Compared with SDCA, its corresponding blocking probability is lower when the network starts to saturate, and only slightly higher when the network is overloaded, as shown in Fig. 20.

VI. Conclusion

In this paper, we present a new distributed and dynamic call admission control scheme. The novelties of the proposed scheme that make the control stable and precise are as follows.

1) We have taken into account the effects of limited capacity and time dependence on the call dropping probability.
2) We have included the nonzero probability of multiple hops from distant cells for longer control periods, which improves the accuracy of the control mechanism.
3) Instead of implementing the control by adjusting the admission threshold, we have computed the acceptance ratio, which is able to spread the new calls uniformly over the control period. In contrast, adjusting the admission threshold tends to block late comers when it increases above the previous value, and early comers when it decreases below the previous value.

The major advantage of SDCA is its insensitivity to the network load. The dropping probability is maintained at a stable level over a wide range of call rates. By comparison, the performance of DCA varies rapidly with the network load, and TR is designed for static control. Though there is a slight deviation from the prescribed P_{QoS} in, say, Figs. 7 and 10, the stability of the algorithm implies that this can be offset easily. We also found that the control algorithm is rather insensitive to the assumptions of exponential holding times, since the length of the control period is set to be less than the dwell time of a call in a cell, observed from Figs. 17 and 18. In situations where the termination and handoff rates change with time, it is possible to further improve the algorithm by using a moving average to estimate them. This approximates the rates as exponential within a control period, whose mean rates change with the moving average from period to period.

The algorithm can operate over a wide range of networks: homogeneous and inhomogeneous call rates and handoff rates, large and small cell capacities, long and short control periods, tight and loose QoS requirements, and erroneous measurements of call rates and handoff rates. Besides being adaptive to changes in the call rates, it also has the potential to be adaptive to changes in the handoff rates using the local approximation for on-line updates. No matrix operations are required.

While the dropping probability is an important parameter for network management, the probability of forced termination may be more relevant to the subscriber. If a mobile moves across K cells during its lifetime, the probability of forced termination is given by $P_f = 1 - (1 - D)^K$ where D is the average dropping probability. Though we have not presented results for the average probability of forced termination, it can be readily obtained by $P_f = hD/(hD + \mu)$.

The major complexity of the algorithm comes from solving (21) by the bisection method. However, since the precision for a stochastic control does not need to be high (here, we use a precision of 0.01 for a_i), this is not a problem for modern computers.

Efforts are in progress to extend the method for call admission control with multiple classes of traffic, each having its own requirements bandwidth, QoS guarantee, and handoff rates. Specifically, we are investigating the call admission control for two types of traffic using a complete partition scheme with a movable boundary [9], and employing the preassignment policy [17].

Appendix A
Transition Probabilities for Uniform Handoff Rates

We note that the transition matrix J defined in Section III-A can be written as

$$J = J_0 + J_1 \tag{32}$$

where

$$\begin{aligned}(J_0)_{ik} &= (h_k + \mu)\delta_{ik} \\ (J_1)_{ik} &= \begin{cases} -h_{ik} & k,\, i = \text{nearest neighbors} \\ 0 & \text{otherwise.}\end{cases}\end{aligned} \tag{33}$$

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

$$\begin{aligned}&[\exp(-Jt)]_{ik} \\ &= \sum_{r=0}^{\infty} \left\{ \frac{(-t)^r}{r!} \cdot [J_0^r + (J_0^{r-1}J_1 + \cdots + J_1 J_0^{r-1}) + \cdots]_{ik} \right\}.\end{aligned} \tag{34}$$

The 0th-order term consists of those terms in (34) which contain no J_1. Hence the 0th-order contribution goes only to $i = k$, with

$$q_0(t) = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} (J_0^r)_{ii} = \exp[-(h+\mu)t]. \tag{35}$$

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

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

$$\begin{aligned}q_1(t) &= \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} \left[(h+\mu)^{r-1} \frac{h}{6} + \cdots + \frac{h}{6} (h+\mu)^{r-1} \right] \\ &= \frac{ht}{6} \exp[-(h+\mu)t].\end{aligned} \tag{36}$$

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

Higher order contributions can be evaluated similarly.

Appendix B
Transition Probabilities for Inhomogeneous Handoff Rates

In the local approximation, the transition probabilities $f_{ik}(t)$ from cell k to i consist of contributions from all possible paths starting from cell k and ending at cell i. Let P be a path of m hops following the sequence $k \to l_1 \to \cdots \to l_{m-1} \to i$. This path makes an mth-order contribution to $f_{ik}(t)$ given by

$$\begin{aligned}q_P(t) = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} &\sum_{s_0 + \cdots s_m = r-m} (h_i + \mu)^{s_m} h_{il_{m-1}} \\ &\cdot \left(h_{l_{m-1}} + \mu\right)^{s_{m-1}} \cdots h_{l_1 k} (h_k + \mu)^{s_0}.\end{aligned} \tag{37}$$

Since the path may visit a cell more than once, we denote the distinct cells along the path by the label j (including k and i), and the number of stops in each cell by m_j. Hence after collecting terms corresponding to the same cell, (37) can be written as

$$\begin{aligned}q_P(t) &= \sum_{r=0}^{I} \infty \frac{(-t)^r}{r!} \cdot \sum_{p_1 + p_2 + \cdots = r-m} \\ &\cdot \left\{ \prod_{j \in P} \left[\sum_{s_1 + \cdots + s_{m_j} = p_j} (h_j + \mu)^{p_j} \right] \cdot h_{il_{m-1}} \cdots h_{l_1 k} \right\}.\end{aligned} \tag{38}$$

Consider the number of terms $N(p, m)$ allowed in the summation $s_1 + \cdots + s_m = p$, where $s_1, s_2, \ldots$ are integers. It is equal to the number of ways of partitioning p indistinguishable objects into m groups (empty groups allowed). Hence $N(p, m) = (p + m - 1)!/p!(m - 1)!$, leading to

$$\sum_{s_1 + \cdots + s_m = p} (h+\mu)^p = \frac{1}{(m-1)!} \frac{\partial^{m-1}}{\partial h^{m-1}} (h+\mu)^{p+m-1}. \tag{39}$$

Substituting into (37), we have

$$\begin{aligned}q_P(t) = \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} &\sum_{\sum_{j \in P} p_j = r-m} \\ &\cdot \prod_{j \in P} \left[\frac{1}{(m_j - 1)!} \frac{\partial^{m_j - 1}}{\partial h_j^{m_j - 1}} (h_j + \mu)^{p_j + m_j - 1} \right] \\ &\cdot h_{il_{m-1}} \cdots h_{l_1 k}.\end{aligned} \tag{40}$$

To simplify this expression, we use two algebraic identities. The first is

$$D(x_1, \cdots, x_r) \equiv \begin{vmatrix} 1 & x_1 & \cdots & x_1^{r-1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & x_r & \cdots & x_r^{r-1} \end{vmatrix} = \prod_{t>s}^{r} (x_t - x_s). \tag{41}$$

Identity (41) can be verified directly by repeatedly applying row transformations to the determinant. The second identity is

$$\sum_{n_1+\cdots+n_r=n} x_1^{n_1} \cdots x_r^{n_r} = \sum_{t=1}^{r} \frac{x_t^{n+r-1}}{\prod\limits_{s\neq t}^{r} (x_t - x_s)}. \tag{42}$$

Identity (42) can be proved by mathematical induction in r. It can be easily verified for $r = 2$. Now assume that it holds for the value r, then for the value $r + 1$, we can write

$$\sum_{n_1+\cdots+n_{r+1}=n} x_1^{n_1} \cdots x_{r+1}^{n_{r+1}} = \sum_{m=0}^{n} \left(\sum_{t=1}^{r} \frac{x_t^{n-m+r-1}}{\prod\limits_{s\neq t}^{r} (x_t - x_s)} \right) x_{r+1}^{m} \tag{43}$$

where $m = n_{r+1}$, and we have applied the theorem for the value r and the condition $n_1 + \cdots + n_r = n - m$. The expression now reduces to a geometric series, and the result is

$$\begin{aligned} &\sum_{n_1+\cdots+n_{r+1}=n} x_1^{n_1} \cdots x_{r+1}^{n_{r+1}} \\ &= \sum_{t=1}^{r} \frac{x_t^{n+r}}{\prod\limits_{s\neq t}^{r+1} (x_t - x_s)} + \sum_{t=1}^{r} \frac{x_{r+1}^{n+1} x_t^{r-1}}{(x_{r+1} - x_t) \prod\limits_{s\neq t}^{r} (x_t - x_s)}. \end{aligned} \tag{44}$$

The first term is just the desired terms up to $t = r$. In the second term, which will be denoted by T_2, we first rearrange the factors appearing in the denominator for a given t as follows:

$$\prod_{s\neq t}^{r} (x_t - x_s) = (-1)^{r-t} \prod_{s\neq t}^{r} (x_> - x_<) \tag{45}$$

where $x_> = x_{\max(s,\,t)}$ and $x_< = x_{\min(s,\,t)}$. The factor $(-)^{r-t}$ appears since we have rearranged $r - t$ terms when $t + 1 \le s \le r$. Multiplying both the denominator and numerator by $\prod_{u>s\neq t}^{r} (x_u - x_s)$, the second term in (44) becomes, on using (41)

$$T_2 = \sum_{t=1}^{r} \frac{(-1)^{r-t} x_{r+1}^{n+1} x_t^{r-1}}{x_{r+1} - x_t} \, \frac{D(x_1, \cdots, x_{t-1}, x_{t+1}, \cdots, x_r)}{D(x_1, \cdots, x_r)}. \tag{46}$$

We can rewrite T_2 as

$$\begin{aligned} T_2 = {} & x_{r+1}^{n+1} \\ & \cdot \begin{vmatrix} 1 & x_1 & \cdots & x_1^{r-2} & x_1^{r-1}/(x_{r+1} - x_1) \\ . & . & & . & . \\ . & . & & . & . \\ . & . & & . & . \\ 1 & x_r & \cdots & x_r^{r-2} & x_r^{r-1}/(x_{r+1} - x_r) \end{vmatrix} \\ & \cdot D(x_1, \cdots, x_r)^{-1}. \end{aligned} \tag{47}$$

This can be verified by expanding the determinant down the last column. Multiplying and dividing the tth row by $x_{r+1} - x_t$, and performing column operations successively, (44) reduces to

$$T_2 = \frac{x_{r+1}^{n+r}}{\prod\limits_{t=1}^{r} (x_{r+1} - x_t)}. \tag{48}$$

Combining with the first term in (44), (42) is proved.

We can now return to (40) which, on using (42), becomes

$$\begin{aligned} q_P(t) = {} & \sum_{r=0}^{\infty} \frac{(-t)^r}{r!} \prod_{j\in P} \frac{1}{(m_j - 1)!} \, \frac{\partial^{m_j - 1}}{\partial h_j^{m_j - 1}} \\ & \cdot \sum_{j\in P} \frac{(h_j + \mu)^{r-m+[m-\sum_j (m_j-1)+1]-1}}{\prod\limits_{t\neq j} (h_j - h_t)} \\ & \cdot h_{il_{m-1}} \cdots h_{l_1 k} \end{aligned} \tag{49}$$

where the number of distinct cells along the path is equal to $m - \sum_j (m_j - 1) + 1$. The summation over r further reduces the expression to

$$\begin{aligned} q_P(t) = {} & (-1)^m \prod_{j\in P} \frac{1}{(m_j - 1)!} \, \frac{\partial^{m_j - 1}}{\partial h_j^{m_j - 1}} \\ & \cdot \sum_{j\in P} \frac{\exp(-(h_j + \mu)t)}{\prod\limits_{t\neq j} (h_j - h_t)(h_j + \mu)^{\sum_j (m_j - 1)}} \\ & \cdot h_{il_{m-1}} \cdots h_{l_1 k}. \end{aligned} \tag{50}$$

APPENDIX C
SOLUTION OF THE DIFFUSION EQUATION

Equation (12) with boundary condition (13) can be solved by Laplace transform, analogous to problems in heat conduction [8]. Let

$$F(x, z) = \int_0^{\infty} dt \, e^{-zt} P(x, t). \tag{51}$$

Using integration by parts and the initial condition (14), (12) can be transformed to

$$\left(z + v \frac{\partial}{\partial x} - D \frac{\partial^2}{\partial x^2} \right) F(x, z) = \delta(x - x_0). \tag{52}$$

Since (52) reduces to a homogeneous linear differential equation for $x < x_0$ and $x > x_0$, $F(x, z)$ can be obtained by piecewise construction, subject to the condition that F is continuous at $x = x_0$. The result is

$$F(x, z) = \begin{cases} Ce^{k_+(x-x_0)} & x < x_0, \\ C\left[(1-r)e^{k_+(x-x_0)} + re^{k_-(x-x_0)}\right] & x_0 < x \le 1 \end{cases} \tag{53}$$

where $k_\pm = (v \pm \sqrt{v^2 + 4Dz})/2D$. C is determined by the continuity condition that $\Delta\partial F/\partial x = -1/D$ at $x = x_0$

$$C = \frac{1}{Dr(k_+ - k_-)} \tag{54}$$

and r is determined by the boundary condition (13)

$$\frac{1-r}{r}\exp[k_+(1-x_0)] = -\frac{v - Dk_-}{v - Dk_+}\exp[k_-(1-x_0)]. \tag{55}$$

We are particularly interested in the solution at $x = 1$. Substituting (54) and (55) into (53), we find

$$F(1, z) = \frac{2}{\sqrt{v^2 + 4Dz} - v}\exp\left[-\frac{\sqrt{v^2 + 4Dz} - v}{2D}(1 - x_0)\right]. \tag{56}$$

The distribution $P(1, t)$ can be obtained by inverse Laplace transform

$$P(1, t) = \int_C \frac{dz}{2\pi i}\, e^{zt} F(1, z) \tag{57}$$

where C runs from $s_0 - i\infty$ to $s_0 + i\infty$ to the right of all singularities of $F(1, z)$.

Case 1, $v > 0$: $F(1, z)$ has a pole at $z = 0$ and a branch cut terminating at $z = -v^2/4D$. Evaluating the contour integral along the branch cut explicitly

$$P(1, t) = \frac{v}{D} + Q(x_0, t) \tag{58}$$

where, after a change of variable to $\lambda = \sqrt{-4Dz - v^2}/2D$

$$\begin{aligned} Q(x_0, t) = &\int_0^\infty \frac{d\lambda}{2\pi}\frac{8D\lambda}{4D^2\lambda^2 + v^2} \\ &\cdot \exp\left[-\frac{t}{4D}(4D^2\lambda^2 + v^2) + \frac{v}{2D}(1 - x_0)\right] \\ &\cdot \left[2D\lambda\cos\lambda(1 - x_0) - v\sin\lambda(1 - x_0)\right]. \end{aligned} \tag{59}$$

Though this is difficult to integrate, we note that solving $\partial Q(x_0, t)/\partial x_0$ is simpler, since

$$\frac{\partial Q(x_0, t)}{\partial x_0} = \frac{\exp\left[-\frac{(1-x_0-vt)^2}{4Dt}\right]}{\sqrt{4\pi Dt}}\,\frac{1 - x_0}{Dt}. \tag{60}$$

Integrating and using $Q(\infty, t) = 0$, we arrive at (15).

Case 2, $v < 0$: In this case, there is no pole at $z = 0$. For the integral along the branch cut, we integrate $\partial Q(x_0, t)/\partial x_0$ and using $Q(-\infty, t) = 0$, we again arrive at (15).

REFERENCES

[1] A. S. Acampora and M. Naghshineh, "An architecture and methodology for mobile-executed handoff in cellular ATM networks," *J. Select. Areas Commun.*, vol. 12, pp. 1365–1375, Oct. 1994.

[2] I. F. Akyildiz, J. McNair, J. Ho, H. Uzunalioglu, and W. Wang, "Mobility management in next-generation wireless systems," *Proc. IEEE*, vol. 87, pp. 1347–1384, Aug. 1999.

[3] F. Barcelo and J. Jordan, "Channel holding time distribution in cellular telephony," in *Proc. Int. Conf. Wireless Communications*, Alta, Canada, July 1997, pp. 125–134.

[4] B. Epstein and M. Schwartz, "Reservation strategies for multimedia traffic in a wireless environment," in *45th IEEE Vehicular Technology Conf. (VTC'95)*, vol. 1, Chicago, IL, July 1995, pp. 165–169.

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

[6] M. Fang and I. Chlamtac, "Teletraffic analysis and mobility Modeling of PCS networks," *IEEE Trans. Commun.*, vol. 47, pp. 1062–1072, July 1999.

[7] M. Fang, private communication, Aug. 2001.

[8] A. L. Fetter and J. D. Walecka, *Theoretical Mechanics of Particles and Continua*. New York: McGraw Hill, 1980.

[9] Y.-R. Haung, Y.-B. Lin, and J. M. Ho, "Performance analysis for voice/data integration on a finite-buffer mobile system," *IEEE Trans. Veh. Technol.*, vol. 49, pp. 367–378, Mar. 2000.

[10] V. Jacobson, "Congestion avoidance and control," in *Proc. ACM SIGCOMM*, Aug. 1988, pp. 314–329.

[11] C. Jedrzycki and V. C. M. Leung, "Probability distribution of channel holding time in cellular telephone systems," in *46th IEEE Vehicular Technology Conf. (VTC'96)*, vol. 1, Atlanta, GA, May 1996, pp. 247–251.

[12] S. M. Jiang, H. K. Tsang, and B. Li, "Subscriber-assisted handoff support in multimedia PCS," *ACM Mobile Computing Commun. Rev.*, vol. 1, no. 3, pp. 29–36, Sept. 1997.

[13] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*. Amsterdam, The Netherlands: North Holland, 1981.

[14] 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 Trans. Networking*, vol. 5, pp. 1–12, Feb. 1997.

[15] B. Li, C. Lin, and S. Chanson, "Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks," *ACM/Baltzer J. Wireless Networks*, vol. 4, no. 4, pp. 279–290, Aug. 1998.

[16] B. Li, L. Yin, K. Y. M. Wong, and S. Wu, "An efficient and adaptive bandwidth allocation scheme for mobile Wireless networks based on intelligent on-line parameter estimations," *ACM/Kluwer J. Wireless Networks*, vol. 7, no. 2, pp. 107–116, Mar./Apr. 2001.

[17] X. Luo, B. Li, I. Thng, Y.-B. Lin, and I. Chlamtac, "A measurement-based preassignment scheme with connection-level QoS support for multi-service mobile networks," *IEEE Trans. Wireless Commun.*, to be published.

[18] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *J. Select. Areas Commun.*, vol. 14, pp. 711–717, May 1996.

[19] E. C. Posner and R. Guerin, "Traffic policies in cellular radio that minimize blocking of handoff calls," in *Proc. 11th ITC*, Kyoto, Japan, Sept. 1985, pp. 294–298.

[20] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[21] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," in *Proc. IEEE INFOCOM'96*, vol. 1, San Francisco, CA, Mar. 1996, pp. 43–50.

[22] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Englewood Cliffs, NJ: Prentice Hall, 1996.

[23] M. Schwartz, "Network management and control issues in multimedia wireless networks," *IEEE Personal Commun.*, vol. 2, pp. 8–16, June 1995.

[24] S. Wu, K. Y. M. Wong, and B. Li, "A new distributed dynamic call admission policy for mobile wireless networks with QoS guarantee," in *Proc. 9th Int. IEEE Symp. Personal, Indoor and Mobile Radio Communications (PIMRC'98)*, vol. 1, Boston, MA, Sept. 1998, pp. 260–264.

Si Wu received 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, all from the Beijing Normal University, Beijing, China.

He worked as a Postdoctoral Research Associate at the Hong Kong University of Science and Technology, the Limburg University Center, Belgium, and the RIKEN Brain Science Institute, Japan, for five years. In 2000, he became a Faculty Member with the Department of Computer Science, Sheffield University, Sheffield, U.K. His research interests include computational neuroscience, machine learning, neural networks, information geometry, and the application of intelligent methods for telecommunication control.

K. Y. Michael Wong received the B.S. degree in physics from the University of Hong Kong in 1978, and the M.S. and Ph.D. degrees in physics from the University of California, Los Angeles, in 1982 and 1986, respectively.

He worked as a Postdoctoral Research Associate with the Imperial College, London, U.K., and the University of Oxford, Oxford, U.K. In 1992, he became a Faculty Member with the the Hong Kong University of Science and Technology, where he is now an Associate Professor in physics. His research interests include stochastic processes and applications in telecommunications, learning theory and neural computation, and complex optimization.

Bo Li (S'89–M'92–SM'99) received the B.S. (*summa cum laude*) and M.S. degrees in computer science from Tsinghua University, Beijing, China, in 1987 and 1989, respectively, and the Ph.D. degree in computer engineering from the University of Massachusetts, Amherst, in 1993.

Between 1994 and 1996, he worked on high-performance routers and ATM switches with IBM Networking System Division, Research Triangle Park, NC. Since then, he has been with the Computer Science Department, Hong Kong University of Science and Technology. He is also an Adjunct Researcher at Microsoft Research, Asia. His current research interests include wireless mobile networking supporting multimedia, video multicast and all optical networks using WDM.

Dr. Li serves and has served on the editorial board for the ACM *Mobile Computing and Communications Review* , the ACM/Kluwer *Journal of Wireless Networks*, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS —Wireless Communication Series (to be named as IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS), IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the SPIE/Kluwer *Optical Networking Magazine*, and the KICS/IEEE *Journal of Communications and Networks*. He served as a Guest Editor for the IEEE *Communications Magazine* Special Issue on Active, Programmable, and Mobile Code Networking (April 2000), the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Protocols for Next Generation Optical WDM Networks (October 2000), the ACM *Performance Evaluation Review* Special Issue on Mobile Computing (December 2000), and the SPIE/Kluwer *Optical Networks Magazine* Special Issue on Wavelength Routed Networks: Architecture, Protocols and Experiments (January/February 2002). In addition, he has been involved in organizing over 30 conferences, in particular, IEEE INFOCOM since 1986.