Dynamic Resource Allocation Schemes During Handoff for Mobile Multimedia Wireless Networks

Parameswaran Ramanathan, *Member, IEEE*, Krishna M. Sivalingam, *Member, IEEE*, Prathima Agrawal, *Fellow, IEEE*, and Shalinee Kishore

***Abstract*—User mobility management is one of the important components of mobile multimedia systems. In a cell-based network, a mobile should be able to seamlessly obtain transmission resources after handoff to a new base station. This is essential for both service continuity and quality of service assurance. In this paper, we present strategies for accommodating continuous service to mobile users through estimating resource requirements of potential handoff connections. A diverse mix of heterogeneous traffic with diverse resource requirements is considered. We investigate static and dynamic resource allocation schemes. The dynamic scheme probabilistically estimates the potential number of connections that will be handed off from neighboring cells, for each class of traffic. The performance of these strategies in terms of connection blocking probabilities for handoff and local new connection requests are evaluated. The performance is also compared to a scheme previously proposed in [15]. The results indicate that using dynamic estimation and allocation, we can significantly reduce the dropping probability for handoff connections.**

***Index Terms*—Dynamic schemes, handoffs, resource allocation, wireless and mobile networks.**

I. Introduction

A wireless network is typically organized into geographical regions called cells [12]. The mobile users in a cell are served by a base station. Before a mobile user can communicate with other user(s) in the network, a connection must usually be established between the users. The establishment and maintenance of a connection in a wireless network is the responsibility of the base station. To establish a connection, a mobile user must first specify its traffic characteristics and quality of service (QoS) needs. This specification may be either implicit or explicit depending on the type of services provided by the network. For example, in a cellular phone network, the traffic characteristics and the QoS needs of voice connections are known *a priori* to the base station, and therefore, they are usually implicit in a connection request.

Manuscript received March 25, 1999. This work was supported in part by Air Force of Scientific Research Grants F-49620-97-1-0471 and F-49620-99-1-0125, National Science Foundation Grant MIP-9526761, and AT&T Labs, Whippany, NJ.

P. Ramanathan is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706 USA (e-mail: parmesh@ece.wisc.edu).

K. M. Sivalingam is with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164 USA.

P. Agrawal is with Telcordia Technologies, Morristown, NJ 07960 USA.

S. Kishore is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Publisher Item Identifier S 0733-8716(99)06109-0.

Future wireless networks, however, will have to provide support for multimedia services where the traffic characteristics and the QoS needs of a connection may not be known *a priori* to the base station. In this case, the mobile user must explicitly specify the traffic characteristics and QoS needs as part of the connection request. Wireless ATM networks are an example of such a network [10], [8], [2]. In either case, the base station determines whether it can meet the requested QoS needs and, if possible, establish a connection.

When a user moves from one cell to another, the base station in the new cell must take responsibility for all the previously established connections. A significant part of this responsibility involves allocating sufficient resources in the cell to maintain the QoS needs of the established connection(s). If sufficient resources are not allocated, the QoS needs may not be met, which in turn may result in premature termination of the connection. Since premature termination of established connections is usually more objectionable than rejection of a new connection request, it is widely believed that a wireless network must give higher priority to the handoff connection requests as compared to new connection requests. Many different admission control strategies have been discussed in literature to provide priorities to handoff requests without significantly jeopardizing the new connection requests [1], [4], [7], [11], [14], [15].

The basic idea of these admission control strategies is to *a priori* reserve resources in each cell to deal with handoff requests. In conventional cellular networks, where the traffic and QoS needs of all connections are the same, the reservation of resources typically occurs in the form of "guard channels," where a new connection request is established if and only if the total available channels or capacity is greater than a predetermined threshold [4], [7], [11], [14], [15]. The strategies differ in how the number of guard channels (i.e., the threshold) is chosen by a base station.

One simple strategy is to reserve a fixed percentage of the base station's capacity for handoff connections. If this percentage is high, adequate capacity will most likely be available to maintain the QoS needs of handoff connections, but, at the expense of rejecting new connections. The advantage of this strategy is, of course, its simplicity because there is no need for exchange of control information between the base stations. A more involved, but possibly better, strategy is for each base station to dynamically adapt the capacity reserved for dealing with handoff requests based on the current connections in the neighboring cells. This will enable the base station to

0733–8716/99$10.00 © 1999 IEEE

approximately reserve the actual resources needed for handoff requests and thereby accept more new connection requests as compared to a fixed scheme. Such dynamic strategies are proposed and evaluated in [11] and [15].

In [11], Naghshineh and Schwartz develop a theoretical model to compute the resource requirements for handoff requests so as to maintain a target handoff blocking probability. This is the probability of not having adequate capacity to allocate to a handoff request. Their model assumes that all connection requests are identical and the analysis is carried out for a simple three cell configuration under stationary traffic conditions. In [15], Yu and Leung also propose a technique to compute the capacity to be reserved for handoff requests so as to either strictly or loosely maintain the handoff blocking probability within a specified target. They also simulate a more realistic multicell wireless network and compare the performance of their strategy with that of a static strategy. To estimate the future probability of blocking, they assume Poisson arrival of new connection requests, Poisson arrival of handoff connection requests, exponential connection duration, and exponential channel holding time. Note that channel holding time for a connection in a cell depends on the unencumbered cell residence time (i.e., cell residence time if the connection is of an infinite duration) and the remaining connection duration. In practice, unencumbered cell residence time may not be exponentially distributed [16], in which case, the strategy proposed in [15] will not be theoretically valid. Also, as in [11], Yu and Leung's model assumes that all connection requests are identical, which is not valid if multimedia services are to be supported by the wireless network.

In contrast, in this paper, we consider a wireless network supporting diverse traffic characteristics of voice, data, and video applications. Since the connections can now differ in the amount of resources (say bandwidth) required to meet their QoS needs, the question is how should a base station dynamically adapt the amount of resources reserved for dealing with handoff requests. The strategy proposed in this paper is an approximation of the ideal strategy described below.

Consider an ideal wireless network in which each base station knows the exact arrival times and resource requirements of all future handoff requests and the completion times and the cell residence times of connections presently in its cell. Now suppose a new connection request comes into a base station at time t and let T be the amount of time this connection will spend in the cell. Further suppose that the objective is to accept all handoff requests. Then, the base station must accept the new connection request if and only if the additional resources needed to accept all incoming handoff requests in the interval $(t, t+T)$ plus the resources needed to support the new request is less than the amount of resources available at time t. This strategy is ideal because a base station can only estimate the arrival times of handoff requests, the resource requirements of handoff requests, and the residence time of connections in the cell. Therefore, in the proposed approach, a base station first estimates T, the expected cell residence time of the new connection request and the expected maximum additional resources needed to accept all incoming handoff requests in the interval $(t, t+T)$. If the estimated maximum additional resources needed to deal with handoffs plus the resources needed to support the new connection requests is less than the resource available at time t, then the new request is accepted.

The blocking probabilities for handoff and new connection requests in the proposed strategy are evaluated using a discrete-event simulator of a cellular network in a metropolitan area. The simulator also implements an extended version of the strategy proposed in [15] and two static schemes. The strategy in [15] is extended to deal with connection requests with different traffic characteristics. A comparison of the blocking probabilities show that the handoff blocking probability is among the smallest for the proposed scheme in different network types and traffic scenarios. The traffic scenarios simulated include the morning and evening rush-hour situations. The simulation also shows that an extended version of the strategy in [15] does not always perform better than a static scheme when connections with diverse traffic requirements are present.

The rest of the paper is organized as follows. Section II presents our assumed model of the wireless network and reviews details of related strategies from literature. Section III describes the proposed strategy. Section IV presents results of an empirical evaluation of the proposed strategy. Section V provides the summary and conclusions.

II. System Model and Related Work

A base station in a cellular network may receive new connection requests from mobile users within its cell as well as handoff requests from mobile users in the neighboring cells. As part of a connection request, a mobile user promises to adhere to certain traffic characteristics and in return seeks some QoS guarantees from the network. The connections may differ in the traffic characteristics (constant bit rate, variable bit rate) and the desired QoS guarantees (e.g., delay bound, loss bound, throughput). In this paper, we assume that the promised traffic characteristics and the desired QoS guarantees can together be represented by a single number called the effective bandwidth of the connection.

Techniques for computing the effective bandwidth for different traffic characteristics and QoS requirements have been discussed elsewhere in literature [5], [6], [9], [3], and is not the focus of this paper. For example, given a traffic envelope (i.e., a bound on the number of bytes generated by the user in any given time interval) and a desired delay bound, Le Boudec discusses an approach for computing the effective bandwidth which completely characterizes the envelope and the delay requirement [3]. Similarly, given stochastic characteristics of the traffic, the buffer size at a network element, and a desired bound on probability of packet loss, many different techniques have been proposed to compute the equivalent effective bandwidth [5], [6], [9].

Given the effective bandwidths of all the active connections in a cell and the effective bandwidth of a new connection request, the QoS requirements of all connections can be guaranteed if the sum of the effective bandwidths including

```
Admission Control
    If incoming request belongs to class τ
    If incoming request is a handoff then
        If available bandwidth > Δ_{h,τ} + φ_τ then
            Accept
        Else
            Reject
        End
    Else   /* it is a new connection request */
        If available bandwidth > Δ_{n,τ} + φ_τ then
            Accept
        Else
            Reject
        End
    End
```

Fig. 1. Admission control scheme for multimedia connections.

the new request is less than or equal to the capacity of the cell. If this simple admission control criterion is used to accept both new and handoff connection requests, then the blocking probability for both types of requests will be the same. Since it is desirable to have smaller blocking probabilities for handoff requests, however, the proposed strategy is based on the admission control scheme shown in Fig. 1.

In Fig. 1 and in the rest of this paper, we assume that the connection requests in the network belong to one of M diverse classes. The classes correspond to different multimedia applications like voice, data, and video which are expected to run on future wireless networks. From the point of view of the wireless network, each class τ is represented by its effective bandwidth ϕ_τ. For admission control, we associate two guard thresholds $\Delta_{h,\tau}$ and $\Delta_{n,\tau}$ with each traffic class τ. A cell accepts an incoming handoff request of class τ if and only if the available bandwidth in that cell is greater than $\Delta_{h,\tau}$ plus the bandwidth of the connection. Otherwise, the handoff request is rejected and the connection is prematurely terminated. Similarly, a request for a new connection in a cell is accepted if and only if the available bandwidth in the cell is greater than $\Delta_{n,\tau}$ plus the bandwidth of the connection. Otherwise, the new connection request is rejected. Since premature termination of an ongoing connection is usually more undesirable than rejection of a new connection request, $\Delta_{n,\tau} \geq \Delta_{h,\tau}$ for each traffic class τ.

The challenge is how to select the values of the guard thresholds such that most, if not all, handoff requests are accepted without significantly jeopardizing the probability of acceptance of a new request. In Section III, we propose a strategy for selecting the values of the guard thresholds. Other strategies have been discussed in the literature. Before describing our strategy we briefly describe three different strategies from the literature. We refer to these strategies as Fixed, Static, and YL97. A comparison of the performance of our strategy relative to these strategies is given in Section IV.

A. Fixed (f) Strategy

In this strategy, each base station sets aside f% of its capacity for dealing with handoff requests. This is achieved by choosing the guard threshold values to be f% of the cell's capacity. Specifically, if $\Gamma_{\mathtt{c}}$ is the capacity of cell c, then the base station in c selects $\Delta_{h,\tau} = 0$ and $\Delta_{n,\tau} = f \cdot \Gamma_{\mathtt{c}}$ for each traffic class τ.

B. Static (k) Strategy

The key limitation of the $\mathtt{fixed}(f)$ strategy is that the threshold values are not directly based on the effective bandwidths of the connection requests. The $\mathtt{static}(k)$ strategy, on the other hand, is cognizant of the effective bandwidths of the handoff requests.

In this strategy, the base station is assumed to be aware of the steady fraction of connection requests for each traffic class τ. This fraction may be determined from historic traffic information available to the base station. Let p_τ denote the fraction of connection requests for class τ. Then, the expected effective bandwidth for a handoff request is $\Sigma_{i=1}^{M}\ p_i\phi_i$. In $\mathtt{static}(k)$ strategy, each base station selects $\delta_{h,\tau} = 0$ and $\Delta_{n,\tau} = k \cdot \Sigma_{i=1}^{M}\ p_i\phi_i$ for each traffic class τ.

Note that, if all connection requests are identical, then this strategy is equivalent to selecting k guard channels.

C. YL97 Strategy

This strategy is based on the scheme presented in [15]. For comparison to the proposed strategy, this strategy has been modified slightly to deal with M classes of traffic. We first give an overview of the strategy proposed in [15] and then discuss our extension to deal with multiple traffic classes.

In this strategy, each base station dynamically adapts the guard threshold values based on current estimates of the rate at which mobiles in the neighboring cells are likely to incur a handoff into this cell. The objective of the adaptation algorithm is to maintain a target block probability for handoff requests, despite temporal fluctuations in the connection request rate into the cell.

The determination of the guard threshold values is based on an analytic model which relates the guard threshold values to the blocking probabilities for handoff and new connection requests. This model requires the following key assumptions.[1]

1) The arrival of new connection requests in a cell forms a Poisson process.
2) The arrival of handoff requests in a cell forms a Poisson process.
3) The time spent by a connection in a cell is exponentially distributed.
4) The change in arrival rates is moderate in the sense that the network reaches steady state between any two changes in the arrival rate.

In this strategy, each base station periodically queries neighboring base stations and computes an estimate of the rate at which handoff connection requests are expected to arrive in the next update period. This estimate is derived from known stochastics of the connection duration times, cell residence times, and mobility patterns. The arrival of new connection requests is also estimated based on local measurements. Using

[1] These assumptions are not required for the strategy proposed in this paper.

```
Extended YL97 Strategy
    Let N_{T,i} = Γ_c/φ_1.
    For i = 1 to M do
        N_{G,i} = f(N_{T,i}, λ_{n,i}, λ_{h,i}, μ_i, B_h, B_n);
        N_{T,i+1} = (N_{T,i} − N_{G,i})φ_i;
    Endfor
    Δ_{h,τ} = 0 for all 1 ≤ τ ≤ M
    Δ_{n,τ} = Σ_{i=1}^{M} N_{G_i} · φ_i for all 1 ≤ τ ≤ M.
End.
```

Fig. 2. Extended version of YL97 scheme to deal with multiple traffic classes.

expressions from queueing analysis, the base station can then estimate the blocking probabilities for handoff and new connection requests as a function of the number of guard channels. From this function, the base station computes the minimum number of guard channels required to meet the target blocking probabilities for handoff requests.

For comparison to the proposed strategy, we extend this strategy to deal with multiple traffic classes. To explain this extension, we need the following notations. Consider a typical cell c. Let $\lambda_{n,\tau}$ and $\lambda_{h,\tau}$, respectively, be the estimated arrival rate of new and handoff connections of class τ in cell c for the next update period. Let μ_τ be the estimated departure rate of class τ connections in cell c. Also let B_h and B_n denote the target blocking probabilities for handoff and new connection requests and let Γ_{cell} be the total capacity of the cell. Furthermore, without loss of generality, assume that the effective bandwidths are such that $\phi_1 > \phi_2 > \cdots > \phi_M$. Fig. 2 shows a pseudocode of the extended version of the YL97 scheme. In this pseudocode, the function $f(\ )$ computes the minimum number of guard channels required to achieve the target handoff blocking probability exactly as in [15].

III. Dynamic ExpectedMax Strategy

Consider a typical cell c. Let t be the time of arrival of a new connection request in cell c. At time t, the base station in cell c sends a query to the base stations in the neighboring cells requesting the information required to compute the guard threshold values. Once the guard threshold values are computed, the admission control scheme described in Fig. 1 is used to determine whether or not to establish the new connection. Presented below is a formal description of the scheme used to compute the guard threshold values.

Ideally, the update of the guard threshold values in the proposed strategy must occur in a cell upon arrival of each new connection request. Because of the associated communication and control overhead, however, it may not be possible in practice to update the threshold values so frequently. Therefore, in practice, base stations may update the guard threshold values once every $K \geq 1$ new connection requests, where K is a design parameter. Larger values of K means less overhead. Since larger K means that the updates will be performed less frequently, however, the performance of the proposed strategy may worsen as compared to the ideal strategy. The effect of the value of K on the performance of the proposed strategy is evaluated using a discrete-event simulator and the detailed results of this evaluation are shown in Section IV. The results basically show that impact on the performance is very small. Therefore, for ease of understanding, in the description of the proposed strategy, we assume that the update is performed upon arrival of each new connection request.

If accepted, let d be the expected duration of the new connection in cell c. Note that the connection will leave cell c either due to completion or due to a handoff out of the cell. Therefore, the expected duration of the new connection in the cell can be estimated based on known stochastics of the unencumbered completion and cell residence times of connections. A technique for estimating the value of d is discussed later in this section. For now, assume that the value of d is known. Let m_τ be the number of class τ connections in cell c which are expected to either complete or incur a handoff out of the cell in the time interval $(t, t+d]$. Likewise, let n_τ be the expected number of connections of class τ in the neighboring cells which will incur a handoff into cell c in the time interval $(t, t+d]$. In practice, the values of m_τ and n_τ must be estimated by the cell and can therefore be inaccurate. For now, however, assume that their values are known exactly. After describing the basic idea of the proposed strategy, we describe a method for estimating the values of m_τ and n_τ.

Define an outgoing τ-event (denoted by O) to be either a completion of a class τ connection or a handoff of a class τ connection from cell c. Similarly, define an incoming τ-event (denoted by I) to be a handoff of a class τ connection into cell c. Note that, a request for a new connection of class τ is not considered an incoming τ-event. This is because the threshold $\Delta_{n,\tau}$ for accepting a new connection request is set based on the expected bandwidth required to deal with the handoff requests; therefore, the computation of $\Delta_{n,\tau}$ depends only on handoffs and completions. Now consider the sequence of events which occur in cell c in the interval $(t, t+d]$. From the definition of m_τ and n_τ, we know that there will be $(m_\tau + n_\tau)$ events in this interval. Let $s \equiv a_1 a_2 \cdots a_{(m_\tau+n_\tau)}$ denote this sequence of events, where

$$a_l = \begin{cases} O, & \text{if } l\text{th event is completion or outgoing handoff} \\ I, & \text{if } l\text{th event is an incoming handoff} \end{cases}$$

for $1 \leq l \leq (m_\tau + n_\tau)$. Furthermore, given s, let $X_{\tau,k}(s)$ denote the net change in the bandwidth allocated to class τ connections from time t to the end of kth event in s. More formally, assuming $X_{\tau,0}(s) = 0$

$$X_{\tau,k}(s) = \begin{cases} X_{\tau,k-1}(s) - \phi_\tau, & \text{if } a_k \equiv O \\ X_{\tau,k-1}(s) + \phi_\tau, & \text{otherwise} \end{cases}$$

for $1 \leq k \leq (m_\tau + n_\tau)$. Define $Y_\tau(s) = \max\{X_{\tau,k}(s)\colon 0 \leq k \leq (m_\tau+n_\tau)\}$. Informally, $Y_\tau(s)$ is the maximum net change in the bandwidth allocated to class τ connections in $(t, t+d]$. Define $S(m_\tau, n_\tau)$ to be the set of all possible sequences of τ-events in $(t, t+d]$, i.e., $S(m_\tau, n_\tau)$ contains sequences of length $S(m_\tau+n_\tau)$ where each element in the sequence belongs to the set $\{I, O\}$ such that there are exactly n_τ I's in each sequence. More formally

$$S(m_\tau, n_\tau) = \{a_1 \cdots a_{(m_\tau+n_\tau)}\colon \{j\colon a_j = I\}| = n_\tau\}.$$

```
ExpectedMax
    Δ_{h,τ} = 0 for all τ
    Δ_{n,τ} = Σ_{j=1}^{M} Ȳ_j for all τ
```

Fig. 3. The proposed `ExpectedMax` strategy.

Since all sequences in $S(m_\tau, n_\tau)$ are equally likely to occur, the probability of occurrence of any particular sequence s is

$$p_\tau(s) = \frac{1}{|S(m_\tau, n_\tau)|}$$

and the expected value of $Y_\tau(s)$ is

$$\overline{Y}_\tau = \sum_{s \in S(m_\tau, n_\tau)} Y_\tau(s) p_\tau(s).$$

Intuitively, $\overline{Y}_\tau$ is the expected maximum net bandwidth that will be needed to deal with class τ handoff connections in the next update period. Therefore, $\Sigma_{\tau=1}^{M} \overline{Y}_\tau$ is the expected maximum net bandwidth to deal with all handoff connections in the next update period. By setting $\Delta_{n,\tau} = \Sigma_{\tau=1}^{M} \overline{Y}_\tau$, this amount of bandwidth is effectively reserved for dealing with the incoming handoff requests in the interval $(t, t+d]$. This is the approach used in `ExpectedMax` strategy (see Fig. 3). This idea is further clarified by the following example.

Example 1: Suppose at time $t, m_\tau = 2$ and $n_\tau = 3$. Then

$$S(2,3) = \{IIIOO, IIOIO, IIOOI, IOIIO, IOIOI, OIIIO, OIIOI, IIOOI, IOOII, OOIII\}.$$

Then

$$\begin{aligned} X_{\tau,0}(IIOIO) &= 0 \\ X_{\tau,1}(IIOIO) &= \phi_\tau \\ X_{\tau,2}(IIOIO) &= 2\phi_\tau \\ X_{\tau,3}(IIOIO) &= \phi_\tau \\ X_{\tau,4}(IIOIO) &= 2\phi_\tau \\ X_{\tau,5}(IIOIO) &= \phi_\tau \end{aligned}$$

and $Y_\tau(IIOIO) = 2\phi_\tau$. Intuitively, it means that if $Y_\tau(IIOIO)$ is reserved for dealing with incoming handoff requests and the actual sequence of events happens to be IIOIO, then all the incoming handoff requests can be accepted. IIOIO, however, is only one out of the ten possible sequence of events and the bandwidth that will be required to accept all handoff requests will differ depending on the actual sequence of events. Table I shows the values of $Y_\tau(s)$ for all $s \in S(2,3)$ and the corresponding probability of occurrence of that s. From this table, it follows that the probability of the cell needing $+3\phi_\tau$ additional bandwidth to accept all incoming handoff requests is 0.1. Similarly, the probability of the cell needing $2\phi_\tau$ additional bandwidth is 0.5, and the probability of the cell needing ϕ_τ additional bandwidth is 0.4. If the base station in the cell assumes that $3\phi_\tau$ bandwidth will be needed, then all incoming handoff requests can be accepted irrespective of the actual sequence. In most cases, however, it will be overallocating for handoff, because the probability of requiring $3\phi_\tau$ is only 0.1. Therefore, in the `ExpectedMax` strategy, the base station assumes that it will only need the expected value of $Y_\tau(s)$ instead of the maximum value of $Y_\tau(s)$. Note that, as a result, all incoming handoff requests are not guaranteed to be accepted. Since $Y_\tau(s)$ is the maximum net bandwidth required if s is the actual sequence of events, reserving the expected value of $Y_\tau(s)$ will result in accepting most handoff requests. In this example, the base station will assume that the bandwidth required to deal with handoffs is $3\phi_\tau \times 0.1 + 2\phi_\tau \times 0.5 + \phi_\tau \times 0.4 = 1.7\phi_\tau$. ■

TABLE I
VALUES OF $Y_\tau(s)$ FOR $s \in S(2,3)$

$s \in S(2,3)$	$Y_\tau(s)$	$p_\tau(s)$
IIIOO	$+3\phi_\tau$	0.1
IIOIO	$+2\phi_\tau$	0.1
IIOOI	$+2\phi_\tau$	0.1
IOIIO	$+2\phi_\tau$	0.1
IOIOI	$+\phi_\tau$	0.1
OIIIO	$+2\phi_\tau$	0.1
OIIOI	$+\phi_\tau$	0.1
IIOOI	$+2\phi_\tau$	0.1
IOOII	$+\phi_\tau$	0.1
OOIII	$+\phi_\tau$	0.1

A. Modification to `ExpectedMax` Strategy for Fairness

In the `ExpectedMax` strategy as described above, the handoff and new connection blocking probabilities will not be the same for the different classes of traffic. More specifically, classes with higher effective bandwidth will have higher handoff and new connection blocking probabilities as compared to those with smaller effective bandwidths. In some situations, it may be desirable to have comparable blocking probabilities irrespective of their effective bandwidths. To achieve fairness in blocking probabilities among all traffic classes, the guard threshold value for each class τ should be chosen as follows:

$$\begin{aligned} \Delta_{h,\tau} &= \phi_{\max} - \phi_\tau \\ \Delta_{n,\tau} &= \phi_{\max} - \sum_{i=1}^{M} \overline{Y}_\tau \end{aligned}$$

where $\phi_{\max} = \max_{1 \le i \le M} \phi_i$.

The basic idea of this modification is to accept connection requests with smaller effective bandwidth if and only if the cell can accept a connection with the largest effective bandwidth. As a result, there will be an increase in the blocking probabilities for some traffic classes (especially, the ones with small effective bandwidth needs) and a decrease in the blocking probabilities of other traffic classes. The end result is that all classes will have comparable handoff blocking probabilities and comparable new connection blocking probabilities. Since, however, the guard threshold values for new connection requests includes the term $\Sigma_{i=1}^{M} \overline{Y}_\tau$, the blocking probabilities for handoff requests will be larger than that for new connection requests.

B. Computational Issues in `ExpectedMax` Strategy

1) Estimation of d: As described earlier, a connection leaves a cell either because it completes or because it incurs

a handoff out of the cell. Therefore, the expected time spent by a connection in a cell can be derived from the probability distribution functions of the duration of class τ connection and the unencumbered cell residence times (i.e., residence time in a cell if the connection is of an infinite duration). Let $F_\tau(t)$ denote the probability distribution function of the duration of class τ connection. Also, let $R_\tau(t)$ denote the probability distribution function of the unencumbered cell residence time of a class τ connection. Let $r_\tau(t)$ denote the probability density function corresponding to $R_\tau(t)$.

Then, the probability distribution of the time spent by a connection in a cell is $1-(1-F_\tau(t))(1-R_\tau(t))$. The value of d can be estimated to be the expected value of the time spent by a connection in a cell computed from this probability distribution function.

Note that, for the special case when the connection duration time is exponentially distributed and the unencumbered cell residence time is exponentially distributed, i.e.,

$$F_\tau(t) = 1 - e^{-\mu t}$$
$$R_\tau(t) = 1 - e^{-\gamma t}$$

the probability distribution function of the time spent by a class τ connection in a cell is $1-e^{-(\mu+\gamma)t}$. Therefore, the value of d can be estimated as $1.0/(\mu+\gamma)$.

2) Estimation of m_τ and n_τ: Consider a connection of class τ which started at time u, entered the cell at time v, and is active at time t. Clearly, $u \le v \le t$. Then, the probability that such a connection will incur a handoff in the interval $(t, t+d]$ can be shown to be

$$\begin{aligned} H_\tau(u,v,t,d) &= \int_{w=t}^{t+d} \frac{\mathrm{P}[\text{cell residence time } = w-v]}{\mathrm{P}[\text{cell residence time} > t-v]} \\ &\quad \cdot \frac{\mathrm{P}[\text{conn. duration} > w-u]}{\mathrm{P}[\text{conn. duration} > t-u]}\, dw \\ &= \int_{w=t}^{t+d} \frac{r_\tau(w-v)}{1-R_\tau(t-v)} \cdot \frac{1-F_\tau(w-u)}{1-F_\tau(t-u)}\, dw \end{aligned}$$

if $u \le v \le t$ and zero otherwise. Likewise, the conditional probability that a connection of class τ will neither incur a handoff in the interval $(t, t+d]$ nor complete in the interval $(t, t+d]$ given that it started at time u, entered the cell at time v, and is active at time t is

$$\begin{aligned} &\overline{HC}_\tau(u,v,t,d) \\ &\quad = \frac{\mathrm{P}[\text{conn. dur.} > t+d-u,\ \text{cell res. time} > t+d-v]}{\mathrm{P}[\text{conn. dur.} > t-u,\ \text{cell res. time} > t-v]} \\ &\quad = \frac{F_\tau(t+d-u)R_\tau(t+d-v)}{F_\tau(t-u)R_\tau(t-v)} \end{aligned}$$

if $u \le v \le t$ and zero otherwise. Finally, the conditional probability that a connection of class τ will either complete or incur a handoff in the interval $(t, t+d]$ given that it started at time u, entered the cell at time v, and is active at time t is $1-\overline{HC}_\tau(u,v,t,d)$.

Let $G_\tau^{\mathtt{c}}$ be the set of all connections of class τ in cell c at time t. Also, for each $c \in G_\tau^{\mathtt{c}}$, let u_c denote the time at which the connection started and v_c denote that connection c entered the cell under consideration. Note that, $v_c = u_c$ if the connection started in cell c. Then, in the ExpectedMax strategy, the base station in cell c estimates m_τ as

$$m_\tau = \left\lfloor \sum_{c \in G_\tau} (1 - \overline{HC}_\tau(u_c, v_c, t, d)) \right\rfloor$$

i.e., m_τ is the expected number of connections to either complete or incur a handoff in time $(t, t+d]$.

The estimation of n_τ requires interaction with neighboring cells. Let $N_{\mathtt{c}}$ denote the set of cells neighboring c. As explained earlier, the base station in cell c sends a message to the base station in each cell $j \in N_{\mathtt{c}}$ requesting the information necessary to estimate n_τ. The base station then estimates

$$n_\tau = \left\lceil \sum_{j \in N_i} n_\tau^j \right\rceil$$

where n_τ^j denotes the value returned from cell $j \in N_{\mathtt{cell}}$. The base station in the neighboring cell j computes as follows. For each connection, $c \in G_\tau^j$, let $q_{j,\mathtt{c}}(c)$ denote the conditional probability that the handoff will be to cell c given that connection c incurs a handoff. Then

$$n_\tau^j = \sum_{c \in G_\tau^j} H(u_c, v_c, t, d) q_{j,\mathtt{c}}(c).$$

The above expressions for estimating m_τ and n_τ hold for any given probability distribution function for connection duration time and cell residence time. For the special case, when the connection duration time is exponentially distributed and the cell residence time is also exponentially distributed, the above expressions become even simpler and are as shown below. These expressions are obtained by substitution and algebra in the above general expressions

$$\begin{aligned} F_\tau(t) &= 1 - e^{-\mu t} \\ R_\tau(t) &= 1 - e^{-\gamma t} \\ H_\tau(u,v,t,d) &= \frac{\gamma}{\gamma+\mu}(1 - e^{-(\gamma+\mu)d}) \\ \overline{HC}_\tau(u,v,t,d) &= e^{-(\gamma+\mu)d}. \end{aligned}$$

3) Efficient computation of $\overline{Y}_\tau$: Recall that

$$\overline{Y}_\tau = \sum_{s \in S(m_\tau, n_\tau)} Y_\tau(s) p_\tau(s).$$

If this expression is evaluated directly, the computational complexity is proportional to the cardinality of the set $S(m_\tau, n_\tau)$. We know that the cardinality of $S(m_\tau, n_\tau)$ is

$$|S(m_\tau, n_\tau)| = \frac{(m_\tau + n_\tau)!}{m_\tau! n_\tau!}.$$

The main problem with this approach is that $|S(m_\tau, n_\tau)|$ can be large when m_τ and n_τ are large. We describe below a scheme to reduce the computational complexity of the above expression.

Define $Z_\tau(s) = \max\{X_{\tau,k}\colon 1 \le k \le (m_\tau + n_\tau)\}$. Since $X_{\tau,0}(s) = 0$ for all s

$$Y_\tau(s) = \max\{0, Z_\tau(s)\}.$$

TABLE II
RECURSIVE EQUATIONS USED IN THEOREM 1

$$f(m_\tau,0,b)=\begin{cases}1 & \text{if } m_\tau=0,\ n_\tau=1,\ b=1\\ 1 & \text{if } m_\tau>0,\ n_\tau=0,\ b=-1\\ f(m_\tau-1,n_\tau,0)+f(m_\tau-1,n_\tau,-1) & \text{if } m_\tau\geq 0,\ n_\tau>0,\ b=-1\\ f(m_\tau-1,n_\tau,1) & \text{if } m_\tau\geq 0,\ n_\tau>0,\ b=0\\ f(m_\tau,n_\tau-1,0)+f(m_\tau,n_\tau-1,-1)+f(m_\tau-1,n_\tau,2) & \text{if } m_\tau>0,\ n_\tau>0,\ b=1\\ f(m_\tau,n_\tau-1,b-1)+f(m_\tau-1,n_\tau,b+1) & \text{if } m_\tau>0,\ n_\tau>0,\ b>1\\ 0 & \text{otherwise.}\end{cases}$$

Define $f(m_\tau,n_\tau,b)$ to be the number of sequences $s\in S(m_\tau,n_\tau)$ for which $Z_\tau(s)=b\phi_\tau$, That is

$$f(m_\tau,n_\tau,b)=|\{s\colon s\in S(m_\tau,n_\tau),Z_\tau(s)=b\phi_\tau\}|.$$

From the definition of $Z_\tau(s)$, the value of b can range from -1 to n_τ and

$$\overline{Y}_\tau=\sum_{b=-1}^{n_\tau}\frac{\max\{0,b\}\cdot\phi_\tau\cdot f(m_\tau,n_\tau,b)}{|S(m_\tau,n_\tau)|}. \tag{1}$$

Theorem 1: Given m_τ and n_τ such that $m_\tau+n_\tau\geq 1$, $f(m_\tau,n_\tau,b)$ is the solution of the recursive equation shown in Table II.

Proof: The proof of this theorem is given in the Appendix. ■

Let $\phi_{\min}=\min\{\phi_\tau\colon 1\leq\tau\leq M\}$. Then, the maximum value of n_τ in cell c is $|N_c|\cdot(\Gamma/\phi_{\min})$, where $|N_c|$ is the number of neighboring cells of c and Γ the bandwidth of the wireless link. Similarly, the maximum value of $m_\tau=(\Gamma/\phi_{\min})$. Therefore, the values of $f(m_\tau,n_\tau,b)$ can be computed offline and stored in a table. The stored values can be used at runtime to compute $\overline{Y}_\tau$ using (1).

IV. EVALUATION OF ExpectedMac STRATEGY

In this section, we compare the performance of ExpectedMax strategy with that of other schemes in literature. The comparison is done using a C-based discrete-event wireless network simulator. The inputs to the simulator are a model of the wireless network and the characteristics/requirements of the multimedia traffic in this network. The outputs of the simulator include the blocking probabilities for handoff and new connection requests.

In this section, we compare the handoff and new connection blocking probabilities of four different strategies. The first strategy, labeled Fixed(5%), is a static scheme in which each base station sets the threshold $\Delta_{n,\tau}$ to be 5% of its capacity for all τ. The second strategy, labeled Static3 is also a static scheme in which each base station sets the threshold $\Delta_{n,\tau}$ to be three times the average bandwidth requirement of the connection requests. This strategy requires knowledge of the relative occurrences of different traffic classes in the network. The third strategy is the extended version of YL97 scheme (see description in Section II) with hard constraint. Finally, the fourth strategy is the proposed ExpectedMax strategy.

The performance of the four strategies was compared for three different type of networks. In all cases, the assumed topology of the wireless network is as shown in Fig. 4. The other common aspects in all the three network types are as follows.

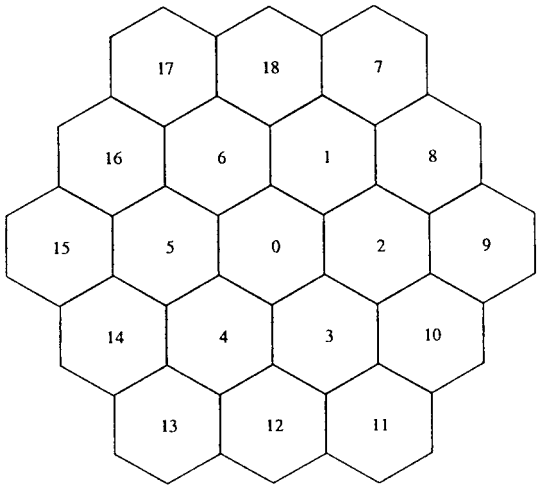


Fig. 4. The simulated wireless network.

1) The arrival of new connection requests of class τ in each cell c is a Poisson process with rate $\lambda_{\tau,i}$. The rate $\lambda_{\tau,i}$ varies with time depending on the scenario.
2) The duration of each class τ connection request is selected from an exponential distribution with rate μ_τ. The duration of a connection is selected when it is first admitted into the network. Once determined, its value is fixed until the connection completes. The base stations, of course, do not use this information to make any decisions because in practice exact duration of a connection will not be known to the network.
3) The residence time of a class τ connection in a cell is chosen when the connection starts and when it incurs a handoff. Consider a connection which enters a cell at time v (i.e., it either started in the cell at time v or it incurred a successful handoff into the cell at time v). Let u' be the selected completion time for the connection. First, a random number w is generated from an exponential distribution with rate γ_τ. If $v+w,>u'$, then the connection completes in the cell at time u'. Otherwise, it incurs a handoff out of the cell at time $v+w$. Since, in practice, a base station will not know the exact time of completion or handoff of a connection, this is assumed to be unknown to the base station.

TABLE III
STEADY-STATE HANDOFF PROBABILITIES BETWEEN CELLS IN MORNING RUSH HOUR SITUATION, FOR NETWORK 1

Handoff Type	Probability
Suburb to suburb	0.25
Suburb to city	0.5
City to suburb	0.033
City to city	0.2
City to downtown	0.5
Downtown to city	0.166

4) When a connection enters a cell (i.e., it either starts or it incurs a handoff into the cell), one of the neighboring cells is picked as a preferred cell for handoff. If the connection incurs a handoff (see discussion above), then handoff occurs to the preferred cell with 0.9 probability and with equal probability to one of the other neighboring cells. Since, in practice, a profile of a connection can be used to estimate the preferred handoff cell, the base station is assumed to be aware of the preferred cell for each connection.

The selection of the preferred cell for handoff is done as follows. As part of the input to the simulator, we specify parameters $q_{i,j}$ for each pair of adjacent cells i and j. $q_{i,j}$ represents the fraction of connections incurring a handoff from cell i which enter cell j. When a connection enters cell i, cell j is picked as a preferred cell for handoff with probability $q_{i,j}$.

A. Network 1

In this network type, we simulate the wireless network of Fig. 4 with cell 0 in a downtown region, cells 1–6 in the city and cells 7–18 in the suburbs. We first consider the morning rush hour scenario in which most users are moving toward the downtown area from the suburbs and the city by selecting the parameters as shown in Table III.

There are three classes of traffic in this network. We refer to them as classes 0, 1, and 2. The parameters for class 2 connections are similar to that of a typical cellular phone conversation. In particular, the bandwidth requirement of a class 2 connection is 64 Kbps and its mean duration is assumed to be 150 s. Class 0 and class 1 require much higher bandwidths and they also last longer on the average. This is because, in practice, users of higher bandwidth connections like video conferencing are typically connected for much longer duration as compared to typical voice connection. Specifically, the bandwidth requirements of classes 0 and 1 connections are, respectively, eight and four times that of a class 2 connection. Moreover, the mean duration of classes 0 and 1 connections are, respectively, 25 and five times that of a class 2 connection.

In each cell, the arrival rate of each class of connection increases in the first half of our simulation and decreases in the second half. This corresponds to a typical increase in the call arrival rate from say 6:00–8:00 a.m. and then a decrease in the call arrival rate from 8:00–10:00 a.m. The increases in the call arrival rate in the various cells do not occur at the same rate. In the first half of the simulation, approximately once every 24 min, the call arrival rate in a cell is increased by a random factor chosen from an uniform distribution between 1.0 and 1.4. Similarly, in the second half of the simulation, approximately once every 24 min, the call arrival rate in a cell is decreased by a random factor chosen from an uniform distribution between 1.0 and 1.4.

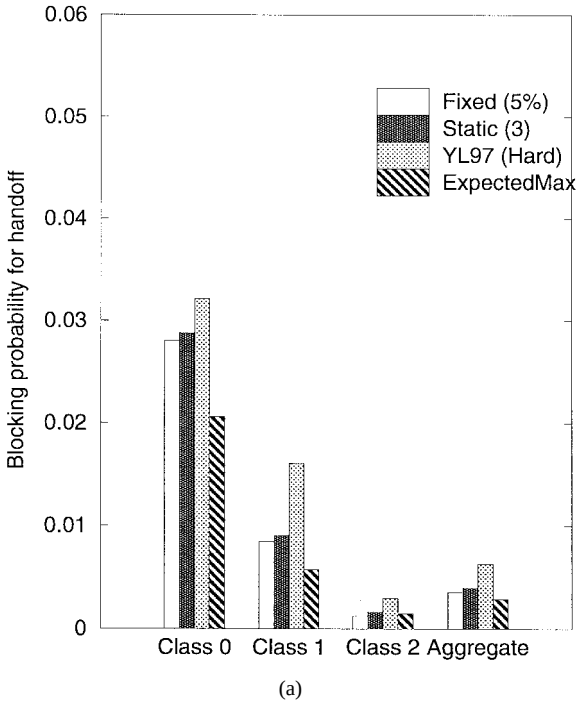
There is usually a difference in the new connection request arrival rate between a downtown cell and a suburb cell. To account for this, the new connection request arrival rate in downtown is assumed to be 40% higher on the average than in the suburb. Likewise, new connection request arrival rate in the city is on the average 20% higher than in the suburb. Similarly, in practice, there is also likely to be a difference between the cell residence times of a connection for downtown, city, and suburb. For instance, cellular phone users in a downtown are more likely to remain in downtown as compared to city or suburb. To account for this the mean unencumbered cell residence times of each connection in downtown and city are respectively assumed to be 100 and 33% longer than that in the suburb.

Furthermore, since the cells differ considerably in the arrival rate of handoff and new connection requests, we assume that the total bandwidth available in the cells differ correspondingly. Specifically, we assume that the total bandwidth available in downtown is twice that of the suburb and 25% more than that of the city. Furthermore, in downtown, we assume that the total bandwidth available is adequate to simultaneously support at most twenty class 0 connections. The difference in the cell capacity can be achieved in practice by allocating more channels to the downtown cell as compared to the city and suburb. For example, 20 different frequencies can be assigned to the downtown cell, 16 frequencies to the city cells, and 10 frequencies to the suburb cells to achieve the above variation in cell capacity.

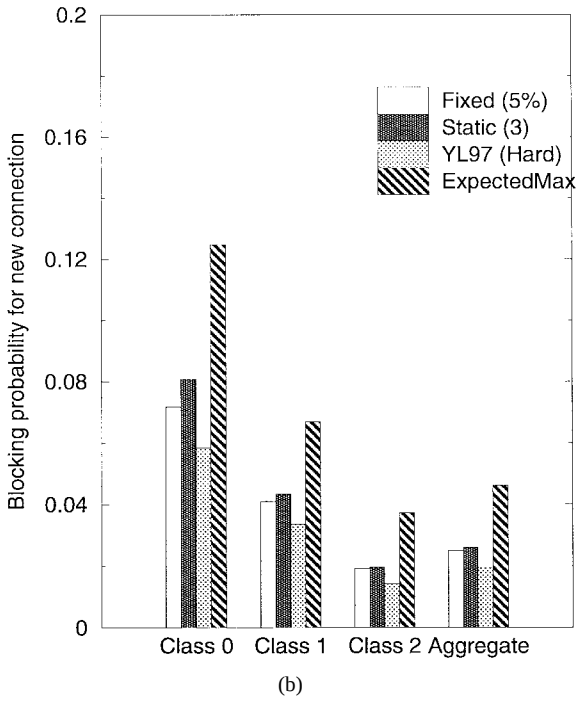
Fig. 5(a) and (b), respectively, shows the blocking probabilities for handoff and new connection requests as a function of mean new connection request arrival rate for the four different strategies. Since this arrival rate increases in the first half and decreases in the second half of the simulation, the mean arrival rate is computed as the total number of connections which arrive during the simulation divided by the duration of the simulation.

First observe that, as expected, the blocking probabilities for handoff and new connection request increase in all strategies with increase in the arrival rate of new connection request. The performance of `Fixed`(5%) and `Static`(3) schemes are better than that of YL97 strategy. Importantly, note that the blocking probabilities for handoff request is the smallest for the proposed `ExpectedMax` strategy. The maximum load for which the target handoff blocking probability of 0.01 is achieved, is largest for our `ExpectedMax` strategy.

Fig. 5(b) shows that the new connection blocking probabilities are lower for all schemes compared to `ExpectedMax`. As mentioned earlier, there is an obvious tradeoff in blocking probabilities for handoff versus new connections. The network designer needs to make a decision regarding the amount of penalty he is willing to accept in terms of higher new connection blocking probability. For instance, the designer can

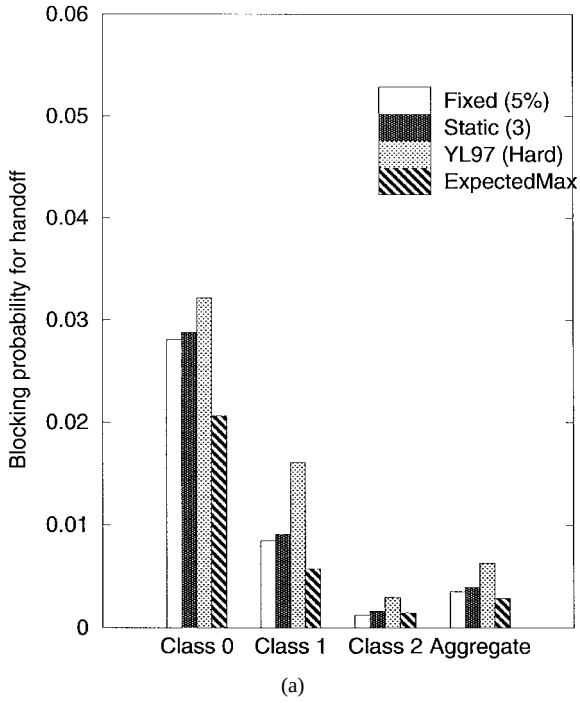


(a)

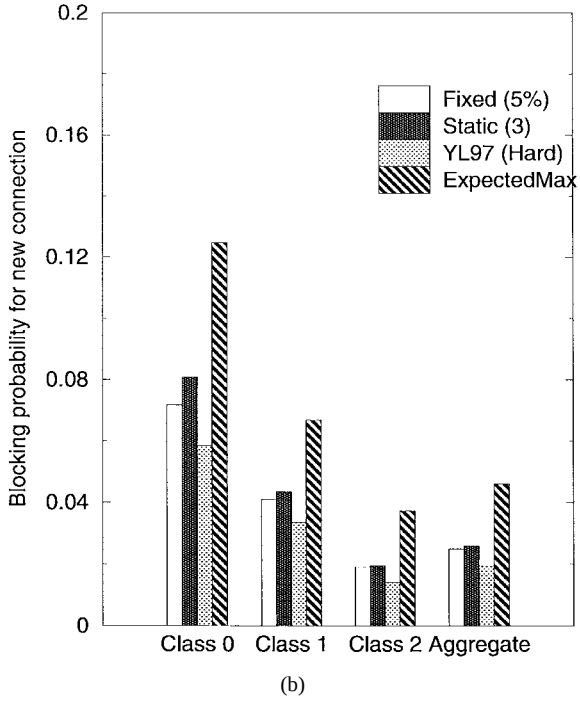


(b)

Fig. 5. Blocking probabilities in different strategies as a function of load in Network 1 with morning rush hour traffic. (a) Handoff and (b) new connection.



(a)



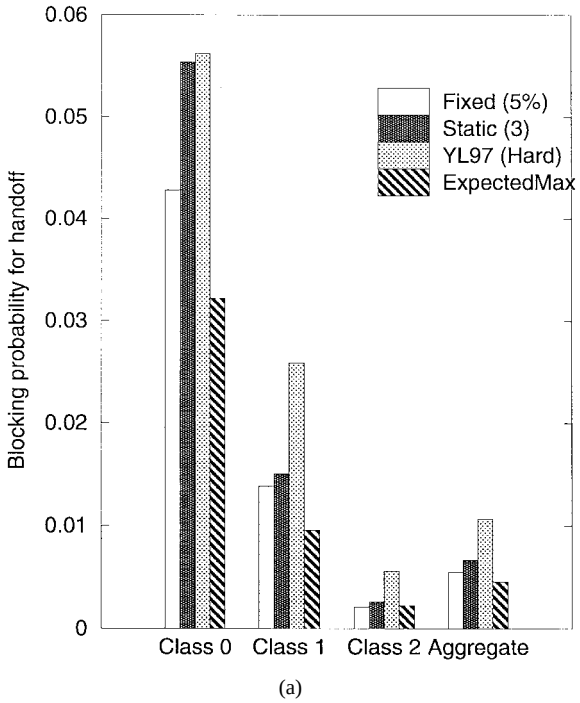
(b)

Fig. 6. Blocking probabilities for the traffic classes under medium load in Network 1 with morning rush hour traffic. (a) Handoff and (b) new connection.

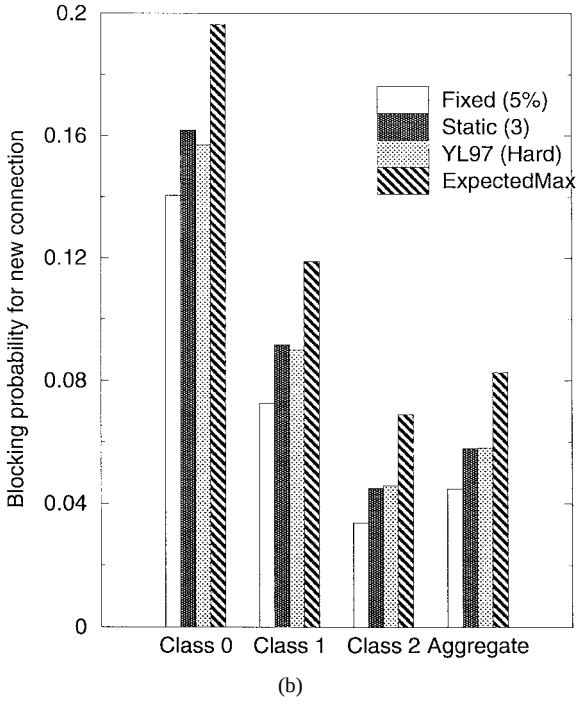
decide on target probabilities of 0.005 and 0.10, respectively, for handoff and new connections. Our scheme targets such a situation and strives to closely estimate handoff resource requests to actual future requests.

Figs. 6 and 7 show the blocking probabilities for the three individual classes, for two different system loads (medium and high). The medium load corresponds to a new connection arrival rate of 0.53 connections/s whereas the high load corresponds to an arrival rate of 0.672 connections/s. The `ExpectedMax` strategy consistently results in lower handoff blocking probabilities for all the three classes in both cases. Note that, class 2 has typically very low blocking probability, while class 0 has higher blocking probability. This is expected since class 2 has lower bandwidth requirements. This is also attractive to network service providers where voice (class 2) typically is the mainstream application compared to video (class 0).

1) Update Frequency: As described earlier, ideally each base station must get updated information about expected handoffs from neighboring cells upon arrival of each new



(a)



(b)

Fig. 7. Blocking probabilities for the traffic classes under high load in Network 1 with morning rush hour traffic. (a) Handoff and (b) new connection.

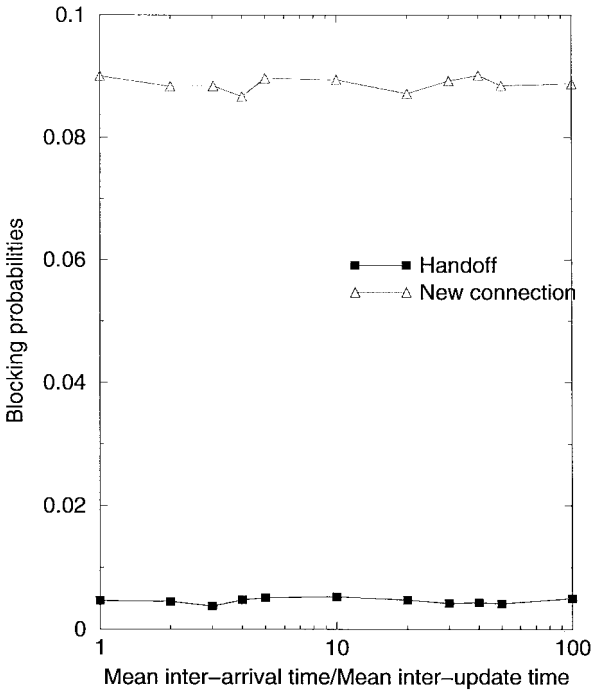


Fig. 8. Blocking probabilities as update is reduced in Network 1 at high load.

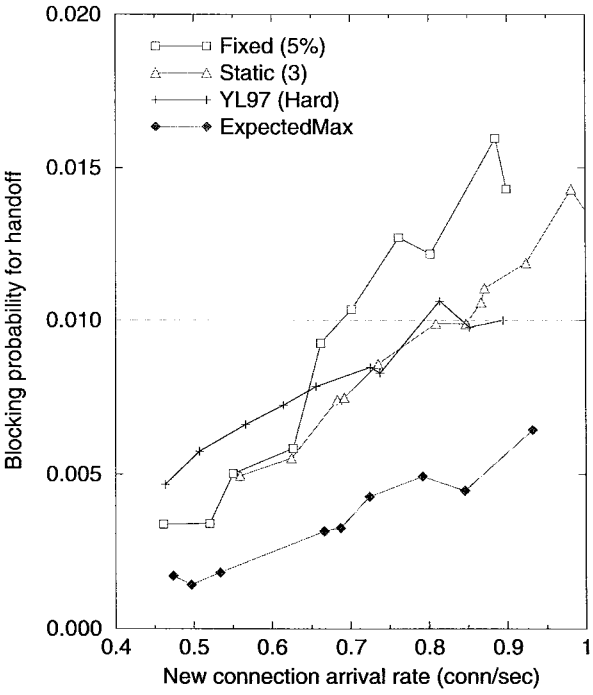
TABLE IV
STEADY-STATE HANDOFF PROBABILITIES BETWEEN CELLS IN EVENING RUSH HOUR SITUATION, FOR NETWORK 1

Handoff Type	Probability
Suburb to suburb	0.45
Suburb to city	0.05
City to suburb	0.25
City to city	0.1
City to downtown	0.05
Downtown to city	0.166

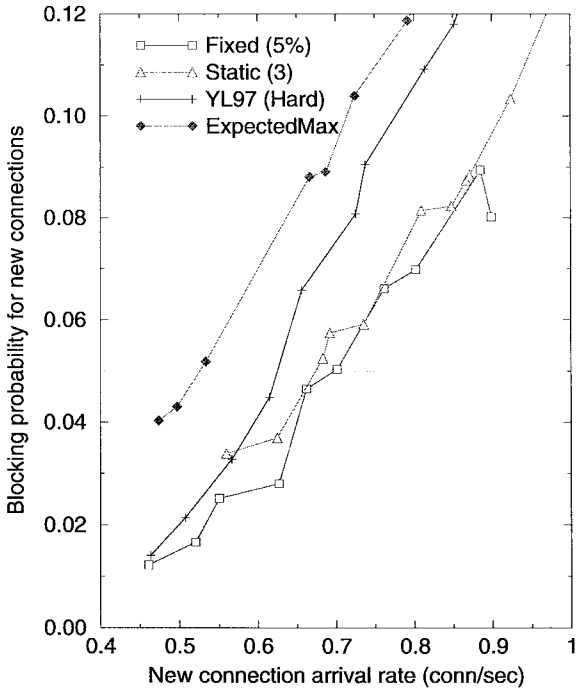
connection request. The overhead of such frequent updates is clearly very high. Our objective is to minimize the frequency of update while achieving accurate estimates of handoff resource requests. Fig. 8 shows the variation in blocking probabilities as the update rate is decreased at one particular load, namely 0.672 conn/s new connection arrival rate. The x-axis in Fig. 8 shows the ratio of the mean new connection inter-arrival rate and the mean inter-update rate. A ratio of one means that, on the average, updated information is obtained from neighboring nodes for each incoming new connection request (i.e., approximately the ideal update rate). A ratio of 100 means that, on the average, updated information is obtained from neighboring nodes once every 100 new connection requests. Observe that, the blocking probabilities are fairly steady even when the update rate is reduced to approximately 1/100 of the ideal update rate. This indicates that the proposed `ExpectedMax` strategy can be implemented without much overhead in terms of frequent updates between base stations.

2) Evening Rush Hour: To provide more perspective, we also simulated the condition where users are moving away from downtown toward city and suburbs. Specifically, the mobility parameters are as shown in Table IV.

Here, we assume that, on the average, arrival rate of new connection requests in downtown is 2.5 times that in suburb, and in the city is 2.0 times that in the suburb. As before, the arrival rates in each cell for traffic type increases in the first half of the simulation and decreases in the second half. The relative behavior of all strategies are very similar to that discussed for morning rush hour situation in Network 1 (see Fig. 9). In particular, the proposed `ExpectedMax` strategy has the least handoff blocking probability. Correspondingly, it has the highest new connection request blocking probability. As stated earlier, since premature termination of established



(a)



(b)

Fig. 9. Blocking probabilities in different strategies as a function of load in Network 1 with evening rush hour traffic. (a) Handoff and (b) new connection.

connection requests is probably more undesirable than rejection of new connection request, the proposed `ExpectedMax` strategy seems to be better even in this situation.

B. Network 2

The wireless network and the mobility pattern in this network are exactly as in Network 1. Instead of three different classes of traffic, however, all connections requests are of the same class in this network. Specifically, the parameters of all the connection requests are that of a typical cellular phone conversation (i.e., class 2 connection of 1). This network was used in the evaluation described in [15] and is included here for comparison.

As in Network 1, the arrival rate of new connection requests increases in the first half of the simulation and then decreases in the second half. The method used to increase and decrease the rate of arrival of connections is exactly as in Network 1. Also, as in Network 1, the average call arrival rate in a downtown is 40% higher than that in suburb. The call arrival rate in the city is on the average 20% higher than in the suburb. Moreover, to account for the differences in the residence times among cells, mean unencumbered cell residence time in downtown (city) is assumed to be twice (1.33 times) that in suburb.

Fig. 10 shows the variation in the blocking probabilities in the different strategies when arrival rate of new connection requests is increased. Unlike in Network 1, the `Static`(3) and the `Fixed`(5%) strategies perform quite well in this network. They have the smallest handoff blocking probability as compared to the dynamic schemes. The main reason is that the average bandwidth per connection is very small (approximately 0.0064) and therefore reserving 5% results in over-reserving resources for handoffs. Since the `Static`(3) reserves for at most three handoff connections, its blocking probability is higher when compared to `Fixed`(5%). The handoff blocking probabilities in the proposed `ExpectedMax` strategy are smaller than that of the YL97 schemes. The difference between the two schemes, however, seems to be much less in this network as compared to in Network 1. The new connection blocking probabilities for the YL97 schemes are much better than that for `ExpectedMax` strategy. The results here seem to indicate that with a low-bandwidth homogeneous traffic network, a static scheme will be sufficient to obtain good performance. As mentioned earlier, static schemes do not require the overhead associated with base station updates. As shown in the previous section, there is significant advantage in multiple class networks, with diverse traffic requirements.

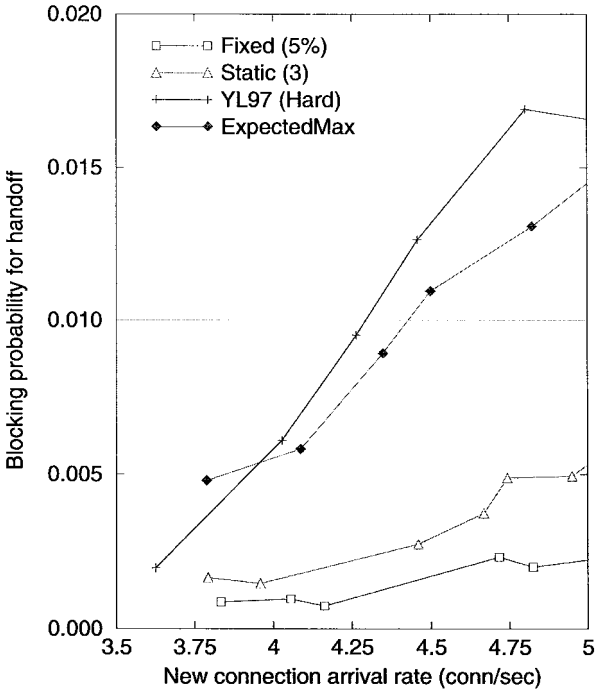
C. Network 3

This network is meant to capture a uniform network where all cells are identical in terms of traffic flow and the probabilities of moving between cells is uniform. That is, a mobile is equally likely to move to any of the neighboring cells. The parameters of the traffic classes are as in Network 1.

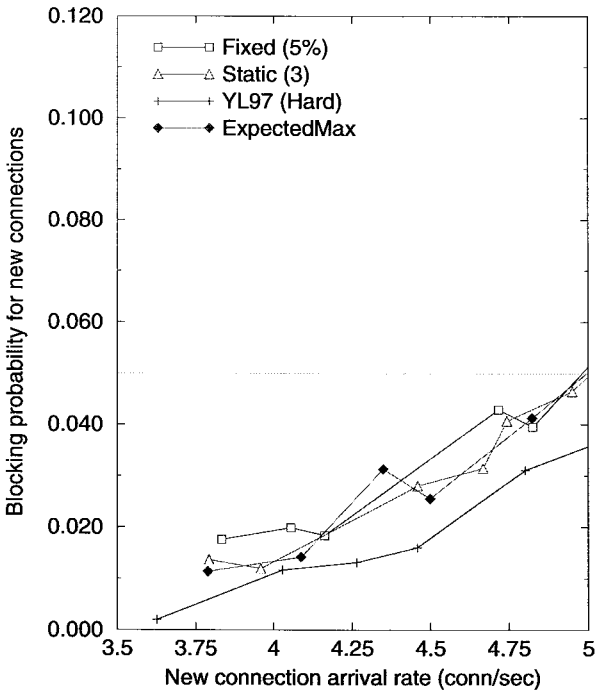
Here again, Fig. 11 shows the variation in handoff and new connection blocking probabilities for different loads. As in Network 1, the handoff blocking probability is the least for the `ExpectedMax` strategy. In this network, YL97 strategies are worse than the `Fixed`(5%) strategies and the `Static`(3) schemes for both handoff and new connection blocking probabilities. This shows that our `ExpectedMax` strategy is well adapted to different network conditions and traffic patterns.

V. Conclusions

This paper addressed the problem of providing resources to mobile connections during handoff between base stations.

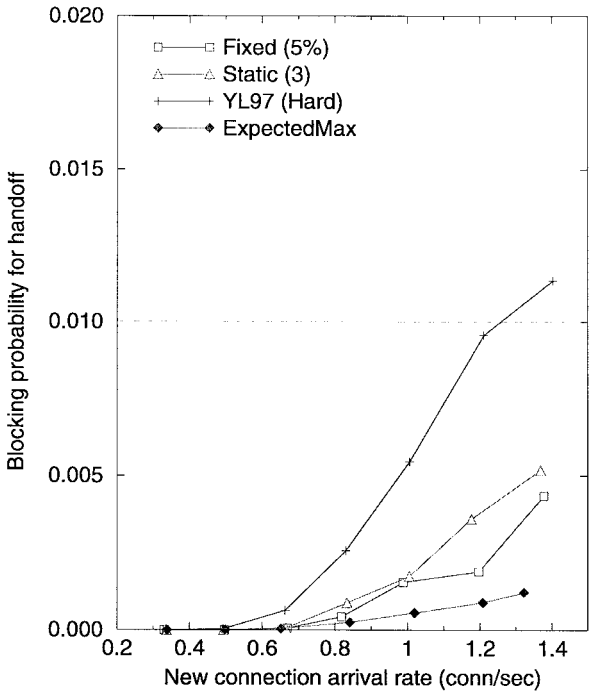


(a)

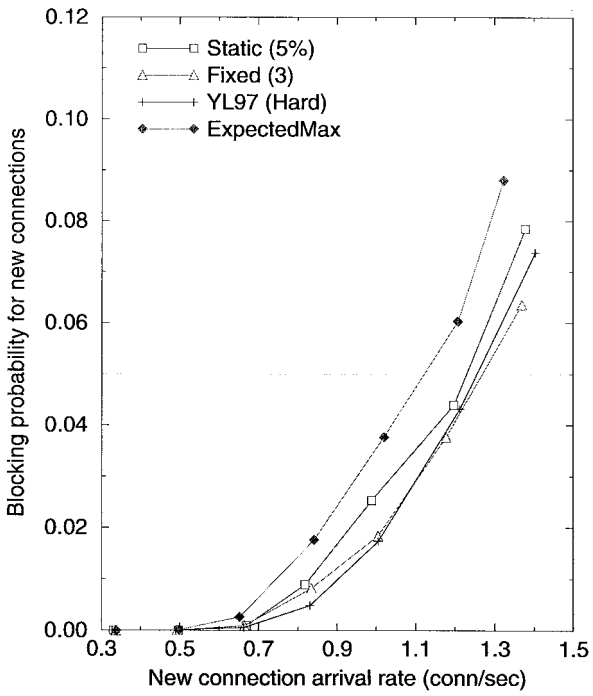


(b)

Fig. 10. Blocking probabilities in different strategies as a function of load in Network 2. (a) Handoff and (b) new connection.



(a)



(b)

Fig. 11. Blocking probabilities in different strategies as a function of load in Network 3. (a) Handoff and (b) new connection.

The network is assumed to be cell-based with support for diverse traffic types. The goal is to estimate the requirements for resources from mobiles that are currently in neighboring cells and that might potentially move to the current cell. This estimate is then used to appropriately reserve resources in the current cell for potential handoff connections. A dynamic strategy that uses the estimated holding times and mobility pattern information is proposed in the paper. The performance of this strategy is compared to two fixed or static schemes that do not use any dynamic information, and to a scheme proposed earlier. The results were studied for three different type of networks. The performance metrics studied are blocking probabilities for handoff and for new connections. A fundamental assumption is that the network designer desires a lower handoff

blocking probability. This requirement might result in higher blocking probability for new connections. A suitable tradeoff is required based on the network service objectives.

The results show that our proposed `ExpectedMax` strategy consistently achieves lower handoff blocking probability than all other schemes. The results also show that the dynamic estimation can be achieved without significant overhead in terms of control communication between base stations. Further work in this area includes translating the high-level resource allocations into scheduling at the multiple access level using an access protocol such as described in [13] to ensure quality-of-service.

Appendix I
Proof of Theorem 1

Proof of Theorem 1: The theorem is proved by considering the individual cases in (2), namely the following.

- *Case 1:* $m_\tau = 0, n_\tau = 1, b = 1$.
- *Case 2:* $m_\tau > 0,\ n_\tau = 0, b = -1$.
- *Case 3*: $m_\tau \geq 0,\ n_\tau > 0, b = -1$.
- *Case 4*: $m_\tau \geq 0, n_\tau > 0,\ b = 0$.
- *Case 5*: $m_\tau > 0,\ n_\tau > 0, b = 1$.
- *Case 6*: $m_\tau > 0, n_\tau > 0,\ b > 1$.

Case 6 is the common case. Cases 1–5 are in some sense boundary cases. Therefore, we prove Case 6 in detail and outline the key proof steps for other cases. Some of the cases need the following definitions and observation.

Let $S_I(m_\tau, n_\tau) \subseteq S(m_\tau,\ n_\tau) = \{s\colon s \in S(m_\tau, n_\tau), s \equiv Is'\}$ be the set of all sequences in $S(m_\tau, n_\tau)$ which start with I. Likewise, let $S_O(m_\tau, n_\tau) \subseteq S(m_\tau,\ n_\tau) = \{s\colon s \in S(m_\tau, n_\tau), s \equiv Os'\}$ be the set of all sequences in $S(m_\tau, n_\tau)$ which start with O. Let $f_I(m_\tau, n_\tau, b)$ be the number of sequences $s \in S_I(m_\tau, n_\tau)$ such that $Z_\tau(s) = b\phi_\tau$. Similarly, let $f_O(m_\tau, n_\tau, b)$ be the number of sequences $s \in S_O(m_\tau, n_\tau)$ such that $Z_\tau(s) = b\phi_\tau$. Observe that

$$f(m_\tau, n_\tau, b) = f_I(m_\tau, n_\tau, b) + f_O(m_\tau, n_\tau, b). \tag{3}$$

Case 6—$m_\tau > 0, n_\tau > 0, b > 1$: By definition any $s \in S_I(m_\tau, n_\tau)$, can be written as Is' for some s'. Thus, for $s \in S_I(m_\tau, n_\tau)$, $X_{\tau,1} = \phi_\tau$ and

$$\begin{aligned} Z_\tau(s) &= \max\left\{0, X_{\tau,1}, \max_{2\leq k\leq |s|} X_{\tau,k}\right\} \\ &= \max\left\{\phi_\tau, \max_{2\leq k\leq |s|} X_{\tau,k}(s)\right\} \\ &= \max\left\{\phi_\tau, \phi_\tau + \max_{1\leq k\leq |s'|} X'_{\tau,k}(s')\right\} \\ &= \max\{\phi_\tau, \phi_\tau + Z_\tau(s')\}. \end{aligned}$$

Therefore, for $b > 1$

$$Z_\tau(s) = b\cdot\phi_\tau \text{ iff } Z_\tau(s') = (b-1)\phi_\tau.$$

Hence, $f_I(m_\tau, n_\tau, b) = f_I(m_\tau, n_\tau - 1, b-1)$.

Similarly, by definition any $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s'. Thus, for $s \in S_O(m_\tau, n_\tau)$, $X_{\tau,1} = -\phi_\tau$ and

$$\begin{aligned} Z_\tau(s) &= \max\left\{0, X_{\tau,1}, \max_{2\leq k\leq |s|} X_{\tau,k}\right\} \\ &= \max\left\{0, \max_{2\leq k\leq |s|} X_{\tau,k}(s)\right\} \\ &= \max\left\{0, -\phi_\tau + \max_{1\leq k\leq |s'|} X'_{\tau,k}(s')\right\} \\ &= \max\{0, -\phi_\tau + Z_\tau(s')\}. \end{aligned}$$

Therefore, for $b > 1$,

$$Z_\tau(s) = b\cdot\phi_\tau \text{ iff } Z_\tau(s') = (b+1)\phi_\tau.$$

Hence, $f_O(m_\tau, n_\tau, b) = f_O(m_\tau - 1, n_\tau, b+1)$. The theorem then follows for this case from (3).

Case 1—$m_\tau = 0$,: Since $m_\tau = 0$ and $n_\tau = 1$, the set $S(m_\tau, n_\tau)$ contains only one element, namely the sequence I. Therefore, $Z_\tau(s) = \phi_\tau$ for all $s \in S(m_\tau, n_\tau)$. That is, $f(0,1,1) = 1$ and $f(0,1,b) = 0$ for all $b \neq 1$.

Case 2—$m_\tau > 0, n_\tau = 0, b = -1$: In this case, the set $S(m_\tau, 0)$ contains only one sequence, namely a sequence of m_τ O's. Therefore, $Z_\tau(s) = X_{\tau,1}(s) = -\phi_\tau$ for all $s \in S(m_\tau, 0)$. That is, $f(m_\tau, 0, -1) = 1$ and $f(m_\tau, 0, b) = 0$ for all $b \neq -1$ and all $m_\tau > 0$.

Case 3—$m_\tau \geq 0, n_\tau > 0, b = -1$: For any $s \in S_I(m_\tau, n_\tau)$, $X_{\tau,1}(s) = \phi_\tau$. Therefore, $Z_\tau(s) \geq \phi_\tau$ for all for all $s \in S_I(m_\tau, n_\tau)$ and $f_I(m_\tau, n_\tau, -1) = 0$.

By definition $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s' and

$$Z_\tau(s) = -\phi_\tau \text{ iff } Z_\tau(s') = 0 \text{ or } Z_\tau(s') = -\phi_\tau.$$

Therefore, $f_O(m_\tau, n_\tau, -1) = f_O(m_\tau - 1, n_\tau, 0) + f_O(m_\tau - 1, n_\tau, -1)$. The theorem follows for this case from (3).

Case 4—$m_\tau \geq 0, n_\tau > 0, b = 0$: For any $s \in S_I(m_\tau, n_\tau)$, $X_{\tau,1} = \phi_\tau$. Therefore, $Z_\tau(s) \geq \phi_\tau$ for all $s \in S_I(m_\tau, n_\tau)$ and $f_I(m_\tau, n_\tau, 0) = 0$.

Similarly, any $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s'. Thus, for $s \in S_O(m_\tau, n_\tau)$

$$Z_\tau(s) = 0 \text{ iff } Z_\tau(s') = \phi_\tau$$

and therefore $f_O(m_\tau, n_\tau, 0) = f_O(m_\tau - 1, n_\tau, 1)$. The theorem follows for this case from (3).

Case 5—$m_\tau > 0, n_\tau > 0, b = 1$: By definition, any $s \in S_I(m_\tau, n_\tau)$, can be written as Is' for some s'. Thus, for $s \in S_I(m_\tau, n_\tau)$

$$Z_\tau(s) = \phi_\tau \text{ iff } Z_\tau(s') = 0 \text{ or } Z_\tau(s') = -\phi_\tau.$$

Therefore, $f_I(m_\tau, n_\tau, 1) = f_I(m_\tau, n_\tau - 1, 0) + f_I(m_\tau, n_\tau - 1, -1)$.

Likewise, any $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s' and therefore

$$Z_\tau(s) = \phi_\tau \text{ iff } Z_\tau(s') = 2\phi_\tau.$$

Hence, $f_O(m_\tau - 1, n_\tau, 1) = f_O(m_\tau - 1, n_\tau, 2)$. The theorem follows for this case from (3).

The theorem follows from Cases 1–6. ■

REFERENCES

[1] P. Agrawal, D. K. Anvekar, and B. Narendran, "Channel management policies for handovers in cellular networks," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 97–110, 1996.

[2] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. Srivastava, and J. A. Trotter, "SWAN: A mobile multimedia wireless network," *IEEE Personal Commun.*, vol. 39, pp. 18–33, Apr. 1996.

[3] J.-Y. Le Boudec, "Network calculus made easy," Tech. Rep. EPFL-DI 96/218, Dec. 1996. [Online]. Available WWW: http://lrcwww.epfl.ch.

[4] J. Daigle and N. Jain, "A queueing system with two arrival streams and reserved servers with application to cellular telephone," in *Proc. INFOCOM*, Apr. 1992, pp. 2161–2167.

[5] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.

[6] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–981, Sept. 1991.

[7] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone system with prioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 3, pp. 77–91, Aug. 1986.

[8] T. Hsing, D. C. Cox, L. F. Chang, and T. Van Landegren, "Wireless ATM: Special issue," *IEEE J. Select. Areas Commun.*, vol. 15, Jan. 1997.

[9] F. P. Kelly, "Effective bandwidths at multi-type queues," *Queueing Syst.*, vol. 9, pp. 5–15, 1991.

[10] M. Naghshineh, Ed., "Wireless ATM: Special issue," *IEEE Personal Commun. Mag.*, vol. 39, Aug. 1996.

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

[12] K. Pahlavan and A. H. Levesque, "Wireless data communications," *Proc. IEEE*, vol. 82, pp. 1398–1430, 1994.

[13] K. M. Sivalingam, J.-C. Chen, P. Agrawal, and M. B. Srivastava, "Design and analysis of low-power access protocols for wireless and mobile ATM networks," *ACM/Baltzer Mobile Networks and Applications*, to be published.

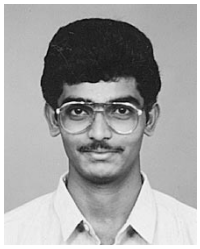
[14] C. H. Yoon and C. K. Un, "Performance of personal portable radio telephone systems with and without guard channels," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 911–917, Aug. 1993.

[15] O. T. W. Yu and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1208–1225, Sept. 1997.

[16] M. M. Zonoozi and P. Dassanayake, "User mobility modeling and characterization of mobility patterns," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1239–1252, Sept. 1997.

Parameswaran Ramanathan (S'85–M'89) received the B.Tech. degree from the Indian Institute of Technology, Bombay, India, in 1984 and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1986 and 1989, respectively.

He is an Associate Professor in the Department of Electrical and Computer Engineering and in the Department of Computer Sciences at the University of Wisconsin, Madison. From 1984 to 1989 he was a Research Assistant in the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. He was an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Wisconsin, Madison, from 1989 to 1995. He was on sabbatical leave in 1997–1998 and spent it at AT&T Labs, Whippany, NJ, and Bellcore, Morristown, NJ. His research interests include the areas of real-time systems, high-speed networks, fault tolerant computing, distributed systems, and parallel algorithms.



Krishna M. Sivalingam (S'92–M'95) received the B.E. degree in computer science and engineering in 1988 from Anna University, Madras, India, and the M.S. and Ph.D. degrees in computer science in 1990 and 1994, respectively, from State University of New York at Buffalo.

He is an Assistant Professor in the School of Electrical Engineering and Computer Science, at Washington State University, Pullman. Earlier, he was an Assistant Professor at University of North Carolina ,Greensboro, from 1994 until 1997. He has conducted research at Lucent Technologies' Bell Labs in Murray Hill, NJ, and at AT&T Labs in Whippany, NJ, where he also served as a consultant during 1997. His research interests include wireless and mobile networks, optical WDM networks, ATM networks, high-speed communication networks such as gigabit ethernet, high-performance distributed computing, and performance evaluation. He is co-editing a book on optical WDM networks.

Dr. Sivalingam has served on the conference committees of IEEE INFOCOM 1997, WOWMOM Workshop 1998 and 1999, ACM Mobicom 1999, MASCOTS 1999, Mobile Data Access 1999, and ICCCN 1999. He is a Guest Editor for an issue of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He was a Presidential Fellow from 1988 to 1991 at State University of New York at Buffalo and is a member of ACM.

Prathima Agrawal (S'74–M'77–SM'85–F'89) received the B.E. and M.E. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles.

She is a Chief Scientist in the Internet Architecture Research Laboratory at Telcordia Technologies (formerly Bellcore), Morristown, NJ. During 1997–1998, she was head of the Networked Computing Technology Department at AT&T Labs in Whippany, NJ. Prior to that, she headed the Networked Computing Research Department at Bell Labs in Murray Hill, NJ. Her research interests are computer networks, mobile computing, parallel processing, and VLSI CAD. She has published over 150 papers and has received or applied for more than 30 U.S. patents.

Dr. Agrawal is a member of the ACM. Presently, she chairs the IEEE Fellow Selection Committee.

Shalinee Kishore received the B.S. and M.S. degrees in electrical engineering from Rutgers University, New Brunswick, NJ, in 1996 and 1999, respectively. She is currently pursuing the Ph.D. degree in electrical engineering at Princeton University, Princeton, NJ.

During the summers of 1992 and 1993, she wrote software for the analysis of fusion reactions at the Princeton Plasma Physics Laboratory. She was an intern at AT&T Bell Laboratories in the summers of 1994 to 1996. In 1997, she held an internship in the Advanced Communications Laboratory of AT&T Labs, Whippany, NJ, working on the design of low-power media access schemes for wireless communications. Her current research interests are in the areas of resource allocation and signal processing for wireless multiple access.

Ms. Kishore received the AT&T Labs Fellowship in 1997.