A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept

David A. Levine, Ian F. Akyildiz, *Fellow, IEEE*,
and Mahmoud Naghshineh, *Member, IEEE*

***Abstract*—The *shadow cluster* concept can be used to estimate future resource requirements and to perform call admission decisions in wireless networks. Shadow clusters can be used to decide if a new call can be admitted to a wireless network based on its quality-of-service (QoS) requirements and local traffic conditions. The shadow cluster concept can especially be useful in future wireless networks with microcellular architectures where service will be provided to users with diverse QoS requirements. The framework of a shadow cluster system is completely distributed, and can be viewed as a message system where mobile terminals inform the base stations in their neighborhood about their requirements, position, and movement parameters. With this information, base stations predict future demands, reserve resources accordingly, and admit only those mobile terminals which can be supported adequately. The shadow cluster concept involves some processing and communication overheads. These overheads have no effect on wireless resources, but only on the base stations and the underlying wireline network. In this paper, it is shown how base stations determine the probabilities that a mobile terminal will be active in other cells at future times, define and maintain shadow clusters by using probabilistic information on the future position of their mobile terminals with active calls, and predict resource demands based on shadow cluster information. In addition, a call admission algorithm is introduced, which uses current traffic and bandwidth utilization conditions, as well as the amount of resources and maximum allowable "dropping probability" being requested. Performance results showing the advantages of the shadow cluster concept are also included in the paper.**

***Index Terms*— Active mobile probability, call admission, resource allocation, shadow cluster.**

I. Introduction

THE BANDWIDTH in a wireless network is perhaps the most precious and scarce resource of the entire communication system. This resource should be used in the most efficient manner. A base station sometimes may need to reserve resources, even if this means denying access to a mobile terminal requesting admission to the network, in order to keep enough resources to support active users currently outside of its coverage area, but who may soon emigrate to its cell. Base stations must maintain a balance between the two conflicting requirements: 1) maintain maximum resource (bandwidth) utilization and 2) reserve enough bandwidth resources so that the maximum rate of unsuccessful incoming handoffs (due to insufficient resources) is kept below an acceptable level. The probability of unsuccessful handoffs can be established in terms of a quality-of-service (QoS) metric, e.g., call dropping probability, that the network agrees to maintain.

An accurate determination of the amount of resources that a base station must reserve (to maintain a certain call dropping probability) is likely to become a very important issue in future wireless networks. In contrast to current systems, future wireless networks will support a wide range of applications with diverse bandwidth requirements. Also, in future systems the demand on wireless bandwidth within a cell may change abruptly in a short period of time [8], as for example, when several video or high data rate users enter or leave a cell at the same time. In contrast, in current systems the bandwidth demand usually varies gradually, and hence it is much easier to handle. Moreover, future wireless networks may provide customized QoS parameters on a per call and/or on a service basis, enabling users to select a level of service according to a pricing plan.

In order to maintain an acceptable call dropping probability rate, several schemes [2], [8], [10], [13] have been proposed to dynamically organize the allocation of bandwidth resources. These schemes consider only limited information from neighboring cells, and do not specifically consider admission control policies as means to prevent congestion. Issues and relationships between resource reservation, channel assignment, call admission, and traffic intensity have been studied previously [7], [12], [14]. Admission control policies which determine the number of new voice or data users for acceptance in a packet radio network are given in [14]. For these policies, voice users are accepted only if a long-term blocking probability is not exceeded, while data users are accepted only if the mean packet delay and the packet loss probability are maintained below certain levels. In [12], a "flexible" channel assignment scheme is proposed based on the analysis of offered traffic distributions or blocking probabilities. A distributed call admission control procedure is proposed in [7], which takes into consideration the number of calls in adjacent cells as well as in the cell where a new

Manuscript received November 10, 1995; revised October 16, 1996; approved by IEEE/ACM Transactions on Networking Editor C. Perkins. This work was supported in part by the National Science Foundation under Grant EID-9018632. This paper was presented in part at MOBICOM'95, Berkeley, CA, November 1995.

D. A. Levine and I. F. Akyildiz are with Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

M. Naghshineh is with IBM T. J. Watson Research Center, Yorktown Heights, NY 10532 USA.

Publisher Item Identifier S 1063-6692(97)01706-8.

call request is being made, in addition to the knowledge of the mean call arrival, call departure, and call handoff rates. None of these schemes consider the individual trends of the users in the wireless network, e.g., position, speed, direction, and bandwidth demands.

In this paper, we describe the so-called "shadow cluster concept" [5], a predictive resource estimation scheme which provides high wireless network utilization by dynamically reserving only those resources that are needed to maintain the call dropping probability requested by the wireless connection. The shadow cluster concept is a solution to the problems of resource reservation and call admission in a wireless network. It is the first scheme that utilizes real-time information about the dynamics, traffic patterns, and bandwidth utilization of the individual mobile terminals in a network. The shadow cluster scheme is *dynamic and proactive*, i.e., the amount of resources to be reserved is determined "on-the-fly," and the control functions on call admissions are aimed at preventing congestion conditions. By using shadow clusters, the number of dropped calls during handoffs can be reduced, and the establishment of new calls that are highly likely to result later in dropped calls can be avoided. Shadow clusters are best suited for wireless networks with small cell sizes (nano, micro, mini), that result in a high number of cell handoffs during the lifetime of the average wireless connection. The shadow cluster concept is completely distributed. It requires processing overheads in base stations as well as some communication between a mobile terminal's base station and its neighbors through the wireline network. The mechanism does not require the use of wireless resources. Since future base stations are likely to be linked by switches which use high-bandwidth optical fiber [1], and since the shadow cluster algorithms' cycle time is likely to be in the order of several seconds [5], the amount of extra information generated by the mechanism should be easily manageable by the wireline network.

This paper is organized as follows. In Section II, we present a general description of the shadow cluster scheme. In Section III, we study the *active mobile probability* concept, used by base stations to inform their neighbors about the probable future locations of active mobile terminals in the network. Also, we explain information structures and methodologies that can be used to compute active mobile probabilities. In Section IV, we describe how shadow clusters can be used to estimate future resource demands in a cell. Using the procedures outlined in Section IV, we develop a call admission algorithm in Section V. In Section VI, we examine the performance of the shadow cluster concept for mobile terminals moving along a highway. Finally, in Section VII, we provide some conclusions about this work.

II. The Shadow Cluster Concept

Consider a microcell wireless network system that can support mobile terminals running applications which demand a wide range of resources. Users can freely roam within the network's coverage area, and experience a large number of handoffs during a typical connection. The wireless network users expect good QoS from the system, e.g., low delays, small call dropping and packet loss probabilities. The wireless network must provide the requested level of service even if an active mobile terminal moves to a congested cell. In this case, the corresponding base station must provide the expected service even if this implies denying network access to new connection requests. Ideally, base stations should deny network access to certain connection requests only when it is *strictly necessary*. This constitutes a problem that could only be optimized if knowledge were available regarding the future movement and call holding times of the active mobile terminals in the wireless network, as well as the future movements and call holding times of the mobile terminals[1] with connection requests. As in related problems, solutions close to optimal can often be obtained by using knowledge of *past events* to predict future behavior.

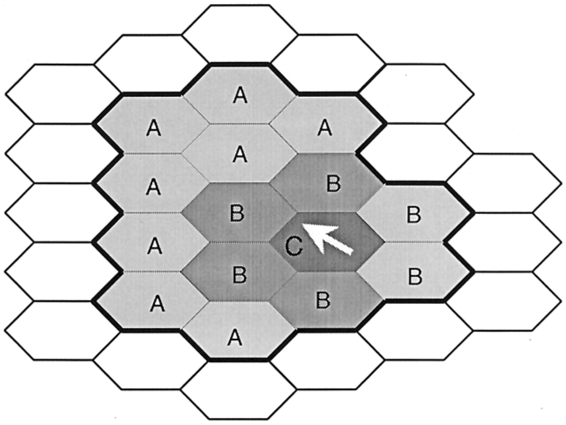


Fig. 1. Shadow cluster produced by an active mobile terminal. A denotes the nonbordering neighbor. B denotes the bordering neighbor, and C denotes the shadow cluster center.

The fundamental idea of the shadow cluster concept is 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*. As an active mobile terminal travels to other cells, the region of influence also moves, following the active mobile terminal to its new location. The base stations (and their cells) currently being influenced are said to form a *shadow cluster*, because the region of influence follows the movements of the active mobile terminal like a shadow, as shown in Fig. 1. The shadow (and therefore the level of influence) is strongest near the active mobile terminal, and fades away depending on factors such as the distance to the mobile terminal, current call holding time and priority, bandwidth resources being used, and the mobile terminal's trajectory and velocity. Because of these factors, the shape of a shadow cluster is usually not circular and can change over time.

We say that the center of a shadow cluster is not the geometric center of the area described by the shadow, but the cell where the mobile terminal is currently located. We also refer to this cell as the mobile terminal's *current home cell*. A *bordering neighbor* is a cell that shares a common border with the shadow cluster's center cell. In contrast, a *nonbordering*

[1] By active mobile terminal, we mean a mobile terminal that has a wireless connection and is consuming wireless resources.

neighbor cell, although being a part of the shadow cluster, does not share a border with the shadow cluster's center cell. In a manner consistent with the above definitions, we also use the terms *current home base station*, *bordering base station*, and *nonbordering base station*. Conceptually, the number and "darkness" of the shadows covering a cell reflect the amount of resources that the cell's base station needs to reserve in order to support the active mobile terminals currently in its own and in neighboring cells. With the information provided by shadow clusters, base stations can determine, for each new call request, whether the request can be supported by the wireless network. In practice, a shadow cluster is a virtual message system where base stations share probabilistic information with their neighbors on the likelihood that their active mobile terminals will move to neighbor cells (while remaining active) in the near future. With the information provided by shadow clusters, base stations project future demands and reserve resources accordingly. Base stations reserve resources by denying network access to new call requests, and by "waiting" for active users to end their calls.

The decision process for the acceptance of a new call request also involves a shadow cluster. Every new call request results in the implementation of a *tentative* shadow cluster. Base stations exchange information on their new call requests, and decide, based on this and other information, which requests should be accepted and which requests should be denied. When implementing shadow clusters, it is important to consider that the amount of information shared among neighbor base stations should be limited, so that the wireline network is not overburdened with control messages. In practice, after a mobile call has been admitted, only a small amount of information needs to be shared.

When an active mobile terminal is handed off to another cell, the shadow cluster moves. After a handoff, base stations within the old shadow cluster are notified about this movement, and the mobile terminal's new current base station has to assume the responsibility of supplying the appropriate information to the base stations within the new shadow cluster. Base stations which were in an old shadow cluster that has just moved away must delete any entries corresponding to the active mobile terminal that established the shadow cluster, and free reserved resources if appropriate. Base stations which become part of the influence region of a shadow cluster must be given appropriate information on the shadow cluster's active mobile terminal, such as the respective QoS requirements, e.g., bandwidth demands, call dropping probabilities, and any other useful information such as the wireless connection's elapsed time, for the establishment of the new shadow cluster. In return for the communication and processing overheads involved, the shadow cluster concept improves some QoS requirements of the network, providing the ability to control the call dropping probability by establishing resource reservation requirements and a call admission policy.

III. Active Mobile Probabilities

In a wireless network with the shadow clusters implemented on it, every base station must inform its neighbors about the future location probabilities of the active mobile terminals currently under its control. For each active mobile terminal, only those base stations which belong to that particular mobile terminal's shadow cluster are informed about these probabilities.

We consider a wireless network where the time is divided in equal intervals τ at $t = t_1, t_2, \cdots, t_m$. Let j denote a base station in the network, where $j \in J$, and J is the set of all base stations in the network. Since there is a one-to-one correspondence between the set of base stations and the set of cells in the wireless network being considered, we use the same notation to denote base stations and their cells in the network. Let x be a mobile terminal with an active wireless connection, where $x \in X$, and X is the set of all active mobile terminals at the present time. Excluding the current home base station, the set of base stations that form the shadow cluster of active mobile terminal x is denoted by $K(x)$, where $k \in K(x)$ is one of the base stations that form the shadow cluster. If active mobile terminal x is currently under the control of base station j, then it is the task of this base station to determine the projected *active mobile probabilities* $\mathbf{P}_{x,j,k}(t) = [P_{x,j,k}(t_1), P_{x,j,k}(t_2), \cdots, P_{x,j,k}(t_m)]$, of that mobile terminal x, currently in cell j, will be active in cell k (and therefore under the control of base station k) at times $t_1, t_2, \cdots, t_m$. Thus, the active mobile probabilities are the projected probabilities that a mobile terminal will remain active in the future *and* at a particular location (within a cell in its shadow cluster).

The probabilities $\mathbf{P}_{x,j,k}(t)$ can be interpreted as the minimum percentage of the total amount of resources currently being used by mobile terminal x that base station j recommends to base station k to have available at times $t = t_1, t_2, \cdots, t_m$ in the event that active mobile terminal x migrates to cell k. As time passes, mobile terminal x's current home base station may recompute and refine the estimates on the probabilities $\mathbf{P}_{x,j,k}(t)$, sending the respective updates to all base stations within the shadow cluster $K(x)$. Active mobile probability estimates for the short term future are likely to be more accurate than the corresponding estimates for the distant future.

Evidently, the probability that active mobile terminal x, currently in cell j, will be in cell k at times $t_1, t_2, \cdots, t_m$ constitutes a random process that is a function of several parameters. The more knowledge about the dynamics of a mobile terminal (past and present) and call holding patterns, the more complex and accurate the probability function is likely to become. For example, if precise knowledge about the current dynamics of a mobile terminal is available, the active mobile probability becomes a function of the position, velocity, and acceleration vectors of the mobile terminal. The active mobile probability function can also be refined with information such as the mobile terminal's past history, and with detailed knowledge of the geographical features of the region where the mobile terminal is currently located. For example, if the mobile terminal's position coincides with that of a highway, both the mobile terminal's general velocity and direction can be predicted to a significant extent.

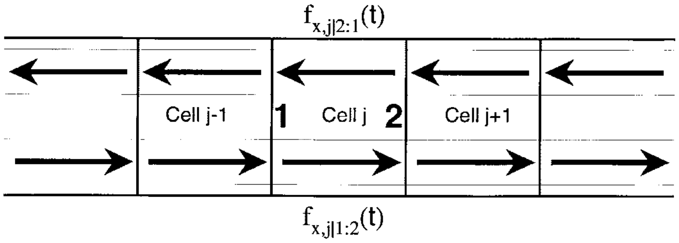


Fig. 2. Numbering of cell's sides for the one-dimensional (1-D) case.

A. Calculation of Active Mobile Probabilities for One-Dimensional Case

For the sake of simplicity, we first calculate the active mobile probabilities $P_{x,j,k}(t)$ for the case where mobile terminals travel along a highway. We assume that several cells of rectangular shape are used to cover the highway, as shown in Fig. 2. We also assume that the exact position of the mobile terminals within each cell is unknown. The two sides where a mobile terminal can enter or leave a cell are numbered (arbitrarily) 1 and 2. We say that a mobile terminal which is currently in cell j is traveling in a "right" direction if the next cell it is going to visit is cell $j+1$. Otherwise, we say that the mobile terminal is going in the "left" direction.

Consider an active mobile terminal x traveling through the highway. We assume that this mobile terminal has been observed in the past, and that probabilistic information on the times this mobile terminal typically spends in the cells along the highway has been collected and is available. We define a *residence time probability density function (pdf) matrix* $\mathbf{f}_{x,j}(t)$ for mobile terminal x in cell j. The matrix has the general form

$$\mathbf{f}_{x,j}(t) = \begin{bmatrix} 0 & f_{x,j|1:2}(t) \\ f_{x,j|2:1}(t) & 0 \end{bmatrix} \tag{1}$$

where $f_{x,j|v:w}(t)$, for $v, w = 1, 2$, is the pdf of the residence time of mobile terminal x in cell j, given that the mobile terminal enters the cell through side v, and leaves the cell through side w, and given that the mobile terminal does not turn around. Thus, there are pdf's for each direction of travel.

We also consider that an *initial handoff probability vector* $\vec{\Upsilon}_{x,j}(t)$ is available, with the general form

$$\vec{\Upsilon}_{x,j}(t) = [\Upsilon_{x,j|:0}(t) \quad \Upsilon_{x,j|:1}(t) \quad \Upsilon_{x,j|:2}(t)] \tag{2}$$

where $\Upsilon_{x,j|:0}(t)$ is the probability that mobile terminal x will remain in cell j, given that the call was initiated while in cell j, while $\Upsilon_{x,j|:w}(t)$, for $w = 1, 2$, is the probability that mobile terminal x will have left cell j by time t through side w, given that the call was initiated while in cell j. The handoff probabilities of this vector are required because unless a mobile terminal has crossed a cell boundary, we are assuming it is not possible to determine the initial direction of travel of the terminal that is requesting admission to cell j. Likewise, we define an *initial residence time pdf vector* $\vec{g}_{x,j}(t)$ with the general form

$$\vec{g}_{x,j}(t) = [g_{x,j|:1}(t) \quad g_{x,j|:2}(t)] \tag{3}$$

where $g_{x,j|:w}(t)$, for $w = 1, 2$, describes the residence time distribution of mobile terminal x in cell j, given that the call is initiated in cell j, and that the mobile terminal exits the cell through side w.

We need to consider another category of pdf's before we can determine the active mobile probabilities of mobile terminal x. Since the active mobile probabilities are the projected probabilities that a mobile terminal will *remain active* and it will be at a specific location, we consider *active pdf's* of the form $h_{x,M(x)}(t)$ which represents the distribution of call lengths for mobile terminal x when using a service with class descriptor $M(x)$. The class descriptor specifies the service being requested/used by the mobile terminal, e.g., video, audio, voice, and fax. $M(x)$ affects the required QoS. We assume that $h_{x,M(x)}(t)$ is independent of the dynamics of mobile terminal x. Moreover, we assume that $h_{x,M(x)}(t)$ can be constructed by measuring the connection times of mobile terminal x when using the service with class descriptor $M(x)$.

With all these considerations in place, we now can compute active mobile probabilities. For a mobile terminal x that is initiating a call of class $M(x)$ while in cell j at time t, the active mobile probability $P_{x,j,j,}(t)$ for this cell can be determined by using the following expression:

$$\begin{aligned} P_{x,j,j}(t) &= [1 - H_{x,M(x)}(t)] \\ &\cdot \left\{ \Upsilon_{x,j|:0}(t) + \sum_{w=1}^{2} [1 - G_{x,j|:w}(t)] \cdot \Upsilon_{x,j|:w}(t) \right\} \end{aligned} \tag{4}$$

for $x = 1, 2, \cdots, |X|$, $j = 1, 2, \cdots, |J|$, and $|M(x)|$ the total number of call classes. Here, $H_{x,M(x)}(t)$ and $G_{x,j|:w}(t)$ are the cumulative distribution functions (cdf's) for $h_{x,M(x)}(t)$ and $g_{x,j|:w}(t)$, respectively. Equation (4) is the result of two independent probabilities: the probability that the call will not end by time t, and the probability that the mobile terminal will still be in cell j at time t. Here, $[1 - H_{x,M(x)}(t)]$ is the probability that the call is not over by time t, and $[1 - G_{x,j|:w}(t)] \cdot \Upsilon_{x,j|:w}$, for $w = 1, 2$, are the probabilities that the mobile terminal is still in cell j at time t, considering two possible travel directions of this mobile terminal.

Next, we obtain the active mobile probabilities $\mathbf{P}_{x,j,k}(t)$ for neighboring cells. Here, we consider the case where a mobile terminal is traveling to the right. We need to determine the *residence time pdf's* $q_{x,j,k}(t)$, for $k > j$, which describe the probable residence times of mobile terminal x in cells j to k, given that the mobile terminal is initially in cell j. We assume that the residence time pdf's for mobile terminal x in other cells are available, and that the pdf's are independent. Thus, we can determine $q_{x,j,k}(t)$ by convolving the pdf's which correspond to the residence times of the cells that mobile terminal x will visit en route to cell k. Note that this is equivalent to finding the pdf of a random variable that is the sum of independent random variables (which describe the time spent in each cell). If mobile terminal x initiates a wireless connection while in cell j, then the residence time pdf $q_{x,j,k}(t)$ becomes

$$\begin{aligned} q_{x,j,k}(t) &= g_{x,j|:2}(t) \circledast f_{x,j+1|1:2}(t) \\ &\circledast \cdots \circledast f_{x,k-1|1:2}(t) \circledast f_{x,k|1:2}(t). \end{aligned} \tag{5}$$

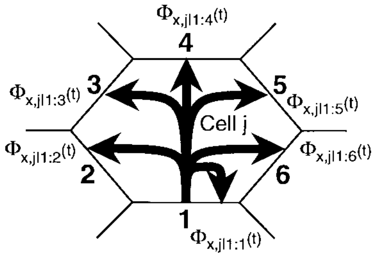


Fig. 3. Numbering of cell's sides and handoff probabilities, 2-D case.

If $Q_{x,j,k}(t)$ is the cumulative pdf for $q_{x,j,k}(t)$, then $Q_{x,j,k}(t)$ is the probability that the mobile terminal will be beyond cell k at time t, and $[Q_{x,j,k-1}(t) - Q_{x,j,k}(t)]$ is the probability that the mobile terminal will be in cell k at time t. Note that $[Q_{x,j,k-1}(t) \geq Q_{x,j,k}(t)]$ for all $t \geq 0$.

Finally, we obtain the active mobile probability $P_{x,j,k}(t)$ that mobile terminal x, currently in cell j, will be in cell k at time t:

$$P_{x,j,k}(t) = [1 - H_{x,M(x)}(t)] \cdot \{[Q_{x,j,k-1}(t) - Q_{x,j,k}(t)] \cdot \Upsilon_{x,j|:2}(t)\}. \quad (6)$$

Note that similar expressions can easily be obtained for the case when a mobile terminal is traveling in a left direction.

B. Calculation of Active Mobile Probabilities for Two-Dimensional Case

We now determine the active mobile probabilities for the general two-dimensional (2-D) case, assuming that the exact position and dynamics of the mobile terminal are unknown. General knowledge of the mobile terminal's last position is again limited to the area where the last handoff occurred. We assume that cells in the wireless network are of hexagonal shape, and the sides in each cell are numbered (arbitrarily) $1, 2, \cdots, 6$, as shown in Fig. 3.

Consider an active mobile terminal x traveling through a wireless network that has been observed several times in the past. We assume that probabilistic information about this mobile terminal is available. This information is in the form of pdf's which describe the times spent by the mobile terminal in different cells, and the handoff probabilities that the mobile terminal will move from a given cell to a neighbor cell. We consider that the users of the mobile terminals are "well-behaved" (users do not travel at "random" through the cells of a wireless network, but rather, they move through paths that can be predicted to some extent [3], [4]). For cell j, a *handoff probability matrix* $\mathbf{\Phi}_{x,j}(t)$ is available for mobile terminal x.[2] This matrix describes the probability that mobile terminal x, given that it enters cell j through a particular side, will remain in the cell, or that it will exit the cell through a specific side. The matrix $\mathbf{\Phi}_{x,j}(t)$ is dependent on t because the handoff behavior of a mobile terminal is likely to be different depending on the day of the week as well as on the time of the day. The handoff probability matrix $\mathbf{\Phi}_{x,j}(t)$ is a 7×6 matrix with elements $\Phi_{x,j|v:0}(t)$, for $v = 1, 2, \cdots, 6$, which is the probability that mobile terminal x will remain in cell j, given that it enters the cell through side v, and $\Phi_{x,j|v:w}(t)$, for $v, w = 1, 2, \cdots, 6$, is the probability that mobile terminal x will leave cell j through side w, given that it enters the cell through side v.

We also define a *residence time pdf matrix* $\mathbf{f}_{x,j}(t)$ for mobile terminal x in cell j, which is the extended form of (1), with elements $f_{x,j|v:w}(t)$, for $v, w = 1, 2, \cdots, 6$, which is the pdf of the residence time of mobile terminal x in cell j, given that the mobile terminal enters the cell through side v, and leaves through side w.

We also consider that an *initial handoff probability vector* $\vec{\Upsilon}_{x,j}(t)$ [expanded form of (2)] is available for mobile terminal x in cell j

$$\vec{\Upsilon}_{x,j}(t) = [\Upsilon_{x,j|:0}(t) \quad \Upsilon_{x,j|:1}(t) \quad \cdots \quad \Upsilon_{x,j|:6}(t)] \quad (7)$$

where $\Upsilon_{x,j|:0}(t)$ is the probability that mobile terminal x will remain in cell j, given that the call was initiated while in cell j, while $\Upsilon_{x,j|:w}(t)$, for $w = 1, 2, \cdots, 6$, is the probability that mobile terminal x will leave cell j through side w, given also that the call was initiated while in cell j.

We also define an *initial residence time pdf vector* $\vec{g}_{x,j}(t)$, which is the extended form of (3):

$$\vec{g}_{x,j}(t) = [g_{x,j|:1}(t) \quad \cdots \quad g_{x,j|:6}(t)] \quad (8)$$

where $g_{x,j|:w}(t)$, for $w = 1, 2, \cdots, 6$, describes the probable residence times of mobile terminal x in cell j, given that the call is initiated in cell j, and that the mobile terminal exits the cell through side w.

As in Section III-A, we also consider *active pdf's* (computed from empirical data) of the form $h_{x,M(x)}(t)$ that represent the distribution of call lengths for mobile terminal x when using a service with class descriptor $M(x)$. We also assume that $h_{x,M(x)}(t)$ is independent of the dynamics of mobile terminal x.

For a mobile terminal x that is initiating a call of class $M(x)$ while in cell j, the active mobile probability for this cell (assuming that the mobile terminal does not move out and back into the cell for the time being considered) can be determined by using the initial residence time pdf's for the mobile terminal, weighted by the corresponding initial handoff probabilities [this is the general form of (4)]

$$P_{x,j,j}(t) = [1 - H_{x,M(x)}(t)] \cdot \left\{ \Upsilon_{x,j|:0}(t) + \sum_{w=1}^{6} [1 - G_{x,j|:w}(t)] \cdot \Upsilon_{x,j|:w}(t) \right\} \quad (9)$$

where $H_{x,M(x)}(t)$ and $G_{x,j|:w}(t)$ are the cdf's for $h_{x,M(x)}(t)$ and $g_{x,j|:w}(t)$, respectively. Here again $[1 - H_{x,M(x)}(t)]$ is the probability that the call will not end by time t, and $[1 - G_{x,j|:w}(t)]$ is the probability that the mobile terminal x is still in cell j at time t, given that it is going to leave the cell through side w. Equation (9) has the term $\Upsilon_{x,j|:0}(t)$ because there is a possibility that the mobile terminal will remain in cell j for the duration of the call.

We now compute active mobile probabilities for a cell other than the current cell. Since potentially there can be

[2] Note that handoff probability matrices are not required for the one-dimensional case since a mobile terminal in a highway will move to the next cell along its direction of travel with probability 1.

an infinite number of different routes which can be taken when traveling between two cells, we must limit the number of possible routes to be used in the calculations. Therefore, only the most common or efficient routes (e.g., those with short travel times, i.e., those minimizing the distance of travel) should be considered. In order to determine the active mobile probability of mobile terminal x for cell k, given that the mobile terminal is currently in cell j and has just initiated a call, we need to consider the following. Mobile terminal x can enter cell k through any of the cell's sides $v = 1, 2, \cdots, 6$. For a particular side v, there can be $r = 1, 2, \cdots$ different routes which can be taken. The probability $R_{x, r|v}$ of taking a particular route can be calculated by multiplying the initial handoff probability $\Upsilon_{x, j|:w(r|v, j)}(t)$ when leaving cell j by the handoff probabilities $\Phi_{x, l|v(r|v, l):w(r|v, l)}(t)$ of the cells along the route, e.g.,

$$R_{x, r|v} = \Upsilon_{x, j|:w(r|v, j)}(t) \prod_{l} \Phi_{x, l|v(r|v, l):w(r|v, l)}(t) \tag{10}$$

where l is a cell along the route $r|v$, and $v(r|v,\ l)$ and $w(r|v,\ l)$ are the entrance and exit sides (in cell l), respectively, for the route along cell l.

In addition, to compute the probability $P_{x, j, k}(t)$ that mobile terminal x is in cell k at time t, we need to determine two residence time pdf's: $q_{x, j, m|(r|v)}(t)$, the residence time pdf for the cells along the route and up to cell m, where cell m is the last cell along route $(r|v)$ before cell k is reached, and $q_{x, j, k|(r|v)}(t)$, the residence time pdf for the cells along the route and up to cell k. Again, assuming independence of the residence time pdf's for all cells, the pdf $q_{x, j, m|(r|v)}(t)$, when using route $r|v$, can be calculated by the convolution of the residence time pdf's along the "route of travel" as follows:

$$\begin{aligned} q_{x, j, m|(r|v)}(t) = & g_{x, j|:w(r|v, j)}(t) \circledast \cdots \\ & \circledast f_{x, l|v(r|v, l):w(r|v, l)}(t) \circledast \cdots \\ & \circledast f_{x, m|v(r|v, m):w(r|v, m)}(t). \end{aligned} \tag{11}$$

The residence time $q_{x, j, k|(r|v):w}(t)$, given that the mobile terminal enters cell k immediately after leaving cell m, and then leaves cell k through side w becomes

$$q_{x, k, k|(r|v):w}(t) = q_{x, j, m|(r|v)}(t) \circledast f_{x, k|v(r|v, k):w}(t) \tag{12}$$

where (12) is the extended form of (5). $[Q_{x, j, m|(r|v)}(t) - Q_{x, j, k|(r|v):w}(t)]$ is the probability that mobile terminal x will be in cell k at time t, given that it takes route $r|v$ from cell j to cell k, and that it will leave cell k through side w.

Finally, we provide a general expression for the active mobile probability $P_{x, j, k}(t)$ that mobile terminal x, currently in cell j, will be active and in cell k at time t, which is the extended form of (6):

$$\begin{aligned} P_{x, j, k}(t) = & \\ & [1 - H_{x, M(x)}(t)] \sum_{v=1}^{6} \sum_{r} R_{x, r|v} \\ & \cdot \left(\sum_{w=1}^{6} [Q_{x, j, m|(r|v)}(t) - Q_{x, j, k|(r|v):w}(t)] \cdot \Phi_{x, k|v:w} \right) \end{aligned} \tag{13}$$

i.e., the active mobile probability that mobile terminal x, currently in cell j, will be in cell k at time t is the result of two probabilities: the probability that the call will continue at time t, and the probability that the mobile terminal will be in cell k at time t. To compute the probability that a mobile terminal x will be in cell k at time t, given that initially it is in cell j, we need to consider the following:

1) all possible sides v where mobile terminal x can enter cell k;
2) all possible routes $r|v$ from cell j to cell k;
3) the probability that the mobile terminal will stay in cell k at time t given that the terminal later exits the cell through side w.

Although the complexity of (13) seems to be high, the calculation of active mobile probabilities $\mathbf{P}_{x, j, k}(t)$ should be reasonable in practice. This is because in most situations, several of the handoff probabilities in $\mathbf{\Phi}_{x, j}(t)$ and $\vec{\Upsilon}_{x, j}(t)$, as well as several of the residence time pdf's in $\mathbf{f}_{x, j}(t)$ and $\vec{g}_{x, j}(t)$, for a given mobile terminal x in cell j will be zero, or will be taken as equal to average values obtained after measuring handoffs and residence time pdf's of several mobile terminals. Most mobile terminal users have well-defined routes and behaviors [4] when driving. Moreover, here we consider the general case where it is possible to cross any cell, entering and exiting through any side, which is not possible in most microcells.

The introduction of vehicle navigation systems as well as the future development of autonomous vehicle navigation and intelligent highway systems [9] will increase the accuracy and ease the determination of active mobile probabilities—the exact traveling routes may be known, and the determination of active mobile probabilities will be much easier.

Note that although we have considered cells of hexagonal shape, the methodology developed above for the determination of active mobile probabilities can easily be modified for cells with an arbitrary number of sides. Next, we proceed to describe in detail how the active mobile probabilities can be used to determine the amount of resources required by a base station, so that most current active mobile terminals within the base station's cell and nearby cells can engage in successful handoffs.

IV. Resource Demand Calculations

Consider a wireless network system where the principal responsibilities of a base station are to provide (with a reasonable probability of success): 1) incoming handoffs from active mobile terminals with different bandwidth requirements and 2) uninterrupted service for new connection requests accepted by the base station.

In this section, we show how the shadow cluster concept can contribute to the achievement of these objectives by predicting the amount of resources that a station will require in its near future.

We consider that every base station has a total of C *bandwidth units* (BU's), where a bandwidth unit is the minimum quota of (uplink) bandwidth resources that can be assigned to any mobile terminal. For example, while a voice call user may require the least bandwidth, i.e., a single BU, a video mobile terminal will require several BU's. The parameter C comprises the number of free bandwidth units C_f and the number C_u of bandwidth units that are currently being used. The number of free and "used" BU's varies over time. The total number of BU's is

$$C = C_u(t) + C_f(t). \qquad (14)$$

Here, we consider that the parameter C is constant over time. For a TDMA or FDMA system, this situation corresponds to a fixed channel allocation (FCA) discipline [15]. However, the approach that we propose next can be extended to include values of C that vary over time (conversely, the approach can be used to provide channel requirement forecasts to a dynamic channel allocation (DCA) scheme [2], [8], [10], [13]).

The number of bandwidth units being used C_{u_j} in a cell j varies over time. When a base station receives handoff requests, it has to honor them as best as it can, trying to minimize the number of calls that will need to be dropped because of insufficient resources. Dropped calls are generally the result of *cell overloading,* a condition characterized by an excessive number of active users that results in an insufficient number of free resources left in a cell receiving a handoff. Since a base station can control neither the number of handoff requests nor the number of call disconnection requests, the best mechanism available to the base station to prevent/reduce overloading is to limit the number of newly accepted connections. Obviously, drastic measures could be taken, i.e., a wireless connection can be either terminated prematurely or dropped during a handoff. This should be avoided as much as possible so as to maintain a good QoS.

Using information provided by shadow clusters, each base station can obtain estimates on the number of BU's that are going to be used in the near future, *considering only the current active mobile terminals in the wireless network.* Based on these estimates, base stations can then decide, using a distributed algorithm, whether or not to accept new connection requests. In order to obtain a resource demand algorithm, we consider the following. Assuming that an active mobile terminal cannot cross a cell boundary multiple times within the time interval τ being considered, we observe that an active mobile terminal that is in a reference cell has three possible outcomes.

1) It can remain active and stay within the reference cell.
2) It can terminate the connection without leaving the cell.
3) It can remain active and move to a neighboring cell.

The number of future active mobile terminals within a reference cell depends also on the active mobile terminals currently within a neighbor cell but moving toward the reference cell.

Based on these observations, we can determine the components for an estimate on the number of BU's $\widetilde{C_{u_j}(t)}$ to be used by base station j at times $t = t_1, t_2, \cdots, t_m$ by

$$\widetilde{C_{u_j}(t)} = C_{u_j}(t_o) - C^{\dagger}_{u_j}(t) + C^{\ddagger}_{u_j}(t) \qquad (15)$$

where $C_{u_j}(t_0)$ the initial number of busy BU's, is a known quantity, $C^{\dagger}_{u_j}(t)$ is the estimate of the number of BU's which will become free by active users who end their calls or emigrate to other cells by time t, and $C^{\ddagger}_{u_j}(t)$ is the estimate of the number of BU's which will become busy due to handoffs by external mobile terminals moving from neighbor cells within the shadow cluster to cell j.

Let $c(x)$ denote the number of BU's being used by active mobile terminal x, and let $X_j \subset X$ be the set of all active mobile terminals within cell j. Then $C^{\dagger}_{u_j}(t)$ can be calculated by

$$C^{\dagger}_{u_j}(t) = \sum_{x \in X_j} \left[1 - P_{x,\,j,\,j}(t)\right] \cdot c(x) \qquad (16)$$

where $P_{x,\,j,\,j}(t)$ can be obtained by using (4) or (9). Likewise, $C^{\ddagger}_{u_j}(t)$ can be calculated by

$$C^{\ddagger}_{u_j}(t) = \sum_{x \notin X_j,\, j \in K(x)} P_{x,\,k,\,j}(t) \cdot c(x) \qquad (17)$$

where $P_{x,\,k,\,j}(t)$ can be obtained by (6) or (13).

Note that the estimate $\widetilde{C_{u_j}(t)}$ of the number of BU's to be used by base station j at time t can be greater than C_j, the total number of BU's available in cell j. Obviously, the actual value of this estimate, $C_{u_j}(t)$, is always smaller than or equal to C_j. If $\widetilde{C_{u_j}(t)} > C_j$ then some calls might need to be dropped in cell j at time t. Note that base stations will not need to send individual values of active mobile probabilities to their neighbors. Rather, base stations will only need to transmit estimates of the number of BU's to be used in the future in other cells. For example, base station k will send estimates to base station j of the form $\sum P_{x,\,k,\,j}(t_1) \cdot c(x)$, $\sum P_{x,\,k,\,j}(t_2) \cdot c(x)$, $\cdots$, $\sum P_{x,\,k,\,j}(t_m) \cdot c(x)$, where $x \in X_k$ and $j \in K(x)$.

Ideally, base station j recomputes active mobile probability estimates at each time step, and determines the values for the expected number of BU's to be used in the next $t_1, t_2, \cdots, t_m$ steps. When these estimates are obtained, a base station j can determine the estimates for the number of free BU's $\widetilde{C_{f_j}(t_1)}$, $\widetilde{C_{f_j}(t_2)}$, $\cdots$, $\widetilde{C_{f_j}(t)}$ for the next $t_1, t_2, \cdots, t_m$ time steps simply by using

$$\widetilde{C_{f_j}(t)} = \begin{cases} C_j - \widetilde{C_{u_j}(t)}, & \text{if } C_j - \widetilde{C_{u_j}(t)} \geq 0 \\ 0, & \text{otherwise.} \end{cases} \qquad (18)$$

Note that the above estimates are based only on the current active mobile terminals in the network, and considering that no new connection requests are accepted for the next $t_1, t_2, \cdots, t_m$ time steps. If the estimates for $\widetilde{C_{u_j}(t)}$ yield values smaller than C, then base station j will probably be able to admit new active mobile terminals to the wireless network. In the following section, we use the estimates developed in (15)–(18), and propose a decision algorithm for the acceptance of new mobile terminals to the wireless network.

V. Call Admission Algorithm Based on Shadow Clusters

The call admission algorithm is implemented in a distributed fashion, with every base station exchanging information with its bordering and nonbordering neighbors periodically. As before, we assume that the time is quantized in slots of length

τ. Also, we assume that new call requests are reported at the beginning of each time slot, and that a decision regarding an admission request is made sometime before the end of the same time slot where the request was received.

The call admission algorithm is applicable to both mobile terminal-initiated calls and calls initiated by users connected to the wireline network (once the target mobile terminal has already been located). Without loss of generality, we explain the call admission algorithm from the perspective of base station j. The sequence of steps that base station j executes in every time slot is the following.

1) Base station j gathers call connection requests from mobile terminals within its cell. Each received request includes a description of the desired QoS call dropping parameter and the number of BU's being requested for the call. The base station checks whether its current resources can support the requested parameters, e.g., for mobile terminal x requesting admission to the wireless network, base station j checks if

$$c(x) \leq C_{f_j}(t_0) \tag{19}$$

where $C_{f_j}(t_0)$ is determined using (18). If this condition is not satisfied, the request is immediately turned down. Note that if several mobile terminals request admission to a particular cell at the same time, and if the sum of the bandwidth resources being requested exceeds the number of BU's that are currently available, i.e., if

$$\sum_x c(x) > C_{f_j}(t_0) \tag{20}$$

then the decision of which connection requests should be turned down is resolved in Steps 5) and 6) of this algorithm.

2) Base station j defines a shadow cluster $K(x)$ for each mobile terminal x making an admission request. In other words, base station j determines how many base stations should be affected by each admission request, and how many active mobile probabilities [(6) and (13)] it should compute (how far into the future) for each base station in the shadow cluster. The extension of the new shadow cluster depends on the velocity of the mobile terminal as well as on the average call length of the mobile terminal making the admission request. The base station also informs its neighbors about the preliminary active mobile probabilities [also computed using (6) and (13)] and the number of BU's $c(x)$ being requested by the mobile terminals within its cell.

3) Within the same time slot, base station j receives preliminary estimates on active mobile probabilities and number of requested BU's from neighbor base stations, which have previously received their own admission requests. Base station j receives this information only if it is within the shadow cluster of a new connection request that is currently in another cell. Based on the active mobile probabilities from its own new requests, and on active mobile probabilities from current requests in neighbor cells, base station j computes *availability estimates* $\Delta_j(t)$ for the future $t = t_1, t_2, \cdots, t_m$ time steps by using the following expression:

$$\Delta_j(t) = \frac{\widetilde{C_{f_j}(t)}}{\sum\limits_{x \in X_j} P_{x,j,j}(t)c(x) + \sum\limits_{x \notin X_j} P_{x,k,j}(t)c(x)} \tag{21}$$

where $\widetilde{C_{f_j}(t)}$ is obtained using (18). Equation (21) yields a value that is directly proportional to the total number of BU's expected to be free in cell j at time t, and inversely proportional to the number of BU's that would be required in the cell if *all of the new call requests in the wireless network at time* t_0, *were admitted to the network*. The higher the value of $\Delta_j(t)$, the more likely a mobile terminal in cell j will have bandwidth resources available at time t. Base station j distributes the availability estimates $\Delta_j(t_1), \Delta_j(t_2), \cdots, \Delta_j(t_m)$ among the base stations which sent data about their own current connection requests earlier. Conversely, base station j receives availability estimates from all base stations which fall within the shadow clusters of base station j's current call requests. Note that if a given base station k is part of two or more shadow clusters defined by connection requests in cell j, then base station j receives only a single set of availability estimates from base station k.

4) After having collected availability estimates (21) from neighboring base stations, base station j computes *survivability estimates* $\Lambda_j(x, \hat{t}_x)$ for each mobile terminal x making a connection request in cell j, where $\hat{t}_x$ is the known average call duration time for mobile terminal x. The survivability estimate of mobile terminal x for the next time steps $t_1, t_2, \cdots \hat{t}_x$ can be calculated from

$$\Lambda_j(x, \hat{t}_x) = \frac{1}{\hat{t}_x} \sum_{t=t_1}^{\hat{t}_x} \left[\Delta_j(t) P_{x,j,j}(t) + \sum_{k \in K(x)} \Delta_k(t) P_{x,j,k}(t) \right]. \tag{22}$$

Equation (22) is a sum of the active mobile probabilities of mobile terminal x weighted by the availability estimates of the different cells that the mobile terminal might visit up to time $\hat{t}_x$. In order to prevent that the survivability estimates are biased in favor of calls with long durations, the sum in (22) is normalized by the inverse of the average duration of the call, i.e., $1/\hat{t}_x$. In (22), equal treatment is given for all connection requests regardless of the number of BU's being requested. However, (22) can be adjusted to give preference to new connection requests with multiple BU's. The higher the value of $\Lambda_j(x, \hat{t}_x)$, the more likely mobile terminal x will survive until time $\hat{t}_x$. Thus, mobile terminals with higher values for their survivability estimate should be accepted more often than mobile terminals with lower values.

5) The preliminary decision on whether to accept or reject a connection request involves a decision function

$\Omega[\Lambda(x, \hat{t}_x), D_{dp}(x)]$, where $D_{dp}(x)$ is the *dropping probability* QoS parameter, i.e., the maximum dropping probability allowed for the call being requested by mobile terminal x. The Ω function is defined to favor the acceptance of mobile terminals with higher overall survivability estimates (22), provided that the requested dropping probability parameter is satisfied. To construct the Ω function, experiments must be conducted to determine the minimum values of the survivability estimate $\Lambda(x, \hat{t}_x)$ that can support certain call dropping probability values $D_{dp}(x)$, given a number of BU's $c(x)$ being requested. Thus, the Ω function may be implemented with a lookup table. The Ω function returns an *acceptance value* that is equal to the difference between the survivability estimate (22) for the mobile terminal making the admission request, and the survivability estimate that is required to support the requested call dropping probability (obtained from experiments). Therefore, positive values of the Ω function indicate that the connection request can be honored.

6) The base station tries to accept all mobile terminals that have a positive acceptance value. The admission process is done in order, starting with the mobile terminal that has the highest acceptance value, and ending with either the mobile terminal that has the lowest (positive) acceptance value, or when there are no more free BU's available for the mobile terminals making admission requests. Other admission order criteria may be defined provided that an accepted mobile terminal has a positive acceptance value.

VI. Performance Evaluation

For the sake of simplicity, we evaluate the performance of the shadow cluster concept for mobile terminals which are moving along a highway. In our simulation model we have the following assumptions.

1) The time is quantized in intervals $\tau = 10$ s.
2) During each time interval, mobile terminals are generated in each cell according to a Bernoulli process. Newly generated mobile terminals can appear anywhere along a cell with equal probability.
3) Mobile terminals can have speeds of: 70, 90, or 105 km/h. The probability of each speed is 1/3, and mobile terminals can travel in either of two directions with equal probability.
4) Mobile terminals can transmit three types of traffic: voice, audio, or video. The probabilities of these types are 0.7, 0.2, and 0.1, respectively.
5) Call holding times are the same for all call types. The call holding times are exponentially distributed with mean value equal to 180 s.
6) All mobile terminals have the same maximum call dropping probability specification.
7) The highway is covered by 10 cells, laid at 1-km intervals.
8) Each cell has a capacity of 40 BU's.
9) The number of BU's required by each call type is: voice $= 1$, audio $= 5$, video $= 10$.

We simulated a total of 4 h of real-time highway traffic, with a constant cell load equal to 360 new calls/h/cell. For simplicity, we assume that the base stations can determine the speeds of the mobile terminals. Since the call holding times of all mobile terminals are exponentially distributed with a mean value equal to 180 s, the pdf that base stations use to determine the probability values for different call lengths is $h(t) = (1/180) \cdot e^{-t/180}$. When a mobile terminal is generated, its initial position within a cell is not known. Since the speed of the mobile terminal is known (but not its direction), the probability $g(t)$ that the mobile terminal remains in the cell where it was generated, at time t, is uniformly distributed, where the maximum time within the cell is d/s, where d is the cell length, and s is the speed of the mobile terminal. Since we assume that the initial direction is not known, the probability of traveling in a right or left direction is $\Upsilon(t) = 0.5$. With the values for $h(t)$, $g(t)$, and $\Upsilon(t)$, base stations can determine active mobile probabilities (for mobile terminals that are either already admitted or requesting admission to the network), using (4) and (6). When a mobile terminal executes a handoff, then the position of the mobile terminal is known, and it is possible to determine the times when the mobile terminal will be in other cells.

In our simulations, we considered that for each time step τ, base stations determine active mobile probabilities for the next 180 s. In other words, since the time interval $\tau = 10$ s, each base station calculates 18 active mobile probabilities for each mobile terminal and for each cell. Note that since for some cells the probability that a mobile terminal will be in those cells at a given time may be zero, several of the calculated active mobile probabilities $\mathbf{P}_{x,j,k}(t)$ of (4) and (6) will be nil. After active mobile probabilities have been determined, a base station j can compute estimates $C^{\dagger}_{u_j}(t)$ [using (16)] of the number of BU's that will become free by active mobile terminals that either end their calls while in cell j or that migrate to other cells. Base station j receives estimates $C^{\ddagger}_{u_j}(t)$ (17) from other base stations on the number of BU's that will be required to support active mobile terminals which may move to cell j at time t. Base station j then sends corresponding estimates to the other base stations. With $C^{\dagger}_{u_j}(t)$ and $C^{\ddagger}_{u_j}(t)$, base station j will estimate [using (18)] the number of BU's that will be available $[\widetilde{C_{f_j}(t_1)}, \widetilde{C_{f_j}(t_2)}, \cdots, \widetilde{C_{f_j}(t)}]$ in the next time steps.

Since in this model mobile terminals are generated according to a Bernoulli process, up to one new mobile terminal may be generated in each cell during each time step τ. If a base station j receives an admission request from newly-generated mobile terminal x, it first checks whether it has enough BU's to support this call, i.e., it checks if $c(x) \leq C_{f_j}(t_0)$. If it has enough resources, the base station computes active mobile probabilities $\mathbf{P}_{x,j,k}(t)$ (4) and (6) for its own and the other cells along the highway, and for a total of 18 time steps in the future [Step 2) of the admission algorithm]. Base station j then shares this probability information with other cells which fall within the shadow clusters of its current mobile terminals, and receives active mobile probabilities from other base stations (with shadow clusters covering cell j) which also had call admission requests in the present time slot. With this information, base station j computes availability estimates $\Delta_j(t)$ for the next 18 time intervals (21), and shares this information with other cells [Step 3) of the admission algorithm]. Base station j then computes the survivability estimate $\Lambda_j(x, \hat{t}_x)$ for the mobile terminal x using (22) [Step 4)].

In our simulations, a new mobile terminal x was accepted into a cell only if its survivability value $\Lambda_j(x, \hat{t}_x)$ (22) was above a *rejection threshold* value. We varied the rejection threshold value to observe its effect on the call admission and call dropping percentages, as well as on the average bandwidth utilization in the cells of the network. The call of a mobile terminal x was dropped (terminated) if the mobile terminal needed to handoff to a cell j that did not have enough BU's to support the call, i.e., if $c(x) > C_{f_j}(t)$ (19). By varying the value of the rejection threshold in the simulations, we were able to determine the minimum survivability values necessary to support different call dropping rates for the conditions in the wireless network stipulated in the assumptions (the Ω function described in Section V can be constructed using this information). In other words, in this case the rejection threshold was equal to the minimum survivability value required by the Ω function to admit a new call into the network.

In Fig. 4, we depict the average bandwidth utilization experienced by the cells in the network, and the percentage of calls that are dropped. The top curve corresponds to the average bandwidth utilization, which is equal to the average number of BU's that are used in all cells in the network, considering the entire simulation time. The maximum bandwidth utilization occurs when the rejection threshold is equal to zero. In this case, the maximum bandwidth utilization is approximately equal to 30 BU's. This value could be higher if no video users were allowed in the network. The bottom curve depicts the percentage of dropped calls in the network. The highest percentage of dropped calls also occurs when the rejection threshold is equal to zero, i.e., when all mobile terminals are accepted regardless of their survivability estimate, as long as there are available BU's in the cells where the mobile terminals make their admission requests. For the simulated cell load, the maximum percentage of dropped calls is almost equal to 14%. By controlling the admission of new calls in terms of the rejection threshold, the shadow cluster mechanism can easily control the percentage of calls that will be dropped. With a rejection threshold equal to 70, the percentage of dropped calls is reduced to a value that is less than 1%. The shadow cluster mechanism allows to tradeoff average bandwidth utilization for percentage of dropped calls. Using the shadow cluster mechanism, it is easy to achieve a reduction in the percentage of dropped calls from 14% to 1%, which results in a reduction of the average bandwidth utilization from 30 BU's, to approximately 25 BU's. In this case, no calls need to be dropped if the rejection threshold is increased to about 90, with a corresponding average bandwidth utilization above 20 BU's. Thus, by carefully determining the mobile terminals which should be admitted to the network, the shadow cluster allows the reduction of the percentage of dropped calls with an acceptable degradation in total bandwidth utilization.

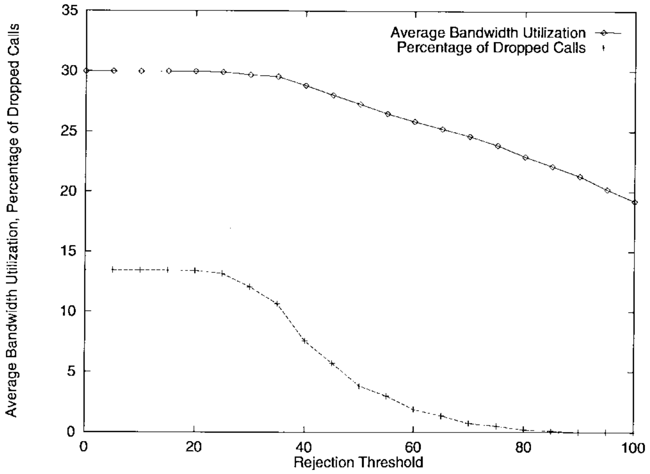


Fig. 4. Average bandwidth utilization and percentage of dropped calls, shadow cluster case.

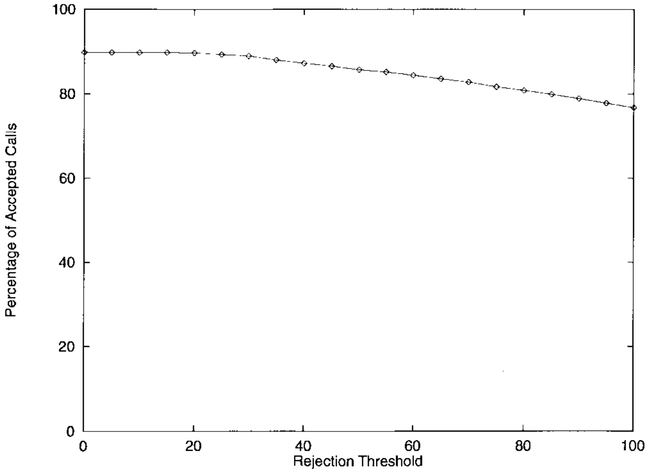


Fig. 5. Percentage of accepted calls, shadow cluster case.

In Fig. 5, we present simulation results for the percentage of calls that are accepted as a function of the rejection threshold. Note that an acceptance percentage equal to 100% is not possible, even when the rejection threshold is equal to zero. This is because regardless of the value of the rejection threshold, it is not possible to accept a new call if there are not enough free BU's for the call [as stipulated in Step 1) of the call admission algorithm]. As expected, in Fig. 5 we can observe that the percentage of accepted calls decreases monotonically as the rejection threshold increases. Note that for a rejection threshold equal to 100 (i.e., when no calls are dropped), the percentage of calls that are accepted is still very high, almost 80%.

In order to compare the performance of the shadow cluster to an alternative scheme, we present in Figs. 6 and 7 the performance results for the same network but when the admission of new mobile terminals is done in a random fashion. In this experiment, mobile terminals were admitted to the network when a random number in the interval (0,1) generated for each mobile terminal was above a *rejection value*. In other words, for a rejection value equal to 0, all mobile terminals were admitted to the network (provided there were enough

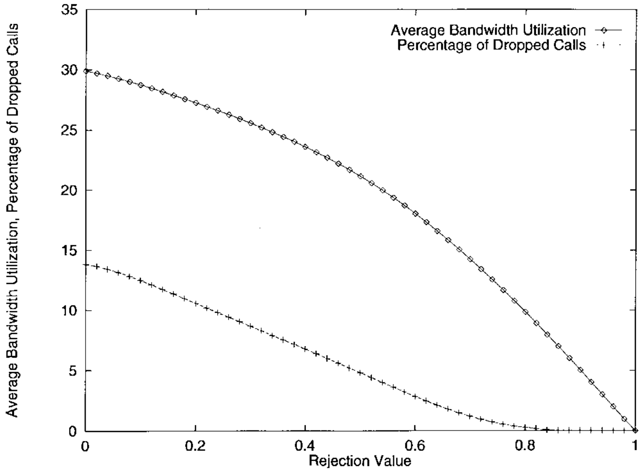


Fig. 6. Average bandwidth utilization and percentage of dropped calls, random admission case.

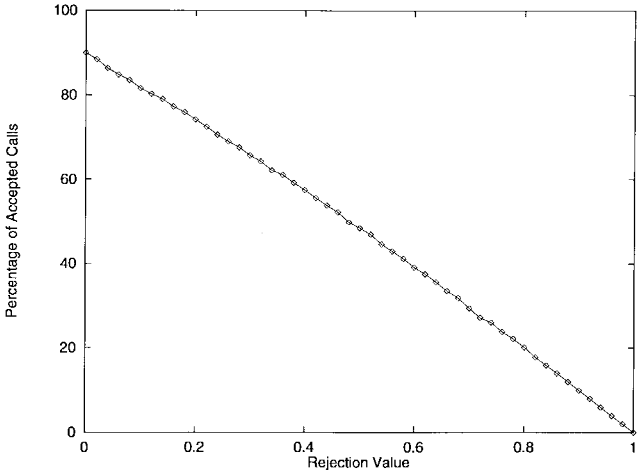


Fig. 7. Percentage of accepted calls, random admission case.

free BU's in the cell to support the new mobile terminal). When the rejection value was z, base stations tried to admit $100 \cdot (1 - z)\%$ of all new call requests to the network.

In Fig. 6, we depict the average bandwidth utilization and the percentage of dropped calls for a network that is using random admission. If we want to limit the number of dropped calls to 1%, then the rejection value must be greater than 0.7, which implies that the average bandwidth utilization falls to approximately 13 BU's. Likewise, if we want to have a negligible percentage of dropped calls, the rejection value must be greater than 0.85, with a corresponding average bandwidth utilization of only about 6 BUs! In comparison, using the shadow cluster algorithms, a call dropping rate of 1% still allows a very high average bandwidth utilization of about 25 BU's, and an average bandwidth utilization greater than 20 BU's can be maintained when reducing the call dropping rate to negligible values.

In Fig. 7, we present the percentage of accepted calls when using random admission. Similar to the shadow cluster case, the maximum percentage of call admissions is approximately 90%. As expected, a rejection value equal to unity results in that no mobile terminals are admitted to the network. Comparing Fig. 7 to Fig. 5 (percentage of accepted calls for the shadow cluster case), we observe that a negligible call dropping rate for the random admission case requires that almost no new mobile terminals are accepted to the network, while for the shadow cluster case, similar call dropping rates are achieved while admitting more than 75% of all call admission requests.

VII. Conclusion

The shadow cluster concept is likely to be very useful in any wireless network with small cells (e.g., nano, micro, mini), irregular and time-varying traffic loads, e.g., time varying "hot spots," and with a high number of cell handoffs per call.

We have provided algorithms that can be used to implement shadow clusters. These algorithms presume knowledge about the probability that a mobile terminal will be active in a given cell and at a particular time. The accuracy in the determination of these probabilities for a particular mobile terminal depends on the amount of knowledge available on the behavioral patterns of the mobile terminal user under study, as well as on the characteristics of its physical surroundings where the mobile terminal travels [6]. The effectiveness of the shadow cluster concept is closely tied to the accuracy in the determination of active mobile probabilities, as well as on the variances in the pdf's presented in Section IV-A. In fact, the smaller the variance on the number of free BU's $\widetilde{C_{f_j}(t)}$ presented in (18), the better the shadow cluster will perform. Thus, a fair quantitative evaluation of the shadow cluster concept would require also an evaluation of the accuracy in the determination, as well as on the probability characteristics, of active mobile probabilities. Besides their applicability in shadow cluster algorithms, we claim that the determination of active mobile probabilities will prove very useful in location update procedures: the number of paging messages and paging delays could greatly be reduced.

The shadow cluster algorithms for resource prediction and call admission are scalable. Thus, the amount of required processing and communication costs can be adjusted as a function of the shadow clusters' size and the length of τ, the time interval. As real data on the traffic characteristics of individual mobile terminals become available, the above variables can be evaluated to determine optimal values that result in a shadow cluster system that is practical and manageable.

We showed the applicability and usefulness of the shadow cluster concept with simulation experiments. With the shadow cluster mechanism, it is simple to control the percentage of calls that will need to be dropped. Because the shadow cluster mechanism selects for admission only those mobile terminals that are likely to complete their calls, the percentage of dropped calls can be reduced in a controlled fashion, while still maintaining high cell throughputs.

Acknowledgment

The authors express their sincere thanks to C. Perkins for his constructive and very detailed comments which shaped up the paper into this format. His extremely keen interest in the development of this paper is greatly appreciated.

References

[1] A. S. Acampora and M. Naghshineh, "An architecture and methodology for mobile-executed hand-off in cellular ATM networks," *IEEE J. Select. Areas Commun.,* vol. 12, pp. 1365–1374, Oct. 1994.

[2] G. Falciasecca, M. Frullone, G. Riva, M. Sentinelli, and A. M. Serra, "Investigation on a dynamic channel allocation for high capacity mobile radio systems," in *38th IEEE Veh. Technol. Conf.,* 1988, pp. 176–181.

[3] S. Katsuki and E. Hato, "A study of drivers' behavior and traffic management," in *Proc. 1994 Veh. Navigation & Inform. Syst. Conf.,* Yokohama, Japan, 1994, pp. 255–258.

[4] R. König, A. Saffran, and H. Breckle, "Modeling of drivers' behavior," in *Proc. 1994 Vehnol. Navigation & Inform. Syst. Conf.,* Yokohama, Japan, 1994, pp. 271–276.

[5] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "The shadow cluster concept for resource allocation and call admission in ATM-based wireless networks," in *Proc. ACM Int. Conf. Mobile Comp. Networking MOBICOM'95,* Berkeley, CA, Nov. 1995, pp. 142–150.

[6] G. Liu, A. Marlevi, and G. Q. Maguire, "A mobile virtual-distributed system architecture for supporting wireless mobile computing and communications," *Wireless Networks,* vol. 2, no. 1, pp. 77–86, Jan. 1996.

[7] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," in *Proc. PIMRC'95,* Toronto, Canada, Sept. 1995.

[8] S. Nanda and D. J. Goodman, "Dynamic resource acquisition: Distributed carrier allocation for TDMA cellular systems," in *Third Generation Wireless Inform. Networks.* Norwell, MA: 1992, pp. 99–124.

[9] A. Niehaus and R. Stengel, "Probability-based decision making for automated highway driving," *IEEE Trans. Veh. Technol.,* vol. 43, pp. 626–634, Aug. 1994.

[10] H. Panzer and R. Beck, "Adaptive resource allocation in metropolitan area cellular mobile radio systems," in *40th IEEE Veh. Technol. Conf.,* 1990, pp. 156–160.

[11] D. Raychaudhuri and N. D. Wilson, "ATM based transport architecture for multiservices wireless personal communication networks," *IEEE J. Select. Areas Commun.,* vol. 12, pp. 1401–1414, Oct. 1994.

[12] J. Tajima and K. Imamura, "A strategy for flexible channel assignment in mobile communication systems," *IEEE Trans. Veh. Technol.,* vol. 37, pp. 92–103, May 1988.

[13] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *IEEE Commun. Mag.,* vol. 29, pp. 42–46, Nov. 1991.

[14] W.-B. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE J. Select. Areas Commun.,* vol. 12, pp. 654–664, May 1994.

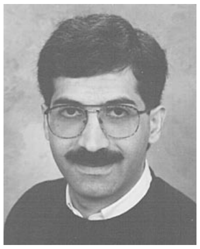
[15] M. Zhang and T. P. Yum, "Comparisons of channel-assignment strategies in cellular mobile telephone systems," *IEEE Trans. Veh. Technol.,* vol. 38, pp. 211–215, Nov. 1989.

David A. Levine received the B.S. degree in electronics and communications from the Universidad La Salle, Mexico City, in 1988, the M.S.E.E. degree from the University of Virginia, Charlottesville, in 1991, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 1996.

While at Georgia Tech, he was a recipient of a scholarship from the National Science Foundation. In the Summer of 1995, he worked at AT&T Bell Laboratories, Whippany, NJ, as a summer intern. He joined BellSouth Telecommunications, Atlanta, GA, in 1996. His current research interests are in wireless networks, optical LAN's, performance evaluation, and computer telephony.

Ian F. Akyildiz (M'86–SM'89–F'96) is a Professor of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA.

Dr. Akyildiz is an ACM Fellow.



Mahmoud Naghshineh (S'87–M'91) received the Vordiplom degree in electrical engineering from RWTH Aachen, Germany, in 1985, the B.S. degree in computer engineering and the M.S. degree in electrical engineering from the Polytechnic University, New York, in 1991 and 1988, respectively, and the Ph.D. degee from Columbia University, New York, in 1994.

He is a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he currently works in the wireless and mobile networks group. He joined IBM in 1988. From 1988 to 1991, he worked on a variety of research and development projects dealing with design and analysis of local-area networks, communication protocols, and fast packet-switched/broadband networks. Since 1991, he has been working in the area of wireless and mobile ATM, wireless access broadband networks, and mobile and wireless local-area networks. He is an Editor of *IEEE Personal Communications Magazine*. He has published numerous technical papers and holds a number of IBM patents in the area of high-speed and wireless/mobile networks.

Dr. Naghshineh is a member of the IEEE Technical Committee on Computer Communications as well as the Technical Committee on Personal Communications. He has served as a member of Technical Program Committee, Session Organizer and Chairperson for many IEEE Conferences and Workshops.