The Predictive User Mobility Profile Framework for Wireless Multimedia Networks

Ian F. Akyildiz, *Fellow, IEEE,* and Wenye Wang, *Member, IEEE*

***Abstract*—User mobility profile (UMP) is a combination of historic records and predictive patterns of mobile terminals, which serve as fundamental information for mobility management and enhancement of quality of service (QoS) in wireless multimedia networks. In this paper, a UMP framework is developed for estimating service patterns and tracking mobile users, including descriptions of location, mobility, and service requirements. For each mobile user, the service requirement is estimated using a mean-square error method. Moreover, a new mobility model is designed to characterize not only stochastic behaviors, but historical records and predictive future locations of mobile users as well. Therefore, our approach incorporates aggregate history and current system parameters to acquire UMP. In particular, an adaptive algorithm is designed to predict the future positions of mobile terminals in terms of location probabilities based on moving directions and residence time in a cell. Simulation results are shown to indicate that the proposed schemes are effective on mobility and resource management by evaluating blocking/dropping probabilities and location tracking costs in wireless networks.**

***Index Terms*—Mobility and resource management, quality of service (QoS), user mobility profile (UMP), wireless network.**

I. Introduction

DIVERSE mobile services and development in wireless networks have stimulated an enormous number of people to employ mobile devices such as cellular phones and portable laptops as their communications means. The most salient feature of wireless networks is mobility support, which enables mobile users to communicate with others regardless of location. It is also the very source of many challenging issues, relating to the mobility and service patterns of mobile terminals (MTs), namely, user mobility profile (UMP). For each mobile user, a UMP consists of detailed information of service requirements and mobility models that is essential to quality of service (QoS) and roaming support. The applications of UMP can be categorized as follows:

- *Development and analysis of handoff algorithms*. One of the most important QoS issues is to design efficient handoff algorithms to reduce handoff-dropping probability caused by bandwidth shortage and mobility when mobile users move from one cell to another [8], [23].
- *Call admission control (CAC) and resource management*. An efficient CAC algorithm demands the knowledge of UMP in order to accommodate the maximum number of users or to yield maximum system throughput [11], [12].
- *Routing optimization*. User mobility information can also be used to assist traffic routing in wireless networks to ease the bottleneck effect in overloaded base stations or access points [13].
- *Location update and paging*. Many mobility management schemes utilize UMP to improve system performance with regards to reducing signaling costs and call loss rates [6], [21].

Since mobility and resource management are critical to supporting mobility and providing QoS in wireless networks, it is very important to describe movement patterns of mobile objects accurately. In [1], the location probability at the time of a call arrival is calculated by assuming that the MTs take the shortest paths when they move from one cell to another with four possible directions. Another prediction method is proposed in [5] in which the next probable cell is determined based on path information. In [13], a hierarchical location-prediction algorithm is described in which a two-level user mobility model is used to represent the movement behavior at global and local levels. The next cell is predicted by considering speed and direction of a user's trajectory. Through estimation of mobile users' trajectory and arrival/departure times in [4], a group of future cells are determined, which constitute the most likely cluster into which a terminal will move. In order to predict resource demands, two methods are proposed in [23]. One of them is based on the Wiener Process which predicts the bandwidth requirement according to current bandwidth usage. The other method is to use time series analysis on the premise that future demand increments are related to past variations.

To summarize, most of the existing methods are aimed at finding the most probable cell [4], [5], [13]. However, when an MT moves quickly in micro-cell networks, the short residence time in a cell may not allow computations in every cell, i.e., next-cell prediction. Also, there are very limited efforts on estimating a group of probable cells or a cluster of cells without considering the historical records [1], [4], [6]. Often, the demand for multimedia services is not taken into account, which is critical for efficient resource management. Furthermore, UMP is not well defined, which should consider the characteristics of users' mobility and service patterns.

Manuscript received January 8, 2002; revised January 21, 2003, and November 10, 2003; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Grossglauser. This work was supported by the National Science Foundation under Grant ANI-0117840.

I. F. Akyildiz is with the Broadband and Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: ian@ece.gatech.edu).

W. Wang is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: wwang@eos.ncsu.edu).

Digital Object Identifier 10.1109/TNET.2004.838604

Therefore, a framework for UMP is proposed in this paper, which includes the following contributions.

- A *zone* concept is proposed to add an additional level of location description to differentiate varying future locations of a mobile user depending on its moving direction, thus reducing computation overhead.
- A new framework of user mobility profile is designed to incorporate service requirements and the mobility model, including long-term and short-term information.
- By using an order-L Markov predictor, the service requirements of a mobile user are predicted based on the most recent k records, aiming to minimize the mean-square error.
- An adaptive prediction algorithm is developed to predict a group of cells into which an MT will move by considering historical records, path information, moving direction and speed, cell residence time, and tradeoff of computation overhead.
- The implementation and the effectiveness of the proposed scheme for mobility and resource management is discussed with respect to system performance and overhead.

The rest of this paper is organized as follows. In Section II, a system model and a new concept of *zone* is for location description. In addition, a new mobility model, including stochastic model, historical records, and predictive trajectory, is described. In Section III, the framework of UMP is defined, and it is categorized into quasi-stationary and dynamic UMP. The estimation and prediction algorithms of service requirements and future location probabilities are presented in Section IV. In Section V, we describe the simulation model and the parameters used in our experiments. The effectiveness, applications, and overhead of the proposed schemes in mobility and resource management are discussed in Section VI, followed by conclusions in Section VII.

II. System Model, Location Description, and Mobility Model

In this section, we describe a system model based on cellular networks, and we present a new concept, *zone partition*, and a mobility model for a more precise location description.

A. System Model

Consider a mobile wireless network with a cellular infrastructure, e.g., General Packet Radio System (GPRS), which may be one of several macro-, micro-, and pico-cell systems. This wireless mobile network provides diverse service applications such as voice, audio, data, and video. A typical network is composed of a wired backbone and a number of base stations (BSs). Each BS is in control of a cell, and a group of BSs are managed by a mobile switching center (MSC) in the circuit-switch domain and a serving GPRS supporting node (SGSN) in the packet-switch domain. The service area is divided into location areas (LAs), and each LA consists of a group of cells. If an MT is moving from one cell to another cell belonging to another MSC, location registration and identity authorization need to be carried out.

Since each LA consists of a number of cells, we propose a *zone* partition concept to improve the granularity of location description. A zone is a subset of an LA, which is composed of a group of adjacent cells. We expect that the MTs in the same zone demonstrate the same, if not similar, movement behavior. For example, in Fig. 1, the coverage of an MSC-A is an LA, and there are two MTs, α and β, which are currently located in the coverage area of MSC-A and may possibly move into the area of MSC-B and MSC-C, respectively. The service area of each MSC is divided into n zones $(n = 7)$. If we know that MT β is moving south, it is most likely that it is going to move into zone 3. Therefore, we incorporate an MT's movement direction and position (zone) to provide a more accurate prediction of the MT's mobility pattern than simply using the LA information.

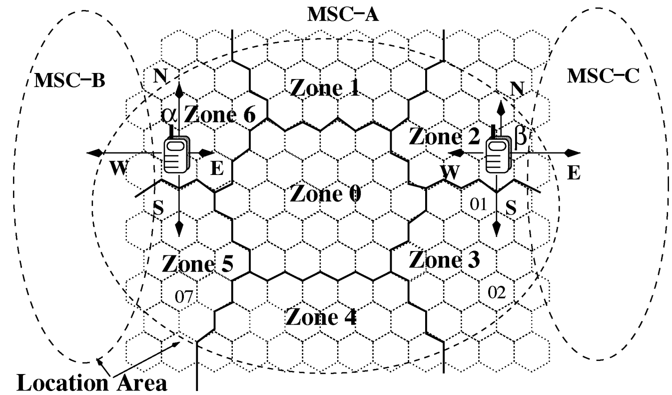


Fig. 1. Zone partition.

B. Location Description

Based on our description in Section II-A, locations can be specified at three levels. In other words, a mobile user's current position can be represented as follows.

- *Location Area*: As shown in Fig. 1, the current location area ID is available in home location register (HLR) and visitor location register (VLR) because an HLR is a centralized database that maintains permanent information of mobile users, whereas VLR keeps the up-to-date locations of visiting mobile terminals.
- *Zone Partition*: The granularity of LA is not sufficient to predict future locations. Thus, *zone* partition is used to describe more accurate information of an MT's current position and to estimate next position due to the continuity of movement.
- *Cell ID*: In order to maintain an active connection for a mobile user, it is most important to know in which cell the mobile user is located. The network knows in which cell an MT resides by sending polling messages, which can be acquired without additional cost during call origination/termination, or through location client service (LCS) management. The details of implementation are introduced in Section VI-C.

The ultimate goal of this work is to estimate the next cells into which a mobile user will possibly move. The prediction of future LAs is beyond the scope of this paper, and more information can be found in [2]. In this study, we focus on how to predict possible future cells. It is worth mentioning that most information about mobile users is subscriptive and aggregate information collected and stored in current wireless systems for location management, such as *destination address* and *stored location area ID* in 3GPP specifications [19].

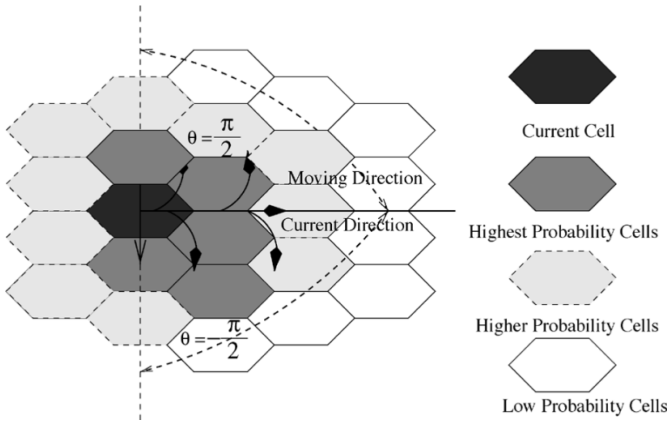


Fig. 2. Moving direction.

C. Mobility Model

There are many mobility models that describe stochastic characteristics of random movement with regard to mobility scale, varying randomness in direction, speed, and residence time, and geographical circumstances [7], [20]. In this paper, we propose a mobility model that considers stochastic behavior of mobile users, e.g., residence time within a cell, historical records, and predictive location patterns in terms of location probabilities. There are five components in this model to characterize the movement of a mobile user.

- The MT's residence time in a cell is represented by a random variable T, which has a Gamma distribution with probability density function (pdf) [2], [14], [15]. Gamma distribution is selected for its flexibility because, given different parameters, a Gamma distribution can be an exponential, an Erlang or a chi-square distribution. The pdf of an MT's residence time with Gamma distribution has Laplace transform $Q_{x,T}(s)$ with the mean value $1/\mu$ and the variance V. Then

$$Q_{x,T}(s) = \left(\frac{\mu\gamma}{s+\mu\gamma}\right)^{\gamma}$$

where

$$\gamma = \frac{1}{V\mu^2}. \tag{1}$$

The mean residence time of this distribution (1) is $E_x[T] = 1/\mu$. We assume that the residence time of an MT in each cell is independent throughout this paper.

- The MT's current direction $\theta_x(t)$ is collected through real-time monitoring, which can be initiated by the serving BS. The MT's current direction is defined as the direction from its previous position to the current position. Here we denote $\theta_x(t)$ as the moving direction of the MT x, which is defined as the degree from the current direction clockwise or counterclockwise, i.e., $-\pi \leq \theta \leq \pi$. Fig. 2 shows the probable cells in the shadow cluster for $\theta = [-\pi/2, \pi/2]$ when the current direction is from west to east.
- Moving speed: The MT is allowed to move away from its current position in any direction, and variation of the MT's direction based on its previous direction is a uniform distribution in the range of $\pm\pi/2$ [9], [22]. The initial velocity of an MT is assumed to be a random variable with Gaussian pdf truncated in the range of $[0, V_{\max}$ km/h], and the velocity increment is taken to be a uniformly distributed random variable in the range of $\Delta v\%$ of the average velocity $\overline{V}$/h.
- Historical records: In this context, we trace the historical records of an MT by using a trace records matrix (TRM) of $L \times M$, where L is the total number of records and M is the total number of cells that an MT has traversed in the period of observation. The element $g_{\alpha\beta}$, $\alpha = 1, 2, \ldots, L$, $\beta = 1, 2, \ldots, M$, of the TRM denotes whether the MT has traversed a cell. The matrix $\mathbf{G}$ can be written as

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1M} \\ \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & g_{\alpha\beta} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ g_{L1} & g_{L2} & \cdots & g_{LM} \end{pmatrix} \tag{2}$$

and $g_{\alpha\beta}$ is given by

$$g_{\alpha\beta} = \begin{cases} 1, & \text{if the MT has entered this cell} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

- Predictive future locations: The probabilities that an MT will be in other cells at time t is denoted as $\vec{\mathbf{P}}_{x,i}(t)$, given that the current serving BS of the MT x is i. The serving BS of an MT is expected to determine *location probabilities* and notify other BSs with this estimation. Then

$$\vec{\mathbf{P}}_{x,i}(t) = [P_{x,i,0}(t)\ P_{x,i,1}(t)\ P_{x,i,2}(t)\ \cdots\ P_{x,i,N}(t)] \tag{4}$$

where N is the total number of cells for the estimation. This number is closely related to the scope of location prediction. In reality, the prediction is valid for a specific time period. Given the moving speed of an object, the maximum distance that an object can travel within a time period can be determined. Thus, the maximum number of cells that cover the maximum distance can also be decided accordingly.

III. User Mobility Profile Framework

In order for wireless networks to support different bandwidth reservation and mobility management strategies, the system must be able to dynamically and adaptively maintain historical records and predictive locations of mobile users online. The historical records can be collected from the user's subscription such as the user's information stored in and the HLR for cellular systems or during the origination or termination of a session for the purpose of quality monitoring and billing. On the contrary, future positions must be predicted based on historical records and mobility-related parameters such as current position, direction, and speed [9], [22]. In the proposed framework, the predictive records are referred to as the location probabilities of the cells in the shadow cluster and the estimated service type during an MT's roaming into other cells. There are two types of data in the new UMP. The first type is called *quasi-stationary UMP*, which represents the MT's information that changes infrequently or that can be obtained from network databases. This includes both subscriptive and historical information. The other type is called *dynamic UMP*, which changes over time or cannot be obtained from network databases.

A. Quasi-Stationary UMP

An MT's *quasi-stationary* information includes an MT's current network characteristics, calling pattern, and service requirements. In wireless networks, a mobile user receives its mobile services by subscribing to a home network. For example, a user can subscribe to a service package, which is the combination of different service classes. Such information will remain unchanged for a long In other words, this information may remain unchanged for a certain period of time. Therefore, such information can be referred to as *quasi-stationary UMP*. We denote this long-term information as a quadruple, $\mathbf{\Lambda}_{\mathbf{x}}^{\mathbf{Q}} = (\Gamma_x, \Omega_x, \Psi_x, \Phi_x)$ for an MT x. Here, we consider that each element of $\mathbf{\Lambda}_{\mathbf{x}}^{\mathbf{Q}}$ is a component of a vector, which indicates a specific character. Thus, Γ_x, Ω_x, Ψ_x, and Φ_x are the elements of vectors $\mathbf{\Gamma}$, $\mathbf{\Omega}$, $\mathbf{\Psi}$, and $\mathbf{\Phi}$, respectively, which are elaborated as follows.

- *Network Configuration* $\mathbf{\Gamma} = \{\Gamma_1, \Gamma_2, \cdots, \Gamma_W\}$. This is a vector that identifies different network architectures, such as pico-cell, micro-cell, macro-cell, and satellite systems. W is the number of the registered networks among which an MT is allowed for seamless roaming. At a specific time instant, an MT must be located in a network, e.g., Γ_x. This information is not required from users; instead, it is recorded in the network entities such as an HLR and VLRs.
- *Time Period* $\mathbf{\Omega} = \{\Omega_1, \Omega_2, \cdots, \Omega_S\}$. This set identifies different time periods of one day. S is the cardinality of the set, i.e., the maximum number of time segments divided within 24 h. In current cellular networks, time segments relating to accounting are categorized into peak time, non-peak time, weekends, and weekdays and are decided by service providers.
- *Service Description* $\mathbf{\Psi} = \{\Psi_1, \Psi_2, \cdots, \Psi_U\}$. This set represents service requirements such as bandwidth requirement and end-to-end delay. U is the total number of service patterns allowed in the network. Each element in this set provides the probability mass function (PMF) over a particular sampling space of service types as described in Section III-B. For example, $\Psi_x \in \mathbf{\Psi}$ is used to represent the PMF of an MT x and its corresponding sample space $\mathbf{A}_x$.
- *Calling Pattern* $\mathbf{\Phi} = \{\Phi_1, \Phi_2, \cdots, \Phi_V\}$, where V is the cardinality of the set. Each element is related to the description of calling events, which can be different from one user to another. However, in most of the existing networks, the calling patterns are defined based on a group of mobile users so that computation complexity can be simplified and scalable. A calling pattern may include the probability distribution of calling arrival time, call holding time, and call origination rate.

B. Dynamic UMP

The quasi-stationary UMP provides long-term information of mobile users. In order to satisfy real-time prediction requirements, we also need to consider data that changes from time to time such as moving direction, speed, and service type. This type of information is referred to as *dynamic UMP*. Our objective is to derive or estimate dynamic UMP information based on quasi-stationary UMP. We focus on the predictive records that are derived from the mobile user's previous service and mobility information. We assume that locations are independent of services in this context because they can be reflected in the network configuration. In particular, we examine location probabilities and service type because these two metrics are the most often used parameters for mobility and resource management.

Let $\mathbf{\Lambda}_{\mathbf{x}}^{\mathbf{D}} = [\alpha_x^o, \vec{\mathbf{P}}_{x,i}(t)]$ denote the dynamic UMP for the MT x at time t. The first element α_x^o is the service type that x will request, and the second item $\vec{\mathbf{P}}_{x,i}(t)$ represents the estimation of location probabilities given that the MT x is currently in cell i. Location probabilities are used to represent the likelihood of an MT's presence in a cell. For single-cell location estimation, there are only two possibilities: 1 or 0, in accordance with whether an MT will be in a cell or not, respectively. Since it is unrealistic to represent an MT's movement by a single random variable, we use location probabilities to describe the result of many factors, such as mobility models, changing directions, and geographic conditions. This method is widely used in the research of handoff, location tracking, and mobility management [8], [18], [21].

In this context, we use PMF to describe different service requirements, which are offered to subscribers in terms of service packages, covering the most popular service requirements of mobile users. In other words, each service pattern has a specific PMF with regard to each service type. For a discrete random variable, service type Y, the PMF $\psi(a)$ of Y over a sample space $\mathbf{A}$ is defined by

$$\begin{aligned} \psi(a) =& Prob\{Y = a\} \\ \sum_{a \in \mathbf{A}} \psi(a) =& 1. \end{aligned} \tag{5}$$

For a particular MT x, the service pattern is denoted by Ψ_x. The PMF can be decided either by the network administrators or by accumulating the mobile user's records. The algorithm for estimating α_x^o will be illustrated in Section IV-B. One of the main tasks of predicting and computing UMP is to determine $\vec{\mathbf{P}}_{x,i}(t)$. The sequence of cells that will be visited by the MT constitutes a random process, with location probabilities $P_{x,i,j}(t)$, which is the probability that an MT x, residing in cell i, will be in cell j at time t. The location probabilities $\vec{\mathbf{P}}_{x,i}(t)$ are calculated by the serving BS based on the MT's historical records, current position, velocity, and moving direction.

IV. Measurement and Prediction Algorithms

A block diagram of the algorithms proposed for service and mobility prediction is illustrated in Fig. 3. We start with the account of the UMP framework by explaining the parameters used for prediction. Then, we discuss the description and prediction of service requirements. Afterwards, we present the prediction algorithm of location probabilities.

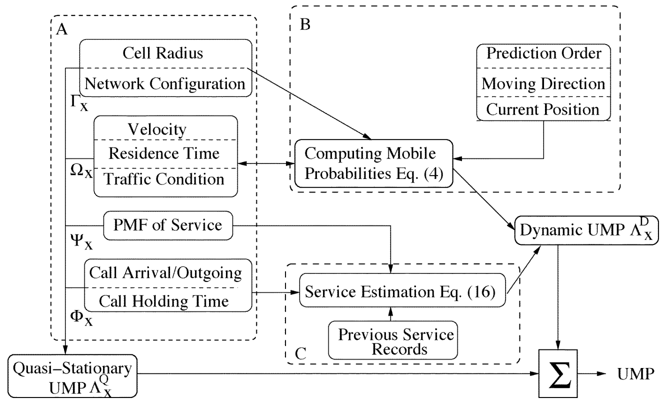


Fig. 3. Flowchart of computing a UMP.

A. *Flowchart of the Algorithm*

First, given the network infrastructure by Γ_x, the cell radius and network configuration are then available. Second, the time period $\Omega_x \in \mathbf{\Lambda_x^Q}$ is one element in Ω. The direct information corresponding to a particular timing period is the moving velocity and the pdf of an MT's residence time in a cell, which depend upon traffic conditions. Then, the service description is given by $\Psi_x \in \mathbf{\Lambda_x^Q}$, with PMF $\psi_x(a)$ and its corresponding sample space $\mathbf{A}_x$. Finally, the calling pattern of the MT is described by $\Phi_x \in \mathbf{\Lambda_x^Q}$, relating to the incoming/outgoing call distribution and call holding time. This information will be used for resource allocation in possible future cells.

In addition to predicting location probabilities, we estimate the next service request based on the PMF of service patterns as well as historical records shown in box C of Fig. 3. The history of service requirements is known to the network because of billing and service management. Thus, this information does not require extra effort and can be utilized efficiently.

B. *Service Description, Measurement, and Estimation*

The aim of our algorithm is to minimize the estimation error between the estimated service type and the real value. The future service pattern is estimated based on an *a priori* PMF of service patterns, $\psi_x(a)$. First, we assume that the service pattern of an MT can be represented by a random variable Y. The PMF of this random variable is $\psi_x(a)$, which is the probability for value $Y = a$. The sample space of service types is denoted by A_x. For example, there are four samples in A_x, i.e., there are four different services which may be requested by an MT, such as voice, data, audio, and video, corresponding to values 1, 2, 3, and 4, respectively. We define a PMF vector $\vec{\mathbf{F}}(\Psi_x)$ as

$$\mathbf{F}(\Psi_x) = [\psi_x(1),\ \psi_x(2),\ \cdots,\ \psi_x(K)] \tag{6}$$

where $\psi_x(\cdot)$ is the probability of each service type for $a \in \mathbf{A}_x$ and $K = |\mathbf{A}_x|$ relating to Ψ_x.

Then, we consider an order-L Markov predictor, which assumes that the service can be predicted from the most recent values in the service history, which is

$$\vec{\mathbf{h}}_x(L) = [h_x(t_1)\ h_x(t_2)\ \cdots\ h_x(t_L)] \tag{7}$$

where L is the length of historical records and $h_x(t_k)$ is the service that the MT x has requested at time t_k.

If we think of the user's service requirement as a random variable Y, and $Y(i,j)$ is a string $Y_iY_{i+1}\ldots Y_j$ representing the sequence of values that Y takes for any $1 \le i \le j \le L$, then the Markov assumption is that Y behaves as follows, for all $a \in \mathbf{A}$:

$$\begin{aligned}\text{Prob}\left(Y_{n+1} = a|Y(1,n) = \vec{\mathbf{h}}_x(L)\right) &= \text{Prob}\left(Y_{n+1} = a|Y(n-L+1,n) = c\right)\\ &= \text{Prob}\left(Y_{i+L+1} = a|Y(i+1,i+k) = c\right)\end{aligned} \tag{8}$$

where the notation $\text{Prob}(Y_{n+1} = a_i|\ldots)$ denotes the probability that Y_i takes the value a_i. The first line indicates that the probability depends on only the most recent L records, while the latter two lines indicate the assumption of a stationary distribution. We denote the probability that the next service type takes the value a as $\hat{\psi}(a)$ as follows:

$$\hat{\psi}\left(a|\vec{\mathbf{h}}_x(L)\right) = Prob\left(Y_{L+1} = a|\vec{\mathbf{h}}_x(L)\right). \tag{9}$$

We also notice that history records $\vec{\mathbf{h}}_x(L)$ and the new estimate a will result in an *estimate PMF*

$$\hat{\psi}\left(y_i|Y_{L+1} = a, \vec{\mathbf{h}}_x(L)\right) = \begin{cases}\frac{N(y_i,L+1)+1}{L+1}, & \text{if } y_i = a\\ \frac{N(y_i,L+1)}{L+1}, & \text{if } y_i \neq a\end{cases} \tag{10}$$

where y_i, for $i = 1, 2 \cdots, K$, is one of the values in the sample space $\mathbf{A_x}$, and $N(c,d)$ denotes the number of times that value c occurs in the number d historic records.

As a result, we can generate an estimate PMF vector, which shows the frequency of each service request for $a \in \mathbf{A}_x$. We denote this PMF as $\hat{\mathbf{F}}(\Psi_x)$ as follows:

$$\begin{aligned}\hat{\mathbf{F}}_x(\Psi_x) = \Big[&\hat{\psi}\left(y_1 = 1|Y_{L+1} = a, \vec{\mathbf{h}}_x(L)\right),\\ &\hat{\psi}\left(y_i = 2|Y_{L+1} = a, \vec{\mathbf{h}}_x(L)\right)\cdots\\ &\hat{\psi}\left(y_K = K|Y_{L+1} = a, \vec{\mathbf{h}}_x(L)\right)\Big].\end{aligned} \tag{11}$$

We predict the next service requirement by minimizing the mean-square error between the prior PMF and the estimate PMF. Consider that the mean-square estimation of the random variable Y is to find an optimal constant a that minimizes e

$$e = E\left\{(Y-a)^2\right\} = \int_{-\infty}^{\infty} (Y-a)^2 f(y)dy \tag{12}$$

where $f(y)$ is the pdf of a continuous random variable Y. For the discrete random variable Y, which follows the PMF, Ψ_x, the above equation is rewritten as

$$e = \sum_{y_i \in \mathbf{A_x}} (y_i - a)^2 \psi_x(y_i) \tag{13}$$

where a is one of the values of $y_1, y_2, \cdots, y_i, \cdots, y_K$.

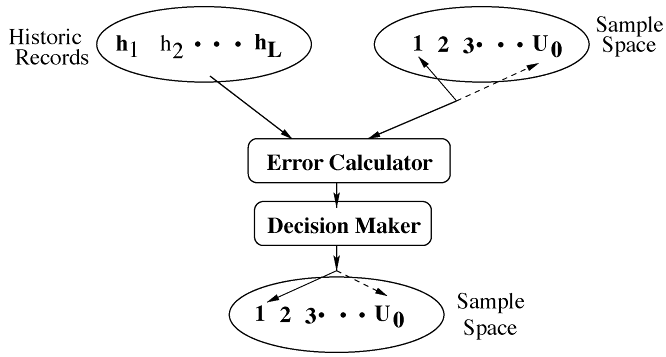


Fig. 4. Estimation of service type.

However, this definition cannot indicate the relationship between the history and the selection of an optimal value, showing that an optimal value is selected independently of the previous records of the random variable. When we consider historical records in our estimation, it is necessary to take into account the historical effect on finding an optimal value. In this context, this effect is represented by the estimate PMF for an estimate value. Accordingly, the objective of the estimation is to find an optimal value α_x^o that can minimize the difference between the prior PMF vector $\vec{\mathbf{F}}(\Psi_x)$ in (6) and the empirical PMF, $\hat{\mathbf{F}}_x(\Psi(b|\vec{\mathbf{h}}_x(L))$, in (11). Therefore, we define mean-square error between these two vectors $\varepsilon_x(a|\vec{\mathbf{h}}_\mathbf{x}(\mathbf{L}))$ as

$$\varepsilon_x\left(a|\vec{\mathbf{h}}_\mathbf{x}(\mathbf{L})\right)=\frac{1}{K}\sum_{i=1}^{K}\left|\psi_x(y_i)-\hat{\psi}\left(y_i|Y_{L+1}=a,\vec{\mathbf{h}}_x(L)\right)\right|^2 \tag{14}$$

where K is the cardinality of $\mathbf{A_x}$. By applying the minimum square error for the estimation, the estimated result of the next service type is

$$\alpha_x^o\vec{\mathbf{h}}_x(L)=\underbrace{\arg\mathrm{Min}}_{\mathbf{A}_x}\left\{\varepsilon_x\left(a|\vec{\mathbf{h}}_\mathbf{x}(\mathbf{L})\right)\right\}. \tag{15}$$

Therefore, the estimation is determined by both the service PMF and historical records.

During the implementation, we consider that an MT's historical records will be updated in the VLR/SGSN and HLR during the connection origination stage. Then, we can proceed to estimate the future service type α_x^o by examining each service type $a \in \mathbf{A}_x$. Each service type in the sample space $\mathbf{A}_x$ is selected sequentially and input to the square error calculator together with the historical records, as shown in Fig. 4. The results of this calculation are then input into a decision maker to obtain the most probable service type with minimum error. Finally, the corresponding service type α_x^o is chosen as the estimated service pattern. This method guarantees that the estimated service pattern fits well with the known PMF. If an MT is not in the progress of a call and its next probable service request is estimated as α_x^o, then the bandwidth that needs to be reserved can be determined by substituting α_x^o in (15) for a in (28).

C. Prediction Algorithm of Location Probabilities

In this study, we develop an algorithm which takes several important factors into account, while providing a suboptimal maximum-likelihood estimation.

1) Assumptions and Parameters: We assume that the MT's residence time in a cell is represented by a random variable T, which has a Gamma distribution with a pdf. First, we need to assume and measure the following parameters described in Section II-C.

- The pdf of the residence time is described by a Laplace transform $Q_{x,T}(s)$ in (1), with a mean of $(1/\mu)$ and variance of V.
- The MT's current position is represented at three levels, as described in Section II-B. The zone partition can be obtained by knowing the cell ID.
- The MT's current direction $\theta_x(t)$ and speed $v(x)$ are collected through real-time monitoring, which can be initiated by the serving BS, as explained in Section VI-C. The initial velocity of an MT is assumed to be a random variable with s Gaussian pdf truncated in the range of $[0, V_{max}$ km/h].
- Historical records are available in the form of a trace records matrix (TRM) of $L \times M$ as in (2).
- *Path database* (PD) is a part of the digital map database, which is obtained by discretizing a map into small segments, and each segment has many routes along with their relationship to others. This information is available for many applications, such as finding directions from one location to another.
- We also assume that an aggregate historical path database $\mathcal{D}_x^H$ is retrievable in the network administration center. Each record in this database is the previous path that the MT x has traversed.

Note that the path database $\mathcal{D}_x^H$ is different from trace matrix $\mathbf{G}$. The former shows the cells in the sequence of an MT's travels, while the latter may be organized in an order of cell IDs. The objective of a TRM is to determine future cells given the trajectory of an MT, whereas the path database objective is to find the similarity of an MT's movement.

2) Prediction Algorithm: There has been some success of varying degrees on determining location probabilities, which is mainly focused on determining the next cell based on the measurement of signal strength without considering road conditions and an MT's historical behavior [13], [17]. In this study, we distinguish those probable cells under the following constraints:

- The maximum distance: for a certain time period ΔT, the maximum distance of an MT's traveling is $\Delta T \times V_{\max}$, where $V_{\max}$ is the upper limit of moving speed. Therefore, given the cell radius, we can obtain the maximum number of cells that an MT can traverse, i.e., the maximum number of cells to be considered in the prediction.
- The future zones: by considering the maximum distance, we can determine a group of cells into which an MT can possibly move. However, this set can be further reduced when we consider moving direction and the MT's current position.

- The prediction level: similar to all of the prediction-based schemes, our algorithm inevitably involves overhead for computations and communications. The higher the prediction levels, the more complicated the computations, which result in more accurate predictions because more cells are considered. The details of prediction level are described in Section IV-C2b.
- The repetitive movement: we assume that an MT has a repetitive moving pattern. Therefore, if a route in a cell has been traversed by an MT before, which is shown in $\mathcal{D}_x^H$, then we consider this cell to have higher location probabilities, i.e., the MT is more likely to move into this cell.
- The path information: to be realistic, we take the path information into account. For a given part of the digital map, we assume that it is covered by a number of cells. Each cell has a set of routes. If a cell consists of a route that is continued from the current cell in which an MT resides, then this cell has a higher location probability. In this way, we combine the effect of historical records as well as path information.

The prediction of location probabilities involves many parameters, which cannot be simply represented by a mathematical predictor. By considering the above constraints, we will achieve a suboptimal estimation with maximum likelihood. In our prediction, first we will find the most probable cells, based on the MT's: 1) current position; 2) moving speed; 3) prediction level; and 4) direction. Then, we will examine each cell in accordance with the last two elements.

a) Estimation of zones: For the sake of simplicity in presentation, we consider a wireless network with hexagonal configuration for the remainder of this paper, although arbitrary shape configurations can also be covered by our solution. Here we denote the position of an MT by coordinates at a specific time in an example of seven zones. In Fig. 5, let $l_x(t) = Z_0$ represent the MT currently in zone 0 and let $\theta_x(t)$ represent the current moving direction of the MT, which is the direction variation derived from the previous direction. This coordinate system is defined with its origin at the current location of the MT, i.e., the MT is always in its origin, and its previous direction is the positive direction of the x axis. The y axis can be obtained by turning 90° counterclock-wise from the x axis. Thus, the MT's position can be represented by $X(\rho_x(t), \theta_x(t))$, as shown in Fig. 5. This coordinate system is dynamic in the sense that its origin and axes' orientation change over time.

Assume an MT is moving from point O toward A; thus, its next position may be in zone 1, Z_1. In general, the future zone, $Case\ k$ $(1 \le k \le n-1)$, can be determined as

$$Z_k = \begin{cases} \left\lceil \frac{\theta_x(t)}{\theta_0} \right\rceil, & \text{if } \theta_x(t) \ge 0 \\ n - 1 - \left\lceil \frac{|\theta_x(t)|}{\theta_0} \right\rceil, & \text{if } \theta_x(t) < 0 \end{cases} \tag{16}$$

where n is the total number of zones and θ_0 is the angle of each zone. By default, we always consider Z_0 to be included in the UMP. It is possible that more than one zone is involved in estimating location probabilities since it is difficult to differentiate

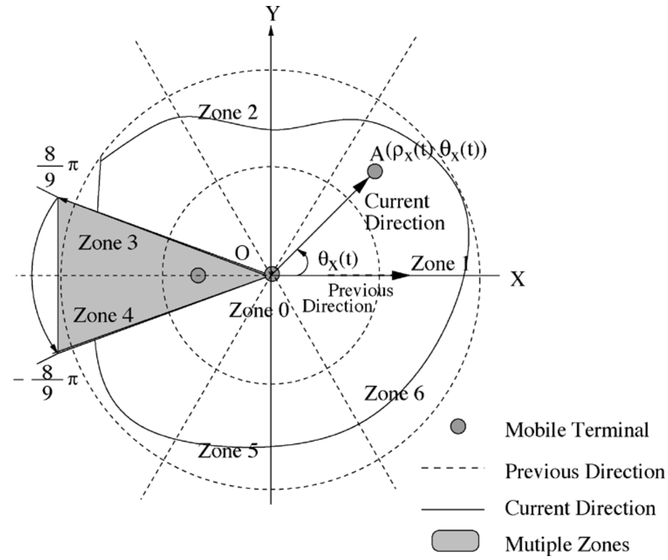


Fig. 5. Coordinate system with zone partition.

an MT's position around the boundary of zones. Thus, we extend possible zones for $Case\ k$ $(n \le k \le 2n-1)$ as follows:

$$\begin{aligned} &|\theta_x(t)| \pmod{\pi} \le \frac{1}{3}\cdot\theta_0 \\ &\quad \rightsquigarrow \text{Case } n \rightarrow Z_{n-1} \text{ and } Z_1 \\ &|\theta_x(t) - \theta_0| \pmod{\pi} \le \frac{1}{3}\cdot\theta_0 \\ &\quad \rightsquigarrow \text{Case } n+1 \rightarrow Z_1 \text{ and } Z_2 \\ &\qquad\vdots \\ &|\theta_x(t) - (n-2)\cdot\theta_0| \pmod{\pi} \le \frac{1}{3}\cdot\theta_0 \\ &\quad \rightsquigarrow \text{Case } 2n-1 \rightarrow Z_{n-2} \text{ and } Z_{n-1}. \end{aligned} \tag{17}$$

In the example shown in Fig. 5, $n = 7$ is the number of zones, and $\theta_0 = \pi/3$ is the angle of each zone. Although the selected zones can shrink the number of probable cells, it is not sufficient to identify the probable cells. Next, we will further approach those probable cells by considering velocity $v_x(t)$ and residence time distribution.

b) Prediction level: Basically, the more cells or larger areas considered in the prediction, the better the approximation that can be reached, requiring more computations. To balance the computation and estimation accuracy, we introduce a new concept called *prediction level* to describe the influenced region of the shadow cluster of an MT. According to the concept of shadow cluster, the influenced cells are a group of cells surrounding the cell in which the MT is residing. Thus, we always start from a *current cell*, which is the center of a shadow cluster. The vicinity of the current cell can be denoted according to its distance away from the center cell. If a cell is adjacent to the current cell, then it is in the *first layer* of the current cell. The cells adjacent to the first-layer cells form the *second layer* of the current cell. If the estimated cells cover only the cells of the first layer, then it is called *first-level prediction*. Similarly, the *second-level prediction* is associated with both first- and second-layer cells.

The definition of the prediction level is related to the accuracy of a prediction and can also be used as a QoS constraint, as well. For the first level of prediction, only six cells are considered, i.e., the computation is not very complicated. On the other hand, if the network should provide location probabilities with the second level of prediction, there are eighteen cells, including the first- and second-layer cells. Correspondingly, the complexity of computation increases, which is demonstrated in Section IV-C. Note that some second-layer cells may have higher location probabilities than some cells in the first layer, based on an MT's movement pattern. One example is that an MT traveling on a highway is more likely to be in the second-layer cells along its movement direction than the first-layer cells behind its trajectory or off the highway.

c) Calculation of the number of cells: Mobile users' velocity can be influenced by circumstances including geographical condition and timing period. For example, in an urban area in which there are many buildings, MTs are forced to travel at a low speed with diverse direction changes. On the contrary, in rural areas, MTs can travel at a higher speed and change direction infrequently. The maximum distance traveled by an MT can be determined by a knowledge of the upper and lower velocity bounds of an MT.

We compute the number of cells that an MT could have traveled during the time window ΔT. We consider that the MT goes through a cell with the mean residence time in a cell. The *average distance*, $\overline{d_x(t)}$, that an MT may travel along one direction, in terms of number of cells, is obtained by

$$\overline{d_x(t)} = \left\lceil \frac{\Delta T}{E_x[T]} \right\rceil . \tag{18}$$

Note that $\overline{d_x(t)}$ may be greater or less than the required prediction level $O_x(t)$ at time t for a terminal x. If the former is the case, the $O_x(t)$ will not affect the computation of probable cells, which means that the estimated region covers the area specified by $O_x(t)$. However, if the latter is the case, then the computed region cannot cover the area that is required by the prediction level. In this scenario, it is necessary to enlarge the calculation range to meet the requirement of $O_x(t)$. We denote $N_x(r : t = \max\{\overline{d_x(t)}, O_x(t)\}, k)$, which is rewritten as $N_x(r : t, k)$ in short, as the number of probable cells in the mobility profiles with order r at time t for $Case\ k$. The simplest scenario is $\overline{d_x(t)} = O_x(t) = 1$; the number of the probable cells for *Case* 1 in (16), $N_x(r = 1 : t, 1)$, can be calculated by the following formula as there are six cells in the first layer:

$$N_x(r = 1 : t, 1) = \left\lceil 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil . \tag{19}$$

The second layer includes two parts, one of which is the first-layer cells and the other is the second-layer cells which fall into the zone, that is

$$N_x(r=2 : t, 1) = \left(1 + \left\lceil 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil \right) + \left(1 + \left\lceil 2 \cdot 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil \right). \tag{20}$$

One additional cell is added to each item in (20) because we consider the worst-case scenario to be two incomplete cells falling into the probable zones. Similarly, we can have a general form for calculating the number of probable cells in the mobility profiles, $N_x(r : t, k)$, shown in (21) at the bottom of the page. For the *Cases* of $n \le k \le 2n - 1$, the number of cells is doubled since there are two possible zones resulting from the moving direction. Notice that the cells included in our illustration are those that fall into the zone partitions and within the average order $\overline{d_x(t)}$ or prediction level $O_x(t)$. The set or sample space of the mobility profiles is denoted as $\mathcal{N}_x(r : t, k)$, and each cell that belongs to this set is denoted by $X \in \mathcal{N}_x(r : t, k)$.

d) Prediction of location probabilities: Thus far, we have determined the total number of probable cells and the set of those cells. Now, we must estimate the location probability for each cell. We observed that, for a particular cell, the number of paths or travel routes is finite, i.e., the MT is moving among finite states. The $P_{x,i,j}(t)$ for $j = 1, 2, \cdots, N_x(r : t, k)$, which is the probability that an MT x currently in cell i will be in cell j during the timing window ΔT is computed by performing the following procedures.

- **Step 1**: Select a value from $0 < p_0 < 1$ as the initial point for computing the location probabilities.
- **Step 2**: Start from the bottom line of the TRM and take the last two nonzero elements of the TRM to make a temporary path $\mathcal{P}$, as shown in Fig. 6(a).
- **Step 3**: Compare the path $\mathcal{P}$ to the equal or close segment in PD as shown in Fig. 6(b). There may be a set of cells that can be the next cell along with path $\mathcal{P}$, which is represented by a set $\mathcal{P}_x^S$. Each element of this set, $X_j \in \mathcal{P}_x^S$, is a probable cell in Fig. 6(c) and which provides a *possible path*, $\widetilde{\mathcal{P}}(X_j)$.
- **Step 4**: Estimate location probabilities, $P_{x,i,j}(t)$, by performing the process shown in Fig. 7. This algorithm starts by examining each cell in the set of possible cells, $X_j \in \mathcal{P}_x^S$, and the total number of cells in this set is $\mathcal{N}_x(r : t, k)$ from (21). If a probable cell X_j is in the first order of prediction and its corresponding path $\widetilde{\mathcal{P}}(X_j)$ can be found in the historical path database $\mathcal{D}_x^H$, then this cell X_j has the highest location probabilities. This emphasizes the importance of prediction constraints and user history. Additionally, as the probable cells get further away from the MT's

$$N_x(r : t, k) = \begin{cases} k - 1 + \left\lceil 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil + \left\lceil 2 \cdot 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil + \cdots + \left\lceil k \cdot 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil, & \text{if } 1 \le k \le n - 1 \\ 2 \cdot \left(k - 1 + \left\lceil 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil + \left\lceil 2 \cdot 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil + \cdots + \left\lceil k \cdot 6 \cdot \frac{|\theta_x(t)|}{2\pi} \right\rceil \right), & \text{if } n \le k \le 2n - 1 \end{cases} \tag{21}$$

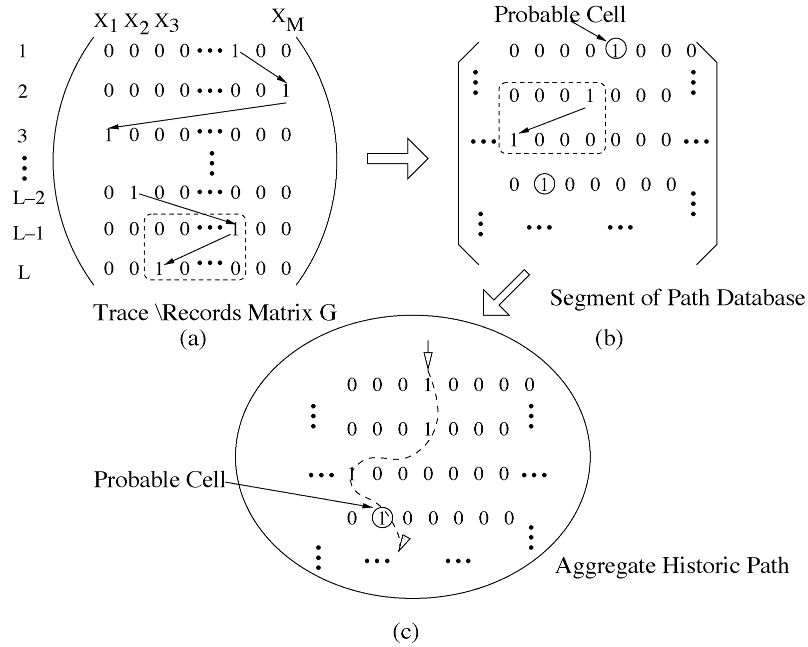


Fig. 6. Prediction of location probabilities.

---

p_0 := initial value of the highest probability
$\mathcal{P}$:= temporary path of TRM
$\mathcal{P}_x^S$:= set of probable cells in PD along path $\mathcal{P}$
$\widetilde{\mathcal{P}}(X_j)$:= possible path of cell $X_j \in \mathcal{P}_x^S$
$\mathcal{N}_x(r:t,k) := \mathcal{N}_x(\max\{\overline{o_x(t)}, O_x(t)\} = r : t,k)$
$P_{x,i,j}(t)$:= location probability at cell j given an MT x
  is currently in cell i
P_0 := threshold of probability computation
while $n \leq N_x(r:t,k)$ **do**
  for all $X_j \in \mathcal{P}_x^S$ **do**
    if $X_j \in \mathcal{N}_x(1:t,k)$ and $\widetilde{\mathcal{P}}(X_j) \in \mathcal{D}_x^H$ **then**
      $P_{x,i,j}(t) := p_0$
    else
      case : $X_j \in \mathcal{N}_x(2:t,k)$ and $\widetilde{\mathcal{P}}(X_j) \in \mathcal{D}_x^H$
        $P_{x,i,j}(t) := \frac{1}{2}p_0$
      case : $X_j \in \mathcal{N}_x(2:t,k)$ and $\widetilde{\mathcal{P}}(X_j) \notin \mathcal{D}_x^H$
        $P_{x,i,j}(t) := \frac{1}{4}p_0$
      case : $X_j \notin (\mathcal{N}_x(1:t,k) \cup \mathcal{N}_x(2:t,k))$ and $\widetilde{\mathcal{P}}(X_j) \in \mathcal{D}_x^H$
        $P_{x,i,j}(t) := \frac{1}{4}p_0$
      others :
      for all $r > 2$ **do**
        while $P_{x,i,j}(t) \leq P_0$ **do**
          case : $X_j \in \mathcal{N}_x(r:t,k)$
            $P_{x,i,j}(t) := (\frac{1}{2})^{r-1}p_0$
          case : $X_j \notin \mathcal{N}_x(r:t,k)$
            $P_{x,i,j}(t) := (\frac{1}{2})^{r}p_0$
        end while
      end for
    end if
  end for
end while

Fig. 7. Estimation of location probabilities (step 4).

current position, and they are not relevant to the historical paths, the location probabilities decrease. This examination continues until all cells in the possible set are scrutinized.

As a result, a sequence of location probabilities is obtained in terms of p_0, where p_0 can be solved by applying the following expression:

$$\underbrace{\sum}_{j\in\left(\mathcal{P}_x^S \bigcup \mathcal{N}_x(r:t,k)\right)} P_{x,i,j}(t) = 1. \tag{22}$$

V. Simulation Model

In this section, assumptions and specifications used in the simulation are described. Instead of assuming in many previous studies, that an MT travels along a highway or one-dimensional (1-D) environment, we consider random behavior in our simulation, which involves much more complicated computation. We consider two scenarios in our service estimation: uniform and nonuniform distributions. As described in Section III-A, i.e., the service description of a quasi-stationary UMP is able to provide the PMF of services as shown in (5).

Suppose there are five types of services supported by the MT's current wireless network, which are audio, video, voice, data, and voicemail. When the uniform PMF is the case, i.e., $\psi_x(a) = 0.2$, the probability of each type of service is equal. As for the nonuniform PMF, we use the same example as in Section III. Each MT is allowed to request any type of service at a particular moment. We will predict the service pattern by using (9)–(15) in Section IV-B.

The time is quantized in intervals $\Delta T = 50$ s, which is a preset time window. This interval is chosen based on cell residence time for each mobile terminal. In reality, different users may have different moving speeds and different mean residence times. However, for a specific cell, the average velocity can be determined for the mobile users in its coverage area. Accordingly, the cell residence time can be calculated by dividing the cell diameter by the average velocity. The calculation time window is much smaller than the cell residence time except in pico-cell systems. In our simulation, the service request from an MT is generated according to a Bernoulli process in each cell and each moment.

We also consider the MT's speed and changes in its moving direction [9], [22]. The MT is allowed to move away from its current position in any direction, and variation from its previous direction is a uniform distribution limited in the range of $\pm(\pi/2)$. The initial velocity of an MT is assumed to be a random variable with a Gaussian pdf truncated in the range of [0,112 km/h] and the velocity increment is taken to be a uniformly distributed random variable in the range of $\pm 40\%$ of the average velocity, 80 km/h. As for the residence time distribution in (1), the values of μ is taken with 1.65 [24].

The most important feature of this simulation is that we use an actual digital map for computing the mobile location probabilities. In the selected segmentation of a map of Atlanta, Georgia, there is an Interstate Highway I-85 and a toll freeway 400 on which most of the city traffic travels. We deliberately chose this segment, 7 km $\times$ 5 km, because it is a combination of diverse environments that provides a number of choices for a traveling MT, thus generating different location probabilities. If we only consider highways, then the location probability prediction is only related to the next cell in a 1-D model. By establishing a grid model, this area is covered by 30 cells covering the area. We find that the maximum number of routes for each cell is 10. Thus, the maximum number of routes in our simulation is 300. Based on the geographical condition of this area, we generate a PD which will be used for predicting location probabilities in Section IV-C2.

A. Effect on Mobility Management

Location update and paging are two fundamental operations for locating an MT. As the demand for wireless services grows rapidly, the signaling traffic caused by location update and paging increases accordingly, which consumes limited available radio resources. Location update is concerned with reporting current locations of the MTs. In a paging process, the system searches for the MT by sending poll messages to the cells close to the last reported location of the MT at the arrival of an incoming call. Delay time and cost are two key factors in the paging issue. Of the two factors, *paging delay*, is very important as the QoS requirement for multimedia services. Paging cost, which is measured in terms of cells to be polled before the called MT is found, is related to the efficiency of bandwidth utilization and should be minimized under delay bound [21].

We will compare the results with the so-called *selective paging* in [1] because the location probabilities are estimated. The cell radius is assumed to be 2 km in our simulation. The full area of the segmentation map is covered by this type of cells, and we assume there is one BS in each cell and all of them are controlled by an MSC, as in Fig. 1. A highest-probability-first (HPF) scheme is introduced in which the sequential polling is performed in decreasing order of probabilities to minimize the mean number of cells being searched [18].

We assume that each LA consists of the same number M of cells in the system. The worst-case paging delay is considered as a delay bound $\mathcal{D}$ in terms of polling cycle. We consider the partition of paging areas (PAs) given that $1 \leq \mathcal{D} \leq M$, which requires grouping cells within an LA into the smaller PAs under delay bound $\mathcal{D}$ [18], [21].Suppose, at a given time, the initial state $\mathbf{P}$ is defined as $\mathbf{P} = [p_1, p_2, \cdots, p_j, \cdots, p_N]$, where p_j is the location probability of the jth cell to be searched in decreasing order of probability. We use triplets $PA^*(i, q_i, n_i)$ to denote the PAs in which i is the sequence number of the PA; q_i is the location probability that the called MT can be found within the ith PA and n_i is the number of cells contained in this PA. An LA can be divided into $\mathcal{D}$ PAs because the delay bound is assumed to be $\mathcal{D}$. Thus, the worst-case delay is guaranteed to be $\mathcal{D}$ polling cycles. The system searches the PAs one after another until the called MT is found.

Accordingly, the *location probability* q_i of the ith PA is $q_i = \sum_{j \in PA^*(i)} p_j$. If the called MT is found in the ith PA, the *average paging cost* under delay bound $\mathcal{D}$, $E[C(\mathcal{D})]$, and average delay, $E[D(\mathcal{D})]$, are computed as follows:

$$E[C(\mathcal{D})] = \sum_{i=1}^{\mathcal{D}} q_i \cdot \sum_{k=1}^{i} n_k$$
$$E[D(\mathcal{D})] = \sum_{i=1}^{\mathcal{D}} i \cdot q_i. \qquad (23)$$

B. Effect on Resource Management

When the estimation of service type is applied to resource management, we define the probability vector $\vec{\mathbf{\Upsilon}}_{x,i}(a,t)$ as

$$\vec{\mathbf{\Upsilon}}_{x,i}(a,t) = [\Upsilon_{x,i,0}(a,t)\ \Upsilon_{x,i,1}(a,t)\ \cdots\ \Upsilon_{x,i,N}(a,t)] \quad (24)$$

where $\Upsilon_{x,i,0}(a,t)$ is the probability that an MT x remains in a cell i given that the MT initiates a call in cell i while this call is an a type of service and it is not over by t. $\Upsilon_{x,i,j}(a,t)$, for $j = 1, 2 \cdots, N$, is the probability that a call is still active in cell j given that this call is initiated in cell i. We also consider the pdf $g(a,t)$, which represents the distribution of call holding time of service type a. With the above consideration and recalling the definition of $\vec{\mathbf{P}}_{x,i}$ of (4) in Section III-B, for an MT x that is initiating a call of service a in cell i, the probability $\Upsilon_{x,0}(a,t)$ can be determined by the following expression:

$$\Upsilon_{x,0}(a,t) = [1 - G(a,t)] \cdot \left[1 - \sum_{j=1}^{N} P_{x,i,j}(t)\right] \qquad (25)$$

where $G(a,t)$ is the cumulative distribution functions (cdfs) for $g(a,t)$ and $P_{x,i,j}(t)$, for $j = 1, 2, \cdots, N$, is the probability that the MT will be in cell j in (25), which is the product of two probabilities because we assume the calling pattern is independent of an MT's movement. $[1 - G(a,t)]$ is the probability that the call is not over by time t, and $[1 - \sum_{j=1}^{N} P_{x,i,j}(t)]$ is the probability that the MT is still in cell i at time t, given that there are N probable cells.

For the other cells in the shadow cluster, we define the cdfs for N probable cells with service a as $\vec{\mathcal{G}}_{x,i}(a,t) =$

$[\mathcal{G}_{x,i,1}(a,t)\ \mathcal{G}_{x,i,2}(a,t)\ \mathcal{G}_{x,i,N}(a,t)]$. Therefore, $\vec{\mathbf{\Upsilon}}_{x,i}(t)$ is determined by

$$\underbrace{\vec{\mathbf{\Upsilon}}_{x,i}(a,t)}_{i\neq 0} = \left[\vec{I_N} - \vec{\mathcal{G}}_{x,i}(a,t)\right] \cdot \mathbf{\Pi} \tag{26}$$

where $\vec{I_N}$ is a vector with N ones and Π is a diagonal matrix as

$$\mathbf{\Pi} = \begin{bmatrix} P_{x,i,1}(t) & 0 & \cdots & 0 \\ 0 & P_{x,i,2}(t) & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & & \cdots & P_{x,i,N}(t) \end{bmatrix}. \tag{27}$$

As a result, each element of $\vec{\mathbf{\Upsilon}}_{x,i}(a,t)$ in (24) is determined. Accordingly, the bandwidth needed in the next probable cells for an MT x in cell i, $\vec{\mathbf{W}}_{x,i}(a,t)$, can be reserved as

$$\vec{\mathbf{W}}_{x,i}(a,t) = a \cdot \vec{\Upsilon}_{x,i}(a,t). \tag{28}$$

We assume that the time is quantized in slots of length ΔT. 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 time slot where the request was received. Every BS gathers call connection requests from MTs in its cell and checks whether its current resources can support the requested equivalent bandwidth. A call is admitted if the sum of the equivalent bandwidth at a link is less than the link capacity, which is also called semi-resource reservation [11]. We use the metrics of handoff call dropping probability for handoff calls and call blocking probability for new arrivals.

VI. Comparison and Evaluation

A. Comparison of Single-Cell and Multicell Prediction

In the proposed UMP framework, we consider the following factors in predicting future locations of mobile users: 1) historical records; 2) path information; 3) statistical model; 4) moving direction and velocity; 5) current position of a mobile object; and 6) shadow cluster, i.e., a set of possible cells. All of these factors impact location prediction as described in previous sections. In this section, we evaluate the proposed scheme with regard to the effect on mobility and resource management.

In addition, we compare the proposed scheme with previous work on location prediction, which can be categorized into single-cell prediction [5], [8], [13], [17] and multicell prediction [1], [4], [12]. In the first category, each of them considers only two or three factors compared to six factors covered in our proposed scheme. Moreover, location estimation is focused on the next cell instead of a group of cells. Therefore, the impact of other factors is ignored and may also overlook the possibilities in other cells. In particular, if there is no new interaction between the system and a mobile user since the last update, multicell prediction will have an important influence on user tracking. Regardless of the reasons causing the interruption of communications, the system can always initiate searching or testing a mobile object based on its profiles in mobile environments. Moreover, it may not be efficient to generate dynamic UMP frequently for micro- or pico-cell networks, e.g., per cell; instead, we can adjust the window time for calculating a UMP to reduce computation and communication cost. The effect of single-cell prediction and multicell prediction are shown in Figs. 10 and 11.

For the second category, i.e., multicell prediction, it has been very difficult to compare the location probabilities quantitatively. Thus, most of the literature will elaborate on the rationale of the proposed schemes and demonstrate the results by the effect on either mobility management or resource management [4], [12], [23]. To find a group of cells in multicell prediction in [4], the most likely cluster (MLC) based on directional probabilities needs to be determined. The method of MLC further approaches and reduces the number of cells defined. The location probabilities, therefore, are defined as the fraction of directional probability in one cell to all of the cells through which a mobile object can traverse during the window time.

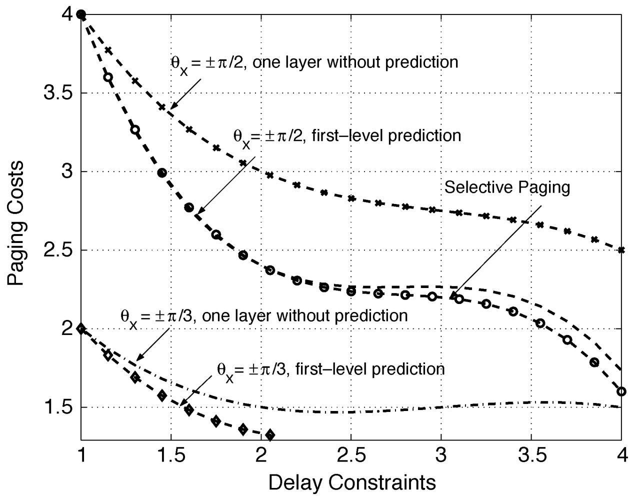
However, there are three major concerns of this method. First, it depends on accurate direction measurements using the Global Positioning System (GPS). In our algorithm, as long as the direction measurement is accurate to 15°, no cells will be missed in the prediction, which is easier to implement in real systems. Second, the MLC solution does not take user history into consideration. Although user history may not be important for macrocell systems, it is critical to micro- and pico-cell systems because mobile objects are very likely to change their directions frequently. Third, there was no consideration of path information in [1] and [4]. No matter how likely a mobile object can move into a region from the analysis based on current measurements, it must move into an area that allows movement continuity, i.e., a path must exist in the next cell. Therefore, we take the path database into account by comparing the available path with the predictive future path. The algorithm described in [1] is labeled "Selective Paging" in Figs. 8 and 9.

B. Numerical Results

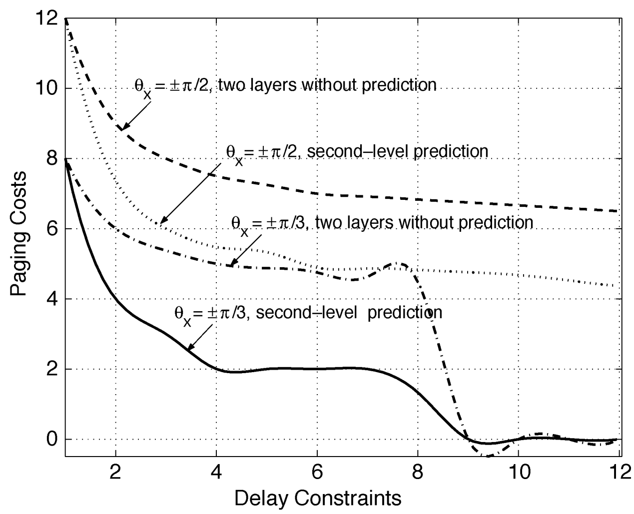
First, we investigate the convergence of mean-square error by using (14). For the uniform PMF, the first convergence point is approximately $L = 65$, while it is approximately $L = 120$ for a nonuniform PMF. Thus, if an MT subscribes a service package which conforms to a uniform PMF, the next probable service type can be predicted more accurately compared to a nonuniform PMF. Accordingly, the bandwidth requirement can be determined. We estimate service type α_o for an MT using the algorithm in Section IV-B. In our experiments, both the uniform and nonuniform PMFs are considered. We compute the mean-square errors using (13) and (14). Then, we determine the service type according to the scheme in Fig. 4.

Then, we predict the probable cells using the algorithm described in Section IV-C. For $\theta_x(t) = \pm\pi/3$ and $\theta_x(t) = \pm\pi/2$, we first determine the probable zones using (16) and (17), limiting the probable cells in a particular region. Next, the number of probable cells is computed using (1), (18), and (21).

There are many ways to evaluate the effectiveness of the prediction algorithm of computing location probabilities [8], [12], [23], as we introduced in Section I. Here, we show the effect of these results on paging issues since it involves both paging costs and paging delays. Paging is the process by which the MSC sends polling message to BSs in its management area to determine the serving cell of the called MT. Paging cost affects

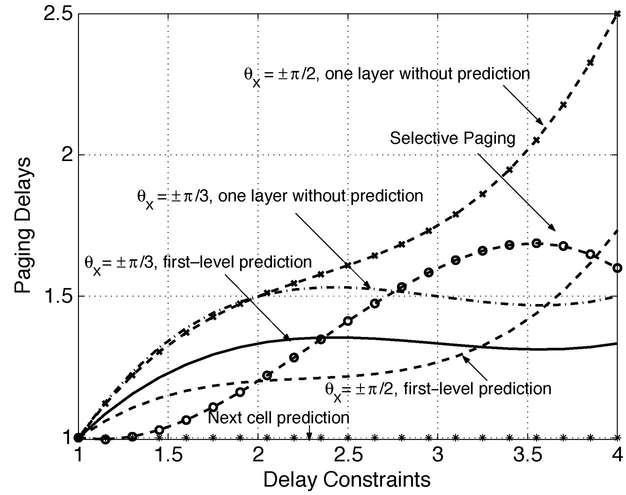


(a)

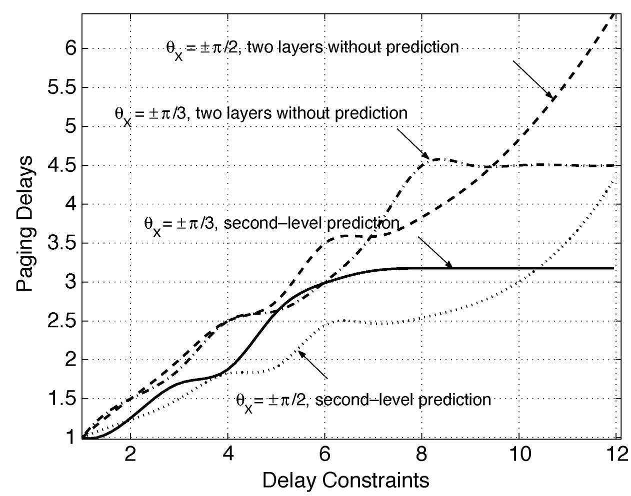


(b)

Fig. 8. Comparison of paging costs. (a) First-level prediction. (b) Second-level prediction.



(a)



(b)

Fig. 9. Comparison of paging delays. (a) First-level prediction. (b) Second-level prediction.

network resources because the paging message is sent via downlink channels; thus, it should be reduced as much as possible. Paging delay is part of call delivery delay, and it is related to QoS requirement. Thus, it should also be reduced so that the call connection can be established quickly.

Paging costs resulting from location probabilities of the first and second levels of prediction are compared to those of uniform distribution assumed in the existing paging schemes [18], [21]. Note that the number of cells is determined by the prediction level and moving direction presented in Section IV-C2, which has a strong impact on the performance of the scheme. Increasing the prediction level and change in direction, by including more cells, increases the likelihood of supporting QoS in future possible cells. However, this will cause more computations and communications. Therefore, we compare paging cost and delay with regard to change in direction and prediction levels. We conduct the experiments on the following cases: $\theta_x = \pi/3$ with first- and second-level prediction; $\theta_x = \pi/3$ without prediction for the cells in one or two layers surrounding the current cell; $\theta_x = \pi/2$ with first- and second-level prediction.

The results of paging costs are given in Fig. 8(a) and (b), in which paging costs are measured by the number of cells to be searched before finding the MT. Since the location probabilities provided in [1] are related to three cells, the paging cost is better than the two-cell prediction and worse than the four-cell prediction, as shown in Figs. 8(a) and 9(a). When the variation of the moving direction is high, causing more probable changes, the improvement of paging costs is more visible in Fig. 8(a). For example, when $\theta_x = \pm(\pi/3)$, the reduction in paging costs due to the location probabilities is not as large as that of $\theta_x = \pm(\pi/2)$. This means that it is more important to predict location probabilities if the MTs are moving randomly, i.e., the movement of the MTs is not uniformly distributed in the location area.

The paging cost of the proposed scheme is very close to the scheme proposed in [1] for the first-level prediction because there are a small number of cells in the shadow cluster. Therefore, the benefits of the proposed scheme are not evident compared to the "Selective Paging" scheme, which also covers the cells adjacent to the current cell. If the prediction level is higher, the paging costs are significantly reduced compared to those without prediction. Specifically, if MTs are moving very fast

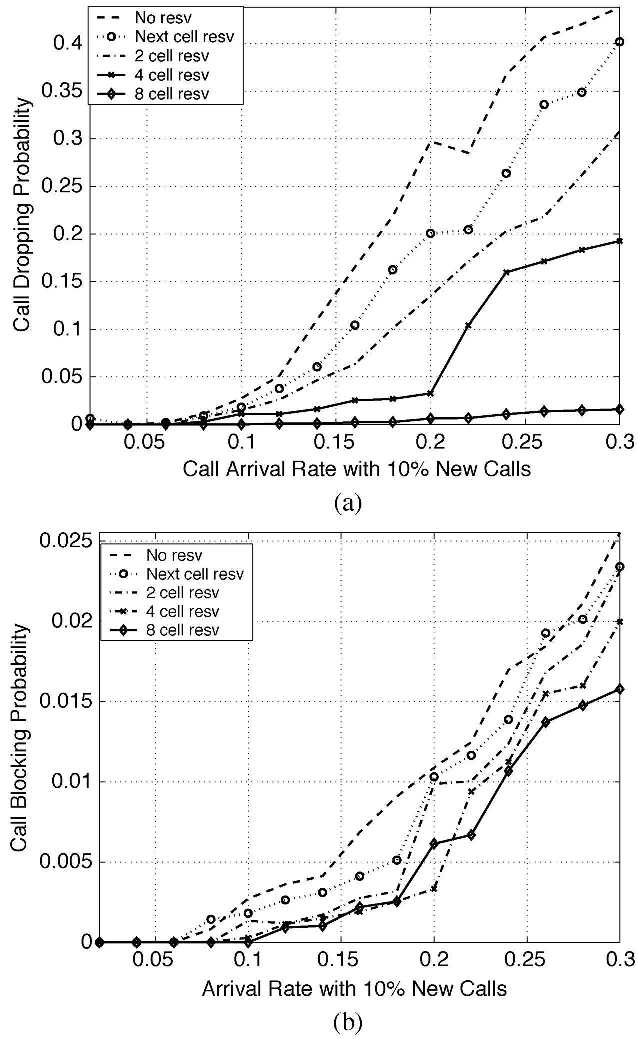


Fig. 10. Dropping and blocking probabilities with 10% new calls. (a) Handoff dropping probabilities. (b) New-call blocking probabilities.

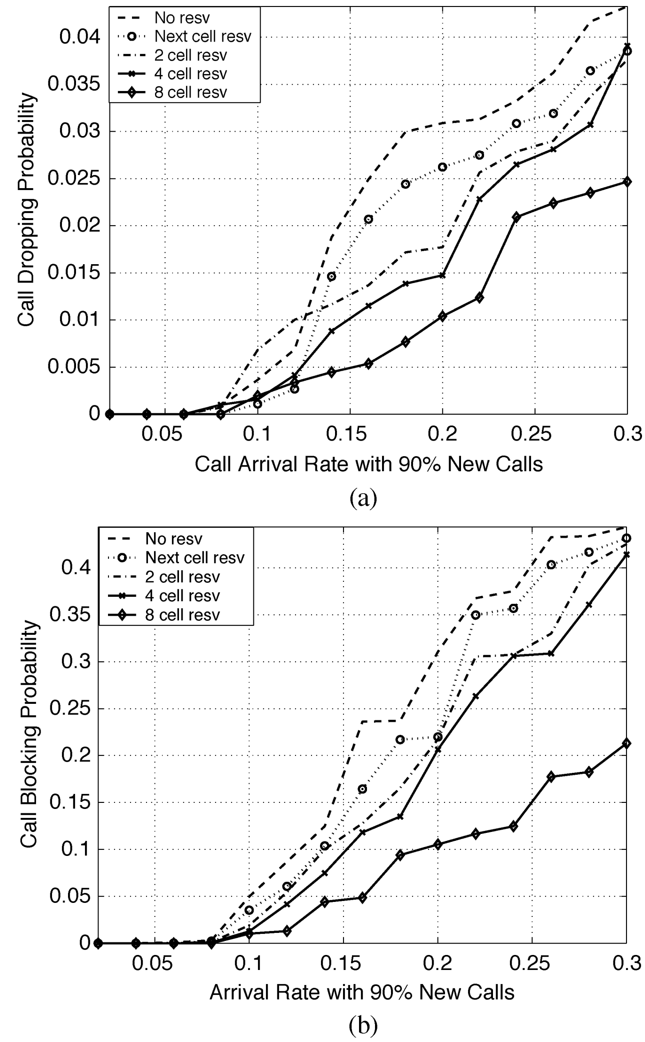


Fig. 11. Dropping and blocking probabilities with 90% new calls. (a) Handoff dropping probabilities. (b) New call blocking probabilities.

and are expected to other cells in a short time, there are more probable cells. As a result, location probability in each cell is smaller. As a result, it is more difficult to locate the MT. Accordingly, the prediction of MTs' location probabilities is more effective and more important.

Paging delays are compared in Fig. 9(a) and (b), which are measured in terms of polling cycles. Each polling cycle is the time from sending a paging request to receiving a response. In Fig. 9(a), the paging delays are greatly reduced in comparison to not using prediction. As the delay constraints increase, the average paging delays increase while the paging costs decrease. We notice that the delays are reduced even more when the delay constraints are greater. Also, the predicted probabilities are more effective on reducing paging delays when the prediction level is higher, as shown in Fig. 9(b). This is especially useful for those MTs moving in wide areas, where there are many paths available instead of highway scenarios.

We also conduct simulations of resource management. The predicted service type and location probabilities are used to reserve the equivalent bandwidth for the mobile terminals. In our simulation, we consider a different new-call ratio as the number of requests of new calls to the total number of call requests. The results shown in Figs. 10 and 11 are the statistics of 20 000 call requests. If the requested bandwidth cannot be allocated, then a new call will be blocked or the handoff call will be dropped.

We consider that handoff calls have higher priority than new calls. Therefore, call dropping/blocking probabilities are related to the new call ratio, which is the fraction of new calls in the total number of connections requested. We compare no-reservation, next-cell reservation, two-cell reservation, and four-cell reservations. In particular, next-cell reservation corresponds to single-cell prediction discussed in Section IV-A. In Figs. 10 and 11, we can observe that call dropping probabilities increase as call arrival rates increase, indicating that the bandwidth release depends on the call departure. Given the fixed call holding time, the more call arrivals there are, the higher call dropping and blocking probabilities are. Meantime, we can see that the handoff dropping probabilities decrease as we reserve bandwidth in more cells, i.e., better than single-cell prediction. The effect of reservation is obviously on the handoff dropping probabilities as opposed to on the new call blocking probabilities. If major call requests come from handoff, then call dropping probabilities are higher, compared to the case in which the new calls are dominant.

C. Discussions on Implementation and Overhead

Finally, we discuss the implementation of our solution in real systems and the overhead of computation and communications.

1) Implementation in Existing Wireless Networks: There are three aspects relating to the implementation and incorporation within existing networks: location description, real-time monitoring of speed and velocity, and collection of historical records.

Location information: There is no extra effort is required to know a mobile user's current LA because of location registration process. In third-generation (3G) wireless systems, LCS is a new feature to provide location information. The proposed user profile framework can be used as an access point to user profile data for the service providers [19]. In addition, given a cell ID, we know to which zone a cell belongs since the zone partition is determined during the system design. There are several ways for wireless systems to collect cell ID, such as during the process of delivering an incoming service or the process of establishing a routing path for an outgoing service. Furthermore, an MT's location can also be obtained through LCS management provided by wireless systems, which may be call-independent. In addition to the information of cell IDs, the LCS also provides the geographical estimation about an MT in terms of universal latitudinal and longitudinal data. Mapping between the MT's position in terms of latitude and longitude and local coordinate system is finished through location client coordinate transformation function.

Measurement and real-time monitoring: The quasi-stationary information are stored in the HLR and VLR, which can be updated in three processes: location update, call origination, and call termination. The measurement of directions and velocity is an issue of real-time monitoring. There have been a few algorithms for mobile velocity and direction estimations [9], [16], [22]. The velocity information can be collected during the handoff because velocity estimation is required to keep handoff delay acceptable. The detailed algorithms about velocity estimation are beyond the scope of this paper.

Other related work also addresses the scalability issue of monitoring and collecting mobile users' information by carefully constructing the databases that maintain the querying records [10]. This work has demonstrated that the BSs, location measurement units (LMUs), and VLRs are able to support discrete information monitoring and collecting. For the continuous monitoring in which the mobile objects are monitored continuously over a time period until the MTs or BSs interrupt, simulation results showed that the existing BSs and MTs are capable of handling even continuous monitoring if the databases and safe regions are designed appropriately.

2) Overhead of Prediction Algorithms: The overhead that will be incurred by the proposed scheme consists of three parts: 1) buffer space to store historical records; 2) communication cost to send prediction results to probable cells for bandwidth reservation; and 3) computation time required to perform the algorithms to obtain prediction results.

As described in Section III-A, the quasi-stationary UMP will be updated during the procedures of location update, call origination, and termination. This information is stored in the VLR/SGSN and HLR; therefore, there are no extra requirements for storing quasi-stationary information. In the proposed framework, each MT keeps a list that stores the cell IDs that it has traversed and the services that it has requested. This list is refreshed when the MT experiences a handoff or incoming or outgoing service. According to the specifications of cellular handsets, each one can store 50–100 records of calls and an additional 50–100 frequent calling users. Each of these records can accommodate 180 characters, which results in a total storage of $180 \times 8 \times 100 = 144\,000$ b to $576\,000$ b. Considering each cell ID is 15 b [3], 50-cell IDs will take just 750 b. The historical records of cell IDs take up to 0.5% auxiliary storage space in a cellular handset, which is a very small amount. Thus, historical records are considered in many schemes [5], [8]. Note that the historical records can also be stored in the serving BS rather than in the terminals; therefore, using historical records will not cause significant overhead in storage.

The UMP information can be forwarded from the previous BS to the current serving BS, which is transmitted via the wired interface between BSs. Therefore, the radio resource used to communicate between the MTs and BSs will not be consumed while delivering user information. Compared to other algorithms introduced in Section I, the proposed scheme will incur extra communication costs for distribution prediction results to other cells. Let us consider that the maximum number of cells to be predicted in our algorithm is $N_x(r : t, k)$ in (21) for terminal x given that the MT is currently in zone k, and the prediction level is r. Then, the total communication cost of informing other cells is obtained as $N_x(r : t, k) \times C_c$, where C_c is the communication cost for one cell, i.e., the cost for all algorithms of next-cell prediction. Therefore, the overhead of the proposed scheme is closely dependent on the number of cells in the mobility profiles, which is also the number we need to reduce as much as possible. However, this communication cost does not take any radio resources; instead, it uses dedicated wirelines between base stations. With the transmission speed in fiber optics at 10 Gb/s, this communication overhead is trivial. But, we obtain benefits in terms of QoS improvement and decreased tracking cost. Moreover, we can further reduce the overhead by grouping users.

VII. Conclusion

We have explored a fundamental issue of providing high-level QoS in wireless networks. Noticing that the key point to QoS-based mobile networks is the knowledge of service requirements and future locations prior to the arrival of mobile objects, we first proposed a novel framework of UMP, considering many important factors associated with the MTs' behavior. The service requirement is estimated by using a mean-square error method based on the historical records and service probability distributions. Moreover, we introduced the concept of zones and prediction levels to shrink the region of probable cells. In the proposed prediction algorithms, we took several important factors, including direction and velocity of mobile objects, historical records, stochastic model of cell residence time, and path information into account. Therefore, the service requirement and future locations are predicted more accurately compared to previous schemes because this is the first time that all of these factors are considered. As examples, we provided the simulation results to demonstrate that the proposed algorithms for mobility and resource management are effective in terms of reducing location tracking cost, delays, and call dropping/blocking probabilities.

REFERENCES

[1] A. Abutaleb and V. O. K. Li, "Paging strategy optimization in personal communication system," *ACM-Baltzer J. Wireless Networks (WINET)*, vol. 3, pp. 195–204, Aug. 1997.

[2] I. F. Akyildiz and W. Wang, "A dynamic location management scheme for next generation multi-tier PCS systems," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 2040–2052, Jan. 2002.

[3] D. P. Agrawal and Q.-A. Zeng, *Introduction to Wireless and Mobile*. Pacific Grove, CA: Systems Thomson Brooks/Cole, 2003.

[4] A. Aljadhai and T. F. Znati, "Predictive mobility support for QoS, provisioning in mobile wireless environments," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1915–1931, Oct. 2001.

[5] A. Bhattacharya and S. K. Das, "LeZi-update: an information-theoretic approach to track mobile users in PCS networks," in *Proc. ACM/IEEE MobiCom*, Aug. 1999.

[6] E. Cayirci and I. F. Akyildiz, "User mobility pattern scheme for location update and paging in wireless systems," *IEEE Trans. Mobile Comput.*, vol. 1, pp. 236–247, July–Sept. 2002.

[7] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for Ad Hoc network research," *Wireless Commun. Mobile Comput.*, vol. 2, no. 5, pp. 483–502, 2002.

[8] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservations for hand-offs in QoS-sensitive cellular networks," in *Proc. ACM SIGCOMM*, Vancouver, BC, Canada, August 1998, pp. 155–166.

[9] M. Hellebrandt, R. Mathar, and M. Scheibenbogen, "Estimation position and velocity of mobiles in a cellular radio network," *IEEE Trans.Veh. Technol.*, vol. 46, pp. 65–71, Feb. 1997.

[10] C. S. Jensen, A. Friis-Christensen, T. B. Pedersen, D. Pfoser, S. Saltenis, and N. Tryfona, "Location-based services—a database perspective," in *Proc. 8th Scandinavian Research Conf. Geographical Information Science*, June 2001, pp. 59–68.

[11] G. S. Kuo, P.-C. Ko, and M.-L. Kuo, "A probabilistic resource estimation and semi-reservation scheme for flow-oriented multimedia wireless networks," *IEEE Commun. Mag.*, vol. 39, pp. 135–141, Feb. 2001.

[12] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Trans. Networking*, vol. 5, pp. 1–12, Feb. 1997.

[13] T. Liu, P. Bahl, and I. Chlamtac, "Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 922–936, Aug. 1998.

[14] Y.-B. Lin, "Reducing location update cost in a PCS network," *IEEE/ACM Trans. Networking*, vol. 5, pp. 25–33, Feb. 1997.

[15] W. Ma and Y. Fang, "Two-level pointer forwarding strategy for location management in PCS networks," *IEEE Trans. Mobile Comput.*, vol. 1, pp. 32–45, Mar. 2002.

[16] R. Narasimhan and D. C. Cox, "Speed estimation in wireless systems using wavelets," *IEEE Trans. Commun.*, vol. 47, pp. 1357–1364, Sept. 1999.

[17] X. Shen, J. W. Mark, and J. Ye, "User mobility profile prediction: an adaptive fuzzy inference approach," *ACM-Baltzer J. Wireless Networks*, vol. 6, pp. 363–374, June 2000.

[18] C. Rose and R. Yates, "Minimizing the average cost of paging under delay constraints," *ACM-Baltzer J. Wireless Networks*, vol. 1, pp. 211–219, Feb. 1995.

[19] "Digital Cellular Telecommunications System (Phase 2+); Location Services (LCS); Location Services Management (GSM 12.71 version 8.0.1 Release 1999)," ETSI, TS 101 513 V8.0.1 (2000-11), Nov. 2000.

[20] W. Wang, "Modeling and management of location and mobility," in *Wireless Information Highways*, D. Katsaros, A. Nanopoulos, and Y. Manalopoulos, Eds. Hershey, PA: IRM Press, 2005, ch. VI.

[21] W. Wang, I. F. Akyildiz, G. Stüber, and B.-Y. Chung, "Effective paging schemes with delay bounds as QoS constraints in wireless systems," *ACM-Baltzer J. Wireless Networks*, vol. 7, no. 5, pp. 455–466, Sept. 2001.

[22] C. Xiao, K. D. Mann, and J. C. Olivier, "Mobile speed estimation for TDMA-based hierarchical cellular systems," *IEEE Trans. Veh. Technol.*, vol. 50, pp. 981–991, July 2001.

[23] T. Zhang, E. V. D. Berg, J. Chennikara, P. Agrawal, J.-C. Chen, and T. Kodama, "Local predictive resource reservation for handoffs in multimedia wireless IP networks," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1931–1941, Oct. 2001.

[24] 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.

Ian F. Akyildiz (M'86–SM'89–F'95) received the B.S., M.S., and Ph.D. degrees in computer engineering from the University of Erlangen-Nuernberg, Erlangen, Germany, in 1978, 1981, and 1984, respectively.

Currently, he is the Ken Byers Distinguished Chair Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, and Director of Broadband and Wireless Networking Laboratory. His current research interests are in Sensor Networks, InterPlaNetary Internet,Wireless Networks and Satellite Networks. He is an Editor-in-Chief of *Computer Networks* and for the newly launched *Ad Hoc Networks Journal*.

Dr. Akyildiz is a Fellow of the Association for Computing Machinery (ACM). He served as a National Lecturer for ACM from 1989 until 1998 and received the ACM Outstanding Distinguished Lecturer Award for 1994. He was also the recipient of the 1997 IEEE Leonard G. Abraham Prize of the IEEE Communications Society for his paper entitled "Multimedia Group Synchronization Protocols for Integrated Services Architectures." He received the 2002 IEEE Harry M. Goode Memorial Award with the citation "for significant and pioneering contributions to advanced architectures and protocols for wireless and satellite networking," the 2003 IEEE Best Tutorial Award from the IEEE Communicatons Society for his paper entitled "A Survey on Sensor Networks," and the 2003 ACM SIGMOBILE Award for his significant contributions to mobile computing and wireless networking.

Wenye Wang (M'98) received the B.S. and M.S. degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 1986 and 1991, respectively,and the M.S.E.E. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, in 1999 and 2002, respectively.

She is now an Assistant Professor with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh. Her research interests are in mobile and secure computing, quality-of-service sensitive networking protocols, mobility, security, and resource management in single- and multi-hop networks.

Dr. Wang is a member of the Association for Computing Machinery. She has served on program committees for IEEE INFOCOM, ICC, and ICCCN. She also serves on the Editorial Board of *Computer Networks*.