Location Management for Next-Generation Personal Communications Networks

Vincent W.-S. Wong and Victor C. M. Leung, The University of British Columbia

Abstract

This article presents a survey on location management algorithms for next-generation personal communications networks. We first describe different static and dynamic location update algorithms. Then we discuss various selective paging strategies. We also present various modeling techniques that have been used for the performance analysis of location update and terminal paging. We conclude by stating a number of open problems that need to be addressed for the deployment of next-generation PCNs.

ver the past few years, there has been tremendous growth in wireless communications. Personal communications service (PCS) subscribers are increasing at an exponential rate and will continue to increase in the near future. The next-generation personal communications network (PCN) is being standardized as part of the International Mobile Telecommunications 2000 (IMT-2000) system [1], whose goal is to unify many diverse systems existing today (including PCS, two-way paging, mobile satellite, etc.) into a seamless radio infrastructure capable of offering a wide range of services.

The current PCNs use a cellular architecture. The geographical coverage area is partitioned into *cells*, each served by a base station. Mobile users and their terminals are connected to the network via the base stations. Cells can have different sizes: picocells are commonly used in indoor environments; microcells are used within cities; macrocells are used in rural areas and to cover highways. Smaller cells use less power for transmission and allow greater frequency reuse. Several base stations are connected to a base station controller, and a number of base station controllers are then connected to a mobile switching center. The connection of the base stations, base station controllers, and mobile switching centers, along with the radio links between the base stations and mobile terminals, form the *access network*.

Location management enables the network to track the locations of users and their terminals between call arrivals.[1] Since mobile users are free to move within the coverage area, the network can only maintain the approximate location of each user. When a connection needs to be established for a particular user, the network has to determine the user's exact location within the cell granularity. The operation of inform-

ing the network about the current location of the mobile user is known as *location update* or *location registration*, and the operation of determining the location of the mobile user is called *terminal paging* or *searching*.

It is well known that there is a trade-off between the costs of location update and paging. If the mobile terminal updates its location whenever it crosses a cell boundary, the network can maintain its precise location, thus obviating the need for paging. However, if the call arrival rate is low, the network wastes its resources by processing frequent update information, and the mobile terminal wastes its power transmitting the update signal. On the other hand, if the mobile terminal does not perform location update frequently, a large coverage area has to be paged when a call arrives, which wastes radio bandwidth. Thus, the central problem of location management is to devise algorithms that minimize the overall cost of location update and paging.

Current PCNs use a location area (LA)-based update algorithm and blanket polling paging strategy. The coverage area is partitioned into a number of LAs, each containing a group of cells. All base stations within the same LA broadcast the identifier (ID) of their LA periodically. Each mobile terminal compares its registered LA ID with the current broadcast LA ID. Location update is triggered if the two IDs are different. Upon a call arrival for a particular mobile terminal, all the cells within its current LA are polled simultaneously, ensuring success within a single step.

Although the LA-based update scheme is widely adopted by current cellular systems and can be extended to next-generation wireless broadband networks, there are a number of inef-

[1] In this article, we use the terms mobile user and mobile terminal interchangeably.

ficiencies associated with this scheme. First of all, for an LA with a large number of cells, a significant amount of radio bandwidth is consumed in paging for each call arrival. This may not be scalable for future wireless broadband networks with a large number of mobile users. Second, mobile terminals located close to an LA boundary may perform excessive location updates as they move back and forth between two LAs, thus increasing the signaling and processing load on the network database. In addition, since each user has its own mobility pattern, it is difficult to choose an LA size which is optimal for all users. Current research on location management focuses on per-user-based algorithms in which location update and terminal paging procedures can be adjusted dynamically based on a user's call and mobility patterns.

In essence, location management for next-generation PCNs has to address the following issues:
- When should a mobile terminal update its location to the network?
- When a call arrives, how should the exact location of the called mobile terminal be determined within a specific time constraint?
- How should user location information be stored and disseminated throughout the network?

The first two questions pertain to the design of location update and paging algorithms, whereas the last question pertains to database architecture and management. In this article we mainly focus on discussion of different distributed location update and paging algorithms. However, we would like to point out that in recent years, different database management schemes have been proposed to reduce user profile lookup time and signaling traffic. These methods utilize techniques such as hierarchical partitioning [2], load balancing [3], and data replication [4], which depend on the mobility and call patterns of mobile users.

There are a number of survey papers on location management in wireless networks (e.g., [5, 6]). In particular, [6] provides a comprehensive survey on various handoff and location management techniques for the public land mobile networks (PLMNs), mobile Internet Protocol (IP), wireless asynchronous transfer mode (ATM), and satellite networks.

The objective of this article is to describe the current research efforts on location management for next-generation PCNs. Particular emphasis is placed on:
- Discussion of various recently proposed location update and paging algorithms
- A survey of different modeling techniques that have been used for performance analysis of location update and terminal paging

This article is complementary to [6]; together, they provide a more complete picture on location management for future PCNs.

The rest of this article is organized as follows. We survey different static and dynamic location update algorithms. Then we discuss various selective paging algorithms. We summarize different teletraffic models for performance analysis. We conclude by stating a number of open problems that need to be addressed for the deployment of next-generation PCNs.

Location Update

In order to reduce its location uncertainty, each mobile terminal has to report its location from time to time. The location update procedure begins with an update message sent by the mobile terminal over the uplink control channel, which is followed by some signaling procedures which update the
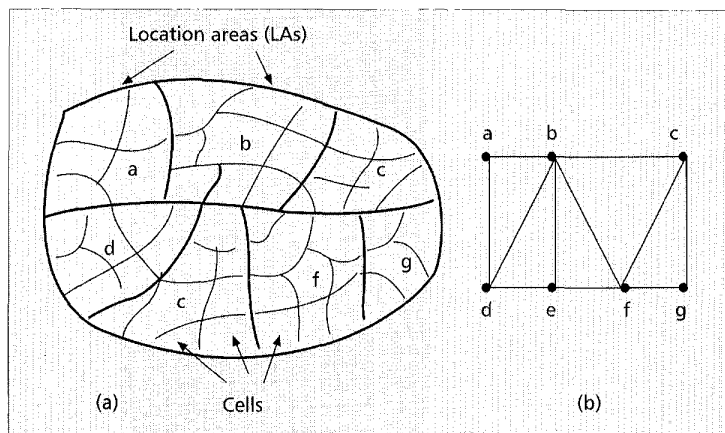
■ Figure 1. *A representation of an actual cellular network topology by a graph: a) the cell and LA topology in a cellular network; b) a graph model showing the interconnections of the LAs.*

database. Location update algorithms can be divided into two main groups: static and dynamic. In a static algorithm, location update is triggered based on the topology of the network. In a dynamic algorithm, location update is based on the user's call and mobility patterns. In this section we summarize various location update schemes proposed in the literature.

Selective LA Update

The rationale behind the *selective LA update scheme* [7] is that a daily commuter may cross a number of LAs on his/her way to and from work. However, he/she may only stay in some LAs for very short periods of time. Rather than performing location update whenever he crosses a new LA, the update process at certain LAs can be skipped.

In [7], an analytical model is introduced in which the interconnections of the LAs are characterized by a graph model. A Markov movement model is used. The residence time in each LA follows a geometric distribution. A genetic algorithm is used to obtain the near-optimal solutions. For low residing probability in certain LAs and high update cost, results show that this scheme incurs a lower location management cost than the conventional LA-based scheme.

For implementation, information regarding the transition probabilities and residence time is required. To estimate the transition probabilities between LAs for a particular user, his/her movements throughout the day can be observed over long periods of time. Since the LA-based update scheme is used in current PCNs, information about the frequency of his/her transition from one LA to another can be retrieved from the database.

Profile-Based

The goal of the *profile-based location update scheme* [8] (also known as the *alternative location strategy* [9]) is to reduce the update cost by taking advantage of the user's mobility pattern. The network maintains a profile for each user, which includes a sequential list of the LAs the user is most likely to be located at in different time periods. This list is sorted from the most to least likely LA where a user can be found. When a call arrives, the LAs on the list are paged sequentially. As long as the mobile terminal moves between LAs within the list, no location update is necessary. Location update is performed only when the mobile terminal moves to a new LA not on the list. The list may be derived from the user's movement history.

As an example, in Fig. 1a, the set of LAs is $\{a, b, c, d, e, f, g\}$. Suppose during the time period (t_1, t_2), the sequential list for the mobile terminal is $\{b, a, e, f\}$. Thus, no update is required as long as the mobile terminal stays within those LAs. Upon a call arrival, the network will page location area b
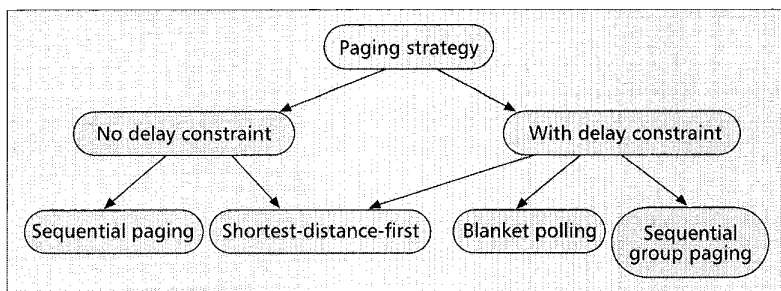
■ Figure 2. *A classification of different paging schemes.*

first, followed by location area *a*, and so on, until the mobile terminal is found or the list exhausted. For implementation, each mobile terminal must maintain a valid sequential list corresponding to a particular time interval. This list has to be updated from time to time.

Movement-Based

In the *movement-based update scheme* [10], each mobile terminal counts the number of boundary crossings between cells incurred by its movements. Location update is performed when this number exceeds a predefined movement threshold *M* (e.g., *M* = 6). This scheme allows the dynamic selection of the movement threshold on a per-user basis.

For implementation, the mobile terminal only needs a counter to count the number of cell boundary crossings. The counter is reset whenever it reaches the movement threshold. The *cell identification code* (CIC) [11] can also be used. With the CIC each cell is assigned a code, which is not necessarily unique. The code is used to identify the cell's orientation relative to other cells within the same LA. Each cell periodically broadcasts its identification code through the downlink control channel. The mobile terminals use this information to facilitate the update decision. For the movement-based scheme in an arbitrary cell topology, only four different codes are necessary [11].

In [12], an analytical model is introduced to determine the optimal movement threshold. The model is applicable for mesh and hexagonal cell configurations under the assumptions of a general cell residence time distribution and symmetric random walk movement pattern. The maximum paging delay constraint is considered, and a shortest-distance-first order paging scheme is used.

Timer-Based

In the *timer-based update scheme* [10], each mobile terminal updates its location every *T* time units (e.g., *T* = 1 hour). This scheme does not require the mobile terminal to record or process location information during the time between updates. For implementation, the timer threshold can be programmed into the mobile terminal by a hardware or software timer.

An analytical model is introduced in [13] to study the timer-based scheme. Assuming Gaussian distribution of user location probability and Poisson call arrivals, the update period that minimizes the cost of location update and paging is derived. Results show that the timer-based scheme performs substantially better than the LA-based scheme.

A variation of the timer-based scheme called the *adaptive threshold scheme* is proposed in [14]. The mobile terminal transmits an update message every *T'* time units, where the parameter *T'* (referred to as the *location registration threshold*) is not a constant, but varies with the current signaling load on the uplink control channel of the base station. Numerical results, under the assumptions of a one-dimensional linear model and random walk mobility patterns, show that the adaptive threshold scheme has better performance than the static timer-based scheme.

Distance-Based

In the *distance-based update scheme*, each mobile terminal tracks the distance it has moved (in number of cells) since the last update and transmits an update signal whenever the distance exceeds a certain threshold. For implementation, the mobile terminal requires some knowledge of cell topology. In order to identify the cells within the distance threshold or along the distance threshold boundary, the mobile terminal needs to download a set of these cell IDs after each location update. Even though CIC can be used [11], its implementation is restricted to some particular paging strategies.

The distance-based scheme has been studied extensively. In [10], the authors compared the movement, timer, and distance-based schemes under the assumptions of random walk mobility movements and a ring topology of cells. Analytical results show that the distance-based scheme gives the lowest location management cost. In [15], the distance-based update scheme is formulated as an optimization problem. The goal is to minimize the expected total cost for update and paging within a time interval. Under a one-dimensional linear model and symmetric random walk movement patterns, the optimal distance threshold is determined by dynamic programming. In [16], an iterative approach is used to compute the optimal distance threshold in a two-dimensional hexagonal model under the assumption of symmetric random walk mobility pattern.

Predictive Distance-Based

In the *predictive distance-based update scheme* [17], the mobile terminal reports both its location and velocity during the update process. Based on the above information, the network determines the probability density function of the mobile's location, which is used to predict the mobile terminal's location in future time. This prediction information is made available to both the network and mobile terminal. The mobile terminal checks its position periodically and performs location update whenever its distance exceeds the threshold distance measured from the predicted location. Upon a call arrival, the network pages the mobile terminal starting from the predicted location (which may be the one that performed the last update) and outward, in a shortest-distance-first order, until the mobile terminal is found.

For performance analysis, a Gauss-Markov process is used to model the user's mobility pattern. The Gauss-Markov model captures the correlation of the mobile's velocity in time, and can represent different user mobility patterns, including the random walk and constant velocity fluid-flow models. Numerical results, under the assumptions of an infinite one-dimensional linear model and Poisson call arrivals, show that the predictive distance-based scheme has better performance than the nonpredictive one.

State-Based

In the *state-based update scheme*, the mobile terminal determines whether to perform location update based on its current *state*. The state information can include the time elapsed or the number of cell crossings since the last update, the cell distance between the current and last registered locations, or some other criteria. Thus, maintaining different state information corresponds to different location update schemes.

In [18], the author analyzed the state-based scheme where the system state includes the current location and the time elapsed since the last update. A time-varying Gaussian process is used to model the user's movement. The suboptimal solution for the average cost of location update and paging under no paging

delay constraint is obtained by a greedy method. Results show that the state-based update scheme achieves 10 percent improvement in average cost compared to the timer-based scheme.

LeZi Update

The idea of the *LeZi update* (pronounced "lazy update") algorithm [19] is based on a compression algorithm proposed by Ziv and Lempel. The LeZi update algorithm can be considered to be a *path-based* update scheme in which *movement history* rather than current location is sent in an update message. The movement history consists of a list of *zone* (i.e., LA or cell) IDs the mobile terminal has crossed after the last update. The network database maintains the movement history in a compact form by a *trie* or *digital search tree*. This trie can be considered part of the user's profile. Upon a call arrival, selective paging based on the trie information is used to locate the mobile terminal.

Terminal Paging

Terminal paging is the process by which the network determines the exact location of a particular mobile terminal. In each *polling cycle* or *search iteration*, polling signals are sent over the downlink control channel to all cells where the mobile terminal is likely to be present. All the mobile terminals listen to the page message, and only the target mobile terminal sends a response message back over the uplink control channel. In each polling cycle, there is a *timeout* period. If the target mobile terminal replies before the timeout, the paging process is terminated. Otherwise, another group of cells is chosen in the next polling cycle.

To avoid call dropping, the mobile terminal must be located within an allowable *time constraint*. The *maximum paging delay* corresponds to the maximum number of polling cycles allowed to locate the mobile terminal. For example, if the maximum paging delay is equal to 1, the mobile terminal has to be located within a single search iteration.

Since radio bandwidth is consumed during the paging process, the paging cost is proportional to the number of polling cycles, as well as the number of cells being polled in each cycle. The paging area depends on the information provided by the location update function. The paging cost can be reduced by predicting the current location of the mobile terminal. In this section we summarize different paging strategies proposed in the literature. A classification of various paging strategies is shown in Fig. 2.

Blanket Polling

In *blanket polling*, all the cells within the LA in which the mobile terminal is located are polled simultaneously when a call arrives. Since the mobile terminal is located within the LA, its location can be determined within a single polling cycle. This paging strategy is currently deployed on top of the LA-based update scheme in existing PCNs. The major drawback of blanket polling is that since the number of cells within a typical LA is large, the paging cost is very high.

Shortest-Distance-First

In this paging strategy, the network pages the mobile terminal starting from the cell where the mobile terminal last updated its location, and moving outward in a shortest-distance-first order. The distance is measured in terms of the number of cells away from the last update location. If a threshold-based update scheme (e.g., distance or movement) is used, the paging or residing area of the mobile terminal is bounded. The mobile terminal can be located within a fixed number of polling cycles. The paging delay constraint can be incorporated by grouping cells at different distances for each polling cycle.
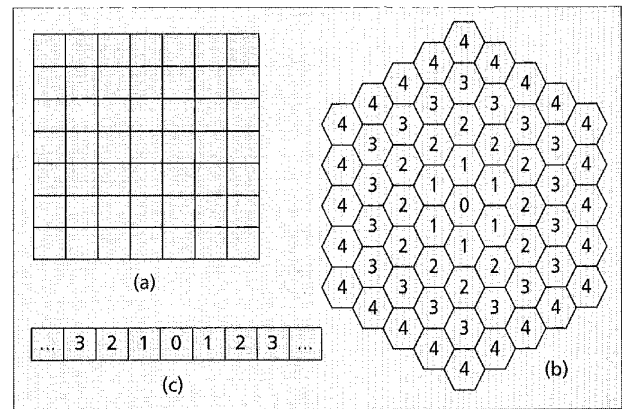

■ **Figure 3.** *Regular cell topology: (a) mesh; (b) hexagonal; (c) linear.*

To illustrate the mechanism of this paging strategy, consider Fig. 3b in which the cell topology is hexagonal. Suppose the cell labeled 0 is where the mobile terminal performed the last update. Assume that the distance-based update scheme is used with a distance threshold equal to 3; that is, location update is performed whenever the mobile terminal moves to any cell labeled 4.

With no paging delay constraint, the polling sequence is {0, 1, 2, 3}; that is, cell 0 is polled first. If no response is received from the mobile terminal after a timeout, all cells labeled 1 are polled in the next polling cycle. This continues until either the mobile terminal sends a response message to the base station or all the cells within the sequence list have been polled.

With delay constraints, cells with different labels or distances may be polled as a group in a polling cycle. For example, if the maximum paging delay is equal to 3, the paging sequence list can be {{0, 1}, {2}, {3}}. That is, cells with label 0 or 1 are polled in the first polling cycle. Cells labeled 2 and 3 are polled in the second and third polling cycles, respectively. Various location update algorithms such as distance and movement-based schemes have used this paging strategy to determine the location management cost for performance comparisons.

Sequential Paging Based on a User's Location Probability

In this paging strategy, the current location of the mobile terminal is predicted based on its location probability distribution. Polling signals are sent only to those cells in which the user is likely to be present. An intuitive result derived in [20] states that:

> "Given the probability distribution on user location, under no paging delay constraint, the paging cost is minimized by sequentially polling the cells in decreasing order of probability."

Clearly, uniform location distribution gives the highest paging cost and delay.

When there is a maximum paging delay constraint, a group of cells can be polled together in each polling cycle. Dynamic programming [21] can be used to determine the optimal group size to minimize paging cost. In [20], the authors obtained the optimal paging sequence resulting in minimum paging cost with average paging delay constraint. The sequential paging strategy has been used for performance analysis of the timer and state-based update schemes. There are a number of open issues related to the deployment of the sequential paging strategy:

- How do we *estimate* the time-varying location probabilities based on measurements?
- Under what conditions do we initiate an *update* of the location probabilities?
- How do we *store* these location probabilities in the network database *efficiently*?

	a.m.			p.m.			
Time	8:10	8:25	8:40	12:15	1:10	6:20	6:30	6:45
LA crossing	a→b	b→e	e→f	f→e	e→f	f→e	e→b	b→a
LA location	b	e	f	e	f	e	b	a
Residence time	15 min	15 min	3.6 hr	55 min	5.2 hr	10 min	15 min	1.4 hr

■ **Table 1.** *User movement between 8:00 a.m. and 8:00 p.m.*

Velocity Paging

The *velocity paging scheme* [22] aims to reduce paging cost by decreasing the size of the paging area. The goal is achieved by grouping users into different *velocity classes*, based on their velocity at the location update instant. When a call arrives, the paging area is dynamically generated based on the user's last registration time and the velocity class index. The velocity paging scheme can be deployed on top of other location update algorithms. To implement this paging strategy, information such as the mobile's last known location, velocity class index, and last registration time is required at the user's database profile.

When the velocity paging scheme is combined with the movement-based update algorithm, numerical results [22] indicate that this combined scheme may not always result in a reduction in cost from the LA-based update scheme with blanket polling. The authors determine the range of cell radius under which this combined scheme should be used based on system parameters.

Ensemble Polling

The paging delay may increase and the paging channels get overloaded if there are a large number of paging requests (i.e., incoming call arrivals). In [23], the authors formulated ensemble polling for mobile terminals as a Markov queuing problem. The paging requests that arrive at the network follow a Poisson distribution. The service rate is assumed to be exponentially distributed. Given the user location probability distribution, various paging strategies including blanket polling, sequential paging, and sequential group paging are analyzed. The average delay between a page request arrival and the determination of the user's location is calculated. Results indicate that blanket polling provides the lowest delay at small load, while sequential paging can sustain a higher paging request rate.

The ensemble polling problem is also investigated in [24]. There are a fixed number of paging channels in each base station. The paging requests follow a Poisson distribution. Sequential paging based on the theory of optimal search is used. Results indicate that this sequential paging strategy incurs a lower signaling load than the blanket polling approach.

Teletraffic Modeling

Teletraffic models are an invaluable tool for network planning and design. In this section we summarize various modeling techniques that have been used for the performance analysis of location update and paging algorithms. Particular attention will be given to the appropriateness of the modeling assumptions made in view of the intended objectives. This is crucial in order to clearly identify the usefulness of the modeling effort accomplished. Indeed, we recall that: "A theory has only the alternative of being right or wrong. A model has a third possibility: it might be right but irrelevant."[2]

---

[2] *From* A Selection of Scientific Quotations, *collected by A. L. Mackay;* M. Ebison, Editor.

Topology Model

The topology model specifies the connectivity between base stations or cells. Regular cell topologies are commonly used to model the coverage area of a cellular network. Mesh or hexagonal cell configurations are used in two-dimensional models, while a linear model is used in the one-dimensional case. Figure 3 shows some of these regular cell topologies. Although these models simplify analytical computation, they do not give an accurate representation of a realistic cellular network topology, where the sizes of cells depend on transmit power, receiver sensitivity, antenna radiation pattern, and propagation environment, and the number of neighboring cells varies from cell to cell.

The topology of a cellular network can be characterized by a graph model [7, 25]. Consider an LA-based cellular system. The interconnection of the LAs can be modeled as a connected graph $G = (N, E)$, where node set N represents the set of LAs and edge set E represents the connectivity between two neighboring LAs. For example, referring to Fig. 1, $N = \{a, b, c, d, e, f, g\}$ and $E = \{(a, b), (a, d), (b, c), (b, d), (b, e),...,(f, g)\}$.

The graph model can be applied to other levels in a hierarchical manner. At the cell level, the node set consists of the identifiers of the cells, and the edge set represents the connectivity of the neighboring cells.

The Residence Time Model

The *residence time* or *dwell time* at a location represents the amount of time the mobile user stays in that location before moving somewhere else. Certain location update schemes such as selective LA and profile-based update algorithms depend on good estimation of the residence time at different LAs. The optimal threshold values in threshold-based update schemes depend on cell residence time. However, it is somewhat interesting to note that the majority of research on location management assumes a geometric (or exponential) cell residence time distribution in performance analysis. The distribution is assumed to be independent and identically distributed (i.i.d.) for all cells. The major limitation of the i.i.d. geometric residence time assumption is that it does not capture an accurate representation of individual user mobility patterns, where a user may stay at certain locations (e.g., his home or office) for a relatively long period of time.

The model in [7] allows different residence times in different LAs. The only drawback is that the residence time has to be geometrically distributed. The model in [12] allows an i.i.d. general cell residence time distribution, but is restricted to a hexagonal cell configuration. Only the model in [25] allows an i.i.d. general cell residence time distribution in an arbitrary cell topology.

The estimation of residence time at the LA and cell levels is also important. To estimate residence time between LAs for a particular user, his/her movements throughout the day must be observed over a long time interval. As an example, consider the movements of a particular user within the coverage area shown in Fig. 1a. Table 1 shows the time when LA boundaries are crossed and the residence time at various LAs. A simple way to estimate the residence times at different LAs is by averaging those values over a long time period (e.g., weeks or months).

Estimation of cell residence time is more complicated. Individual mobile terminals have to collect the IDs of the cells travelled as well as the time of those cell crossings. This information will then be sent to the network for processing.

Mobility Models

Mobility models play an important role in examining different issues in wireless networks, including resource allocation, handoff, and location management. In general, the mobility models depend on the speed, direction, or movement history

of the mobile users [26]. Here, we summarize different mobility models used to describe aggregate and individual user movement behavior. A classification of various mobility models is shown in Fig. 4.

■ Figure 4. *The classification of different mobility models.*

The Fluid Flow Model — This model characterizes aggregate movement behavior as the flow of a fluid. Mobile users are assumed to move at an average velocity of v, and their direction of movement is uniformly distributed over $[0, 2\pi]$. Assuming that the mobile users are uniformly populated with a density of ρ and the location area boundary is of length L, the rate of users moving out of LA C is given by

$$C = \frac{\rho v L}{\pi}. \tag{1}$$

The above model is accurate for boundary crossing rate in a symmetric grid of streets (i.e., Manhattan-style). This model has been used to study the profile-based update scheme. One of the limitations of this model is that it describes aggregate traffic and is difficult to apply to scenarios where individual movement patterns are desired.

The Gravity Model — This model has been used extensively in the area of transportation research to model human movement behavior. In this model, the amount of traffic $T_{i,j}$ moving from region i to region j is described by
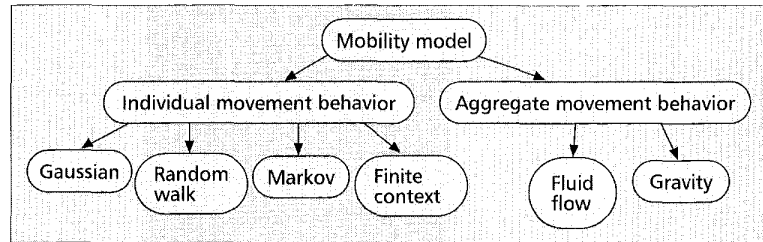
$$T_{i,j} = K_{i,j} \cdot P_i \cdot P_j, \tag{2}$$

where P_i is the population in region i, and $\{K_{i,j}\}$ are parameters that have to be determined for all possible region pairs (i, j). Variations of the gravity model have been used in [27] to describe the *national* and *international mobility models*. The national mobility model characterizes aggregate movement behavior between the 10 largest metropolitan areas in the United States. The international mobility model characterizes aggregate movement behavior between the United States and 10 other countries. Both models are constructed based on air passenger traffic data.

Time-varying Gaussian User Location Distribution — This arises as a result of isotropic random user motion with *drift*, defined as the mean velocity in a given direction; it can be used to model directed traffic such as vehicles along a highway. The probability distribution function in a one-dimensional model is given by

$$f(x,t) = \frac{1}{\sqrt{\pi D t}} e^{\frac{-(x-vt)^2}{Dt}}, \tag{3}$$

where x is the location variable, t is the time elapsed since last contact with the mobile terminal, v is the mean drift velocity, and D is a constant. When the time is partitioned into small intervals and $v = 0$, the above distribution function approximately corresponds to the symmetric random walk model [28]. Gaussian user location distribution has been used to study the timer and state-based update schemes.

The Symmetric Random Walk Model — This model is most commonly used to describe individual movement behavior. When the mobile user leaves a cell, there is an equal probability that he/she will move to any neighboring cell. In general, this probability is equal to the reciprocal of the number of neighboring cells. For a one-dimensional linear model, the probability is 1/2. For a two-dimensional hexagonal model, the probability is 1/6. This model has been used for the performance analysis of vari-

ous update/paging algorithms [15, 16]. Although the random walk model simplifies the analysis, one of its limitations is that the direction of the mobile user is not taken into account.

The Markov Model — This attempts to capture the direction of a user's movement pattern by assigning different probabilities to different neighboring cells. Suppose a mobile user is located in cell i. In each time interval, the user will either remain in the cell with probability $P_r(i|i)$ or move to a neighboring cell j with probability $P_r(j|i)$. Thus, the user's preference is characterized by the probability function. The cell residence time follows a geometric distribution. This model has been used for performance analysis of the selective LA and threshold-based [10] update schemes. One of the limitations of this approach is that there is no concept of movement history or trip of a particular mobile user.

The Markov Model with Movement History — In this model, when a mobile terminal leaves a cell, the probability that it will move to a particular neighboring cell is based on the set of cells through which the mobile terminal has traveled. The model incorporates the concept of a trip in which the mobile user may follow a particular path to a destination.

The same idea has also been applied to the *finite context model* [19] in which the probability that a user is in a particular *zone* (i.e., cell or LA) can be predicted by a history of zones reported in the previous updates. Entropy is used to measure the location uncertainty of the mobility model. The rule of thumb is *the lower the uncertainty of the model, the richer the model.* Results based on the comparison on entropy show that the finite context model is more information-rich than the Markov and random walk models.

Mobility Traces — Mobility traces which record actual movement behavior for certain segments of the population, are certainly more realistic than mobility models. They indicate the aggregate movement behavior and can be used to verify assumptions used in the mobility models. However, publicly available mobility traces for a PCN are hard to come by.

In [30], the authors analyze a seven-week trace of the Metricom metropolitan-area packet radio wireless network. The findings indicate that the locations mobile users frequently visit are close to each other. In addition, the users' movement pattern is Gaussian distributed around the radius of the network. Although these results are only valid for this particular network, lessons can be learned about the techniques in gathering and analyzing the data.

The Stanford University mobile activity traces (also known as SUMATRA [29]) is a publicly available trace generator which includes simulation data of the Bay Area location information traces for voice calls. Future work on location management can use these traces for performance comparisons.

The Call Model

Poisson distribution is commonly used to model incoming call arrivals in PCNs; that is, the time between call arrivals follows an exponential distribution. Although the Poisson assumption is true for aggregate call arrivals in telephone networks, the

call arrival rate for an individual user may not be Poisson distributed. It may depend on time of day and day of the week. Since the threshold values of different threshold-based update algorithms depend on the individual call arrival rate, a realistic time-varying call model should be used in order to achieve better results. The time-varying call model can be constructed based on the call arrival data in the user's billing record.

Conclusions

We conclude this article by stating a number of open issues that need to be resolved for deployment of next-generation wireless PCNs.

User Profile — Several location update and paging strategies improve network performance by predicting the user's location based on his/her velocity, probability distribution, and so on. This information has to be stored in the user profile. An efficient way to collect, store, update, and disseminate user profile information is crucial for the deployment of these update and paging schemes.

Class-Based Schemes — Although the proposed per-user-based update and paging schemes give better performance than the conventional LA-based update scheme with blanket polling strategy, their implementation is more complex. To facilitate the deployment of these dynamic schemes, it may be useful to partition users into several classes or groups based on their call and mobility patterns. Future work should focus on per-class-based location update and paging schemes which are scalable and simple to implement.

Performance Comparison Tools — Although various location update and paging schemes have been proposed in the literature, the relative performance of these schemes is not clear (e.g., predictive distance-based vs. adaptive timer-based). This is mainly due to the fact that different models and assumptions are used to evaluate different schemes. Since it is impractical to develop testbeds or field trials simply for performance comparisons, it is desirable to have a generic analytical or simulation model that can analyze various update and paging schemes in a realistic environment.

Heterogeneous Networks — As part of the IMT-2000 system, the mobile terminal will be able to roam among different networks (e.g., between PCS and satellite, or between WATM and Mobile IP). Careful design is necessary in order to reduce the signaling and processing overhead for location update and call delivery. An efficient way to disseminate and update user information between heterogeneous networks remains an open issue.

References

[1] IEEE Pers. Commun., Special Issue on IMT-2000: Standards Efforts of the ITU, vol. 4, no. 4, Aug. 1997.
[2] Y. Bejerano and I. Cidon, "An Efficient Mobility Management Strategy for Personal Communication Systems," Proc. ACM/IEEE MobiCom'98, Dallas, TX, Oct. 1998, pp. 215–22.
[3] G. Krishnamurthi, M. Azizoglu, and A. Somani, "Optimal Location Management Algorithms for Mobile Networks," Proc. ACM/IEEE MobiCom'98, Dallas, TX, Oct. 1998, pp. 223–32.
[4] J. Jannink et al., "Efficient and Flexible Location Management Techniques for Wireless Communication Systems," ACM/Baltzer J. Wireless Networks, vol. 3, Oct. 1997, pp. 361–74.
[5] S. Tabbane, "Location Management Methods for Third-Generation Mobile Systems," IEEE Commun. Mag., vol. 35, no. 8, Aug. 1997, pp. 72–84.
[6] I. F. Akyildiz et al., "Mobility Management in Next-Generation Wireless Systems," Proc. IEEE, vol. 87, no. 8, Aug. 1999, pp. 1347–84.
[7] S. K. Sen, A. Bhattacharya, and S. K. Das, "A Selective Location Update Strategy for PCS Users," ACM/Baltzer J. Wireless Networks, vol. 5, no. 5, Sept.1999, pp. 313–26.
[8] G. P. Pollini and C.-L. I, "A Profile-Based Location Strategy and Its Performance," IEEE JSAC, vol. 15, no, 8, Oct. 1997, pp. 1415–24.
[9] S. Tabbane, "An Alternative Strategy for Location Tracking," IEEE JSAC, vol. 13, no. 5, 1995, pp. 880–92.
[10] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile Users: To Update or Not to Update?" ACM/Baltzer J. Wireless Networks, vol. 1, no. 2, July 1995, pp. 175–95.
[11] Z. Naor and H. Levy, "Cell Identification Codes for Tracking Mobile Users," Proc. IEEE INFOCOM '99, New York, NY, Mar. 1999.
[12] I. F. Akyildiz, J. Ho, and Y.-B. Lin, "Movement-Based Location Update and Selective Paging for PCS Networks," IEEE/ACM Trans. Net., vol. 4, no. 4, Aug. 1996, pp. 629–38.
[13] C. Rose, "Minimizing the Average Cost of Paging and Registration: A Timer-Based Method," ACM/Baltzer J. Wireless Networks, vol. 2, no. 2, June 1996, pp. 109–16.
[14] Z. Naor and H. Levy, "Minimizing the Wireless Cost of Tracking Mobile Users: An Adaptive Threshold Scheme," Proc. IEEE INFOCOM '98, San Francisco, CA, Mar./Apr. 1998, pp. 720–27.
[15] U. Madhow, M. Honig, and K. Steiglitz, "Optimization of Wireless Resources for Personal Communications Mobility Tracking," IEEE/ACM Trans. Net., vol. 3, no. 4, Dec. 1995, pp. 698–707.
[16] J. Ho and I. F. Akyildiz, "Mobile User Location Update and Paging under Delay Constraints," ACM/Baltzer J. Wireless Networks, vol. 1, no. 4, Dec. 1995, pp. 413–25.
[17] B. Liang and Z. Haas, "Predictive Distance-based Mobility Management for PCS Networks," Proc. IEEE INFOCOM '99, New York, NY, Mar. 1999.
[18] C. Rose, "State-Based Paging/Registration: A Greedy Technique," IEEE Trans. Vehic. Tech., vol. 48, no. 1, Jan. 1999, pp. 166–73.
[19] A. Bhattacharya and S. K. Das, "LeZi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks," Proc. ACM/IEEE MobiCom '99, Seattle, WA, Aug. 1999, pp. 1–12.
[20] C. Rose and R. Yates, "Minimizing the Average Cost of Paging under Delay Constraints," ACM/Baltzer J. Wireless Networks, vol. 1, no. 2, July 1995, pp. 211–19.
[21] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley, 1994.
[22] G. Wan and E. Lin, "A Dynamic Paging Scheme for Wireless Communication Systems," Proc. ACM/IEEE MobiCom '97, Budapest, Hungary, 1997, pp. 195–203.
[23] C. Rose and R. Yates, "Ensemble Polling Strategies for Increased Paging Capacity in Mobile Communications Networks," ACM/Baltzer J. Wireless Networks, vol. 3, no. 2, May 1997, pp. 159–67.
[24] R. Rezaiifar and A. Makowski, "From Optimal Search Theory to Sequential Paging in Cellular Networks," IEEE JSAC, vol. 15, no, 7, Sept. 1997, pp. 1253–64.
[25] W. S. V. Wong, "Stochastic Control of Inter-Switch Handoff and Location Update in Wireless Cellular Networks," Ph.D. thesis, Univ. of BC, Canada, Apr. 2000.
[26] M. M. Zonoozi and P. Dassanayake, "User Mobility Modeling and Characterization of Mobility Patterns," IEEE JSAC, vol. 15, no. 7, Sept. 1997, pp. 1239–52.
[27] D. Lam, D. Cox, and J. Widom, "Teletraffic Modeling for Personal Communications Services," IEEE Commun. Mag., vol. 35, no. 2, Feb. 1997, pp. 79–87.
[28] C. Rose and R. Yates, "Location Uncertainty in Mobile Networks: A Theoretical Framework," IEEE Commun. Mag., vol. 35, no. 2, Feb. 1997, pp. 94–101.
[29] The Sumatra homepage: http://www-db.stanford.edu/sumatra
[30] D. Tang and M. Baker, "Analysis of a Metropolitan-Area Wireless Network," Proc. ACM/IEEE MobiCom'99, Seattle, WA, Aug. 1999, pp. 13–23.

Biographies

VINCENT W.-S. WONG [M] (vincent.wong@ieee.org) received his B.Sc. degree from the University of Manitoba in 1994, M.A.Sc. degree from the University of Waterloo in 1996, and Ph.D. degree from the University of British Columbia in 2000, all in electrical engineering. Currently, he is working at PMC-Sierra, Inc. in Burnaby, British Columbia, Canada. His research interests include location and handoff management in wireless cellular networks, IP over DWDM, and OAM management in optical transport networks.

VICTOR C. M. LEUNG [SM] (vleung@ieee.org) received a B.A.Sc. (Hons.) degree in electrical engineering from the Univ. of British Columbia (U.B.C.) in 1977, and was awarded the APEBC Gold Medal as head of the graduating class in the Faculty of Applied Science. He attended graduate school at U.B.C. on a Natural Sciences and Engineering Research Council Postgraduate Scholarship and obtained a Ph.D. degree in electrical engineering in 1981. From 1981 to 1987 he was a Senior Member of Technical Staff at MPR Teltech Ltd., specializing in the planning, design, and analysis of satellite communication systems. He also held a part-time position as visiting assistant professor at Simon Fraser University in 1986 and 1987. In 1988 he was a lecturer in the Department of Electronics at the Chinese Univ. of Hong Kong. He joined the Department of Electrical and Computer Engineering at U.B.C. in 1989, where he is a professor and a member of the Centre for Integrated Computer Systems Research. In January 2000 he was appointed to the TELUS Mobility Research Chair in Advanced Telecommunications Engineering at U.B.C. He is also a project leader in the Canadian Institute for Telecommunications Research, a Network of Centres of Excellence funded by the Canadian Government. His research interests are in the areas of architectural and protocol design and performance analysis for computer and telecommunication networks, with applications in satellite, mobile, personal communications and high speed networks.