Performance Evaluation of Forwarding Strategies for Location Management in Mobile Networks

ING-RAY CHEN[1], TSONG-MIN CHEN[2] AND CHIANG LEE[2]

[1]*Department of Computer Science, Virginia Tech, Northern Virginia Center, 7054 Haycock Road,*
Falls Church, VA 22043, USA
[2]*Department of Computer Science and Information Engineering, National Cheng Kung University,*
Tainan, Taiwan
Email: irchen@cs.vt.edu

This paper presents a methodology for evaluating the performance of forwarding strategies for location management in a personal communication services (PCS) mobile network. A forwarding strategy in the PCS network can be implemented by two mechanisms: a forwarding operation which follows a chain of databases to locate a mobile user and a resetting operation which updates the databases in the chain so that the current location of a mobile user can be known directly without having to follow a chain of databases. In this paper, we consider the PCS network as a server whose function is to provide services to the mobile user for 'updating the location of the user as the user moves across a database boundary' and 'locating the mobile user'. We use a Markov chain to describe the behavior of the mobile user and analyze the best time when forwarding and resetting should be performed in order to optimize the service rate of the PCS network. We demonstrate the applicability of our approach with hexagonal and mesh coverage models for the PCS network and provide a physical interpretation of the result.

Received June 6, 1997; revised June 22, 1998

1. INTRODUCTION

In a personal communication services (PCS) network, it is possible for a mobile user to be called by other mobile users or to call other mobile users, as it moves. In recent years, various location management strategies based on the concept of forwarding have been proposed and studied in the literature [1, 2] with the goal of minimizing the network signaling and database loads. The basic mechanisms involved in implementing these forwarding strategies are 'forwarding' and 'resetting' operations. The 'forwarding' operation means that when a mobile user moves across a database boundary (e.g. a registration area boundary [1] or a base station boundary [2]), only a pointer is set up between the two involved databases, while the 'resetting' operation means that all databases along the forwarding chain for locating a mobile user are updated all at once so that after the update the current location of the mobile user is known directly by consulting a single database instead of a chain of databases. Naturally, if resettings are done frequently, say, on a per move basis, then the cost of locating a mobile user per call would be small since only a single database needs to be consulted to know the location of the called mobile user; however, the network signaling cost due to frequently updating all involved databases would be large. Conversely, if resettings are done infrequently then the cost of locating a user per call would be large since the system has to follow a potentially long chain of databases to locate the called mobile user. Of course, the updating cost for keeping the location information of the called mobile user up-to-date in this case would be small because reset operations are performed only once in a while.

The main research issue in previous studies is in analyzing 'how often' the system should perform a reset in response to a move made by a mobile user crossing a database boundary, so that a specified 'cost' measure may be minimized, essentially by trading-off the costs involved in locating users and updating user location information [3, 4]. Various cost measures have been proposed in previous studies and have resulted in different solutions to the research issue. Rao *et al.* [2] considered a per-user cost measure expressed in terms of the 'average cost' of a call, assuming that after every *k* forwarding steps a reset will be performed. The average cost of a call in their cost model includes two components: a signaling cost term for setting-up the pointers and updating the database changes due to movements of the mobile user, and a service cost term for all the nodes (e.g. switches) along the forwarding chain to allocate resources to service the call. It is not clear why, after the location of the called mobile user is found, all the nodes along the forwarding chain need to allocate resources to service the call, since presumably only the nodes (switches) connecting the calling and called parties need to do so. It is also not clear how their cost measure can be extended to a more general case in which multiple calls may be placed simultaneously for the same mobile user. Another related work is by Bar-Noy and Kessler [5] who considered a cost measure in terms of the expected number of forwarding steps required to locate a mobile user for a given update rate. Their analysis, however, is mainly for comparing the performance of various dynamic updating strategies in a ring cellular topology without using the concept of forwarding. Thus, a search for locating a mobile user through a chain of forwarding pointers was not considered.

In [1], Jain *et al.* proposed a per-user cost measure to characterize the benefit of forwarding over non-forwarding mechanisms. The cost measure considers the possibility of multiple calls and is defined as the ratio of 'the cost of maintaining a mobile user's location and locating the mobile user between two consecutive database crossings assuming that after every k forwarding steps a reset will be performed' to the same quantity with $k = 1$. A database unit in their study corresponds to a registration area (RA) in an IS-41 [6] or a GSM [7] standard. By using the common channel signaling (CCS) network with a Signaling System No. 7 (SS7) protocol as a case study, they discovered that, when k is fixed to a certain value, the forwarding mechanism will be beneficial only to mobile users with their call-to-mobility ratio (CMR) smaller than a certain threshold value. The contribution of their work was that they provided a framework for applying the forwarding mechanism in existing standards such as IS-41 and GSM, and also demonstrated the advantage of forwarding over non-forwarding mechanisms in the CCS-SS7 network, particularly for some asymptotic cases in which certain cost terms are dominating. However, they did not address the issue of how to determine the optimal k value when the system is given the CMR of a mobile user. Moreover, when estimating the cost of locating a mobile user in their analysis, they made the simplifying assumption that all calls travel through one half of the databases in the forwarding chain before the address of the called mobile user is found. This simplifying assumption may make their analysis less trustworthy.

In this paper, we develop a Markov model to describe the behavior of a mobile user as it moves while being called by other mobile users in a PCS network. We assign 'rewards' to the states of the Markov chain and then calculate the probabilistic 'average' reward. By utilizing the Markov chain, we formulate the problem of finding the best time to perform resetting as an optimization problem, i.e. finding the optimal value of k under which the average 'reward' representing the specified cost measure is optimized.

The rest of the paper is organized as follows. Section 2 gives a more detailed description of the system model. Section 3 describes the stochastic analysis model. We view the PCS network as the server and the operations for which the server must provide services to a mobile user include 'updating the location of the user as the user moves across a database boundary' and 'locating the user'. We discuss several ways of assigning 'rewards' to the states of a mobile user so as to yield different 'cost measures' which we are trying to optimize, and show that different cost measures can result in different optimal k values. Section 4 discusses two coverage models for constructing the PCS network and demonstrates the applicability of our model. The first case considers user movements under the hexagonal coverage model, while the second case considers user movements under the mesh coverage model. In each case, we show how our model can be parametrized for the resulting PCS network and how the optimal k value based on the cost measure assigned may be determined. Finally, Section 5 summarizes the paper and outlines some possible future research areas.

2. SYSTEM DESCRIPTION

We first state the system model and assumptions. We assume that a signaling network consisting of possibly several levels of databases is used for locating mobile users and setting up calls. We differentiate high-level databases from low-level databases. For the purpose of our analysis, we are particularly interested in two adjacent database levels which may use the resetting and forwarding technique for tracking mobile users. We shall call the high-level database the 'home' database. The home database will direct an incoming call to the first low-level database unit on the forwarding chain. We shall call the databases on the forwarding chain at the low level 'visitor' databases. To allow us to easily distinguish the visitor databases at the lower level, we shall call the first one on the chain v_0 and the last one v_i for a forwarding chain with a length of i. When a call is placed, the system will first go to the home database and then follow the visitor databases along the forwarding chain to locate the mobile user. We assume that a mobile user can cross visitor database boundaries freely as it is being called. The time that a particular mobile user stays within a visitor database before moving to another one is characterized by an exponential distribution with an average rate of σ. Such a parameter can be estimated using the approach described in [1, 8] on a per-user basis. The interarrival time between two consecutive calls to a particular mobile user, regardless of the current location of the user, is also assumed to be exponentially distributed with an average rate of λ. A mobile user is thus characterized by its CMR, defined as λ/σ.

We assume that the forwarding chain is not reset when a mobile user is called. We also assume that all calls to a particular mobile user will go to the current home database. Consequently, if there are two or more requests waiting at the home database to locate a mobile user, the home database can locate the called mobile user and then return the location of the called mobile user to all pending requests simultaneously. Another assumption is that when a mobile user moves across a visitor database boundary, a pointer between the two involved visitor databases will be set up before the mobile user can possibly move across another visitor database boundary. This assumption implies that the time to set up a pointer is much shorter than the dwell time within a visitor database, so that the time to set up a pointer is so small compared with the dwell time that it will not affect the distribution of the dwell time. Of course, during the time period in which the pointer connection is set up, call requests can still arrive.

Figure 1 shows a PCS network as discussed in [9]. Here, the home location register (HLR) and visitor location registers (VLRs) contain databases for tracking mobile users, but the intermediate switches such as regional signal transfer points (RSTPs) and local signal transfer points (LSTPs) are only used for connecting the HLR with VLRs. Separate RSTPs are connected by a public switched telephone network (PSTN). The round trip average communication cost between a VLR and the HLR is represented by T and the round trip average communication cost between two neighboring VLRs is represented by τ.

When applying the forwarding and resetting technique to this network structure, forwarding pointers can be set up among VLRs. Consequently, the HLR corresponds to the 'home' database while a VLR corresponds to a 'visitor' database.

The question we are interested in solving is to find out the best k value (the length of the forwarding chain) under which a cost measure specified in terms of a 'reward' metric can be optimized. We solve the problem by first developing a generic Markov model that describes the behavior of a mobile user subject to the forwarding and resetting technique. Then, we apply the Markov model to two different coverage models (hexagonal and mesh) by parametrizing (giving values to) the model parameters based on the specific structures under consideration so that there is no need to modify the Markov model or the solution technique.

3. STOCHASTIC ANALYSIS MODEL

Notation

λ: the arrival rate of calls to a particular mobile user.

σ_f: the mobility rate of a particular mobile user which moves to a new VLR.

σ_b: the mobility rate of a particular mobile user which returns to the last-visited VLR.

σ: the mobility rate of a particular mobile user; $\sigma = \sigma_f + \sigma_b$.

CMR: the call-to-mobility ratio of a particular mobile user, i.e. λ/σ.

μ_p: the execution rate to set up, delete, or travel a pointer between two visitor databases.

μ_i: the execution rate to find v_i for a forwarding chain with i pointers.

k: the number of forwarding steps after which a reset operation is performed.

m_k: the execution rate to reset a forwarding chain with k pointers so that, after the reset operation is performed, v_k becomes v_0 and the home database as well as all v_j, $0 \leq j \leq k$, are informed of the change.

P_j: the probability that the system is in a particular state in equilibrium.

X: the throughput of the signaling network in servicing 'updating the location of a mobile user as the user moves across a database boundary' and 'locating a mobile user'.

The state of a mobile user as it crosses database boundaries while being called can be described by two state components: (i) the number of forwarding steps; and (ii) a binary quantity indicating whether or not it is in the state of being called. Figure 2 shows a Markov model describing the behavior of a mobile user wherein a state is represented by (a, b) where a is either I (standing for IDLE) or C (standing for CALLED), while the other component b indicates the number of forwarding steps that have been made since the last reset operation. Initially, the mobile user is in the state $(I, 0)$, meaning that it is not being called and the number of forwarding steps is zero. Below, we explain briefly how we construct the Markov model.

1. If the mobile user is in the state (I, i) and a call arrives, then the new state is (C, i) in which the number of forwarding steps remains at i but the mobile user is now in the state of being called. This behavior is modeled by the (downward) transition from state (I, i) to state (C, i), $0 \leq i < k$, with a transition rate of λ.
2. If the mobile user is in the state (C, i) and another call arrives, then the mobile user will remain at the same state, since the mobile user remains in the state of being called and the number of forwarding steps also remains at i. This behavior is described by a hidden transition from state (C, i) back to itself with a transition rate of λ. This type of transition is not shown in Figure 2 since it does not need to be considered when solving a Markov chain [10]. Note that this implies that in state (C, i) the number of requests accumulated to locate the mobile user may be greater than one.
3. If the mobile user is in the state (C, i), the signaling network can service all pending calls simultaneously with a service rate of μ_i. After the service, the new state is (I, i) since all calls have been serviced while the number of forwarding steps remains at i. We use the subscript 'i' in μ_i to refer to the service rate of the signaling network to locate a mobile user when the number of forwarding steps is i. It should be emphasized that all pending calls are serviced at once with this service rate. This behavior is described by the (upward) state transition from state (C, i) to state (I, i) with a transition rate of μ_i.
4. Regardless of whether the mobile user is in the state of being idle or having been called, if the mobile user moves across a visitor database boundary to a new VLR, then the new visitor database, i.e. v_{i+1}, must determine if a pointer connection or a reset operation has to be performed. This behavior is modeled by a transition from state (I, i) to state $(I, i + 1)^*$ (if the mobile user is idle) or from state (C, i) to state $(C, i + 1)^*$ (if the mobile user is in the state of being called), with a mobility rate of σ_f. If the mobile user moves back to the last-visited VLR, i.e. $v_{i+1} = v_{i-1}$,

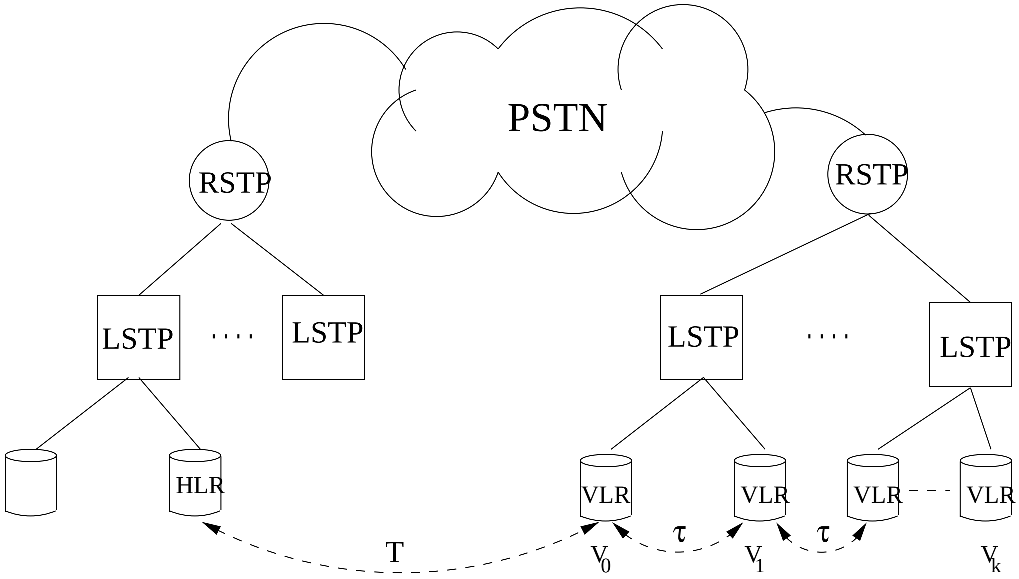

FIGURE 1. Forwarding and resetting technique applying to the PCS network.

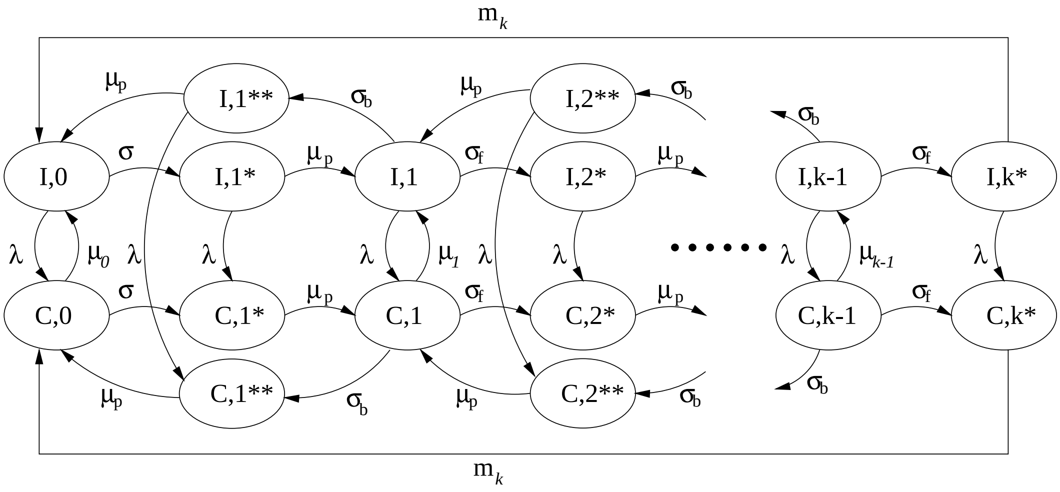

FIGURE 2. A Markov model for describing forwarding strategies.

then this behavior can be modeled by a transition from state (I, i) to state $(I, i)^{**}$ (if the mobile user is idle) or from state (C, i) to state $(C, i)^{**}$ (if the mobile user is in the state of being called), with a mobility rate of σ_b. Note that here we assume that the movement 'cycle', if one ever exists, involves only two VLRs.

The subsequent action performed by v_{i+1} depends on whether or not $v_{i+1} = v_{i-1}$ and, if not, whether or not the number of forwarding steps has reached k.

(a) If v_{i+1} is the last-visited VLR, i.e. $v_{i+1} = v_{i-1}$, then the pointer connection between v_i and v_{i-1} is deleted and the length of the forwarding chain is reduced to $i - 1$. This behavior is modeled by the (horizontal, left) transition from state $(I, i)^{**}$ to state $(I, i - 1)$ (if the mobile user is idle) or from state $(C, i)^{**}$ to state $(C, i - 1)$ (if the mobile user is in the state of being called), with a rate of μ_p.

(b) If v_{i+1} is not the last-visited VLR and $0 \le i < k - 1$, then the new visitor database v_{i+1} simply sets up a pointer connection between v_i and v_{i+1}. This behavior is modeled by the (horizontal, right) transition from state $(I, i + 1)^*$ to state $(I, i + 1)$ (if the mobile user is idle) or from state $(C, i + 1)^*$ to state $(C, i + 1)$ (if the mobile user is in the state of being called), with a rate of μ_p. Note that we assume that when a mobile user moves across a visitor database boundary, it will make sure that the pointer connection is established properly between the two involved databases before it can cross another visitor database boundary.

(c) If v_{i+1} is not the last-visited VLR and $i = k - 1$, then the new visitor database v_{i+1} knows that the mobile unit has made k boundary moves. Therefore, a reset operation will be invoked. This behavior is modeled by the (wrap-around) transition from state $(I,k)^*$ to state $(I,0)$ (if the mobile user is idle) or from state $(C,k)^*$ to state $(C,0)$ (if the mobile user is in the state of being called), with a rate of m_k. The subscript 'k' in m_k refers to the fact that the reset cost depends on the magnitude of k.

Note that all competing events in a state can occur concurrently. For example, in state $(C,1)$ there are actually three competing events which can occur concurrently, namely: (i) another call arrival which makes the system stay at the same state; (ii) a move by the mobile user crossing a visitor database boundary which makes the system transit to state $(C,2)^*$ or $(C,1)^{**}$; and (iii) a service completion of all calls which makes the system transit to state $(I,1)$. The possibility of the system moving from a given state to a neighboring state depends on the relative magnitude of the transition rates of the corresponding competing events; only one state transition is possible at a time.

The Markov chain shown in Figure 2 is ergodic [10], which means that all states have a non-zero probability. The probability that the system is found in a particular state in equilibrium also depends on the relative magnitude of the outgoing and incoming transition rates. For example, if the call arrival rate λ is much greater than the mobility rate σ, then the probability that the system is found to stay in state (C,i) would be much greater than in states $(I,i+1)^*$ and $(I,i)^{**}$ since it is more likely for state (I,i) to make a transition into state (C,i) than into states $(I,i+1)^*$ and $(I,i)^{**}$. Since the probability that the system stays in a particular state depends on the relative magnitude of the transition rates, it implies that the best value of k for performing a reset operation can vary on a case-by-case basis, as it also depends on the relative magnitude of the transition rates.

One way to determine the best k value is to view the signaling network as the server and the operations for which the server must provide services to a mobile user include 'updating the location of the user as the user moves across a database boundary' and 'locating the user'. In this view, a natural performance measure which we can maximize is the 'throughput' of the signaling network with respect to servicing the above two types of operations. Now, considering the Markov model in Figure 2, we observe that not all states contribute to this performance metric. In other words, when the mobile user is neither being called nor moving across a database boundary, it does not require the service of the signaling network. Therefore, when calculating the throughput of the system in servicing a mobile user, these idle states must be excluded. Specifically, states (I,i), $0 \le i \le k-1$, in Figure 2 are to be excluded. Let X represent the average throughput of the server in servicing the above two types of operations, denoting the performance metric we attempt to maximize. Also let P_j represent the probability (fraction of the time) that the system is found to be staying at state j in equilibrium. Then,

$$\begin{aligned} X = {} & \left(\sum_{i=0}^{k-1} P_{(C,i)} \times \mu_i \right) \\ & + \left(\sum_{i=1}^{k-1} (P_{(I,i)^*} + P_{(I,i)^{**}} + P_{(C,i)^*} + P_{(C,i)^{**}}) \times \mu_p \right) \\ & + (P_{(I,k)^*} + P_{(C,k)^*}) \times m_k. \end{aligned} \tag{1}$$

Here the first term represents the throughput of the signaling system in 'locating the user', while the second and third terms represent the throughput of the signaling network in 'updating the location of the user as the user moves across a database boundary'. Note that the second term accounts for the pointer connection operation between two involved visitor databases, while the third term accounts for the reset operation which is performed upon every k moves.

Equation (1) above yields the average effective throughput of the signaling network as a function of k. For a given set of parameter values, we can first compute the values of P_j for all states and then use Equation (1) to determine the best value of k that maximizes X. Of course, different signaling network structures may give different parameter values and thus may yield different optimal k values. When applying the Markov model developed in this section, we have to parametrize it based on the specific characteristics of the signaling network under consideration. In this paper, we use a software package called SHARPE [11] to solve P_js and subsequently compute X for a parametrized Markov model.

4. PERFORMANCE ANALYSIS

In this section, we first discuss the parametrization process, i.e. how to estimate values for the parameters of the Markov model in Figure 2. Then, we illustrate how this parametrization process can be applied to two separate PCS networks constructed by the hexagonal and mesh coverage models. Our goal is to determine the optimal k under different environment settings for such PCS networks.

4.1. Parametrization

Assume that the communication time (round trip) between the HLR at the higher level and a VLR at the lower level is exponentially distributed with an average time of T. Also assume that the communication time (round trip) between two neighboring VLRs at the lower level is exponentially distributed with an average time of τ (see Figure 1). Furthermore, assume that when performing a reset operation, v_k first communicates with the HLR to indicate that it is now the new v_0 for the called mobile user. The HLR confirms the change by replying a message to v_k and also sends a cancellation message to the old v_0, which in turn initiates a propagation of the cancellation message along the forwarding chain to all base stations to delete all obsolete pointers and also to accumulate the service cost (e.g. credits) information. After the last step is done, v_{k-1} informs the HLR to confirm the completion

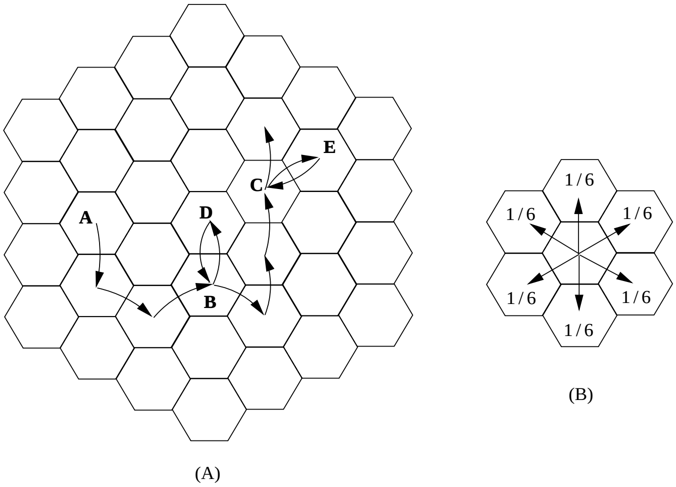

FIGURE 3. Hexagonal coverage model.

of the reset operation. As we can see, a reset operation implemented in this way involves twice the amount of VLR–HLR communication time needed for locating a user.

Given the above assumptions, we can parametrize the Markov chain as follows:

1. The execution rate to set up, delete or travel a pointer connection between two visitor databases is given by

$$\mu_p = \frac{1}{\tau}. \tag{2}$$

2. The execution rate to find v_i and then to locate the mobile user can be estimated as

$$\mu_i = \frac{1}{T + i\tau}. \tag{3}$$

3. The execution rate to perform a reset operation can be estimated as

$$m_k = \frac{1}{2T + k\tau}. \tag{4}$$

Here we note that $2T$ is used for estimating m_k because the VLRs at the lower level have to communicate with the HLR at the higher level twice.

4.2. Hexagonal coverage model

When modeling a mobile PCS network, it is frequently assumed that the coverage area of the PCS network can be divided into cells of the same size. In the hexagonal coverage model [8], cells are assumed to be hexagonally shaped, with each cell having six neighbors (see Figure 3A). In this section, we illustrate how we can use the Markov model developed in this paper to determine the optimal number of forwarding steps before performing a reset for the hexagonal coverage model. Figure 3A shows an example of a mobile user movement. The mobile user is initially registered at (visitor database) A. Suppose that the optimal number of forwarding steps is three. Then the mobile user will perform a reset operation as it visits B, since it has made three moves across the visitor database boundaries since the last reset operation was performed. When the mobile user subsequently moves from B to D, the number of forwarding steps is updated to one. When the mobile user later returns from D to B, however, the number of forwarding steps becomes zero again since B is the last visited VLR. In this case, the forwarding pointer from D to B is not established and the stale forwarding pointer from B to D is deleted.

In this section, we shall parametrize our model based on the optimal hexagonal structure determined in [8] and the PCS network shown in Figure 1 as discussed in [9]. Our result enhances the previous results in [8, 9] in that we allow the optimal number of forwarding steps to be determined if the forwarding and resetting technique is to be used in tracking mobile users under the same mobile environment structure and setting.

In Figure 1, the HLR and VLRs contain databases for tracking mobile users, but the intermediate switches such as LSTPs and RSTPs are only used for connecting the HLR with VLRs. At the lowest level, each of the hexagonally-shaped cells in Figure 3A can be viewed as corresponding to an RA. Adopting the hexagonal coverage model, the optimal number of RAs covered by a VLR is determined to be $N = 3n^2 - 3n + 1$ where n is equal to either two or three [8]. We shall call such a VLR an n-layer VLR which itself is a hexagonal region. Going into the second lowest level of Figure 1, we can again view each hexagonally-shaped cell in Figure 3A as corresponding to a VLR and therefore an n-level LSTP will contain $3n^2 - 3n + 1$ VLRs. This view continues as we recursively go up to the higher levels of the PCS network shown in Figure 1, until the RSTP level is reached. Note that only VLRs and the HLR contain the databases. Each n-layer entity (i.e. an n-layer VLR, LSTP or RSTP), when viewed as a hexagonal composite, contains $6N$ or $6(3n^2-3n+1)$ edges (counting internal edges twice to account for the reverse traffic) with the number of boundary edges around it being $12n - 6$, where n is the number of layers. We will analyze the case when $n = 2$ which is considered to be the optimal case [8].

Let the optimal forwarding steps of the system be k. When a mobile user makes a move within the same VLR, only the VLR database is updated and no pointer forwarding is needed. However, as the mobile user moves across the VLR boundary and the number of forwarding steps is less than k, a pointer connection between two involved VLRs must be established to track the user. After the move, the mobile user may be within the same LSTP (we call it an intra-LSTP movement), out of the same LSTP but within the same RSTP (we call it an intra-RSTP movement) or out of the same RSTP, the probabilities of which are estimated below.

1. The probability that a mobile user moves within the same LSTP, that is, the probability of an intra-LSTP movement, when the mobile unit moves across the VLR boundary, is given by

$$q_l = \frac{6(3n^2 - 3n + 1) - (12n - 6)}{6(3n^2 - 3n + 1)} = \frac{3n^2 - 5n + 2}{3n^2 - 3n + 1}. \tag{5}$$

2. Assume we use two-layer RSTPs, the probability that a mobile user moves out of an LSTP but still within the same RSTP, that is, the probability of an intra-RSTP movement, when the mobile unit moves across the VLR boundary, is given by

$$\begin{aligned} q_r &= \frac{7 \times 6(3n^2 - 3n + 1) - (6 \times (12n - 6)/2)}{7 \times 6(3n^2 - 3n + 1)} \\ &= \frac{21n^2 - 27n + 10}{7(3n^2 - 3n + 1)}. \end{aligned} \tag{6}$$

Let δ be the average cost of setting up the forwarding pointer between two VLRs inside the same LSTP, ϵ be the average cost of setting up the forwarding pointer between two VLRs inside the same RSTP but in different LSTPs and ω be the average cost of setting up the forwarding pointer between two VLRs in two different RSTPs. Again, let T represent the average communication cost (round trip) between a VLR and the HLR and let τ represent the average communication cost (round trip) between two neighboring VLRs. Also, we consider the assumption that the HLR of a mobile user and any VLR which the mobile user may visit are in two separate RSTPs connected by a PSTN [4]. This assumption is justified for users moving among different cities or even just among different locations in a city in a more dense area. Thus, we have

$$T = \omega \tag{7}$$

$$\tau = q_l \times \delta + (q_r - q_l) \times \epsilon + (1 - q_r) \times \omega. \tag{8}$$

To quantify these two parameters further, assume that the communication costs for traversing various network elements are as follows. Let L and R represent the costs of processing and routing a message by an LSTP and an RSTP, respectively. Let C_{VL} be the cost of transmitting a message between a VLR and its LSTP. Let C_{LR} be the cost of transmitting a message between an LSTP and its RSTP. Finally, let C_H be the communication cost (round trip) to pass through a PSTN for the case when two VLRs are located in two distinct RSTPs. Then, we can calculate δ, ϵ and ω as follows:

$$\delta = 4C_{VL} + 2L \tag{9}$$

$$\epsilon = 4(C_{VL} + C_{LR}) + 2(L + R) \tag{10}$$

$$\omega = 4(C_{VL} + C_{LR}) + 2(L + R) + C_H. \tag{11}$$

If in the mobile environment the communication cost is a dominant factor, the cost of processing and routing (L and R) can be ignored in the above computation. Plugging into the equation for τ using the expressions for δ, ϵ and ω, we obtain

$$T = 4(C_{VL} + C_{LR}) + C_H \tag{12}$$

$$\begin{aligned} \tau = q_l \times (4C_{VL}) + (q_r - q_l) \times (4C_{VL} + 4C_{LR}) \\ + (1 - q_r) \times (4C_{VL} + 4C_{LR} + C_H). \end{aligned} \tag{13}$$

In the hexagonal coverage model, we assume that the mobile user moves to one of its neighbors with equal probability, i.e. 1/6, as illustrated in Figure 3B. We therefore calculate the mobility rates of a user moving to a new VLR and returning to the last-visited VLR, respectively, as follows:

$$\sigma_f = \tfrac{5}{6}\sigma \tag{14}$$

$$\sigma_b = \tfrac{1}{6}\sigma. \tag{15}$$

Below we consider an example for which the network consists of two-layer VLRs, LSTPs, RSTPs and HLR, and C_{VL}, C_{LR} and C_H are 1, 0.5 and 6 s, respectively. This yields δ as 4 s, ϵ as 6 s and ω as 12 s.

The probabilities of intra-LSTP and intra-RSTP movements are given, respectively, by

$$q_l = \frac{3 \times 2^2 - 5 \times 2 + 2}{3 \times 2^2 - 3 \times 2 + 1} = 0.57 \tag{16}$$

$$q_r = \frac{21 \times 2^2 - 27 \times 2 + 10}{7 \times (3 \times 2^2 - 3 \times 2 + 1)} = 0.82. \tag{17}$$

Therefore, τ and T are given as

$$T = 4(1 + 0.5) + 6 = 12 \text{ s} \tag{18}$$

$$\begin{aligned} \tau &= 0.57 \times 4 + (0.82 - 0.57) \times 6 + (1 - 0.82) \times 12 \\ &= 5.94 \text{ s}. \end{aligned} \tag{19}$$

Based on the estimated values of τ and T above, we can then parametrize μ_p, μ_i and m_k for the Markov model under the hexagonal coverage model using Equations (2)–(4). Here one should note that μ_p, μ_i and m_k are essentially network-dependent parameters, so their values will be changed as we consider a different network model. Later in Section 4.3, we will apply the same parametrization procedure to estimate the values of the same three network-dependent parameters under the mesh coverage model.

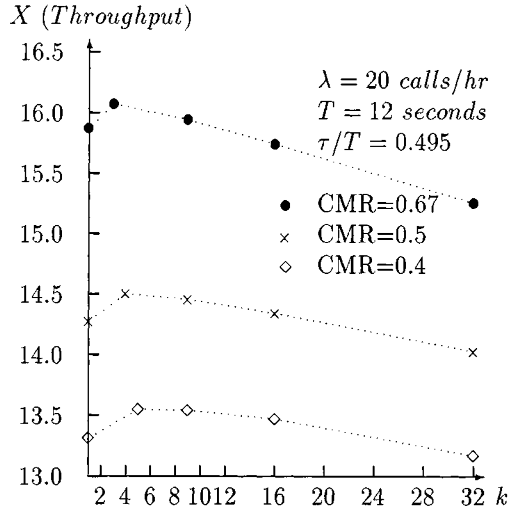

FIGURE 4. Finding the optimal number of forwarding steps under the hexagonal coverage model.

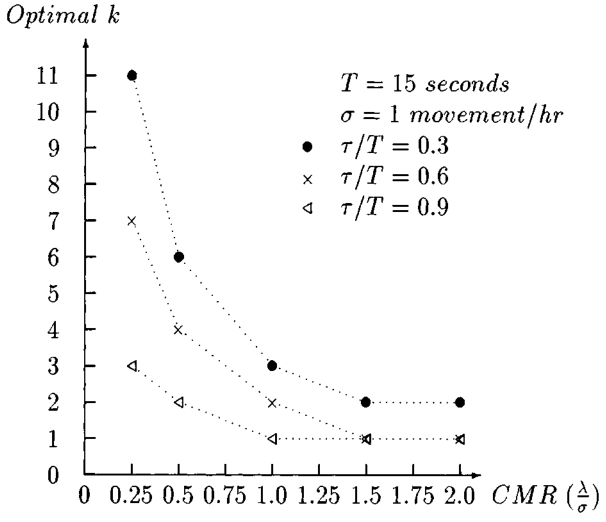

FIGURE 5. Optimal k under various τ/T and CMR values for the hexagonal coverage model.

We now analyze the effect of mobile-user dependent parameters, i.e. λ and σ, by changing their values. Figure 4 shows a case where the arrival rate of calls to a particular mobile user is fixed at $\lambda = 20$ calls/h while the mobility rate of the mobile user is $\sigma = 30$, 40 or 50 movements/h, so that the value of the CMR parameter (λ/σ) is 0.67 (top curve), 0.5 (middle curve) or 0.4 (bottom curve). The data shown in Figure 4 were obtained by solving the Markov model shown in Figure 2 for each k value ranging from 2 to 32, using the SHARPE software package [11] to first obtain the P_j for each state j and then compute X based on Equation (1). From Figure 4, we observe that there exists an optimal k value for each CMR value. For example, the optimal k value is 3 when $\lambda = 20$ calls/h and $\sigma = 30$ movements/h, or when CMR $= 0.67$. In this case, after the mobile user moves across three VLR boundaries, it should perform a reset operation to inform the HLR of its current VLR location so that the average effective throughput of the signaling network can be optimized. Figure 4 also displays the optimal k values at other CMR values.

The above performance analysis for the hexagonal model is done with a particular set of values for (C_{VL}, C_{LR}, C_H), which results in τ/T being fixed at a particular value (i.e. 0.495 in Figure 4). In reality the values of C_{VL}, C_{LR} and C_H may change as the network structure changes. The effect of different values of C_{VL}, C_{LR} and C_H can be analyzed by directly considering the effect of τ/T on k. Figure 5 summarizes the effects of τ/T and CMR on k. It should be mentioned that τ/T affects k via network-dependent parameters (μ_p, μ_i and m_k), while CMR affects k via mobile-user dependent parameters (λ and σ). In Figure 5, the X-coordinate indicates the value of the CMR parameter while the Y-coordinate indicates the optimal k value associated with a particular set of values for τ/T and CMR.

We first discuss the effect of CMR on k. As we can see from Figure 5, when the CMR value becomes larger or, equivalently stated, when the mobile user is called more often than it crosses database boundaries, the system would yield a better throughput with a smaller k value. On the other hand, with a smaller CMR value, the system would perform better with a larger k value. For example, with τ/T fixed at 0.3 (the bullet curve), the optimal k is 3 when CMR $= 1$ and then becomes 11 when CMR $= 0.25$. The interpretation of this result is clear: when the arrival rate is low compared with the mobility rate (i.e. when CMR is low), it is better that we hold the resetting operation such that the amortized cost per move for updating the user location information is minimized. In other words, since the user is not called very often relative to its moves, it is worthless to minimize the search time per call by performing resetting operations frequently. These trends correlate well with those reported in [1]. Our analysis here, however, is one step further as it predicts exactly what the optimal value of k should be so as to optimize a specified performance (or cost) measure.

Figure 5 also demonstrates the effect of τ/T. Specifically, when τ/T is a small ratio, or when the home-to-visitor communication cost is much higher than the visitor-to-visitor communication cost, it is better that k is a large value; otherwise, it is better that k is a small value. For example, with CMR fixed at 0.5 in Figure 5, the optimal k is 6 when τ/T is 0.3, and then drops to 2 when τ/T is 0.9. The physical interpretation of this result is also obvious, i.e. resetting operations can be performed more frequently when the home-to-visitor communication cost is not too high compared with the visitor-to-visitor communication cost. In the extreme case when $\tau \approx T$, the benefit of forwarding is small and the system is better off performing reset operations very frequently except when CMR is very small.

In Figure 6, we study the effect of the magnitude of the call arrival rate (λ) on k. Figure 6 uses the same coordinate system as in Figure 5. However, we deliberately lower the

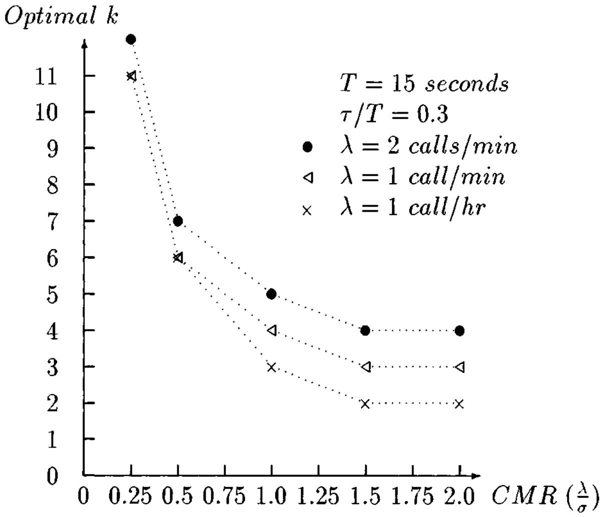

FIGURE 6. Optimal k under various λ and CMR values for the hexagonal coverage model.

arrival rate of λ from 2 calls/min (top curve) to 1 call/min and finally down to 1 call/h (bottom curve). Note that since the x-coordinate is the CMR parameter, the magnitude of σ changes the same way as λ changes in these three cases. However, the service time for locating the user, which is related to T and τ, is fixed in these three cases. Consequently, the bottom curve in Figure 6 represents the case in which the service rate (for executing the operations provided by the PCS network) is relatively faster than the operation arrival rate. (Recall that the two operations are 'updating the user location' and 'locating the user'.)

Figure 6 shows that although all three curves follow the same trend, i.e. k increases as CMR decreases (the same trend as the result obtained in Figure 5), the curves with a low arrival rate (the bottom curve with $\lambda = 1$ call/h) has a sharper slope than the curves with a high arrival rate (e.g. the top curve with $\lambda = 2$ calls/min). Furthermore, the optimal k value decreases as the arrival rate decreases for the same CMR and τ/T values, except when CMR is very small. A possible physical interpretation of the above results is that as the arrival rate decreases, the possibility that the PCS network can serve a 'bulk' of calls accumulated at the HLR simultaneously also decreases. In other words, it is more likely that one call will be serviced at a time before the next call arrives. Because of the lack of bulk services, X in effect is more close to the traditional throughput measure in terms of the number of operations serviced per unit time, with 'an operation' being a move across a visitor database boundary or a single call, but not a 'bulk call' operation. Consequently, the bottom curve is more sensitive to the optimal k value. The lack of 'bulk call' operations when the operation arrival rate is low also reduces the benefit associated with amortization. As a result, the optimal k value also decreases.

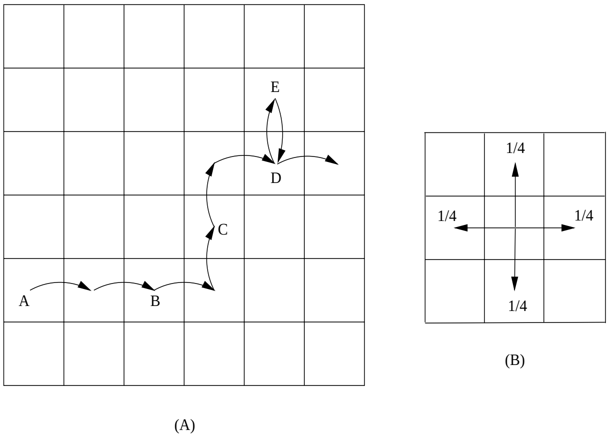

FIGURE 7. Mesh coverage model.

4.3. Mesh coverage model

In this section, we consider the mesh coverage model as shown in Figure 7A. In the mesh coverage model, cells are assumed to be square shaped, with each cell having four neighbors (see Figure 7B). Suppose that the mobile user was initially registered at A and that the optimal number of forwarding steps is two. Then, the system shall perform a reset operation as the mobile user visits B, C and D, since two movements have been made across the visitor database boundaries since the last reset operation was performed. However, if the mobile user after having visited E returns to D again, then there is no need to perform a reset operation since D–E–D forms a cycle. In this case the number of steps is updated to zero at D and no reset operation with the HLR is required, even though the number of the forwarding steps since the last reset operation is two. Figure 7A illustrates the above scenarios.

We use the mesh coverage model to construct the PCS network shown in Figure 1 as follows. We assume that an LSTP contains e^2 VLRs, where e is a natural number, and that an RSTP contains m^2 LSTPs, where m is also a natural number. Let the optimal number of forwarding steps in this PCS network be k. As in Section 4.2, we consider the following three cases when the mobile user moves across the VLR boundary and the number of the forwarding steps is less than k: (i) it may be still within the same LSTP (we call it an intra-LSTP movement); (ii) it may be not in the same LSTP but still within the same RSTP (we call it an intra-RSTP movement); or (iii) it may be not in the same RSTP. Again, in all these cases a pointer connection between two involved VLRs must be established to track the user. Below, we estimate the probabilities of these three cases for the mesh coverage model.

1. The probability of an intra-LSTP movement when the mobile user moves across the VLR boundary is given by

$$p_l = \frac{4e^2 - 4e}{4e^2} = \frac{e-1}{e}. \tag{20}$$

2. Assume an RSTP contains m^2 LSTPs, the probability that a mobile user moves out of a LSTP but still within the same RSTP, that is, the probability of an intra-RSTP movement, when the mobile unit moves across the VLR boundary, is given by

$$p_r = \frac{4(m^2 \times e^2) - 4me}{4m^2\, e^2} = \frac{me - 1}{me}. \qquad (21)$$

Using the same definition for δ, ϵ, ω, T and τ as in Section 4.2, we have

$$T = \omega \qquad (22)$$

$$\begin{aligned}\tau &= p_l \times \delta + (p_r - p_l) \times \epsilon + (1 - p_r) \times \omega \\ &= \frac{e-1}{e}\delta + \frac{m-1}{me}\epsilon + \frac{1}{me}\omega. \qquad (23)\end{aligned}$$

In the mesh coverage model, we assume that the mobile user moves to one of its neighbors with equal probability, i.e. 1/4, as shown in Figure 7B. We can estimate the mobility rates of the mobile user moving to a new VLR and returning back to the last-visited VLR, respectively, as follows:

$$\sigma_f = \tfrac{3}{4}\sigma; \quad \sigma_b = \tfrac{1}{4}\sigma. \qquad (24)$$

We now consider the four-level PCS network in Figure 1 for the case when each LSTP contains $3^2 = 9$ VLRs ($e = 3$) and each RSTP contains $3^2 = 9$ LSTPs ($m = 3$), and also when C_{VL}, C_{LR} and C_H are 1, 0.5 and 6 s, respectively. This yields δ as 4 s, ϵ as 6 s and ω as 12 s. Based on these values, the probabilities of intra-LSTP and intra-RSTP movements are calculated as

$$p_l = \frac{4 \times 3^2 - 4 \times 3}{4 \times 3^2} = 0.67 \qquad (25)$$

$$p_r = \frac{4(3^2 \times 3^2) - 4 \times 3 \times 3}{4 \times 3^2 \times 3^2} = 0.89. \qquad (26)$$

Therefore, T and τ are given as

$$T = 12 \text{ s} \qquad (27)$$

$$\begin{aligned}\tau &= 0.67 \times 4 + (0.89 - 0.67) \times 6 + (1 - 0.89) \times 12 \\ &= 5.46 \text{ s}. \qquad (28)\end{aligned}$$

Now using Equations (2)–(4), we can parametrize μ_p, μ_i and m_k in the Markov model to analyze the behavior of the PCS network constructed from the mesh coverage model. Below we show only one result for the mesh coverage model for simplicity since the results obtained under the mesh coverage model are very similar to those obtained under the hexagonal coverage model. Figure 8 shows the resulting optimal k values (y-coordinate) as a function of CMR values (x-coordinate) and τ/T values for the mesh coverage model. Not surprisingly, the trend shown in Figure 8 is very similar to that shown in Figure 5 for the hexagonal coverage model. The reason for the similarity is that the same physical interpretation can apply regardless of which coverage model has been used to construct the PCS network. There are some subtle differences between the two sets of curves in Figures 5 and 8. For example, with the ratio of τ/T fixed at 0.9 and the value of CMR fixed at 0.25, the optimal k value is 3 under the hexagonal coverage model and is 5 under the mesh coverage model. The differences are conceivable since the coverage model inherently affects the parameter values for the same set of environment settings. Nevertheless, these two coverage models have shown good correlations with each other in predicting the optimal k value for all environment-setting conditions that we have tested. Therefore, the same physical interpretation of the results for the hexagonal model can also be applied here.

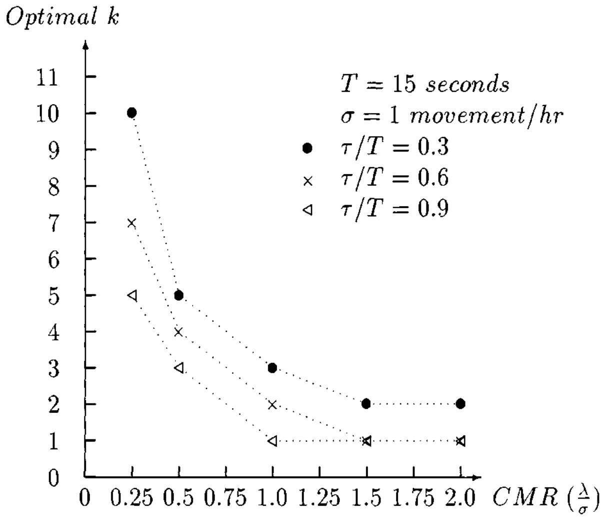

FIGURE 8. Optimal k under various τ/T and CMR values for the mesh coverage model.

5. SUMMARY

In the paper, we developed a Markov model to describe the behavior of a mobile user as it moves while being called in a PCS network subject to the forwarding and resetting technique for tracking mobile users. A designer can define a performance or cost measure by assigning rewards or penalties to states of the Markov model, and then find out the best number of forwarding steps based on our model so as to maximize the specified performance metric (as done in this paper) or minimize the specified cost measure. In this paper, we demonstrated how the Markov model can be parametrized and applied to a PCS network structure (an HLR and its VLRs) with the hexagonal coverage model and the mesh coverage model based on the concept of ‘reward optimization’. We found that the optimal value of k depends not only on the values of CMR and τ/T, but also on the relative ratio of the call arrival rate and call service rate. We studied some possible cases and gave a physical interpretation of the result.

Some future research areas related to this paper include (i) analyzing the effect of combining the forwarding and caching strategies [9] for location management and (ii) applying a similar modeling technique to determine the best number of forwarding steps when the PCS network is able to provide patron services [12].

REFERENCES

[1] Jain, R., Lin, Y. B., Lo, C. and Mohan, S. (1995) A forwarding strategy to reduce network impacts of PCS. *14th Annual Joint Conf. of the IEEE Computer and Communications Societies (IEEE INFOCOM '95)*, April, Boston, MA, USA, pp. 481–489.

[2] Rao, S., Gopinath, B. and Kurshan, D. (1992) Optimizing call management of mobile units. *3rd IEEE Int. Symp. Personal, Indoor and Mobile Communications*, pp. 225–229.

[3] Krishna, P., Vaidya, N. H. and Pradhan D. K. (1994) Location management in distributed mobile environment. *3rd Int. Conf. Parallel and Distributed Information Systems*, Austin, Texas, USA, pp. 81–88.

[4] Lin, Y. B. (1997) Reducing location update cost in a PCS network. *IEEE/ACM Trans. Networking*, **5**, 25–33.

[5] Bar-Noy A. and Kessler, I. (1994) Mobile users: to update or not to update? *13th Annual Joint Conf. of the IEEE Computer and Communications Societies (IEEE INFOCOM '94)*, Toronto, Ontario, Canada, June, pp. 570–576.

[6] TIA/EIA-IS-41.6 (1996) *Cellular Radio Telecommunication Intersystem Operations*. Electronic Industries Alliance, Arlington, VA, USA.

[7] Mouly, M. and Pautet, M. B. (1992) *The GSM System for Mobile Communications*, 49 rue Louise Bruneau, Palaiseau, France.

[8] Lai, W. R. and Lin, Y. B. (1996) Mobility database planning for PCS. *1996 Workshop on Distributed System Technologies and Applications*, Tainan, Taiwan, pp. 263–269.

[9] Jain, R., Lin, Y. B., Lo, C. and Mohan, S. (1994) A caching strategy to reduce network impacts of PCS. *IEEE J. Selected Areas Commun.*, **12**, 1434–1444.

[10] Kleinrock, L. (1975) *Queueing Systems, Vol. 1: Theory*. Wiley, New York.

[11] R. Sahner, R., Trivedi, K. S. and Puliafito, A. (1996) *Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the SHARPE Software Package*. Kluwer Academic, Norwell, MA.

[12] Cho, G. and Marchall, L. F. (1995) An efficient location and routing scheme for mobile computing environments. *IEEE J. Selected Areas Commun.*, **13**, 868–879.