Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Modeling and Analysis of Forwarding and Resetting Strategies for Location Management in Mobile Environments

Ing-Ray Chen, Tsong-Min Chen and Chiang Lee*
Institute of Information Engineering
National Cheng Kung University
Tainan, Taiwan

Abstract

"Forwarding and resetting" is a technique for reducing the signaling cost of the network for location management of mobile users who can move from one place to another while being called. With forwarding, the system follows a chain of databases to locate a mobile user; with resetting, the system 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 formulate the problem of finding the best time to perform resetting as an optimization problem. We consider the signaling network as a 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 use a Markov chain to describe the behavior of a mobile user and analyze the best time when forwarding and resetting should be performed in order to optimize the service rate of the mobile network with respect to the above two types of operations. We demonstrate the applicability of our approach with a case study and provide a physical interpretation of the result.

Index Terms — Mobile computing, location management, forwarding and resetting, Markov models, performance evaluation, distributed systems.

1 Introduction

In a distributed mobile environment, a mobile user can call or be called by other mobile users while it moves. Over the past few years, various location management strategies based on the concept of forwarding and resetting have been proposed and studied in the literature [4, 10] with a goal of minimizing the network signaling and database loads. Both the "forwarding" and "resetting" mechanisms refer to the action taken by the system in response to a move by a mobile user. "Forwarding" means that when a mobile user moves across a database boundary (e.g., a registration area boundary [4] or a base station boundary [10]), only a pointer is set-up between the two involved databases, while "resetting" 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 for 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 for 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.

A 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 acrossing 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 [6, 8]. Various cost measures have been proposed in previous studies and have resulted in different solutions to the research issue. Rao, Gopinath and Kurshan [10] 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

*This work is supported in part by the National Science Council of R.O.C. under Grant No. NSC-86-2745-E-006-020.

work is by Bar-Noy and Kessler [1] 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.

Recently, Jain et al. [4] 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 [2] or a GSM [9] 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) greater than a certain threshold value. The contribution of their work is that they provide 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 (VLRs) in the forwarding chain before the address of the called mobile user is found. This simplifying assumption may yield 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, also assuming that the system will perform a reset after the mobile user crosses the database boundary k times. Our Markov model does not assume any specific architecture so that it can be applied at the base station (BS) level as having been considered in [10], or at the registration area (RA) level as having been considered in [4]. Furthermore, the Markov model is developed without any specific cost measure in mind. A specific cost measure is obtained by first assigning "rewards" to the states of the Markov chain and then calculating the probabilistic "average" reward. By utilizing the Markov chain, we therefore formulate the problem of finding the best time to perform resetting as an optimization problem, that is, finding the optimal value of k under which the average "reward" representing the specified cost measure is optimized. Our Markov model takes into account the mobility and calling events of the mobile user and keeps track of the states of the mobile user as these events occur; thus, we do not make any simplifying assumption regarding the number of forwarding steps a call must follow to locate a user as having been done in [4].

The rest of the paper is organized as follows. Section 2 gives a more detailed background of the problem we are trying to solve and states the system assumptions. Section 3 develops a Markov model that describes the states of a mobile user as it is being called while it moves. We view the signaling network as a 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 a method of assigning "rewards" to the states of a mobile user so as to yield a performance measure to reflect the throughput of the server with respect to servicing the above two types of operations. Section 4 shows a case study to demonstrate the utility of our model. Finally, Section 5 summarizes the paper and outlines some possible future research areas.

2 Background, System Model and Assumptions

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 as the "home" database. The home database will give a direction to the first low-level database unit on the forwarding chain. We shall call the databases on the forwarding chain at the low-level as "visitor" databases. To allow us to easily distinguish the visitor databases at the lower-level, we shall call the first one on the chain as v_0 and the last one as 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 (e.g., a BS [10] or an RA [4]) 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 [4, 7] 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 thus is characterized by its call-to-mobility ratio, defined as λ/σ.

We assume that the forwarding chain for locating a mobile user does not form a cycle and 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 (eventually) go to

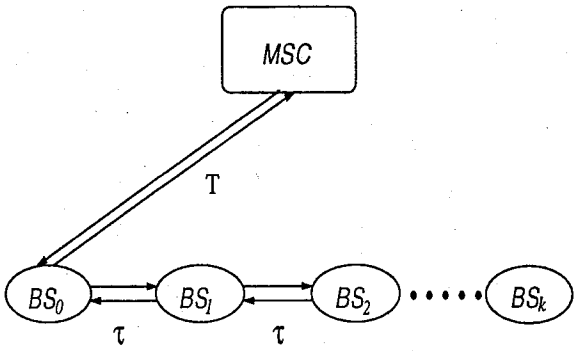

Figure 1: Forwarding and Resetting Technique Applying to MSC and BS levels.

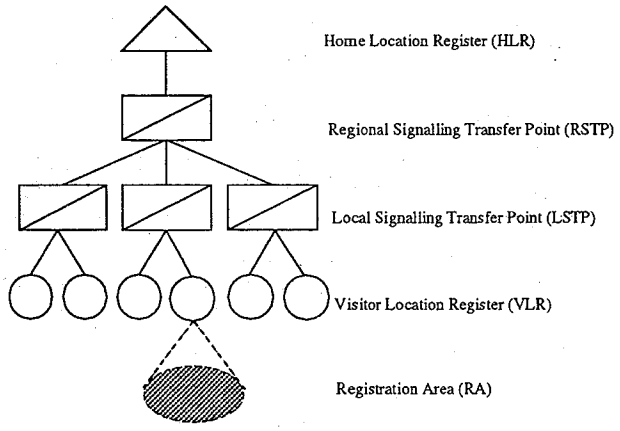

Figure 2: Forwarding and Resetting Technique Applying to HLR and VLR levels.

the current home database.[1] Consequently, if there are more than one request 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 all at once 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. Of course, during the time period in which the pointer connection is set-up, call requests can arrive.

Our model does not specifically assume any architecture. The modeling and analysis technique developed in the paper can be applied to any two adjacent levels in a tree-like signaling network. To facilitate our discussion, however, we shall refer to the two structures shown in Figures 1 and 2. Figure 1 depicts a simple two-level structure between a mobile switching center (MSC) and its base stations (BS) as discussed in [10]. This structure corresponds to a lower-level segment of the signaling network. When applying the forwarding and resetting technique to this structure, a mobile user can move across the BS boundaries and set-up a chain of forwarding pointers among BSs. In this case, the MSC corresponds to the "home" database while a BS corresponds to a "visitor" database. Figure 2, on the other hand, shows a possible higher-level segment of the signaling network as discussed in [3]. In this case, the home location register (HLR) and visitor location registers (VLRs) contain databases for tracking mobile users, but the intermediate switches such as RSTPs and LSTPs are only used for connecting the HLR with VLRs. When applying the forwarding and resetting technique to this 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, without specifically referring to any structure. Then we can apply the Markov model to the above two different structures by parameterizing (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. Later in Section 4, we will illustrate how to apply our Markov model to the structure shown in Figure 1.

3 Stochastic Analysis of Forwarding and Resetting

Notation

- λ : the arrival rate of calls to a particular mobile user.
- σ : the mobility rate of a particular mobile user.
- CMR : the call-to-mobility ratio of a particular mobile user, i.e., λ/σ.
- μ_p : the execution rate to set-up 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,

[1] We deliberately call the home database as the *current* home database since the resetting and forwarding technique can also be applied to the next two higher-levels.

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: (a) the number of forwarding steps; and (b) a binary quantity indicating whether or not it is in the state of being called. Figure 3 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 has been made since the last reset operation. Initially, the mobile user is in the state of $(I, 0)$, 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 of (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 of (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 3 since it does not need to be considered when solving a Markov chain [5]. Note that this implies that in state (C, i) the number of requests accumulated to locate the mobile user may be greater than 1.

3. If the mobile user is in the state of (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 all 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.

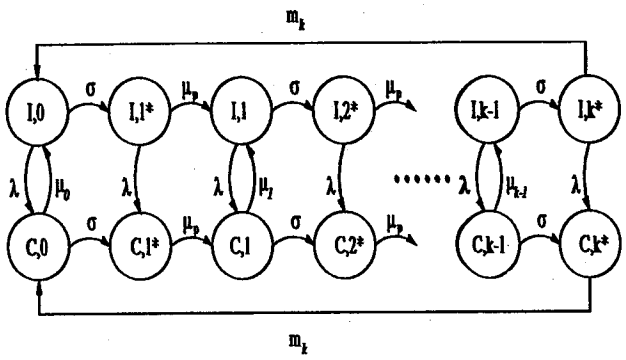

Figure 3: A Markov Model for Describing Forwarding and Resetting.

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

   (a) If $0 \leq i < k-1$, 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) 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.

   (b) If $i = k-1$, 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, (a) another call arrival which makes the system stay at the same state; (b) a move by the mobile user crossing a visitor database boundary which makes the system transit to state $(C,2)^*$; and (c) 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 3 is ergodic [5], 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 transitions 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 state $(I,i+1)^*$ since it is more likely for state (I,i) to make a transition into state (C,i) than into state $(I,i+1)^*$. Since the probability that the system stays at 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 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[2] 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 3, 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 3 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 percentage of time the system is found to be staying at state j in equilibrium. Then,

$$X = \left(\sum_{i=0}^{k-1} P_{(C,i)} \times \mu_i\right) + \left(\sum_{i=0}^{k-1} (P_{(I,i)^*} + P_{(C,i)^*}) \times \mu_p\right) + (P_{(I,k)^*} + P_{(C,k)^*}) \times m_k \quad (1)$$

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 parameterize it based on the specific characteristics of the signaling network under consideration.

4 Performance Analysis: An Example

In this section, we illustrate the utility of our model with the signaling network structure shown in Figure 1. Assume that the communication time (round trip) between the MSC at the higher level and a BS at the lower-level is exponentially distributed with an average time of T. Also assume that the communication time (round trip) between two neighboring BSs at the lower-level is exponentially distributed with an average time of τ (see Figure 1). Furthermore, we assume that when performing a reset operation, BS_k first communicates with the MSC to indicate that it is now the new BS_0 for the called mobile user. The MSC confirms the change by replying a message to BS_k and also sends a cancellation message to the old BS_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 accumulate the service cost (e.g., credits) information. After the last step is done, BS_{k-1} informs the MSC to confirm the completion of the reset operation. As we can see, a reset operation implemented this way involves twice the amount of the signaling cost than locating a user.

Given the above assumptions, we can parameterize the

[2] If we were to choose a cost measure instead of a performance measure, then we would have to minimize it instead of to maximize it.

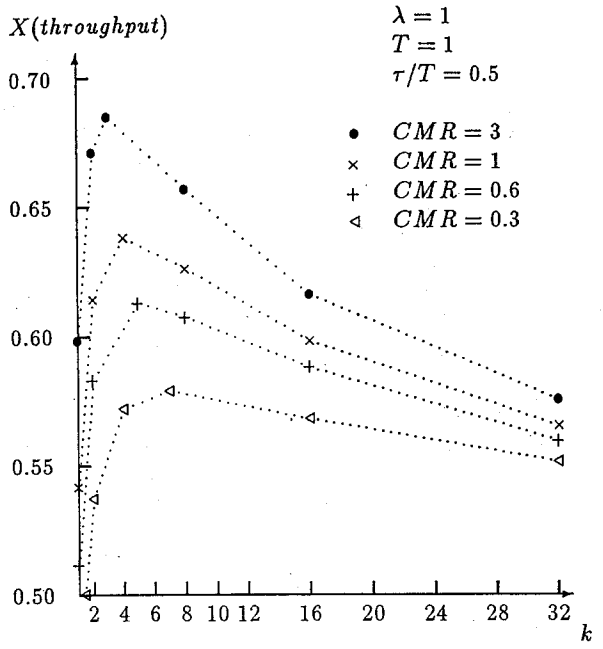

Figure 4: Effect of CMR on the Optimal Value of k to Maximize X.

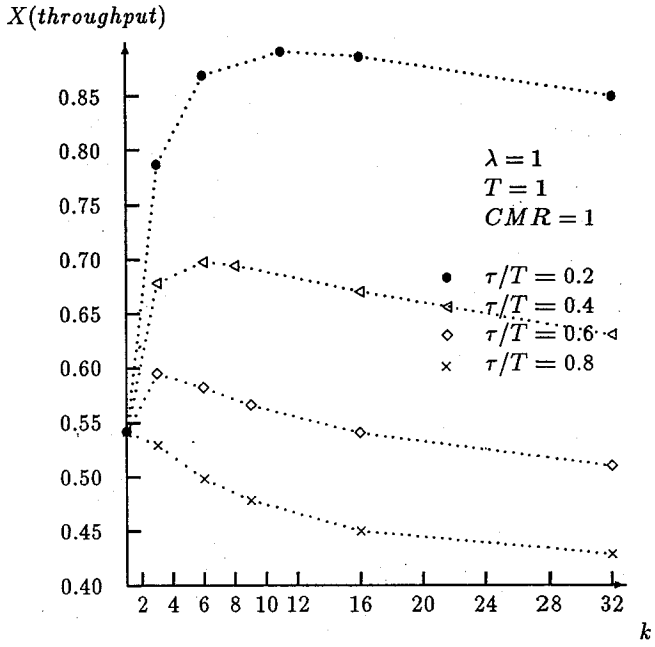

Figure 5: Effect of τ/T on the Optimal Value of k to Maximize X.

Markov chain as follows:

1. The execution rate to set-up or travel a pointer connection between two visitor databases is simply given by
$$\mu_p = \frac{1}{\tau} \quad (2)$$
2. The execution rate to find BS_i and then to locate the mobile user can be estimated as
$$\mu_i = \frac{1}{T + i\tau} \quad (3)$$
3. The execution rate to perform a reset operation can be estimated as
$$m_k = \frac{1}{2T + k\tau} \quad (4)$$

Here we note that $2T$ is used for estimating m_k because the BSs at the lower-level have to communicates with the MSC at the higher-level two times.

Figure 4 shows a plot of X against k, considering the case when $\tau = 0.5T$, for various CMR (i.e., λ/σ) values. The data on the figure are obtained by first solving the Markov model using the SHARPE software package [11] to obtain the P_j for each state j and then by computing X based on Equation 1. As expected, Figure 4 shows that when the CMR value is large, that is, when the mobile user is called more often than it crosses BS database boundaries, the system yields a better throughput with a small k value. For example, when $CMR = 3$, the optimal k value is 3. On the other hand, with a small CMR value, the system performs better with a large k value, e.g., when $CMR = 0.3$, the optimal k value is 7. These trends correlate well with those reported in [4]. 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 shows the effect of the ratio of τ/T on the optimal value of k, when the CMR of the called mobile user is assumed to be fixed at a constant value. Recall that the ratio of τ/T represents the signaling cost (service time) difference between the "visitor-to-visitor" communication cost and the "home-to-visitor" communication cost. Figure 5 shows that when the home-to-visitor communication cost is much higher than the visitor-to-visitor communication cost, that is, when τ/T is a small ratio, it is better that k is a large value; otherwise, it is better that k is a small value. In the extreme case that $T \approx \tau$, the benefit of forwarding is very small and the system is better-off by performing reset operations very frequently, i.e., $k \approx 1$.

Figures 4 and 5 above thus study the effects of CMR (λ/σ) and τ/T, respectively, on the optimal value of k, because Figure 4 fixes τ/T while varies the CMR values, and Figure 5 does the opposite. Another effect we are interested in studying is the magnitude relationship between the two ratios CMR and τ/T, since it conceivably will also affect the shape of the curve. Similar to Figures 4 and 5, Figures 6 and 7 also investigate the effects of CMR and τ/T on the

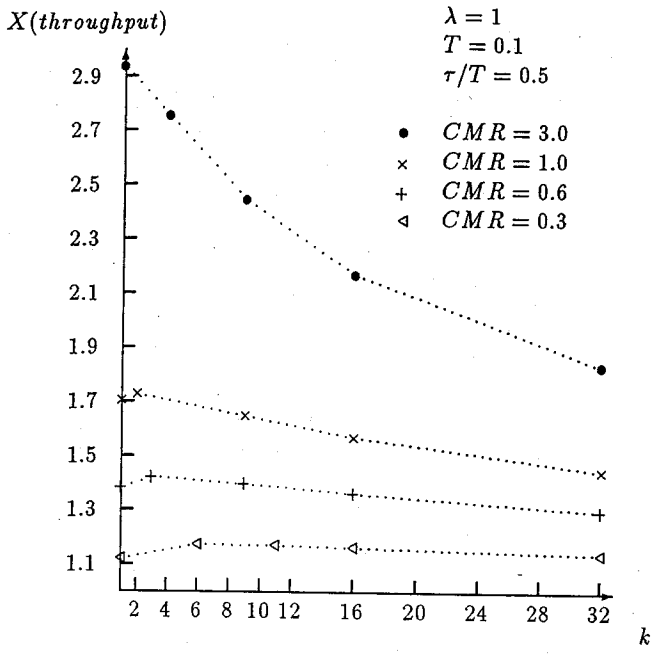

Figure 6: Optimal Value of k to Maximize X for Various CMR Values When the Service Rate is an Order of Magnitude Higher Than the Call Arrival Rate.

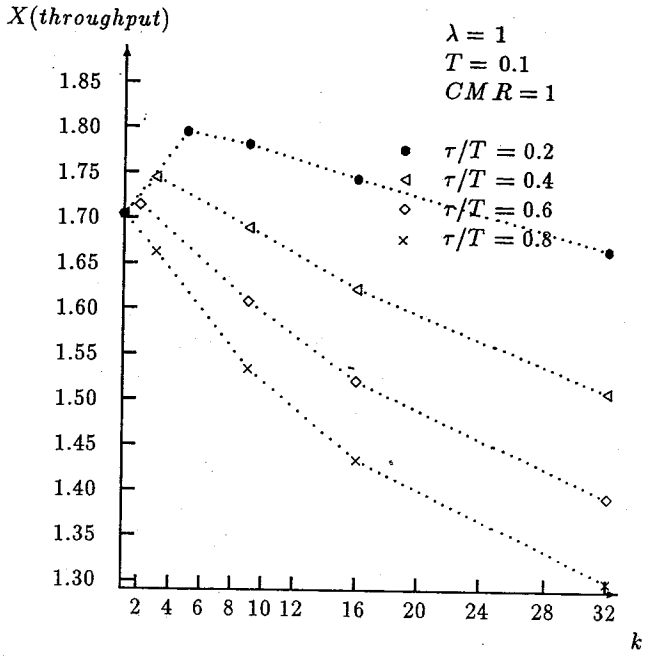

Figure 7: Optimal Value of k to Maximize X for Various τ/T When the Service Rate is an Order of Magnitude Higher Than the Call Arrival Rate.

optimal value of k, respectively, except that $\lambda = 10T = 1$ instead of $\lambda = T = 1$ as in Figures 4 and 5. It represents the case in which the service rate for locating the mobile user, i.e., μ_i based on Equation 3, is about an order of magnitude higher than the call arrival rate λ. This is opposed to the case in Figures 4 and 5 in which μ_i is just at the same order of magnitude as λ.

By comparing Figures 4 and 5 (μ_i is at the same order of magnitude as λ) with Figures 6 and 7 (μ_i is an order of magnitude higher), we see that the trends exhibited in these two sets are consistent with each other, with the optimal k value shifting toward a smaller value in Figures 6 and 7. Furthermore, the throughput X is more sensitive to the value of k in Figures 6 and 7, i.e., the slope of the curve in Figures 6 and 7 is higher. Note that the curves in Figure 6 actually have higher slopes than the corresponding curves in Figure 4 since the scale on the Y coordinate is larger. A possible physical interpretation of the result is that when the service rate per call is an order of magnitude higher than the call arrival rate, it is less likely that the server will serve a "bulk" of calls accumulated at the MSC simultaneously. 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 servered 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, in Figures 6 and 7, X is more sensitive to the the k value. The fact that "bulk call" operations are not serviced much in Figures 6 and 7 (as opposed to Figures 4 and 5) is also reflected by the difference in the magnitude of X in these two sets of figures for the same arrival rate.

One possible reason why Figures 6 and 7 have smaller optimal k values than Figures 4 and 5 for the same CMR and τ/T ratio values is as follows: when the cost of a reset operation is relatively high per move (as in Figures 4 and 5), it is better that the system only performs the reset operation once in a while so that the amortized cost per move is minimized. On the other hand, when the cost of a reset operation is relatively low (as in Figures 6 and 7), the benefit obtained due to amortization reduces. As a result, the optimal k value also decreases in Figures 6 and 7 when compared with Figures 4 and 5.

¿From comparing Figures 5 and 7, we can also make the following conclusion: For any two network structures, the structure with a smaller τ/T ratio can have a higher optimal k value than the structure with a larger τ/T ratio. This result implies that the structure shown in Figure 2 is likely to have a higher optimal k value than the structure shown in Figure 1, because, unlike the latter, the former structure has many intermediate-level routing switches between the home

database (the HLR in Figure 2) and the visitor databases (VLRs in Figure 2) and thus should have a smaller τ/T ratio.

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 mobile environment 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 what the best number of forwarding steps should be using our model so as to maximize the specified performance metric (as having been done in this paper) or minimize the specified cost measure. Our work complements previous works in that we allow the best number of forwarding steps to be determined analytically. Furthermore, since the Markov model is generic in nature, it can be applied to various segments of the signaling network, e.g., between an MSC and its BSs, between a HLR and its VLRs, or between a VLR and its RAs (or MSCs). In this paper, we demonstrated how it can be parameterized and applied to a two-level structure including an MSC and its BSs 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 (a) performing an experimental evaluation on existing signaling networks such as CCS-SS7; (b) combining the forwarding and resetting technique with the mobility database planning technique [7] to determine the best number of RAs that should be covered by a VLR in order to optimize a specified cost or performance measure; (c) considering the system traffic to the signaling network generated by all mobile users as they are calling or being called while moving, and designing and evaluating system-wide (instead of per-user) location management policies for minimizing the total system traffic.

References

[1] A. Bar-Noy and I Kessler, "Mobile users: to update or not to update?" *13th Annual Joint Conference of the IEEE Computer and Communications Societies, (IEEE INFOCOM '94)*, 1994, Vol. 2, pp. 570-576.

[2] EIA/TIA, *Cellular Radio Telecommunication Intersystem Operations,* Technical Report IS-41 (Revision B), EIA/TIA, July 1991.

[3] R. Jain, Y.B. Lin, C. Lo and S. Mohan, "A caching strategy to reduce network impacts of PCS," *IEEE Journal on Selected Areas in Communications,* Vol. 12, No. 8, Oct. 1994, pp. 1434-1444.

[4] R. Jain, Y.B. Lin, C. Lo and S. Mohan, "A forwarding strategy to reduce network impacts of PCS," *14th Annual Joint Conference of the IEEE Computer and Communications Societies, (IEEE INFOCOM '95)*, 1995, Vol. 2, pp. 481-489.

[5] L. Kleinrock, *Queueing Systems, Vol. 1: Theory,* John Wiley and Sons, 1975.

[6] P. Krishna, N.H. Vaidya and D.K. Pradhan, "Location management in distributed mobile environment," *3rd Inter. Conf. Parallel and Distributed Information Systems,* 1994, pp. 81-88.

[7] W.R. Lai and Y.B. Lin, "Mobility database planning for PCS," *1996 Workshop Distributed System Technologies and Applications,* Tainan, Taiwan, 1996, pp. 263-269.

[8] Y.B. Lin, "Reducing location update cost in a PCS network," To appear in IEEE/ACM Transactions on Networking.

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

[10] S. Rao, B. Gopinath and D. Kurshan, "Optimizing call management of mobile units," *3rd IEEE Inter. Symp. Personal, Indor and Mobile Communications,* 1992, pp. 225-229.

[11] R. Sahner, K. S. Trivedi and A. Puliafito, *Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the SHARPE Software Package,* Kluwer Academic, 1995.