Comparing the PCS Location Tracking Strategies

Yi-Bing Lin and Shu-Yuen Hwang

Abstract—**The cache scheme has been proposed to reduce the location tracking overhead of a personal communications services (PCS) network. In the previous papers, we studied the cache scheme under the assumptions that the home location register (HLR) access time is constant and the portable residence times have an exponential distribution. This paper compares the cache scheme with a basic scheme (such as IS-41). We generalize the previous models by considering the queueing effect of the HLR (i.e., we model the HLR by an M/G/1 queue) and by considering an arbitrary distribution for the portable residence times. Our study shows that the cache scheme is likely to outperform the basic scheme when 1) the net traffic to the HLR in the basic scheme saturates and the hit ratio in the cache scheme is larger than zero, 2) the portable mobility is low with respect to the call arrival rate, and 3) the variance of the HLR service time distribution is large (for a fixed mean service time). We also indicate an intuitive result that the cache hit ratio is high for a high call arrival rate and low portable mobility. For a fixed mean portable residence time, we show that a higher cache hit ratio is expected for a portable residence distribution with larger variance.**

I. Introduction

PERSONAL communications services (PCS) provides information (e.g., voice, image, and data) exchange between nomadic end users independent of time, location, access arrangements, or equipment capabilities. In a PCS network, it is necessary to locate the "portables," or subscribers, that move from place to place. A two-level hierarchical strategy [1]–[3] is usually adopted to track the portables. A "location" is meant to be a registration area (RA) that consists of one or more radio port coverage areas. Every RA is associated with a visitor location register (VLR). A VLR may serve one or more RA's. For demonstration purposes, this paper assumes that every VLR serves one RA. Every portable is associated with a database called the home location register (HLR) that stores the current location (i.e., the address of the VLR) of the portable. The HLR record is modified when the portable moves from one RA to another. Consider the example illustrated in Fig. 1.

Suppose that a portable moves from RA R_1 to RA R_2. The portable is registered at VLR_2 (the VLR of R_2), and the new address is reported to the HLR (see Step 1 in Fig. 1). This action is referred to as the "move" operation. The move operation may be followed by a "deregistration" operation to remove the obsolete record in VLR_1 (the VLR of R_1). This action is referred to as "explicit" deregistration [1]. To deregister in protocols such as IS-41, the HLR sends a deregistration message to VLR_1. In [4], the deregistration message is sent directly from VLR_2 to VLR_1 to bypass the signaling network. The deregistration operation may not be performed immediately after a move operation. In "timeout" deregistration, the obsolete VLR entries are cancelled periodically [1]. In "implicit" deregistration, the deregistration operation is performed when the VLR database is full [1], [5]. Comparisons of the three deregistration strategies are discussed elsewhere [6]. In this paper, we consider explicit deregistration, and deregistration is assumed to be part of the move operation (see Step 2 in Fig. 1).

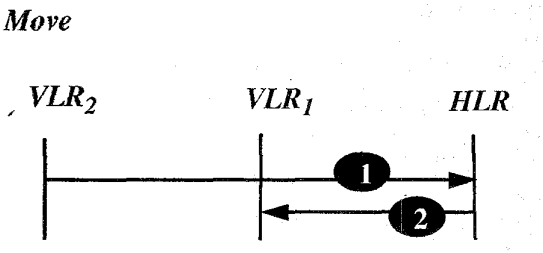

Fig. 1. The move operation (including deregistration).

The procedure to locate a portable (for call delivery) is referred to as the "find" operation. The find operation for most protocols such as IS-41 is described as follows (see Fig. 2):

1) The incoming call to a portable is delivered to an end office (EO).
2) From the dialed mobile identification number (MIN), EO identifies that the call is for a PCS user. EO queries the HLR for the portable's location.
3) The HLR queries the VLR where the portable was last registered.
4) The VLR queries the mobile switching center (MSC), in which the portable is located, to determine whether the portable is capable of receiving the call. If so, the MSC returns a routable address TLDN to the VLR.
5) The VLR forwards the TLDN back to EO via the HLR.

After EO receives the TLDN, it routes the call to the MSC where the portable is located. The detailed signaling message flows of the move and find operations for IS-41 are described in [7]. The above find procedure is called the "basic find" operation to distinguish from the find operation of a newly proposed protocol, to be described next. Note that paging is not required in the find operation (paging is required when the connection between the caller and the callee is established). The MSC and the VLR are assumed to be co-located within a service switching point, and we will use the terms VLR, RA, and MSC interchangeably.

Manuscript received November 30, 1993; revised April 19, 1994. This work was supported by NSC (R.O.C.) project NSC85–2213–E-009–064.

The authors are with the Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan.

Publisher Item Identifier S 0018-9545(96)00184-3.

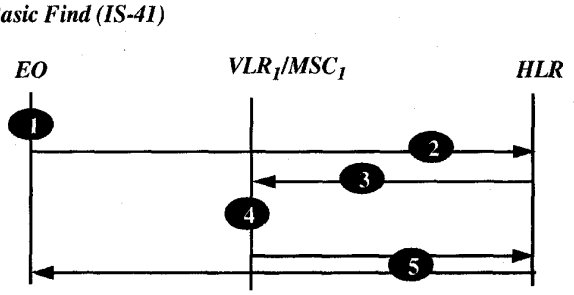

Fig. 2. The basic find operation (the call delivery for IS-41).

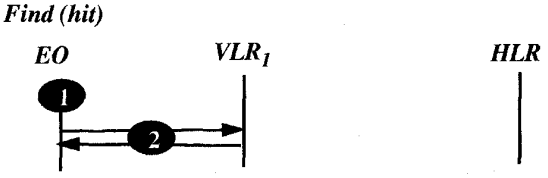

Fig. 3. The find operation in the cache scheme (cache hit).

Studies [8] indicated that the overhead due to the find and move operations is significant. When the frequency of the incoming calls is low with respect to the portable mobility (i.e., the rate that a portable moves to new RA's), it is important to reduce the move cost. Several approaches have been proposed to reduce either the cost of a single move operation [9], [10] or to reduce the number of move operations [11].

On the other hand, if the incoming call frequency is higher than the portable mobility, it is important to reduce the find cost. We have [4] proposed a "cache scheme" to reduce the find cost. The idea is to store the locations of frequently accessed portables in a local database (i.e., cache) within an EO. When a call arrives, the location of the called portable is identified in the cache to avoid sending query messages to the HLR. The steps to locate a portable using cache information are described as follows (see Fig. 3):

1) An incoming call arrives at EO. The cache record indicates that the portable resides in RA R_1.
2) VLR_1 (the VLR of R_1) is queried, which in turn queries MSC_1 to find the routable address for the portable. Then VLR_1 returns the address back to EO.

In the above example, EO contains the current portable location, and a cache hit occurs when the portable is tracked using the cached information. However, due to the mobility of portables, the location information in the cache may be obsolete. Consider Fig. 4. If the portable moves from R_1 to R_2, then the location information in EO is obsolete. If the obsolete cache information is used to locate the portable, then the procedure shown in Fig. 3 fails (the failure is referred to as a "cache miss"), and after the failure the basic find operation (in Fig. 2) is used to locate the portable (see Steps 3–5 in Fig. 4). After the correct location is identified (the end of Step 5 in Fig. 4), the cache of the EO is updated to record the current location of the portable. When a cache miss occurs, the find cost of the cache scheme is higher than the basic find cost. Thus, the cached information should be used only as a hint to

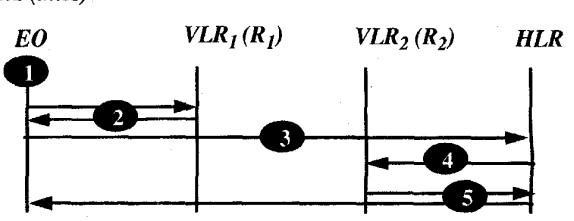

Fig. 4. The find operation in the cache scheme (cache miss).

locate a portable. This paper assumes that the cache is large enough to accommodate all frequently accessed portables. (If the cache size is too small, then some replacement policy is required for replacement.) Several heuristics [4], [12] have been proposed to determine whether the cache information should be used. These studies modeled the portable residence times by an exponential distribution, as proposed in [13], and the time to access the HLR is assumed to be a constant (i.e., there is no queueing at the HLR). This paper extends the previous models to study the impact of the queueing effect at the HLR. We also derive the hit ratio of the cache scheme by assuming a general portable residence time distribution. The notation used in this paper is listed in Appendix A.

II. Modeling the HLR Database

This section models the queueing effect of the HLR. To simplify and strengthen the results, we consider only two types of traffic to the HLR: the traffic due to the find and the move operations. We note that it is trivial to accommodate other types of traffic in our model. Assume that the portables are homogeneous such that the net arrival traffic to the HLR can be approximated by a Poisson process. The mean service time of a VLR is c_1. We assume that the access rate of a VLR is sufficiently small such that the mean time to access the VLR (i.e., the response time) is c_1. The service times of an HLR are represented by a random variable S. The HLR is modeled as a queueing bottleneck, and the mean response time may be much longer than $E[S]$. We ignore the message sending delay. (In the previous studies [4], [12], we considered the message-sending delay. However, the HLR queueing effect was not modeled.) It is not difficult to include the message delays in our model. The response time to complete a query (i.e., the service time plus the waiting time) at an HLR can be derived using the standard technique [14]

$$c_2 = \frac{\lambda^* E[S^2]}{2(1-\lambda^* E[S])} + E[S]. \tag{1}$$

Let N be the expected number of EO's that issue calls to a portable, and n be the number of portables registered in an HLR. Suppose that the calls directed from an EO to a portable are λ. Then $\lambda^* = Nn\lambda + n\eta$ for the basic scheme, and the find cost is

$$C_{basic} = c_1 + E[S] + \frac{(Nn\lambda + n\eta)E[S^2]}{2\{1-(Nn\lambda + n\eta)E[S]\}}. \tag{2}$$

For the cache scheme, let p_m be the probability that a cache miss occurs, and $\lambda^* = Nnp_m\lambda + n\eta$. The find cost is

$$C_{cache} = c_1 + p_m\left\{c_1 + E[S] + \frac{(Nnp_m\lambda + n\eta)E[S^2]}{2\{1-(Nnp_m\lambda+n\eta)E[S]\}}\right\}. \quad (3)$$

We consider HLR service times that have a Gamma distribution with the density function

$$f_G(\mu, \gamma, t) = \frac{\mu^\gamma t^{\gamma-1}e^{-\mu t}}{\Gamma(\gamma)} \quad (4)$$

where

$$\Gamma(\gamma) = \int_{\tau=0}^{\infty} e^{-\tau}\tau^{\gamma-1}\,d\tau, \quad \gamma > 0$$

μ is called the "scale" parameter and γ is called the "shape" parameter. We are interested in the Gamma distribution because it has one very important property—the distribution has no specific characteristic shape. In fact, depending upon the values of the parameters, it can be shaped to represent many distributions as well as shaped to fit sets of measured data that cannot be characterized as a particular distribution other than as a Gamma distribution with certain shaping parameters. By controlling the shape parameter γ, we can study the impact of the variance of the service time when comparing the performance of the basic and the cache schemes. The first two moments of the gamma distribution are $E[S] = \gamma/\mu$ and $E[S^2] = [\gamma(\gamma+1)]/\mu^2$. If $E[S] = c_1$, then $\mu = \gamma/c_1$ and $E[S^2] = [(\gamma+1)/\gamma]c_1^2$. Let $\alpha = Nn\lambda c_1$, and $\beta = n\eta c_1$, then the offered load to the HLR in the basic scheme is $\rho = \alpha + \beta$. From (2)

$$C_{basic} = c_1\left[2 + \frac{(\gamma+1)\rho}{2\gamma(1-\rho)}\right] \quad (5)$$

and from (3)

$$C_{cache} = c_1\left[1 + 2p_m + \frac{(\gamma+1)(\alpha p_m^2 + \beta p_m)}{2\gamma(1-\beta-\alpha p_m)}\right]. \quad (6)$$

From (5) and (6), we can derive the condition that the cache scheme outperforms the basic scheme in terms of the find operation

$$\begin{aligned} & C_{cache} < C_{basic} \\ \Leftrightarrow\ & 2\gamma(2p_m - 1) + \frac{(\gamma+1)(\alpha p_m^2 + \beta p_m)}{1-\beta-\alpha p_m} < \frac{(\gamma+1)\rho}{(1-\rho)} \\ \Leftrightarrow\ & Ap_m^2 + Bp_m - C < 0 \end{aligned} \quad (7)$$

where

$$\begin{aligned} A &= \alpha(1-3\gamma)(1-\rho) \\ B &= [4(1-\beta)\gamma + 2\alpha\gamma + \beta(\gamma+1)](1-\rho) + \alpha(1+\gamma)\rho \\ C &= 2(1-\beta)\gamma(1-\rho) + (1-\beta)(1+\gamma)\rho. \end{aligned}$$

Note that $\rho < 1$. (If $\rho \geq 1$ then the HLR is saturated, and $c_2 \to \infty$ in the basic scheme.) Based on the value of the shape parameter γ, there are three cases.

Case 1 ($\gamma > 1/3$): In this case, $A = -A_1 < 0$, $B > 0$, and $C > 0$, and (7) is rewritten as

$$\begin{aligned} & -A_1p_m^2 + Bp_m - C < 0 \\ \Leftrightarrow\ & A_1p_m^2 - Bp_m + C > 0 \\ \Leftrightarrow\ & p_m < p_m^+ = \frac{B - \sqrt{B^2 - 4A_1C}}{2A_1} \end{aligned} \quad (8)$$

or

$$p_m > \frac{B + \sqrt{B^2 - 4A_1C}}{2A_1}. \quad (9)$$

Note that the right-hand side of (9) is larger than 1, which contradicts the fact that $p_m \leq 1$. Thus, (8) is the sufficient and necessary condition that $C_{cache} < C_{basic}$.

When $\gamma \to \infty$, $E[S^2] = c_1^2$, the first two moments of the Gamma distribution are the same as the constant distribution, and

$$p_m^+ = \frac{1}{6\alpha(1-\rho)}\left\{(1-\rho)(4-3\beta) + \alpha(2-\rho) - \sqrt{\begin{matrix}[(1-\rho)(4-3\beta)+\alpha(2-\rho)]^2 \\ -12\alpha(1-\beta)(1-\rho)(2-\rho)\end{matrix}}\right\}.$$

When $\gamma = 1$, the Gamma distribution is an exponential, and

$$p_m^+ = \frac{1}{2\alpha(1-\rho)}\{\alpha + (2-\beta)(1-\rho) - \sqrt{[\alpha+(2-\beta)(1-\rho)]^2 - 4\alpha(1-\beta)(1-\rho)}\}.$$

If $\gamma = 3$, then $E[S^2] = 4c_1^2/3$, and the first two moments of the Gamma distribution are the same as the uniform distribution with the range $[0, 2c_1]$.

Case 2 ($\gamma = 1/3$): In this case, $A = 0$, $B > 0$, $C > 0$. From (7)

$$\begin{aligned} C_{cache} < C_{basic} &\Leftrightarrow p_m < p_m^+ \\ &= \frac{(1+\rho)(1-\beta)}{4(1-\rho)+\alpha(1+\rho)}. \end{aligned} \quad (10)$$

Case 3 ($\gamma < 1/3$): In this case, $A > 0$, $B > 0$, and $C > 0$, and (7) is rewritten as

$$\begin{aligned} & Ap_m^2 + Bp_m - C < 0 \\ \Leftrightarrow\ & p_m < p_m^+ = \frac{-B + \sqrt{B^2 + 4AC}}{2A} \end{aligned} \quad (11)$$

or

$$p_m > \frac{-B - \sqrt{B^2 + 4AC}}{2A}. \quad (12)$$

Since the left-hand side of (12) is less than 0 and $p_m \geq 0$, (11) is the sufficient and necessary condition that the cache scheme outperforms the basic scheme in terms of the find cost.

Fig. 5 plots p_m^+ as the function of the offered load ρ to the HLR in the basic scheme (where $\beta = 0.2$). We shall study p_m^+ based on (10) because of its simplicity. The conclusions drawn from (10) also hold for (8) and (11).

From (10), it is easy to see that $p_m^+ \to 1$ as $\rho \to 1$. When $\rho = 1$, the HLR server saturates in the basic scheme, and the response time $c_2 \to \infty$. In such a case, if $p_m < 1$ in the cache scheme, $c_2 < \infty$ and the cache scheme always outperforms the basic scheme (see Fig. 5).

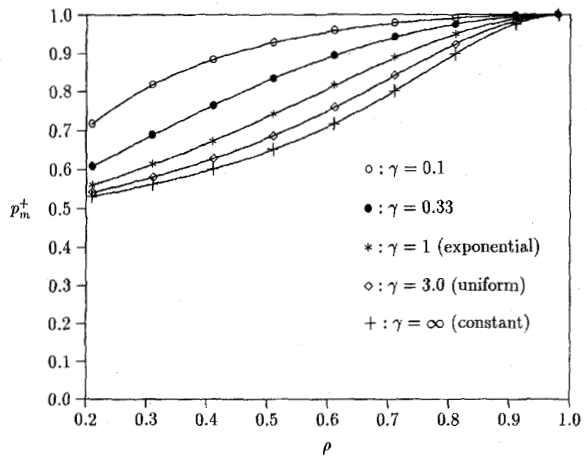

Fig. 5. The maximum p_m (i.e., p_m^+) which ensures that the find cost of the cache scheme is always lower than the basic scheme ($\beta = 0.2$).

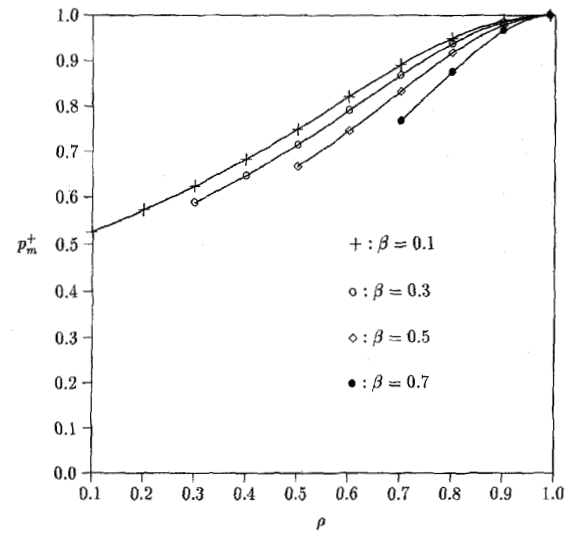

Fig. 6. The impact of β on p_m^+ (the HLR service times are exponentially distributed).

Fig. 5 also indicates that for a fixed $E[S] = c_1$, p_m^+ is a decreasing function of γ (and thus an increasing function of $E[S^2]$). In other words, for a fixed mean, if the variance of the HLR service times is large, then the cache scheme is more likely to outperform the basic scheme in terms of the **find** cost.

Replacing α by $\rho - \beta$, (10) can be rewritten as

$$p_m^+ = \frac{(1+\rho)(1-\beta)}{4(1-\rho)+(\rho-\beta)(1+\rho)}. \tag{13}$$

For a fixed $\rho < 1$ value, (13) is a decreasing function of β (see Fig. 6). Note that the β component of the offered load is due to the move operations. Thus, if a large portion of the traffic to the HLR is for the move operations (which cannot be reduced by the cache scheme), then a small p_m value (or a high cache hit ratio) is required for the cache scheme to outperform the basic scheme.

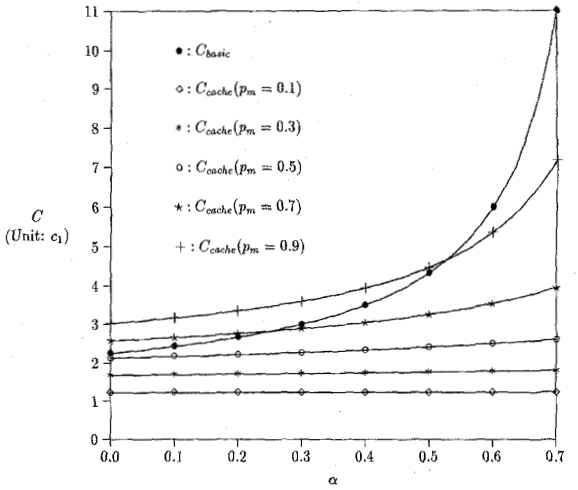

Fig. 7. The costs C_{basic} and C_{cache} (the HLR service times are exponentially distributed, and $\beta = 0.2$).

Based on (5) and (6), Fig. 7 plots the find cost for both the basic and the cache schemes. The information conveyed in the figure is similar to Fig. 5: The cache scheme is likely to outperform the basic scheme in terms of the find cost when α (or the call arrival rate λ) is large or p_m (the cache miss ratio) is small.

The cost c_2 merits further discussion. From (5) and (6), the mean waiting times of the HLR in the basic scheme and the cache scheme, respectively, are

$$c_{2,\,basic} - c_1 = \frac{(\gamma+1)\rho}{2\gamma(1-\rho)}\, c_1$$

and

$$c_{2,\,cache} - c_1 = \frac{(\gamma+1)(\alpha p_m+\beta)}{2\gamma(1-\beta-\alpha p_m)}\, c_1.$$

The numbers of the move messages sent to the HLR are the same for both the basic and the cache schemes. However, the find message traffic (to the HLR) for the basic scheme is larger than the cache scheme. Thus the waiting time for a move HLR query in the basic scheme is longer than the cache scheme. In other words, the move cost for the cache scheme is always no larger than the cost for the basic scheme. Fig. 8 illustrated the impacts of α, β, γ, and p_m on $c_2 - c_1$ (the waiting time at the HLR). The results are consistent with our intuition: For the move operation, the cache scheme is likely to significantly outperform the basic scheme when α (the call arrival rate λ to a portable) is large, β (the moving rate η of a portable) is small, γ is small (the variance of the HLR service time is large), or p_m (the cache miss ratio) is small.

III. Derivation of the Hit Ratio

In the previous papers [4], [12] we proposed a model to derive the hit ratio for Poisson call arrivals and the exponential portable residence times. This section extends the previous model to accommodate an arbitrary portable residence time distribution.

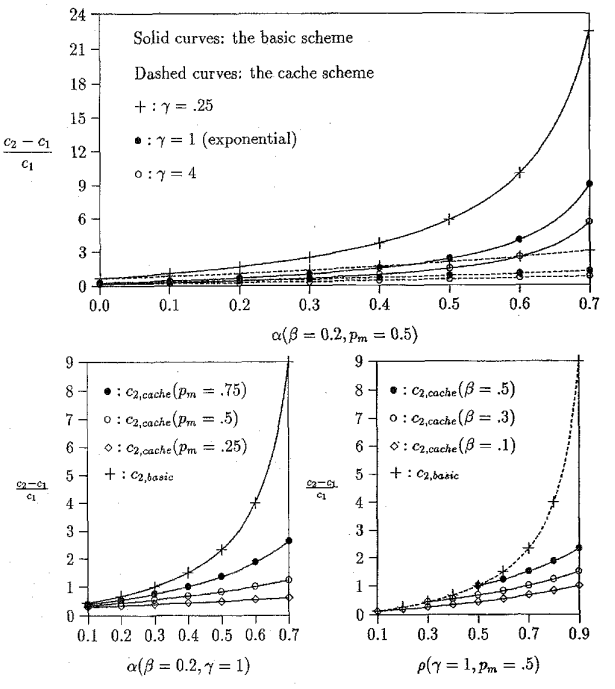

Fig. 8. Effects of different parameters on the mean HLR waiting time $c_2 - c_1$ (the HLR service time distribution is a Gamma with mean $E[S] = c_1$).

Let t_c be an independent and identically distributed random variable, which represents the time interval between two consecutive phone calls directed from an EO to a remote portable. We consider t_c with an exponential distribution (i.e., the call arrival stream is a Poisson stream) with mean $1/\lambda$. The density function is $f_c(t_c) = \lambda e^{-\lambda t_c}$. Let t_M be an independent and identically distributed random variable which represents the time interval that the portable resides in an RA. Let t_m be the time interval between the previous phone call to the portable and the time when the portable moves out of the RA. The relationship among t_c, t_M, and t_m is given in Fig. 9. Let $f_M(t)$ and $f_m(t)$ be the density function of t_M and t_m, respectively. Assume that the distribution function of t_M is *nonlattice*[1] and $\eta = 1/E[t_m] < \infty$. Because the call arrival stream forms a Poisson process, an incoming call is a random observer of t_M, and the density function for t_m can be derived using the standard "excess life" technique (see Proposition 3.4.5 in [15])

$$f_m(t_m) = \frac{1}{E[t_M]} \int_{t=t_m}^{\infty} f_M(t)\, dt. \tag{14}$$

For arrival calls originating from the same EO, after a previous call to the portable, the cache entry records the actual location of the portable when the call is issued. After a time period t_c, the next call is issued from the EO again. If $t_c < t_m$ then the portable has not moved out of the RA when the next call arrives. Thus, the cache contains the current portable location, and a cache hit occurs. If $t_c > t_m$ the portable has moved out of the RA, and a cache miss occurs. Thus, the hit ratio p_h (or

[1] A nonnegative random variable is said to be lattice if it takes on only integral multiples of some nonnegative number.

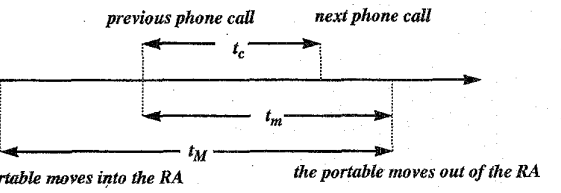

Fig. 9. The relationship among t_c, t_M, and t_m.

the probability of a cache hit) is

$$\begin{aligned} p_h &= \Pr[t_c < t_m] \\ &= \int_{t_m=0}^{\infty} \int_{t_c=0}^{t_m} f_m(t_m) f_c(t_c)\, dt_c\, dt_m \\ &= \int_{t_m=0}^{\infty} \eta[1 - F_M(t_m)] \int_{t_c=0}^{t_m} \lambda e^{-\lambda t_c}\, dt_c\, dt_m \\ &= \int_{t_m=0}^{\infty} \eta(1 - e^{-\lambda t_m})[1 - F_M(t_m)]\, dt_m \\ &= 1 - \frac{\eta}{\lambda} + \eta \int_{t_m=0}^{\infty} F_M(t_m) e^{-\lambda t_m}\, dt_m \\ &= 1 - \frac{\eta}{\lambda} - \left(\frac{\eta}{\lambda}\right) e^{-\lambda t_m} F_M(t_m)\Big|_{t_m=0}^{\infty} \\ &\quad + \left(\frac{\eta}{\lambda}\right) \int_{t_m=0}^{\infty} f_M(t_m) e^{-\lambda t_m}\, dt_m \\ &= 1 - \left(\frac{\eta}{\lambda}\right) [1 - f_M^*(s)|_{s=\lambda}] \end{aligned} \tag{15}$$

where

$$f_M^*(s) = \int_{t=0}^{\infty} f_M(t) e^{-st}\, dt$$

is the Laplace transform of $F_M(t)$. Using (15), the hit ratio can be easily computed for a general portable residence time distribution (the Laplace pairs for many functions are already available [16], [17]).

We consider three portable residence time distributions: the Gamma distribution, the uniform distribution, and the constant distribution. We are particularly interested in the Gamma distribution because in many cases it represents empirical reality more closely than the exponential distribution does. Consider the hazard rate function of the Gamma portable residence time distribution (4) with mean $1/\eta$

$$\begin{aligned} h_M(t) &= \frac{f_M(t)}{1 - F_M(t)} \\ &= \frac{(\gamma\eta)^{\gamma} t^{\gamma-1} e^{-(\gamma\eta)t}}{\Gamma(\gamma) - \int_{\tau=0}^{t} (\gamma\eta)^{\gamma} \tau^{\gamma-1} e^{-(\eta\gamma)\tau}\, d\tau}. \end{aligned}$$

Suppose that a portable has stayed in an RA for the time period t. Then $h_M(t)\, dt$ is the probability that the portable will move out of the RA in $(t, t + dt)$. If $\gamma = 1$, the Gamma is an exponential. For this case, $h_M(t) = f_M(t)$, and the portable residence time as measured by t does not influence the probability of movement in the next short time interval $(t, t + dt)$. On the other hand, $\gamma > 1$ implies that the hazard rate increases with age (see [18, p. 73] for the Gamma hazard rate function curves); i.e., at time t, if the portable has stayed

in an RA for a long time, it is more likely that the portable will move out of the RA during $(t, t + dt)$. While $\gamma < 1$ implies that the hazard rate decreases with age. For $\gamma < 1$, the Gamma right tail is longer than that of the exponential (extremely positively skewed), while if $\gamma > 1$ the positive skewness is less pronounced.

For a Gamma residence time distribution with mean $E[t_M] = 1/\eta$, we have

$$f^*_{\mathbf{M}}(s) = \left(\frac{\gamma\eta}{s+\gamma\eta}\right)^{\gamma}$$

and

$$p_h = 1 - \left(\frac{\eta}{\lambda}\right)\left[1 - \left(\frac{\gamma\eta}{\lambda+\gamma\eta}\right)^{\gamma}\right].$$

For a constant residence time $t_M = 1/\eta$

$$f^*_{\mathbf{M}}(s) = e^{-(s/\eta)}$$

and

$$p_h = 1 - \left(\frac{\eta}{\lambda}\right)[1 - e^{-(\lambda/\eta)}].$$

For a uniform residence time distribution with mean $E[t_M] = 1/\eta$

$$f^*_{\mathbf{M}}(s) = \frac{\eta[1 - e^{-(2s/\eta)}]}{2s}$$

and

$$p_h = 1 - \left(\frac{\eta}{\lambda}\right)\left\{1 - \frac{\eta[1 - e^{-(2\lambda/\eta)}]}{2\lambda}\right\}.$$

Fig. 10 plots p_h as a function of η (where the unit of η is λ). The figure shows the intuitive result that a high hit ratio p_h is expected if the portable mobility is low with respect to the call arrival rate (i.e., a small η/λ). The figure also indicates that for the same $E[t_M]$ value, p_h is large for a portable residence time distribution with high variation. The phenomenon is known as the "inspection paradox" [15]: In a line of renewal intervals (the portable residence intervals), it is more likely that an inspection point (i.e., a call arrival) falls in a larger interval as opposed to a shorter one. Consider four consecutive movements of a portable with eight calls to the portable during the four movements. For a portable residence time distribution with small variation, the lengths of the residence intervals are roughly the same, and the calls are likely to fall into the intervals evenly. In the worst case, every interval has two calls, and $p_h = 0.5$. In Fig. 10, the variance for the constant distribution is zero, and the lowest p_h values are observed. One the other hand, for a portable residence time distribution with large variation, calls are more likely to fall in the long intervals. In the best case, there is a long interval and three short intervals such that all eight calls fall in the long interval, and $p_h = 0.875$. In Fig. 10, the Gamma distribution with $\gamma = 0.25$ has a large variance, and the highest p_h values (among the other curves) are observed. In reality, the portable residence times tend to have large variation. For example, a PCS subscriber may travel from an initial RA to a destination RA, and there are several RA's between the initial and the destination RA's. The time intervals that the subscriber stays in the immediate RA's are much shorter than the intervals in the initial and the destination RA's. The result is that a high p_h is expected which allows the cache scheme to perform better.

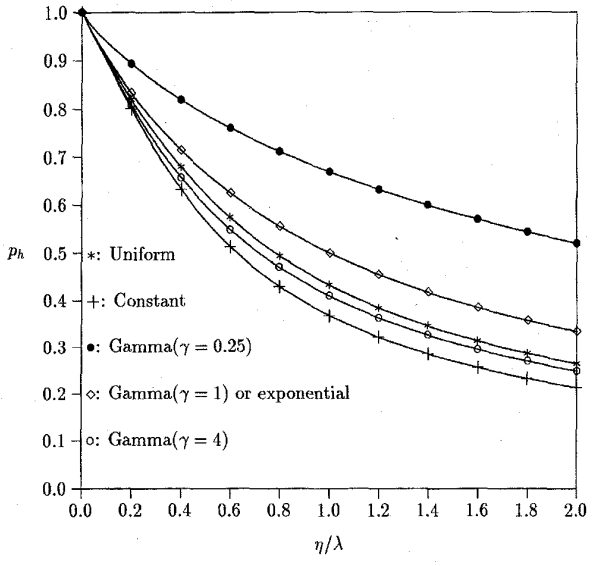

Fig. 10. The hit ratio (the mean portable residence time is $E[t_M] = 1/\eta$).

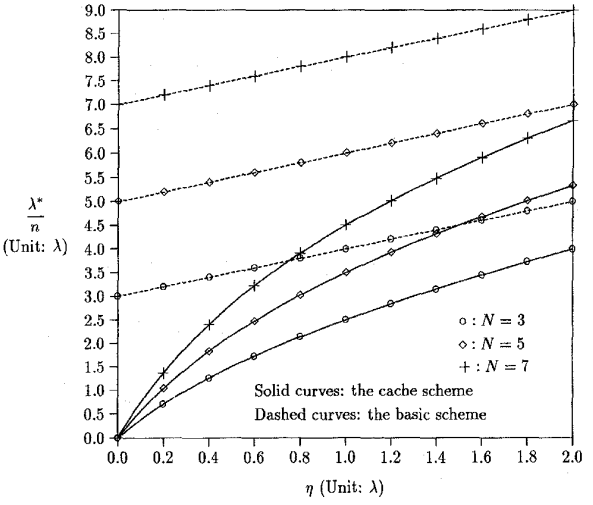

Fig. 11. Query traffic per portable to the HLR (the portable residence times are exponentially distributed).

The net query stream to an HLR is λ^*_{basic} and λ^*_{cache} in the basic and the cache schemes, respectively. Then the query traffic generated by a portable is

$$\frac{\lambda^*_{basic}}{n} = N\lambda + \eta$$

and

$$\frac{\lambda^*_{cache}}{n} = Np_m\lambda + \eta = N\eta[1 - f^*_M(s)|_{s=\lambda}] + \eta.$$

Fig. 11 plots λ^*/n as a function of η for the exponential portable residence time distribution. The figure illustrates the intuitive result that the cache scheme generates much less HLR traffic than the basic scheme when the portable mobility is low.

IV. Conclusion

The cache scheme has been proposed to reduce the location tracking overhead of a PCS network. In the previous papers [4], [12], we studied the cache scheme under the assumptions that the HLR access time is constant and the portable residence times have an exponential distribution. This paper generalizes the previous models by considering the queueing effect of the HLR (i.e., we model the HLR by an M/G/1 queue) and considering an arbitrary distribution for the portable residence times. The cache scheme was compared with a basic scheme (e.g., IS-41).

Our study indicated the following:

- The find cost for the cache scheme is lower than the basic scheme if 1) the net traffic to the HLR in the basic scheme saturates and the hit ratio in the cache scheme is larger than zero, 2) the portable mobility is low, and 3) for a fixed mean service time, the variance of the HLR service time distribution is large.
- The move cost for the cache scheme is always lower than the basic scheme. The improvement of the cache scheme over the basic scheme is significant for a large call arrival rate, low portable mobility, or if the variance of the HLR service time distribution is large for a fixed mean service time.
- It is intuitive that the cache hit ratio is high for a high call arrival rate and low portable mobility. For a fixed mean portable residence time, we showed that a higher cache hit ratio is observed for a portable residence distribution with larger variance.

Appendix A
Notation

This section lists the notation used in this paper.

1) $\alpha = Nn\lambda c_1$: the offered load (to the HLR in the basic scheme) due to the find operations.
2) $\beta = n\eta c_1$: the offered load (to the HLR in both the basic and the cache schemes) due to the move operations.
3) c_1: the mean service time of an HLR or a VLR.
4) c_2: the mean response (waiting + service) time of an HLR.
5) C_{basic}, C_{cache}: the find costs for the basic and the cache schemes, respectively.
6) η: the moving rate (or the mobility) of a portable.
7) $f_M^*(s) = \int_{t=0}^{\infty} f_M(t)e^{-st}\,dt$: the Laplace transform of $F_M(t)$.
8) $f_c(t)$: the density function for the random variable t_c.
9) $f_m(t)$: the density function for the random variable t_m.
10) $f_M(t)$: the density function for the random variable t_M.
11) $F_M(t) = \int_{\tau=0}^{t} f_M(\tau)\,d\tau$: the distribution function for the random variable t_M.
12) γ: the shape parameter of a Gamma distribution.
13) λ^*: the arrival rate of queries to an HLR.
14) λ: the rate of call termination.
15) μ: the scale parameter of a Gamma distribution.
16) n: the number of portables registered in an HLR.
17) N: the expected number of EO's that issue calls to a portable.
18) p_h: the hit ratio or the probability that the current location of a portable is found in the cache of an EO.
19) $p_m = 1 - p_h$: the miss ratio.
20) p_m^+: an upper bound of p_m such that the find cost of the cache scheme is lower than the basic scheme. In other words, if $p_m < p_m^+$, then the find cost for the cache scheme is lower than the basic scheme.
21) $\rho = \alpha + \beta$: the offered load to an HLR in the basic scheme.
22) S: a random variable that represents the service times of an HLR.
23) t_c: an independent and identically distributed random variable that represents the time interval between two consecutive calls directed from an EO to a portable. $E[t_c] = 1/\lambda$.
24) t_m: the time interval between the previous phone call to a portable and the time when the portable moves out of the RA.
25) t_M: an independent and identically distributed random variable that represents the portable residence times. $E[t_M] = 1/\eta$.

Acknowledgment

The authors would like to thank J. R. Cruz and the reviewers for their valuable comments.

References

[1] Bellcore, "Generic criteria for version 0.1 wireless access communications systems (WACS) and supplement," Bellcore, Tech. Rep. TR-INS-001313, Issue 1, 1994.
[2] EIA/TIA, "Cellular radio-telecommunications intersystem operations: Automatic roaming," EIA/TIA, Tech. Rep. IS-41.3-B, 1991.
[3] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications.* Palaiseau, France: M. Mouly, 1992.
[4] R. Jain, Y.-B. Lin, C. N. Lo, and S. Mohan, "A caching strategy to reduce network impacts of PCS," *IEEE J. Select. Areas Commun.,* vol. 12, no. 8, pp. 1434–1445, 1994.
[5] Y.-B. Lin and A. Noerpel, "Implicit deregistration in a PCS network," *IEEE Trans. Veh. Technol.,* vol. 43, no. 4, pp. 1006–1010, 1994.
[6] Y.-B. Lin and S.-Y. Hwang, "Deregistration strategies for PCS networks," submitted for publication to *IEEE/ACM Trans. Networking,* 1993.
[7] Y.-B. Lin and S. K. DeVries, "PCS network signaling using SS7," *IEEE Personal Commun. Mag.,* pp. 44–55, June 1995.
[8] C. N. Lo, R. S. Wolff, and R. C. Bernhardt, "Expected network database transaction volume to support personal communication services," in *First Int'l. Conf. Universal Personal Commun.,* 1992.
[9] T. Imielinski and B. R. Badrinath, "Querying in highly mobile distributed environment," in *Proc. 18th VLDB Conf.,* 1992.
[10] R. Jain and Y.-B. Lin, "Performance modeling of an auxiliary user location strategy in a PCS network," *ACM-Baltzer Wireless Networks,* vol. 1, pp. 197–210, 1995.
[11] Y.-B. Lin, "Reducing location update cost in a PCS network," submitted for publication, 1993.
[12] ——, "Determining the user locations for personal communications networks," *IEEE Trans. Veh. Technol.,* vol. 43, no. 3, pp. 466–473, 1994.
[13] W. C. Wong, "Packet reservation multiple access in a metropolitan microcellular radio environment," *IEEE J. Select. Areas Commun.,* vol. 11, no. 6, pp. 918–925, 1993.
[14] L. Kleinrock, *Queueing Systems: Vol. I–Theory.* New York: Wiley, 1976.
[15] S. M. Ross, *Stochastic Processes.* Wiley, 1983.
[16] E. J. Muth, *Transform Methods with Applications to Engineering and Operations Research.* Prentice-Hall, 1977.
[17] E. J. Watson, *Laplace Transforms and Applications.* Birkhauserk, 1981.
[18] N. A. J. Hastings and J. B. Peacock, *Statistical Distributions.* New York: Wiley, 1975.

Yi-Bing Lin received the B.S.E.E. degree from National Cheng Kung University in 1983, and the Ph.D. degree in Computer Science from the University of Washington in 1990.

From 1990 to 1995, he was with the Applied Research Area at Bell Communications Research (Bellcore), Morristown, NJ. In 1995 he was appointed full Professor in the Department of Computer Science and Information Engineering, National Chiao Tung University. His current research interests include design and analysis of personal communications services network, distributed simulation, and performance modeling.

Dr. Lin is a subject area Editor of the *Journal of Parallel and Distributed Computing,* an Associate Editor of the *International Journal in Computer Simulation,* an Associate Editor of *ACM Transactions on Modeling and Computer Simulation* magazine, a member of the editorial board of the *International Journal of Communications,* a member of the editorial board of *Computer Simulation Modeling and Analysis,* the Guest Editor of *ACM-Baltzer Wireless Networks,* the Guest Editor of IEEE TRANSACTIONS ON COMPUTERS, Program Chair for the Eighth Workshop on Distributed and Parallel Simulation, and General Chair for the Ninth Workshop on Distributed and Parallel Simulation.

Shu-Yuen Hwang received the B.S. and M.S. degrees in electrical engineering from National Taiwan University in 1981 and 1983, respectively, and the M.S. and Ph.D. degrees in computer science from the University of Washington in 1987 and 1989, respectively.

He is currently Professor and Director of the Department of Computer Science and Information Engineering, National Chiao-Tung University. His research interests include computer vision, artificial intelligence, computer simulation, and mobile computing.