Analysis and Simulation of a Cache-based Auxiliary User Location Strategy for PCS

Harry Harjono, Ravi Jain and Seshadri Mohan*

Bellcore, 445 South Street, Morristown, NJ 07960-6438

Abstract: The strategies commonly proposed to locate mobile users maintain a system of home and visited databases (Home Location Registers or HLR, and Visitor Location Registers or VLR) to keep track of user locations. We have previously proposed an auxiliary location strategy to augment the basic strategy; the auxiliary strategy is based on locally caching and reusing previously obtained user location information. This paper compares the basic and auxiliary strategies based on both analysis and simulation.

1 Introduction

The vision of nomadic personal communications is the ubiquitous availability of services to facilitate exchange of information (voice, data, video, image, etc.) between nomadic end users independent of time, location, access arrangements, or equipment capabilities. To realize this vision, it is necessary to locate users that move from place to place. The strategies commonly proposed are two-level hierarchical strategies, which maintain a system of home and visited databases (Home Location Registers or HLR, and Visitor Location Registers or VLR) to keep track of user locations. Studies indicate that the signaling traffic and database query and update rates associated with user mobility are likely to grow to levels well in excess of that associated with conventional calls [1], [2]. In [8], we have proposed an auxiliary location strategy to augment the basic two-level location strategy such as that proposed in IS-41 [4] and to significantly reduce the signaling messages and the query and update rates to databases. This scheme, which we call the cache-based auxiliary location scheme, relies on the observation that if a user receives many calls from a particular registration area between successive moves, then it should be possible to save and reuse that user's location information, obtained previously, rather than query the user's HLR. This should in turn result in fewer signaling messages and database queries, and reduced call setup time. This paper reports the results, based on both analysis and simulation.

2 A Reference Model for PCS

Figure 1 illustrates the reference model used for the comparative analysis. The model assumes that the Home Location Register (HLR) resides in a *Service Control Point* (SCP) connected to a *Regional Signal Transfer Point* (RSTP). The SCP is a storehouse of the AIN service logic, i.e., functionality used to perform the processing required to provide advanced services, such as speed calling, outgoing call screening etc., in the AIN architecture. The RSTP and the *Local STP* (LSTP) are packet switches, connected together by various links such as A-links or D-links, that perform the signaling and routing functions of the Signaling System 7 (SS7) network. Such functions include, for example, Global Title Translation for routing messages between the AIN switching system, which is also referred to as the *Service Switching Point* (SSP), and SCP and IS-41 messages [4]. Several SSPs may be connected to an LSTP.

For our purposes, the geographical area served by a PCS system is partitioned into a number of ***radio port coverage areas*** (or ***cells*** in cellular terms) each of which is served by a ***base station*** which communicates with mobile stations in that cell. A ***registration area*** (also known in the cellular world as ***location area***) is composed of a number of cells. The base stations of all cells in a registration area are connected by wireline links to a ***mobile switching center*** (MSC). We assume that each registration area is served by a single VLR. The MSC of a registration area is responsible for maintaining and accessing the VLR, and switching between radio ports. The VLR associated with a registration area is responsible for maintaining a subset of the user information contained in the HLR.

The basic location strategy is easily explained as follows. When a call is placed to a PCS user, the switch detects the call. If the called user is in the same registration area as the calling user, the called's record exists in the switch (i.e, in the VLR) and no further database lookups are needed. Otherwise, the switch queries the called party's HLR, which in turn requests the called user's VLR. The VLR returns the user's location to the HLR, which forwards it the calling switch. This is basically the IS-41 strategy described in [4] recaptured in terms of the reference model in Figure 1.

* The work was done while the author was with Bellcore during June - August 1993; the author is with Columbia University

3 A Cache-based Auxiliary Location Strategy

In the basic location strategy, current practices require that every call to user in a different registration area requires an access to the HLR. If business considerations permit, other efficient strategies could be used if the called party never moves out of the registration area or moves very infrequently. For example, the VLR information the first time a call is made to a user and cache that information in the local switch. Subsequent calls to that user from any user within the registration area covered by the switch will use the VLR information in the cache, thereby avoiding a location query to the HLR. A cache 'miss' occurs if the called user has moved out. The VLR would return a message indicating that the user is not in its registration area. In this case, the switch will proceed to query the HLR as in the basic location strategy. We see that the 'cache-based' scheme could result in substantial savings over the basic scheme if the cache 'hit' ratio is large, where cache hit ratio is defined as the number of times a user is found in the registration area pointed to by the cache to the total number of cache inquiries for that user. The hit ratio can be related to the users call and mobility pattern, quantified by the user's call-to-mobility ratio. We define the call-to-mobility ratio as the average number calls made to a user per unit time divided by the average number of moves between registration areas made by that user per unit time.

4 Analysis

Reference [8] analyzes the basic strategy based on the simple fluid flow model for mobility described in [7] and the following set of assumptions.

- 128 total registration areas;
- square registration area of size $(7.575\text{km})^2 = 57.4\text{sq. km}$, with border length L = 30.3km;
- mean density of terminals = ρ = 390 per sq. km
- total number of terminals = $128 \times 57.4 \times 390$ = 2.87 million
- average call origination rate = average call termination (delivery) rate = 1.4 / hr / terminal;
- Average speed of a mobile, v = 5.6 km per hour.;
- fluid flow mobility model

The results are summarized in the following table.

Table 1 IS-41 Query and Update Rates (per second) to HLR and VLR

Activity	HLR Update	VLR Update	HLR Queries	VLR Queries
Mobility-related activities at registration	749	5.85	749	5.85
Mobility-related activities at deregistration		5.85		
Call Origination				8.7

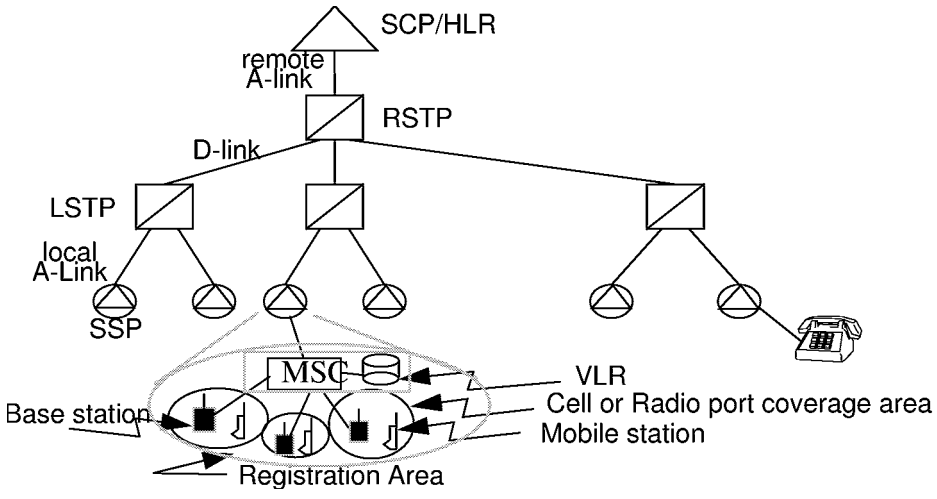

Figure 1 Example of a Reference Model for PCS.

Table 1 IS-41 Query and Update Rates (per second) to HLR and VLR

Activity	HLR Update	VLR Update	HLR Queries	VLR Queries
Call Delivery			1116	8.7
Total (per VLR)	749	11.7	1865	23.25
Total (Network)	749	1497.6	1865	2976

The cache-based scheme proposed in [8] relies on the observation that if a user receives many calls from a particular registration area between successive moves, then it should be possible to save and reuse that user's location information, obtained previously, rather than query the user's HLR. Consequently, of the four quantities listed in Table 1, the cache-based scheme will only affect the rate of HLR queries that take place at call delivery. The reduction in HLR queries is related to the cache hit ratio at the calling switch. For example (for the example scenario assumed at the beginning of the section), let us assume that 60% of the users receive their calls uniformly from the 9 registration areas and the VLRs at these registration areas experience a cache hit probability of 80% for these users and that caching is not applied to the remaining 40% of the users. The HLR query rate will reduce to

$$\frac{(0.5 \times 2.87 \times 10^6) \times 1.4 \times 1.2}{3600} = 670 \text{persec.}$$

as compared to HLR query rate of 1116 per second for the basic scheme.

For any cache hit ratio, p, greater than zero, the caching scheme will require less HLR queries compared to the basic scheme. Should caching be then used regardless of the user call and mobility pattern? Recall that, during call origination, the originating switch initially uses the cache information to determine the VLR that served the called party the last time a call was made to that user and, should a cache miss occur, the switch must then be invoke basic location strategy. Consequently, there is an overhead involved for every cache miss. The switch should therefore use the cache information only for those users with large enough hit ratio exceeding a certain threshold. We define r the threshold P_T as and derive an expression for it in [8] in terms of the cost of processing at the various network elements and of transmitting messages over the different links.

$$P_T = \frac{C_{\text{Hit}}}{C_{\text{Basic}}} = \frac{4A_l + 4D + 4L + 2R + V_Q - q(4D + 2L + 2R)}{4A_l + 4D + 4A_r + 4L + 4R + H_Q + V_Q} \quad \text{(EQ 1)}$$

where

C_{Hit} = Cost of the caching strategy when a hit occurs

C_{Basic} = Cost of basic strategy

A_l = Cost of transmitting a location request response on an A-link between SSP and LSTP

D = Cost of transmitting a location request a D link

A_r = Cost of transmitting a location response on A link between RSTP and STP

L = Cost of processing a location request by LSTP

R = Cost of processing a location request by RSTP

H_Q = Cost of a query to the HLR to get the VLR address

V_Q = Cost of a VLR query to obtain the current routing address

(Eq 1) specifies that the hit ratio for any given user measured at a switch must exceed the threshold given by p_T for the caching scheme to start producing savings over the basic scheme. The hit ratio for any given user may vary from one switch to another depending on the volume of calls made to that user from the switch. To evaluate (Eq 1), we assume that each of the parameters in the equation dominates the threshold independently. resulting in Table 2. The columns under HLR query rate are computed assuming values given at the beginning of the section for the call rate and the total number of users served by the HLR. We also assume that the mobility rate is such that the hit ratio p for all the users measured at any switch equals or exceeds the threshold under column 2. If the hit ratio exceeds the threshold the query rate will decrease. If the hit ratio exceeds the threshold only for a fraction of the users then the savings in the query rate will correspondingly decrease.

We infer from Table 2 caching will not be fruitful for users that move if using a local link or querying a VLR is the dominating factor (the cases corresponding to the first and the last rows in the table). On the other hand, for those users do not move (for example, in a fixed telephone network) caching will be beneficial. When the remote link cost (row 2) or the cost of querying the HLR (second row from last) is the dominating factor (row 2), caching will be worthwhile for any user with a hit ratio greater than 0. The remaining three cases listed in the table indicate that caching will be beneficial only to those users that receive very many calls between moves or that have a large call-to-mobility ratio.

Table 2 HLR Query Rate and Minimum Hit Ratios for Each Dominant Parameter in (Eq 1) for q = 0.04

Dominant Cost Term	Hit Ratio Threshold p_T	HLR Query Rate for p = 0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
A_l	1	1116	1116	1116	1116	0
A_r	0	1116	837	558	279	0
D	$1-q$	1116	1116	1116	1116	44.64
L	$1-\frac{q}{2}$	1116	1116	1116	1116	22.32
R	$1-\frac{q}{2}$	1116	1116	1116	1116	22.32
H_Q	0	1116	837	558	279	0
V_Q	1	1116	1116	1116	1116	0

5 Simulation Results

This section summarizes the simulation set up and results obtained from the set up comparing the basic and the caching strategies. The reference model shown in Figure 1 was specialized with the following parameters.

- Number of SCPs = 1; Number of RSTPs = 1; Number of LSTPs = 3; Number of SSPs = 9
- Number of Users = 900
- SCP, RSTP, LSTP, and SSP are modeled as queues with the following average service time in milliseconds; the numbers within the parentheses represent standard deviation of the service time. The service time distributions are assumed Gaussian.
- Service Time: SCP = 100(50), RSTP, LSTP = 10.6 (4.2), SSP = 42 (16)
- Mean call generation rate = 3 calls/hr./user; time between calls is exponentially distributed with mean 1/3 hr.

Figure 2 summarizes the results for the caching scheme. The results are normalized with respect to the HLR queries for the basic scheme. The parameter T in the figure is the threshold that each SSP maintains. The SSP applies caching to only those users for whom the hit ratio measured by the SSP exceeds the threshold T maintained by the SSP. Thus T determines how aggressively we will use caching. Observe that the caching scheme with T set to 1 reduces to the basic strategy. With T set to 0, the caching scheme is applied to all users regardless of the CMR or the hit ratio at each SSP. The figure plots the normalized HLR query rate versus CMR with T as the free parameter, which is varied from 0 to 1 in increments of 0.1. With each curve, the mean call rate is fixed at 3 calls/hr/user and the user's average mobility rate is varied. Mobility rate is defined as the number of user moves between registration areas per unit time. We assume that the time between moves is exponentially distributed with average equal to the inverse of the mobility rate. The figure shows that the HLR query rate is reduced compared to the basic scheme for any CMR greater than 0 and for any threshold. In general, the savings are greater the smaller the threshold. Thus the smallest T, T= 0, yields the maximum possible benefit in terms of HLR query rate. As CMR increases arbitrarily large, that is, the users become less and less mobile and eventually become fixed users, the HLR query rate tends towards 0. If HLR query rate is the criterion to be optimized, then the threshold should always be set to 0. However, this would increase the call setup time for those users that are highly mobile or have a low CMR. This is illustrated in Figure 3.

Figure 3 shows the normalized call setup time versus CMR, again with T as the free parameter. We observe that, in general, for large enough CMR, the reduction in setup time is the greatest when caching is always used (that is, T = 0). We also observe that, for T = 1, the caching scheme becomes the basic scheme, with call setup time equal to that for the basic scheme. Interestingly, the figure illustrates that at low CMR, the threshold should be made sufficiently large so that frequent cache misses could be avoided. For example, for T = 0, the setup time can increase by 7% over that for the basic scheme.

Figure 3 suggests that each SSP should possibly adapt the threshold relative to the CMR. This implies that knowledge of CMR should be made available to the SSPs. Such knowledge of CMR could be downloaded to SSPs from a central source periodically. Alternatively, each SSP could measure locally the CMR (local CMR or LCMR) and adapt the threshold based on the LCMR. Techniques for measurement of the LCMR and its relationship to the threshold are beyond the scope of this paper and are explored in [8].

6 Conclusions

We have analyzed and simulated a cache-based auxiliary strategy to augment basic user location strategies such as the one specified in IS-41. The caching strategy achieves reduced HLR query rate and call setup time compared to the basic strategy by the use of increased storage and local processing at the switches. Depending on the CMR, the call setup time with caching can be reduced by as much as 35% compared to the basic scheme and the HLR query volume reduced by up to 70%. However, the call setup time and the HLR query rate cannot be simultaneously reduced under all conditions. The trade-off between the two is determined by the cache threshold maintained by each SSP. Schemes for adapting the threshold based on locally observed call-to-mobility ratio is an interesting topic for future study.

Acknowledgments:

The authors would like to thank D. O. Hakim, A.K. Knapp, and M. Kramer for reviewing the document.

References

[1] K. Meiller-Hellstern and E. Alonso, "The Use of SS7 and GSM to Support High Density Personal Communications," Proc. ICC '92.

[2] C. N. Lo, R. S. Wolff and R. C. Bernhardt, "Expected Network Database Transaction Volume to Support Personal Communications Services", *First International Conference on Universal Personal Communications*, Dallas, Texas, September 1992.

[3] R.K. Berman and J.H. Brewster, "Perspectives on the AIN Architecture," *IEEE Communications Magazine*, Feb. 1992, Vol.1, No.2, pp. 27 - 32.

[4] EIA/TIA IS-41.3 (Revision B), Cellular Radiotelecommunications Intersystem Operations: Automatic Roaming, July 1991.

[5] Personal Communications Services (PCS) Network Access Service Alternatives, *Bellcore Special Report*, SR-INS-002245, Issue 1, April 1992.

[6] PCS Network Access Services to PCS Providers, *Bellcore Special Report* SR-TSV-002459, Issue 2, Oct. 1993..

[7] R. Thomas, H. Gilbert, and G. Mazziotto, "Influence of the Mobile Station on the Performance of a Radio Mobile Cellular Network," Proc. 3rd Nordic Seminar, Paper 9.4, Copenhagen, September 1988.

[8] R. Jain, J. Lin, C.N. Lo, S. Mohan, "A Caching Strategy to Reduce Network Impacts of PCS," submitted for publication, Sep. 1993.

[9] S. Mohan and R. Jain, "Two User Location Strategies for Personal Communications Services," IEEE Personal Communications, Feb. 1994.

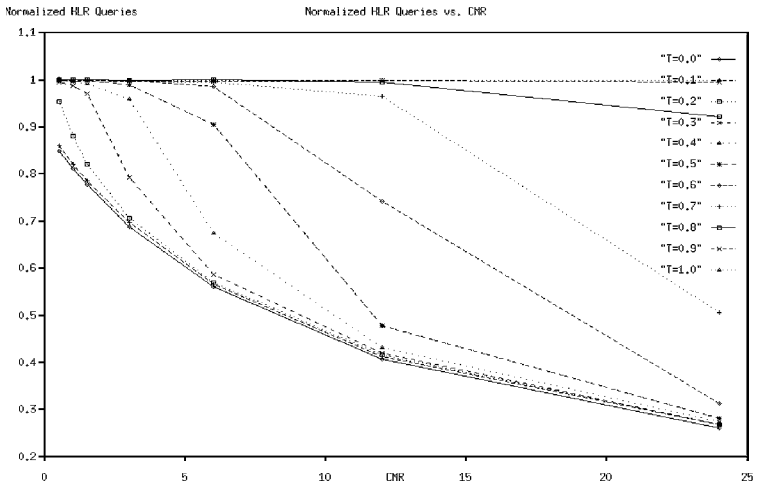

Figure 2 Normalized HLR Queries versus CMR

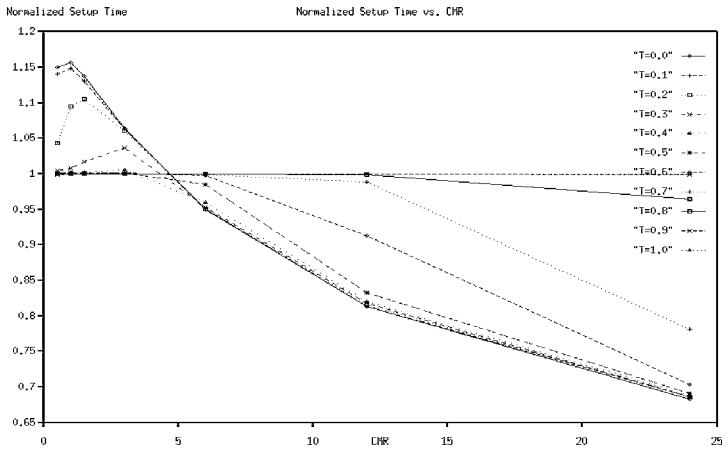

Figure 3 Normalized Setup Time versus CMR