Jain, R.; Lin, Y. & Mohan, S. “Location Strategies for Personal Communications Services”
Mobile Communications Handbook
Ed. Suthan S. Suthersan
Boca Raton: CRC Press LLC, 1999

Location Strategies for Personal Communications Services

Ravi Jain
Bell Communications Research

Yi-Bing Lin
Bell Communications Research

Seshadri Mohan[1]
Bell Communications Research

20.1 Introduction
20.2 An Overview of PCS
Aspects of Mobility—Example 20.1 • A Model for PCS
20.3 IS-41 Preliminaries
Terminal/Location Registration • Call Delivery
20.4 Global System for Mobile Communications
Architecture • User Location Strategy
20.5 Analysis of Database Traffic Rate for IS-41 and GSM
The Mobility Model for PCS Users • Additional Assumptions • Analysis of IS-41 • Analysis of GSM
20.6 Reducing Signalling During Call Delivery
20.7 Per-User Location Caching
20.8 Caching Threshold Analysis
20.9 Techniques for Estimating Users' LCMR
The Running Average Algorithm • The Reset-K Algorithm • Comparison of the LCMR Estimation Algorithms
20.10 Discussion
Conditions When Caching Is Beneficial • Alternative Network Architectures • LCMR Estimation and Caching Policy
20.11 Conclusions
Acknowledgment
References

[1]Address correspondence to: Seshadri Mohan, MCC-1A216B, Bellcore, 445 South St, Morristown, NJ 07960; Phone: 973-829-5160, Fax: 973-829-5888, e-mail: smohan@bellcore.com.+

20.1 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, or access arrangements. 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 mobility databases, home location registers (HLR) and visitor location resisters (VLR), to keep track of user locations. Two standards exist for carrying out two-level hierarchical strategies using HLRs and VLRs. The standard commonly used in North America is the EIA/TIA Interim Standard 41 (IS 41) [6] and in Europe the Global System for Mobile Communications (GSM) [15, 18]. In this chapter, we refer to these two strategies as *basic* location strategies.

We introduce these two strategies for locating users and provide a tutorial on their usage. We then analyze and compare these basic location strategies with respect to load on mobility databases and signalling network. Next we propose an auxiliary strategy, called the *per-user caching* or, simply, the *caching* strategy, that augments the basic location strategies to reduce the signalling and database loads.

The outline of this chapter is as follows. In Section 20.2 we discuss different forms of mobility in the context of personal communications services (PCS) and describe a reference model for a PCS architecture. In Sections 20.3 and 20.4, we describe the user location strategies specified in the IS-41 and GSM standards, respectively, and in Section 20.5, using a simple example, we present a simplified analysis of the database loads generated by each strategy. In Section 20.6, we briefly discuss possible modifications to these protocols that are likely to result in significant benefits by either reducing query and update rate to databases or reducing the signalling traffic or both. Section 20.7 introduces the caching strategy followed by an analysis in the next two sections. This idea attempts to exploit the spatial and temporal locality in calls received by users, similar to the idea of exploiting locality of file access in computer systems [20]. A feature of the caching location strategy is that it is useful only for certain classes of PCS users, those meeting certain call and mobility criteria. We encapsulate this notion in the definition of the user's call-to-mobility ratio (CMR), and local CMR (LCMR), in Section 20.8. We then use this definition and our PCS network reference architecture to quantify the costs and benefits of caching and the threshold LCMR for which caching is beneficial, thus characterizing the classes of users for which caching should be applied. In Section 20.9 we describe two methods for estimating users' LCMR and compare their effectiveness when call and mobility patterns are fairly stable, as well as when they may be variable. In Section 20.10, we briefly discuss alternative architectures and implementation issues of the strategy proposed and mention other auxiliary strategies that can be designed. Section 20.11 provides some conclusions and discussion of future work.

The choice of platforms on which to realize the two location strategies (IS-41 and GSM) may vary from one service provider to another. In this paper, we describe a possible realization of these protocols based on the advanced intelligent network (AIN) architecture (see [2, 5]), and signalling system 7 (SS7). It is also worthwhile to point out that several strategies have been proposed in the literature for locating users, many of which attempt to reduce the signalling traffic and database loads imposed by the need to locate users in PCS.

20.2 An Overview of PCS

This section explains different aspects of mobility in PCS using an example of two nomadic users who wish to communicate with each other. It also describes a reference model for PCS.

20.2.1 Aspects of Mobility—Example 20.1

PCS can involve two possible types of mobility, terminal mobility and personal mobility, that are explained next.

Terminal Mobility: This type of mobility allows a terminal to be identified by a unique terminal identifier independent of the point of attachment to the network. Calls intended for that terminal can therefore be delivered to that terminal regardless of its network point of attachment. To facilitate terminal mobility, a network must provide several functions, which include those that locate, identify, and validate a terminal and provide services (e.g., deliver calls) to the terminal based on the location information. This implies that the network must store and maintain the location information of the terminal based on a unique identifier assigned to that terminal. An example of a terminal identifier is the IS-41 EIA/TIA cellular industry term mobile identification number (MIN), which is a North American Numbering Plan (NANP) number that is stored in the terminal at the time of manufacture and cannot be changed. A similar notion exists in GSM (see Section 20.4).

Personal Mobility: This type of mobility allows a PCS user to make and receive calls independent of both the network point of attachment and a specific PCS terminal. This implies that the services that a user has subscribed to (stored in that user's service profile) are available to the user even if the user moves or changes terminal equipment. Functions needed to provide personal mobility include those that identify (authenticate) the end user and provide services to an end user independent of both the terminal and the location of the user. An example of a functionality needed to provide personal mobility for voice calls is the need to maintain a user's location information based on a unique number, called the universal personal telecommunications (UPT) number, assigned to that user. UPT numbers are also NANP numbers. Another example is one that allows end users to define and manage their service profiles to enable users to tailor services to suit their needs. In Section 20.4, we describe how GSM caters to personal mobility via smart cards.

For the purposes of the example that follows, the terminal identifiers (TID) and UPT numbers are NANP numbers, the distinction being TIDs address terminal mobility and UPT numbers address personal mobility. Though we have assigned two different numbers to address personal and terminal mobility concerns, the same effect could be achieved by a single identifier assigned to the terminal that varies depending on the user that is currently utilizing the terminal. For simplicity we assume that two different numbers are assigned.

Figure 20.1 illustrates the terminal and personal mobility aspects of PCS, which will be explained via an example. Let us assume that users Kate and Al have, respectively, subscribed to PCS services from PCS service provider (PSP) A and PSP B. Kate receives the UPT number, say, 500 111 4711, from PSP A. She also owns a PCS terminal with TID 200 777 9760. Al too receives his UPT number 500 222 4712 from PSP B, and he owns a PCS terminal with TID 200 888 5760. Each has been provided a personal identification number (PIN) by their respective PSP when subscription began. We assume that the two PSPs have subscribed to PCS access services from a certain network provider such as, for example, a local exchange carrier (LEC). (Depending on the capabilities of the PSPs, the access services provided may vary. Examples of access services include translation of UPT number to a routing number, terminal and personal registration, and call delivery. Refer to Bellcore, [3], for further details). When Kate plugs in her terminal to the network, or when she activates it, the terminal registers itself with the network by providing its TID to the network. The network creates an entry for the terminal in an appropriate database, which, in this example, is entered in the terminal mobility database (TMDB) A. The entry provides a mapping of her terminal's TID, 200 777 9760, to a routing number (RN), RN1. All of these activities happen without Kate being aware of them. After activating her terminal, Kate registers herself at that terminal by entering her UPT number (500 111 4711) to inform the network that all calls to her UPT number are to be delivered to her

at the terminal. For security reasons, the network may want to authenticate her and she may be prompted to enter her PIN number into her terminal. (Alternatively, if the terminal is equipped with a smart card reader, she may enter her smart card into the reader. Other techniques, such as, for example, voice recognition, may be employed). Assuming that she is authenticated, Kate has now registered herself. As a result of personal registration by Kate, the network creates an entry for her in the personal mobility database (PMDB) A that maps her UPT number to the TID of the terminal at which she registered. Similarly, when Al activates his terminal and registers himself, appropriate entries are created in TMDB B and PMDB B. Now Al wishes to call Kate and, hence, he dials Kate's UPT number (500 111 4711). The network carries out the following tasks.

1. The switch analyzes the dialed digits and recognizes the need for AIN service, determines that the dialed UPT number needs to be translated to a RN by querying PMDB A and, hence, it queries PMDB A.
2. PMDB A searches its database and determines that the person with UPT number 500 111 4711 is currently registered at terminal with TID 200 777 9760.
3. PMDB A then queries TMDB A for the RN of the terminal with TID 200 777 9760. TMDB A returns the RN (RN1).
4. PMDB A returns the RN (RN1) to the originating switch.
5. The originating switch directs the call to the switch RN1, which then alerts Kate's terminal. The call is completed when Kate picks up her terminal.

Kate may take her terminal wherever she goes and perform registration at her new location. From then on, the network will deliver all calls for her UPT number to her terminal at the new location. In fact, she may actually register on someone else's terminal too. For example, suppose that Kate and Al agree to meet at Al's place to discuss a school project they are working on together. Kate may register herself on Al's terminal (TID 200 888 9534). The network will now modify the entry corresponding to 4711 in PMDB A to point to B 9534. Subsequent calls to Kate will be delivered to Al's terminal.

The scenario given here is used only to illustrate the key aspects of terminal and personal mobility; an actual deployment of these services may be implemented in ways different from those suggested here. We will not discuss personal registration further. The analyses that follow consider only terminal mobility but may easily be modified to include personal mobility.

20.2.2 A Model for PCS

Figure 20.2 illustrates the reference model used for the comparative analysis. The model assumes that the 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 (see Bellcore, [2] and Berman and Brewster, [5]). The RSTP and the local STP (LSTP) are packet switches, connected together by various links such A links or D links, that perform the signalling functions of the 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 [6]. Several SSPs may be connected to an LSTP.

The reference model in Fig. 20.2 introduces several terms which are explained next. We have tried to keep the terms and discussions fairly general. Wherever possible, however, we point to equivalent cellular terms from IS-41 or GSM.

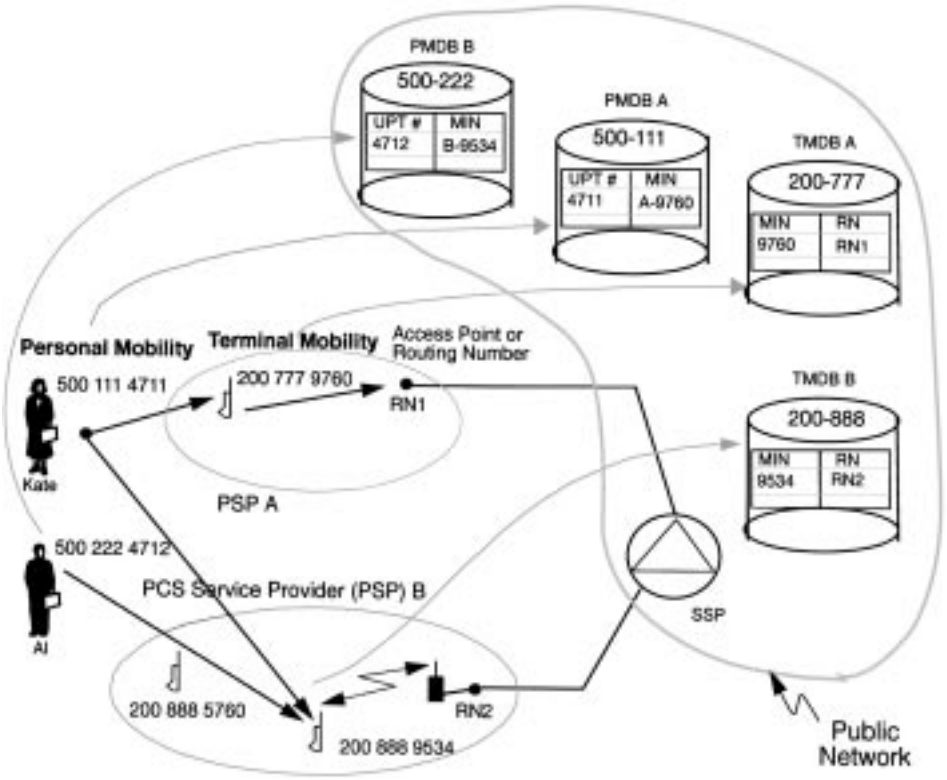


FIGURE 20.1: Illustrating terminal and personal mobility.

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 radio port (or, equivalently, base station) that communicates with PCS terminals 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 for 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.

Terminal registration process is initiated by terminals whenever they move into a new registration area. The base stations of a registration area periodically broadcast an identifier associated with that area. The terminals periodically compare an identifier they have stored with the identifier to the registration area being broadcast. If the two identifiers differ, the terminal recognizes that it has moved from one registration area to another and will, therefore, generate a registration message. It also replaces the previous registration area identifier with that of the new one. Movement of a terminal within the same registration area will not generate registration messages. Registration messages may also be generated when the terminals are switched on. Similarly, messages are generated to deregister them when they are switched off.

PCS services may be provided by different types of commercial service vendors. Bellcore, [3]

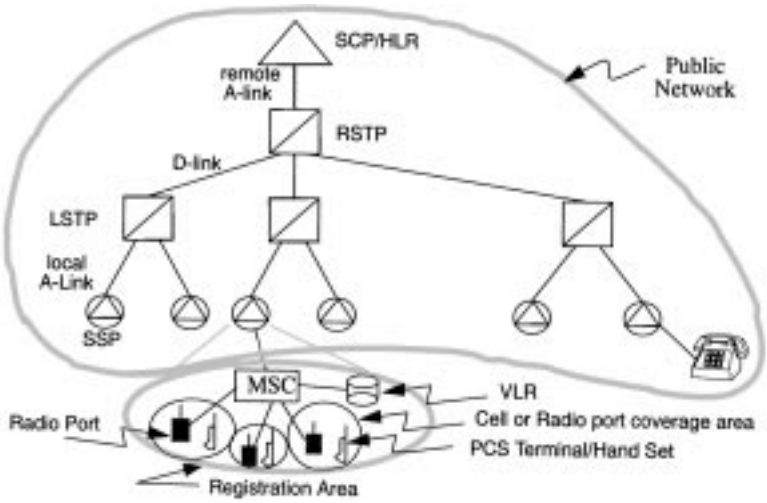


FIGURE 20.2: Example of a reference model for a PCS.

describes three different types of PSPs and the different access services that a public network may provide to them. For example, a PSP may have full network capabilities with its own switching, radio management, and radio port capabilities. Certain others may not have switching capabilities, and others may have only radio port capabilities. The model in Fig. 20.2 assumes full PSP capabilities. The analysis in Section 20.5 is based on this model and modifications may be necessary for other types of PSPs.

It is also quite possible that one or more registration areas may be served by a single PSP. The PSP may have one or more HLRs for serving its service area. In such a situation users that move within the PSP's serving area may generate traffic to the PSP's HLR (not shown in Fig. 20.2) but not to the network's HLR (shown in Fig. 20.2). In the interest of keeping the discussions simple, we have assumed that there is one-to-one correspondence between SSPs and MSCs and also between MSCs, registration areas, and VLRs. One impact of locating the SSP, MSC, and VLR in separate physical sites connected by SS7 signalling links would be to increase the required signalling message volume on the SS7 network. Our model assumes that the messages between the SSP and the associated MSC and VLR do not add to signalling load on the public network. Other configurations and assumptions could be studied for which the analysis may need to be suitably modified. The underlying analysis techniques will not, however, differ significantly.

20.3 IS-41 Preliminaries

We now describe the message flow for call origination, call delivery, and terminal registration, sometimes called location registration, based on the IS-41 protocol. This protocol is described in detail in EIA/TIA, [6]. Only an outline is provided here.

20.3.1 Terminal/Location Registration

During IS-41 registration, signalling is performed between the following pairs of network elements:

- New serving MSC and the associated database (or VLR)

- New database (VLR) in the visited area and the HLR in the public network
- HLR and the VLR in former visited registration area or the old MSC serving area.

Figure 20.3 shows the signalling message flow diagram for IS-41 registration activity, focusing only on the essential elements of the message flow relating to registration; for details of variations from the basic registration procedure, see Bellcore, [3].

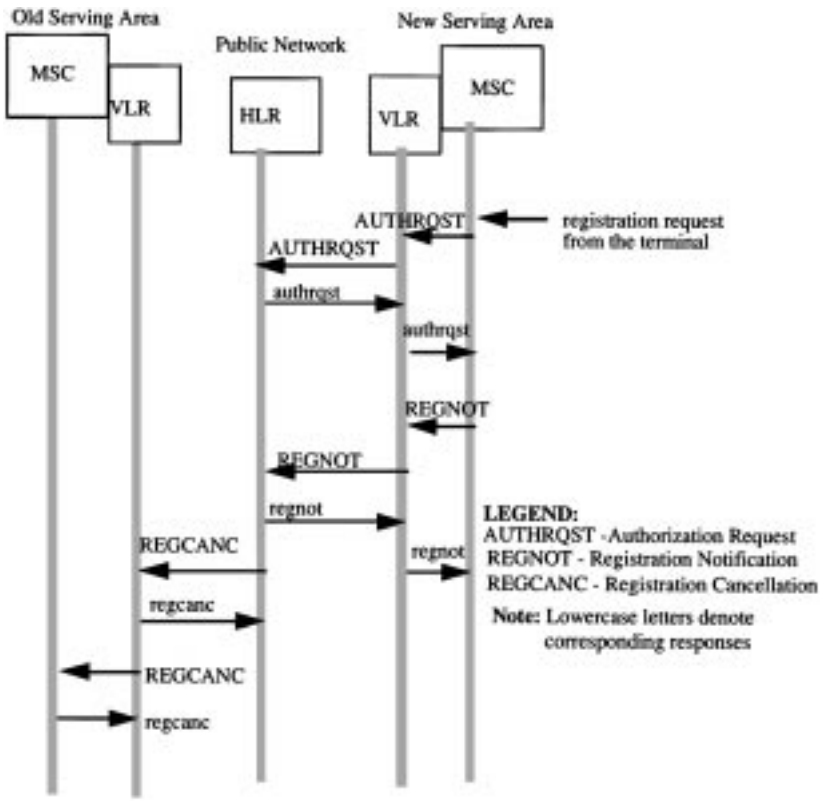


FIGURE 20.3: Signalling flow diagram for registration in IS-41.

The following steps describe the activities that take place during registration.

1. Once a terminal enters a new registration area, the terminal sends a registration request to the MSC of that area.
2. The MSC sends an authentication request (AUTHRQST) message to its VLR to authenticate the terminal, which in turn sends the request to the HLR. The HLR sends its response in the authrqst message.
3. Assuming the terminal is authenticated, the MSC sends a registration notification (REGNOT) message to its VLR.
4. The VLR in turn sends a REGNOT message to the HLR serving the terminal. The HLR updates the location entry corresponding to the terminal to point to the new serving

MSC/VLR. The HLR sends a response back to the VLR, which may contain relevant parts of the user's service profile. The VLR stores the service profile in its database and also responds to the serving MSC.

5. If the user/terminal was registered previously in a different registration area, the HLR sends a registration cancellation (REGCANC) message to the previously visited VLR. On receiving this message, the VLR erases all entries for the terminal from the record and sends a REGCANC message to the previously visited MSC, which then erases all entries for the terminal from its memory.

The protocol shows authentication request and registration notification as separate messages. If the two messages can be packaged into one message, then the rate of queries to HLR may be cut in half. This does not necessarily mean that the total number of messages are cut in half.

20.3.2 Call Delivery

The signalling message flow diagram for IS-41 call delivery is shown in Fig. 20.4. The following steps describe the activities that take place during call delivery.

1. A call origination is detected and the number of the called terminal (for example, MIN) is received by the serving MSC. Observe that the call could have originated from within the public network from a wireline phone or from a wireless terminal in an MSC/VLR serving area. (If the call originated within the public network, the AIN SSP analyzes the dialed digits and sends a query to the SCP.)
2. The MSC determines the associated HLR serving the called terminal and sends a location request (LOCREQ) message to the HLR.
3. The HLR determines the serving VLR for that called terminal and sends a routing address request (ROUTEREQ) to the VLR, which forwards it to the MSC currently serving the terminal.
4. Assuming that the terminal is idle, the serving MSC allocates a temporary identifier, called a temporary local directory number (TLDN), to the terminal and returns a response to the HLR containing this information. The HLR forwards this information to the originating SSP/MSC in response to its LOCREQ message.
5. The originating SSP requests call setup to the serving MSC of the called terminal via the SS7 signalling network using the usual call setup protocols.

Similar to the considerations for reducing signalling traffic for location registration, the VLR and HLR functions could be united in a single logical database for a given serving area and collocated; further, the database and switch can be integrated into the same piece of physical equipment or be collocated. In this manner, a significant portion of the messages exchanged between the switch, HLR and VLR as shown in Fig. 20.4 will not contribute to signalling traffic.

20.4 Global System for Mobile Communications

In this section we describe the user location strategy proposed in the European Global System for Mobile Communications (GSM) standard and its offshoot, digital cellular system 1800 (DCS1800). There has recently been increased interest in GSM in North America, since it is possible that early deployment of PCS will be facilitated by using the communication equipment already available from

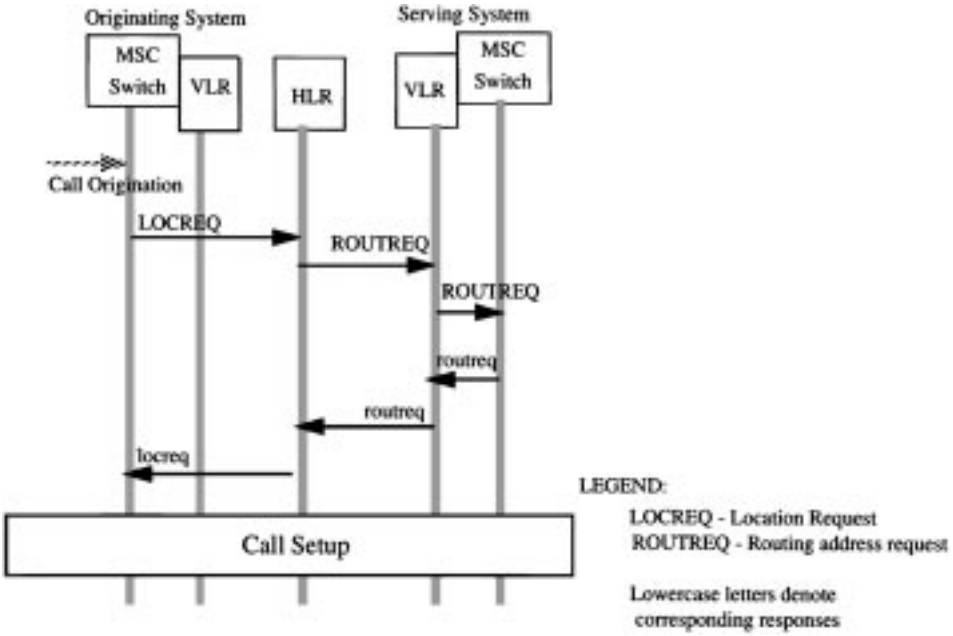


FIGURE 20.4: Signalling flow diagram for call delivery in IS-41.

European manufacturers who use the GSM standard. Since the GSM standard is relatively unfamiliar to North American readers, we first give some background and introduce the various abbreviations. The reader will find additional details in Mouley and Pautet, [18]. For an overview on GSM, refer to Lycksell, [15].

The abbreviation GSM originally stood for Groupe Special Mobile, a committee created within the pan-European standardization body Conference Europeenne des Posts et Telecommunications (CEPT) in 1982. There were numerous national cellular communication systems and standards in Europe at the time, and the aim of GSM was to specify a uniform standard around the newly reserved 900-MHz frequency band with a bandwidth of twice 25 MHz. The phase 1 specifications of this standard were frozen in 1990. Also in 1990, at the request of the United Kingdom, specification of a version of GSM adapted to the 1800-MHz frequency, with bandwidth of twice 75 MHz, was begun. This variant is referred to as DCS1800; the abbreviation GSM900 is sometimes used to distinguish between the two variations, with the abbreviation GSM being used to encompass both GSM900 and DSC1800. The motivation for DCS1800 is to provide higher capacities in densely populated urban areas, particularly for PCS. The DCS1800 specifications were frozen in 1991, and by 1992 all major GSM900 European operators began operation.

At the end of 1991, activities concerning the post-GSM generation of mobile communications were begun by the standardization committee, using the name universal mobile telecommunications system (UMTS) for this effort. In 1992, the name of the standardization committee was changed from GSM to special mobile group (SMG) to distinguish it from the 900-MHz system itself, and the term GSM was chosen as the commercial trademark of the European 900-MHz system, where GSM now stands for global system for mobile communications.

The GSM standard has now been widely adopted in Europe and is under consideration in several other non-European countries, including the United Arab Emirates, Hong Kong, and New Zealand.

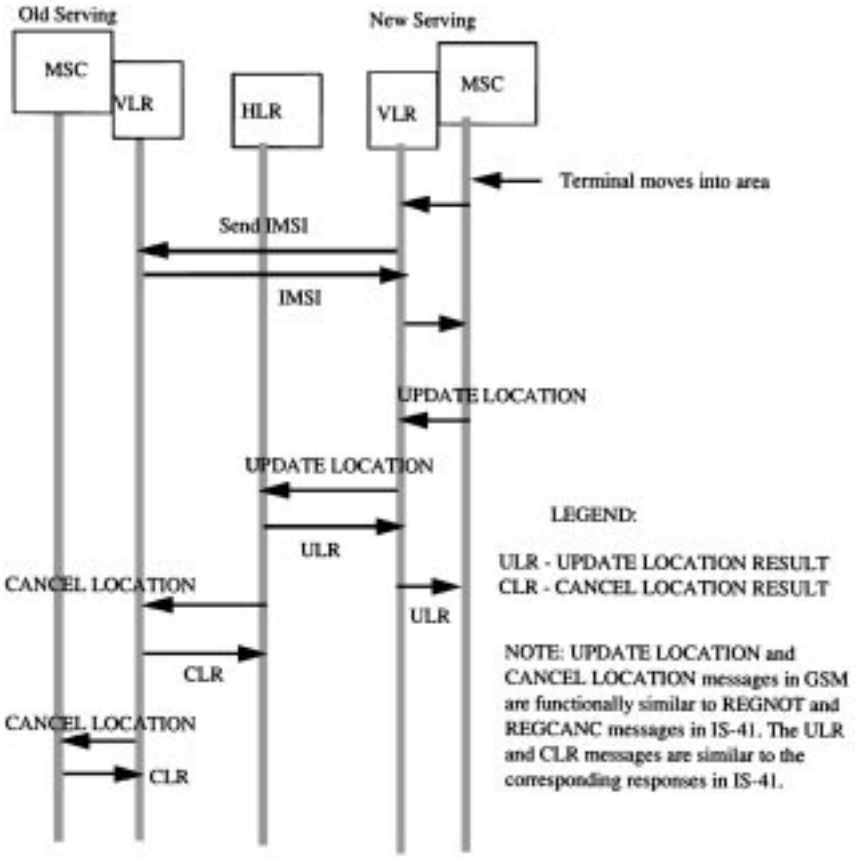


FIGURE 20.5: Flow diagram for registration in GSM.

In 1992, Australian operators officially adopted GSM.

20.4.1 Architecture

In this section we describe the GSM architecture, focusing on those aspects that differ from the architecture assumed in the IS-41 standard.

A major goal of the GSM standard was to enable users to move across national boundaries and still be able to communicate. It was considered desirable, however, that the operational network within each country be operated independently. Each of the operational networks is called a public land mobile network (PLMN) and its commercial coverage area is confined to the borders of one country (although some radio coverage overlap at national boundaries may occur), and each country may have several competing PLMNs.

A GSM customer subscribes to a single PLMN called the home PLMN, and subscription information includes the services the customer subscribes to. During normal operation, a user may elect to choose other PLMNs as their service becomes available (either as the user moves or as new operators enter the marketplace). The user's terminal [GSM calls the terminal a mobile station (MS)] assists the user in choosing a PLMN in this case, either presenting a list of possible PLMNs to the user using

explicit names (e.g., DK Sonofon for the Danish PLMN) or choosing automatically based on a list of preferred PLMNs stored in the terminal's memory. This PLMN selection process allows users to choose between the services and tariffs of several competing PLMNs. Note that the PLMN selection process differs from the cell selection and handoff process that a terminal carries out automatically without any possibility of user intervention, typically based on received radio signal strengths and, thus, requires additional intelligence and functionality in the terminal.

The geographical area covered by a PLMN is partitioned into MSC serving areas, and a registration area is constrained to be a subset of a single MSC serving area. The PLMN operator has complete freedom to allocate cells to registration areas. Each PLMN has, logically speaking, a single HLR, although this may be implemented as several physically distributed databases, as for IS-41. Each MSC also has a VLR, and a VLR may serve one or several MSCs. As for IS-41, it is interesting to consider how the VLR should be viewed in this context. The VLR can be viewed as simply a database off loading the query and signalling load on the HLR and, hence, logically tightly coupled to the HLR or as an ancillary processor to the MSC. This distinction is not academic; in the first view, it would be natural to implement a VLR as serving several MSCs, whereas in the second each VLR would serve one MSC and be physically closely coupled to it. For GSM, the MSC implements most of the signalling protocols, and at present all switch manufacturers implement a combined MSC and VLR, with one VLR per MSC [18].

A GSM mobile station is split in two parts, one containing the hardware and software for the radio interface and the other containing subscribers-specific and location information, called the subscriber identity module (SIM), which can be removed from the terminal and is the size of a credit card or smaller. The SIM is assigned a unique identity within the GSM system, called the international mobile subscriber identity (IMSI), which is used by the user location strategy as described the next subsection. The SIM also stores authentication information, services lists, PLMN selection lists, etc., and can itself be protected by password or PIN.

The SIM can be used to implement a form of large-scale mobility called SIM roaming. The GSM specifications standardize the interface between the SIM and the terminal, so that a user carrying his or her SIM can move between different terminals and use the SIM to personalize the terminal. This capability is particularly useful for users who move between PLMNs which have different radio interfaces. The user can use the appropriate terminal for each PLMN coverage area while obtaining the personalized facilities specified in his or her SIM. Thus, SIMs address personal mobility. In the European context, the usage of two closely related standards at different frequencies, namely, GSM900 and DCS1800, makes this capability an especially important one and facilitates interworking between the two systems.

20.4.2 User Location Strategy

We present a synopsis of the user location strategy in GSM using call flow diagrams similar to those used to describe the strategy in IS-41.

In order to describe the registration procedure, it is first useful to clarify the different identifiers used in this procedure. The SIM of the terminal is assigned a unique identity, called the IMSI, as already mentioned. To increase confidentiality and make more efficient use of the radio bandwidth, however, the IMSI is not normally transmitted over the radio link. Instead, the terminal is assigned a temporary mobile subscriber identity (TMSI) by the VLR when it enters a new registration area. The TMSI is valid only within a given registration area and is shorter than the IMSI. The IMSI and TMSI are identifiers that are internal to the system and assigned to a terminal or SIM and should not be confused with the user's number that would be dialed by a calling party; the latter is a separate number called the mobile subscriber integrated service digital network (ISDN) number (MSISDN),

and is similar to the usual telephone number in a fixed network.

We now describe the procedure during registration. The terminal can detect when it has moved into the cell of a new registration area from the system information broadcast by the base station in the new cell. The terminal initiates a registration update request to the new base station; this request includes the identity of the old registration area and the TMSI of the terminal in the old area. The request is forwarded to the MSC, which, in turn, forwards it to the new VLR. Since the new VLR cannot translate the TMSI to the IMSI of the terminal, it sends a request to the old VLR to send the IMSI of the terminal corresponding to that TMSI. In its response, the old VLR also provides the required authentication information. The new VLR then initiates procedures to authenticate the terminal. If the authentication succeeds, the VLR uses the IMSI to determine the address of the terminal's HLR.

The ensuing protocol is then very similar to that in IS-41, except for the following differences. When the new VLR receives the registration affirmation (similar to regnot in IS-41) from the HLR, it assigns a new TMSI to the terminal for the new registration area. The HLR also provides the new VLR with all relevant subscriber profile information required for call handling (e.g., call screening lists, etc.) as part of the affirmation message. Thus, in contrast with IS-41, authentication and subscriber profile information are obtained from both the HLR and old VLR and not just the HLR.

The procedure for delivering calls to mobile users in GSM is very similar to that in IS-41. The sequence of messages between the caller and called party's MSC/VLRs and the HLR is identical to that shown in the call flow diagrams for IS-41, although the names, contents and lengths of messages may be different and, hence, the details are left out. The interested reader is referred to Mouly and Pautet, [18], or Lycksell, [15], for further details.

20.5 Analysis of Database Traffic Rate for IS-41 and GSM

In the two subsections that follow, we state the common set of assumptions on which we base our comparison of the two strategies.

20.5.1 The Mobility Model for PCS Users

In the analysis that follows in the IS-41 analysis subsection, we assume a simple mobility model for the PCS users. The model, which is described in [23], assumes that PCS users carrying terminals are moving at an average velocity of v and their direction of movement is uniformly distributed over $[0, 2\pi]$. Assuming that the PCS users are uniformly populated with a density of ρ and the registration area boundary is of length L, it has been shown that the rate of registration area crossing R is given by

$$R = \frac{\rho v L}{\pi} \tag{20.1}$$

Using Eq. (20.1), we can calculate the signalling traffic due to registration, call origination, and delivery. We now need a set of assumptions so that we may proceed to derive the traffic rate to the databases using the model in Fig. 20.2.

20.5.2 Additional Assumptions

The following assumptions are made in performing the analysis.

- 128 total registration areas

- Square registration area size: $(7.575\ \text{km})^2 = 57.5\ \text{km}^2$, with border length $L = 30.3$ km
- Average call origination rate = average call termination (delivery) rate = 1.4/h/terminal
- Mean density of mobile terminals = $\rho = 390/\text{km}^2$
- Total number of mobile terminals = $128 \times 57.4 \times 390 = 2.87 \times 10^6$
- Average call origination rate = average call termination (delivery) rate = 1.4/h/terminal
- Average speed of a mobile, $v = 5.6$ km/h
- Fluid flow mobility model

The assumptions regarding the total number of terminals may also be obtained by assuming that a certain public network provider serves 19.15×10^6 users and that 15% (or 2.87×10^6) of the users also subscribe to PCS services from various PSPs.

Note that we have adopted a simplified model that ignores situations where PCS users may turn their handsets on and off that will generate additional registration and deregistration traffic. The model also ignores wireline registrations. These activities will increase the total number of queries and updates to HLR and VLRs.

20.5.3 Analysis of IS-41

Using Eq. (20.1) and the parameter values assumed in the preceding subsection, we can compute the traffic due to registration. The registration traffic is generated by mobile terminals moving into a new registration area, and this must equal the mobile terminals moving out of the registration area, which per second is

$$R_{\text{reg, VLR}} = \frac{390 \times 30.3 \times 5.6}{3600\pi} = 5.85$$

This must also be equal to the number of deregistrations (registration cancellations),

$$R_{\text{dereg, VLR}} = 5.85$$

The total number of registration messages per second arriving at the HLR will be

$$R_{\text{reg, HLR}} = R_{\text{reg, VLR}} \times \text{ total No. of registration areas} = 749$$

The HLR should, therefore, be able to handle, roughly, 750 updates per second. We observe from Fig. 20.3 that authenticating terminals generate as many queries to VLR and HLR as the respective number of updates generated due to registration notification messages.

The number of queries that the HLR must handle during call origination and delivery can be similarly calculated. Queries to HLR are generated when a call is made to a PCS user. The SSP that receives the request for a call, generates a location request (LOCREQ) query to the SCP controlling the HLR. The rate per second of such queries must be equal to the rate of calls made to PCS users.

This is calculated as

$$\begin{aligned} R_{\text{CallDeliv, HLR}} &= \text{call rate per user} \times \text{total number of users} \\ &= \frac{1.4 \times 2.87 \times 10^5}{3600} \\ &= 1116 \end{aligned}$$

For calls originated from a mobile terminal by PCS users, the switch authenticates the terminal by querying the VLR. The rate per second of such queries is determined by the rate of calls originating

in an SSP serving area, which is also a registration area (RA). This is given by

$$R_{\text{CallOrig, VLR}} = \frac{1116}{128} = 8.7$$

This is also the number of queries per second needed to authenticate terminals of PCS users to which calls are delivered:

$$R_{\text{CallDeliv, VLR}} = 8.7$$

Table 20.1 summarizes the calculations.

TABLE 20.1 IS-41 Query and Update Rates to HLR and VLR

Activity	HLR Updates/s	VLR Updates/s	HLR Queries/s	VLR queries/s
Mobility-related activities at registration	749	5.85	749	5.85
Mobility-related activities at deregistration		5.85		
Call origination				8.7
Call delivery			1116	8.7
Total (per RA)	5.85	11.7	14.57	23.25
Total (Network)	749	1497.6	1865	2976

20.5.4 Analysis of GSM

Calculations for query and update rates for GSM may be performed in the same manner as for IS-41, and they are summarized in Table 20.2. The difference between this table and Table 20.1 is that in GSM the new serving VLR does not query the HLR separately in order to authenticate the terminal during registration and, hence, there are no HLR queries during registration. Instead, the entry (749 queries) under HLR queries in Table 20.1, corresponding to mobility-related authentication activity at registration, gets equally divided between the 128 VLRs. Observe that with either protocol the total database traffic rates are conserved, where the total database traffic for the entire network is given by the sum of all of the entries in the last row total (Network), i.e.,

$$\text{HLR updates} + \text{VLR updates} + \text{HLR queries} + \text{VLR queries}$$

From Tables 20.1 and 20.2 we see that this quantity equals 7087.

The conclusion is independent of any variations we may provide to the assumptions in earlier in the section. For example, if the PCS penetration (the percentage of the total users subscribing to PCS services) were to increase from 15 to 30%, all of the entries in the two tables will double and, hence, the total database traffic generated by the two protocols will still be equal.

20.6 Reducing Signalling During Call Delivery

In the preceding section, we provided a simplified analysis of some scenarios associated with user location strategies and the associated database queries and updates required. Previous studies [13, 16]

TABLE 20.2 GSM Query and Update Rates to HLR and VLR

Activity	HLR Updates/s	VLR Updates/s	HLR Queries/s	VLR Queries/s
Mobility-related activities at registration	749	5.85		11.7
Mobility-related activities at deregistration		5.85		
Call origination				8.7
Call delivery			1116	8.7
Total (per VLR)	749	11.7	1116	29.1
Total (Network)	749	1497.6	1116	3724.8

indicate that the signalling traffic and database queries associated with PCS due to user mobility are likely to grow to levels well in excess of that associated with a conventional call. It is, therefore, desirable to study modifications to the two protocols that would result in reduced signalling and database traffic. We now provide some suggestions.

For both GSM and IS-41, delivery of calls to a mobile user involves four messages: from the caller's VLR to the called party's HLR, from the HLR to the called party's VLR, from the called party's VLR to the HLR, and from the HLR to the caller's VLR. The last two of these messages involve the HLR, whose role is to simply relay the routing information provided by the called party's VLR to the caller's VLR. An obvious modification to the protocol would be to have the called VLR directly send the routing information to the calling VLR. This would reduce the total load on the HLR and on signalling network links substantially. Such a modification to the protocol may not be easy, of course, due to administrative, billing, legal, or security concerns. Besides, this would violate the query/response model adopted in IS-41, requiring further analysis.

A related question which arises is whether the routing information obtained from the called party's VLR could instead be stored in the HLR. This routing information could be provided to the HLR, for example, whenever a terminal registers in a new registration area. If this were possible, two of the four messages involved in call delivery could be eliminated. This point was discussed at length by the GSM standards body, and the present strategy was arrived at. The reason for this decision was to reduce the number of temporary routing numbers allocated by VLRs to terminals in their registration area. If a temporary routing number (TLDN in IS-41 or MSRN in GSM) is allocated to a terminal for the whole duration of its stay in a registration area, the quantity of numbers required is much greater than if a number is assigned on a per-call basis. Other strategies may be employed to reduce signalling and database traffic via intelligent paging or by storing user's mobility behavior in user profiles (see, for example, Tabbane, [22]). A discussion of these techniques is beyond the scope of the paper.

20.7 Per-User Location Caching

The basic idea behind per-user location caching is that the volume of SS7 message traffic and database accesses required in locating a called subscriber can be reduced by maintaining local storage, or cache, of user location information at a switch. At any switch, location caching for a given user should be employed only if a large number of calls originate for that user from that switch, relative to the user's mobility. Note that the cached information is kept at the switch from which calls originate, which may or may not be the switch where the user is currently registered.

Location caching involves the storage of location pointers at the originating switch; these point to

the VLR (and the associated switch) where the user is currently registered. We refer to the procedure of locating a PCS user a *FIND* operation, borrowing the terminology from Awerbuch and Peleg, [1]. We define a basic *FIND*, or *BasicFIND*( ), as one where the following sequence of steps takes place.

1. The incoming call to a PCS user is directed to the nearest switch.
2. Assuming that the called party is not located within the immediate RA, the switch queries the HLR for routing information.
3. The HLR contains a pointer to the VLR in whose associated RA the subscriber is currently situated and launches a query to that VLR.
4. The VLR, in turn, queries the MSC to determine whether the user terminal is capable of receiving the call (i.e., is idle) and, if so, the MSC returns a routable address (TLDN in IS-41) to the VLR.
5. The VLR relays the routing address back to the originating switch via the HLR.

At this point, the originating switch can route the call to the destination switch. Alternately, *BasicFIND*( ) can be described by pseudocode as follows. (We observe that a more formal method of specifying PCS protocols may be desirable).

```
BasicFIND( ){
    Call to PCS user is detected at local switch;
    if called party is in same RA then return;
    Switch queries called party's HLR;
    Called party's HLR queries called party's current VLR, V;
    V returns called party's location to HLR;
    HLR returns location to calling switch;
}
```

In the *FIND* procedure involving the use of location caching, or *CacheFIND*( ), each switch contains a local memory (cache) that stores location information for subscribers. When the switch receives a call origination (from either a wire-line or wireless caller) directed to a PCS subscriber, it first checks its cache to see if location information for the called party is maintained. If so, a query is launched to the pointed VLR; if not, *BasicFIND*( ), as just described, is followed. If a cache entry exists and the pointed VLR is queried, two situations are possible. If the user is still registered at the RA of the pointed VLR (i.e., we have a *cache hit*), the pointed VLR returns the user's routing address. Otherwise, the pointed VLR returns a *cache miss.*

```
CacheFIND( ){
    Call to PCS user is detected at local switch;
    if called is in same RA then return;
    if there is no cache entry for called user
    then invoke BasicFIND( ) and return;
    Switch queries the VLR, V, specified in the cache entry;
    if called is at V, then
        V returns called party's location to calling switch;
    else {
        V returns "miss" to calling switch;
        Calling switch invokes BasicFIND( );
    }
}
```

When a cache hit occurs we save one query to the HLR [a VLR query is involved in both *CacheFIND*( ) and *BasicFIND*( )], and we also save traffic along some of the signalling links; instead of four message transmissions, as in *BasicFIND*( ), only two are needed. In steady-state operation, the cached pointer for any given user is updated only upon a miss.

Note that the *BasicFIND*( ) procedure differs from that specified for roaming subscribers in the IS-41 standard EIA/TIA, [6]. In the IS-41 standard, the second line in the *BasicFIND*( ) procedure is omitted, i.e., every call results in a query of the called user's HLR. Thus, in fact, the procedure specified in the standard will result in an even higher network load than the *BasicFIND*( ) procedure specified here. To make a fair assessment of the benefits of *CacheFIND*( ), however, we have compared it against *BasicFIND*( ). Thus, the benefits of *CacheFIND*( ) investigated here depend specifically on the use of caching and not simply on the availability of user location information at the local VLR.

20.8 Caching Threshold Analysis

In this section we investigate the classes of users for which the caching strategy yields net reductions in signalling traffic and database loads. We characterize classes of users by their CMR. The CMR of a user is the average number of calls to a user per unit time, divided by the average number of times the user changes registration areas per unit time. We also define a LCMR, which is the average number of calls to a user from a given originating switch per unit time, divided by the average number of times the user changes registration areas per unit time.

For each user, the amount of savings due to caching is a function of the probability that the cached pointer correctly points to the user's location and increases with the user's LCMR. In this section we quantify the minimum value of LCMR for caching to be worthwhile. This caching threshold is parameterized with respect to costs of traversing signalling network elements and network databases and can be used as a guide to select the subset of users to whom caching should be applied. The analysis in this section shows that estimating user's LCMRs, preferably dynamically, is very important in order to apply the caching strategy. The next section will discuss methods for obtaining this estimate.

From the pseudocode for *BasicFIND*( ), the signalling network cost incurred in locating a PCS user in the event of an incoming call is the sum of the cost of querying the HLR (and receiving the response), and the cost of querying the VLR which the HLR points to (and receiving the response). Let

α = cost of querying the HLR and receiving a response
β = cost of querying the pointed VLR and receiving a response

Then, the cost of *BasicFIND*( ) operation is

$$C_B = \alpha + \beta \tag{20.2}$$

To quantify this further, assume costs for traversing various network elements as follows.

A_l = cost of transmitting a location request or response message on A link between SSP and LSTP
D = cost of transmitting a location request on response message or D link
A_r = cost of transmitting a location request or response message on A link between RSTP and SCP
L = cost of processing and routing a location request or response message by LSTP
R = cost of processing and routing a location request or response message by RSTP
H_Q = cost of a query to the HLR to obtain the current VLR location

V_Q = cost of a query to the VLR to obtain the routing address

Then, using the PCS reference network architecture (Fig. 80.2),

$$\alpha = 2(A_l + D + A_r + L + R) + H_Q \tag{20.3}$$
$$\beta = 2(A_l + D + A_r + L + R) + V_Q \tag{20.4}$$

From Eqs. (20.2)–(20.4)

$$C_B = 4(A_l + D + A_r + L + R) + H_Q + V_Q \tag{20.5}$$

We now calculate the cost of *CacheFIND*( ). We define the *hit ratio* as the relative frequency with which the cached pointer correctly points to the user's location when it is consulted. Let

p = cache hit ratio
C_H = cost of the *CacheFIND*( ) procedure when there is a hit
C_M = cost of the *CacheFIND*( ) procedure when there is a miss

Then the cost of *CacheFIND*( ) is

$$C_C = p\,C_H + (1-p)C_M \tag{20.6}$$

For *CacheFIND*( ), the signalling network costs incurred in locating a user in the event of an incoming call depend on the hit ratio as well as the cost of querying the VLR, which is stored in the cache; this VLR query may or may not involve traversing the RSTP. In the following, we say a VLR is a *local* VLR if it is served by the same LSTP as the originating switch, and a *remote* VLR otherwise. Let

q = Prob (VLR in originating switch's cache is a local VLR)
δ = cost of querying a local VLR
ϵ = cost of querying a remote VLR
η = cost of updating the cache upon a miss

Then,

$$\delta = 4A_l + 2L + V_Q \tag{20.7}$$
$$\epsilon = 4(A_l + D + L) + 2R + V_Q \tag{20.8}$$
$$C_H = q\delta + (1-q)\epsilon \tag{20.9}$$

Since updating the cache involves an operation to a fast local memory rather than a database operation, we shall assume in the following that $\eta = 0$. Then,

$$C_M = C_H + C_B = q\delta + (1-q)\epsilon + \alpha + \beta \tag{20.10}$$

From Eqs. (20.6), (20.9) and (20.10) we have

$$C_C = \alpha + \beta + \epsilon - p(\alpha + \beta) + q(\delta - \epsilon) \tag{20.11}$$

For net cost savings we require $C_C < C_B$, or that the hit ratio exceeds a *hit ratio threshold* p_T, derived using Eqs. (20.6), (20.9), and (20.2),

$$p > p_T = \frac{C_H}{C_B} = \frac{\epsilon + q(\delta - \epsilon)}{\alpha + \beta} \tag{20.12}$$
$$= \frac{4A_l + 4D + 4L + 2R + V_Q - q(4D + 2L + 2R)}{4A_l + 4D + 4A_r + 4L + 4R + H_Q + V_Q} \tag{20.13}$$

Equation (20.13) specifies the hit ratio threshold for a user, evaluated at a given switch, for which local maintenance of a cached location entry produces cost savings. As pointed out earlier, a given user's hit ratio may be location dependent, since the rates of calls destined for that user may vary widely across switches.

The hit ratio threshold in Eq. (20.13) is comprised of heterogeneous cost terms, i.e., transmission link utilization, packet switch processing, and database access costs. Therefore, numerical evaluation of the hit ratio threshold requires either detailed knowledge of these individual quantities or some form of simplifying assumptions. Based on the latter approach, the following two possible methods of evaluation may be employed.

1. Assume one or more cost terms dominate, and simplify Eq. (20.13) by setting the remaining terms to zero.
2. Establish a common unit of measure for all cost terms, for example, *time delay*. In this case, A_l, A_r, and D may represent transmission delays of fixed transmission speed (e.g., 56 kb/s) signalling links, L and R may constitute the sum of queueing and service delays of packet switches (i.e., STPs), and H_Q and V_Q the transaction delays for database queries.

In this section we adopt the first method and evaluate Eq. (20.13) assuming a single term dominates. (In Section 20.9 we present results using the second method). Table 20.3 shows the hit ratio threshold required to obtain net cost savings, for each case in which one of the cost terms is dominant.

TABLE 20.3 Minimum Hit Ratios and LCMRs for Various Individual Dominant Signalling Network Cost Terms

Dominant Cost Term	Hit ratio Threshold, p_T	LCMR Threshold, $LCMR_T$	LCMR Threshold ($q = 0.043$)	LCMR Threshold ($q = 0.25$)
A_l	1	∞	∞	∞
A_r	0	0	0	0
D	$1 - q$	$1/q - 1$	22	3
L	$1 - q/2$	$2/q - 1$	45	7
R	$1 - q/2$	$2/q - 1$	45	7
H_Q	0	0	0	0
V_Q	1	∞	∞	∞

In Table 20.3 we see that if the cost of querying a VLR or of traversing a local A link is the dominant cost, caching for users who may move is never worthwhile, regardless of users' call reception and mobility patterns. This is because the caching strategy essentially distributes the functionality of the HLR to the VLRs. Thus, the load on the VLR and the local A link is always increased, since any move by a user results in a cache miss. On the other hand, for a fixed user (or telephone), caching is always worthwhile. We also observe that if the remote A links or HLR querying are the bottlenecks, caching is worthwhile even for users with very low hit ratios.

As a simple average-case calculation, consider the net network benefit of caching when HLR access and update is the performance bottleneck. Consider a scenario where $u = 50\%$ of PCS users receive $c = 80\%$ of their calls from $s = 5$ RAs where their hit ratio $p > 0$, and $s' = 4$ of the SSPs at those RAs contain sufficiently large caches. Assume that caching is applied only to this subset of users and to no other users. Suppose that the average hit ratio for these users is $p = 80\%$, so that 80% of the

HLR accesses for calls to these users from these RA are avoided. Then the net saving in the accesses to the system's HLR is $H = (u\,c\,s'\,p)/s = 25\%$.

We discuss other quantities in Table 20.3 next. It is first useful to relate the cache hit ratio to users' calling and mobility patterns directly via the LCMR. Doing so requires making assumptions about the distribution of the user's calls and moves. We consider the steady state where the incoming call stream from an SSP to a user is a Poisson process with arrival rate λ, and the time that the user resides in an RA has a general distribution $F(t)$ with mean $1/\mu$. Thus,

$$LCMR = \frac{\lambda}{\mu} \tag{20.14}$$

Let t be the time interval between two consecutive calls from the SSP to the user and t_1 be the time interval between the first call and the time when the user moves to a new RA. From the random observer property of the arrival call stream [7], the hit ratio is

$$p = Pr[t < t_1] = \int_{t=0}^{\infty} \lambda e^{-\lambda t} \int_{t_1=t}^{\infty} \mu\,[1 - F\,(t_1)]\;\mathrm{d}t_1\,\mathrm{d}t$$

If $F(t)$ is an exponential distribution, then

$$p = \frac{\lambda}{\lambda + \mu} \tag{20.15}$$

and we can derive the *LCMR threshold*, the minimum LCMR required for caching to be beneficial assuming incoming calls are a Poisson process and intermove times are exponentially distributed,

$$LCMR_T = \frac{p_T}{1 - p_T} \tag{20.16}$$

Equation (20.16) is used to derive LCMR thresholds assuming various dominant costs terms, as shown in Table 20.3.

Several values for $LCMR_T$ in Table 20.3 involve the term q, i.e., the probability that the pointed VLR is a local VLR. These values may be numerically evaluated by simplifying assumptions. For example, assume that all of the SSPs in the network are uniformly distributed amongst l LSTPs. Also, assume that all of the PCS subscribers are uniformly distributed in location across all SSPs and that each subscriber exhibits the same incoming call rate at every SSP. Under those conditions, q is simply $1/l$. Consider the case of the public switched telephone network. Given that there are a total of 160 local access transport area (LATA) across the 7 Regional Bell Operating Company (RBOC) regions [4], the average number of LATAs, or l, is 160/7 or 23. Table 20.3 shows the results with $q = 1/l$ in this case.

We observe that the assumption that all users receive calls uniformly from all switches in the network is extremely conservative. In practice, we expect that user call reception patterns would display significantly more locality, so that q would be larger and the LCMR thresholds required to make caching worthwhile would be smaller. It is also worthwhile to consider the case of a RBOC region with PCS deployed in a few LATA only, a likely initial scenario, say, 4 LATAs. In either case the value of q would be significantly higher; Table 20.3 shows the LCMR threshold when $q = 0.25$.

It is possible to quantify the net costs and benefits of caching in terms of signalling network impacts in this way and to determine the hit ratio and LCMR threshold above which users should have the caching strategy applied. Applying caching to users whose hit ratio and LCMR is below this threshold results in net increases in network impacts. It is, thus, important to estimate users' LCMRs accurately. The next section discusses how to do so.

20.9 Techniques for Estimating Users' LCMR

Here we sketch some methods of estimating users' LCMR. A simple and attractive policy is to not estimate these quantities on a per-user basis at all. For instance, if the average LCMR over all users in a PCS system is high enough (and from Table 20.3, it need not be high depending on which network elements are the dominant costs), then caching could be used at every SSP to yield net system-wide benefits. Alternatively, if it is known that at any given SSP the average LCMR over all users is high enough, a cache can be installed at that SSP. Other variations can be designed.

One possibility for deciding about caching on a per-user basis is to maintain information about a user's calling and mobility pattern at the HLR and to download it periodically to selected SSPs during off-peak hours. It is easy to envision numerous variations on this idea.

In this section we investigate two possible techniques for estimating LCMR on a per-user basis when caching is to be deployed. The first algorithm, called the *running average* algorithm, simply maintains a running average of the hit ratio for each user. The second algorithm, called the *reset-K* algorithm, attempts to obtain a measure of the hit ratio over the recent history of the user's movements. We describe the two algorithms next and evaluate their effectiveness using a stochastic analysis taking into account user calling and mobility patterns.

20.9.1 The Running Average Algorithm

The running average algorithm maintains, for every user that has a cache entry, the running average of the hit ratio. A running count is kept of the number of calls to a given user, and, regardless of the *FIND* procedure used to locate the user, a running count of the number of times that the user was at the same location for any two consecutive calls; the ratio of these numbers provides the measured running average of the hit ratio. We denote the measured running average of the hit ratio by p_M; in steady state, we expect that $p_M = p$. The user's previous location as stored in the cache entry is used only if the running average of the hit ratio p_M is greater than the cache hit threshold p_T. Recall that the cache scheme outperforms the basic scheme if $p > p_T = C_H/C_B$. Thus, in steady state, the running average algorithm will outperform the basic scheme when $p_M > p_T$.

We consider, as before, the steady state where the incoming call stream from an SSP to a user is a Poisson process with arrival rate λ, and the time that the user resides in an RA has an exponential distribution with mean $1/\mu$. Thus $LCMR = \lambda/\mu$ [Eq. (20.14)] and the location tracking cost at steady state is

$$C_C = \begin{cases} p_M C_H + (1 - p_M) C_B, & p_M > p_T \\ C_B, & \text{otherwise} \end{cases} \tag{20.17}$$

Figure 20.6 plots the cost ratio C_C/C_B from Eq. (20.17) against *LCMR*. (This corresponds to assigning uniform units to all cost terms in Eq. (20.13), i.e., the second evaluation method as discussed in Section 20.8. Thus, the ratio C_C/C_B may represent the percentage reduction in user location time with the caching strategy compared to the basic strategy.) The figure indicates that in the steady state, the caching strategy with the running average algorithm for estimating LCMR can significantly outperform the basic scheme if *LCMR* is sufficiently large. For instance with $LCMR \sim 5$, caching can lead to cost savings of 20–60% over the basic strategy.

Equation (20.17) (cf., solid curves in Fig. 20.6) is validated against a simple Monte Carlo simulation (cf., dashed curves in Fig. 20.6). In the simulation, the confidence interval for the 95% confidence level of the output measure C_C/C_B is within 3% of the mean value. This simulation model will later be used to study the running average algorithm when the mean of the movement distribution changes from time to time [which cannot be modeled by using Eq. (20.17)].

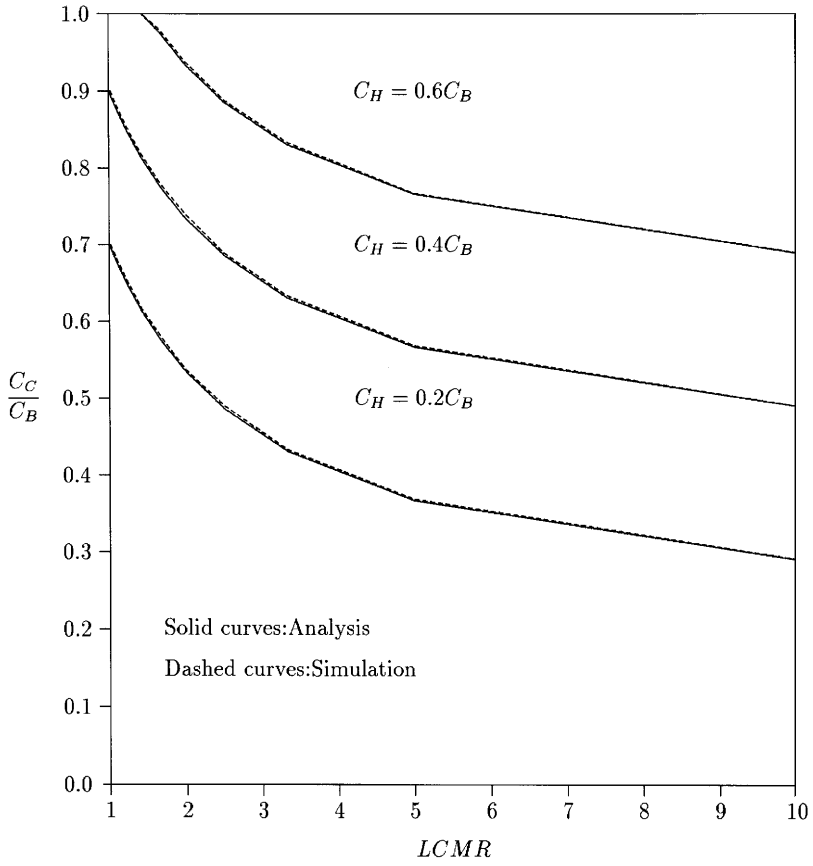


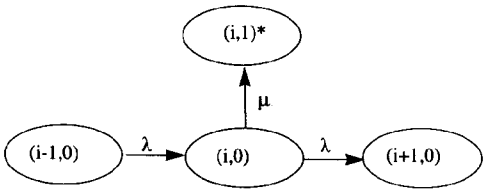
FIGURE 20.6: The location tracking cost for the running average algorithm.

One problem with the running average algorithm is that the parameter p is measured from the entire past history of the user's movement, and the algorithm may not be sufficiently dynamic to adequately reflect the recent history of the user's mobility patterns.

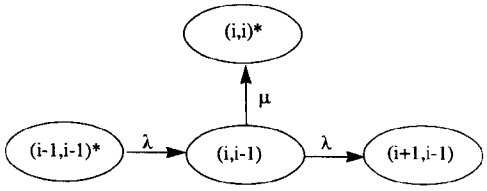
20.9.2 The Reset-K Algorithm

We may modify the running average algorithm such that p is measured from the recent history. Define every K incoming calls as a *cycle.* The modified algorithm, which is referred to as the reset-K algorithm, counts the number of cache hits n in a cycle. If the measured hit ratio for a user, $p_M = n/K \geq p_T$, then the cache is enabled for that user, and the cached information is always used to locate the user in the next cycle. Otherwise, the cache is disabled for that user, and the basic scheme is used. At the beginning of a cycle, the cache hit count is reset, and a new p_M value is measured during the cycle.

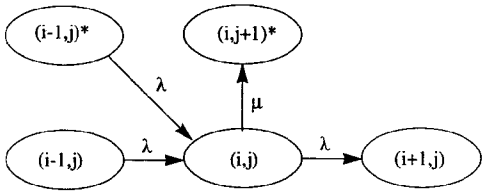
To study the performance of the reset-K algorithm, we model the number of cache misses in a cycle by a Markov process. Assume as before that the call arrivals are a Poisson process with arrival rate λ and the time period the user resides in an RA has an exponential distribution with mean $1/\mu$. A pair (i, j), where $i > j$, represents the state that there are j cache misses before the first i incoming



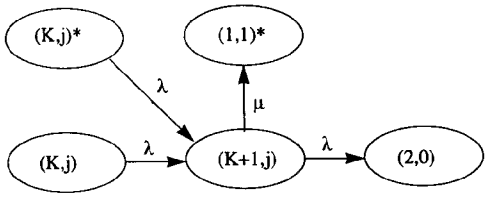
(a) Transitions for state (i,0) (2 < i < K+1)



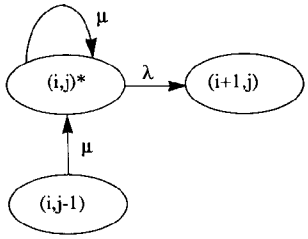
(b) Transitions for state (i,i-1)(1 < i < K+1)



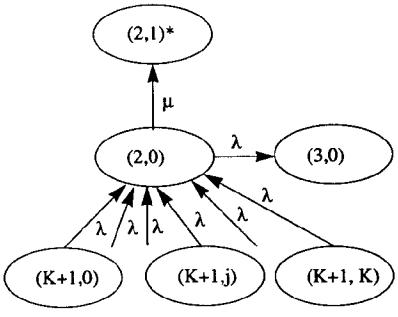
(c) Transitions for state (i,j) (2< i < K+1, 0 < j < i-1)



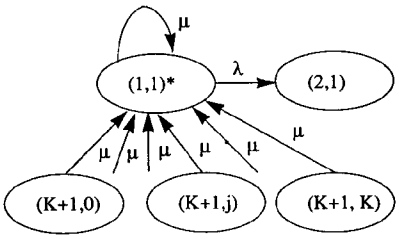
(d) Transitions for state (K+1,j) (0 < j < K+1)



(e) Transitions for state (i,j)* (0 < j ≤ i, 1 < i < K



(f) Transitions for state (2,0)



(g) Transitions for state (1,1)*

FIGURE 20.7: State transitions.

phone calls in a cycle. A pair $(i, j)^*$, where $i \geq j \geq 1$, represents the state that there are $j-1$ cache misses before the first i incoming phone calls in a cycle, and the user moves between the ith and the $i+1$ phone calls. The difference between (i, j) and $(i, j)^*$ is that if the Markov process is in the state (i, j) and the user moves, then the process moves into the state $(i, j+1)^*$. On the other hand, if the process is in state $(i, j)^*$ when the user moves, the process remains in $(i, j)^*$ because at most one cache miss occurs between two consecutive phone calls.

Figure 20.7(a) illustrates the transitions for state $(i, 0)$ where $2 < i < K+1$. The Markov process moves from $(i-1, 0)$ to $(i, 0)$ if a phone call arrives before the user moves out. The rate is λ. The process moves from $(i, 0)$ to $(i, 1)^*$ if the user moves to another RA before the $i+1$ call arrival. Let $\pi(i, j)$ denote the probability of the process being in state (i, j). Then the transition equation is

$$\pi(i, 0) = \frac{\lambda}{\lambda + \mu}\pi(i-1, 0), \qquad 2 < i < K+1 \tag{20.18}$$

Figure 20.7(b) illustrates the transitions for state $(i, i-1)$ where $1 < i < K+1$. The only transition into the state $(i, i-1)$ is from $(i-1, i-1)^*$, which means that the user always moves to another RA after a phone call. [Note that there can be no state $(i-1, i-1)$ by definition and, hence, no transition from such a state.] The transition rate is λ. The process moves from $(i, i-1)$ to $(i, i)^*$ with rate μ, and moves to $(i+1, i-1)$ with rate λ. Let $\pi^*(i, j)$ denote the probability of the process being in state $(i, j)^*$. Then the transition equation is

$$\pi(i, i-1) = \frac{\lambda}{\lambda + \mu}\pi^*(i-1, i-1), \qquad 1 < i < K+1 \tag{20.19}$$

Figure 20.7(c) illustrates the transitions for state (i, j) where $2 < i < K+1, 0 < j < i-1$. The process may move into state (i, j) from two states $(i-1, j)$ and $(i-1, j)^*$ with rate λ, respectively. The process moves from (i, j) to $(i, j+1)^*$ or $(i+1, j)$ with rates μ and λ, respectively. The transition equation is

$$\pi(i, j) = \frac{\lambda}{\lambda + \mu}\left[\pi(i-1, j) + \pi^*(i-1, j)\right], \qquad 2 < i < K+1, \qquad 0 < j < i-1 \tag{20.20}$$

Figure 20.7(d) illustrates the transitions for state $(K+1, j)$ where $0 < j < K+1$. Note that if a phone call arrives when the process is in (K, j) or $(K, j)^*$, the system enters a new cycle (with rate λ), and we could represent the new state as $(1, 0)$. In our model, we introduce the state $(K+1, j)$ instead of $(1, 0)$, where

$$\sum_{0 \leq j \leq K} \pi(K+1, j) = \pi(1, 0)$$

so that the hit ratio, and thus the location tracking cost, can be derived [see Eq. (20.25)]. The process moves from $(K+1, j)$ [i.e., $(1, 0)$] to $(1, 1)^*$ with rate μ if the user moves before the next call arrives. Otherwise, the process moves to $(2, 0)$ with rate λ. The transition equation is

$$\pi(K+1, j) = \frac{\lambda}{\lambda + \mu}\left[\pi(K, j) + \pi^*(K, j)\right], \qquad 0 < j < K+1 \tag{20.21}$$

For $j = 0$, the transition from $(K, j)^*$ to $(K+1, 0)$ should be removed in Fig. 20.7(d) because the state $(K, 0)^*$ does not exist. The transition equation for $(K+1, 0)$ is given in Eq. (20.18). Figure 20.7(e) illustrates the transitions for state $(i, j)^*$ where $0 < j < i, 1 < i < K+1$. The

process can only move to $(i, j)^*$ from $(i, j-1)$ (with rate μ). From the definition of $(i, j)^*$, if the user moves when the process is in $(i, j)^*$, the process remains in $(i, j)^*$ (with rate μ). Otherwise, the process moves to $(i+1, j)$ with rate λ. The transition equation is

$$\pi^*(i, j) = \frac{\mu}{\lambda}\pi(i, j-1), \qquad 0 < j \le i, \qquad 1 < i < K+1, \qquad i \ge 2 \tag{20.22}$$

The transitions for $(2, 0)$ are similar to the transitions for $(i, 0)$ except that the transition from $(1, 0)$ is replaced by $(K+1, 0), \ldots, (K+1, K)$ [cf., Fig. 20.7(f)]. The transition equation is

$$\pi(2, 0) = \frac{\lambda}{\lambda+\mu}\left[\sum_{0\le j\le K} \pi(K+1, j)\right] \tag{20.23}$$

Finally, the transitions for $(1, 1)^*$ is similar to the transitions for $(i, j)^*$ except that the transition from $(1, 0)$ is replaced by $(K+1, 0), \ldots, (K+1, K)$ [cf., Fig. 20.7(g)]. The transition equation is

$$\pi^*(1, 1) = \frac{\mu}{\lambda}\left[\sum_{0\le j\le K} \pi(K+1, j)\right] \tag{20.24}$$

Suppose that at the beginning of a cycle, the process is in state $(K+1, j)$, then it implies that there are j cache misses in the previous cycle. The cache is enabled if and only if

$$p_M \ge p_T = \frac{C_H}{C_B} \Rightarrow 1 - \frac{j}{K} \ge \frac{C_H}{C_B} \Rightarrow 0 \le j \le \left\lceil K\left(1 - \frac{C_H}{C_B}\right)\right\rceil$$

Thus, the probability that the measured hit ratio $p_M < p_T$ in the previous cycle is

$$Pr\,[p_M < p_T] = \frac{\displaystyle\sum_{\lceil k[1-(C_H/C_B)]\rceil < j \le K} \pi(K+1, j)}{\displaystyle\sum_{0\le j\le K} \pi(K+1, j)}$$

and the location tracking cost for the reset-K algorithm is

$$\begin{aligned} C_C \;=\; & C_B Pr\,[p_M < p_T] + (1 - Pr\,[p_M < p_T]) \\ & \times \left\{ \sum_{0\le j\le K} \left(\frac{(K-j)C_H}{K} + \frac{j(C_H + C_B)}{K}\right) \left[\frac{\pi(K+1, j)}{\displaystyle\sum_{0\le i\le K} \pi(K+1, i)}\right] \right\} \end{aligned} \tag{20.25}$$

The first term Eq. (20.25) represents the cost incurred when caching is disabled because the hit ratio threshold exceeds the hit ratio measured in the previous cycle. The second term is the cost when the cache is enabled and consists of two parts, corresponding to calls during which hits occur and calls during which misses occur. The ratio in square brackets is the conditional probability of being in state $\pi(K+1, j)$ during the current cycle.

The numerical computation of $\pi(K+1, j)$ can be done as follows. First, compute $a_{i,j}$ and $b_{i,j}$ where $\pi(i, j) = a_{i,j}\pi^*(1, 1)$ and $\pi^*(i, j) = b_{i,j}\pi^*(1, 1)$. Note that $a_{i,j} = 0 (b_{i,j} = 0)$ if $\pi(i, j)[\pi^*(i, j)]$ is not defined in Eqs. (20.18)–(20.24). Since

$$\sum_{i,j} \left[\pi(i, j) + \pi^*(i, j)\right] = 1$$

we have

$$\pi^*(1,1) = \frac{1}{\sum_{i,j}(a_{i,j} + b_{i,j})}$$

and $\pi(K+1, j)$ can be computed and the location tracking cost for the reset-K algorithm is obtained using Eq. (20.25).

The analysis is validated by a Monte Carlo simulation. In the simulation, the confidence interval for the 98% confidence level of the output measure C_C/C_B is within 3% of the mean value. Figure 20.8 plots curves for Eq. (20.25) (the solid curves) against the simulation experiments (the dashed curves) for $K = 20$ and $C_H = 0.5C_B$ and $0.3C_B$, respectively. The figure indicates that the analysis is consistent with the simulation model.

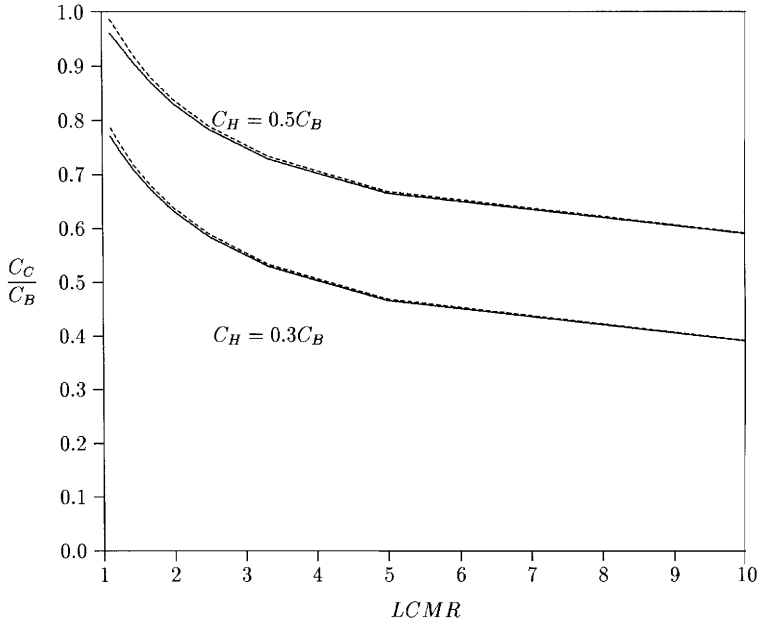


FIGURE 20.8: The location tracking costs for the reset-K algorithm; $K = 20$.

20.9.3 Comparison of the LCMR Estimation Algorithms

If the distributions for the incoming call process and the user movement process never change, then we would expect the running average algorithm to outperform the reset-K algorithm (especially when K is small) because the measured hit ratio p_M in the running average algorithm approaches the true hit ratio value p in the steady state. Surprisingly, the performance for the reset-K algorithm is roughly the same as the running average algorithm even if K is as small as 10. Figure 20.9 plots the location tracking costs for the running average algorithm and the reset-K algorithm with different K values.

The figure indicates that in steady state, when the distributions for the incoming call process and

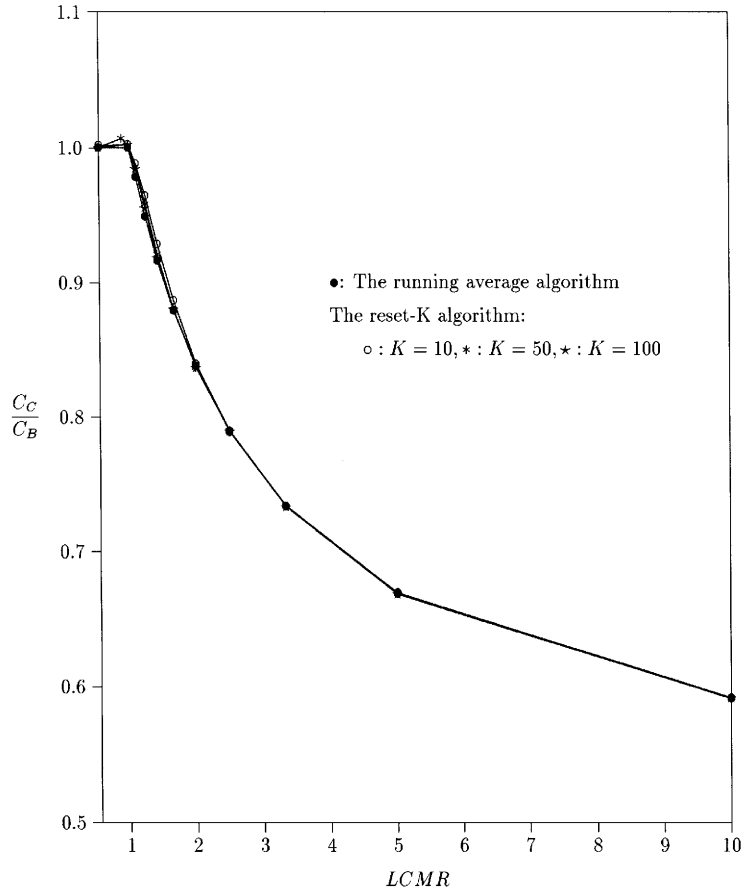


FIGURE 20.9: The location tracking costs for the running average algorithm and the reset-K algorithm; $C_H = 0.5C_B$.

the user movement process never change, the running average algorithm outperforms reset K, and a large value of K outperforms a small K but the differences are insignificant.

If the distributions for the incoming call process or the user movement process change from time to time, we expect that the reset-K algorithm outperforms the running average algorithm. We have examined this proposition experimentally. In the experiments, 4000 incoming calls are simulated. The call arrival rate changes from 0.1 to 1.0, 0.3, and then 5.0 for every 1000 calls (other sequences have been tested and similar results are observed). For every data point, the simulation is repeated 1000 times to ensure that the confidence interval for the 98% confidence level of the output measure C_C/C_B is within 3% of the mean value. Figure 20.10 plots the location tracking costs for the two algorithms for these experiments. By changing the distributions of the incoming call process, we observe that the reset-K algorithm is better than the running average algorithm for all C_H/C_B values.

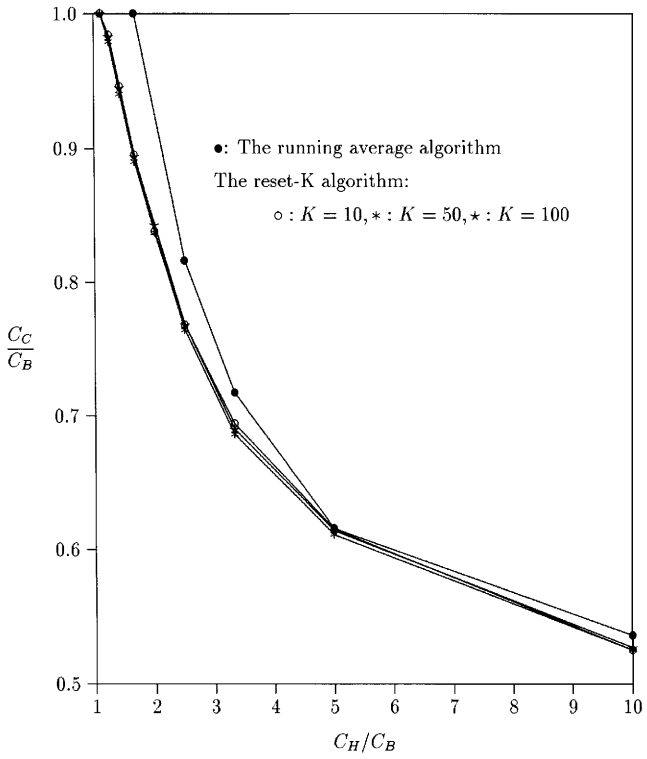


FIGURE 20.10: Comparing the running average algorithm and the reset-K algorithm under unstable call traffic.

20.10 Discussion

In this section we discuss aspects of the caching strategy presented here. Caching in PCS systems raises a number of issues not encountered in traditional computer systems, particularly with respect to architecture and locality in user call and mobility patterns. In addition, several variations in our reference assumptions are possible for investigating the implementation of the caching strategies. Here we sketch some of the issues involved.

20.10.1 Conditions When Caching Is Beneficial

We summarize the conditions for which the auxiliary strategies are worthwhile, under the assumptions of our analysis.

The caching strategy is very promising when the HLR update (or query load) or the remote A link is the performance bottleneck, since a low *LCMR* (*LCMR* > 0) is required. For caching, the total database load and signalling network traffic is reduced whenever there is a cache hit. In addition, load and traffic is redistributed from the HLR and higher level SS7 network elements (RSTP, D links) to the VLRs and lower levels where excess network capacity may be more likely to exist. If the VLR is the performance bottleneck, the caching strategy is not promising, unless the VLR capacity is upgraded.

The benefits of the caching strategy depend on user call and mobility patterns when the D link, RSTP, and LSTP are the performance bottlenecks. We have used a Poisson call arrival model and

exponential intermove time to estimate this dependence. Under very conservative assumptions, for caching to be beneficial requires relatively high *LCMR* (25–50); we expect that in practice this threshold could be lowered significantly (say, $LCMR > 7$). Further experimental study is required to estimate the amount of locality in user movements for different user populations to investigate this issue further. It is possible that for some classes of users data obtained from active badge location system studies (e.g., Fishman and Mazer, [8]) could be useful. In general, it appears that caching could also potentially provide benefits to some classes of users even when the D link, the RSTP, or the LSTP are the bottlenecks.

We observe that more accurate models of user calling and mobility patterns are required to help resolve the issues raised in this section. We are currently engaged in developing theoretical models for user mobility and estimating their effect on studies of various aspects of PCS performance [10].

20.10.2 Alternative Network Architectures

The reference architecture we have assumed (Fig. 20.2) is only one of several possible architectures. It is possible to consider variations in the placement of the HLR and VLR functionality, (e.g., placing the VLR at a local SCP associated with the LSTP instead of at the SSP), the number of SSPs served by an LSTP, the number of HLRs deployed, etc. It is quite conceivable that different regional PCS service providers and telecommunications companies will deploy different signalling network architectures, as well as placement of databases for supporting PCS within their serving regions [19]. It is also possible that the number and placement of databases in a network will change over time as the number of PCS users increases.

Rather than consider many possible variations of the architecture, we have selected a reference architecture to illustrate the new auxiliary strategies and our method of calculating their costs and benefits. Changes in the architecture may result in minor variations in our analysis but may not significantly affect our qualitative conclusions.

20.10.3 LCMR Estimation and Caching Policy

It is possible that for some user populations estimating the LCMR may not be necessary, since they display a relatively high-average LCMR. For some populations, as we have shown in Section 20.9, obtaining accurate estimates of user LCMR in order to decide whether or not to use caching can be important in determining the net benefits of caching.

In general, schemes for estimating the LCMR range from static to dynamic and from distributed to centralized. We have presented two simple distributed algorithms for estimating LCMR, based on a long-range and short-range running calculation; the former is preferable if the call and mobility pattern of users is fairly Tuning the amount of history that is used to determine whether caching should be employed for a particular user is an obvious area for further study but is outside the scope of this chapter.

An alternative approach is to utilize some user-supplied information, by requesting profiles of user movements (e.g., see Tabbane, [22] and to integrate this with the caching strategy. A variation of this approach is to use some domain knowledge about user populations and their characteristics.

A related issue is that of cache size and management. In practice it is likely that the monetary cost of deploying a cache may limit its size. In that case, cache entries may not be maintained for some users; selecting these users carefully is important to maximize the benefits of caching. Note that the cache hit ratio threshold cannot necessarily be used to determine which users have cache entries, since it may be useful to maintain cache entries for some users even though their hit ratios have temporarily fallen below the threshold. A simple policy that has been found to be effective in

computer systems in the least recently used (LRU) policy [20] in which cache entries that have been least recently used are discarded; LRU may offer some guidance in this context.

20.11 Conclusions

We began this chapter with an overview of the nuances of PCS, such as personal and terminal mobility, registration, deregistration, call delivery, etc. A tutorial was then provided on the two most common strategies for locating users in PCS, in North American interim standard IS-41 and the Pan-European standard GSM. A simplified analysis of the two standards was then provided to show the reader the extent to which database and signalling traffic is likely to be generated by PCS services. Suggestions were then made that are likely to result in reduced traffic.

Previous studies [12, 13, 14, 16] of PCS-related network signalling and data management functionalities suggest a high level of utilization of the signalling network in supporting call and mobility management activities for PCS systems. Motivated by the need to evolve location strategies to reduce signalling and database loads, we then presented an auxiliary strategy, called per-user caching, to augment the basic user location strategy proposed in standards [6, 18].

Using a reference system architecture for PCS, we quantified the criteria under which the caching strategy produces reductions in the network signalling and database loads in terms of users' LCMRs. We have shown that, if the HLR or the remote A link in an SS7 architecture is the performance bottleneck, caching is useful regardless of user call and mobility patterns. If the D link or STPs are the performance bottlenecks, caching is potentially beneficial for large classes of users, particularly if they display a degree of locality in their call reception patterns. Depending on the numbers of PCS users who meet these criteria, the system-wide impacts of these strategies could be significant. For instance, for users with $LCMR \sim 5$ and stable call and move patterns, caching can result in cost reduction of 20–60% over the basic user location strategy *BasicFIND*( ) under our analysis. Our results are conservative in that the *BasicFIND*( ) procedure we have used for comparison purposes already reduces the network impacts compared to the user location strategy specified in PCS standards such as IS-41.

We have also investigated in detail two simple on-line algorithms for estimating users' LCMRs and examined the call and mobility patterns for which each would be useful. The algorithms allow a system designer to tune the amount of history used to estimate a users' LCMR and, hence, to attempt to optimize the benefits due to caching. The particular values of cache hit ratios and LCMR thresholds will change with variations in the way the PCS architecture and the caching strategy is implemented, but our general approach can still be applied. There are several issues deserving further study with respect to deployment of the caching strategy, such as the effect of alternative PCS architectures, integration with other auxiliary strategies such as the use of user profiles, and effective cache management policies.

Recently, we have augmented the work reported in this paper by a simulation study in which we have compared the caching and basic user location strategies [9]. The effect of using a time-based criterion for enabling use of the cache has also been considered [11]. We have proposed elsewhere, for users with low CMRs, an auxiliary strategy involving a system of forwarding pointers to reduce the signalling traffic and database loads [10], a description of which is beyond the scope of this chapter.

Acknowledgment

We acknowledge a number of our colleagues in Bellcore who have reviewed several previous papers by the authors and contributed to improving the clarity and readability of this work.

References

[1] Awerbuch, B. and Peleg, D., Concurrent online tracking of mobile users. In *Proc. SIGCOMM Symp. Comm. Arch. Prot.,* Oct. 1991.

[2] Bellcore., Advanced intelligent network release 1 network and operations plan, Issue 1. Tech. Rept. SR-NPL-001623. Bell Communications Research, Morristown, NJ, Jun. 1991.

[3] Bellcore., Personal communications services (PCS) network access services to PCS providers, Special Report SR-TSV-002459, Bell Communications Research, Morristown, NJ, Oct. 1993a.

[4] Bellcore., Switching system requirements for interexchange carrier interconnection using the integrated services digital network user part (ISDNUP). Tech. Ref. TR-NWT-000394. Bell Communications Research. Morristown, NJ, Dec. 1992c.

[5] Berman, R.K. and Brewster, J.H., Perspectives on the AIN architecture. *IEEE Comm. Mag.,* 1(2), 27–32, 1992.

[6] Electronic Industries Association/Telecommunications Industry Association., Cellular radio telecommunications intersystem operations. Tech. Rept. IS-41. Rev. B. Jul. 1991.

[7] Feller, W., *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, New York, 1966.

[8] Fishman, N. and Mazer, M., Experience in deploying an active badge system. In *Proc. Globecom Workshop on Networking for Pers. Comm. Appl.,* Dec. 1992.

[9] Harjono, H., Jain, R., and Mohan, S., Analysis and simulation of a cache-based auxiliary location strategy for PCS. In *Proc. IEEE Conf. Networks Pers. Comm.*, 1994.

[10] Jain, R. and Lin Y.-B., An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS. *ACM Journal on Wireless Info. Networks (WINET),* 1(2), 1995.

[11] Lin, Y.-B., Determining the user locations for personal communications networks. *IEEE Trans. Vehic. Tech.,* 466–473, Aug. 1994.

[12] Lo, C., Mohan, S., and Wolff, R., A comparison of data management alternatives for personal communications applications. Second Bellcore Symposium on Performance Modeling, SR-TSV-002424, Bell Communications Research, Morristown, NJ, Nov. 1993.

[13] Lo, C.N., Wolff, R.S., and Bernhardt, R.C., An estimate of network database transaction volume to support personal communications services. In *Proc. Intl. Conf. Univ. Pers. Comm.*, 1992.

[14] Lo, C. and Wolff, R., Estimated network database transaction volume to support wireless personal data communications applications. In *Proc. Intl. Conf. Comm.,* May 1993.

[15] Lycksell, E., GSM system overview. Tech. Rept. Swedish Telecom. Admin., Jan. 1991.

[16] Meier-Hellstern, K. and Alonso, E., The use of SS7 and GSM to support high density personal communications. In *Proc. Intl. Conf. Comm.*, 1992.

[17] Mohan, S. and Jain, R., Two user location strategies for PCS. *IEEE Pers. Comm. Mag.,* Premiere issue. 42–50, Feb. 1994.

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

[19] Russo, P., Bechard, K., Brooks, E., Corn, R.L., Honig, W.L., Gove, R., and Young, J., In rollout in the United States. *IEEE Comm. Mag.,* 56–63, Mar. 1993.

[20] Silberschatz, A. and Peterson, J., *Operating Systems Concepts.* Addison-Wesley, Reading, MA, 1988.

[21] Tabbane, S., Comparison between the alternative location strategy (AS) and the classical location strategy (CS). Tech. Rept. Rutgers Univ. WINLAB. Rutgers, NJ, Jul. 1992.

[22] Tabbane, S., Evaluation of an alternative location strategy for future high density wireless communications systems. Tech. Rept. WINALAB-TR-51, Rutgers Univ. WINLAB. Rutgers, NJ, Jan. 1993.

[23] Thomas, R., Gilbert, H., and Mazziotto, G., Influence of the mobile station on the performance of a radio mobile cellular network. In *Proc. 3rd Nordic Seminar.* Sep. 1988.