ABSTRACT

Personal communications services (PCS) support mobile terminals (MTs) which are free to travel within the service coverage area. In order to effectively locate an MT when a called is initiated, location management schemes are used to keep track of the locations of the MTs. The current approach to location management requires each MT to report its location to the network periodically. The location information is then stored in databases. When a call is initiated, the network determines the current location of the called MT through a database lookup and paging procedure. In this article, a currently available standard for location management is described. Recent research results on location management are surveyed in detail.

On Location Management for Personal Communications Networks

Ian F. Akyildiz and Joseph S.M. Ho, Georgia Institute of Technology

Cellular communication has been experiencing a rapid growth in recent years. Since its introduction in the early 1980s, cellular communication has been evolving from a costly service with limited availability toward an affordable alternative to wired telephone service. This wide acceptance of cellular communication has led to the development of a new generation of mobile communication network called personal communications services (PCS), which can support a larger mobile subscriber population while providing various types of services unavailable to traditional cellular systems. The introduction of different types of services and the establishment of new service providers will result in an unprecedented predicted growth in the number of mobile subscribers from 15 million currently to around 60 million by the year 2005. Both the research and business communities are equally excited by this predicted growth in the number of subscribers. The recent government auction of frequency bandwidth for the emerging PCS has resulted in fierce competition between both established and startup communication companies. In the research arena, a large number of studies are being performed. These research efforts range from the development of location management schemes, to the design of multiple access and channel allocation schemes, to the design of PCS network architectures. In this article, we focus on surveying the location management mechanisms in PCS systems.

In ordinary wireline networks, such as the telephone network, there is a fixed relationship between a terminal and its location. Changing the location of a terminal generally involves network administration, and it cannot easily be performed by a user. Incoming calls for a particular terminal are always routed to its associated location because there is no distinction between a terminal and its location. In contrast, PCS networks support mobile terminals (MTs) that are free to travel, and the network access point of an MT changes as it moves around the network coverage area. As a result, the ID of an MT does not implicitly provide the location information of that MT. A location-tracking mechanism is needed for effective delivery of incoming calls. The current methods (IS-41, GSM MAP) [1, 2] for location management require each MT to report its location to the network periodically. The network stores the location information of each MT in location databases, and this information is retrieved during call delivery. Current methods for location management employ a centralized database architecture, while tracking and searching of MTs involve the transmission of signaling messages among various components of a signaling network. As the number of mobile subscribers increases, this scheme will become inefficient, and new and improved schemes that can effectively support a continuously increasing subscriber population are needed.

This article is organized as follows. In the second section, we describe the network architecture of current PCS systems. The third section describes the current location management standard. The fourth section discusses the recent research efforts in location management. The conclusion and suggestions for future research are given in the final section.

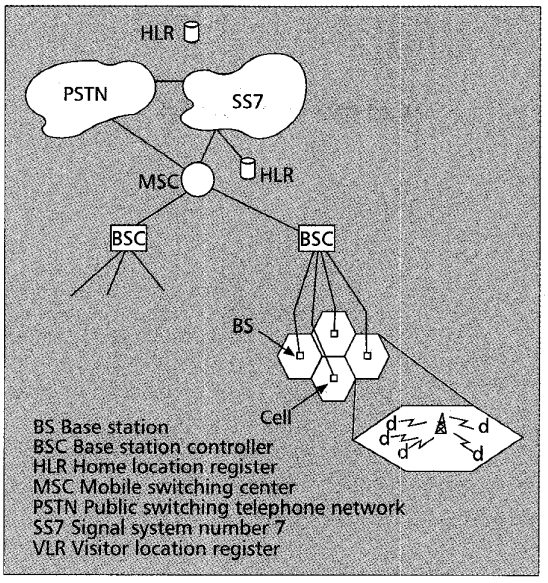
PERSONAL COMMUNICATIONS NETWORK ARCHITECTURE

Wireless bandwidth is inherently a scarce resource, and the number of wireless channels available for mobile communications is very limited. In order to support a large number of PCS subscribers with a relatively small number of wireless channels, most current PCS networks adopt a "cellular" architecture. Under this scheme, the service coverage area is divided into cells, and each cell is assigned a number of wireless channels. The number of active connections within each cell cannot exceed the number of channels available. However, the same channels may be reused in another cell as long as the two cells are separated by a sufficiently large distance to limit the interference. There is a base station installed in each cell, and MTs within a cell communicate with the base station through wireless links. The boundary of a cell is determined by the transmission range of its base station, and neighboring cells may overlap each other in areas close to their boundaries. In this article we assume that cells are hexagonal in shape. This assumption is accurate for the *macrocell* environment, where cells are relatively large (with diameters of kilometers) and the transmission range of a base station can be closely approximated by a circular region. In the *microcell* environment, where the size of cells is relatively small (with diameters of hundreds of meters), the transmission range of the base stations is highly affected by the terrain of the surrounding area, and a large portion of each cell may overlap its neighboring cells. In some crowded urban areas, a cell may cover a street segment between two intersections. Nevertheless, for presentation purposes, we assume cells to be hexagonal in shape throughout this article.

Figure 1 demonstrates the general architecture of a PCS

0163-6804/96/$05.00 © 1996 IEEE

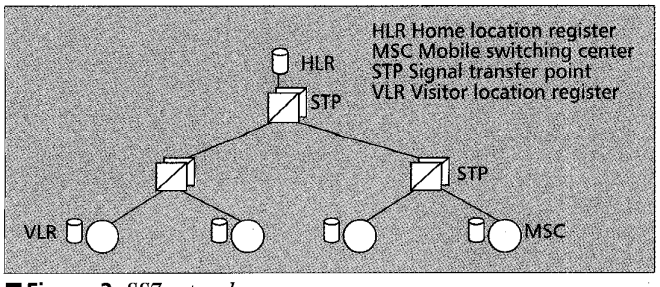
network. It can be seen in Fig. 1 that a number of base stations are connected to a base station controller (BSC). The primary function of the BSC is to manage the radio resources of its base stations, such as by performing handoff and allocating radio channels. The BSC is connected to a mobile switching center (MSC), which provides typical switching functions and coordinates location registration and call delivery. The MSC is connected to both the backbone wireline network (such as the public switched telephone network, or PSTN) and the signaling network (e.g., the Signaling System No. 7—SS7—network, described later in this section). Current schemes for location management are based on a two-level data hierarchy such that two types of database, the home location register (HLR) and the visitor location register (VLR), are involved in tracking an MT. In general, there is an HLR for each PCS network, and a user is permanently associated with an HLR in his/her subscribed PCS network. Information about each user, such as the types of services subscribed, billing information, and location information, are stored in a user profile located at the HLR. The number of VLRs and their placement vary among networks. Each VLR stores the information of the MTs (downloaded from the HLR) visiting its associated area. We will describe in the third section how the HLR and VLRs are used in tracking an MT.



■ **Figure 1.** *PCS network architecture.*

Network management functions of PCS, such as call processing and location registration, are achieved by the exchange of signaling messages through a signaling network. SS7 [3–5] is the protocol used for signaling exchange and the signaling network is referred to as the *SS7 network*. Figure 2 shows the SS7 network, which connects the HLR, the VLRs, and the MSCs in a PCS network. The signal transfer points (STPs), as shown in Fig. 2, are responsible for routing signaling messages within the SS7 network. For reliability, the STPs are installed in pairs, as shown in Fig. 2.

Each VLR is associated with one or more MSCs, and the VLR communicates with an MSC through the STP using the SS7 protocol. For simplicity, here we assume that each VLR is collocated with one MSC, and a direct communication link between the VLR and MSC is available. Depending on the network configuration, there may exist one or more HLRs in the PCS network. In Fig. 2 we assume a single HLR which is connected to the SS7 network through an STP. Note that the IS-41 standard also allows interconnection of the HLR, VLRs, and MSCs by X.25 links. Also, some PCS network architecture may have SS7 connections between the BSC and VLR.



■ **Figure 2.** *SS7 network.*

Location Management

There are two standards for location management currently available: Electronic/Telecommunications Industry Associations (EIA/TIA) Interim Standard 41(IS-41) [1] and the Global System for Mobile Communications (GSM) mobile application part (MAP) [6]. The IS-41 scheme is commonly used in North America for Advanced Mobile Phone System (AMPS), IS-54, IS-136, and Personal Access Communication System (PACS) networks, while the GSM MAP is mostly used in Europe for GSM and Digital Cellular Service at 1800 MHz (DCS-1800) networks. Both standards are based on a two-level data hierarchy, as described in the previous section and Fig. 1. Location management includes two major tasks: *location registration* and *call delivery*. Location registration procedures update the location databases (HLR and VLRs) and authenticate the MT when up-to-date location information of an MT is available. Call delivery procedures locate the MT based on the information available at the HLR and VLRs when a call for an MT is initiated. The IS-41 and GSM MAP location management strategies are very similar. While GSM MAP is designed to facilitate personal mobility and to enable user selection of network provider, there are a lot of commonalities between the two standards. Because of space limitations, the presentation of this article is based primarily on the IS-41 standard. Interested readers may refer to [2] for detailed descriptions of the GSM MAP mobility management strategy.

Location Registration

In order to correctly deliver calls, the PCS network must keep track of the location of each MT. As described previously, location information is stored in two types of database, VLR and HLR. As the MTs move around the network coverage area, the data stored in these databases may no longer be accurate. To ensure that calls can be delivered successfully, a mechanism is needed to update the databases with up-to-date location information. We call this updating process *location registration*. Location registration is initiated by an MT when it reports its current location to the PCS network. We call this reporting process *location update*. Current PCS systems adopt an approach such that the network coverage area is partitioned into registration areas (RAs) (note that GSM refers to this as the *location area*), and an MT performs a location update whenever it enters a new RA. Each RA consists of a number of cells; and, in general, all base stations belonging to the same RA are connected to the same MSC.

When an MT enters an RA, if the new RA belongs to the same VLR as the old RA, the record at the VLR is updated to record the ID of the new RA. Otherwise, if the new RA belongs to a different VLR, a number of extra steps are required to:

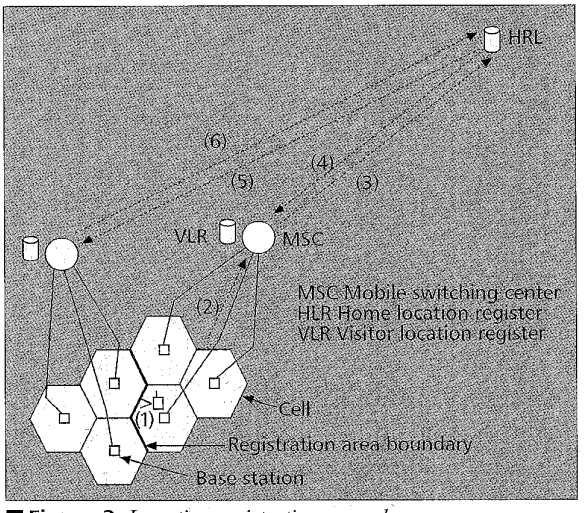
- Register the MT at the new serving VLR
- Update the HLR to record the ID of the new

serving VLR
- Deregister the MT at the old serving VLR

Figure 3 shows the location registration procedure when an MT moves to a new RA. The following is the ordered list of tasks that are performed during location registration:

1. The MT enters a new RA and transmits a location update message to the new base station.
2. The base station forwards the location update message to the MSC through a wired link, which launches a registration query to its associated VLR.
3. The VLR updates its record on the location of the MT. If the new RA belongs to a different VLR, the new VLR determines the address of the HLR of the MT from its *mobile identification number* (MIN). This is achieved by a table lookup procedure called *global title translation*. The new VLR then sends a location registration message to the HLR; otherwise, location registration is complete.
4. The HLR performs the required procedures to authenticate the MT and records the ID of the new serving VLR of the MT. The HLR then sends a registration acknowledgment message to the new VLR.
5. The HLR sends a registration cancellation message to the old VLR.
6. The old VLR removes the record of the MT and returns a cancellation acknowledgment message to the HLR.

Depending on the distance between the current and home locations of the MT, in steps 3, 4, 5, and 6 the signaling messages may have to go through several intermediate STPs before reaching their destinations. Therefore, the location registration may generate a significant traffic load to the SS7 network. As the number of mobile subscribers keeps increasing, the delay for completing a location registration may increase. A number of methods for reducing these signaling costs are discussed later.



■ **Figure 3.** *Location registration procedures.*

Call Delivery

Two major steps are involved in call delivery: determining the serving VLR of the called MT, and locating the visiting cell of the called MT. Locating the serving VLR of the MT involves the following database lookup procedure (Fig. 4):

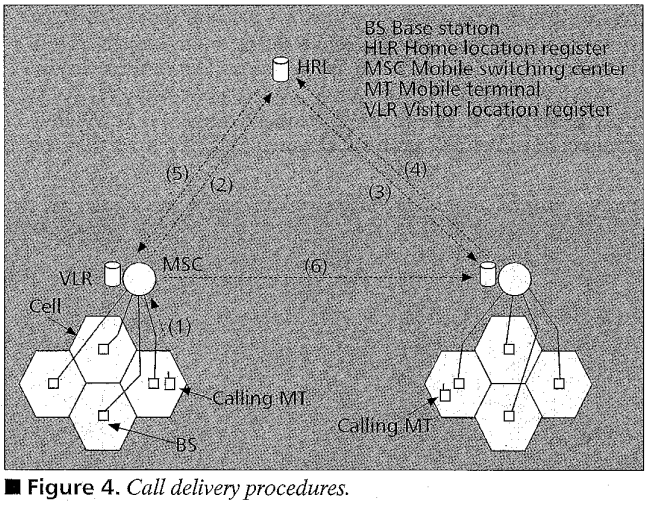
1. The calling MT sends a call initiation signal to the serving MSC of the MT through a nearby base station.
2. The MSC determines the address of the HLR of the called MT by global title translation and sends a *location request* message to the HLR.
3. The HLR determines the serving VLR of the called MT and sends a route *request message* to the VLR. This VLR then forwards the message to the MSC serving the MT.
4. The MSC allocates a temporary identifier called temporary local directory number (TLDN) to the MT and sends a reply to the HLR together with the TLDN.
5. The HLR forwards this information to the MSC of the calling MT.
6. The calling MSC requests a call setup to the called MSC through the SS7 network.

The procedure described above allows the network to set up a connection from the calling MT to the serving MSC of the called MT. Since each MSC is associated with an RA, and there are more than one cell in each RA, a mechanism is therefore necessary to determine the cell location of the called MT. In current PCS networks, this is achieved by a *paging* (or *alerting*) procedure such that polling signals are broadcast to all cells within the residing RA of the called MT. On receiving the polling signal, the MT sends a reply which allows the MSC to determine its current residing cell.

Recent Research on Location Management

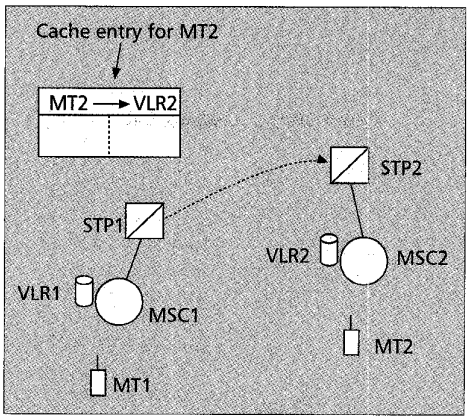
Location Registration and Call Delivery

Location registration involves updating location databases when current location information is available. On the other hand, call delivery involves the querying of location databases to determine the current location of a called MT. These can be costly processes, especially when the MT is located far from its assigned HLR. For example, if the MT is currently roaming in San Francisco and its HLR is in Atlanta, a location registration message is transmitted from San Francisco to Atlanta whenever the MT moves to a new RA that belongs to a different VLR. Under the same scenario, when a call for the MT is originated from a nearby MT in San Francisco, the MSC of the calling MT must first query the HLR in Atlanta before it finds out that the called MT is located in the same area as the caller. This may not be a significant problem in today's cellular network since the penetration is relatively low. As the number of mobile subscribers keeps increasing, the volume of signaling traffic generated by location management is becoming extremely high [7, 8]. Methods for reducing the signaling traffic are therefore needed. Research in this area generally falls into two categories. First, extensions to the IS-



■ **Figure 4.** *Call delivery procedures.*

41 location management strategy are developed which aim to improve the IS-41 scheme while keeping the basic database network architecture unchanged. This type of solution has the advantage of easy adaptation to the current PCS networks without major modification. These schemes are based on the centralized database architecture inherited from the IS-41 standard. Another category of research results in completely new database architectures which require a new set of schemes for location registration and call delivery. Most of these schemes are based on distributed database architectures. We will discuss selected results under each category in the following subsections.



■ **Figure 5.** *Per-user location caching scheme.*

Centralized Database Architectures — In [9] a *per-user location caching* strategy is introduced. The basic idea is that the volume of signaling and database access traffic for locating an MT can be reduced by maintaining a cache of location information at a nearby STP. Whenever the MT is accessed through the STP, an entry is added to the cache which contains a mapping from the ID of the MT to that of its serving VLR. When another call is initiated for an MT, the STP first checks if a cache entry exists for the MT. If no cache entry for the MT exists, the IS-41 call delivery scheme described earlier is used to locate the MT. If a cache entry exists, the STP will query the VLR as specified by the cache. If the MT is still residing under the same VLR, a *hit* occurs, and the MT is found. If the MT has already moved to another location which is not associated with the same VLR, a *miss* occurs, and the IS-41 call delivery scheme is used to locate the MT.

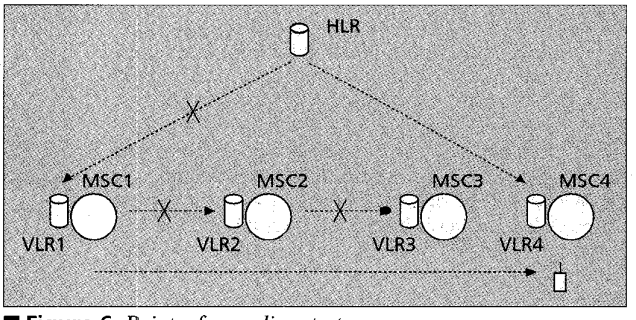
Figure 5 demonstrates the operation of per-user location caching. When a call is initiated from MT 1 to MT 2, as indicated in Fig. 5, the system can locate MT 2 by using the cached information at STP1. As a result, MT 2 is successfully located without querying the HLR of MT 2. As compared to the IS-41 scheme demonstrated in Fig. 4, per-user location caching allows the STP to location the VLR of the called MT after only one cache database lookup. This is true, however, only when the cached location information of the called MT is valid (a hit). The cost of per-user location caching is higher than the IS-41 scheme when a miss occurs. Based on the system parameters, the minimum hit ratio required to produce a performance gain using per-user location caching needs to be determined.

In [9], the authors define the local call-to-mobility ratio (LCMR) as the average number of calls to an MT from a given originating STP divided by the average number of times the user changes VLR per unit time. The minimum LCMR necessary to attain the minimum hit ratio is obtained. In order to reduce the number of misses, it is suggested in [10] that cache entries should be invalidated after a certain time interval. Based on the mobility and call arrival parameters, a T-threshold scheme is introduced in [10] which determines the time when a particular cached location information should be cleared such that the cost for call delivery can be reduced.

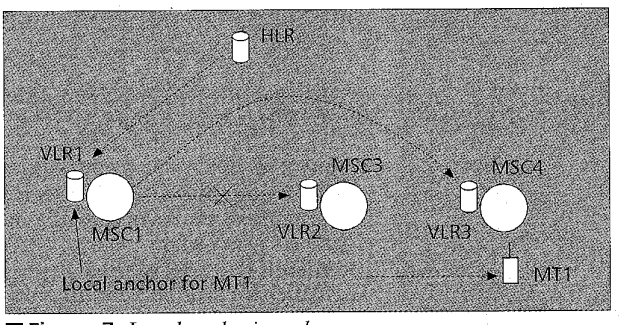
A *user profile replication* scheme is proposed in [11]. Based on this scheme, user profiles are replicated at selected local databases. When a call is initiated for a remote MT, the network first determines if a replication of the called MT's user profile is available locally. If the user profile is found no HLR query is necessary, and the network can locate the called MT based on the location information available at the local database. Otherwise, the network locates the called MT following the IS-41 procedures. When the MT moves to another location, the network updates all replications of the MT's user profile. This results in higher signaling overhead for location registration.

Depending on the mobility rate of the MT and the call arrival rate from each location, this method may significantly reduce the signaling and database access overhead for local management. A scheme is also introduced in [11] which determines the replication for each MT. Based on their scheme, the replication decision is made by a centralized system which must collect the mobility and calling parameters of the whole user population from time to time. This may not be feasible in current PCS networks because of the large number of PCS network providers involved. Besides, generating and distributing the replication decision for a large user population is a computation-intensive and time-consuming process which may incur a significant amount of network bandwidth. Future research should focus on the development of distributed user profile replication mechanisms.

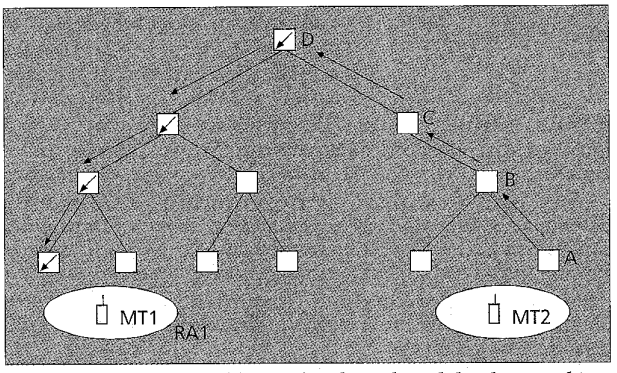
A *pointer forwarding* strategy is introduced in [12]. The basic idea is that instead of reporting a location change to the HLR every time the MT moves to an area belonging to a different VLR, the reporting can be eliminated by simply setting up a forwarding pointer from the old VLR to the new VLR. When a call for the MT is initiated, the network locates the MT by first determining the VLR at the beginning of the pointer chain and then follows the pointers to the current serving VLR of the MT. To minimize the delay in locating an MT, the length of the pointer chain is limited to a predefined maximum value, K. When the length of the pointer chain reaches K, additional forwarding is not allowed, and location change must be reported to the HLR when the next movement occurs. Figure 6 demonstrates the operation of pointer forwarding. Pointers are set up from VLR 1 to VLR 2 and from VLR 2 to VLR 3 as the MT moves from MSC 1 to MSC 2 and from MSC 2 to MSC 3, respectively. For $K = 2$, the pointer chain cannot be extended any further. An additional movement from MSC 3 to MSC 4 will result in a location registration at the HLR. The original pointers are deleted, and the HLR records the ID of the current serving VLR of the MT. It is demonstrated that, depending on the mobility and call arrival parameters and the value of K, this scheme may not always result in a reduction in cost from the original IS-41 scheme. The authors determine the conditions under which



■ **Figure 6.** *Pointer forwarding strategy.*



■ **Figure 7.** *Local anchoring scheme.*



■ **Figure 8.** *Distributed hierarchical tree-based database architecture.*

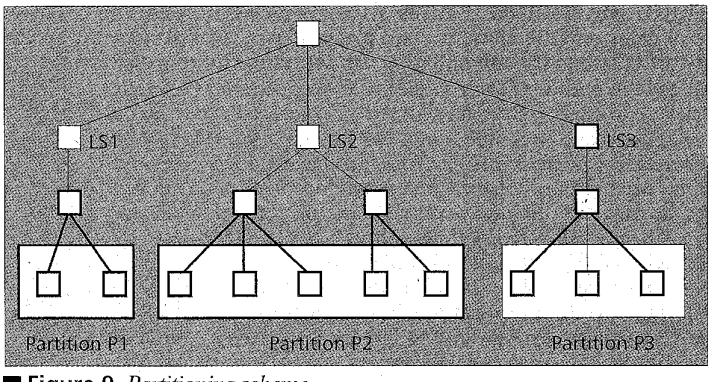
the pointer forwarding scheme should be used based on the system parameters.

A *local anchoring* scheme is introduced in [13]. Under this scheme, signaling traffic due to location registration is reduced by eliminating the need to report location changes to the HLR. A VLR close to the MT is selected as its *local anchor*. Instead of transmitting registration messages to the HLR, location changes are reported to the local anchor. Since the local anchor is close to the MT, the signaling cost incurred in location registration is reduced. The HLR keeps a pointer to the local anchor. When an incoming call arrives, the HLR queries the local anchor of the called MT which, in turn, queries the serving VLR to obtain a routable address to the called MT. Figure 7 demonstrates the local anchoring scheme. Assuming that the local anchor of MT1 is VLR 1, the location change is reported to VLR 1 (instead of the HLR) when MT 1 moves from VLR 2 to VLR 3. The authors introduce two schemes for selecting the local anchor for an MT: *static* and *dynamic* local anchoring. Under static local anchoring, the serving VLR of an MT during its last call arrival becomes its local anchor. The local anchor is changed when the next call arrival occurs. Static local anchoring completely eliminates the need to report location changes to the HLR. However, similar to the location caching and the pointer forwarding strategies, static local anchoring may not always result in performance improvement. In a similar way, dynamic local anchoring changes the local anchor to the serving VLR when a call arrives. However, the network also makes a decision whether the local anchor for an MT should be changed to the new serving VLR after each movement based on the mobility and call arrival parameters. It is demonstrated that the cost for dynamic local anchoring is always lower than or equal to that of the original IS-41 scheme.

Distributed Database Architectures — A distributed database architecture for location registration is proposed in [14]. The two-level HLR/VLR database architecture as described in the IS-41 standard is replaced by a large number of location databases. These location databases are organized as a tree with the root at the top and the leaves at the bottom. The MTs are associated with the leaf (lowest-level) location databases, and each location database contains location information for the MTs residing in its subtree. Figure 8 demonstrates the operation of the proposed scheme. Given that an MT, MT 1, is located at RA 1, an entry exists for MT 1 in each database along the path from its current location to the root of the tree.

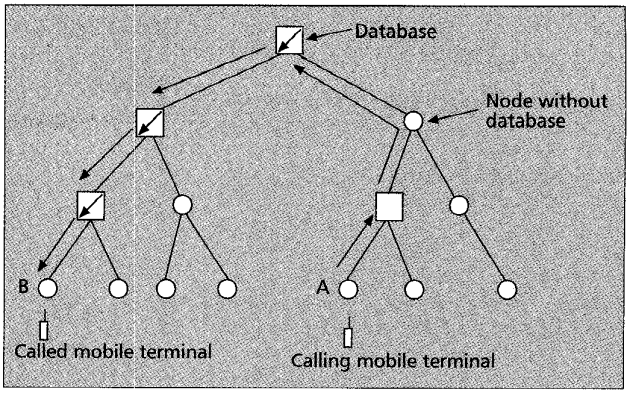
The entries for MT 1 at these databases are as shown in Fig. 8. When a call is initiated, the network locates the called MT by following its database entries; for example, if a call for MT 1 is initiated by MT 2, as shown in Fig. 8. The call request is received by node A. Since the database of node A does not have an entry for MT 1, the call request is forwarded to node B and so on. When the request finally reaches node D, an entry for MT 1 is found and the location of MT 1 determined after another three database lookups, as demonstrated in Fig. 8. When an MT moves to an RA that belongs to a different leaf database, the corresponding databases are updated to indicate the correct location of the MT. When compared to schemes based on a centralized database architecture, such as the IS-41 scheme, the proposed scheme reduces the distance traveled by signaling messages. However, this scheme increases the number of database updates and queries, and thus increases the delay in location registration and call delivery.

A *partitioning scheme* for the fully distributed database hierarchy is introduced in [15]. Since the mobility pattern of MTs varies among locations, partitions can be generated by grouping location servers among which the MT moves frequently. Based on the scheme introduced in [15], location registration is performed only when the MT enters a partition. Figure 9 shows the partitions for a particular PCS network. Partition P2 consists of five location servers that have a least common ancestor location server LS 2. When an MT moves into partition P2, LS 2 is updated to indicate that the MT is residing in its subtree. No location registration is performed when the MT moves to another location server within the same partition. This scheme minimizes the number of location registrations in areas where the mobility rate of the MTs is high. Simulation results demonstrate that the partitioning scheme is effective in reducing the signaling message cost. However, the cost reduction depends on the mobility and call arrival patterns as well as the method used for searching the subtree. Further study is needed to determine the effectiveness of this scheme under various parameters.



■ **Figure 9.** *Partitioning scheme.*

In [16], another distributed database architecture similar to that discussed in [14] is introduced. Here, MTs may be located at any node of the tree hierarchy (not limited to the leaf nodes). The root of the tree contains a database, but it is not necessary for other nodes to have databases installed. These databases store pointers for MTs. If an MT is residing at the subtree of a database, a pointer is set up in this database pointing to the next database along the path to the MT. If there is no other database along this path, the pointer points to the residing node of the MT. When a call for an MT is initiated at a node on the tree, the called MT can be located by following the pointers of the MT. Figure 10 shows the operation of this scheme. We assume that a call is initiated at node A and the called MT is located at node B. The path for searching the called MT is given in Fig. 10. If a database that does not contain a pointer for the called MT is reached, the next database along the path to the root is queried. Given the system parameters, such as the rate of movement between location areas, the authors introduce a method for determining the database placement which reduces the number of database accesses and updates.



■ **Figure 10.** *Distributed database architecture.*

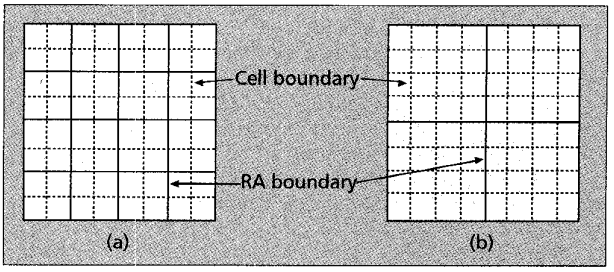
Location Update and Terminal Paging

As discussed previously, current PCS networks partition their coverage areas into a number of registration areas (RAs). Each RA consists of a group of cells, and each MT performs a location update when it enters an RA. When an incoming call arrives, the network locates the MT by simultaneously paging all cells within the RA. There are a number of inefficiencies associated with this location update and paging scheme:

- Excessive location updates may be performed by MTs that are located around RA boundaries and make frequent movements back and forth between two RAs.
- Requiring the network to poll all cells within the RA each time a call arrives may result in an excessive volume of wireless broadcast traffic.
- The mobility and call arrival patterns of MTs vary, and it is generally difficult to select an RA size that is optimal for all users. An ideal location update and paging mechanism should be able to adjust on a per-user basis.

The RA-based location update and paging scheme is a static scheme because it cannot be adjusted based on the parameters of an MT from time to time. Recent research efforts focuses primarily on dynamic location update mechanisms which perform location update based on the mobility of the MTs and the frequency of incoming calls. We will describe a number of dynamic location update and paging schemes in this section.

A method for calculating the optimal RA size given the respective costs for location update and cell polling is introduced in [17]. The authors consider a mesh cell configuration with square-shaped cells. Each RA consists of $k \times k$ cells arranged in a square, and the value of k is selected on a per-user basis according to mobility and call arrival patterns and cost parameters. As an example, we assume that there are two MTs, MT 1 and MT 2, which have different mobility and call arrival patterns such that the values of k for MT 1 and MT 2 are 2 and 4, respectively. Based on the scheme introduced in [17], Figs. 11a and b show the RAs for MT 1 and MT 2, respectively. This mechanism performs better than the static scheme in which RA size is fixed. However, it is generally not easy to use different RA sizes for different MTs because the MTs must be able to identify the boundaries of RAs which are continuously changing. The implementation of this scheme is complicated when cells are hexagonal or, in the worst case, when irregular cells are used.

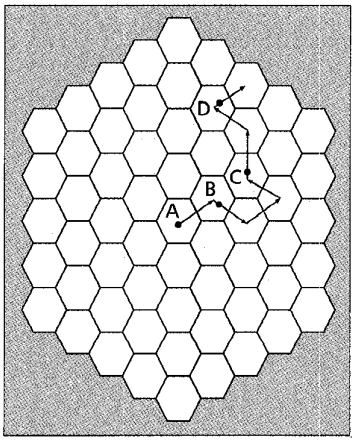


■ **Figure 11.** *Registration area for a)* k = *2; and b)* k = *4.*

Three location update schemes are examined in [18]:

Time-based — An MT performs location updates periodically at a constant time interval ΔT. Figure 12 shows the path of an MT. If a location update occurred at location A at time 0, subsequent location updates will occur at locations B, C, and D if the MT moves to these locations at times ΔT, $2\Delta T$, and $3\Delta T$, respectively.

Movement-based — An MT performs a location update whenever it completes a predefined number of movements across cell boundaries (this number is referred to as the *movement threshold*). Figure 13 shows the same path as in Fig. 12. Assuming a movement threshold of 3 is used, the MT performs location updates at locations B and C, as shown in Fig. 13.



■ **Figure 12.** *Time-based location update scheme.*

Distance-based — An MT performs a location update when its distance from the cell where it performed the last location update exceeds a predefined value (this distance value is referred to as the *distance hreshold*). Figure 14 shows the same path as in Fig. 12. A location update is performed at location B where the distance of the MT from location A exceeds the threshold distance (the distance from location A to the thick solid line shown in Fig. 14 is equal to the threshold distance).

The performance of the above schemes are evaluated in [18] based on a simplified one-dimensional movement

model. Results demonstrate that the distance-based scheme produces the best performance, but its implementation incurs the highest overhead. For the time-based and movement-based schemes, the MT has to keep track of the time elapsed and the number of movements performed, respectively, since the last location update. This can be achieved by simply implementing a timer or a movement counter at the MT. The distance-based scheme, however, assumes that the MTs have knowledge of the distance relationships among all cells. The network must be able to provide this information to each MT in an efficient manner.

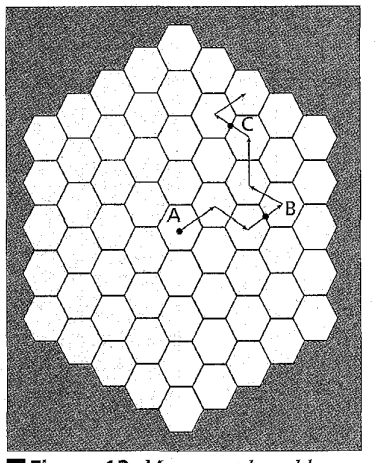


Figure 13. *Movement-based location update scheme.*

A distance-based location update scheme is considered in [19]. The authors introduce an iterative algorithm that can generate the optimal threshold distance which results in the minimum cost. When an incoming call arrives, cells are paged in a shortest-distance-first order such that cells closest to the cell where the last location update occurred are polled first. The delay in locating an MT is therefore proportional to the distance traveled since the last location update. Results demonstrated that, depending on the mobility and call arrival parameters, the optimal movement threshold varies widely. This demonstrates that location update schemes should be per-user-based and dynamically adjusted according to the current mobility and call arrival pattern of the user. However, the number of iterations required for this algorithm to converge varies depending on the mobility and call arrival parameters considered. Determining the optimal threshold distance may require significant computation at the MT.

A dynamic time-based location update scheme is introduced in [20]. The location update time interval is determined after each movement based on the probability distribution of the call interarrival time. This scheme does not make any specific assumptions on the mobility pattern of the MTs, and the shortest-distance-first paging scheme as described in [19] is used. It is demonstrated that the results obtained are close to the optimal results given in [21]. The computation required by this scheme is low, and it is therefore feasible for application in MTs that have limited computing power. Similar to the scheme described in [19], the drawback is that paging delay is not constrained. The time required to locate an MT is directly proportional to the distance traveled since the last location update

In [22], paging subject to delay constraints is considered. The authors assume that the network coverage area is divided into location areas, and the probability that an MT is residing in a location area is given. It is demonstrated that when delay is unconstrained, the polling cost is minimized by sequentially searching the location areas in decreasing order of probability of containing the MT. For constrained delay, the authors obtain the optimal polling sequence that results in the minimum polling cost. However, the authors assume that the probability distribution of user location is provided. This probability distribution may be user-dependent. A location update and paging scheme that facilitates derivation of this probability distribution is needed in order to apply this paging scheme. Besides, the trade-off between the costs of location update and paging is not considered in [22].

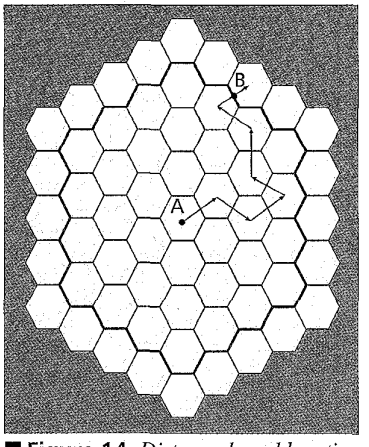


Figure 14. *Distance-based location update scheme.*

Recently, location update and paging subject to delay constraints was considered in [21]. Similar to [19], the authors consider the distance-based location update scheme. However, paging delay is constrained such that the time required to locate an MT is smaller than or equal to a predefined maximum value. When an incoming call arrives, the residing area of the MT is partitioned into a number of subareas. These subareas are then polled sequentially to locate the MT. By limiting the number of polling areas to a given value such as N, the time required to locate a mobile is smaller than or equal to the time required for N polling operations. Given the mobility and call arrival parameters, the threshold distance, and the maximum delay, an analytical model is introduced that generates the expected cost of the proposed scheme. An iterative algorithm is then used to locate the optimal threshold distance that results in lowest cost. It is demonstrated that the cost is lowest when the maximum delay is unconstrained. However, by slightly increasing the maximum delay from its minimum value of 1, the cost is significantly lowered.

Another scheme using the movement-based location update is reported in [23]. Similar to [21], paging delay is confined to a maximum value. Movement-based location update schemes have the advantage of simple implementation. The MTs do not need to know the cell configuration of the network. The scheme introduced in [23] is feasible for use in current PCS networks.

Conclusions and Future Research

In this article, we discussed a number of research results in mobility management. While it is true that each of the proposed schemes can improve the IS-41 mobility management strategy to a certain extent, it is difficult to select a scheme that clearly outperforms the others under all system parameters. In most cases, the performance of the proposed schemes exceeds that of the IS-41 only under certain mobility and call arrival parameters. When a different set of parameters is used, the performance may be changed significantly. It is, however, possible for us to make several general observations on the design of mobility management schemes. In the following subsections, we will discuss these observations for

- Location registration and call delivery
- Location update and paging

Location Registration and Call Delivery

As described previously, recent research efforts in location registration and call delivery are based on either the centralized or distributed database architectures. The centralized approach records the location information of all MTs in the centralized HLR. Signaling messages are exchanged between the current location of an MT and the HLR during location registration and call delivery. As the num-

ber of MTs increases, the signaling traffic may significantly degrade the performance of the PCS network. One undesirable consequence is that the connection setup delay may become very high. On the other hand, an advantage of the centralized approach is that the number of database updates and queries for location registration and call delivery is relatively small. This minimizes the delay due to database accesses. The distributed database approach has the advantage that database accesses are localized. An update or query to a faraway database is executed only when necessary. However, the number of database accesses required for location registration and call delivery is significantly increased from the centralized approach. Careful design is needed to ensure that database accesses will not significantly increase the signaling delay.

Based on these observations, it is likely that the ideal architecture should lie between the centralized and fully distributed approaches. In fact, in order to attain better cost effectiveness, most of the ongoing research efforts either try to:

- Increase the distribution of location information under a centralized database architecture, such as the results reported in [12, 13]
- Limit the distribution of location information in a distributed database architecture, such as the results reported in [15, 16]

Besides, mobility and call arrival patterns vary among users, it is highly desirable that the location registration and call delivery procedures can be adjusted dynamically on a per-user basis. Dynamic schemes usually require the on-line collection and processing of data. This may consume significant computing power, and careful design is necessary so that the computation can effectively be supported by the network.

Future research in location registration and call delivery should focus on the design of network architectures that combine, to a certain degree, the centralized and fully distributed approaches. Dynamic schemes for limiting or enhancing the distribution of location information on a per-user basis should be considered.

Location Update and Paging

As discussed in previously, there are two types of location update and paging schemes: static and dynamic. Static schemes have the disadvantage that they cannot be adjusted according to the parameter of the individual user. For example, under the RA-based location update scheme, the RA size most suitable for one user may be ineffectual for another user. Most of the recent research efforts focus on the development of dynamic location update and paging schemes. Dynamic schemes allow on-line adjustments based on the characteristics of each individual MT. For example, when the distance-based location update scheme is used, a different distance threshold can be assigned to each MT based on its mobility and call arrival pattern. However, some of these schemes require information, such as the distance between cells, that is not generally available to the MTs. Besides, the operation of dynamic schemes may require significant computing power. Implementation of a computation-intensive scheme in an MT may not be feasible.

Future research should focus on the design of dynamic location update and paging schemes that are simple to implement. Most of the schemes are based on simplified assumptions. For example, the random walk mobility model is used in a number of proposed schemes such that the direction of travel of each MT is uniformly distributed. We believe that most of these schemes can be improved by considering more realistic assumptions.

References

[1] EIA/TIA, "Cellular Radio-Telecommunications Intersystem Operations," Tech. Rep. IS-41 Revision C, 1995.

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

[3] Y. B. Lin and S. K. DeVries, "PCS Network Signaling Using SS7," *IEEE Commun. Mag*, June 1995., pp. 44–55.

[4] A. R. Modarressi and R. A. Skoog, "Signaling System 7: A Tutorial," *IEEE Commun. Mag.*, vol. 28, no. 7, July 1990, pp. 19–35.

[5] D. R. Wilson, "Signaling System No. 7, IS-41 and Cellular Telephony Networking," *Proc. IEEE*, vol. 80, no. 4, Apr. 1992, pp. 664–52.

[6] S. Mohan and R. Jain, "Two User Location Strategies for Personal Communications Services," *IEEE Pers. Commun.*, 1st Qtr. 1994, pp. 42–50.

[7] C. N. Lo, S. Mohan, and R. S. Wolff, "Performance Modeling and Simulation of Data Management for Personal Communications Applications," *Proc. IEEE PIMRC '92*, Nov. 1992, pp. 1210–14.

[8] C. N. Lo, R. S. Wolff, and R. C. Bernhardt, "An Estimate of Network Database Transaction Volume to Support Personal Communications Services," *Proc. Int'l. Conf. on Universal Pers. Commun.*, 1992, pp. 236–41.

[9] R. Jain, Y. B. Lin, and S. Mohan, "A Caching Strategy to Reduce Network Impacts of PCS," *IEEE JSAC*, vol. 12, no. 8, Oct. 1994, pp. 1434–44.

[10] Y. B. Lin, "Determining the User Locations for Personal Communications Services Networks," *IEEE Trans. Vehicular Tech.*, vol. 43, no. 3, Aug. 1994, pp. 466–73.

[11] N. Shivakumar and J. Widom, "User Profile Replication for Faster Location Lookup in Mobile Environments," *Proc. ACM MOBICOM '95*, Nov. 1995, pp. 161–69.

[12] R. Jain and Y. B. Lin, "An Auxiliary User Location Strategy Employing Forwarding Pointers to Educe Network Impact of PCS," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, July 1995, pp. 197–210.

[13] J. S. M. Ho and I. F. Akyildiz, "Local Anchor Scheme for Reducing Location Tracking Costs in PCNs," to appear in *IEEE/ACM Trans. Networking*, Oct. 1996.

[14] J. Z. Wang, "A Fully Distributed Location Registration Strategy for Universal Personal Communication Systems," *IEEE JSAC*, vol. 11, no. 6, Aug. 1993, pp. 850–60.

[15] B. R. Badrinath, T. Imielinski and A Virmani, "Locating Strategies for Personal Communication Network," *Proc. Workshop on Networking of Pers. Commun. Applications*, Dec. 1992.

[16] V. Anantharam *et al.*, "Optimization of a Database Hierarchy for Mobility Tracking in a Personal Communications Network," *Performance Evaluation*, vol. 20, no. 1–3, May 1994, pp. 287–300.

[17] H. Xie, S. Tabbane and D. Goodman, "Dynamic Location Area Management and Performance Analysis," *Proc. IEEE VTC '93*, May 1993, pp. 536–39.

[18] A. Bar-Noy, I. Kessler and M. Sidi, "Mobile Users: To Update or Not to Update?" *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, July 1995, pp. 175–86.

[19] U. Madhow, M. L. Honig and K. Steiglitz, "Optimization of Wireless Resources for Personal Communications Mobility Tracking," *Proc. IEEE INFOCOM '94*, June 1994, pp. 577–84.

[20] I. F. Akyildiz and J. S. M. Ho, "Dynamic Mobile User Location Update for Wireless PCS Networks," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, July 1995, pp. 187–96.

[21] J. S. M. Ho and I. F. Akyildiz, "A Mobile User Location Update and Paging Mechanism Under Delay Constraints," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 4, Dec. 1995, pp. 413–25.

[22] C. Rose and R. Yates, "Minimizing the Average Cost of Paging Under Delay Constraints," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, July 1995, pp. 211–19.

[23] I. F. Akyildiz, J. S. M. Ho and Y. B. Lin, "Movement Based Location Update and Selective Paging Schemes," to appear in *IEEE/ACM Trans. Networking*, Aug. 1996.

Biographies

IAN F. AKYILDIZ [F '96] received his B.S., M.S., and Ph.D. degrees in computer engineering from the University of Erlangen-Nuernberg, Germany, in 1978, 1981, and 1984, respectively. Currently, he is a full professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. He has held visiting professorships at the Universidad Tecnica Federico Santa Maria, Chile, Universite Pierre et Marie Curie (Paris VI) and Ecole Nationale Superieure Telecommunications in Paris, France. His current research interests are in ATM and wireless networks.

JOSEPH S. M. HO [SM] received the B.S.E.E and M.S.E.E degrees from the University of Washington at Seattle in 1987 and 1989, respectively. He received the Ph.D. degree from Georgia Institute of Technology in 1996. His current research interests include design and analysis of PCS networks, traffic control for ATM networks, performance evaluation, and parallel simulation. He is a member of Tau Beta Pi.