Mobility Management in Next-Generation Wireless Systems

IAN F. AKYILDIZ, FELLOW, IEEE, JANISE MCNAIR, STUDENT MEMBER, IEEE, JOSEPH S. M. HO, MEMBER, IEEE, HÜSEYIN UZUNALIOĞLU, MEMBER, IEEE, AND WENYE WANG, STUDENT MEMBER, IEEE

This paper describes current and proposed protocols for mobility management for public land mobile network (PLMN)-based networks, mobile Internet protocol (IP), wireless asynchronous transfer mode (ATM), and satellite networks. The integration of these networks will be discussed in the context of the next evolutionary step of wireless communication networks. First, a review is provided of location management algorithms for personal communication systems (PCS) implemented over a PLMN network. The latest protocol changes for location registration and handoff are investigated for Mobile IP, followed by a discussion of proposed protocols for wireless ATM and satellite networks. Finally, an outline of open problems to be addressed by the next generation of wireless network service is discussed.

***Keywords*—** *Handoff, IMT 2000, location management, mobile Internet protocol (IP), mobility management, paging, personal communication systems (PCS), public land mobile network (PLMN), public switched telephone network (PSTN), satellite, wireless asynchronous transfer mode (ATM).*

I. INTRODUCTION

The commercial proliferation of cellular voice and limited data service has created a great demand for mobile communications and computing. Current voice, fax, e-mail, and paging services will give way to data transfer, video conferencing, image transfer, and video delivery. Achieving such an advanced level of tetherless mobile multimedia service requires the development of a wireless network that can provide not only the integrated services, but which can also provide dynamic relocation of mobile terminals. As a result, next-generation mobile communication systems are currently being researched worldwide. Third generation systems, such as the International Mobile Telecommunication System 2000 (IMT 2000) network, the Universal Mobile Telecommunication System (UMTS), and the Future Public Land Mobile Telecommunication System (FPLMTS) are based on a combination of integrated fixed and wireless mobile services that form a global personal communication network [40]. In recent years, these systems have been considered for standardization by the International Telecommunication Union (ITU). The next generation of wireless communication is also based on a global system of fixed and wireless mobile services, but it extends global service to include the integration of heterogeneous services across network providers and network backbones as well as geographical regions. Thus, the future will ensure interoperability between various wireless networks across the globe [82].

The demand for mobile service has motivated research in updating existing high-speed wireline (fixed) communication networks with wireless communication techniques. Several alternative backbone networks that are fostering current research activity are: the public land mobile networks (PLMN), mobile Internet protocol (Mobile IP) networks, wireless asynchronous transfer mode (WATM) networks, and low Earth orbit (LEO) satellite networks. Standardizing these efforts is the goal of working groups within the European Telecommunications Standards Institute (ETSI), the International Telecommunication Union-Telecommunications Standardization Sector (ITU-T), the Internet Engineering Task Force (IETF), and the WATM Working Group of the ATM Forum.

Regardless of the network, one of the most important and challenging problems for tetherless communication and computing is mobility management [74]. Mobility management enables telecommunication networks to locate roaming terminals for call delivery and to maintain connections as the terminal is moving into a new service area. Thus, mobility management supports mobile terminals, allowing users to roam while simultaneously offering them incoming calls and supporting calls in progress.

In this paper, we describe current and proposed protocols for mobility management for the PLMN, Mobile

Manuscript received September 20, 1998; revised May 19, 1999. This work was supported by the Department of Defense (DoD), National Security Agency under Grant MDA904-97-C-1105-0003.

I. F. Akyildiz, J. McNair, and W. Wang are with the Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

J. S. M. Ho is with Nortel Networks, Wireless Solutions, Richardson, TX 75082 USA.

H. Uzunalioğlu is with Bell Laboratories, Lucent Technologies, Holmdel, NJ 07733 USA.

Publisher Item Identifier S 0018-9219(99)06112-5.

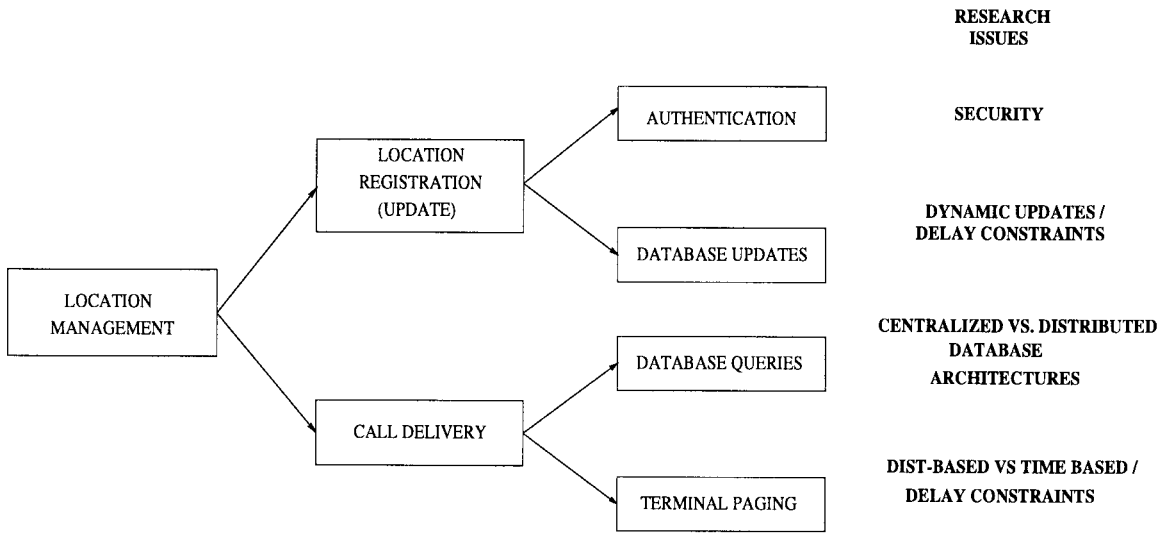


Fig. 1. Location management operations.

IP, WATM, and satellite networks and develop a discussion for mobility management for the next generation of wireless communication. This paper attempts a detailed investigation of mobility management protocols for each network. An abridged version of this paper can be found in [13]. Section II defines the driving concepts behind mobility management. Then, Section III presents the future wireless network architecture. Section IV reviews the location management process for PLMN-based networks, while Section V investigates the latest protocol changes for Mobile IP. In Section VI we provide a selection of proposed protocols for WATM, while a discussion of mobility management for satellite networks is presented in Section VII. Finally, the paper concludes with a discussion of open problems faced by the next generation of wireless systems.

II. Mobility Management

Mobility management contains two components: location management and handoff management.

A. Location Management

Location management is a two-stage process that enables the network to discover the current attachment point of the mobile user for call delivery, as shown in Fig. 1. The first stage is location registration (or location update). In this stage, the mobile terminal periodically notifies the network of its new access point, allowing the network to authenticate the user and revise the user's location profile. The second stage is call delivery. Here the network is queried for the user location profile and the current position of the mobile host is found.

Current techniques for location management involve database architecture design and the transmission of signaling messages between various components of a signaling network. As the number of mobile subscribers increases, new or improved schemes are needed to support effectively a continuously increasing subscriber population. Other issues include: security; dynamic database updates; querying delays; terminal paging methods; and paging delays. Fig. 1 associates these research issues with their respective location management operation. Since location management deals with database and signaling issues, many of the issues are not protocol dependent and can be applied to various networks such as PLMN-based networks, the public switched telephone network (PSTN), ISDN, IP, Frame Relay, X.25, or ATM networks, depending on the requirements.

B. Handoff Management

Handoff (or handover) management enables the network to maintain a user's connection as the mobile terminal continues to move and change its access point to the network. The three-stage process for handoff first involves initiation, where either the user, a network agent, or changing network conditions identify the need for handoff. The second stage is new connection generation, where the network must find new resources for the handoff connection and perform any additional routing operations. Under network-controlled handoff (NCHO), or mobile-assisted handoff (MAHO), the network generates a new connection, finding new resources for the handoff and performing any additional routing operations. For mobile-controlled handoff (MCHO), the mobile terminal finds the new resources and the network approves. The final stage is data-flow control, where the delivery of the data from the old connection path to the new connection path is maintained according to agreed-upon service guarantees. The handoff management operations are presented in Fig. 2.

Handoff management includes two conditions: intracell handoff and intercell handoff. Intracell handoff occurs when the user moves within a service area (or cell) and experiences signal strength deterioration below a certain threshold that results in the transfer of the user's calls to new radio channels of appropriate strength at the same base station (BS). Intercell handoff occurs when the user moves into an adjacent cell and all of the terminal's connections must be transferred to a new BS. While performing handoff,

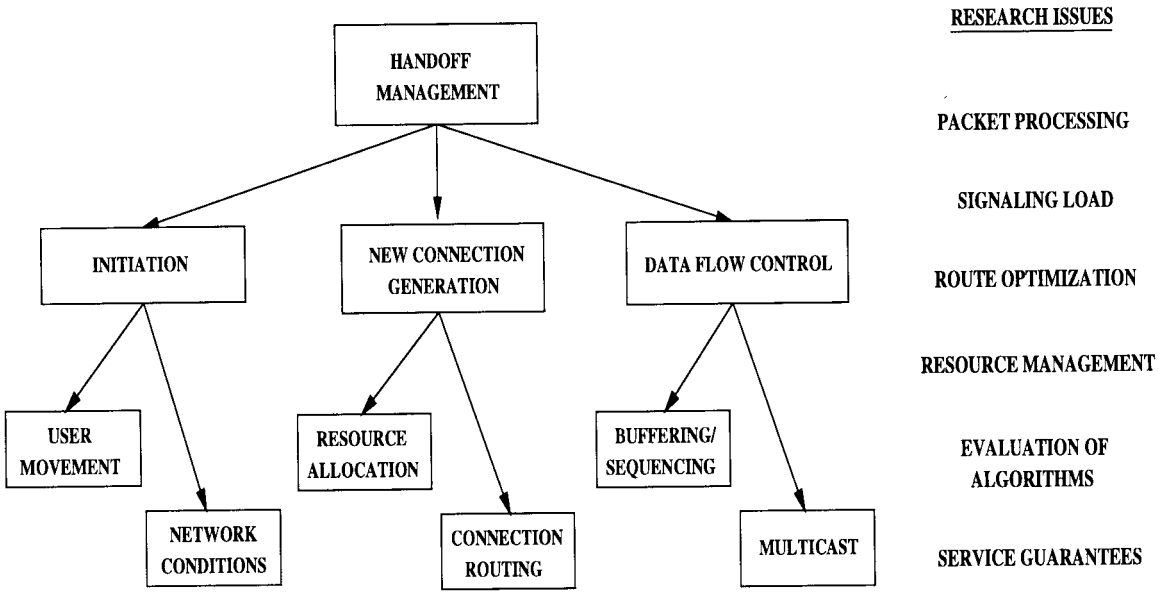


Fig. 2. Handoff management operations.

the terminal may connect to multiple BS's simultaneously and use some form of signaling diversity to combine the multiple signals. This is called soft handoff. On the other hand, if the terminal stays connected to only one BS at a time, clearing the connection with the former BS immediately before or after establishing a connection with the target BS, then the process is referred to as hard handoff [40]. Handoff management research concerns issues such as: efficient and expedient packet processing; minimizing the signaling load on the network; optimizing the route for each connection; efficient bandwidth reassignment; evaluating existing methods for standardization; and refining quality of service for wireless connections. Fig. 2 lists these issues for the handoff management operations.

Although this paper describes individual protocols for location and handoff management according to the type of backbone network, the next generation promises to allow these networks—and thus, their mobile operations—to interoperate. Unlike the location management protocols, the reliance of handoff protocols on routing, resource management, and packet delivery make these algorithms very network protocol dependent.

An additional dependency to be considered for interoperability is the dependency of the user on the local wireless network interfaces and infrastructure. Future communications networks will require the development of a standardized wireless network architecture, enabling any user to employ a somewhat common methodology in order to access regional, national, and global services as efficiently as those on the local level. In Section III, we discuss the basic building blocks of future wireless network architectures.

III. Future Wireless Network Architecture

The next evolutionary step toward personal communication services will provide an architectural and structural basis that will allow evolving networks to implement free circulation of terminals, personal mobility, and network service portability [82]. In this section we describe some of the architecture and system specifications currently outlined for the IMT 2000 by the ITU. These specifications include: a hierarchical cell structure (HCS); global roaming; and an expanding radio spectrum [29].

A. Hierarchical Cell Structure

HCS will cover all of the proposed operating environments of the mobile user. It will support radio environments that range from high capacity picocells, to urban terrestrial micro- and macrocells, to large satellite cells, as shown in Fig. 3. Due to the potential of satellite links performing as traffic congestion relief and global extensions to terrestrial networks, network capacity will potentially increase—supporting more subscribers and greater traffic volumes without requiring additional radio spectrum for the terrestrial networks [69].

The mobile user will access the wireless network using a device called a mobile terminal (MT). This terminal will use radio channels to communicate BS's (also referred to as base transceiver stations) to gain access to the terrestrial (PLMN, ATM, Internet) network. In the satellite network, the MT will communicate with fixed Earth stations (FES), which govern wireless traffic for satellite terminals or with the satellite itself [90]. Dual mode terminals will communicate over both the terrestrial and satellite networks.

Each cell will have one dedicated BS and a corresponding broadcast channel. Channel use is managed by the BS, which converts the network signaling traffic and data traffic to the radio interface for communication with the MT. The BS will also transmit paging messages to the MT and measure the link quality to perform handoffs to other cells. The MT will be able to roam freely within an area consisting of multiple cells called the location area (LA).

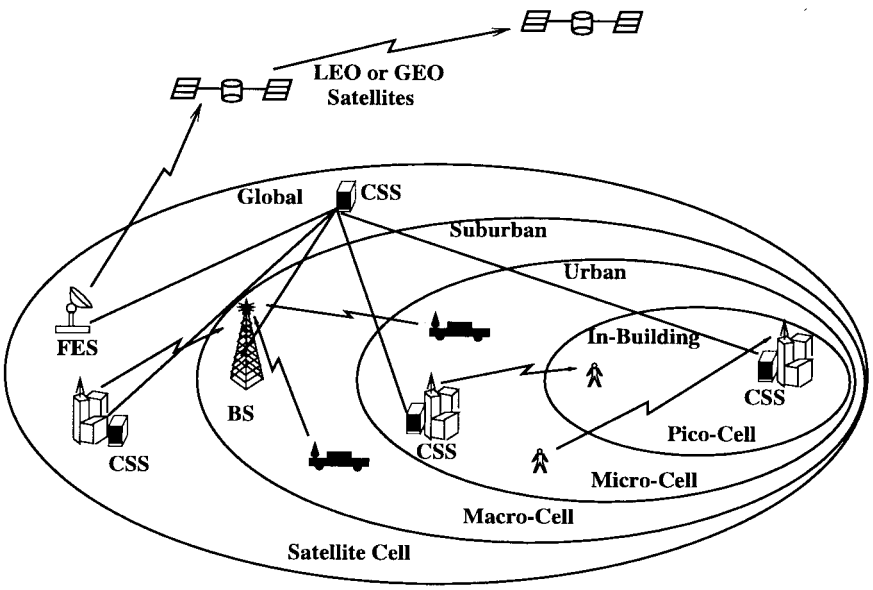


Fig. 3. Next-generation heterogeneous network services.

While within this area, the MT will not be required to update the location information stored in fixed network location distributed databases. However, the network will keep track of the current LA for each mobile that is powered on. While in stand-by mode, the MT will be tuned to the serving broadcast channel associated with the serving cell/BS. In order to make a call, the MT will tune to one of the channels assigned to the serving cell.

Finally, a cell site switch (CSS) will govern one or more BS's. This switch will provide access to the serving mobile network (PLMN, Internet, ATM, or satellite). The CSS will also manage the radio resources provide mobility management control functions, such as location update and handoff to manage global roaming.

B. Global Roaming

The next-generation wireless networks will begin to implement terminal mobility, personal mobility, and service provider portability. Terminal mobility refers to the ability of the network to route calls to the MT regardless of its point of attachment to the network, while personal mobility is the ability of the user to access their personal services independent of the their attachment point or terminal. ITU specifications call for a universal personal telecommunication (UPT) number that will distinguish the user from the terminal itself [82]. Service provider portability allows the user and/or the MT to move beyond regional mobile networks. The user will be able to receive their personalized end-to-end services regardless of their current network—within the limits of the visited network's service offerings. For example, an environmental research scientist with Mobile IP services in Atlanta will be able to travel to the rural rain forests of South America and still receive at least a subset of his/her personal services via the resident satellite network. This idea is illustrated in Fig. 4. The wireless user terminals are connecting to the unified wireless network via their resident networks, which then must identify the data as belonging to another network type, and then provide access, via an IMT2000 subsystem, to the user's prescribed network services. In areas where access to their home networks is readily available, the terminals can use direct connections to their home backbone services. This freedom requires future wireless networks to interoperate and transport heterogeneous traffic over both wireless and wireline networks. Although this area has already been investigated for Mobile IP over ATM, the scope of the future opens opportunities for any combination of wireless network data being transmitted over any other wireline backbone.

This level of global mobile freedom will also require the coordination of a wide range of service providers, compatibility of backbone networks, and network operator agreements. Whereas such agreements are currently governed by commercial contracts, the next generation will facilitate this process by developing global roaming agreements between different countries, regions and service providers, and by increasing available radio spectrum based on these international agreements.

C. Radio Spectrum

The ITU is encouraging national regulators to follow its guidelines in order to promote harmonized utilization of the radio spectrum and to facilitate the development of global personal communication systems (PCS). In the past, frequencies have been allocated only in limited amounts for specific services such as paging, cellular, mobile data, and private mobile radio. The future radio spectrum will include all of these categories and will standardize a pool of frequencies which could be managed dynamically to meet global market needs. This will include the technological developments required to make such dynamic spectrum allocation from the spectrum pool.

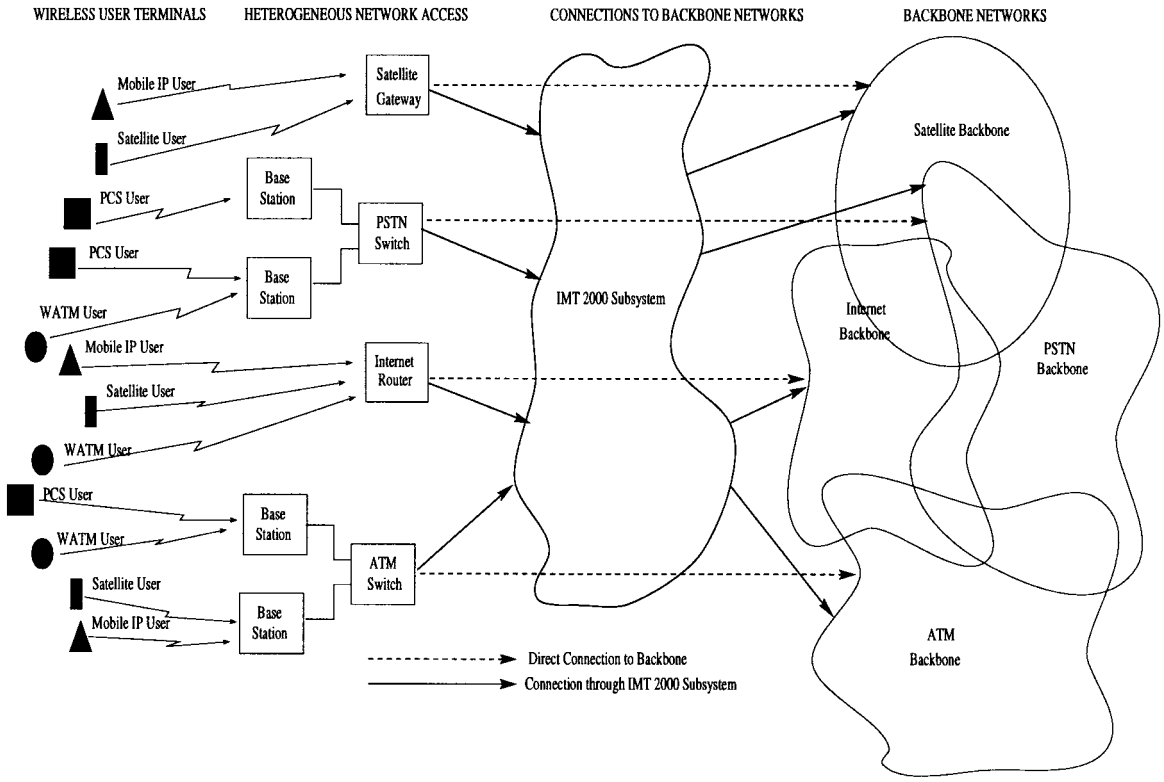


Fig. 4. Next-generation network service portability.

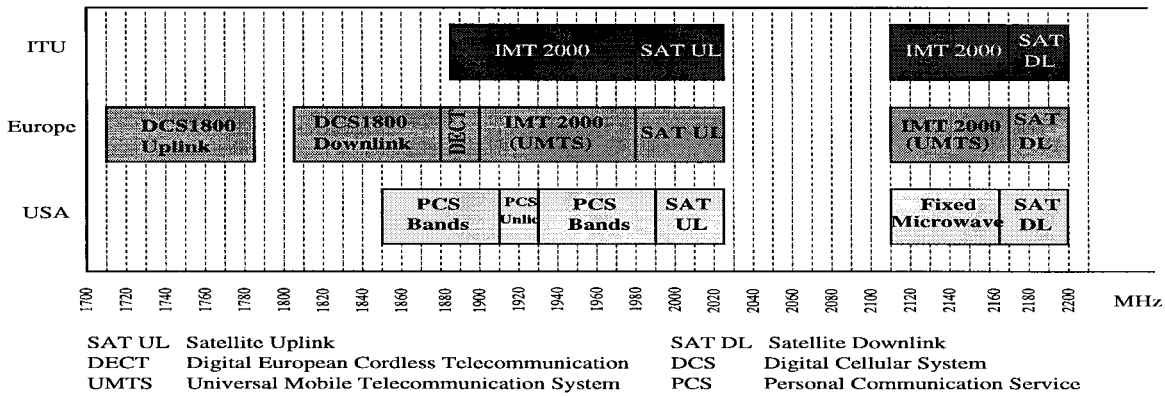


Fig. 5. Frequency allocation.

The preliminary frequency allocation, as determined in the 1992 World Administrative Radio Conference, is shown in Fig. 5. A 170-MHz section of bandwidth was reserved for terrestrial use, while 60-MHz bandwidth was reserved for satellite. The total spectrum was 1885–2025 MHz and 2110–2200 MHz, while the satellite band was 1980–2010 MHz and 2170–2200 MHz. (The frequency gaps between 2025–2170 MHz and beyond 2200 MHz are reserved for other services such as remote sensing, cable TV relay service, electronic news gathering, and space research and operation.) In 1995, the ITU World Radio Conference changed the frequency assignments. The satellite allocation for Region 2 (the Americas and the Caribbean) was revised to the 1990–2025-MHz and 2160–2200-MHz frequency bands. This change will make it difficult for the U.S. service providers to support mobile terminals from other regions that use the mobile satellite service. The revised assignment will remain in effect until the next scheduled conference in 1999.

Since the U.S. Federal Communications Commission (FCC) has not yet decided on the allocation of additional spectrum for next generation cellular systems, existing cellular and PCS service providers will have to allocate some of their current spectrum as needed. Start up of IMT 2000 bands are proposed for Japan by the year 2000 and also for Europe by the year 2002. All frequencies, except those for the Personal Handiphone System (PHS) and Digital European Cordless Telecommunications (DECT),

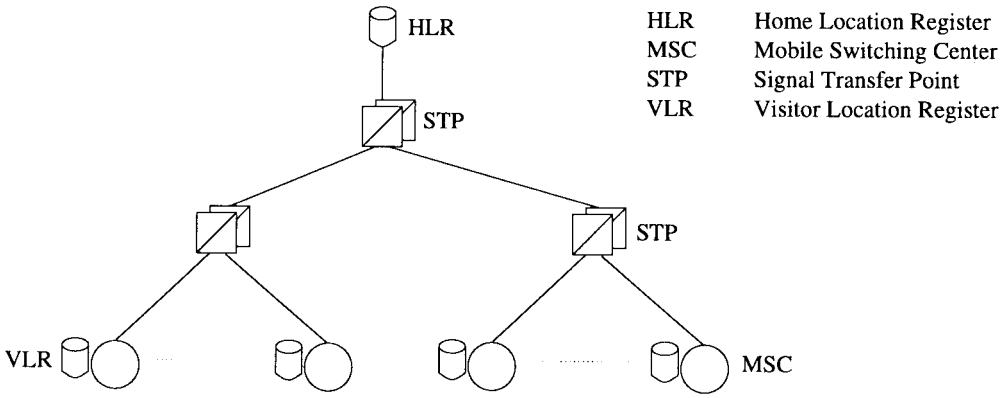


Fig. 6. SS7 signaling network.

will be available for use in Japan by 2003 and in Europe by 2005. Licensing rules and procedures for IMT 2000 are under development in Europe.

The agreement among nations, regions, and network service providers to work toward global mobility will require the interworking of location information and authentication procedures—on the user and the terminal level, as well as network cooperation for the delivery of services as the MT's move. An investigation of current location and handoff management processes opens the door for research into unifying these processes under the next generation. Next we address specific schemes for location management for PCS, with consideration of a PLMN-based backbone network. We then discuss both recent research and open problems for PCS networks.

IV. Mobility Management for the PLMN

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 the network administration and it cannot easily be performed by a user. Incoming calls for a particular terminal are always routed to its associated location as there is no distinction between a terminal and its location. In contrast, mobile terminals (MT's) are free to travel and thus 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 the MT and the call delivery process becomes more complex. The current methods (IS-41, GSM MAP) for PLMN location management strategies require each MT to register its location with the network periodically. In order to perform the registration, update, and call delivery operations described above, the network stores the location information of each MT in the location databases. Then the information can be retrieved for call delivery.

A. Current Location Management Protocols

Current schemes for PLMN location management are based on a two-level data hierarchy such that two types of network location 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 network and a user is permanently associated with an HLR in his/her subscribed network. Information about each user, such as the types of services subscribed and location information, are stored in a user profile located at the HLR. The number of VLR's and their placements vary among networks. Each VLR stores the information of the MT's (downloaded from the HLR) visiting its associated area.

Network management functions, such as call processing and location registration, are achieved by the exchange of signaling messages through a signaling network. Signaling System 7 (SS7) [67], [79], [122] is the protocol used for signaling exchange, and the signaling network is referred to as the SS7 network. The type of CSS currently implemented for the PLMN is known as a mobile switching center (MSC). Fig. 6 shows the SS7 signaling network which connects the HLR, the VLR's, and the MSC's in a PLMN-based network. The signal transfer points (STP's) as shown in Fig. 6 are responsible for routing signaling messages within the SS7 network. For reliability reason, the STP's are installed in pairs.

There are currently two commonly used standards for location management in the PLMN: the Electronic and Telephone Industry Associations EIA/TIA Interim Standard 41 (IS-41) [1], [80] and the Global System for Mobile Communications (GSM) Mobile Application Part (MAP) [4], [80], [81]. The IS-41 scheme is commonly used in North America for the Advanced Mobile Phone System (AMPS) [2], IS-54 [3], IS-136, and the Personal Access Communication System (PACS) networks, while the GSM MAP is mostly used in Europe for GSM and Digital Cellular System-1800 (DCS-1800) and Personal Communication Service-1900 (PCS-1900) networks. Both standards are based on the two-level database hierarchy.

As mentioned previously, location management includes two major tasks: location registration (or update) and call delivery (see Fig. 1). For PLMN, the location registration procedures update the location databases (HLR and VLR's) and authenticate the MT when up-to-date location information of an MT is available. The call delivery procedures locate the MT based on the information available at the HLR and the VLR's when a call for an MT is initiated. The IS-41 and the GSM MAP location management strategies

are very similar to each other. 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 limitation, the presentation of this paper is based primarily on the IS-41 standard. Interested readers may refer to [4], [80], [81] or detailed descriptions of the GSM MAP mobility management strategy.

1) Location Registration: In order to correctly deliver calls, the PLMN must keep track of the location of each MT. As described previously, location information is stored in two types of databases, VLR and HLR. As the MT's 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, the databases are periodically updated through the process called location registration.

Location registration is initiated by an MT when it reports its current location to the network. We call this reporting process location update. Current systems adopt an approach such that the MT performs a location update whenever it enters a new LA. Recall that each LA consists of a number of cells and, in general, all BTS's belonging to the same LA are connected to the same MSC.

When an MT enters an LA, if the new LA belongs to the same VLR as the old LA, the record at the VLR is updated to record the ID of the new LA. Otherwise, if the new LA belongs to a different VLR, a number of extra steps are required to: 1) register the MT at the new serving VLR; 2) update the HLR to record the ID of the new serving VLR; and 3) deregister the MT at the old serving VLR. Fig. 7 shows the location registration procedure when an MT moves to a new LA. The following is the ordered list of tasks that are performed during location registration.

1) The MT enters a new LA and transmits a location update message to the new BS.
2) The BS forwards the location update message to the MSC which launches a registration query to its associated VLR.
3) The VLR updates its record on the location of the MT. If the new LA 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.

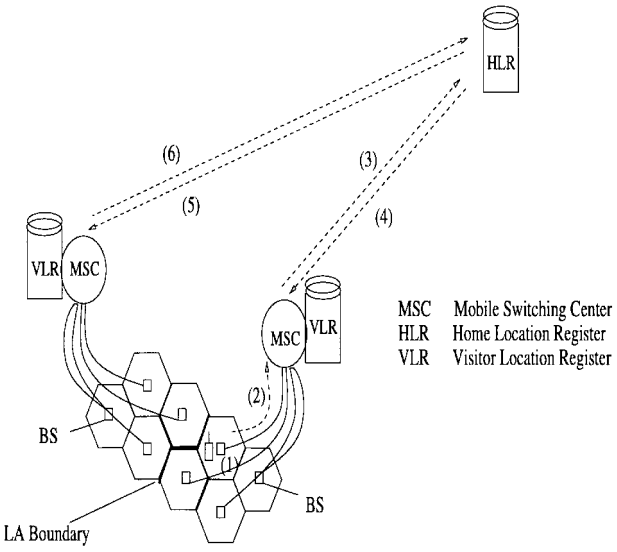


Fig. 7. Location registration procedures.

Depending on the distance between the current and the home locations of the MT, in steps 3)–6), the signaling messages may have to go through several intermediate STP's before reaching their destinations. For example, a user who subscribes to wireless services in Atlanta will normally be assigned to an HLR located in the Atlanta area. When this user is roaming in London, each location update performed by his/her mobile phone will result in the transmission of four transatlantic SS7 messages [messages (3)–(6) as shown in Fig. 7]. These messages may transverse a number of STP's in the SS7 network before reaching their destinations, which generate additional load to the network elements and the transmission links. The location registration may, therefore, result in 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 the signaling cost are discussed in Section IV-B.

2) Call Delivery: Two major steps are involved in call delivery: 1) determining the serving VLR of the called MT and 2) locating the visiting cell of the called MT. Locating the serving VLR of the MT involves the following database lookup procedure (see Fig. 8).

1) The calling MT sends a call initiation signal to the serving MSC of the MT through a nearby BS.
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 forward 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 forward this information to the MSC of the calling MT.

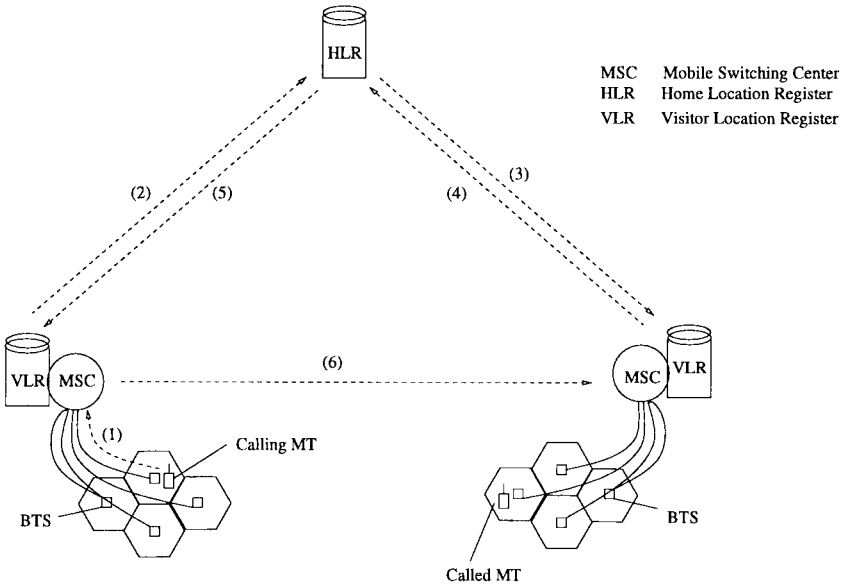


Fig. 8. Call delivery procedures.

6) The calling MSC requests a call set up 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 LA and there are more than one cells in each LA, a mechanism is therefore necessary to determine the cell location of the called MT. In current PLMN networks, this is achieved by a paging (or alerting) procedure, such that polling signals are broadcast to all cells within the residing LA of the called MT. On receiving the polling signal, the MT sends a reply which allows the MSC to determine its current residing cell. As the number of MT's increases, sending polling signals to all cells in an LA whenever a call arrives may consume excessive wireless bandwidth. We describe a number of proposed paging mechanisms for reducing the paging cost in Section IV-C.

B. Location Registration and Call Delivery Research

Location registration involves the updating of 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 (refer to Fig. 1). These can be costly processes, especially when the MT is located far away from its assigned HLR. For example, if the MT is currently roaming at 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 LA 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 at Atlanta before it finds out that the called MT is located in the same area as the caller. As the number of mobile subscribers keeps increasing, the volume of signaling traffic generated by location management is extremely high [70], [71]. Methods for reducing the signaling traffic are therefore needed.

Research in this area generally falls into two categories. In the first category, extensions to the IS-41 location management strategy are developed which aim at improving 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 PLMN networks without major modification. These schemes are based on centralized database architectures inherited from the IS-41 standard. Another category of research results lies in completely new database architectures that require a new set of schemes for location registration and call delivery. Most of these schemes are based on distributed database architectures. Some additional research efforts involve: the reverse virtual call set up—a new scheme for delivering mobile-terminated calls [49], an optimal routing scheme based on the ratio of source messaging to location update rates [124], and a single registration strategy for multitier PCS systems [66]. In what follows, we discuss centralized versus distributed database architectures. A discussion of these two database architectures can be found in Section IV-D1.

1) Centralized Database Architectures: This solution consists of the two-tier IS-41-based database structure with additional optimizations that aim to reduce the location management cost.

a) Dynamic hierarchical database architecture: The first centralized database architecture is the dynamic hierarchical database architecture presented in [44]. The proposed architecture is based on that of the IS-41 standard with the addition of a new level of databases called directory registers (DR's). Each DR covers the service area of a number of MSC's. The primary function of the DR's is to compute periodically and store the location pointer configuration for the MT's in its service area. Each MT has its unique pointer configuration and three types of location pointers are available at the DR:

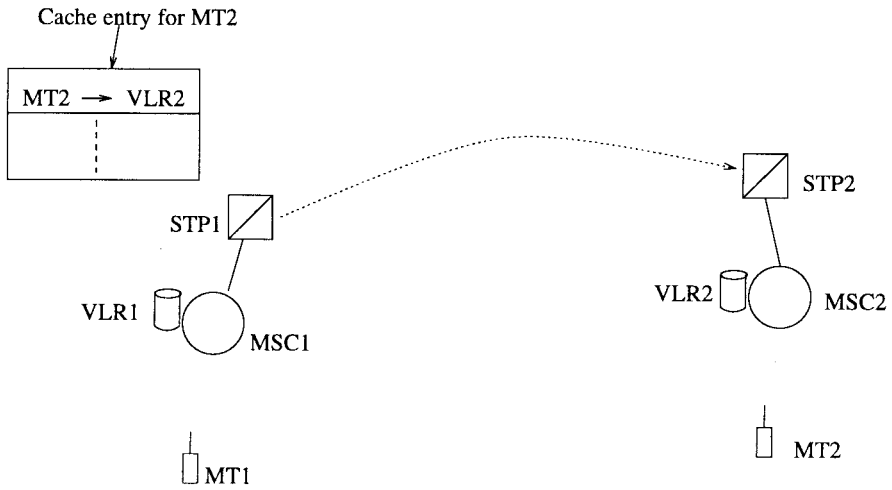


Fig. 9. Per-user location caching scheme.

1) a local pointer is stored at an MT's serving DR which indicates the current serving MSC of the MT;
2) a direct remote pointer is stored at a remote DR which indicates the current serving MSC of the MT;
3) an indirect remote pointer is stored at a remote DR which indicates the current serving DR of the MT.

In addition, the HLR of the MT may be configured to store a pointer to either the serving DR or the serving MSC of the MT. In some cases, it may be more cost effective not to set up any pointers, and the original IS-41 scheme will be used.

The functionality of the proposed scheme can better be described by the following scenarios. Suppose that the HLR of a given MT is located in New York and it is currently roaming in San Francisco. If a significant number of the incoming calls for the MT are originated from Los Angeles, a direct or indirect remote pointer can be set up for the MT in the DR at the Los Angeles area. When the next call is initiated for this MT from Los Angeles, the calling MSC first queries the DR and the call can be immediately forwarded to San Francisco without requiring a query at the HLR, which is located in New York. This reduces the signaling overhead for call delivery. On the other hand, the HLR can be set up to record the ID of the serving DR (instead of the serving MSC) of the MT. When the MT moves to another MSC within the same LA in San Francisco area, only the local pointer at the serving DR of the MT has to be updated. Again, it is not necessary to access the HLR in New York. This reduces the signaling overhead for location registration. The advantage of this scheme is that it can reduce the overhead for both location registration and call delivery.

b) Per-user location caching: The basic idea of the per-user location caching strategy [51] 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 as described in Section IV-A2 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.

Fig. 9 demonstrates the operation of per-user location caching. When a call is initiated from MT1 to MT2 as indicated in Fig. 9, the system can locate MT2 by using the cached information at STP1. As a result, MT2 is successfully located without querying the HLR of MT2. As compared to the IS-41 scheme as demonstrated in Fig. 8, per-user location caching allows the STP to locate 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 that is required to produce performance gain using per-user location caching is determined.

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 that is necessary to attain the minimum hit ratio is obtained. In order to reduce the number of misses, it is suggested in [68] that cache entries should be invalidated after a certain time interval. Based on the mobility and call arrival parameters, the author [68] introduces a T-threshold scheme which determines the time when a particular cached location information should be cleared such that the cost for call delivery can be reduced.

c) User profile replication: Based on the user profile replication scheme [99], 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

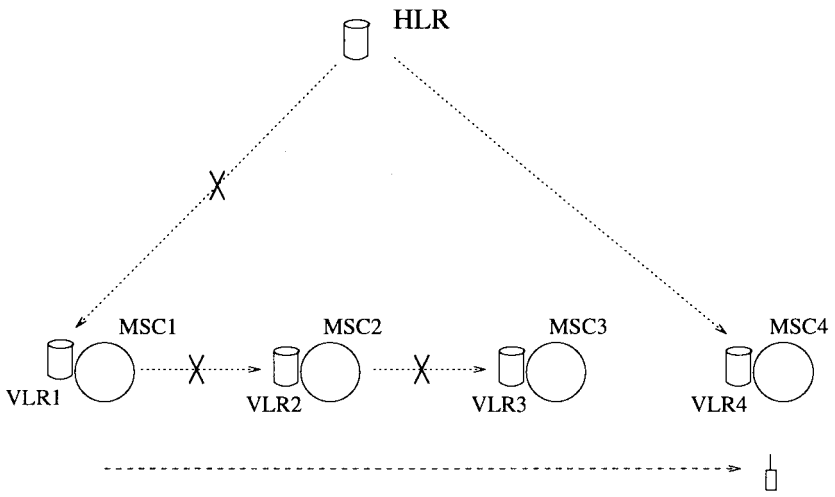


Fig. 10. Pointer forwarding strategy.

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. The authors then introduce a scheme for determining the replication scheme 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 PLMN networks because of the large number of network providers involved. Besides, generating and distributing the replication decision for a large user population is a computationally intensive and time-consuming process which may incur significant amount of network bandwidth. Future research should focus on the development of distributed user profile replication mechanisms.

d) Pointer forwarding: The basic idea of the pointer forwarding strategy [50] 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 following 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. Fig. 10 demonstrates the operation of pointer forwarding. Pointers are set up from VLR1 to VLR2 and from VLR2 to VLR3 as the MT moves from MSC1 to MSC2 and from MSC2 to MSC3, respectively. For $K = 2$, the pointer chain cannot be extended any further. An additional movement from MSC3 to MSC4 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 as compared to the original IS-41 scheme. The authors determine the conditions under which the pointer forwarding scheme should be used based on the system parameters.

e) Local anchoring: Under the local anchoring scheme [45], 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. Fig. 11 demonstrates the local anchoring scheme. Assuming that the local anchor of MT1 is VLR1, location change is reported to VLR1 (instead of the HLR) when MT1 moves from VLR2 to VLR3. 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 discussed previously, 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 decision on 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

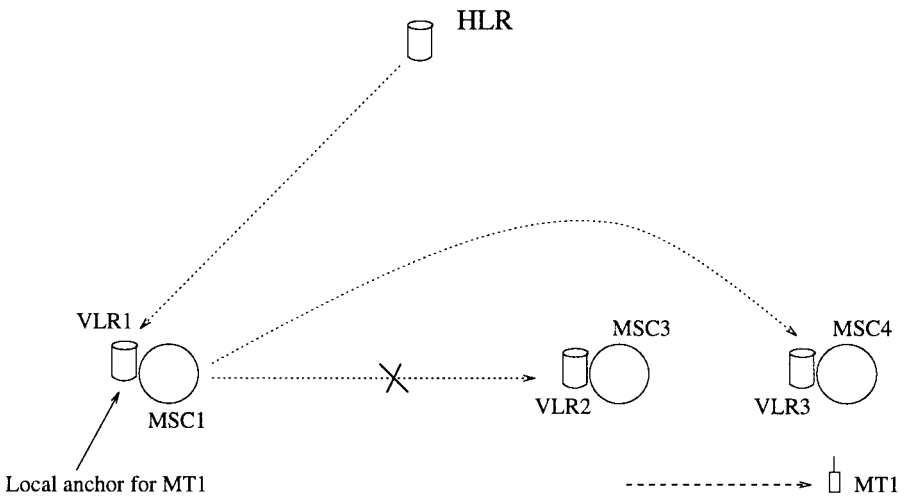


Fig. 11. Local anchoring scheme.

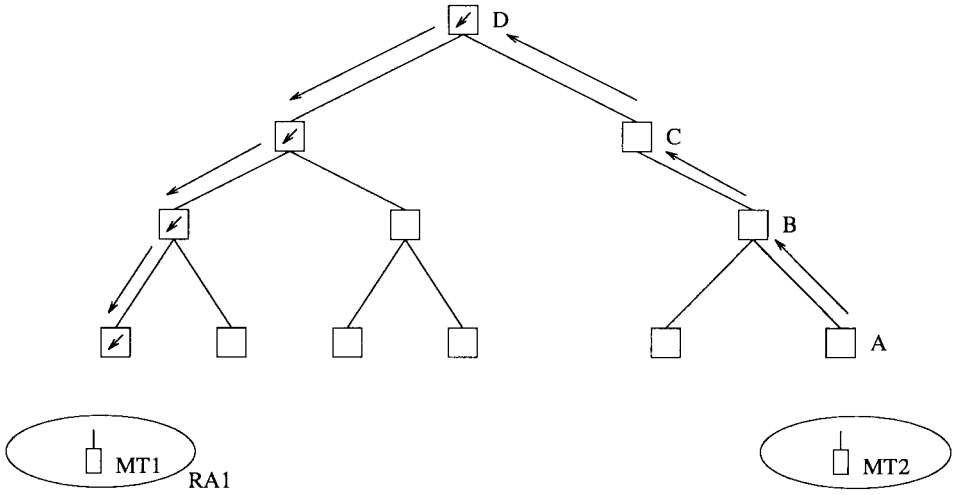


Fig. 12. Distributed hierarchical tree based database architecture.

the cost for dynamic local anchoring is always lower than or equal to that of the original IS-41 scheme.

2) Distributed Database Architectures: This type of solution consists of multiple databases distributed throughout the network coverage area.

a) A fully distributed registration scheme: For this scheme, the author proposes a distributed database architecture for location registration [118]. 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 MT's are associated with the leaf (lowest level) location databases and each location database contains location information of the MT's that are residing in its subtree. Fig. 12 demonstrates the operation of the proposed scheme. Given that an MT MT1 is located at LA1, an entry exists for MT1 in each database along the path from its current location to the root of the tree. The entries for MT1 at these databases are as shown in Fig. 12. When a call is initiated, the network locates the called MT by following its database entries. For example, if a call for MT1 is initiated by MT2 as shown in Fig. 12. The call request is received by node A. Since the database of node A does not have an entry for MT1, the call request is forwarded to node B and so on. When the request finally reaches node D, an entry for MT1 is found and the location of MT1 is determined after another three database lookups as demonstrated in Fig. 12. When an MT moves to an LA that belongs to a different leave 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.

b) Partitioning: Here, the authors introduce a partitioning scheme for the fully distributed database hierarchy [22]. Since the mobility pattern of the MT's varies among locations, partitions can be generated by grouping location servers among which the MT moves frequently. Based on the scheme introduced in [22], location registration is

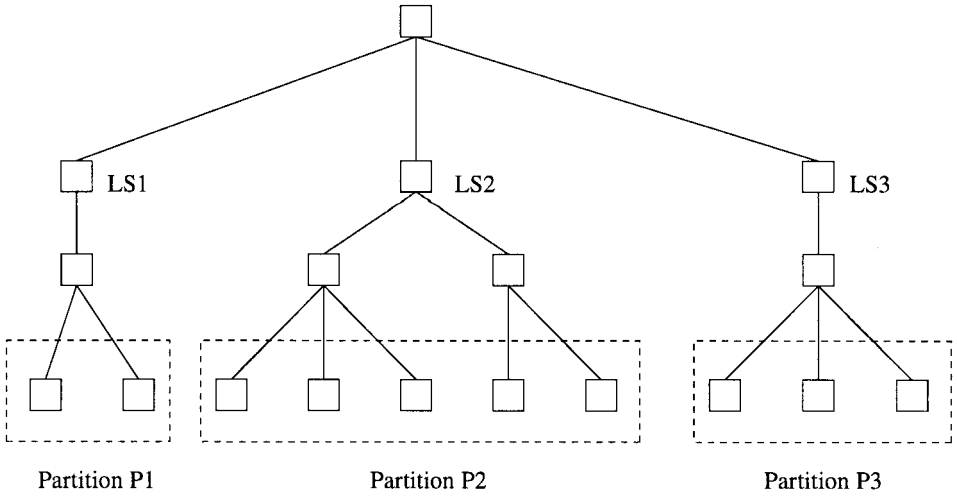


Fig. 13. Partitioning scheme.

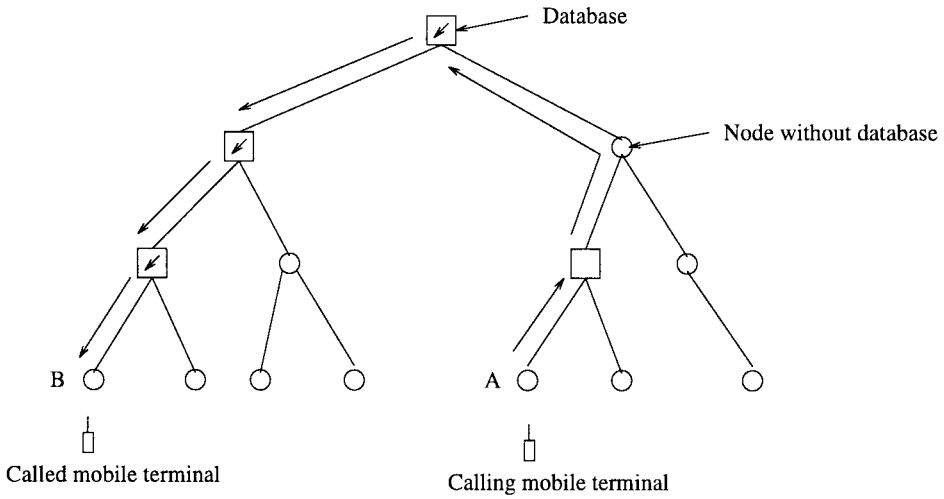


Fig. 14. Distributed database architecture.

performed only when the MT enters a partition. Fig. 13 shows the partitions for a particular PLMN network. Partition P2 consists of five location servers and they have a least common ancestor location server LS2. When an MT moves into partition P2, location server LS2 is updated indicating 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 registration in areas where the mobility rate of the MT's is high. Simulation results demonstrated 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.

c) Database hierarchy: Another distributed database architecture, similar to the fully distributed registration scheme [118] is introduced in [20]. Here, MT's 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 MT's. 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 more 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. Fig. 14 shows the operation of this scheme. 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. 14. 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 LA's, the authors introduce a method for determining the database placement which reduces the number of database accesses and updates.

C. Location Update and Terminal Paging Research

As discussed in Section IV-A1, current PCS networks partition their coverage areas into a number of LA's. Each

LA consists of a group of cells and each MT performs a location update when it enters an LA. When an incoming call arrives, the network locates the MT by simultaneously paging all cells within the LA. There are a number of inefficiencies associated with this location update and paging scheme.

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

In addition, the LA-based location update and paging scheme is a static scheme as it cannot be adjusted based on the parameters of an MT from time to time.

Recent research efforts attempt to reduce the effects of these inefficiencies. Excessive location updates are addressed by [96] and [23]. A timer-based strategy that uses a universal timeout parameter is presented in [96], while a tracking strategy for mobile users in PCS networks based on cell topology is explored and compared with the time-based strategy in [23]. For excessive polling, a one-way paging network architecture and the interfaces among paging network elements are examined in [63]. Additional schemes attempt to reduce the cost of finding a user when the MT moves during the paging process [93], [125]. Many recent efforts focus primarily on dynamic location update mechanisms which perform location update based on the mobility of the MT's and the frequency of incoming calls. We describe a number of dynamic location update and paging schemes in this section.

1) Location Update Schemes: The standard LA-based location update method does not allow adaptation to the mobility characteristics of the MT's. The following techniques allow dynamic selection of location update parameters, resulting in lower cost.

a) Dynamic LA management: This scheme introduces a method for calculating the optimal LA size given the respective costs for location update and cell polling [123]. The authors consider a mesh cell configuration with square-shaped cells. Each LA consists of $k \times k$ cells arranged in a square, and the value of k is selected on a per-user basis according to the mobility and call arrival patterns and the cost parameters. As an example, we assume that there are two MT's, MT1 and MT2, which have different mobility and call arrival patterns such that the values of k for MT1 and MT2 are two and four, respectively. Based in the scheme introduced in [123], Fig. 15(a) and (b) shows the LA's for MT1 and MT2, respectively. This mechanism performs better than the static scheme in which LA size is fixed. However, it is generally not easy to use different LA sizes for different MT's as the MT's must be able to identify the boundaries of LA's which are continuously changing. The implementation of this scheme is complicated when cells are hexagonal shaped, or in the worst case, when irregular cells are used.

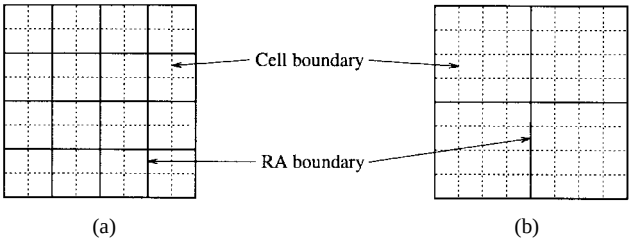


Fig. 15. Registration area for (a) $k = 2$ and (b) $k = 4$.

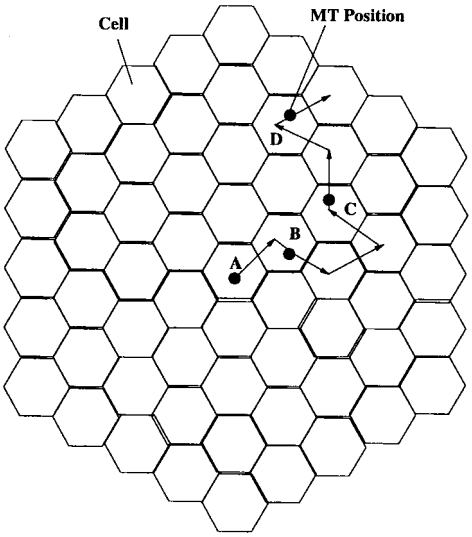


Fig. 16. Time-based location update scheme.

b) Three dynamic update schemes: Three location update schemes are examined in [24].

1) *Time Based:* An MT performs location updates periodically at a constant time interval ΔT. Fig. 16 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 $\Delta T, 2\Delta T$, and $3\Delta T$, respectively.
2) *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). Fig. 17 shows the same path as in Fig. 16. Assuming a movement threshold of three is used, the MT performs location updates at locations B and C as shown in Fig. 17.
3) *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 threshold). Fig. 18 shows the same path as in Fig. 16. A location update is performed at location B where the distance of the MT from location A exceeds the threshold

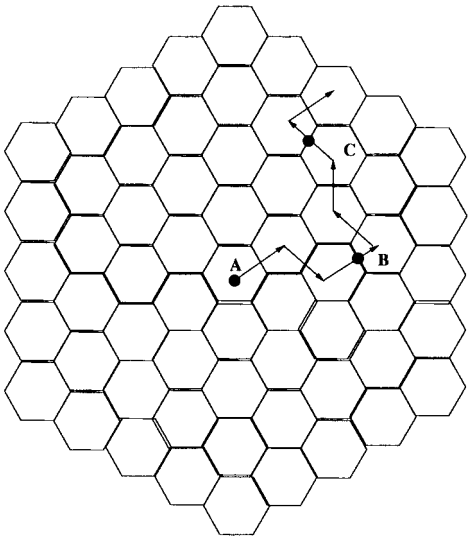


Fig. 17. Movement-based location update scheme.

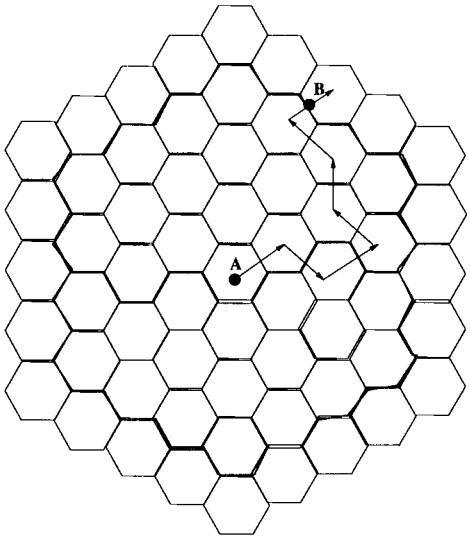


Fig. 18. Distance-based location update scheme.

distance (the distance from location A to the thick solid line as shown in Fig. 18 is equal to the threshold distance).

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

c) Distance-based update: A distance-based location update scheme is considered in [73]. The authors introduce an iterative algorithm that can generate the optimal threshold distance that 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-users based and should be 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.

d) Dynamic time-based update: A dynamic time-based location update scheme is introduced in [16]. 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 MT's, and the shortest-distance-first paging scheme as described in [73] is used. It is demonstrated that the results obtained are close to the optimal results given in [46]. Computation required by this scheme is low and they are, therefore, feasible for application in MT's that have limited computing power. Similar to the scheme described in [73], the drawback of this scheme 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.

2) Terminal Paging Schemes: In general, there is a tradeoff between paging cost and paging delay. The following techniques are two terminal paging methods that intended to minimize the paging cost under a given delay requirement.

a) Paging under delay constraints: Paging subject to delay constraints is considered in [97]. The authors assume that the network coverage area is divided into LA's, and the probability that an MT is residing in a LA is given. It is demonstrated that when delay is unconstrained, the polling cost is minimized by sequentially searching the LA's 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 tradeoff between the costs of location update and paging is not considered in [97].

b) Update and paging under delay constraints: Location update and paging subject to delay constraints is considered in [46]. Similar to [73], 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 the lowest cost. It is demonstrated that the cost is the lowest when the maximum delay is unconstrained. However, by slightly increasing the maximum delay from its minimum value of one, the cost is significantly lowered. Another scheme using the movement-based location update is reported in [15]. Similar to [46], paging delay is confined to a maximum value. Movement-based location update schemes have the advantage that implementation is simple. The MT's do not have to know the cell configuration of the network. The scheme introduced in [15] is feasible for use in current PLMN networks.

D. Open Problems

Each of the proposed schemes for PLMN mobility management can improve the IS-41 strategy to a certain extent. However, 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. In the following sections, we discuss these observations for: 1) location registration and call delivery and 2) location update and paging.

1) Location Registration and Call Delivery: As described in Section IV-B, recent research efforts in location registration and call delivery are based on either the centralized or the distributed database architectures. The centralized approach records the location information of all MT's 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 number of MT's increases, the signaling traffic may significantly degrade the performance of the PLMN 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 far away database is executed only when necessary. However, the number of database accesses required for location registration and call delivery is significantly increased from that of 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 the fully distributed approach. In fact, in order to attain better cost effectiveness, most of the on-going research efforts either try to: 1) increase the distribution of location information under a centralized database architecture such as the results reported in [45] and [50] or 2) limit the distribution of location information in a distributed database architecture such as the results reported in [20] and [22]. 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 online collection and processing of data. This may consume significant computing power, and careful design is necessary so that the computation can be effectively 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 the fully distributed approaches. In addition, methods for determining the mobility level and the call arrival statistics for an MT in real-time must be developed. Dynamic schemes for limiting or enhancing the distribution of location information on a per-user basis should be considered.

2) Location Update and Terminal Paging: As discussed in Section IV-C, 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 individual user. For example, under the LA-based location update scheme, the LA size most suitable for one user may be ineffective for another user. Most of the recent research efforts focus on the development of dynamic location update and paging schemes. Dynamic schemes allow online 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 MT's. 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 discussed in Section IV-C 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.

Thus, remaining open problems for the PLMN-based backbone are the following.

1) *Location Information:* Research work should consider the development of dynamic schemes that limit or enhance the distribution of location information on a per-user basis.
2) *Hybrid Database Architecture:* Ongoing research efforts attempt to reach some middle ground between centralized database architectures and distributed database architectures.
3) *Update and Paging:* Future research should focus on the design of dynamic location update and paging schemes that are simple to implement.

Because of the differences in network organization, certain problems discussed for the PLMN, such as the use of databases, do not apply to the Internet. In Section V, we encounter some of the organizational differences between backbone networks as we investigate current research and standardization efforts supporting terminal mobility under Mobile IP. Note that the term mobile node (MN) is used instead of mobile terminal in order to follow Mobile IP conventions.

V. Mobility Management for Mobile IP

Standards for terminal mobility over the Internet have been developed by the IETF and outlined in Request for Comments (RFC's) 2002–2006 [86]. Within the wireline IP, fixed terminals communicate differently depending upon their subnetwork location. Terminals on the same subnetwork can send packets directly, while terminals belonging to different subnetworks must send their packets through IP nodes, or routers, which perform switching functions [42]. The mobility-enabling protocol for the Internet, Mobile IP, promises to enable terminals to move from one subnetwork to another as packets are being sent, without interrupting this process [52]. Variations in Mobile IP include versions 4 (IPv4) and 6 (IPv6). Compared with IPv4, IPv6 can provide more addresses and mobility support. Thus, the procedures in this section are based largely on IPv6, except where noted.

An MN is a host or router that changes its attachment point from one subnet to another without changing its IP address. The MN accesses the Internet via a home agent (HA) or a foreign agent (FA). The HA is an Internet router on the MN's home network, while the FA is a router on the visited network. The node at the other end of the connection is called the correspondent node (CN). A simple Mobile IP architecture is illustrated in Fig. 19. In this example, the CN sends packets to the MN via the MN's HA and the FA. [Note that the term mobile node (MN) is used instead of mobile terminal (MT) in order to follow Mobile IP conventions.]

As mentioned previously, network organization introduces some differences in the way mobility management is handled over the Internet. For example, Mobile IP allows MN's to communicate their current reachability information to their home agent without the use of databases [84]. As a result, Mobile IP defines new operations for location and handoff management.

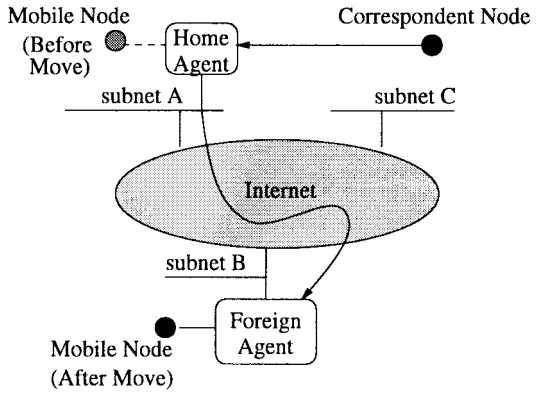


Fig. 19. Mobile IP architecture.

1) *Discovery*—How an MN finds a new Internet attachment point when it moves from one place to another.
2) *Registration*—How an MN registers with its HA, an Internet router on the MN's home network.
3) *Routing and Tunneling*—How an MN receives datagrams when it is away from home [86].

Registration operations include mobile agent discovery, movement detection, forming care-of-addresses, and binding updates, while handoff operations include routing and tunneling. Fig. 20 illustrates the analogous relationships between the location management operations for Mobile IP and those previously described in Fig. 1 and Section IV.

A. Location Registration

When visiting any network away from home, each MN must have an HA. The MN registers with its home agent in order to track the MN's current IP address. There are two IP addresses associated with each MN, one is for locating and the other one is for identification. The new IP address associated with an MN while it visits a foreign link is called its care of address (CoA). The association between the current CoA and the MN's home address is maintained by a mobility binding, so that packets destined for the MN may be routed using the current CoA regardless of the MN's current point of attachment to the Internet. Each binding has an associated lifetime period, negotiated during the MN's registration, and after which time the registration is deleted. The MN must reregister within this period in order to continue service with this CoA [52].

Depending upon its method of attachment, the MN sends location registration messages directly to its HA, or through an FA which forward the registration to the HA [30]. In either case, the MN exchanges registration request and registration reply messages based on IPv4, as described below and shown in Fig. 21.

1) The MN registers with its HA using a registration request message (the request may be relayed to the HA by the current FA).
2) The HA creates or modifies a mobility binding for that MN with a new lifetime.

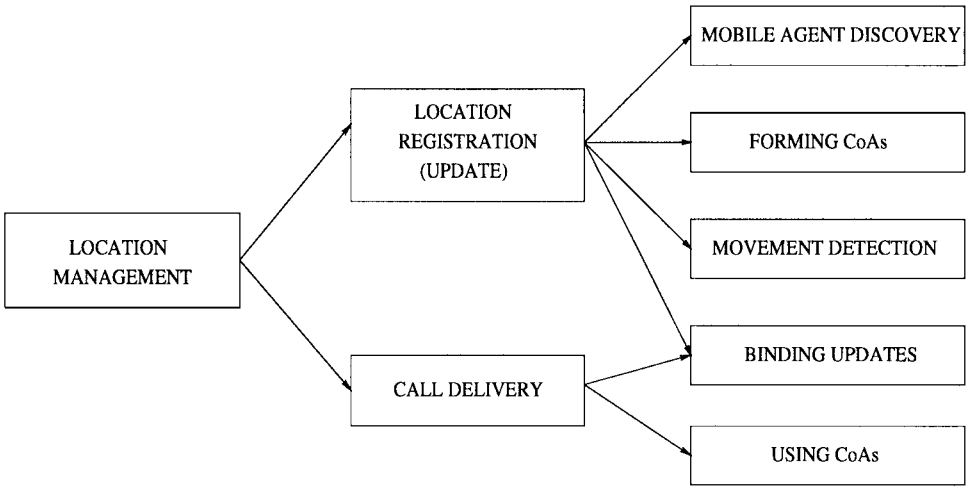


Fig. 20. Mobile IP location management.

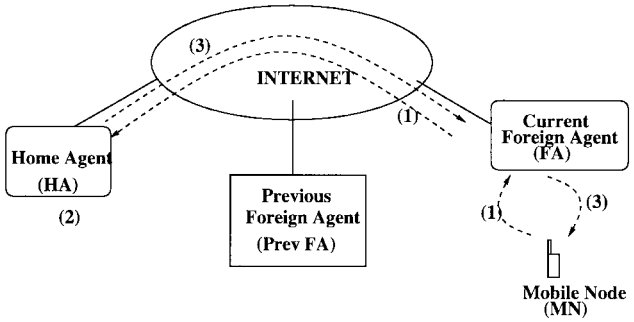


Fig. 21. Mobile IP location registration.

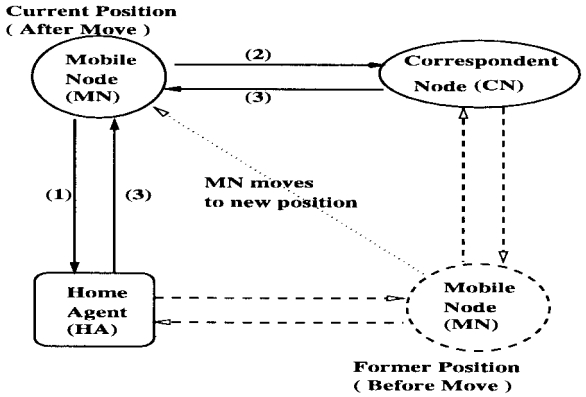


Fig. 22. Mobile IP location management operations.

3) The appropriate mobile agent (HA or FA) returns a registration reply message. The reply message contains the necessary codes to inform the mobile node about the status of its request and to provide the lifetime granted by the HA [86].

In IPv6, the FA's in Fig. 21 no longer exist. The entities formerly serving as FA's are now thought of merely as access points (AP's).

1) Movement Detection: For the other backbone networks discussed in this paper, the movement of the user is determined by updates performed when the user moves into a new LA. Since Mobile IP does not use LA's to periodically update the network, we discuss a new feature to determine whether the MN has moved to a new subnet after changing its network AP's. Mobile agents make themselves known by sending agent advertisement messages. The primary movement detection method for Mobile IPv6 uses the facilities of IPv6 Neighbor Discovery. Two mechanisms used by the MN to detect movement from one subnet to another are the advertisement lifetime and the network prefix.

a) Advertisement lifetime: The first method of detection is based upon the lifetime field within the main body of the Internet control message protocol (ICMP) router advertisement portion of the agent advertisement. A mobile node records the lifetime received in any agent advertisements, until that lifetime expires. If the MN has not maintained contact with its FA, the MN must attempt to solicit a new agent [84].

b) Network Prefix: The second method uses the network prefix—a bit string that consists of some number of initial bits of an IP address—to detect movement. In some cases, an MN can determine whether or not a newly received Agent Advertisement was received on the same subnet as the MN's current CoA. If the prefixes differ, the MN can assume that it has moved. This method is not available if the MN is currently using an FA's CoA.

After discovering that MN is on a foreign network, it can obtain a new CoA for this new network from the prefix advertised by the new router and perform location update procedures. For the PLMN, registration was implemented using database storage and retrieval. In Mobile IP, the MN's registration message creates or modifies a mobility binding at the home agent, associating the MN's home address with its new CoA for the specified binding lifetime. The procedure is outlined below and shown in Fig. 22.

1) The MN registers a new CoA with its HA by sending a binding update.
2) The MN notifies its CN of the MN's current binding information.

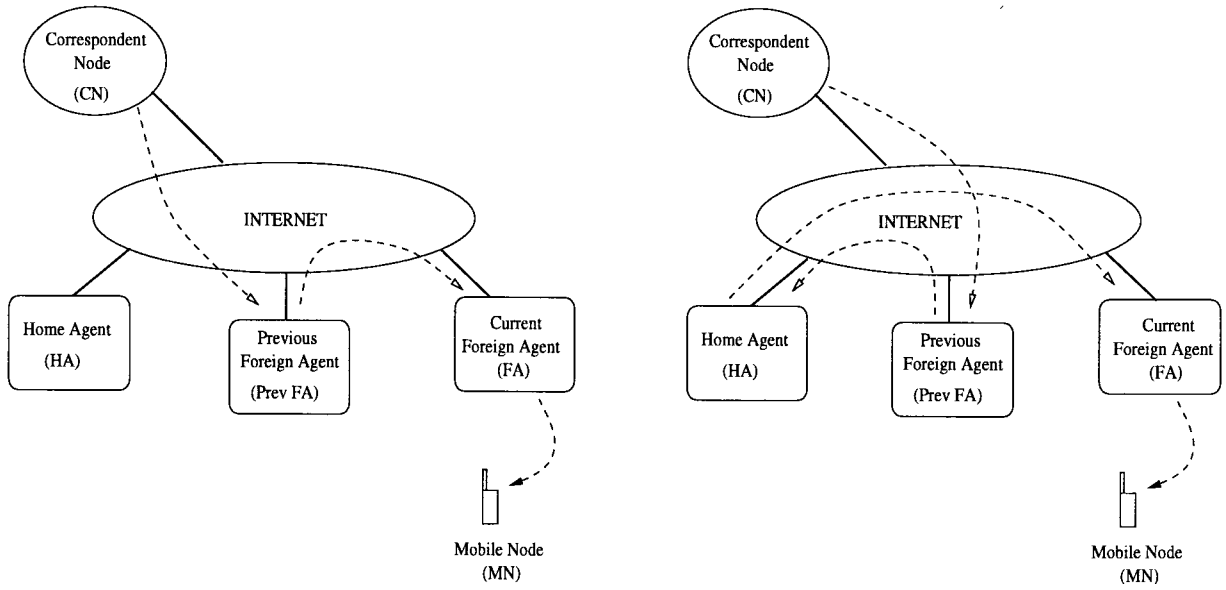


Fig. 23. Mobile IP smooth handoff: (a) fresh binding at previous FA and (b) no fresh binding at previous FA.

3) If the binding update is allowed to expire, the CN and the HA send a binding request to the MN to get the MN's current binding information.

The MN responds to the binding request with its new binding update. After receiving the new CoA, the CN and HA send a binding acknowledgment to the MN.

Once the registration process is complete, call delivery consists of reaching the MN via the new CoA's. Furthermore, a wireless network interface may actually allow an MN to be reachable on more than one link at a time (i.e., within wireless transmitter range of routers on more than one separate link). This establishment of coexisting wireless networks can be very helpful for smooth handoff.

B. Handoff Management

1) Smooth Handoff: A smooth handoff for MN's that are changing their point of attachment to the Internet is crucial for maintaining quality of service (QoS) guarantees. Current routing optimization schemes in IPv4 allow the previous foreign agent (or agents) to maintain a binding for their former mobile visitors, showing a current CoA for each. Then, as packets are sent to the old CoA, the corresponding previous foreign agents can forward the packets to the current CoA of the MN, as demonstrated in Fig. 23(a). As a result, an MN is able to accept packets at its old CoA while it updates its home agent and correspondent nodes with a new CoA on a new link.

If the previous FA does not have a fresh binding—the binding lifetime has expired—the previous FA forwards the packets to the home agent of the MN, which sends the packets to the CoA from the MN's last location registration update, as shown in Fig. 23(b). This can potentially create unnecessary traffic if the HA's binding still refers to the previous FA. Alternatively, the previous FA can invoke the use of special tunnels which forward the packets, but also indicate the need for special handling at the HA.

When special tunnels are used, the datagrams that are sent to the HA are encapsulated with the FA's CoA address as the source IP address. Upon reception of the newly encapsulated datagrams, the HA compares the source IP address with the MN's most recent CoA. Thus, if the two addresses match, the packets will not be circled back to the FA. However, if the addresses do not match, the HA can decapsulate the packets and forward them to the MN's current CoA, as shown in Fig. 23(b) [86]. [Note: In IPv6, the smooth handoff procedure is based on routers (IPv6 nodes) instead of FA's].

2) Routing and Tunneling: This process of routing datagrams for an MN through its HA often results in the utilization of paths that are significantly longer than optimal. Route optimization techniques for Mobile IP employ the use of tunnels, such as the special tunnels mentioned for smooth handoff, to minimize the inefficient path use. For example, when the HA tunnels a datagram to the CoA, the MN's home address is effectively shielded from intervening routers between its home network and its current location. Once the datagram reaches the agent, the original datagram is recovered and is delivered to the MN.

Currently, there are two protocols for routing optimization and tunnel establishment: route optimization in Mobile IP [85] and the tunnel establishment protocol [31].

a) Route optimization in Mobile IP: The basic idea of route optimization is to define extensions to basic Mobile IP protocols that allow for better routing, so that datagrams can travel from a correspondent node to a mobile node without going to the home agent first [85]. These extensions provide a means for nodes to cache the binding of an MN and then tunnel datagrams directly to the CoA indicated in that binding, bypassing the MN's home agent. In addition,

extensions allow for direct forwarding to the MN's new CoA for cases such as datagrams that are in flight when an MN moves and datagrams that are sent based on an out-of-data cached binding.

b) Tunnel establishment protocol: In this protocol, Mobile IP is modified in order to perform between arbitrary nodes [31]. Upon establishing a tunnel, the encapsulating agent (HA) transmits PDU's to the tunnel endpoint (FA) according to a set of parameters. The process of creating or updating tunnel parameters is called tunnel establishment. Generally the establishment parameters will include a network address for the MN. In order to use tunnel establishment to transmit PDU's, the home agent must determine the appropriate tunnel endpoint (FA) for the MN. This is done by consulting a table that is indexed by the MN's IP address. Each table entry contains the address of the appropriate tunnel endpoint, as well as any other necessary tunnel parameters. After receiving the packets, the foreign agent may then make use of any of a variety of methods to transmit the decapsulated PDU's so that it can be received by the MN. If the MN resides at this particular FA, no further network operations are required.

C. Open Problems

1) Simultaneous Binding: Since an MN can maintain several CoA's at one time, the HA must be prepared to tunnel packets to several endpoints. Thus, the HA is instructed to send duplicate encapsulated datagrams to each CoA. After the MN receives the packets from the CoA's, it can invoke some process to remove the duplicates. If necessary, the duplicate packets may be preserved in order to aid signal reconstruction. Due to the slow incorporation of wireless local area network (WLAN) technology into the marketplace, simultaneous binding has not yet been made available [86].

2) Regionalized Registration: Currently, three major concepts have been identified as potential methods for limiting location update and registration cost. First, there is a need for schemes that manage the local connectivity available to the MN and also to manage the buffering of datagrams to be delivered. Through this, the network can glean the benefits of smooth handoffs without implementing route optimization procedures [30]. Second, a multicast group of foreign agents is needed in order to allow the MN to use a multicast IP address as its CoA. Third, a hierarchy of foreign agents can be used in agent advertisement in order to localize the registrations to the lowest common FA of the CoA at the two points of attachments. To enable this method, the MN has to determine the tree-height required for its new registration message, and then arrange for the transmission of this message to reach each level of the hierarchy between itself and the lowest common ancestor of its new and previous CoA [87].

3) Security: As mentioned for the PLMN in Section IV, the authentication of the mobile becomes more complex as the MN's address loses its tie to a permanent access point. This allows for a greater opportunity for impersonating an MN in order to receive services. Thus security measures for the registration and update procedures—specifically protecting the CoA's and HA's must be implemented in order to police terminal use [86]. Some authentication schemes for the MN, the HA, and the FA can be found in [105].

Recently, there are some discussions regarding Mobile IP with respect to the third-generation IMT 2000 system. A high-level IP mobility architecture is described in [26] in which the diverse nature of today's wireless and wireline packet data networks is explored. To support the seamless roaming among the heterogeneous networks, the mobility management based on Mobile IP concepts is extended to the current third generation IMT2000 wireless architecture. Another concept, referred to as Simple Mobile IP (SMIP) [54], aims to find a more simplistic approach to support the mobility of users, compared with the asymmetric triangular approach proposed in IPv6. SMIP employs a more symmetric and distributed solution for location management based on MN connections to fixed network routers that have added mobility functions.

Although Mobile IP did not employ databases and LA's as discussed for the PLMN in Section IV, these issues will be visited again for WATM networks. In addition, the use of network trees and/or hierarchies, as seen in Section IV-B2, will also be considered. Thus, although a change in backbone network can introduce new considerations, many of the concepts used for PCS networks—and some for Mobile IP—will apply in some form to the WATM and satellite networks. Thus, one goal of future networks should be to exploit the common procedures and concepts in order to achieve interoperability. In Section VI, we discuss some of these commonalities and also address additional concerns for mobility management for WATM.

VI. Mobility Management for WATM

Mobility management for WATM deals with transitioning from ATM cell transport based upon widely available resources over wireline to cell transport based upon the limited and relatively unreliable resources over the wireless channel. Thus, it requires the investigation of important issues such as latency, message delivery, connection routing, and QoS [126]. The ATM Forum, through the WATM Working Group, has focused its efforts on developing basic mechanisms and protocol extensions for location and handoff management that address these issues. The Forum has specified that new procedures must be compatible with the current ATM standards in order to be implemented with relative ease and efficiency [88]. As a result, many of the procedures are also compatible with PCS, satellite, and to a lesser degree Mobile IP concepts. In this section, we outline selected proposed solutions for location management, terminal paging, and handoff.

A. Location Management Research

As shown in Fig. 24, proposed protocols for WATM implement location management using three techniques: location servers; location advertisement; and terminal paging.

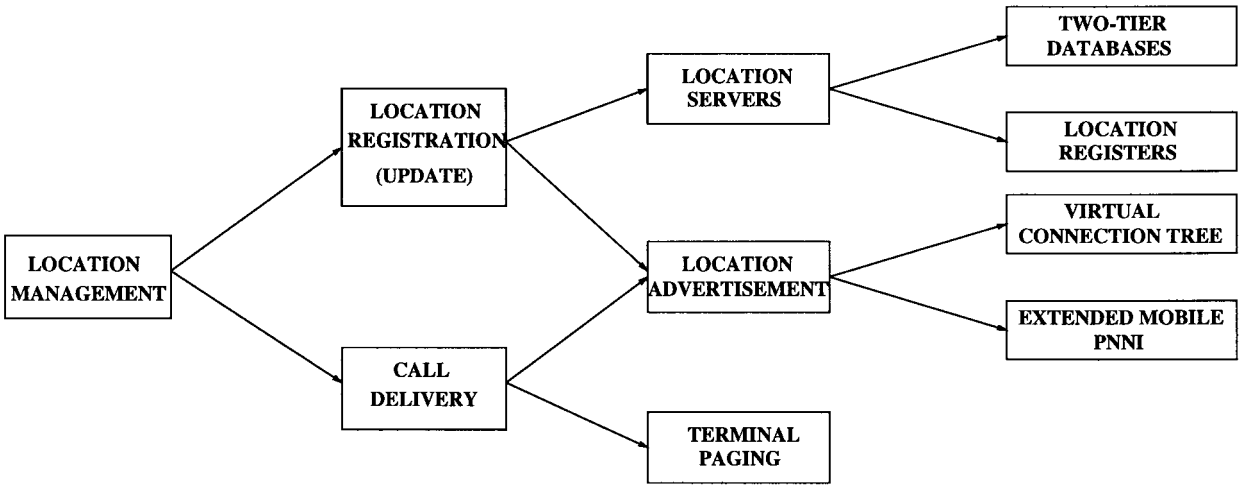


Fig. 24. ATM location management techniques.

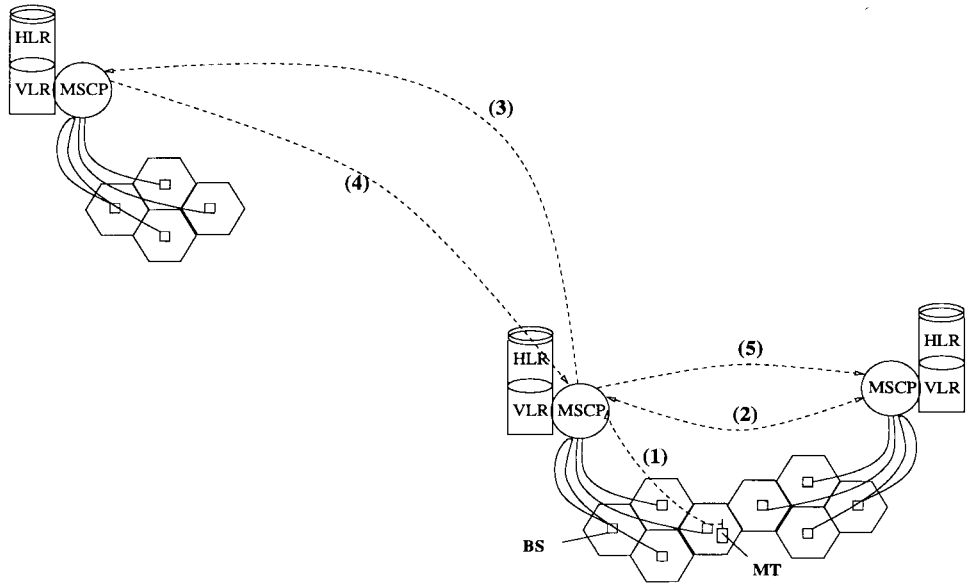


Fig. 25. Two-tier database scheme.

Location servers refer to the use of databases to maintain records of the attachment point of MT's within the network. These protocols are based on concepts and issues similar to those faced in Sections IV-B1 and IV-B2. As discussed in these two sections, the storage and retrieval process can generate excessive signaling and querying operations. Location advertisement avoids the use of databases by passing location information throughout the network on broadcast messages. Terminal paging is employed to locate MT's within the service area of its attachment point, as discussed previously in Section IV-C.

1) Location Servers Techniques: As mentioned above, location servers are the databases used to store and retrieve a record of the current position of the mobile. They require querying operations, as well as signaling protocols for storage and retrieval [43]. WATM server protocols employ the IS-41/GSM MAP-based techniques that were explored for the PLMN backbone in Section IV. The first method makes familiar use of the HLR/VLR database structure of Sections IV-A, IV-A1, and IV-A2. The second algorithm, location registers (LR's), uses a hierarchy of databases similar to those described in Section IV-B1.

a) Two-tier database: The architecture for the two-tier database scheme described in [18] uses bilevel databases that are distributed to zones throughout the network (see Fig. 25). The zones—analogous to the LA's of Section IV—are each maintained by a zone manager. The zone manager—analogous to the mobility service control point (MSCP) of the future wireless architecture—controls the zone's location update procedures. The home tier (HLR) of the zone's database stores location information regarding MT's that are permanently registered within that zone, while the second tier (VLR) of the zone's database stores location information on MT's that are visiting the zone. Each MT has a home zone, i.e., a zone in which it is permanently registered.

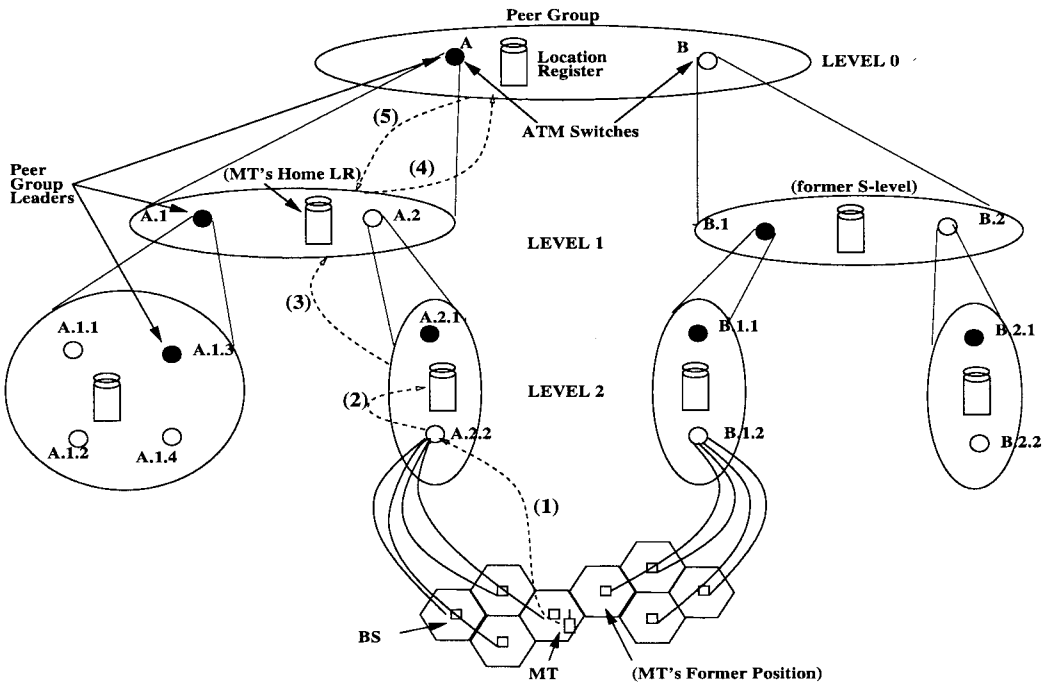


Fig. 26. WATM location registers scheme.

Just as location updates occurred for MT's entering a new LA, location registration for this method is performed according to the present zone of the MT. The process itself is the same as that presented for the PLMN in Section IV-A1, except that one VLR does not serve several LA's. One VLR must be dedicated to each zone and serve only the MT's attached there. Upon entering a new zone, the MT detects the new zone identity broadcast from the BS's. The steps for registration, demonstrated in Fig. 25, are as follows.

1) The MT transmits a registration request message to the new MSCP that contains its user identification number (UID), authentication data, and the identity of the previous zone.
2) The current MSCP determines the home zone of the MT from the previous zone.
3) The current and home MSCP's authenticate the user and update the home user profile with the new location information.
4) The home zone sends a copy of the profile to the current zone, which stores the profile in the VLR tier of its database.
5) The current MSCP sends a purge message to the previous zone so that the user's profile is deleted from the previous zone's VLR tier.

Call delivery is achieved by routing the call to the last known zone first. If the MT has moved and has been purged, the call is immediately forwarded to the home zone. The home zone's HLR is queried for the current location of the MT, which is forwarded back to the calling switch. The calling switch can then set up a connection to the current serving switch of the MT.

The advantage of the two-tier scheme is that it keeps the number of queries low—requiring at most two database lookups for each incoming call to find the MT. However, the use of a centralized HLR may caused increased signaling traffic and unnecessary connection set-up delays if the MT makes several localized moves for an extended period of time. A more localized approach may reduce the need for long-distance queries and thereby reduce connection set-up delays.

b) LR Hierarchy: The LR scheme in [113] distributes location servers throughout a hierarchical private network-to-network interface (PNNI) architecture, as shown in Fig. 26. The PNNI procedure is based on a hierarchy of peer groups, each consisting of collections of ATM switches. Each switch can connect to other switches within its peer group. Special switches, designated peer group leader, can also connect to a higher ranking leader in the "parent" peer group. Each peer group also has its own database, or LR, used to store location information on each of the MT's being serviced by the peer group.

The PNNI organization allows the network to route connections to the MT without requiring the parent nodes to have exact location information. Only the lowest referenced peer must record the exact location, and the number of LR updates then corresponds to the level of mobility of the MT. For example, a connection being set up to a MT located at switch $A.2.2$ in Fig. 26 is first routed according to the highest boundary peer group and switch A. Peer A can then route the connection to its "child" peer group, level $A.x$, to switch $A.2$. Finally, the connection is routed by $A.2$ to the lowest peer group level to switch $A.2.2$. Which resolves the connection to the MT.

Thus, for movement within the $A.2$ peer group, the location update procedure can be localized to only the LR

of that peer group. However, a movement from peer group $B.1$ to peer group $A.2$ requires location registration of a larger scope, and the maintenance of a home LR to store a pointer to the current parent peer position of the MT. To limit signaling for the larger scale moves to the minimum necessary level, the authors implement two scope-limiting parameters, S and L. The S parameter indicates a higher level peer group boundary for LR queries, while the L parameter designates the lowest group. In Fig. 26, the current S level is level one, while the L level is level two.

When the MT performs a location update by sending a registration notification message to the new BS, this message is relayed to the serving switch, which then stores the MT's location information in the peer group's LR. When the MT powers on or off, this message is relayed up the hierarchy until it reaches the preset boundary S. The S-level register records the entry and then relays the message to the MT's home LR. For movement from position $B.1.2$ to position $A.2.2$, the registration procedure, shown in Fig. 26, is as follows.

1) The MT sends a registration notification message to the new BS/switch
2) The new switch stores the MT in the peer group's LR.
3) The peer group then relays the new location information to the higher level LR's for routing, stopping at the first common ancestor of the former and current peer groups.
4) In this example, the former S level is not a common ancestor, so a new S level is designated and the location information stops propagating at the new S level, level 0.
5) The MT's home LR (located at group $A.x$) is notified of the new S-level location for the MT.

After the updates are complete, the new switch sends a purge message to the previous switch so that the former location can be removed from the LR's.

Call delivery is less complicated for this method, since the procedure takes advantage of the hierarchical organization. An incoming call request can be routed to the last known peer group or switch via the S-level LR. If the mobile has moved, the last known switch propagates a location request, querying the upstream LR's until the mobile endpoint's address is recognized by an LR that has a pointer to the mobile's current position. Then the request is sent to the L-level LR for that peer group, which resolves the query and sends the location information back to the calling switch.

Finally, if the call request reaches the S level before being recognized by an LR, the S-level LR forwards the location request directly to the home switch. Since the home LR keeps track of S-level changes for its mobile, the home switch can forward the request directly to the correct S-level switch, whose LR points to the current peer group position of the MT.

2) Location Advertisement Techniques: Although the location server methods provide the advantages of simplicity, decreased computation costs, and flexibility, the method can still require a substantial signaling and database querying load. This load can be reduced by using location advertisement. For Mobile IP, advertisement was described as a router notifying the MN of its new attachment point. For WATM, advertisement refers to the notification of appropriate network nodes of the current location of the MT. The first method, Mobile PNNI, uses the PNNI architecture described above by removing the LR's and by taking advantage of an internal broadcast mechanism [113]. The second method, Destination-Rooted Virtual Connection Trees, advertises location information via provisioned virtual paths [114]. The third and final method, Integrated Location Resolution, extends the signaling framework of ATM with location information elements that incorporate location resolution into the connection set-up process [8].

a) Mobile PNNI: This scheme, described in [113], uses the status notification procedures of the PNNI network to achieve automatic registration, tracking and locating. The PNNI protocol calls for the exchange of PNNI Topology State Packets (PTSP's) between ATM switches in the same peer group, between the peer group leader and higher level peer groups, and between the peer group leader and its lower level groups. The PTSP's are generated by the peer group leaders and contain information about the topology of the group and the load on each peer switch. In addition, the PTSP's contain "reachability" information, i.e., address and parent peer group information, for each network endpoint. By exploiting PNNI, some level of location information for each mobile endpoint can be propagated throughout the entire network [34].

The registration procedure occurs without the use of a database, since the location information is contained in these PTSP packets. Instead, a mobile is assigned a home switch which tabulates the route to the mobile's current switch location. The routing table is updated whenever the mobile powers on/off or moves to a new switch. In addition, the home switch is responsible for advertising the mobile's new location to the network, by updating the reachability information in the PTSP's.

Registration and update procedures are demonstrated in Fig. 27 and are implemented in the following two steps.

1) The mobile sends a registration message to the home switch.
2) The home switch and the current peer group leader propagate the new location information in the PTSP's.

The home switch must send a message to the former switch to begin forwarding packets. Once the reachability information has had time to propagate, the former switch is notified to stop forwarding packets.

Since PNNI reachability only identifies the parent peer group of the endpoints, the extended mobile PNNI algorithm includes a scheme to flood the exact location of the mobile through the network, but only to a certain peer level. This update region is referred to by the authors as a neighborhood in [113].

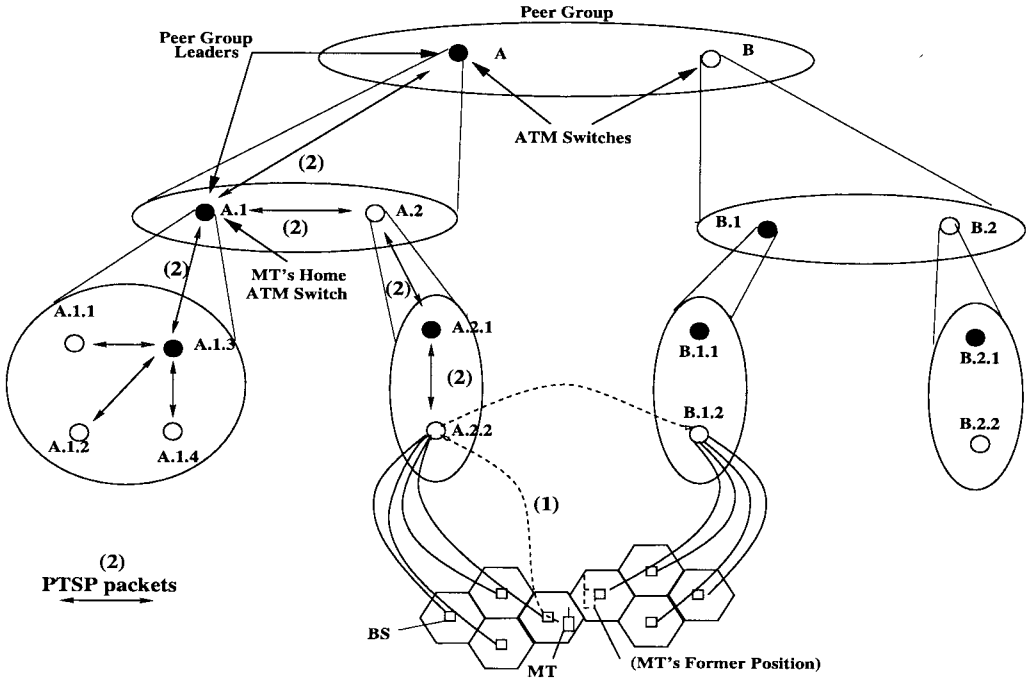


Fig. 27. Extended mobile PNNI.

If an MT moves to a new switch in the same neighborhood as its home switch, the exact location will be flooded to the home switch, without additional signaling load. Switches outside of the neighborhood will only receive the default, parent peer group information for that endpoint and can then use aggregate information to route calls according to the method outlined in the LR Scheme in Section VI-A1. Until the update propagates through the entire neighborhood, the mobile must send its new location information to the previous switch so that its connections can be forwarded.

If a mobile endpoint moves to a new switch that is not in its home neighborhood, the mobile must register its new location with the home switch for storage and update. Again, the previous switch must be notified, and calls must initially be forwarded.

The call delivery phase of Extended Mobile PNNI includes no prior connection set up, since each switch can route the call based on the reachability information it has received. A call from inside the same neighborhood can be immediately routed to the correct switch, provided the latest update has occurred. The same is true for call to a mobile that is in its home neighborhood. For all other possibilities, the home switch must route the call to the current switch. In all cases, if the latest reachability update has not had time to propagate, the last known switch will forward the call to the current switch.

b) Destination-rooted virtual connection tree: The network architecture for the destination-rooted virtual connection tree scheme provided in [114] and shown in Fig. 28 is a collection of portable base stations (PBS's) connected via provisioned virtual paths forming a connection tree. The PBS's are equipped with switching and limited buffering capabilities. The trees are based on the mobility indications of the MT. Each PBS maintains a running list of resident MT's in its coverage area.

The registration process, also shown in Fig. 28, can begin under the power on/off condition, or when the MT moves into a new service area. When the MT powers on or off, the MT sends a message to its local (current) PBS. The PBS simply adds the MT to or deletes the MT from the service list. However, when the MT moves into the PBS's region, the new PBS must send a deregistration message to the old PBS on behalf of the MT and then enter the MT's identity information into its current list.

Call delivery consists of advertising the mobile terminal's identity via a broadcast message from the PBS of the calling terminal. No paging on the air interface is required. The local PBS responds to the broadcast and initiates connection procedures. If there is no response, the connection is rejected based on the assumption that the MT is not registered. For mobile-to-fixed communication, the PBS performs as a switch with routing tables for the fixed endpoint.

c) Integrated location resolution: The integrated scheme presented in [8] modifies the signaling operations of the ATM call set-up process to include indications of the called terminal's current location. In this scheme, the MT is assigned a home switch which governs all information on the mobile's current position within the network. When the MT moves away from the home switch, the MT must first register its presence with the new, foreign switch. Then the foreign switch must notify the MT's home switch of the MT's new foreign address. Thus, the location update operation consists of one update propagated to the MT's home switch.

For the call delivery phase of location management, all initial call requests are routed immediately to the MT's

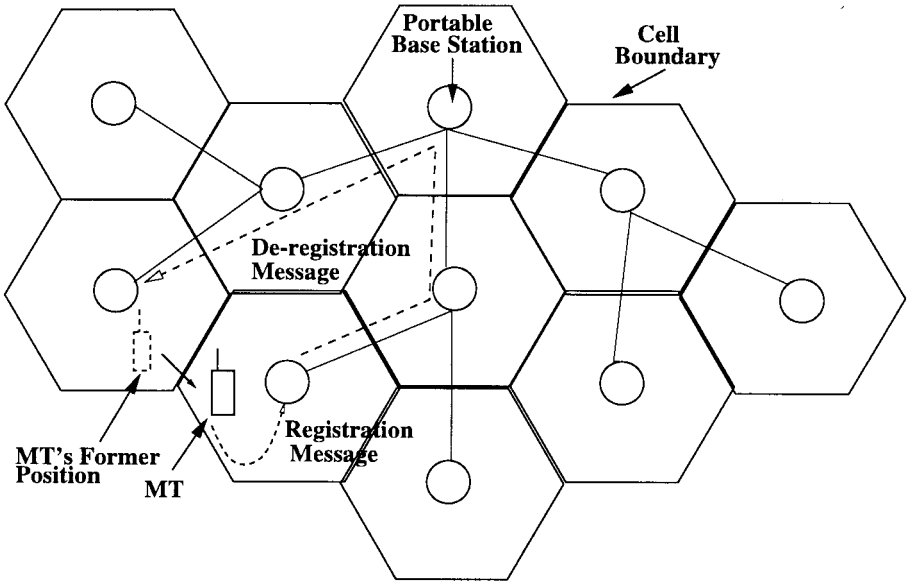


Fig. 28. Destination-rooted virtual connection tree.

home switch with a connection SETUP message. Since this scheme assumes that the home switch of each MT is an ME-ATM switch, the SETUP message can include a flag indicating that an upstream entity is an ME-ATM switch. However, since the calling endpoint does not know if the called terminal is mobile or not, any one of the following conditions can exist:

1) the called terminal is a static, or fixed terminal that is permanently attached to its home switch;
2) the called terminal is mobile but is currently attached to its home switch;
3) the called terminal is mobile and it is currently away from is home switch.

At the called terminal's NNI/UNI boundary, the home switch must determine the current condition of the terminal. If the called terminal is fixed, the home switch sends a CONNECT message to the originating switch. If the called terminal is a mobile attached to its home switch, the home switch sends a CONNECT message to the originating switch that also identifies the connection as mobile to any interested intermediate switches. This allows any ME-ATM switches in the path to prepare for future possible handoffs.

The call delivery scheme for a mobile terminal that is currently away from its home switch is shown in Fig. 29 and is implemented as follows.

1) The calling endpoint issues a connection SETUP message to the MT's home switch.
2) The home switch determines that the terminal is away from home.
3) The MT's home switch sends a RELEASE message to the originating switch. This message must indicate the MT's current foreign address and also identify the connection as mobile.
4) A switch in the original SETUP path establishes a new path for the connection to the MT and sends a new SETUP message to the MT's foreign address. This new message must include the MT's home address.
5) Mobility-enhanced switches in this new path can prepare for future possible handoffs.

If in any of these cases the calling endpoint is an MT, the home switch for the calling MT can identify the connection as mobile and include the calling MT's home address during the connection SETUP phase.

B. Terminal Paging Research

Because this area has not yet been explored for WATM applications, only one algorithm is considered here.

1) Velocity Paging Scheme: The velocity paging scheme outlined in [117] attempts to categorize the travel of the MT into a velocity class. The class is then used to generate a paging zone—a list of cells to be paged. The velocity classes described for the scheme characterize a range of velocities that has been previously demonstrated by the MT. This quantity can be obtained in two ways. The first way employs a distance-based registration, which is also seen in Section IV-C1, wherein the MT registers whenever its distance from the last registered cell has passed a threshold value. The velocity class can then be formed by using the distance threshold value divided by the time between consecutive registration actions. Then an average speed for the MT has been determined. The other way to find the velocity class of a mobile is to take advantage of the movement-based registration procedure, also mentioned in Section IV-C1. The movement-based procedure counts the number of times the MT has traversed through a cell. Then, once that number has reached a threshold, the MT must register. The number of movements, divided by the time passed between registrations and multiplied by a velocity time unit would yield the quantity that we desire. The procedure for paging a mobile terminal is outlined next.

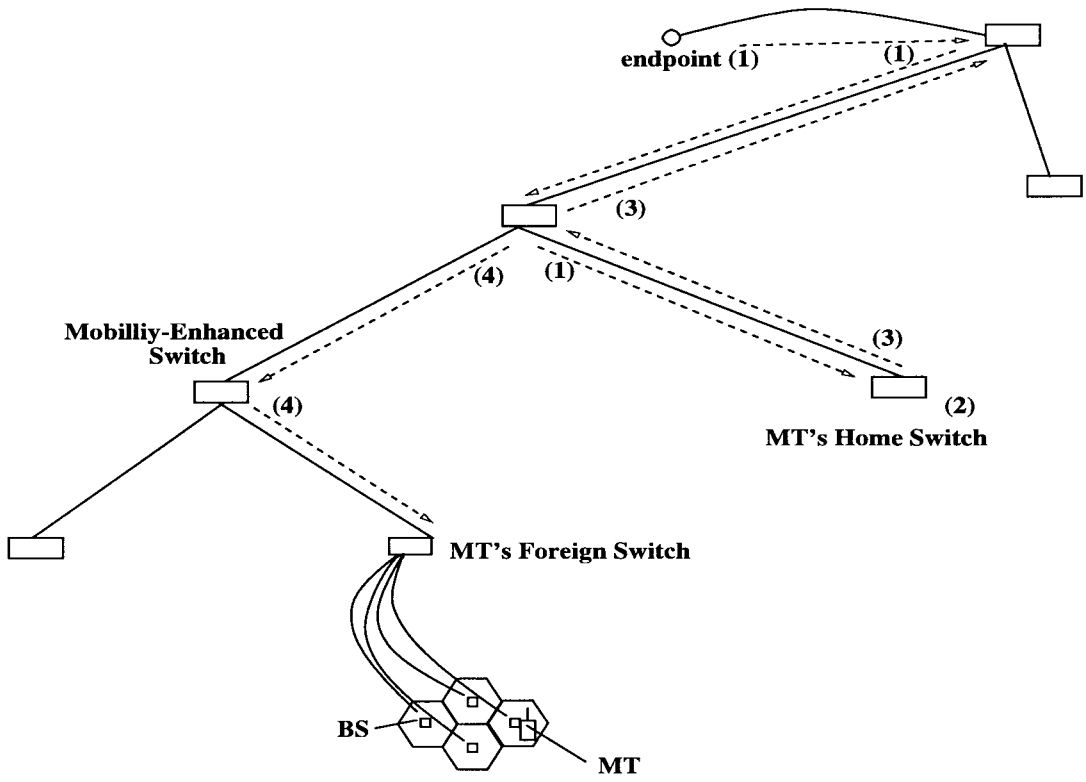


Fig. 29. Integrated location resolution.

When the system wants to deliver a call to a standby MT, the system must query a location server. The server supplies the MT's movement profile, based on the two choices above. It supplies the velocity class index for the MT, the MT's last known location, and the last registration time. The system then uses this information to calculate the maximum distance that the MT could have traveled within the given constraints. Finally, the candidate cells that are within that maximum distance are the first group to be paged.

Alternatives to this scheme, described in [117], include recording the direction of the MT, paging in that direction initially, and then branching out.

C. Handoff Management Research

Handoff management controls the process of maintaining each connection at a certain level of QoS as the MT moves into different service areas [62]. As illustrated in Fig. 30, current proposed protocols can be grouped into four categories: full connection rerouting; route augmentation; partial connection rerouting; and multicast connection rerouting. Full connection rerouting maintains the connection by establishing a completely new route for each handoff—as if it were a brand new call. Route augmentation simply extends the original connection via a hop to the MT's next location. Partial connection rerouting re-establishes certain segments of the original connection, while preserving the remainder. Finally, multicast connection rerouting combines the former three techniques but includes the maintenance of potential handoff connection routes to support the original connection, reducing the time spent in finding a new route for handoff [17].

1) Full Connection Rerouting: Full connection re-establishment is the most optimal and the simplest rerouting technique in that all of the VC's in the connection path from the source to the previous switch are cleared [89]. Then new VC's are established from the source to the new switch. It can be implemented by treating the connection as a newly admitted call, or by employing network elements that perform the mobility functions for that connection, independent of the switches.

a) Interworking devices (IWD's)—Full connection reroute: The connection reroute algorithm for IWD's outlined in [89] manages handoff through the use of external processors that isolate the mobility management from the fixed network. An overlay of these IWD's is placed throughout the fixed network—one IWD corresponding to each switch as shown in Fig. 31(a). The switches control multiple BS's and provide the connection to the fixed network. A connection then extends from the originating terminal to its local switch is routed through the network and terminates at the mobile endpoint via the mobile's switch, but all handoffs are organized via the IWD.

The procedures outlined in [89] for full connection reroute are outlined in Fig. 31(a) and (b) and are listed below.

1) After registration, the MT informs the target switch of the identity of the original connection, including a unique identifier for the connection and the NSAP of the old IWD.

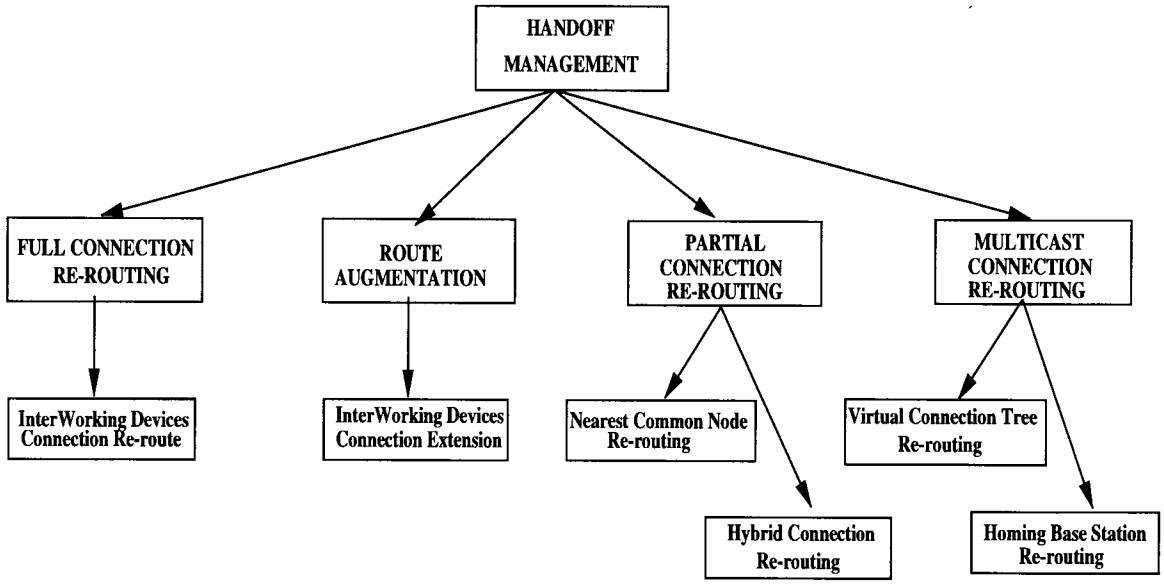


Fig. 30. WATM handoff management techniques.

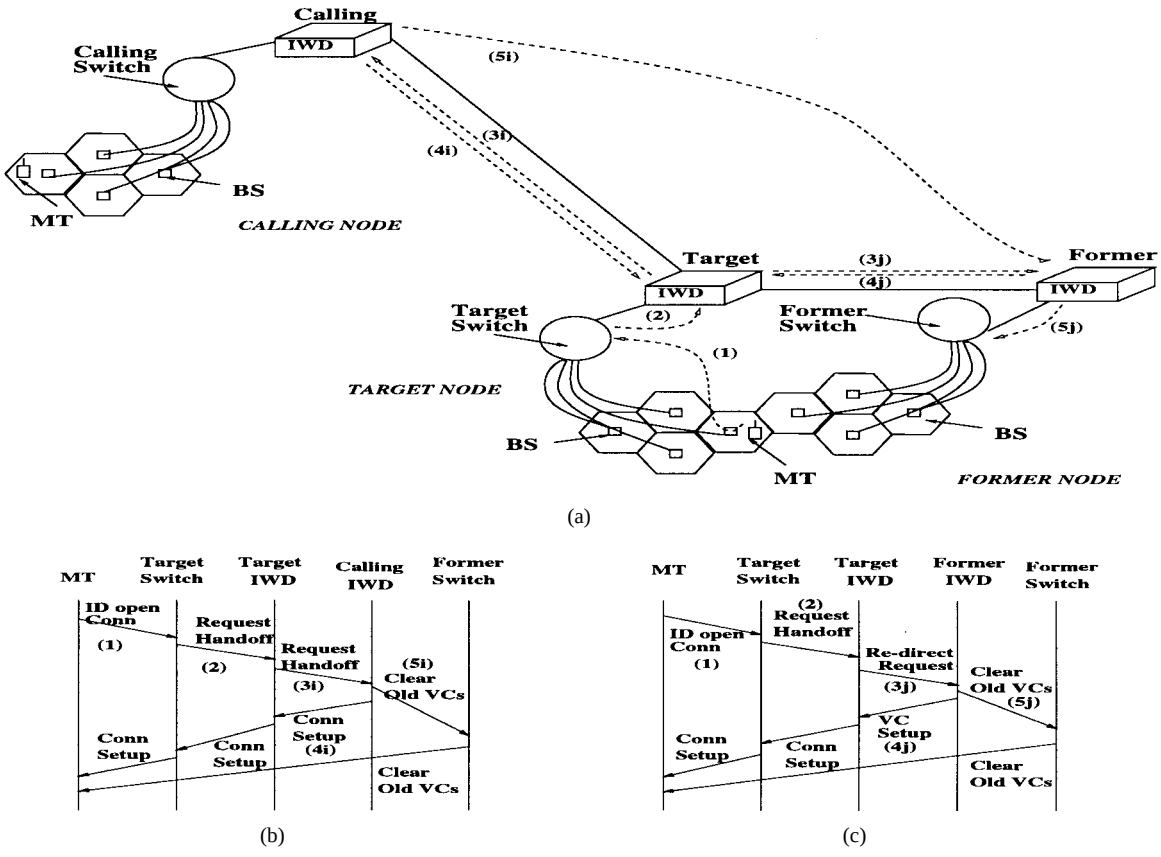


Fig. 31. IWD's handoff scheme: (a) architecture; (b) connection reroute signal flow; and (c) connection extension signal flow.

2) The target switch then sends the handoff request message to its IWD.
3) The target IWD then forwards this request to the calling IWD. This message is sent via the IWD overlay and the connection ID.
4) The calling IWD and target IWD set up a new connection to the MT.
5) To clear the original path, the calling IWD sends a clear message along the old path toward the old switch path toward the old switch.

2) Route Augmentation: Route augmentation does not achieve the optimal route of the full connection technique,

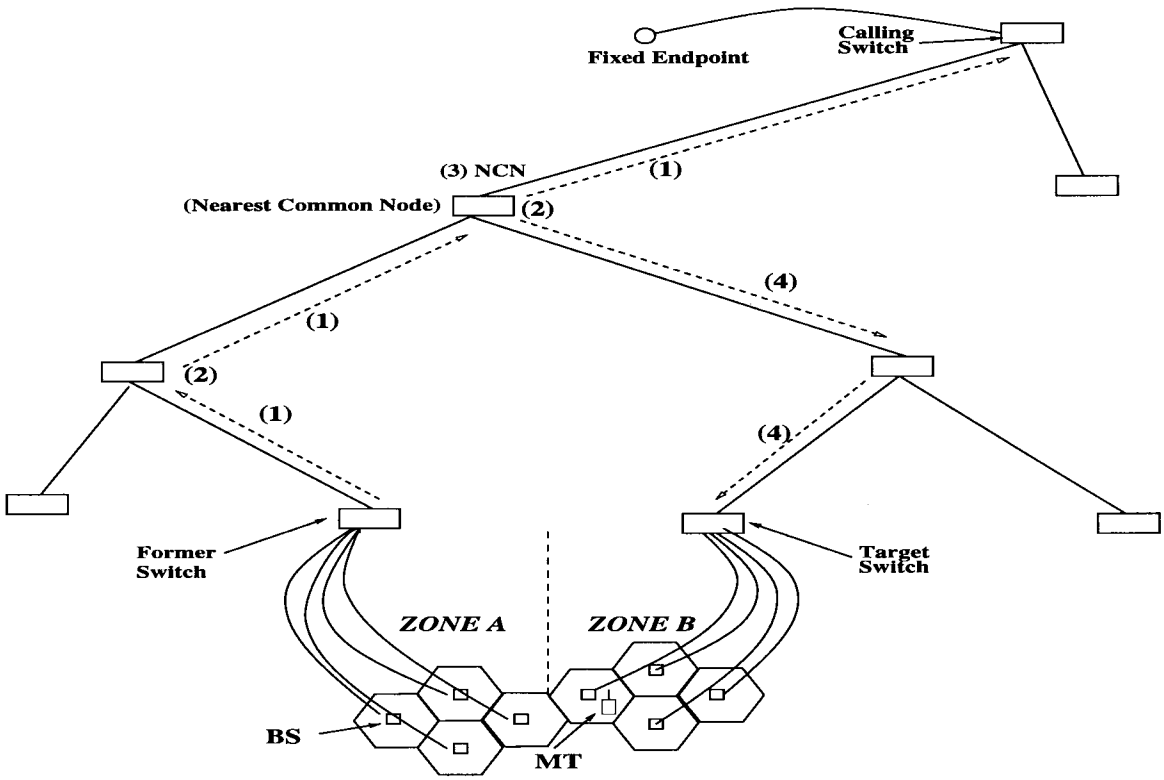


Fig. 32. NCNR procedure.

but it does provide a level of speed and simplicity that can potentially reduce handoff latency and signaling costs.

a) IWD's—Route augmentation: In this scheme outlined in [89], an MT's connection extends from the originating terminal to its local switch, is routed through the network, and terminates at the mobile endpoint via the mobile's switch. The handoff operation then uses the IWD's as anchors through which to route the connection.

We now describe the handoff procedure corresponding to the IWD's architecture and signal flow of Fig. 31(a) and (c).

1) Again the MT informs the target switch of the identity of the original connection, including a unique identifier for the connection and the NSAP of the former IWD. This identifies the former IWD as the anchor.
2) The target switch then forwards the handoff request message to its IWD.
3) The target IWD sends a redirect request to the former IWD.
4) The former IWD responds by bridging the VC connection to the target IWD. The target IWD can notify the switch to set up the new connection to the MT.
5) Finally, the former switch clears its former connections to the MT.

The full connection technique establishes an optimal route at the expense of signaling and resource management. The route augmentation technique simplifies the process, but it can become terribly inefficient if the connection route begins to loop about itself.

3) Partial Connection Rerouting: The partial connection rerouting technique attempts to route a connection more efficiently by preserving some portions of the original route for resource management and simplicity and rerouting other portions for optimality.

a) Nearest common node algorithm: The nearest common node rerouting (NCNR) algorithm presented in [17] routes connections according to the LA (referred to by the authors as zones). Handoff within the LA—or zone—constitutes a VC table update. For handoff between zones, the algorithm attempts to bridge the connection at the nearest WATM network node that is common to both of the zones involved in the handoff transaction. In the tree topology, common refers to two nodes branching from the same point. In a hierarchy, the common point of two zones is a higher node which uses separate paths to access each zone.

Handoff from zone A to zone B begins with the MSCP of zone A checking for a direct physical link between the two LA's. If either zone is a "parent" of the other, the parent zone acts as an anchor for the handoff procedure. Both A and B participate in connecting to the MT until the radio link transfer completes. Then, if A is the parent, A becomes an ATM switch in the connection path. If B is the parent, B deletes A from the connection path altogether.

If A and B are not directly linked, the handoff, illustrated in Fig. 32, is conducted as follows.

1) The MSCP for the mobile terminal transmits a handoff start message to the calling endpoint. This message includes the ATM addresses of the MT, the target switch, and the former switch.

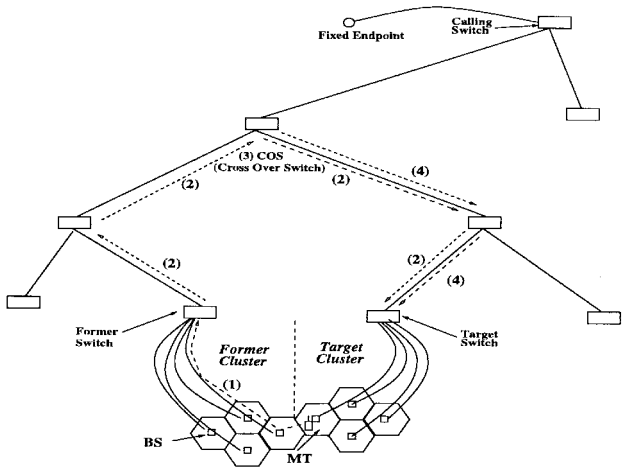


Fig. 33. Hybrid connection rerouting procedure.

2) Each switch in this path uses the ATM address to determine if it is the NCN. (The NCN will have three different egress ports for the three addresses.)
3) Since the above step will repeat itself until the NCN is determined, the switch that recognizes itself as the NCN stops this process by setting the NCN bit in the handoff start message.
4) The NCN initiates a connection to the target zone by forwarding a reroute message to the switches in that path.

The target zone receives the reroute message and replies with an acknowledgment to the former zone, thereby completing the reroute process. Until the radio-level handoff is complete, the NCN forward the connection data to both A and B. Afterward, the NCN clears its connection to A.

b) Hybrid connection algorithm: The hybrid connection algorithm [103] handoff protocol begins with the MT moving into an overlap region where it receives a beacon from a new BS. The target BS must determine whether the handoff is intracluster or intercluster. A cluster is simply a collection of BS's connected to a common cluster switch. If the handoff is intracluster, the cluster switch can perform as the crossover switch (COS). Otherwise, a COS discovery process must be initiated. This procedure is shown in Fig. 33 and is outlined below.

1) The MT sends a handoff hint message to the old BS.
2) The former BS sends a handoff invoke message (containing the MT's connection list) to the target BS.
3) The COS switch is determined.
4) Partial paths are set up from the COS to the target BS.

By the time the MT leaves the overlap region and fully enters the new region, the new path has been established. The MT can then send a greet message to the target BS and the BS can send a redirect to the COS in order become a part of the new connection path. Finally, the COS informs the old BS to disconnect the old partial paths. For handoff due to link failure, some cells will be lost until the MT detects the failure. Then buffering at the MT and at the COS is used to regain the appropriate handoff connection.

Partial connection rerouting provides better resource use while reducing signaling, but it requires algorithms for preserving cell transport through buffering and cell sequencing. It also requires the computation of the NCN/COS. The method described next incorporates the use of multiple handoff paths in order to minimize cell loss and buffering needs.

4) Multicast Connection Rerouting: Multicast reestablishment combines the ideas discussed above in a hybrid fashion but also introduces the idea of maintaining the potential handoff connections in addition to the original connection. Then, under handoff, there is little network time spent in selecting a new route since several are already available [28].

a) Virtual connection tree algorithm: The virtual connection tree algorithm in [6] is based on a hierarchical collection of ATM switching nodes attached to the fixed network, with links extending to BS's. The root of the tree is a fixed switching node connecting to the backbone network, while the leaves are the BS's. Each mobile connection is assigned a set of virtual connection numbers (VCN's) that are used to identify a set of paths from the root to one leaf. Only one path is operational at a time. Then the call can occur from the MT attached to the leaf through the root of the tree and on to the fixed network, or to the root of some other connection tree.

Thus, the authors outlined handoff procedures for two cases: handoff within the same tree and handoff between trees. Fig. 34 demonstrates the handoff within the same tree procedure.

1) The MT transmits cells with a new VCN, corresponding to the previously set-aside path for the target BS
2) The target BS then activates the VCN path by using it to transmit packets from the MT to the root of the tree.
3) When the cells arrive at the root with a new VCN, the routing table at the root switch is updated with the new BS location of the MT. Incoming cells can then routed to the MT via the new path.

For handoff between trees, the mobile's connection has reached the tree boundary, and it has become necessary to leave the tree altogether. The MT must seek admission to the new connection tree, as if it were a new call to be admitted to the network.

By using this method, connections of mobile terminals within a geographical mobile access region or tree may be handed over to any other BS within that area without involving the network processor. The larger the region, the greater the likelihood that the mobile will remain within the area for the duration of a call.

The next algorithm is an example of a concept that can be translated between Mobile IP and WATM. As seen in

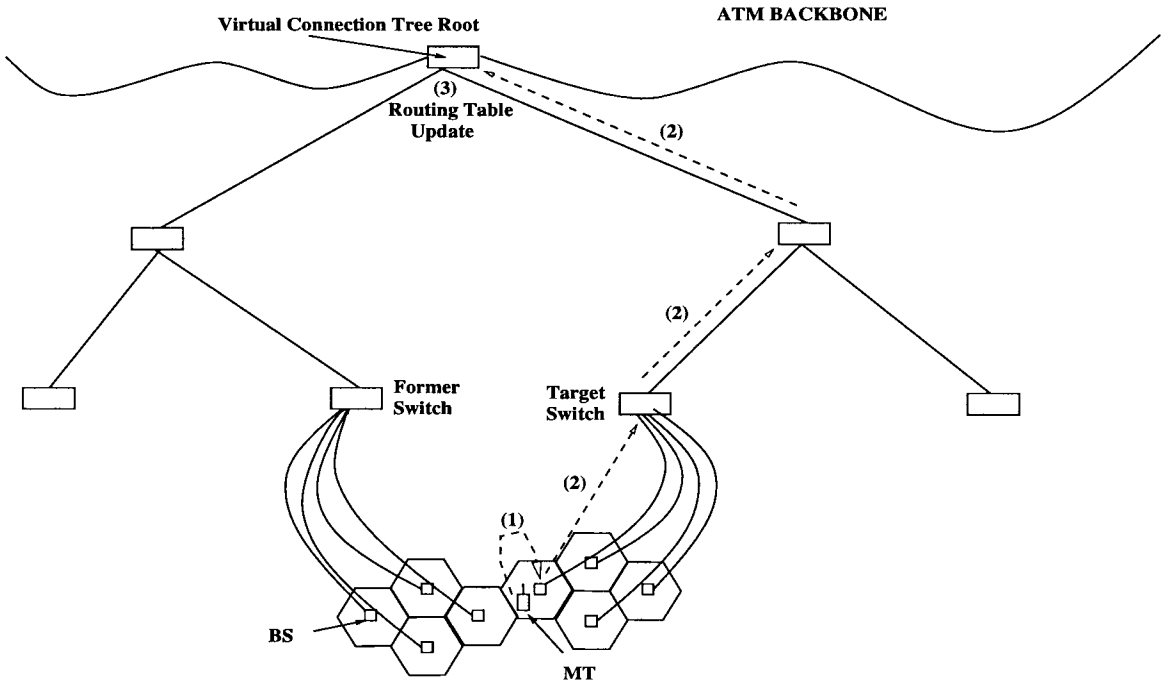


Fig. 34. Virtual connection tree algorithm.

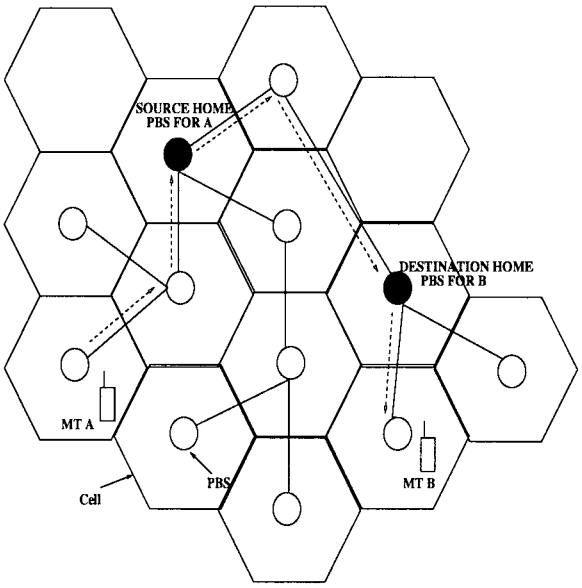


Fig. 35. Homing algorithm handoff routing procedure.

Section V-B1, sending packets to the MT via a home station can ensure that the data arrive at the correct location during handoff.

b) Homing algorithm: The homing algorithm presented in [114] again uses a collection of PBS's connected in a tree of provisioned virtual paths. A home PBS is designated for each mobile terminal accessing the tree. The home station serves as an anchor for each endpoint of the mobile connection and manages the cell transfer during handoff. For MT A transmitting to MT B, cells are first transmitted to the source home PBS for A. Then the cells are routed to the destination home PBS for B via pre-established VP's. The home PBS for B delivers the cells directly to B's mobile terminal.

Handoff for MT A is then executed as shown in Fig. 35 and as listed below.

1) Cells are rerouted from MT A to the home station for A.
2) Next, the cells are transmitted from the source home for A to the destination home for B.
3) Finally, the home station for B forward the cells to the current location for MT B.

The source-to-destination home connection remains relatively stable due to the anchoring scheme. Thus, if both A and B have moved to new PBS's, the cells are first sent to the home station, which is able to forward the cells to the new PBS location. Finally, to improve network utilization, the home PBS's are periodically updated, according to the movement patterns of the MT. For example, if MT A remained at its new PBS for a prescribed period of time, the network would undergo some process to designate the new PBS as the new home PBS for A.

D. Open Problems—ATM Forum Activity

Future research for WATM involves developing signaling protocols and methods for evaluating proposed schemes for performance with call dropping rates, call blocking rates, cell sequencing, and latency [55]. A scheme to integrate the location resolution operations of WATM with the signaling functions of connection set up is presented in [8]. This method is an adaptation of an IETF Mobile IP scheme that attempts to interwork Mobile IP under WATM. The WATM working group is also continuing to evaluate existing protocols that combine techniques from different

schemes and also allow some flexibility of method in order to develop a standard for WATM [88].

1) Location Management: The two protocols under current consideration to be standardized by the ATM Forum are the Extended Location Registers Scheme and the Extended Mobile PNNI Scheme [36]. Addressing is a concern for standardizing the location registers approach. A strategy is needed to assign permanent addresses to mobile terminals. One suggestion is to split the ATM Network Service Access Point (NSAP) address space into mobile NSAP's and fixed terminal NSAP's in order to allow a calling party's switch to generate a location request before initiating connection set up [115]. Another is to create a set of addresses that identifies the level of mobility. Small-scale mobility would represent a limited area defined by the peer group of the endpoint's current node. Large-scale would represent mobility between peer groups. Continuing issues include security in tracking and locating the mobile terminal, whether to have a continuous active VC from the current node to the home node, and the efficiency of continuous flush and re-advertisement [25].

2) Handoff Management: The ATM Forum has also been working to combine the main concepts of the proposed handoff schemes in order to find a feasible yet efficient method of rerouting for handoff [88]. Both path rerouting and path extension require two phases—fast extension and then route optimization. One extreme is the establishment of an entirely new connection—providing an optimal route, but at the expense of latency. The extreme at the other end is connection extension, which is simply a static partial connection reroute policy, where the previous BS is always the designated NCN or COS [116]. This method provides the fastest extension but is a wasteful use of network resources. A handoff signaling framework is needed that can accommodate a number of connection rerouting mechanisms [88]. One consideration is a multicast technique where the mobile would connect to two or more access points at a time in order to help maintain seamlessness for handoff and to possibly reduce buffering needs [100].

Other challenging issues that remain for the ATM Forum are the following [88].

1) *QoS:* Define service classes for mobile connections, devise mechanisms to adapt to QoS degradation, and develop new signaling features, such as QoS renegotiation.
2) *Rerouting Connections:* Develop algorithms for finding new route options, create signaling protocols for reconfiguring the connection path, and determine the feasibility of proposed solutions.
3) *Point to Multipoint:* Develop protocols that address rerouting a mobile terminal's point-to-multipoint connections.
4) *Mobile-to-Mobile Handoff:* Further address changes to existing protocols in order to support connection routing and QoS for a mobile-to-mobile connection.
5) *Optimization:* Develop efficient methods that allow an existing mobile connection to be periodically rerouted for an optimal connection path.

Terrestrial wireless networks such as PCS, Mobile IP, and WATM provide mobile communication services with limited geographic coverage. In recent years several LEO satellite systems have been proposed to provide global coverage to a more diverse user population. In the following section we describe the mobility management concerns for these satellite networks.

VII. Mobility Management for Satellite Networks

LEO satellite systems can support both the areas with terrestrial wireless networks and areas which lack any wireless infrastructure. In the former case, a satellite system could interact with terrestrial wireless network to absorb the instantaneous traffic overload of the terrestrial wireless network. In other words, mobile users would alternatively access a terrestrial or a satellite network through dual-mode handheld terminals. In the latter application area, the LEO satellites would cover regions where the terrestrial wireless systems are economically infeasible to build due to rough terrain or insufficient user population. Among the proposed LEO satellite systems, Iridium is already providing voice and low bit rate data services to its users. Next-generation LEO satellite networks such as Teledesic will provide broad-band access to their users.

LEO satellites are usually defined for those with altitudes between 500 and 1500 km above the Earth's surface [27], [72], [77]. This low altitude provides small end-to-end delays and low power requirements for both the satellites and the handheld ground terminals. In addition, intersatellite links (ISL) make it possible to route a connection through the satellite network without using any terrestrial resources. These advantages come along with a challenge; in contrast to geostationary (GEO) satellites, LEO satellites move in reference to a fixed point on the Earth. Due to this mobility, the coverage region of a LEO satellite is not stationary. A global coverage at any time is still possible if a certain number of orbits and satellites are used. Coverage area of a single satellite consists of small-sized cells, which are called as spotbeams. Different frequencies are used in different spotbeams to achieve frequency reuse in the satellite coverage area. In the following subsections, we present the state-of-the-art and open research areas for Satellite location and handoff management.

A. Location Management Research

As mentioned in Section IV, in ordinary wireline networks there is a fixed relationship between a terminal and its location. In contrast, wireless networks support terminals that are free to travel, and the network access points of the mobile terminals change as they move around the system. As we have discussed for PCS, Mobile IP, and WATM, terrestrial wireless networks require an MT to periodically report its location. To locate an MT within an LA, a paging mechanism is used such that polling signals are simultaneously sent to all cells within the LA. The called MT will reply to the polling signal and its exact location can be determined. Section IV-C and Section VI-

B presented a number of location update mechanisms that have been proposed to reduce the signalling cost and call delay related to location update and paging in terrestrial wireless networks [15], [56], [64], [117]. The LEO satellite network environment brings more challenging problems because of the movement of satellite footprints. For example, an LA cannot be associated with the coverage area of a satellite because of very fast movement of a LEO satellite. Thus, current research concerns the development of new LA definitions for satellite networks as well as the signalling issues mentioned for all of the location management protocols. In [78] and [90], LA's are defined using (gateway, spotbeam) pairs. However, the very fast movement of the spotbeams results in excessive signalling for location updates. In [19], LA's are defined using only gateways. However, the paging problem has not been addressed in [19].

B. Handoff Management Research

To ensure that ongoing calls are not disrupted as a result of satellite movement, calls should be transferred or handed off to new spotbeams or satellites. If a handoff is between two spotbeams served by the same satellite, handoff is intrasatellite. Small size of spotbeams causes frequent intrasatellite handoffs, which are also referred to as spotbeam handoffs. If the handoff is between two satellites, it is referred to as intersatellite handoff. Another form of handoff occurs as a result of the change in the connectivity pattern of the network. Satellites near to polar regions turn off their links to other satellites in the neighbor orbits. Ongoing calls passing through these links need to be rerouted. This type of handoff is referred to as link handoff. Frequent link handoffs result in a high volume of signaling traffic. Moreover, some of the ongoing calls would be blocked during connection rerouting caused by link handoffs. In the following, we present the state-of-the-art in satellite handoff management. More details can be found in [9].

1) Intersatellite Handoff Algorithms: Since two satellites are involved in an intersatellite handoff, the connection route should be modified to include the new satellite into the connection route. Thus, the same connection routing issues discussed for WATM in Section VI-C are again encountered for satellite networks. The route change can be achieved by augmenting the existing route with the new satellite or rerouting the connection completely. Route augmentation is simple to implement, however the resulting route is not optimal. Complete rerouting achieves optimal routes at the expense of signaling overhead.

In [110] and [112], a handoff rerouting algorithm, referred to as footprint handover rerouting protocol (FHRP), has been proposed to handle the intersatellite handoff problem in the LEO satellite networks. The FHRP is a hybrid algorithm that consists of the augmentation and the footprint rerouting phases. In the augmentation phase, a direct link from the new end satellite to the existing connection route is found. This way, the route can be updated with minimum signaling delay and at a low signaling cost. In case there is no such link with the required capacity, a new route is found using the optimum routing algorithm. In the footprint rerouting (FR) phase, connection route is migrated to a route that has the same optimality feature with the original route. The goal of the rerouting is to establish an optimum route without applying the optimum routing algorithm after a number of handoffs. This property is significant because, in the ideal case, the routing algorithm computes a single route for each connection.

2) Spotbeam Handoff Algorithms: Spotbeam handoff occurs frequently due to small size of the spotbeams. As an example, a user terminal is covered with a spotbeam with average duration of 38 s, while it stays in the footprint of a single satellite for an average duration of 10 min [91]. Frequent spotbeam handoffs would cause blocking of the handoff call if no ground-satellite channel is available in the new spotbeam. A number of handoff policies have been recently proposed. Since blocking a handoff call is less desirable than blocking of a new call request, spotbeam handoff algorithms gives higher priority to handoff calls. Some possible prioritization techniques can be based on queueing of handoff requests [92] and the use of guard channels for handoff calls [47]. In addition, a number of call admission control algorithms [76], [109] have been proposed to determine if a newly arriving call should be admitted into the network.

a) Handoff queueing: Handoff queueing algorithms [92] rely on overlapped coverage regions of the spotbeams involved in the handoff process. When a user terminal is in the overlapped coverage region, handoff process is initiated. If there is a channel available in the new spotbeam, this channel is allocated to the user terminal. Otherwise, the handoff request is queued. When a channel becomes available, one of the calls in the handoff queue is served. A handoff call is blocked if no channel is allocated for the call in the new spotbeam when the power level received from the current spotbeam falls below the minimum power level that is required for a successful data transfer. Handoff queueing reduces the handoff call blocking ratio, however its performance depends on the new call arrival rate and the size of the overlapped coverage region. In the worst case, high call arrival rates or small overlapped coverage regions would result in a large value of handoff call blocking rate. A modification to [92] is to use dynamic channel allocation in addition to the queueing of the handoff calls [91]. This algorithm performs well for low-to-moderate traffic levels. However, it requires channel reassignment after each call departure, which occurs very often because of the frequent handoffs, resulting in extreme overheads for the LEO satellite networks.

b) Guard channels: Guard channels are used to ensure that some number of channels is reserved for handoff calls even when the new call arrival rate is high. In a system with guard channels, new call requests are rejected if the number of busy channels is larger than a certain threshold. The difference between the system capacity, in number of channels, and the threshold value is equal to the number of guard channels. The handoff call blocking rate could

be reduced by increasing the number of guard channels. However, reservation of some channels for handoff calls increases the blocking rate for new arrivals. Hence, there is a tradeoff between the handoff call blocking and new call blocking. This tradeoff has not been investigated for broadband traffic. Moreover, multimedia traffic requires certain QoS, such as cell loss ratio, delay, and delay jitter to be guaranteed by the network. Guard channel [47] and handoff queueing [92] algorithms do not have any mechanisms to guarantee users' QoS expectations.

c) Call admission control: These algorithms compute certain performance parameters such as handoff call blocking probability to decide if a newly arriving call can be admitted into the network. In the connection admission control algorithm presented in [76], when a new call request arrives at a spotbeam, it is associated with a list of possible neighboring spotbeams that the user is going to visit with some probability in the future. A mobility reservation metric is updated for each spotbeam in the neighbor list. This algorithm reserves bandwidth in the neighbor spotbeams to decrease the handoff blocking rate. Although the handoff call blocking probability is decreased, the algorithm can not guarantee any upper bound. Moreover, because of the *ad hoc* nature of the reservation metric, the algorithm is conservative, i.e., it underutilizes the network. Finally, the algorithm is evaluated for a medium Earth orbit (MEO) satellite network. MEO satellites are located at altitudes of 5000–13 000 km. As a result of their high altitudes, MEO satellite networks have very low handoff probabilities. The applicability to a LEO network, which has a high handoff probability, is questionable. The geographical connection admission control (GCAC) algorithm [109] has been introduced to limit the handoff call blocking probability for the spotbeam handoffs in the LEO satellite networks. Upon a new call arrival, the GCAC algorithm estimates the future handoff blocking performance of the users to decide whether the newly arriving call can be admitted into the network without increasing the blocking probability for the existing calls while providing the same blocking guarantee to the new user. The new call request is accepted if the handoff blocking probability averaged over the contention area is less than the target blocking probability. The GCAC algorithm assumes that the exact user locations are known by the network. The orbit dynamics and the spotbeam geometry are utilized to estimate the performance metrics. The performance evaluation results show that the GCAC algorithm achieves a bounded handoff blocking probability without penalizing the new calls. Moreover, the GCAC algorithm adapts to the distribution of user terminals over the coverage area and can handle nonuniform traffic distribution.

3) Link Handoffs: The topology of LEO satellite networks changes with time due to intersatellite links that are temporarily switched off. Each LEO satellite has up and down wireless links for communication with ground terminals and intersatellite links for communication with neighbor satellites. There are two types of ISL's; intraplane

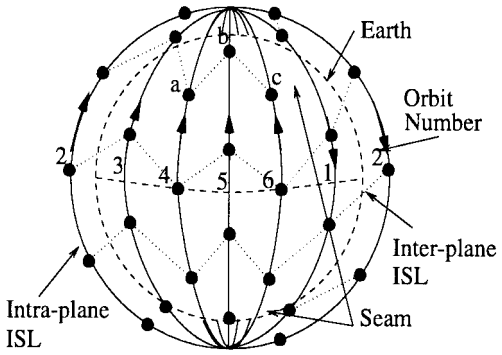


Fig. 36. LEO satellite network with seam.

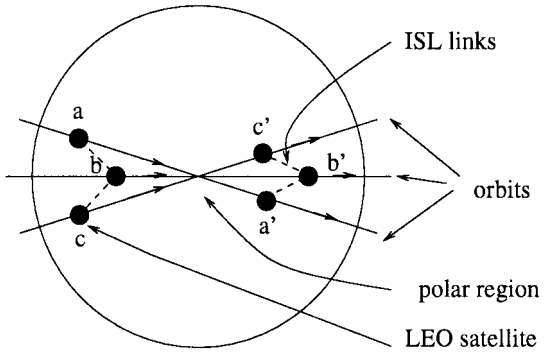


Fig. 37. LEO satellites in polar region (top view).

ISL's connecting satellites within the same orbit and inter-plane ISL's connecting satellites in adjacent orbits. Intra-plane ISL's can be maintained permanently. On the other hand, inter-plane ISL's would be temporarily switched off due to the change in distance and viewing angle between satellites in neighbor orbits. In the analysis reported in [120] for the Iridium system, it is concluded that only ISL's between latitudes of approximately 60° north or south would be maintained between counter-rotating orbits. This is labeled as seam in the example network model depicted in Fig. 36. When the satellites go into seam, they temporarily switch off their ISL's to the neighbor orbits, resulting in a dynamic network topology. The second type of topology change in LEO satellite network occurs due to the satellites temporarily switching off the ISL's as they cross the polar regions [119]. Fig. 37 depicts the satellites passing through a pole. The drawing reflects the top view, i.e., looking at the pole from a viewing position above the satellites. Satellites a, b, and c (also shown in Fig. 37) are moving toward the pole. Satellite b's left and right neighbors are satellites a and c, respectively. After passing the pole, the neighbors of satellite b swap their positions. The new satellite positions are labeled as a', b', and c' in the figure. During the transition, the ISL links $a - b$ and $b - c$ are turned off.

Any connection is subject to rerouting if it is passing through a link that will be turned off before the connection is over. This event is referred to as link handoff. If the number of connections that need to be rerouted due to link handoff is large, the resulting rerouting attempts cause

signaling overhead in the network. The number of rerouting attempts can be reduced if the dynamic topology of the network is taken into account when connection routes are determined during call set up. The routing problem in LEO satellite networks has been addressed in [119] with an emphasis on setting up routes between pairs of satellites to minimize the rerouting attempts during link handoffs, i.e., optimization was performed for the routes between two satellites. Optimization process results in a unique route with minimum number of link handoffs during a system period[1] for each satellite pair. All end-user connections that are served by the same satellite pair use the same unique route. This algorithm reduces the link handoff frequency; however, it can also congest some of the links, while it underutilizes some others. An optimal route between two satellite nodes is not necessarily optimum for a connection between two ground terminals since the handoffs between the ground terminals and the satellites result in changing satellite end nodes for the connection. Realistically, the optimization is needed for the route between two ground terminals.

In [32], a LEO satellite network is modeled as a finite state automaton (FSA) by dividing the period of the satellite network into equal-length intervals. In the FSA model, each state corresponds to one of these intervals, and in each state, the network is considered as having a fixed topology. Each satellite has a fixed number of ISL's. The algorithm determines the optimum link assignment for each satellite in a given state using simulated annealing. This study does not address reducing the number of rerouting attempts due to link handoffs. In contrast, more connections would need to be rerouted during the state changes of the FSA model since the optimization process uses only the traffic pattern.

In [111] and [108], a routing protocol, referred to as probabilistic routing protocol (PRP), has been proposed to reduce the number of rerouting attempts during a link handoff. The algorithm removes all the ISL's that will experience a link handoff during the lifetime of a connection from consideration for routing during the route establishment phase of a new call. However, since the call holding time is a random variable, the connection lifetime cannot be determined exactly. Instead the PRP finds the time duration in which the route will be used by the user terminals with a certain probability that is referred to as target probability. As a result, the route does not experience any link handoff with a probability larger than the target probability. The performance evaluation results show that a tradeoff exists between the value of the target probability and the new call blocking probability.

C. Open Problems

The specific research directions identified for the future of mobility management for satellite networks are as follows.

1) *Broad-Band Traffic:* Development of handoff algorithms applicable to broad-band traffic. The algorithm should guarantee that the user's QoS is not violated during the handoff process.
2) *Minimize Signaling:* Design of a handoff rerouting algorithm that utilizes LEO satellite network dynamics to minimize signaling traffic during the rerouting phase.
3) *Routing:* Development of a routing algorithm that minimizes the number of rerouting attempts during a link handoff. Especially the routing problem for a connectionless network protocol like IP should be studied for the LEO satellite network environment.
4) *Update and Paging:* Design of a location update and paging algorithm that achieves a balance between location update frequency and paging delay.
5) *Evaluation of Algorithms:* Development of a simulation library to evaluate mobility management algorithms in LEO satellite networks.

Each of the four backbone networks discussed for this section has specific mobility management issues that must be addressed. However, many of the issues are not backbone specific. For example, the paging concerns for PCS networks are the same for WATM and satellite. Advertisement and the use of hierarchies can be explored for all networks. Smooth handoffs and connection rerouting are also required by each of the networks. Future networks must capitalize on these common issues in order to bring about interoperation that can be implemented in as simplified a method as possible. Next we explore some of the issues that are introduced by bringing heterogeneous networks and their services together under one unifying infrastructure.

VIII. Research Issues for Integrated Wireless Networks

As mentioned in Section I, the next generation of wireless communication networks promise to bring together mobility—not only without geographical constraints—but also without being tied to one particular backbone network, as shown in Fig. 4. The user that has been receiving home services in an urban setting with access to the Internet will still be able to access services from completely rural sites, which may only have access to satellite. As another example, consider the terminal that roams between an ATM service region and a Mobile IP service region. The authors in [8] examined such a case and present an overview of interworking Mobile IP with mobile ATM. When a mobile terminal moves into an ATM service region, the mobility management is governed by WATM, but this process is transparent to Mobile IP. When the terminal crosses into a new region, say Mobile IP region, the mobility is supported by Mobile IP. This will require a gateway at which the translation of the Mobile IP datagrams into WATM functionality will occur. Likewise, any network equipped for unified operation must be able to support such qualities as intercarrier or intersystem handoff, personal mobility, and location management for a heterogeneous network. Next we explore the following mobility-related research issues for next-generation wireless networks: hardware/software for-

[1] System period is defined as the time interval where a satellite circulates the Earth.

mat transformations; location management; and intersystem handoff.

A. Software Radio

To implement service provider portability, the MT must be able to communicate in more than one system. MT's may operate in multiple modes, with separate transmitter/receiver pairs, such as the satellite/terrestrial multimode terminals, or the MT's may be reconfigured to operate in each new system. An emerging advance in technology is a terminal that is able to download settings from a network server in order to obtain a set of standards and then reconfigure itself to communicate in the new system [61]. The ability to reconfigure the radio interface or the radio protocol stack by software is referred to as software radio [106]. Software radio download techniques allow networks to offer new, personalized, or operator-customized services to a user without requiring an upgrade in the user's terminal equipment. It facilitates roaming and flexible service support, as well as terminal and BS evolution. Although software radio is still in development, work is beginning on the building of tools, libraries, and environments to support product development and software portability [106], [127].

B. Location Management

As technology enables MT's to communicate in various systems, the networks will need to be able to track and update the location of terminals that may or may not be registered within the current system. The stand-alone cases described in the previous sections demonstrate that there are two problems related to location management: location update and paging, both of which consume the limited radio resources in the mobile system. Thus, it is necessary to achieve a tradeoff between cost and performance. For location update, an optimal scheme must be determined to minimize the signaling cost for location update when an MT is roaming among different systems with individual protocols and signaling formats. For the paging system, paging delay and paging cost are two key factors. The paging delay may be more important because of the QoS requirement of the multimedia services. Based on previous location management schemes and the virtue of the heterogeneous wireless network, we are working on corresponding location update and paging schemes [10]. For paging, we are developing a distributed paging scheme with delay bound which is desirable for QoS specifications of next-generation wireless systems [11].

C. Intersystem Handoff

Along with optimizing schemes for location management, the MT must simultaneously be prepared to transfer connections between systems with a low probability of losing the call. Intersystem mobility techniques must distribute handoff management so that calls are routed with optimum efficiency and minimum delay, and so that independent subnetworks can maintain service to their subscribers. To implement intersystem handoff, several major issues must be resolved [41], [65]. First, as mentioned above, the MT must be able to communicate in more than one system. Second, a technique is needed to measure and compare signals from different air interfaces and with different power levels. Third, transmission and signaling facilities must exist between the switches of each system. Last, a new set of handoff procedures and messages must be developed in order to prepare the MT for connection transfers and to maintain a satisfactory level of QoS for the user. Research efforts have addressed the issue of optimizing a homogeneous system for next-generation handoff [37], [38], [59], [107]. However, research results are now being produced for the heterogeneous case [12].

Finally, future research must focus on bringing together consistent as well as reliable services so that the user will not experience service degradation from backbone to backbone. Future communication networks must be able to handle network administrative functions involved in obtaining international or regional permissions for carrying a mobile terminal into a visited country [60].

D. Addressing and Identification

Future networks must identify several types of entities. The mobility properties of these entities necessitate dynamic bindings between their addresses and names. The name and address of a network end point will cease to identify the terminal or the user. The dynamic binding of terminals to their point of attachment to the network will change frequently because of global roaming.

Within the network, a mobile terminal may transition from the PSTN to the Internet, from the Internet to ATM, from ATM to satellite, or from/to any other combination thereof. Thus, a unifying transport method must be developed. This transport method will be required to track crucial mobile characteristics, such as addresses and identification. The use of well-defined and standardized user/terminal identities are needed to manage the following location and handoff operations [82]:

1) determination of the home network or database of a roaming terminal;
2) identification of an MT on the radio control path for update/registration;
3) identification of the MT for terminal paging and location advertisement;
4) identification of the MT when exchanging location or routing information between different network types;
5) identification of the MT for updating or retrieving the user profile.

In addition to the location and handoff operations, another issue is security. Providing access for every type of network under various mobile environments and the resulting complexity of the system leaves mobility-related procedures very vulnerable. In addition, since location information about the user will be extensively used, the providers with unlimited access to management information must come under scrutiny in order to maintain overall privacy

and confidentiality. Technical strategies must be developed for achieving reliable authentication and yet a level of untraceability for roaming subscribers—even against the providers of the systems.

E. Database Issues

The most challenging issue for location management for the next generation of communication systems is database storage and retrieval. For example, consider the mobile terminal whose home network is the PSTN, but whose current visited network is WATM based. In order to register its current location with its home network, the terminal must be registered as a visitor on the network level with the ATM network, possibly requiring database update/query, and then the ATM network must send this updated profile to the mobile's home network for further database update. Likewise, for a WATM network that employs location advertisement or terminal paging, call delivery from the PSTN to the WATM network will require increased signaling as well. Thus, increase of signaling related to mobility will result in a large increase in the number of database transactions, paging operations, and broadcasts [74].

F. Routing Issues

Since the offered services will support point-to-point, multipoint, and point-to-multipoint communication between fixed and/or mobile terminals, connection rerouting will become a major consideration for service availability for the roaming terminal. Maintaining a connection will be more complex, since different networks provide different connection information. For example, ATM networks do not provide cell ordering numbers, but cell sequencing becomes important for wireless communication. Also, the tunneling and routing procedures of handoff for the Internet may undergo problems while the MT is traveling through a satellite-only environment.

Users must be able to negotiate bandwidth, transmission quality, and delay based upon the service profile of the user and the service offerings from the user's service provider. The user will be able to choose services available in that region from a menu or based on their subscription profile. Next-generation services will be continuous and seamless whether the user is on a ship, in a vehicle, or on a plane. It is expected that a variety of services will be available in rural areas as well as dense urban areas. Services requiring higher transmission rates will be accessible in high-density areas such as business offices, while lower speed services will be accessible in all other environments. The services must be provided with continuous availability on a global basis regardless of the user's roaming situation, geographic location, or access within the limits of the visited network.

G. Standardization

Considering the proposed unification, there is a great need for common management capabilities in order to handle the network-level mobility requirements [29]. However, standardization seems to be the slowest area of progress. While there is general agreement among Japan, Europe, the United States, and other countries that global standards are in everyone's best interests, some difficulties still exist in achieving the necessary cooperation between regional and international bodies. On the network service provider level, many may choose to specialize operations in order to serve their own customers. In order to obtain the projected level of tetherless global communication, each of these entities must be able to coexist and cooperate within one infrastructure [29].

As in previous generations, the next generation of wireless systems will be gradually implemented over the current infrastructures. As a result, the most likely scenarios will begin with Mobile IP interworking with ATM or WATM, and PLMN-based terrestrial networks interworking with satellite networks as traffic congestion relief. However, as research continues to explore options for integrating network services, the boundaries that prohibit global freedom for wireless communication will continue to disappear.

REFERENCES

[1] EIA/TIA, "Cellular radio-telecommunications intersystem operations," EIA/TIA, Tech. Rep. IS-41 Revision C, 1995.

[2] ——, "Mobile station-land station compatibility specification," EIA/TIA, Tech. Rep. 553, 1989.

[3] ——, "Cellular system dual-mode mobile station–Base station compatibility standard," EIA/TIA, Tech. Rep. IS-54, 1992.

[4] ETSI/TC, "Mobile application part (MAP) specification, version 4.8.0," Tech. Rep., Recommendation GSM 09.02, 1994.

[5] A. Acampora, "Wireless ATM: A perspective on issues and prospects," *IEEE Personal Commun.*, vol. 3, pp. 8–17, Aug. 1996.

[6] ——, "An architecture and methodology for mobile-executed handoff in cellular ATM networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1365–1375, Oct. 1994.

[7] A. Acharya, J. Li, F. Ansari, and D. Raychaudhuri, "Mobility support for IP over wireless ATM," *IEEE Commun. Mag.*, vol. 36, pp. 84–88, Apr. 1998.

[8] A. Acharya, J. Li, B. Rajagopalan, and D. Raychaudhuri, "Mobility management in wireless ATM networks," *IEEE Commun. Mag.*, vol. 35, pp. 100–109, Nov. 1997.

[9] I. F. Akyildiz, H. Uzunalioğlu, and M. D. Bender, "Handover management in low earth orbit (LEO) satellite networks," *ACM-Baltzer J. Mobile Networks Applicat. (MONET)*, to be published.

[10] I. F. Akyildiz, G. L. Stuber, and W. Wang, "New location management schemes for next generation wireless systems," Broadband and Wireless Networking Lab., Georgia Inst. Technol., Atlanta, Tech. Rep., Feb. 1999.

[11] I. F. Akyildiz and W. Wang, "Paging algorithms for next generation wireless systems," Georgia Inst. Technol., Atlanta, Tech. Rep., June 1999.

[12] I. F. Akyildiz and J. McNair, "Handoff techniques for next generation wireless systems," Broadband and Wireless Networking Lab., Georgia Inst. Technol., Atlanta, Tech. Rep., June 1999.

[13] I. F. Akyildiz, J. McNair, J. S. M. Ho, H. Uzunalioğlu, and W. Wang, "Mobility management in current and future communication networks," *IEEE Network Mag.*, vol. 12, pp. 39–49, Aug. 1998.

[14] I. F. Akyildiz and J. S. M. Ho, "On location management for personal communications networks," *IEEE Commun. Mag.*, vol. 34, pp. 138–145, Sept. 1996.

[15] I. F. Akyildiz, J. S. M. Ho, and Y. B. Lin, "Movement-based location update and selective paging for PCS networks," *IEEE/ACM Trans. Networking*, vol. 4, pp. 629–636, Aug. 1996.

[16] 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, pp. 187–196, July 1995.

[17] B. Akyol and D. Cox, "Re-routing for handoff in a wireless ATM network," *IEEE Personal Commun.*, vol. 3, pp. 26–33, Oct. 1996.

[18] ____, "Handling mobility in a wireless ATM network," in *Proc. IEEE INFOCOM '96*, vol. 3, no. 8, pp. 1405–1413.
[19] F. Ananasso and F. D. Priscoli, "Issues on the evolution toward satellite personal communication networks," in *Proc. GLOBECOM'95*, pp. 541–545.
[20] V. Anantharam, M.L. Honig, U. Madhow, and V.K. Wei, "Optimization of a database hierarchy for mobility tracking in a personal communications network," *Performance Evaluation*, vol. 20, no. 1-3, pp. 287–300, May 1994.
[21] E. Ayanoglu, K. Eng, and M. Karol, "Wireless ATM: Limits, challenges, and proposals," *IEEE Personal Commun.*, vol. 3, pp. 19–34, Aug. 1996.
[22] B. R. Badrinath, T. Imielinski, and A. Virmani, "Locating strategies for personal communication network," in *Proc. Workshop Networking of Personal Communications Applications*, Dec. 1992.
[23] A. Bar-Noy, I. Kessler, and M. Sidi, "Topology-based tracking strategies for personal communication networks," *ACM-Baltzer J. Mobile Networks and Applications (MONET)*, vol. 1, no. 1, pp. 49–56, 1996.
[24] ____, "Mobile users: To update or not to update?" *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, pp. 175–186, July 1995.
[25] G. Bautz and J. Johnsson, "Proposal for location management in WATM," in ATM Forum/96-1516, Vancouver, Canada, Dec. 1996.
[26] C. B. Becker, B. Patil, and E. Qaddoura, "IP mobility architecture framework," Internet Engineering Task Force, Internet draft, draft-ietf-mobileip-ipm-arch-00.txt, Mar. 1999.
[27] J. M. Benedetto, "Economy-class ion-defying IC's in orbit," *IEEE Spectrum*, vol. 35, pp. 36–41, Mar. 1998.
[28] S. Biswas and A. Hopper, "A representative based architecture for handling mobility in connection oriented radio networks," in *Proc. ICUPC '95*, pp. 848–852.
[29] K. Buchanan, R. Fudge, D. McFarlane, T. Phillips, A. Sasaki, and H. Xia, "IMT 2000: Service provider's perspective," *IEEE Personal Commun.*, vol. 4, pp. 8–13, Aug. 1997.
[30] R. Caceres and V. Padmanabhan, "Fast and scalable handoffs for wireless networks," in *Proc. ACM/IEEE MOBICOM'96*, pp. 56–66.
[31] P. Calhoun and C. Perkins, "Tunnel establishment protocol," Internet Engineering Task Force, Internet draft, draft-ietf-mobileip-calhoun-tep-00.txt, 21 Nov. 1997.
[32] H. S. Chang, B. W. Kim, C. G. Lee, Y. H. Choi, S. L. Min, H. S. Yang, and C. S. Kim, "Topological design and routing for low-earth orbit satellite networks," in *Proc. IEEE GLOBECOM 95*, pp. 529–535.
[33] S. Dolev, D. K. Pradhan, and J. L. Welch, "Modified tree structure for location management in mobile environments," *Comput. Commun.*, vol. 19, no. 4, pp. 335–345, 1996.
[34] G. Dommetry, M. Veeraraghavan, and M. Singhal, "Route optimization in mobile ATM networks," in *Proc. ACM/IEEE MOBICOM '97*, pp. 43–54.
[35] F. Dosiere, T. Zein, G. Maral, and J. P. Boutes, "A model for the handover traffic in low earth-orbiting (LEO) satellite networks for personal communications," *Int. J. Satellite Commun.*, vol. 11, pp. 145–149, May–June 1993.
[36] D. Dykeman, I. Iliadis, P. Scotton, L. Frelechoux, and S. Ray, "PNNI routing support for mobile networks," in ATM Forum/97-0766, Paris, France, Sept. 1997.
[37] N. Efthymiou, Y. F. Hu, and R. Sheriff, "Performance of inter-segment handover protocols in an integrated space/terrestrial-UMTS environment," *IEEE Trans. Veh. Technol.*, vol. 47, pp. 1179–1199, Nov. 1998.
[38] A. El-Hoiydi, "Radio independence in the network architecture of the universal mobile telecommunication system," in *Proc. IEEE GLOBECOM '98*, pp. 830–835.
[39] K. Y. Eng, M. J. Karol, M. Veeraraghavan, E. Ayanoglu, C. B. Woodworth, P. Pancha, and R. A. Valenzuela, "A wireless broadband ad-hoc ATM local-area network," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, pp. 161–174, 1995.
[40] G. Fleming, A. Hoiydi, J. de Vriendt, G. Nikolaidis, F. Piolini, and M. Maraki, "A flexible network architecture for UMTS," *IEEE Personal Commun. Mag.*, vol. 5, pp. 8–15, Apr. 1998.
[41] V. K. Garg and J. E. Wilkes, "Interworking and interoperability issues for North American PCS," *IEEE Commun. Mag.*, vol. 34, pp. 94–99, Mar. 1996.
[42] E. Guarene, P. Fasano, and V. Vercellone, "IP and ATM integration perspectives," *IEEE Commun. Mag.*, vol. 36, pp. 74–80, Jan. 1998.
[43] L. van Hauwermeiren, L. Vercauteren, A. Saidi, and T. Landegem, "Requirements for mobility support in ATM," in *Proc. IEEE GLOBECOM '94*, pp. 1691–1695.
[44] J. S. M. Ho and I. F. Akyildiz, "Dynamic hierarchical database architecture for location management in PCS networks," *IEEE/ACM Trans. Networking*, vol. 5, no. 5, pp. 646–661, Oct. 1997.
[45] ____, "Local anchor scheme for reducing signaling cost in personal communication networks," *IEEE/ACM Trans. Networking*, vol. 4, no. 5, pp. 709–726, Oct. 1996.
[46] ____, "A mobile user location update and paging mechanism under delay constraints," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 4, pp. 413–425, Dec. 1995.
[47] D. Hong and S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 35, pp. 77–92, Aug. 1986.
[48] L.-R. Hu and S. Rappaport, "Adaptive location management scheme for global personal communications," in *Proc. IEEE Communications*, vol. 144, no. 1, 1997, pp. 54–60.
[49] C.-L. I, G. P. Pollini, and R. D. Gitlin, "PCS mobility management using the reverse virtual call setup algorithm," *IEEE/ACM Trans. Networking*, vol. 5, pp. 13–24, Feb. 1997.
[50] R. Jain and Y. B. Lin, "An auxiliary user location strategy employing forwarding pointers to reduce network impact of PCS," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, pp. 197–210, July 1995.
[51] R. Jain, Y. B. Lin, and S. Mohan, "A caching strategy to reduce network impacts of PCS," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1434–1444, Oct. 1994.
[52] D. B. Johnson and C. Perkins, "Mobility support in IPv6," Internet Engineering Task Force, Internet draft, draft-ietf-mobileip-ipv6-04.txt, Nov. 1997.
[53] D. Johnson and D. Maltz, "Protocols for adaptive wireless and mobile networking," *IEEE Personal Commun.*, vol. 3, pp. 34–42, Feb. 1996.
[54] M. Johnsson, "Simple mobile IP," Internet Engineering Task Force, Internet-draft, Ericsson, draft-ietf-mobileip-simple-00.txt, Mar. 1999.
[55] C. Kalmanek, P. Mishra, M. Srivastava, and B. Rajagopalan, "Benchmark methodology for evaluation of wireless ATM VC rerouting schemes," in ATM Forum/97-0852, Paris, France, Sept. 1997.
[56] S. J. Kim and C. Y. Lee, "Modeling and analysis of the dynamic location registration and paging in microcellular systems," *IEEE Trans. Veh. Technol.*, vol. 45, pp. 82–89, Feb. 1996.
[57] P. Krishna, N. Vaidya, and D. K. Pradhan, "Static and adaptive location management in mobile wireless networks," *Comput. Commun.*, vol. 19, no. 4, pp. 321–334, 1996.
[58] A. S. Krishnakumar, "ATM without strings: An overview of wireless ATM," in *Proc. IEEE Conf. Personal Wireless Commun. (ICPWC '96)*, pp. 216–221.
[59] Y. H. Kwon, D. K. Kim, J. H. Jung, M. K. Choi, D. K. Sung, H. K. Yoon, and W. Y. Han, "Effect of soft handoffs on the signalling traffic in IMT-2000 networks," in *Proc. IEEE GLOBECOM '98*.
[60] F. Leite, R. Engelman, S. Kodama, H. Mennenga, and S. Towaij, "Regulatory considerations relating to IMT 2000," *IEEE Personal Commun.*, pp. 14–19, Aug. 1997.
[61] P. Lettieri and M. Srivastava, "Advances in wireless terminals," *IEEE Personal Commun. Mag.*, vol. 6, pp. 6–19, Feb. 1999.
[62] B. Li, S. Jiang, and D. Tsang, "Subscriber-assisted handoff support in multimedia PCS," *Mobile Comput. Commun. Rev.*, vol. 1, no. 3, pp. 29–36, Sept. 1997.
[63] Y. B. Lin "Paging systems: Network architectures and interfaces," *IEEE Network*, vol. 11, pp. 56–61, July–Aug. 1997.
[64] ____, "Reducing location update cost in a PCS network," *IEEE/ACM Trans. Networking*, vol. 5, pp. 25–33, Feb. 1997.
[65] Y.-B. Lin and I. Chlamtac, "Heterogeneous personal communication services: Integration of PCS systems" *IEEE Commun. Mag.*, vol. 34, pp. 106–113, Sept. 1996.
[66] Y. B. Lin, F. C. Li, A. Noerpel, and I. P. Kun, "Performance modeling of multitier PCS system," *Int. J. Wireless Information Networks*, vol. 3, no. 2, pp. 67–78, 1996.

[67] Y. B. Lin and S. K. DeVries, "PCS network signaling using SS7," *IEEE Commun. Mag.*, vol. 33, pp. 44–55, June 1995.
[68] Y. B. Lin, "Determining the user locations for personal communications services networks," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 466–473, Aug. 1994.
[69] G. Losquadro, A. Aerospazio, and R. Sheriff, "Requirements of multiregional mobile broadband satellite networks," *IEEE Personal Commun. Mag.*, vol. 5, pp. 26–30, Apr. 1998.
[70] C. N. Lo, S. Mohan, and R. S. Wolff, "Performance modeling and simulation of data management for personal communications applications," in *Proc. IEEE PIMRC '92*, pp. 1210–1214.
[71] C. N. Lo, R. S. Wolff, and R. C. Bernhardt, "An estimate of network database transaction volume to support personal communications services," in *Proc. Int. Conf. Universal Personal Communications*, 1992, pp. 236–241.
[72] E. Lutz, "Issues in satellite personal communication systems," *ACM J. Wireless Networks*, vol. 4, no. 2, pp. 109–124, May 1998.
[73] U. Madhow, M. L. Honig, and K. Steiglitz, "Optimization of wireless resources for personal communications mobility tracking," in *Proc. IEEE INFOCOM '94*, pp. 577–584.
[74] J. Markoulidakis, G. Lyberopoulos, D. Tsirkas, and E. Sykas, "Mobility modeling in third-generation mobile telecommunications systems," *IEEE Personal Commun.*, vol. 4, pp. 41–56, Aug. 1997.
[75] M. Marsan, C.-F. Chiasserini, R. Lo Cigno, M. Munafo, and A. Fumagalli, "Local and global handovers for mobility management in wireless ATM networks," *IEEE Personal Commun.*, vol. 4, pp. 16–24, Oct. 1997.
[76] I. Mertzanis, R. Tafazolli, and B. G. Evans, "Connection admission control strategy and routing considerations in multimedia (non-Geo) satellite networks," in *Proc. IEEE VTC'97*, pp. 431–435.
[77] B. Miller, "Satellite free mobile phone," *IEEE Spectrum*, vol. 35, pp. 26–35, Mar. 1998.
[78] A. Misra and L. Tassiulas, "Optimizing paging and registration costs for location tracking in satellite-based personal communications," in *Proc. WOSBIS'97*, pp. 49–57.
[79] A. R. Modarressi and R. A. Skoog, "Signaling system 7: A tutorial," *IEEE Commun. Mag.*, vol. 28, pp. 19–35, July 1990.
[80] S. Mohan and R. Jain, "Two user location strategies for personal communications services," *IEEE Personal Commun.*, vol. 1, pp. 42–50, 1994.
[81] M. Mouly and M. B. Pautet, "The GSM system for mobile communications," M. Mouly, Palaiseau, France, Tech. Rep. 1992.
[82] R. Pandya, D. Grillo, E. Lycksell, P. Mieybegue, H. Okinaka, and M. Yabusaki, "IMT-2000 standards: Network aspects," *IEEE Personal Commun.*, pp. 20–29, Aug. 1997.
[83] C. E. Perkins, *Mobile IP: Design Principles and Practices* (Addison-Wesley Wireless Communications Series). Reading, MA: Addison Wesley, 1998.
[84] ——, "IP mobility support version 2," Internet Engineering Task Force, Internet draft, draft-ietf-mobileip-v2-00.text, Nov. 1997.
[85] C. Perkins and D. Johnson, "Route optimization in mobile IP," Internet Engineering Task Force, Internet draft, draft-ietf-mobileip-optom-07.txt, Nov. 20, 1997.
[86] C. Perkins, "Mobile IP," *IEEE Commun. Mag.*, vol. 35, pp. 84–99, May 1997.
[87] ——, "Mobile-IP local registration with hierarchical foreign agents," Internet Engineering Task Force, Internet draft, draft-perkins-mobileip-hierfa-00.txt, Feb. 1996.
[88] B. Rajagopalan, "An overview of ATM forum's wireless ATM standards activities," *ACM Mobile Comput. Commun. Rev.*, vol. 1, no. 3, Sept. 1997.
[89] ——, "Mobility management in integrated wireless-ATM networks," *ACM-Baltzer J. Mobile Networks Applicat. (MONET)*, vol. 1, no. 3, pp. 273–285, 1996.
[90] E. del Re, "A coordinated European effort for the definition of a satellite integrated environment for future mobile communications," *IEEE Commun. Mag.*, vol. 34, pp. 98–104, Feb. 1996.
[91] E. del Re, R. Fantacci, and G. Giambene, "Call blocking performance for dynamic channel allocation technique in future mobile satellite systems," *Proc. Inst. Elect. Eng., Commun.*, vol. 143, no. 5, pp. 289–296, 1996.
[92] ——, "Handover requests queueing in low earth orbit mobile satellite systems," in *Proc. 2nd Europ. Workshop Mobile/Personal Satcoms*, 1996, pp. 213–232.
[93] C. Rose, "State-based paging/registration: A greedy technique," *IEEE Trans. Veh. Technol.*, vol. 48, pp 166–173, Jan. 1999.
[94] C. Rose and R. Yates, "Ensemble polling strategies for increased paging capacity in mobile communication networks," *ACM/Baltzer Wireless Networks J.*, vol. 3, no. 2, pp. 159–167, Sept. 1997.
[95] ——, "Location uncertainty in mobile networks: A theoretical framework," *IEEE Commun. Mag.*, vol. 35, pp. 94–101, Feb. 1997.
[96] C. Rose "Minimizing the average cost of paging and registration: A timer-based method," *ACM-Baltzer J. Wireless Networks*, vol. 2, no. 2, pp. 109–116, 1996.
[97] C. Rose and R. Yates, "Minimizing the average cost of paging under delay constraints," *ACM-Baltzer J. Wireless Networks*, vol. 1, no. 2, pp. 211–219, July 1995.
[98] P. Shieh, T. Tedijanto, and R. Rennison, "Handoff schemes to support mobility in wireless ATM," in ATM Forum/96-1622, Vancouver, Canada, Dec. 1996.
[99] N. Shivakumar and J. Widom, "User profile replication for faster location lookup in mobile environments," in *Proc. ACM/IEEE MOBICOM'95*, pp. 161–169.
[100] D. Steer and J. Chow, "Requirements for 'soft' handover," in ATM Forum/97-0696, Paris, France, Sept. 1997.
[101] S. Tabbane, "Location management methods for 3rd generation mobile systems," *IEEE Commun. Mag.*, vol. 35, pp. 72–78, Aug. 1997.
[102] C.-K. Toh, "A unifying methodology for handovers of heterogeneous connections in wireless ATM networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 1, pp. 12–30, Jan. 1997.
[103] ——, "A hybrid handover protocol for local area wireless ATM networks," *ACM-Baltzer J. Mobile Networks Applicat. (MONET)*, vol. 1, no. 3, pp. 313–334, 1996.
[104] ——, "The design and implementation of a hybrid handover protocol for multimedia wireless LAN's," in *Proc. ACM/IEEE MOBICOM '95*, pp. 49–61.
[105] G. Troxel and L. Sanchez, "Rapid authentication for mobile IP," Internet Engineering Task Force, Internet draft, draft-ietf-mobileip-ra-00.txt, Dec. 1997.
[106] W. H. W. Tuttlebee, "Software-defined radio: Facets of a developing technology," *IEEE Personal Commun. Mag.*, vol. 6, pp. 38–44, Apr. 1999.
[107] K. Ushiki and M. Fukazawa, "A new handover method for next generation mobile communication systems," in *Proc. IEEE GLOBECOM '98*, Nov. 1998, pp. 1118–1123.
[108] H. Uzunalioğlu, M. D. Bender, and I.F. Akyildiz, "A routing algorithm for low earth orbit satellite networks with dynamic connectivity," *ACM-Baltzer J. Wireless Networks (WINET)*, to be published.
[109] H. Uzunalioğlu, J. W. Evans, and J. Gowens, "A connection admission control algorithm for low earth orbit satellite networks," in *Proc. IEEE ICC'99*, pp. 1074–1078.
[110] H. Uzunalioğlu, I. F. Akyildiz, Y. Yesha, and W. Yen, "Footprint handover rerouting protocol for low earth orbit satellite networks," *ACM-Baltzer J. Wireless Networks (WINET)*, to be published.
[111] H. Uzunalioğlu, "Probabilistic routing protocol for low earth orbit satellite networks," in *Proc. IEEE ICC'98*, pp. 89–93.
[112] H. Uzunalioğlu, W. Yen, and I. Akyildiz, "A connection handover protocol for LEO satellite ATM networks," in *Proc. ACM/IEEE MOBICOM '97*, pp. 204–214.
[113] M. Veeraraghavan and G. Dommetry, "Mobile location management in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1437–1454, Oct. 1997.
[114] M. Veeraraghavan, M. Karol, and K. Eng, "Mobility and connection management in a wireless ATM LAN," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 50–68, Jan. 1997.
[115] M. Veeraraghavan and G. Dommetry, "Location management update," in ATM Forum/96-1701, Vancouver, Canada, Dec. 1996.
[116] M. Veeraraghavan, M. Karol, and K. Eng, "A combined handoff scheme for mobile ATM networks," in ATM Forum/96-1700, Vancouver, Canada, Dec. 1996.

[117] G. Wan and E. Lin, "A dynamic paging scheme for wireless communication systems," *ACM/IEEE MOBICOM '97*, pp. 195–203.
[118] J. Z. Wang, "A fully distributed location registration strategy for universal personal communication systems," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 850–860, Aug. 1993.
[119] M. Werner, C. Delucchi, H.-J. Vogel, G. Maral, and J.-J. De Ridder, "ATM-based routing in LEO/MEO satellite networks with intersatellite links," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 69–82, Jan. 1997.
[120] M. Werner, A. Jahn, E. Lutz, and A. Bottcher, "Analysis of system parameters for LEO/ICO-satellite communication networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 371–381, Feb. 1995.
[121] N. W. Whinnett, "Handoff between dissimilar systems: General approaches and air interface issues for TDMA systems," in *Proc. IEEE Vehicular Technology Conf., VTC '95*, vol. 2, pp. 953–957.
[122] D. R. Wilson, "Signaling system no. 7, IS-41 and cellular telephony networking," *Proc. IEEE*, vol. 80, pp. 664–652, Apr. 1992.
[123] H. Xie, S. Tabbane and D. Goodman, "Dynamic location area management and performance analysis," in *Proc. IEEE VTC '93*, pp. 536–539.
[124] R. Yates, C. Rose, B. Rajagopalan, and B. Badrinath, "Analysis of a mobile-assisted adaptive location management strategy," *ACM-Baltzer J. Mobile Networks Applicat. (MONET)*, vol. 1, no. 2, pp. 105–112, 1996.
[125] A. Yener and C. Rose, "Highly mobile users and paging: Optimal polling strategies," *IEEE Trans. Veh. Technol.*, vol. 47, pp. 1251–1257, Nov. 1998.
[126] R. Yuan, S. K. Biswas, L. J. French, J. Li, and D. Raychaudhuri, "A signaling and control architecture for mobility support," *ACM-Baltzer J, Mobile Networks Applicat. (MONET)*, vol. 1, no. 3, pp. 287–298, Jan. 1996.
[127] *IEEE Commun. Mag. (Special Issue on Globalization of Software Radio)*, vol. 37, pp. 82–123, Feb. 1999.

Ian F. Akyildiz (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in computer engineering from the University of Erlangen-Nuernberg, Erlangen, Germany, in 1978, 1981, and 1984, respectively.

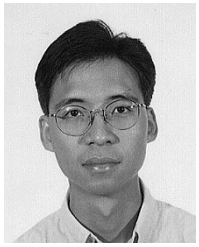
Currently, he is a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, where he serves as the elected Chair of the Telecommunications Area and Director of the Broadband and Wireless Networking Laboratory. He has held visiting professorships at the Universidad Tecnica Federico Santa Maria, Chile, Universite Pierre et Marie Curie (Paris VI), Ecole Nationale Superieure Telecommunications in Paris, France, Universidad Politecnico de Cataluna, Barcelona, Spain, and Universidad Illes Baleares, Palma de Mallorca, Spain. He has published over 200 technical papers in journals and conference proceedings. His current research interests are in wireless networks, satellite networks, ATM networks, Internet, and multimedia communication systems.

Dr. Akyildiz is an Editor for IEEE/ACM TRANSACTIONS ON NETWORKING, *Computer Networks and ISDN Systems Journal*, *ACM-Springer Journal for Multimedia Systems*, *ACM-Baltzer Journal of Wireless Networks*, and the *Journal of Cluster Computing*. He was an Editor for IEEE TRANSACTIONS ON COMPUTERS from 1992 to 1996, and he was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Networks in the Metropolitan Area. He was the Program Chair of the Ninth IEEE Computer Communications Workshop in October 1994, and he also served as the Program Chair for the ACM/IEEE MOBICOM'96 and IEEE INFOCOM'98 conferences. He is an ACM Fellow. He received the Don Federico Santa Maria Medal for his services to the Universidad of Federico Santa Maria in Chile. He served as a National Lecturer for ACM from 1989 until 1998 and received the ACM Outstanding Distinguished Lecturer Award in 1994 and the 1997 IEEE Leonard G. Abraham Prize.

Janise McNair (Student Member, IEEE) received the B.S. and M.S. degrees from the University of Texas, Austin, in 1991 and 1994, respectively. She is currently pursuing the Ph.D. degree at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta.

She is a Research Assistant in the Broadband and Wireless Networking Laboratorym Georgia Institute of Technology. Her research interests include wireless multimedia networks, mobility management, and satellite networks.

Ms. McNair was a National Science Foundation Fellow from 1991 to 1994.



Joseph S. M. Ho (Member, IEEE) received the B.S.E.E. and M.S.E.E. degrees from the University of Washington, Seattle, in 1987 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 1996.

He is currently a Senior Member of Scientific Staff at Nortel Networks, Richardson, TX. His current research interests include design and analysis of IP-based wireless packet data networks, development of algorithms and network architecture for supporting mobility and quality of service (QoS), modeling and analysis of multimedia data traffic, and performance evaluation. He has published extensively and has several pending U.S. and international patents in the wireless networking area.

Dr. Ho is a member of Tau Beta Pi.

Hüseyin Uzunalioğlu (Member, IEEE) received the B.S. and M.S. degrees from the Middle East Technical University, Ankara, Turkey, in 1990 and 1992, respectively, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, in 1998, all in electrical engineering.

He worked as a Researcher in the Georgia Tech Research Institute from 1995 to 1998. He is currently with Network Planning Solutions at Bell Laboratories, Lucent Technologies, Holmdel, NJ. His current research interests are in traffic and mobility management for broad-band networks.

Dr. Uzunalioğlu is a recipient of a research scholarship from Middle East Technical University.

Wenye Wang (Student Member, IEEE) received the B.S and M.S degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 1986 and 1991, respectively. She is currently pursuing the Ph.D degree in the Broadband and Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta.

Her research interests include resource allocation for wireless multimedia systems and location management for IMT-2000 systems.

Ms. Wang is member of ACM.