

Chapter 2

Mobility Management Reference Models

This chapter is dedicated to the mobility management protocol reference model and the network reference model. These models were proposed by the authors based on a summary of various mobility management technologies.

2.1 Protocol Reference Model

The protocol reference model for mobility management is illustrated in Fig. 2.1. The model is composed of three planes: the data plane, the control plane and the management plane. The data plane illustrates the protocol layers that are involved in mobility management, including the physical layer, the data link layer, the network layer, the transport layer and the application layer. The control plane presents the key control functions in mobility management, including the security mechanism, location management, handover control, and interoperability control. The management plane handles the network management-related functions, including configuration management, fault management, performance management, accounting management, and security management. The functions and typical technologies of these planes will be introduced in Sects. 2.1.1, 2.1.2 and 2.1.3.

2.1.1 Data Plane

The data plane in the protocol reference model follows the revised Transmission Control Protocol/Internet Protocol-layered architecture (TCP/IP), which is composed of the physical layer, the data link layer, the network layer, the transport layer, and the application layer, from bottom to top. The physical layer can provide mobility-related physical signal measurements to be used for mobility management

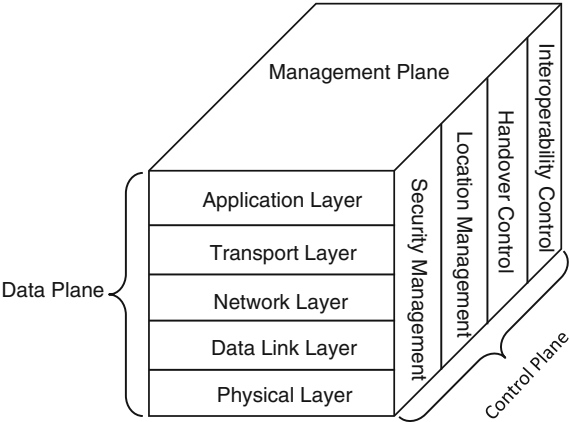


Fig. 2.1 Protocol reference model for mobility management. Reprinted from ref. [1], Copyright 2007, with permission from Journal of communications

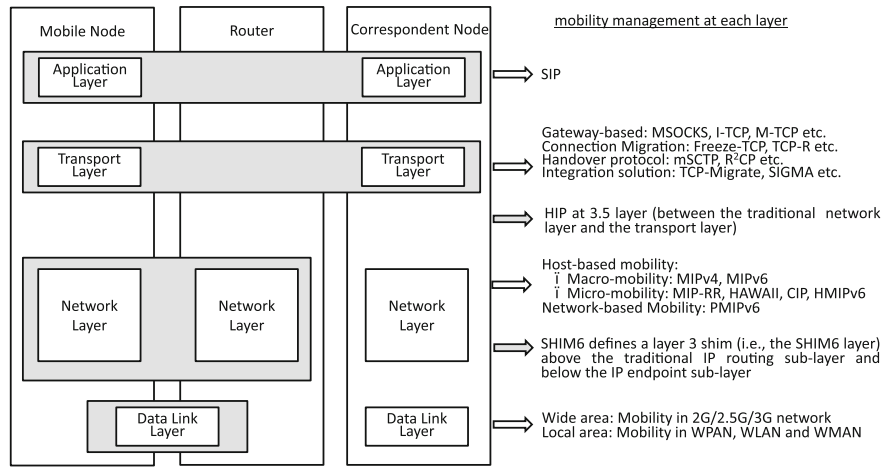


Fig. 2.2 Typical mobility management at the different layers

optimization. Typical mobility management protocols exist at each of the other layers.

Figure 2.2 shows the mobility management protocols and technologies at each layer.

Typical mobility management solutions used at each layer are shown on the right side of Fig. 2.2 and are described for each layer below.

- Mobility management at the data link layer handles the mobility within a subnet area, for example, the mobility in wide area 2G/2.5G/3G cellular mobile communication networks, or mobile access within networks such as the wireless

personal area network (WPAN), the wireless local area network (WLAN) and the wireless metropolitan area network (WMAN).

- Mobility management protocols at the network layer can be classified as host-based mobility supporting or network-based mobility supporting solutions, depending on whether the mobile node participates in mobility management or not. Host-based mobility supporting protocols can be further divided into macro-mobility supporting protocols and micro-mobility supporting protocols.
- The transport layer provides end-to-end mobility support. Mobility management technologies at the transport layer can be classified into gateway-based solutions, connection migration solutions, handover protocols, and integrated solutions.
- The application layer also provides end-to-end mobility support. The Session Initiation Protocol (SIP) is the typical mobility support protocol at the application layer.

These typical protocols and technologies at the different layers will be briefly introduced in Sect. 2.4 and described in more detail in Chaps. 5–8.

Along with the more classical and mature mobility management technologies mentioned above, some new technologies are also proposed based on new technical principles, e.g., ID (identifier)/locator separation solutions with inherent mobility support capability. In general, a permanent identifier associated with dynamically changed locators and the related mapping functions are defined in these solutions. Host Identity Protocol (HIP) [2] and Site Multi-homing by IPv6 Intermediation (SHIM6) [3] are the representative solutions here. HIP introduces a new namespace between the traditional network layer and the transport layer. It defines the host ID to separate the identifier and locator roles of IP addresses. This enables communication continuity when the MN's IP address changes. SHIM6 defines a Layer 3 shim (i.e., the SHIM6 layer) above the traditional IP routing sublayer and below the IP endpoint sublayer. It uses a fixed endpoint identifier above the SHIM6 layer and uses the dynamic IP addresses as locators below the SHIM6 layer. SHIM6 does not define the new namespace in the same way as HIP. Instead, it uses the IP address used in the establishment of the initial session as the fixed endpoint identifier.

For the different mobility types mentioned in Sect. 1.2.2, mobility management technologies at the data link layer and the network layer are often used to support the terminal mobility and network mobility, while the technologies at the transport layer and the application layer often provide personal mobility and service mobility support. Accordingly, terminal mobility and network mobility are called low-level mobility, while personal mobility and service mobility are known as high-level mobility.

Table 2.1 describes the basic functions of each protocol layer in mobility management technologies.

As illustrated in Table 2.1, all of the layers, with the exception of the physical layer, have typical mobility support protocols. Table 2.2 presents a comparison between the typical technologies at the different layers.

Table 2.1 Basic functions of data plane

Protocol layer	Basic functions in mobility management
Physical layer	<ul style="list-style-type: none"> • Provides mobility management-related physical signal detection and measurement, which can be used for function and performance optimization
Data link layer	<ul style="list-style-type: none"> • Provides terminal mobility within an IP subnet • Provides necessary information about link status and L2 (Layer 2) handover starting/finishing event notification, which can be used for function and performance optimization
Network layer	<ul style="list-style-type: none"> • Provides mobility independent of the lower-layer protocols and physical transmission media, and transparent to the upper layers • Mainly supports terminal mobility and network mobility • Provides L3 (Layer 3) handover starting/finishing event notification to the upper layers for handover performance optimization
Transport layer	<ul style="list-style-type: none"> • Provides end-to-end mobility support
Application layer	<ul style="list-style-type: none"> • Provides various types of mobility support, especially for high-level mobility (personal mobility and service mobility)

Table 2.2 Comparison of typical mobility protocols at each layer

Protocol layer	Typical technology	Advantages	Disadvantages
Data link layer	Mobility management technology in cellular communication networks	<ul style="list-style-type: none"> • Supports mobility within an IP subnet, without modifications to network layer, transport layer or application layer 	<ul style="list-style-type: none"> • Mobility is limited to the area of the IP subnet only, i.e., the user could not move from one IP subnet to another without other mobility management technologies
Network layer	MIP	<ul style="list-style-type: none"> • Provides mobility independent of the lower-layer protocols and physical transmission media and is transparent to the upper layers 	<ul style="list-style-type: none"> • High latency incurred by registration procedure if MN is far from a Home Agent (HA); • Scalability problem, i.e., high signaling overhead incurred by frequent registration when number of mobile nodes (MNs) increases; • Requires deployment of HAs in home network, as well as foreign agent (FA) in visiting network for MIPv4; • Requires modifications to network layer of MN;

(continued)

Table 2.2 (continued)

Protocol layer	Typical technology	Advantages	Disadvantages
			<ul style="list-style-type: none">• Unfit for scenario of continuous movement of MN because of performance degradation
Transport layer	mSCTP	<ul style="list-style-type: none">• Follows the end-to-end semantics and does not rely on the support of network infrastructures like routers for mobility support	<ul style="list-style-type: none">• Requires some modification to the transport layer of the MN;• Cannot provide mobility support to the applications based on other transport layer protocols
Application layer	SIP	<ul style="list-style-type: none">• No modifications to transport layer and network layer of MN and correspondent node required;• supports personal mobility and service mobility based on some extensions	<ul style="list-style-type: none">• Requires deployment of application-related servers;• Cannot provide session continuity for other application types;• High handover latency

As shown in Table 2.2, the mobility management technologies at each layer have their own particular advantages and disadvantages. Indeed, no individual technology could satisfy the functional and performance requirements of general mobility. Advanced mobility management technology based on cross-layer principles can integrate the advantages of the different layer technologies and thus provide comprehensive mobility support with good performance. This is the reason why the physical layer is included in the data plane, despite the fact that there are no mobility protocols at this layer. Such cross-layer optimization is also illustrated in Table 2.1. The physical layer, the data link layer and the network layer can all provide mobility-related information and event notifications for mobility management performance, and for handover performance optimization in particular.

2.1.2 Control Plane

The basic functions of the control plane in the mobility management protocol reference model include security mechanism, location management, handover control, and interoperability control.

- Security mechanism handles functions related to Authentication, Authorization, and Accounting (AAA) , as well as the user data and privacy protection in mobility.
- Location management is responsible for storing, updating, and retrieving the location information of the mobile objects.
- Handover control enables session continuity when the access point changes.
- Interoperability control is the particular function in mobility management for a heterogeneous network environment driven by the diversity of the access technologies.

For a more detailed description of these critical control functions in mobility management, please refer to Sect. 2.3.

2.1.3 Management Plane

The management plane in the mobility management protocol reference model is in charge of the essential network management functions. This involves the network management protocols and functions. The traditional network management functions, known as fault, configuration, accounting, performance, and security management (FCAPS), are also included here.

(1) Basic management plane functions

Fault management: the process of monitoring and management of abnormal operations, including maintaining the fault log, locating the problems, isolating the problems, diagnostic testing, and fixing the problems, if possible.

Configuration management: the process of finding and setting up critical configurations on the network devices, collecting, storing, modifying and monitoring the configuration information, and providing configuration information to other related systems. This involves initialization and deletion of managed objects, setting up of suitable configurations for regular operations, and collection of status information.

Accounting management: this function involves tracking each individual's usage and grouping of the network resources to ensure that the users have sufficient resources.

Performance management: this function is useful for system/resource performance evaluation, including: collection and monitoring of performance statistics, maintaining the historic records of the system performance, and measurement of the performance of the network hardware, software, and media.

Security management: this refers to the various protection functions for system security, including maintaining the security log, providing audit trails, and sounding alarms, and distribution of security information to other systems.

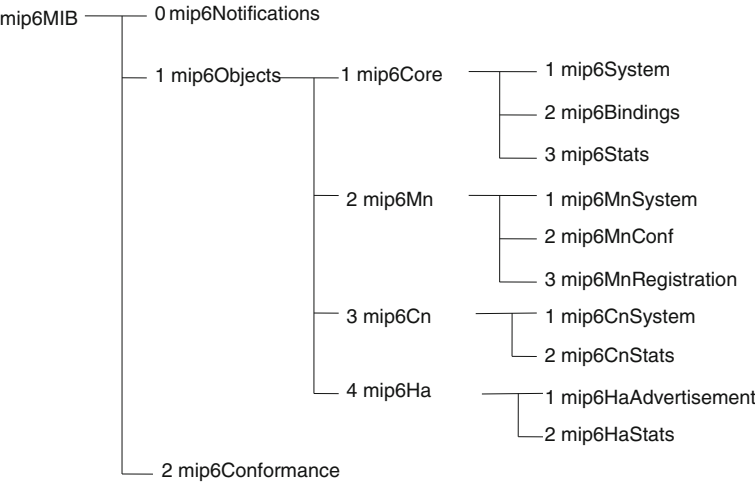


Fig. 2.3 Structural description of MIB definition for MIPv6. Reprinted from IETF RFC4295 [4]

(2) Network management protocol and MIB

Taking Simple Network Management Protocol (SNMP) as an example, it is one of the most commonly used network management protocols. SNMP was defined by the IETF and is expected to realize simple but efficient network management. The management information base (MIB) is the database that records all kinds of information about the managed objects. The MIB extension for MIPv6 is defined by RFC4295 [4], which can be used to monitor and control the MIPv6 entities such as the MN, the HA, and the CN. The detailed monitored information includes functions of MIPv6 entities, MIPv6 service traffic, binding data, and the update history at the MN, CN, and HA.

A structural description of the MIB definition for MIPv6 is defined in [4]. As a part of the whole management information base, the structural description of the mip6Notifications group is shown in Fig. 2.3.

2.2 Network Reference Model

According to the above protocol reference model, the network reference model, and the functional entities of mobility management are abstracted, as shown in Fig. 2.4. Depending on where the mobile object is subscribed, the network is divided into one home network and multiple visiting (foreign) networks. The home network and the visiting networks may then be further divided into subareas (e.g., Location Areas (LAs) and Paging Areas (PAs) in cellular systems). There are four function entities: mobility management servers (including the corresponding databases),

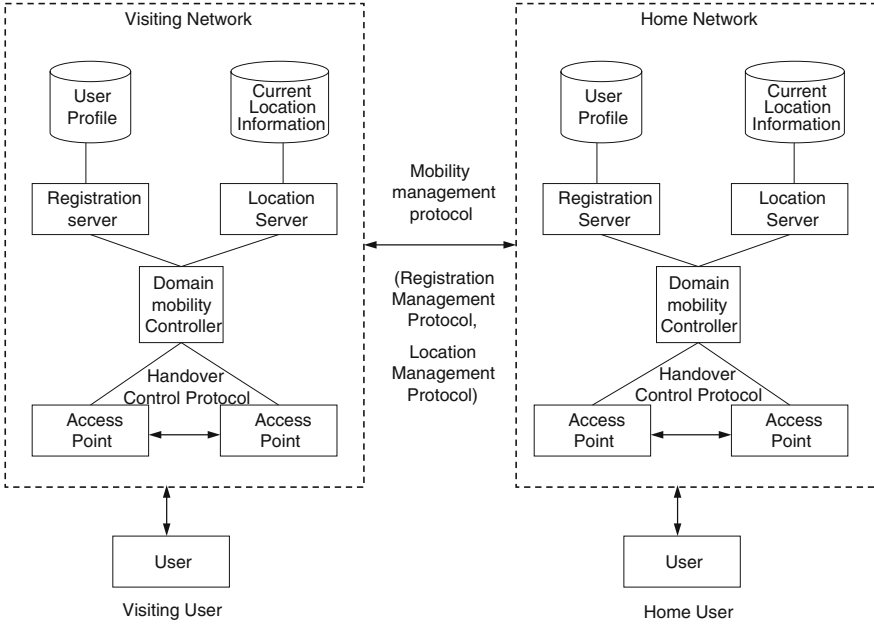


Fig. 2.4 Mobility management network reference model and functional entities. Reprinted from ref. [1], Copyright 2007, with permission from Journal of communications

network access points, the domain mobility controllers, and the corresponding mobility management protocols.

2.2.1 Mobility Management Server

The mobility management servers are responsible for providing the mobility management functions, including the registration servers and the location servers.

The registration servers maintain the users' access authentication information (including network access authentication and service access authentication) and service profile information, but they may be separated into different entities during implementation.

Taking the GSM network as an example, the home location register (HLR) and the authentication center (AuC) cooperate to act as the registration server. HLR records the local user information such as user identity, registration information, roaming capability, subscribed services, and supplementary services. AuC is used to store root key (K_i) and generate a triple [randomly generated number (RAND), signed response (SRES) and the cryptographic key (K_c)] needed to authenticate the user and provide ciphering protection. The HLR stores the triples for each user and transmits them to the Mobile Switching Center (MSC)/Visitor Location Register

(VLR) when required. The MSC/VLR sends RAND to the mobile station, and the mobile station uses RAND, the root key K_i , and the same algorithm as the AuC to compute SRES. The resulting SRES is sent back to MSC/VLR for authentication.

Another example is the Hypertext Transfer Protocol (HTTP)-based authentication mechanism in SIP. It adopts a stateless, challenge-response-based authentication. The challenge and the credentials are exchanged between the user agent client (UAC) and the different SIP servers.

The location servers are responsible for recording, updating, finding, and deregistering the users' current location information. The registration servers and location servers in a visiting network only store the temporary information of any visiting users roaming to their network. When the user roams out of this visiting network, the corresponding information is deleted. The registration server in the home network stores the users' permanent (or semipermanent) information. The location server in the home network also communicates with the location server in the visiting network to update the current user location information.

In MIPv4, the HA and FA behave as the location servers in the home network and the visiting network, respectively. The FA maintains the current location information (i.e., care-of address (CoA)) and the data are routed to the MN accordingly. The HA maintains the home of address (HoA) along with the binding of the HoA with the CoA. A registration request message in MIPv4 is used to update the binding.

2.2.2 Access Point

The access points provide functions including communication access, number/address mapping, and the corresponding handover control and location management related functions. The base station subsystem (BSS) in cellular networks and the AP in WLAN are examples of access points. Handover control will occur between adjacent access points, when the user movement leads to a change in the access points.

The GSM network is used as an example again. Acting as the access point, the BSS also takes part in the handover and location management along with its wireless access function. If the handover target cell is within the same BSS, the handover will be controlled by the base station controller (BSC). Otherwise, the handover request will be sent to the target BSC through the MSC. For location retrieval, the BSC is responsible for paging within all the cells under its coverage. For location updates, when the mobile station moves into the coverage of a BSS in another MSC, it will detect the change in wireless signal quality and the location area code (LAC). The current BSS and MSC will then send a location update request to the new VLR. Subsequent location operations will then be completed by the new VLR and HLR.

2.2.3 Domain Mobility Controller

The network is partitioned into multiple domains for convenience in mobility management. The mobility across different domains is handled by the domain mobility controller. The MSC/gateway MSC (GMSC) in cellular networks, and the HA and FA in MIP are all examples of domain mobility controllers. The domain mobility controller takes responsibility for: (1) handover control between the different access points within a single domain, and across the domains between different visiting networks; (2) the domain gateway (i.e., the exchange of registration and location information between the home network and the visiting network).

The definition of a domain may vary between different mobility management technologies. In GSM networks, the coverage of an individual MSC, or the cell groups under the same BSC, could be defined as a domain. The MSC is in charge of inter-MSC handovers and inter-BSC handovers within the same MSC. The MSC also takes part in the location update and paging functions.

Here, we use the MIPv4 network as another example. HAs and FAs can be considered as the domain mobility controllers of the subnets where they are located. When the MN leaves its home network, the HA is responsible for forwarding the data packets for the MN to the FA through tunnels. After the FA receives the data packets, it depacketizes the data and then sends them to the MN. Thus, the handover procedure is complete. The HA and FA also exchange messages with the MN for location registration.

2.2.4 Mobility Management Protocol

The mobility management protocols are responsible for communications between the functional entities to realize authentication, information exchange (e.g., user information and location information), and control functions. There are four types of protocols: the registration management protocol, the location management protocol, the handover control protocol, and the interoperability protocol. For example, the registration management protocol is responsible for transmission and copying of the service profile data from the home network to the visiting network to guarantee the consistency of the user service environment.

The GSM network is again a good example. In GSM networks, there are typically three types of mobility management protocols: (1) the Mobile Application Part (MAP) of Signaling System No. 7 (SS7), which is used for communications between the MSC, HLR and VLR and is involved in registration management and location management; (2) the Base Station System Application Part (BSSAP), which is used for signaling between the MSC and the BSS and is involved in handover control; and (3) Um interface signaling, which is used for signaling between the mobile station and the BS and is involved in paging and location registration.

Taking the MIPv4 network as another example, its mobility management protocols include agent discovery and Registration Request messages. Agent discovery is realized through agent advertisement messages, which are periodically multi-cast or broadcast by the HAs and FAs. A registration request message is used by an MN to register its new CoA to the HA. The binding of the HoA and the CoA can be set in the HA for location management. For MIPv6, mobility detection, and binding update messages are used for similar functions.

2.3 Critical Control Functions in Mobility Management

2.3.1 Security Mechanism

The threats in mobile environment exist in many aspects such as loss/theft of mobile devices, data interception and tampering, malware, vulnerable applications, compromised devices, Web browser exploration, and OS vulnerability [5, 6]. Security mechanism in mobility management handles the following functions:

(1) Critical data privacy protection

The critical data in mobility management include the user identification information, the user location information, and the signaling messages involved in mobility management. Therefore, two major functions are necessary to provide critical data privacy: (A) To provide the privacy of user identification information and location information; (B) To provide the privacy of signaling messages exchanged between different entities participating in mobility management. The privacy of user data is not included here.

(2) Registration and authentication management

This function is responsible for the management of user registration information and service profile information. AAA is used for: authentication and authorization for network access of the mobile users and mobile terminals; authentication and authorization for service access of the mobile users and mobile terminals; accounting to track the user's utilization of network resources for analysis, audit, and billing. Service profile management handles the registration, update, lookup, and deregistration of the service profile information.

(3) Signaling message integrity

This function is used to provide integrity for the various signaling messages involved in mobility management to keep them from being tampered.

(4) Mobile service non-repudiation

This function is used to keep the mobile objects from fraud of the internal legitimate users in the mobility management system, rather than to prevent threats

from external unknown attackers. Mobile service non-repudiation means both service source and service data receipt are undeniable.

2.3.2 Location Management

In mobility scenarios, the mobile objects (mobile users or mobile terminals) frequently move from one place to another. To track the location of the mobile object for effective mobility support, the location information must be stored in some specific entities (e.g., the location server) in the network and be retrieved when required. As one of the main control functions of mobility management, location management is responsible for storing, updating, and retrieving the locations of the mobile objects.

(1) Functions of location management

There are two important functions in location management: location update (also called location registration) and location retrieval (also called paging).

Location update refers to the event where the mobile object notifies the system of a location change such as the location area (LA) update in 2G cellular systems. When a location update occurs, the system will update the location database and determine the routing for the newly arrived call or data accordingly. The important points in location update are how the mobile object detects the mobility, and when and how to report the object's current location. **Paging** refers to the procedure in which the system finds the locations of mobile objects, with the aim of determining the locations of the objects effectively and precisely, such as the paging function in 2G systems.

Various methods have been proposed for effective location management. In fact, analysis of the extreme conditions in location information storage and location update policy is helpful in understanding the trade-off between location update and paging.

For location information storage, there are two extreme conditions [7]. In condition (a), the location information is stored at each entity in the network, thus making it easy to acquire location information with low paging overhead. However, the location update overhead is extremely high in this case, because every database maintained at each of the different entities must be updated when a mobile object changes its location. In condition (b), the location information is not stored in any of the network entities. Thus, location of a mobile object requires a search of the whole network (e.g., paging through broadcasting). The paging overhead is therefore extremely high, while the location update overhead is very low.

Two extreme conditions also exist for the location update policy [8]. Condition (a) is the "always update" policy, where the location database is updated when any mobile object changes its location, which incurs frequent location update operations and the associated high overhead. However, paging at this time becomes very easy

because the locations recorded in the database always contain the most accurate and latest location information. Condition (b) is the “never update” policy, in which the location database is never updated, even when the mobile object changes its location. This policy incurs an extremely high paging overhead but without any location update overhead.

Based on the above analysis, it is easy to understand the contradiction in the system resource usage of location update and paging. More frequent location updates can improve the paging efficiency but incurs correspondingly greater signaling overhead. Research in location management schemes is aimed at achieving a trade-off between the two.

In general, the various location management policies try to find a balance among availability, precision, and currency [9]. Availability relates to the location information storage policy, ranging from storing location information at every network entity to not storing location information at any of the network entities. Precision relates to the recorded location information content (i.e., whether the location is a possible location or a precise location). Currency relates to when the location information should be updated. For example, for mobile objects with low mobility but with high incoming call probability, the location update need not be performed each time when the object moves.

(2) Databases in location management

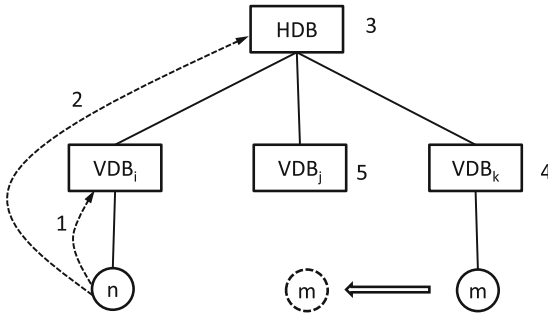
Location management requires a location database. The HLR and the VLR in 2G systems, the HA in the MIP system and the location server in the SIP system are all examples of these databases.

The location database may use several different database structures.

A. Hierarchical database

In the hierarchical database, the location information of the mobile objects is recorded in the databases of different hierarchies. A two-level database structure is normally used, in which two types of databases, the home database (HDB) and the visiting database (VDB), are deployed [7]. Every network is configured with one HDB and multiple VDBs. The HDB retains the registered user-related information (e.g., the user ID, access authority, password, and current location). Each user is associated with one HDB. The VDB holds the related data of the users resident in the associated location area. The data in the VDB are often part of the data in the HDB. The HLR and the VLR in the 2G system, the HA and the FA in MIPv4, and the home location function (HLF) and the visitor location function (VLF) in H.323 could all be considered as typical examples of two-level databases.

The basic location operations of two-level databases can be summarized as follows. (i) location retrieval: if a user in LA i calls the mobile user m , then the location retrieval operation will be conducted first in the VDB of LA i . The operation will then continue in the HDB of user m only if m 's data cannot be found in the VDB of LA i . (ii) Location update: when user m moves from LA k to LA j , m 's location data in his HDB will be updated. At the same time, the m -related data



Location retrieval: 1-2 in this figure describe the operation when user n in LA i calls user m in LA k .
 1: location retrieving firstly in VDB_i ;
 2: location retrieving is continued in HDB of user m when m 's data can't be found in VDB_i ;

Location update: 3-5 in this figure describe the operation when user m moves from LA k to LA j .
 3: m 's location data in his HDB is updated;
 4: m 's location data in VDB_k is deleted;
 5: m 's location data is added in VDB_j .

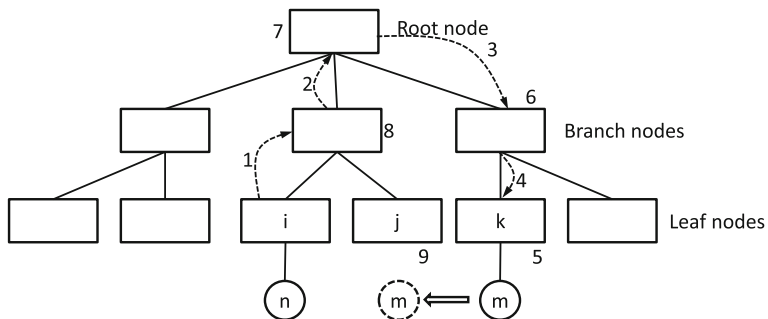
Fig. 2.5 Location operations in a hierarchical database

will be deleted from the VDB of LA k , while being added to the VDB of LA j [7]. Figure 2.5 illustrates the location operations in a hierarchical database.

For a two-level location database, the location of the HDB does not change generally. The location retrieval operation often begins in the HDB. Therefore, for a user that is always resident in distant foreign networks, long-distance communication is required to access their associated HDB whenever they roam from one visiting LA to another adjacent visiting LA. This type of operation cannot use the chain-like characteristics of the multiple adjacent LAs and thus incurs heavy overheads and bandwidth consumption. This is the primary disadvantage of the two-level location database [7].

B. Tree structure database

In the tree structure database, the tree is used to describe the network locations, and the location information of a mobile object is maintained in the corresponding leaf node and branch node. An upended tree is used in the tree structure database. Each leaf node is associated with an individual LA, and a location database is deployed at the leaf node for this LA, recording the location information of all users resident in it. The database is also deployed at the branch node and at the root node to record the location information of users in all the LAs belonging to its subtree. Also, a pointer exists in the database at every non-leaf node, pointing to the database of its child nodes [7].



Location retrieval: 1-4 in this figure describe the operation when user n in LA i calls user m in LA k .

- 1-2: searching m 's location data from i upward to find LCA(i, k);
- 3-4: retrieving operations from the LCA(i, k) downward to k .

Location update: 5-9 in this figure describe the operation when user m moves from LA k to LA j .

- 5-6: m 's location data in related nodes are deleted;
- 7: m 's location data in LCA(k, j) (i.e., the database pointer) is updated;
- 8-9: m 's location data is added in related nodes.

Fig. 2.6 Location operations in the tree structure database

The basic location operations of the tree structure database can be summarized as follows. (i) Location retrieval: if the mobile user in LA i calls mobile user m in LA k , then the location retrieval operation will look for the node containing the location information of user m from node i upward. Obviously, it is simply the least common ancestor (LCA) node of i and k . The querying then goes from this LCA node to node k and user m will be found. (ii) Location update: if the mobile user m moves from LA k to LA j , then the location databases along the path from node k upward to the LCA node of k and j , and then downward to node j will all be updated [7]. Figure 2.6 illustrates the location operations in the tree structure database.

For the scenario mentioned above, in which the user moves between the adjacent LAs frequently, the tree structure database can decrease the communication overhead greatly because it can operate without any need for long-distance communication and only requires access to several adjacent databases. This is the apparent advantage of the tree structure compared with the two-level database. However, the disadvantage of the tree structure database is the large management overhead because of the large number of databases. It also has more stringent performance requirements in terms of capacity and speed for the upper-level nodes [7].

C. Centralized database

In the centralized database, the current location information for all mobile users is maintained in a single, centralized database [7], rather than the two-level database and the tree structure database. Obviously, the advantages of the centralized

database include its simple structure, easy maintenance, simple location operations, and easy data consistency and integrity. However, it also suffers from the common problems of a centralized system, such as scalability and high reliability, stability, capacity, and access speed requirements for the centralized database system.

(3) Analysis of the location management overhead

In location management, the entire network is often partitioned into multiple LAs and paging areas (PAs). The LA is the basic unit for location tracking. When the mobile terminals move within the area of the same LA, it requires no location update operations. However, when a mobile terminal moves across the border of the LA, it should update its location information in the system. In cellular systems, the LA is usually composed of multiple cells. The coverage and shape of the LA could be constant or could change dynamically according to the actual network environments. The PA is the area within which the system searches for the mobile terminal via broadcasting. Depending on the different paging policies, the PA could be limited to a single cell or composed of all the cells in the system.

In future mobile communication systems, tremendous numbers of mobile users will move at random in a large area (even globally) and may communicate at any moment. Therefore, the deployment of the location servers and the planning of the LAs and PAs are very important for fast and precise callee location finding. The planning and optimization of the LAs and PAs are correlative with each other. Depending on the different targets and policies of LAs and PAs, a PA's coverage could be smaller than or equal to that of an LA's. The larger the LA is, the lower the location update frequency becomes, and thus the smaller the update signaling overhead becomes, but the larger that a PA is, the larger the paging signaling overhead becomes.

Performance evaluation parameters for location management include the update signaling overhead, the update delay, the paging signaling overhead, the paging delay, and the system efficiency.

The total cost function for location management can be described as follows:

$$C_{LA} = C_{LU} + C_{PAG}$$

where C_{LA} is the total cost of location management, C_{LU} is the location update cost (including the update signaling overhead and the update delay), and C_{PAG} is the cost of paging (including the paging signaling overhead and the paging delay). The most optimized LA and PA planning schemes aim to minimize this total cost function. In fact, location update and paging are in conflict with each other in terms of system resource consumption, and thus the various location management schemes try to achieve a trade-off between them. Table 2.3 compares the location update and location retrieval and gives an analysis of their overheads.

(4) Location management policies

To achieve the trade-off between the location update and location retrieval, various improved location management schemes have been proposed. According to

Table 2.3 Comparison of location update and location retrieval

Function	Originator	Overhead	Overhead versus LA and PA coverage
Location update	Mobile object	Incurred by update signaling, which consumes the wireless channel resources and increases the load and processing delay of the network and location databases	The larger the LA is, the lower the update frequency becomes, and thus the lower the overhead becomes
Location retrieval	Network	Incurred by paging signaling, which consumes the wireless resources	The smaller the PA is, the lower the paging overhead becomes, and thus the higher the paging success rate becomes with lower paging latency

the update policy of the location database, these location management schemes can be classified into two types: static policy-based schemes and dynamic policy-based schemes. Static policies include the basic policy, the pointer extending policy, the anchor-based policy, the anchor- and pointer-integrated policy, and the circle searching policy. Dynamic policies include the time-based location management policy, the mobility-based location management policy, and the distance-based location management policy. More details of these different policies can be found in reference [7].

Further, most existing mobility management schemes are considered to be passive, because they passively track the location changes of the mobile objects and maintain the network connectivity for them. This passive feature means that these schemes are unlikely to satisfy the requirements of future mobile users. Therefore, some predictive mobility management schemes were proposed. These schemes model the mobility patterns and use prediction algorithms for effective prediction of the possible location of the mobile users. This is the basis of predictive mobility management.

2.3.3 Handover Control

Handover control is used to maintain session continuity for the user when the access point changes during movement (i.e., the ongoing communication provided by the current access point is transferred to another access point rather than being interrupted). As one of the critical control functions and technologies in mobility management, handover control has attracted considerable research effort for both traditional mobile communications and future ubiquitous and heterogeneous network environments.

(1) **Classification of handover**

Depending on different criteria, we can obtain diverse handover types, as shown in Table 2.4.

- (A) Based on the mobility range, handover could be classified into intranetwork handover and internetwork handover. We can see examples of such classification in different networks, such as intra-access service network (intra-ASN) handover and inter-ASN handover in worldwide interoperability for microwave access (WiMAX) networks. Also, in 2G systems, intranetwork handover could be further subdivided into intracell handover, intrabase station controller (BSC) handover, intramobile switching center (MSC) handover, and inter-MSC handover. In IP networks, intranetwork handover can be subdivided into intra-access router (AR) handover, intra-access network (AN) handover, and inter-AN handover.
- (B) Depending on the homogeneity or heterogeneity of the network types involved, handover could be classified into horizontal handover and vertical handover [10, 11]. The difference is that horizontal handover is performed between different wireless access points that use the same technology, while vertical handover involves changing the access technology. For example, handover within the Global System for Mobile Communication (GSM) network is a horizontal handover, but a handover between a Universal Mobile Telecommunication System (UMTS) network and a WLAN is a vertical handover.

Table 2.4 Handover classifications

Classification criteria	Classifications
Mobility range	Intranetwork handover
	Internetwork handover
Network types involved (homogeneous or heterogeneous)	Horizontal handover
	Vertical handover
Handover performance requirements	Fast handover
	Smooth handover
	Seamless handover
Handover purpose	Rescue handover
	Confinement handover
	Traffic handover
Handover procedure	Hard handover
	Soft handover
	Softer handover
Handover necessity	Forced handover
	Unforced handover
User control (permitted or not permitted)	Active handover
	Passive handover

- (C) Based on the handover performance requirements, handover could be classified into fast handover, smooth handover, and seamless handover [12]. Smooth handover aims to minimize the packet loss during the handover procedure. Fast handover aims to minimize the handover delay. Seamless handover requires high performance in both packet loss and handover delay. This is the handover type that means no change in user experience in terms of service capability, security, and quality of service (QoS).
- (D) Based on different purposes, there are rescue handover, confinement handover, and traffic handover [13]. Rescue handover is handover when the mobile station leaves the area covered by one cell and moves into another area. It is the quality of the transmission that determines the handover necessity, where the quality is indicated by the error rate, the received signal strength, the interference level, and the propagation delay. Confinement handover is the handover performed when the mobile station would suffer less interference if it changed cell (interference is caused in part by other mobile stations in the cell). The mobile station continuously listens to other cells to measure the quality of a connection to the latter. Each mobile station is also synchronized with several base transceiver stations (BTSs) to be ready in case of handover. Traffic handover occurs when the number of mobile stations is too large for a cell and the neighboring cells can accommodate new mobile stations. This decision requires knowledge of the charge of other BTSs.
- (E) Depending on the handover procedure (i.e., whether the old link is released before or after the new link is established), we have hard handover, soft handover, and softer handover [14]. This classification originates from the handover technology in traditional cellular systems. A hard handover is one in which the channel in the source cell is released first, and only then is the channel in the target cell engaged. The connection to the source is thus broken before or 'as' the connection to the target is made; for this reason, these handovers are also known as break-before-make. In traditional mobile communications, the handover between base stations (BSs) or sectors using different frequencies could only adopt hard handover. This process could also be used for handover between BSs or sectors belonging to different operators or different systems. A soft handover is the one in which the channel in the source cell is retained and is used for a while in parallel with the channel in the target cell. In this case, the connection to the target cell is established before the connection to the source cell is broken, and thus, this handover is called make-before-break. Soft handover is implemented through the control of the MSCs. It is only used for handover between different BSs using the same frequency. Soft handovers may involve using connections to more than two cells: connections to three, four, or more cells can be maintained by a single mobile station at the same time. When a call is in a soft handover state, the signal that is the best of all used channels can be used for the call at any given moment, or all the signals can be combined to produce a clearer version of the signal. The latter is more advantageous, and when such signal combining is performed in both the downlink (the forward link) and the uplink (the reverse

link), the handover is termed a softer handover. Softer handovers are possible when the cells involved in the handovers have a single cell site. In general mobility management for heterogeneous networks, the hard handover and soft handover are no longer limited to the cellular communication field. They have been extended to various mobility management technologies. For example, soft handover schemes were proposed for MIP and SIP improvement to lower the packet losses during handover for handover performance optimization.

- (F) Depending on the handover necessity, handover could be classified into forced handover and unforced handover [15]. Forced handover is often triggered by the events regarding network interface availability. It occurs when there is only one available interface left. Handover is necessary here to avoid communication interruption. Unforced handover usually happens when multiple interfaces are available simultaneously. The target of unforced handover is to improve the QoS.
- (G) Depending on user control allowance, we also have active handover and passive handover [16]. Active handover allows the user to participate in handover control; the handover is based on the user preference.

(2) **The major technical issues of handover control**

The major technical issues in handover control are the handover rule, the handover control mode, resource allocation during the handover, and communication link transfer.

A. **Handover rule**

The handover rule determines when and under what conditions the handover actions are triggered. In traditional handover for cellular systems, the handover rules are generally determined based on the measured wireless signal strength and the channel quality [17]. For vertical handover across heterogeneous access networks, the handover decision should involve more factors with respect to the user, the application, the network, and the terminal.

Figure 2.7 shows an example of the handover rule in a traditional cellular system—the handover rule for soft handover in a code division multiple access (CDMA) system [17]. It illustrates the variation in the pilot signal strength of a mobile station during soft handover from base station A to base station B. The mobile station measures the pilot signal strength of the adjacent cells while it maintains communication through BS A. When the signal strength of one cell (e.g., BS B in this example) becomes higher than a certain threshold (the upper limit in Fig. 2.7), it will report to the system. The system then tells BS B to establish communication with the mobile terminal and the soft handover is triggered. At this time, the mobile station receives signals from both A and B in parallel. When the mobile station detects that the signal strength of the source cell (BS A in this example) has fallen below another threshold (the lower limit in Fig. 2.7), it starts a timer. If the signal strength of the source cell remains lower than the threshold for a predetermined period, then the mobile station will disconnect from the source cell. The soft handover procedure then terminates.

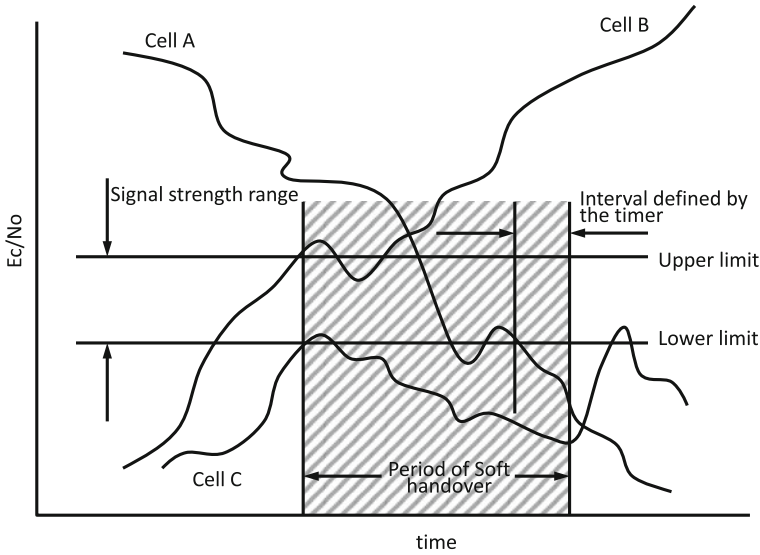


Fig. 2.7 Handover rule for soft handover in CDMA system. Reprinted from ref. [17], Copyright 1999, with permission from the author

The vertical handover decision is a typical multi-criteria decision problem. The decision is made based on various network-related, application-related, and user-related factors. Network-related factors include access technology, coverage range, signal strength, available bandwidth, error rate, actual delay, and other network features and status information. Application factors describe the current application features and requirements. User-related factors include the user's preferences, terminal capabilities, and user mobility characteristics. More details regarding the vertical handover decision will be presented in Sect. 3.1.2.

B. Handover control mode

The handover control mode includes the handover detection and control strategies. It determines who is responsible for handover-related information collection (e.g., measurement of the wireless channel quality), how to collect and report this information, and who initiates the handover.

Based on the roles of the network and the mobile terminals, there are three commonly used handover control modes: network-controlled handover (NCHO), mobile-assisted handover (MAHO), and mobile-controlled handover (MCHO) [18].

In the NCHO mode, the surrounding access points (e.g., the BSs) measure the signal from the mobile station. The network initiates the handover process when certain handover criteria are met. The first-generation analog cellular communication systems, including the Total Access Communication System (TACS) and the Advanced Mobile Phone System (AMPS), used the NCHO mode, where the MSC

controls the handover by following a centralized control mode. NCHO supports intercell handover only.

In the MAHO mode, the network asks the mobile station to measure the signals from the surrounding access points (e.g., the BSs). The network makes the handover decision based on the reports from the mobile station. 2G cellular communication systems such as GSM and CDMA adopted the MAHO mode. Generally, the MAHO mode could support both intercell and intracell handover.

In the MCHO mode, the mobile station continuously monitors the signals of the surrounding access points (e.g., the BSs). The mobile station initiates the handover process when certain handover criteria are met. Systems including WLAN, Digital European Cordless Telecommunications (DECT), and Personal Access Communications System (PACS) use the MCHO mode. The MCHO mode is also capable of supporting intercell and intracell handover.

C. Resource allocation during handover

Resource allocation handles the related resource assignment during handover (e.g., channel assignment in cellular systems, CoA assignment, and IP address binding in MIP systems). Taking the cellular system as an example, the handover call will be terminated (known as forced termination) if there is no available channel on the selected access point. If non-prioritized scheme (NPS) is adopted, a handover call and a new arrival call will be treated in the same way. If there is no available channel, the handover call will be blocked and remains on the original channel until the call is complete or the channel becomes unavailable. Such a scheme may incur forced termination, which is regarded as degradation of the service quality because forced termination is considered to be more serious than new call blocking in terms of the user experience.

Some trade-offs among service quality, spectrum efficiency, and implementation complexity should actually be considered as part of the channel assignment algorithm design [19, 20]. A good algorithm is expected to achieve high spectrum efficiency and low complexity while achieving the required service quality (link quality, new call block rate, and forced termination rate). In this respect, other schemes including reserved channel scheme (RCS), queuing priority scheme (QPS), and subrating scheme (SRS) are considered to be better than NPS.

In RCS, a number of channels in each BS are preserved for handover calls, while the other operations are similar to those of NPS. In other words, the channels are divided into two groups: preserved channels that serve only handover calls, and other channels that serve both new arrival calls and handover calls.

In QPS, a handover call is treated with a higher priority than new calls, i.e., it is easier for a handover call to be allocated a channel. QPS is based on the fact that the adjacent cells in a PCS network overlap each other. Thus, there is a considerable area where a call can be handled by the BS in either of the adjacent cells, which is called the handover area. The time that a mobile station spends in the overlapped area is referred to as the degradation interval. If no channel is available for the handover call after the mobile station moves out of the handover area, i.e., when the

degradation interval expires, the call is forced to terminate. In this scheme, when a channel is released, the BS first checks whether the waiting queue is empty. If not, the released channel is assigned to a handover call in the queue. The next handover to be served is selected based on the queuing policy, for example, a simple first-in first-out (FIFO) policy and a measure-based priority scheme (MBPS) [19].

SRS creates a new channel on a blocked BS for a handover access attempt by subrating an existing call. Subrating is the process of temporarily dividing an occupied full-rate channel into two channels at half the original rate; one serves the existing call while the other serves the handover request [20].

D. Communication link transfer

Communication link transfer is the procedure in which the communication link is transferred from the existing access point to the target access point. The network must bridge the link to the new access point and then drop the link to the old access point.

The link transfer may be made from one channel to another channel on the same access point, for example, intracell handover in the time division multiple access (TDMA) system, or from one access point to another access point attached to the same network control entity, such as inter-BS handover under the same BSC, or inter-BSC handover under the same MSC. The link may also be between access points belonging to different systems (e.g., handover between heterogeneous access networks). Based on the link transfer strategy, we have the hard handover and soft handover classifications.

The handover control function in the cellular mobile communication system is a mature technology. As shown above, it is used as a good example in the introduction of the major functions of the handover process. It should be noted that handover control functions in future heterogeneous network environments are likely to be different from those in cellular systems. Table 2.5 compares the handover control functions in cellular systems with those for heterogeneous networks.

(3) Performance evaluation of handover algorithm

An excellent handover scheme could maximize the system capacity. However, there are several other handover performance evaluation parameters, including the handover success rate, the call drop rate, the new arrival call blocking rate, the average handover frequency, the handover delay, and the forced termination rate.

The handover success rate is the percentage of successful handovers relative to the total number of handover attempts. The call drop rate is the ratio of the number of dropped calls to the total number of successful calls, where “the total number of successful calls” includes both newly initiated calls and handover calls. The new arrival call blocking rate is the ratio of the number of blocked new calls because of wireless resource shortages to the total number of new calls. It should be noted that this parameter does not involve the call blocking incurred for other reasons (e.g., in the shadow area). The average handover frequency is the average handover time throughout the entire session. The handover delay refers to the communication

Table 2.5 Comparison of handover control functions in cellular system with those in heterogeneous networks

Handover control function	Handover in cellular system	Handover in heterogeneous networks
Handover rules	Mainly based on measurements on wireless links; Commonly used measurement methods for link quality: WEI (word error indicator), QI (quality indicator) and RSSI (received signal strength indicator)	Multi-criteria decision problem involving user-related, application-related, network-related, and terminal-related factors
Handover control mode	MCHO, NCHO and MAHO	Commonly used control modes: network initiated, mobile terminal initiated, network and mobile terminal co-initiated
Resource allocation	Mainly refers to wireless channel assignment, including NPS, RCS, QPS, and SRS	For example, CoA assignment and IP address binding in MIP
Communication link transfer	Executed at data link layer	Executed at network layer, transport layer, or application layer

interrupt period caused by the handover. The forced termination rate is the ratio of the calls that were successfully initiated but were then terminated because of handover failure to the total number of calls.

2.3.4 Interoperability Control

Given the diversity of available access technologies and the abundant mobility objects and types, interoperability control is the control function required for mobility management in heterogeneous networks. Here, the interoperability control function involves two aspects: (1) handling the gap between the different access technologies for mobility-related issues; (2) cross-layer methodology in mobility management.

(1) Handling the gaps between different access technologies for mobility-related issues

Users with multiple interfaces (using multi-interface terminals or multiple heterogeneous terminals) while moving across heterogeneous access networks will be the prevalent scenario in the future. In this case, the user is located within the overlapped area covered by heterogeneous access technologies. In such a scenario, the users expect to enjoy their subscribed service seamlessly, and thus, the heterogeneous access networks are required to cooperate with each other to provide the best possible service experience based on their complementary features.

Therefore, interoperability control is used here to handle the gaps between the different access technologies for mobility-related issues.

The best network selection could be considered as an example of interoperability control. A user located within the overlapping area of multiple heterogeneous access networks could connect to the networks via different network interfaces. Selection of the best network for both new services and handover services is an important issue for the best service experience. Here, the best network selection should be the result of a decision based on multiple factors about the networks, users, and applications. The AAA mechanism across heterogeneous networks is another issue in interoperability control. A new AAA mechanism should be designed to be able to adapt flexibly to multi-network or multi-operator environments, to ensure security, and to provide a billing mechanism for intercarrier roaming users. Another inevitable interoperability control issue is QoS adaptation across the heterogeneous access technologies. QoS control mechanisms are defined individually in different wireless access technologies. For example, a cellular system has a strict QoS guarantee mechanism, and wireless broadband access networks like WLAN and WiMAX have also defined their own QoS control schemes. When the user moves across these heterogeneous access technologies, the differences in the QoS control mechanisms and the unpredictable nature of the wireless link quality consequently result in obvious fluctuations in QoS. In this case, a QoS adaptation mechanism is required for seamless mobility support in heterogeneous network environments.

(2) Cross-layer methodology in mobility support

The cross-layer methodology requirements in mobility management stem from two different aspects. On the one hand, mobility management technology at an individual protocol layer could not meet the abundant functionality and performance requirements. Application of cross-layer design is a good choice for integration of these technologies for complementation. On the other hand, the traditional protocol layer design strictly follows the hierarchical layered structure. However, in wireless and mobile network environments, serious performance degradation will be introduced by wireless channel quality instability, scarcity of radio resources and variation in the wireless link features caused by mobility across the different access technologies. A cross-layer design methodology is therefore required for coordination and cooperation between the various technologies at the different layers to achieve better mobility management functionality and performance.

Applications of cross-layer design in mobility management include:

A. Cross-layer integration

Cross-layer integration is based on the complementarity of the various mobility management technologies belonging to the different layers. This cross-layer integration is applied to the network entities and to the signaling design.

Cross-layer integration of the MIP at the network layer and the SIP at the application layer is a good example of this type [21–23]. The complementary capabilities of MIP and SIP can be summarized as the following [21]: (i) The MIP provides transparent mobility to the upper layers and hides the changes in IP address to the application layer of the MN and the correspondent node (CN). The SIP provides the application layer mobility. (ii) The MIP is usually based on TCP, while the SIP is often based on the User Datagram Protocol (UDP); (iii) The MIP is often used to support terminal mobility, while the SIP is used to support high-level mobility such as personal mobility and service mobility.

Accordingly, the integration of MIP and SIP could be used in the design of network entities. Both MIP and SIP require the deployment of some servers in the network, for example, the HA (and the FA in MIPv4) in MIP and the registrar, the location server, the redirect server and the proxy server in SIP. For the integrated scheme, similar functions in different protocols could be integrated at suitable network entities [21]. The MIP/SIP integration could also be used in the design of signaling procedures [21, 23]. Signaling procedures exist for similar functions in MIP and SIP. For example, the MN will execute a location update operation in both MIP and SIP when its location changes. The MN registers its new location information with different entities (HA in MIP and the home registrar in SIP), carrying similar information in the registration messages. This procedure of “moving once, registering twice” brings serious redundancy functionality and excess signaling. This will result in low system efficiency, high packet loss, and handover delay. Therefore, cross-layer integration should be applied to redundant function and signaling optimization.

Another example occurs at the network layer and the transport layer (i.e., the cross-layer integration of mSCTP and MIP). mSCTP provides end-to-end mobility at the transport layer and has a handover control function but does not have a location management function. Integration of mSCTP and MIP would expect to use MIP for location management (i.e., locating the mobile terminal that is being called), and then, the mSCTP association between the two endpoints can be established. After that, mSCTP is used to provide handover for the communication endpoints [24].

B. Cross-layer information interaction

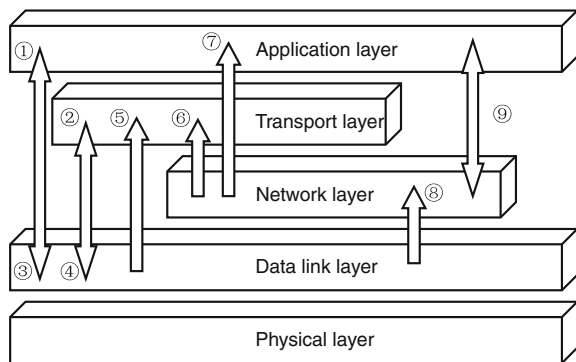
Cross-layer information interaction is often used for mobility management, and especially for handover performance optimization [25]. Figure 2.8 illustrates the possible cross-layer information interactions at the mobile node.

- ①② The link layer conveys the link actions (e.g., frame loss and delay) and link status (e.g., current available bit rate) to the application layer and the transport layer. Then, the application can differentiate the various cases for different handling, thus avoiding the performance degradation that results from these cases simply being treated as congestion.

- ③④ The application layer and the transport layer inform the data link layer about the delay requirements and the tolerable error rate, so that the data link layer can select suitable modulation, error correction, and check mechanisms accordingly.
- ⑤⑥ The data link layer and the network layer notify the TCP at the transport layer about the handover commencement and handover completion events. The TCP can then adjust its congestion control mechanism accordingly, including the congestion window (CWND), the round trip time (RTT), the retransmission timeout (RTO) and the congestion control algorithms (e.g., the slow start algorithm and the congestion avoidance algorithm), for TCP performance optimization.
- ⑦ The application layer adjusts the data sending rate according to the handover notification from the MIP to reduce packet losses during handover and optimize the handover performance.
- ⑧ Data link-layer information is provided for MIP optimization of the mobility detection handover performance.
- ⑨ For cross-layer integration of MIP and SIP, some information exchange and coordination is necessary. For example, IP address acquisition at the new location could be complemented at the network layer and is then notified to the SIP. Such an information exchange process could improve the wireless resource usage.

It should be noted that the trade-off between the cross-layer complexity and resulting performance improvement is a crucial point in cross-layer design for mobility management.

Fig. 2.8 Cross-layer handover optimization at the mobile node



the HLR is used to record the subscriber’s service subscription data and to back up the authority data stored in the AuC. The HLR and VLR form location servers with a two-level database structure. The HLR stores the current subscriber MSC/VLR location information, and the VLR stores the current subscriber LA information. The typical mobile management protocol used is GSM MAP. It is used to exchange messages among MSC, HLR, and VLR and is involved in registration management and location management functions.

According to the mobility management protocol reference model above, from the data plane viewpoint, the cellular mobile network supports mobility primarily in the physical layer (mainly regarding wireless signal measurement) and the link layer. From the control plane viewpoint, mobility management in the cellular network provides the security mechanism, location management, and handover control functions. Table 2.6 summarizes the key points of these control functions. A more detailed description of the mobility management functions in cellular mobile communication networks can be found in Chap. 5.

2.4.2 Network Layer: MIP

MIP, including MIPv4 [26, 27] and MIPv6 [28, 29], supports mobility at the network layer. In the MIP network, an MN, such as a mobile host or a mobile router, can change its network access point, and maintain ongoing communications. Along with the MN, MIPv4 defined two network entities for mobility support: the HA and the FA, as shown in Fig. 2.10.

According to the network reference model, the HA works as a domain mobility controller, a registration server and a location server. The FA is a domain mobility controller. Both the HA and the FA need to maintain related records of MNs, which is similar to the behavior of the HLR and VLR in cellular mobile networks. The

Table 2.6 Mobility management functions in cellular mobile communication networks

Control function	Key points
Security mechanism	<ul style="list-style-type: none">• AAA function is completed through signaling interaction between the SIM (subscriber identity module) card and AuC;• The AAA procedure mainly involves the mobile station (SIM), HLR/AuC, and MSC/VLR. The related information of the subscribers is stored at these network elements
Location management	<ul style="list-style-type: none">• Two-level database composed of HLR and VLR;• Location information is also stored in the SIM card;• Supports location update and paging
Handover control	<ul style="list-style-type: none">• Supports four handover types: intracell handover, intraBSC (RNC) handover, inter-BSC within same MSC/VLR handover and inter-MSC/VLR handover;• Supports hard handover, soft handover, and softer handover

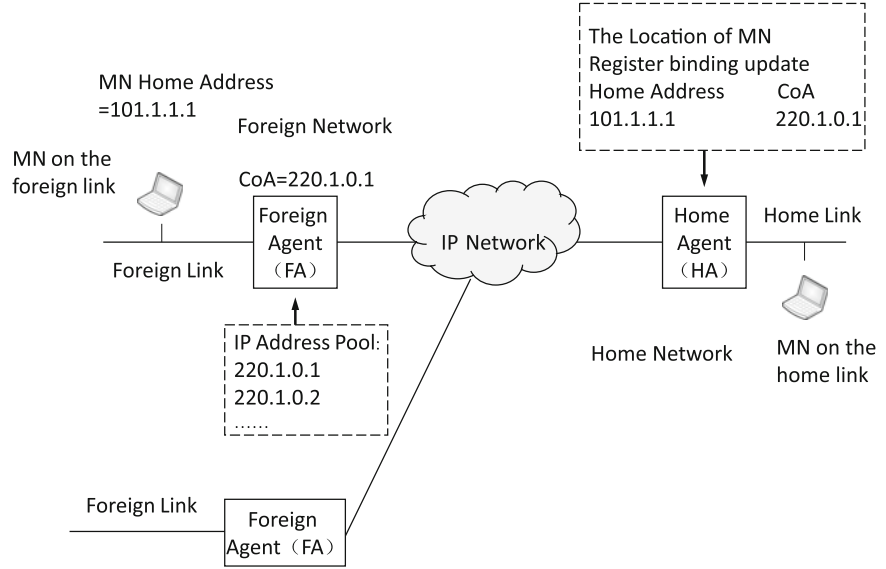


Fig. 2.10 Functional entities in MIPv4

mobility management protocols include agent discovery (an extension based on the Internet Control Message Protocol (ICMP)) and location registration messages.

From the data plane perspective, MIP supports mobility at the network layer. This mobility support is independent of the data link-layer protocols and the physical transmission medium and is transparent to the upper layers and the mobile users.

From the control plane perspective, MIP provides security management, location management, and handover control functions, as shown in Table 2.7. More detailed descriptions of the MIP control functions can be found in Chap. 6.

Table 2.7 Mobility management functions of MIP

Control function	Key points
Security mechanism	<ul style="list-style-type: none">• Provides Security Association (SA) between different network entities, e.g., between MN and HA, between MN and FA, and between FA and HA
Location management	<ul style="list-style-type: none">• HA maintains binding information of MNs' HoA and CoA and records location information of MNs;• Location registration in MIPv4 and binding update in MIPv6 are used for location update functions;• Paging function is not supported in standard MIP, but may be supported in some extension protocols
Handover control	<ul style="list-style-type: none">• Handover occurs when the MN moves across subnets;• Supports fast handover, smooth handover, and seamless handover

2.4.3 Transport Layer: MSCTP

SCTP is a transport-layer protocol defined by the IETF in RFC 2960 [30], which inherits some congestion and traffic control mechanisms from TCP and has some new features, such as multi-homing, multi-streaming, and selective acknowledgment (SACK) to provide a reliable transport service for the upper layers. Mobile SCTP (mSCTP) is an SCTP extension protocol with a dynamic address reconfiguration (DAR) extension [31]. The multi-homing feature, together with the DAR extension, make mSCTP the typical mobility support protocol at the transport layer. Because mobility supported by mSCTP is based on the end-to-end semantics, only the mobility nodes act as mobility entities, while no modifications are needed for any of the network infrastructure devices.

According to the control plane of the protocol reference model, mSCTP has no security mechanism or location management functions and supports only handover control. The key points of the control functions are shown in Table 2.8. More detailed information about mSCTP can be found in Chap. 7.

2.4.4 Application Layer: SIP

Although SIP was initially proposed as an application-layer control protocol for multimedia sessions [32], it has an inherent personal mobility support capability and can be extended to support terminal mobility and service mobility. Therefore, SIP becomes the typical mobility support protocol at the application layer.

Table 2.8 Mobility management functions of mSCTP

Control function	Key points
Security mechanism	<ul style="list-style-type: none"> • Provides critical data privacy, registration, authentication and signaling integrity to a certain extent, based on TLS/SCTP, SCTP/IPSec, and S-SCTP
Location management	<ul style="list-style-type: none"> • No location management function and must be integrated with SIP, MIP, DDNS or RSerPool to support location management functions
Handover control	<ul style="list-style-type: none"> • Multi-homing feature enables the mSCTP endpoint to be configured with multiple IP addresses, one of which will be chosen as the primary address; • DAR extension provides the capability for address addition, address deletion and primary address reconfiguration to be performed dynamically; • The seamless handover procedure can be described as follows: with the movement of the MN, the newly acquired IP address is added as a secondary address, is then set as the primary address used for data transmission, and is deleted when it becomes unavailable

SIP defines the user agent (UA) and various SIP servers (including the proxy server, the redirect server, the location server and the registrar), as shown in Fig. 2.11. The proxy server acts as the domain mobility controller. The functions of the location server and the registrar are obvious. The functions of the mobility management are included in the SIP extensions.

According to the protocol reference model, the SIP is located at the application layer in the data plane. It has security mechanism, location management, and very limited handover control functions from the control plane viewpoint, as listed in Table 2.9. A more detailed description of the SIP control functions can be found in Chap. 8.

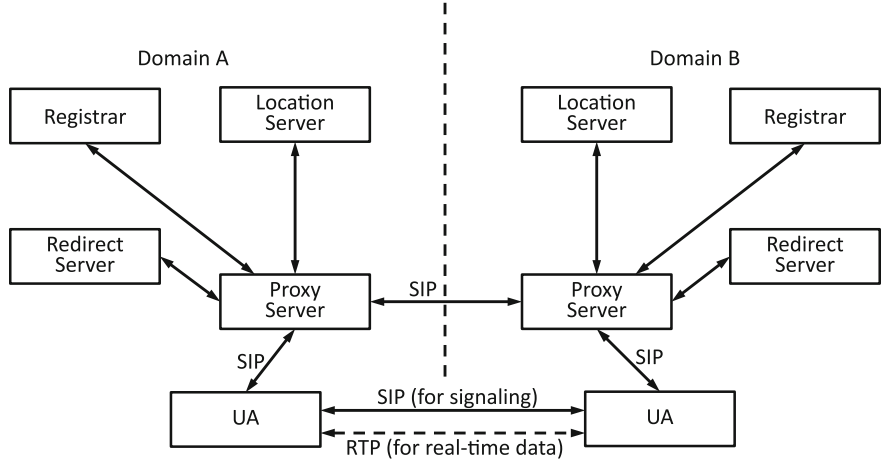


Fig. 2.11 Architecture of the SIP system

Table 2.9 Mobility management functions of SIP

Control function	Key points
Security mechanism	<ul style="list-style-type: none">• Registrar is responsible for registration and authentication;• The mobile terminals use the information carried in REGISTER messages for registration and authentication;• HTTP digest authentication method is adopted;• Transport-layer security (TLS) is used for security transmission
Location management	<ul style="list-style-type: none">• User location information is updated through location registration;• Location server is responsible for maintaining user-related location information;• Redirect server and proxy server are responsible for finding the current location of the user based on interaction with the location server;• No paging support in the basic protocol, but paging can be realized based on cross-layer integration of SIP and L2 paging;
Handover control	<ul style="list-style-type: none">• Supports handover through re-INVITE message;• Handover function is rather limited, only supporting single-side handover based on UDP

References

1. Chen S, Shi Y, Hu B (2007) Mobility management theory and technology. *J Commun* 28 (10):123–133
2. IETF RFC5201 (2008) Host identify protocol
3. IETF RFC5533 (2009) Shim6: level 3 multihoming shim protocol for IPv6
4. IETF RFC4295 (2006) Mobile IPv6 management information base
5. Jain AK, Shanbhag D (2012) Addressing security and privacy risks in mobile applications. *IT Professional* 14(5):28–33. doi: [10.1109/MITP.2012.72](https://doi.org/10.1109/MITP.2012.72)
6. La Polla M, Martinelli F, Sgandurra D (2013) A survey on security for mobile devices. *IEEE Commun Surv Tutor* 15(1):446–471. doi:[10.1109/SURV.2012.013012.00028](https://doi.org/10.1109/SURV.2012.013012.00028)
7. Yihua Zhu (2005) Mobility management in wireless mobile networks. Posts & Telecom Press, Beijing
8. Bar-Noy A, Kessler I (1993) Tracking mobile users in wireless communications networks. In: *Proceedings of IEEE INFOCOM*, pp 1232–1239
9. Pitoura E, Samaras G (2001) Locating objects in mobile computing. *IEEE Trans Knowl Data Eng* 13(4):571–592. doi:[10.1109/69.940733](https://doi.org/10.1109/69.940733)
10. Chakravorty R, Vidales P, Patnapongpibul L, Subramanian K, Pratt I, Crowcroft J (2003) On inter-network handover performance using mobile IPv6. University of Cambridge Computer Laboratory, Technical Report. <http://www.cl.cam.ac.uk/coms/publications.htm>. Accessed 28 March 2014
11. Siddiqui F, Zeadally S (2006) Mobility management across hybrid wireless networks: trends and challenges. *Comput Commun* 29(9):1363–1385
12. IETF RFC3753 (2004) Mobility related terminology
13. Mouly M, Pautet MB, Foreword By-Haug T (1992) *The GSM system for mobile communications*. Telecom Publishing, Washington, DC
14. UMTS (Universal Mobile Telecommunications System) handover. <http://www.umtsworld.com/technology/handover.htm>. Accessed 28 March 2014
15. Xie J, Wang X (2008) A survey of mobility management in hybrid wireless mesh networks. *IEEE Netw* 22(6):34–40. doi:[10.1109/MNET.2008.4694172](https://doi.org/10.1109/MNET.2008.4694172)
16. Chen Y, Kowalik K, Davis M (2009) MeshScan: performance of passive handoff and active handoff. In: *IEEE international conference on wireless communications and signal processing, WCSP 2009*, pp 1–5. doi: [10.1109/WCSP.2009.5371523](https://doi.org/10.1109/WCSP.2009.5371523)
17. Derong Chen, Jiaru Lin (1999) *Digital mobile communication systems*. Beijing University of Posts and Telecommunications Press, Beijing
18. Ekiz N, Salih T, Küçüköner S et al. (2006) An overview of handoff techniques in cellular networks. *Int J Inf Technol* 2(2)
19. Kaur D, Kumar N (2013) Performance analysis of handoff in CDMA cellular system. *Int J Comput Technol* 9(3):1119–1126
20. <http://wireless.cs.tku.edu.tw/~cychang/2G.pdf>. Accessed 28 Mar 2014
21. Wang Q, Abu-Rgheff MA, Akram A (2004) Design and evaluation of an integrated mobile IP and SIP framework for advanced handoff management. *IEEE Int Conf Commun* 7:3921–3925. doi:[10.1109/ICC.2004.1313287](https://doi.org/10.1109/ICC.2004.1313287)
22. Le L, Li G (2007) Cross-layer mobility management based on mobile IP and SIP in IMS. In: *IEEE international conference on wireless communications, Networking and mobile computing, WiCom 2007*, pp 803–806. doi: [10.1109/WICOM.2007.207](https://doi.org/10.1109/WICOM.2007.207)
23. Prior R, Sargento S (2007) SIP and MIPv6: cross-layer mobility. In: *12th IEEE symposium on computers and communications, ISCC 2007*, pp 311–318. doi: [10.1109/ISCC.2007.4381552](https://doi.org/10.1109/ISCC.2007.4381552)
24. IETF draft-sjkoh-msctp-01 (2005) Mobile SCTP (mSCTP) for IP handover support
25. Wang Q, Abu-Rgheff MA (2003) A multi-layer mobility management architecture using cross-layer signalling interactions. In: *5th European personal mobile communications conference*. doi: [10.1049/cp:20030253](https://doi.org/10.1049/cp:20030253)
26. IETF RFC5944 (2010) IP Mobility support for IPv4, Revised

27. IETF RFC3344 (2002) IP Mobility support for IPv4
28. IETF RFC6275 (2011) Mobility support in IPv6
29. IETF RFC3775 (2004) Mobility support in IPv6
30. IETF RFC2960 (2000) Stream control transmission protocol
31. IETF draft-ietf-tsvwg-addip-sctp-08 (2004) Stream control transmission protocol (sctp) dynamic address reconfiguration
32. IETF RFC 3261 (2002) SIP: session initiation protocol

<http://www.springer.com/978-3-662-52724-5>

Mobility Management

Principle, Technology and Applications

Chen, S.; Shi, Y.; Hu, B.; Ai, M.

2016, XXXI, 310 p. 111 illus., 17 illus. in color.,

Hardcover

ISBN: 978-3-662-52724-5