Abstract

This article first surveys existing protocols for supporting IP mobility and then proposes an extension to the Mobile IP architecture, called TeleMIP. Our architecture attempts to achieve smaller handoff latency by localizing the scope of most location update messages within an administrative domain or a geographical region. TeleMIP is intended for use in evolving third-generation wireless networks, and introduces a new logical entity, called the mobility agent, which provides a mobile node with a stable point of attachment in a foreign network. While the MA is functionally similar to conventional foreign agents, it is located at a higher level in the network hierarchy than the subnet-specific FAs. Location updates for intradomain mobility are localized only up to the MA; transmission of global location updates are necessary only when the mobile changes administrative domains and/or geographical regions. By permitting the use of private or locally scoped addresses for handling intradomain mobility, TeleMIP allows efficient use of public address space. Also, by reducing the frequency of global update messages, our architecture overcomes several drawbacks of existing protocols, such as large latencies in location updates, higher likelihood of loss of binding update messages, and loss of inflight packets, and thus provides better mobility support for real-time services and applications. The dynamic creation of mobility agents (in TeleMIP) permits the use of load balancing schemes for the efficient management of network resources.

TeleMIP: Telecommunications-Enhanced Mobile IP Architecture for Fast Intradomain Mobility

SUBIR DAS, ARCHAN MISRA, AND PRATHIMA AGRAWAL,
TELCORDIA TECHNOLOGIES
SAJAL K. DAS, UNIVERSITY OF TEXAS AT ARLINGTON

The rapid growth of wireless networks and services, fueled by next-generation mobile communications systems research, has ushered in the era of ubiquitous computing. Lightweight portable computers, IP-based (office and home) appliances, and the popularity of the Internet are providing strong incentives to service providers to support seamless user mobility. Realizing commercially viable IP mobility support over the current cellular infrastructure, however, remains a research challenge. In particular, for real-time multimedia (audio, video, and text) communications, user mobility poses several challenges.

Wireless access to telecommunications services has traditionally been provided through wide-area cellular systems, which in turn are connected to the public telecommunications network backbones such as the public switched telephone network (PSTN). It is expected that future wireless communications systems will be more heterogeneous and that every mobile user will be able to gain access to the Internet backbone by attaching his or her computer to a wireless access point (Fig. 1). A telecommunications architecture that supports IP mobility will enable service providers to offer high-quality broadband multimedia services to mobile users in a cost-effective way [1]. Although neither the Internet nor the telecommunication networks are currently designed to support high-bandwidth real-time multimedia services, a series of new technologies for third-generation (3G) wireless systems are being developed to make things a reality. These technologies include International Mobile Telecommunications (IMT-2000), Universal Mobile Telecommunications System (UMTS), General Packet Radio Service (GPRS), Enhanced Data Rate for GSM Evolution (EDGE), code-division multiple access (CDMA-2000), and wideband CDMA (WCDMA).

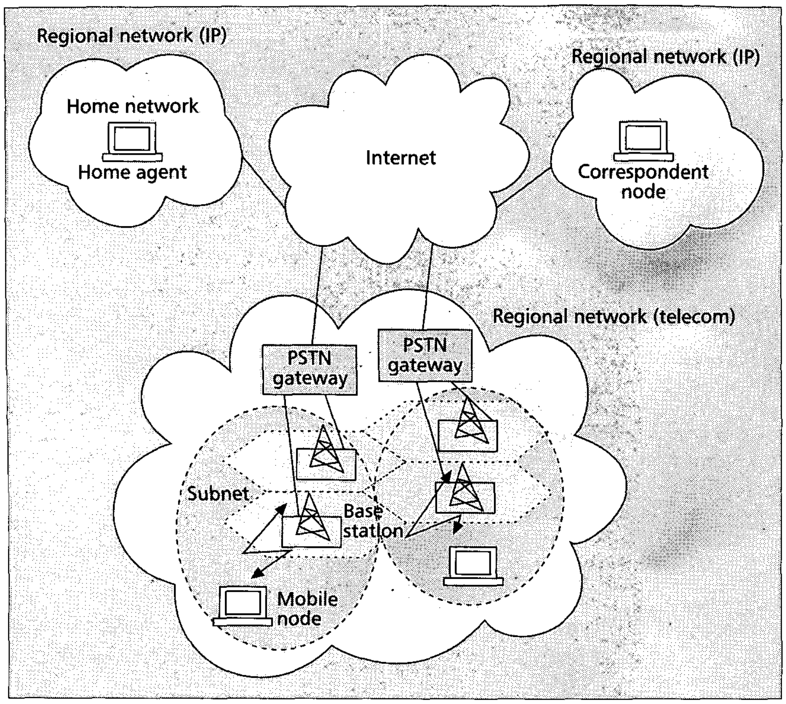
Cellular Mobility

Mobility management in cellular networks is achieved in a different way than in IP-based networks. More precisely, mobility management enables telecommunication networks to locate roaming terminals for call delivery and to maintain connections as the terminal moves into a new service area [2, 3]. It consists of two components:

- *Location management*, including *update* and *paging*, enables the network to discover the current attachment point of the mobile user for call delivery.
- *Handoff 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. This is performed in three steps: *initiation*, *connection generation*, and *data flow control*.

There are two kinds of handoff: *intracell* handoff and *intercell* handoff. Intracell handoff occurs when the user moves within a service area or cell, and changes radio channels under the same base station (BS). On the other hand, intercell handoff occurs when the user moves into an adjacent service area or cell for which all of the mobile's connections are transferred to a new BS. If the mobile node (MN) connects to multiple BSs simultaneously (e.g., in CDMA) while performing the handoff, it is called a *soft* as opposed to a *hard* handoff [3].

1070-9916/00/$10.00 © 2000 IEEE



■ **Figure 1.** *An internetworking scenario.*

Wireless Data Services

Cellular systems now work in the circuit-switching mode and have been designed mainly for voice communications. Data networking and multimedia services necessitate the use of packetized transmission. It is expected that wireless multimedia services will be available in the near future due to an exciting forthcoming data service for Global System for Mobile Communications (GSM) networks, GPRS. It refers to a high-speed packet data technology with a data transmission speed of 144 kb/s that supports Internet communication protocols, such as IP and X.25 [4]. Since GPRS is packet-based, it uses radio resources only when data is being sent or received, and hence multiple users can share the same radio channel very efficiently. To implement GPRS, network operators need to install new hardware, including a packet-based mobile switching center (MSC) called a servicing GPRS service node (SGSN), along with its visiting location register (VLR) and other platforms. The SGSN is the node within the GSM infrastructure that sends and receives data to and from the mobile stations. The SGSN communicates with the Gateway GPRS Support Node (GGSN) to maintain connections with other networks such as the Internet, X.25 networks or private networks (Fig. 2).

EDGE is another high-speed mobile data standard which is effectively enhanced channel coding for GPRS. This allows transmission speeds of up to 384 kb/s. It is expected that EDGE will provide a migration path from GPRS to UMTS.

Mobility Classifications

Third-generation wireless networks such as IMT-2000 will provide terminal mobility, personal mobility, and service provider portability [5]. 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 [6].

User mobility in a cellular architecture that supports IP mobility can be broadly classified into three categories:

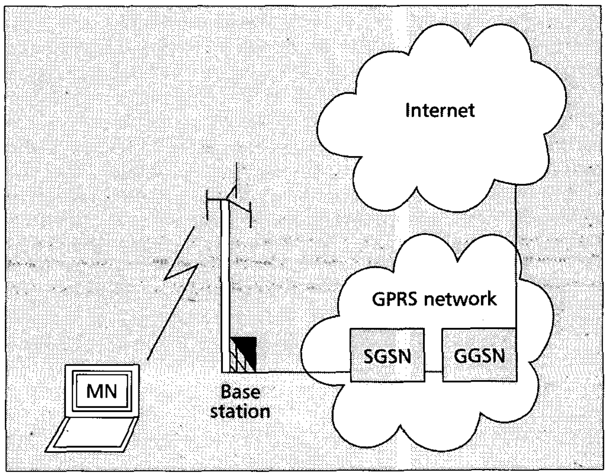
- **Micro-mobility** is the movement of an MN within or across different BSs within a subnet and occurs very rapidly. Currently, management of micro-mobility is accomplished using link-layer support (layer 2 protocol).
- **Macro-mobility** (or **intradomain mobility**) is the movement of an MN across different subnets within a single domain or region, and occurs relatively less frequently. This is currently handled by Internet mobility protocols (layer 3) such as Mobile IP. This will be the focus of the TeleMIP architecture proposed in this article, where we will use the terms macro-mobility and intradomain mobility interchangeably.
- **Global mobility** (or **interdomain mobility**) is the movement of an MN among different administrative domains or geographical regions. At present, this is also handled by layer 3 techniques such as Mobile IP.

In general, the goal of mobility management is to ensure continuous and seamless connectivity during micro- and macro-mobility, which occur over relatively short timescales. Global mobility, on the other hand, usually involves longer timescales — the goal there is often to ensure that mobile users can reestablish communication after a move rather than to provide continuous connectivity.

Our Motivations and Contributions

Several frameworks have been proposed recently to support seamless network access to mobile users [7–16]. Among these proposals, Mobile IP [7] and Session Initiation Protocol (SIP) [16] are currently standardized by the Internet Engineering Task Force (IETF). Mobile IP supports application-transparent IP mobility, while SIP provides an application-layer signaling protocol for creating, modifying, and terminating sessions with one or more participants in both wireline and wireless networks.

As discussed later, the basic Mobile IP protocol was designed to provide a near-term solution for MNs without requiring protocol upgrades in stationary correspondent nodes (CNs) and routers. However, it does not consider the integration of additional functions such as authentication and billing, which are critical for successful adoption in commercial networks. Implementations of Mobile IP are available from vari-



■ **Figure 2.** *A network with a GPRS system.*

ous commercial and research organizations (e.g., MosquitoNet [17], SUN Microsystem [18], University of Singapore [19], Helsinki University of Technology [20]).

On the other hand, SIP [16] is an application-layer control (signaling) protocol that can establish, modify, and terminate multimedia sessions or calls. Recently an architecture was proposed for SIP mobility support [21] to avoid certain problems with Mobile IP. However, SIP mobility cannot support TCP connections and is also not suitable for micro- or macro-mobility. For long-lived TCP connections, the authors in [21] suggested using Mobile IP similar to another approach in [22]. Therefore, we will not further discuss or compare SIP mobility with other IP-based mobility management protocols.

Although the Mobile IP solution meets the goals of operational transparency and handoff support, it is not optimized for managing macro-mobility (or intradomain mobility) in commercial cellular networks. In particular, we shall see that a larger number of location update messages and the latency involved in communicating these update messages to remote nodes make it unsuitable for supporting real-time applications on the Internet. It has thus become necessary to modify the basic Mobile IP architecture to obtain a more scalable solution that is consistent with the evolving cellular architecture and that also supports uninterrupted operation of real-time applications. This has led to the development of protocols like HAWAII [12] and Cellular IP [14]. However, through a comprehensive survey and detailed comparisons in the next section, we will demonstrate that all of these solutions have limitations while dealing with intradomain or macro-mobility in the telecommunications world, and none is best suited for all services and applications. In fact, there are several issues and open problems which need further investigation in order to achieve interoperability (e.g., IP with cellular mobility). Therefore, the search for better protocols to support IP mobility continues, and motivates our work.

In this article, we present an architecture, Telecommunication Enhanced Mobile IP (*TeleMIP*), that supports fast handoffs by localizing the scope of most location update messages within an administrative domain or a geographical region. The proposed architecture is intended for use in evolving 3G wireless networks, and introduces a new logical entity, called the *mobility agent* (MA), which provides an MN with a stable point of attachment in a foreign network. While the MA is functionally similar to conventional foreign agents (FAs), it is located at a higher level in the network hierarchy than subnet-specific FAs. This scheme supports fast subnet handoff and real-time tracking within a domain since most location updates are transported only up to the MA. Mobility updates to the home agent (HA) and/or CN are necessary only when the mobile changes administrative domains and/or geographical regions. By restricting the scope of most location updates, we can lower the large latencies in location updates, the likelihood of losing such messages, and the extent of loss of inflight packets. Our mobility management scheme is especially oriented toward supporting uninterrupted real-time applications. Since address space availability was perceived to be a significant limitation on the scalability of the existing solutions, we have tried to ensure that most care-of addresses in our TeleMIP architecture have private or local scope and hence need not deplete the global IPv4 address space.

The rest of the article is organized as follows. We present a comprehensive survey of existing mobility protocols, their operations and limitations. We describe our proposed mobility architecture and advantages over existing schemes, and present our conclusions.

Existing Solutions to IP Mobility

This is an overview of the basic Mobile IP architecture followed by its extensions for emerging 3G cellular networks.

Mobile IP

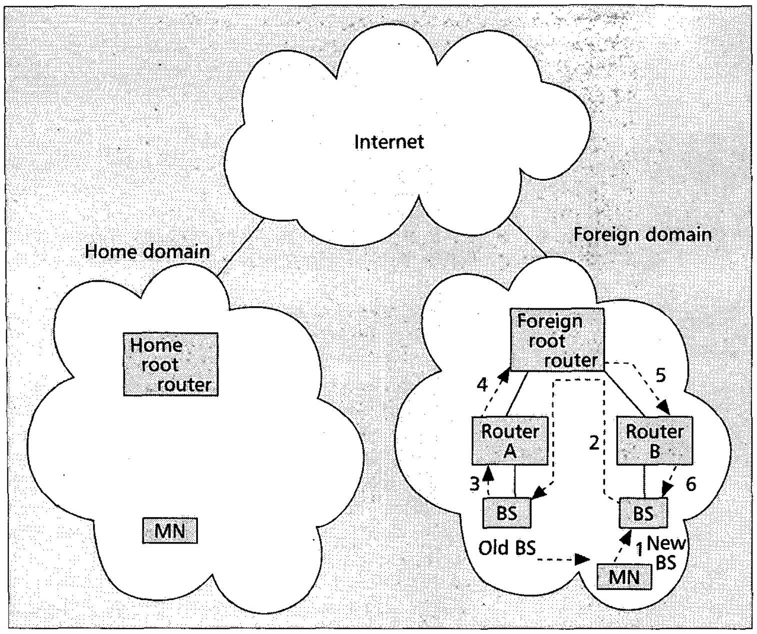
Mobile IP [7] provides an IP-based mobility solution that allows MNs to maintain network connectivity while retaining their permanently assigned IP addresses. In particular, it enables the mobility of a user to be transparent to all executing applications. This is essentially achieved by providing the mobile with an address (in addition to its permanent address) that is topologically consistent. This address is referred to in the foreign network as the *care-of address*, and ensures that packets are forwarded using conventional IP routing to the mobile's current location in the foreign network. The basic Mobile IP specification allows for two distinct methods of operation:

- The first mode of operation uses an FA while visiting the foreign network (a network other than the MN's home network). The FA provides the mobile with a binding (IP address) that is consistent with the addressing scheme deployed in the foreign network. An MN can connect to the foreign network by registering the IP address of the FA with its HA, statically assigned to the MN in its home network.
- The second mode of operation does not require any agent support in the foreign network but requires MNs to obtain a temporary IP address therein. The MN usually obtains this address from a specified pool using protocols such as Dynamic Host Configuration Protocol (DHCP) [23], and then uses its own collocated care-of address in the foreign network.

The collocated address mechanism [7] allows the MN to have direct control over the path of its own packets, and also does not rely on the existence of additional agents in the foreign network. While this may currently seem to be an advantage, we shall shortly argue that an agent-reliant mobility management scheme may be more advantageous in an integrated commercial telecommunications infrastructure.

The basic operational mode of Mobile IP architecture gives rise to the phenomenon of *triangular routing*: while packets from the MN usually follow a direct path to the CNs, packets from the CNs are rerouted via the MN's home network to its point of attachment in a foreign network, from where they are forwarded to the MN's current location. Enhancements have been suggested [24] to avoid triangular routing by essentially transmitting binding messages directly to CNs. While this form of route optimization will result in significant bandwidth savings by eliminating unnecessary path traversals, especially as the number of MNs increases, we shall see later that it can give rise to significantly high latency during the location update process. However, we believe that route optimization, at least in some modified form, is essential for supporting real-time communications in any future mobility-enhanced network infrastructure.

Another approach to IP mobility support has been proposed in the new version of the Internet Protocol, namely IPv6 [10], which supports mobility management as an integral part of the protocol standards and does not require the presence of special agents in foreign networks. Location updates are directly transmitted from the mobiles to the CNs using binding update messages (a mobile maintains a list of current CNs), thereby reducing the role of the HA in the mobile communication process. We shall see, however, that the transmission of binding updates directly to CNs can also result in a large update latency and can become a critical impediment to successful support of real-time applications. Moreover, the deployment of IPv6 infrastructure is still a futuristic goal and



■ **Figure 3.** *Mobility using HAWAII and the corresponding path setup.*

is likely to involve a prolonged period of coexistence with the current IPv4 infrastructure. Investigation and improvement of IPv4-based mobility mechanisms thus continues to be an area of practical concern.

Macro-Mobility Extensions to Mobile IP

In recent years, various solutions have been proposed to handle other problems related to Mobile-IP-based mobility management, such as firewall traversal [25], or reverse or bidirectional tunnelling [26]. However, these enhancements are still not particularly suitable for supporting intradomain mobility in cellular wireless networks. They lack support for fast handoff control, real-time location update, registration, and configuration. Moreover, the importance of application-transparent mobility has diminished in present scenarios since many applications (e.g., Web browsing) are now able to internally handle network-level mobility.

An extension to Mobile IP has been proposed in [27, 28], which uses hierarchical FAs to handle intradomain or macro-mobility. In this architecture, BSs are assumed to be network routers; hence, it is not compatible with current cellular architectures, in which BSs are simply layer 2 forwarding agents. Moreover, deploying a hierarchy of FAs brings with it complex operational and security issues (especially in a commercial multiprovider environment) and requires multiple layers of packet processing on the data transport path. The presence of multiple layers of mobility-supporting agents also significantly increases the possibility of communication failure, since it does not exploit the inherent robustness of Internet routing protocols.

Several IETF proposals [29, 30] have also explored the possibility of using hierarchical FAs for seamless mobility within a domain, but have not been actively pursued in the recent past. The need for hierarchical agents in an Internet mobility architecture remains an open issue. While it does not appear to be a critical consideration in the immediate future, it is possible that hierarchical mobility management will become more attractive as the IP security infrastructure matures and deployment of mobile multimedia terminals gets much larger.

A draft on Mobile IP regional tunnel management [31] was recently proposed in the IETF. The proposal provides a scheme for performing registrations locally in the visited (foreign) domain, thereby reducing the number of signaling messages forwarded to the home network as well as lowering the signaling latency that occurs when an MN moves from one FA to another. The draft addresses one of the important drawbacks of conventional Mobile IP. The suggested enhancement to the registration scheme uses a gateway FA (GFA), which lies one level higher in the FA hierarchy, provides a stabler global care-of address to the MN, and is very similar to our suggested modifications for supporting fast intradomain mobility management (albeit with some differences). Unlike the approach in [31], our solution is not merely a protocol but a more comprehensive architectural framework for supporting intradomain mobility in cellular wireless networks.

Given the concerns of enabling IP-based mobility in a commercial environment, a few other protocols have been proposed to extend Mobile IP to better support micro- and macro-mobility in next-generation cellular environments, as discussed below.

HAWAII

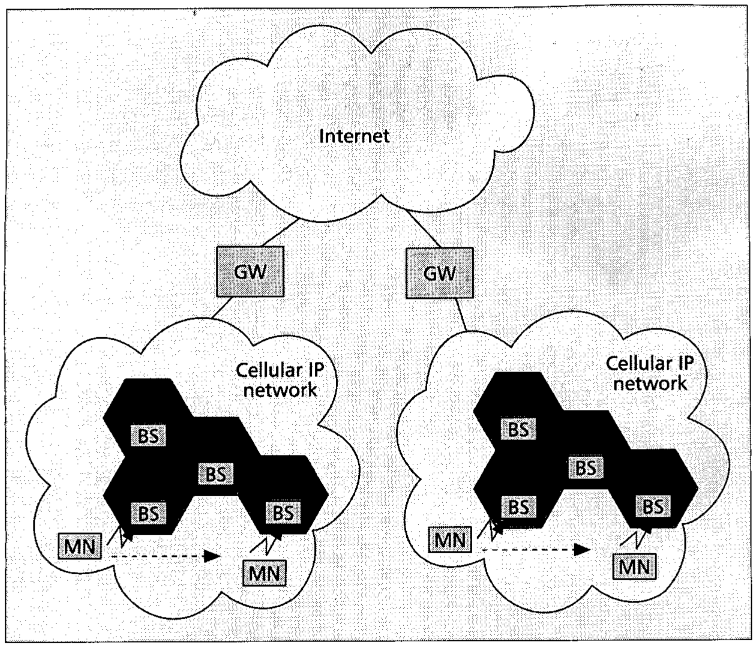
The Handoff-Aware Wireless Access Internet Infrastructure (HAWAII) [12] proposes a technique for using a separate binding protocol to handle intradomain mobility (i.e., micro- and macro-mobility according to our definition) while using Mobile IP for interdomain mobility. It suggests the use of a two-layer hierarchy for mobility management. When the MN moves into a foreign domain, it is assigned a collocated care-of address from that domain, and the MN retains its care-of address unchanged while moving within the foreign domain. Thus, the movement of the MN within a domain is transparent to the HA. This protocol uses path setup messages to establish and update host-based routing entries for MNs in some specific routers within the domain; other routers not in the path are kept in the dark about the MN's new care-of address.

When a CN sends packets to a roaming user, it uses the MN's home IP address. The HA intercepts the packets and sends the encapsulated packet to the MN's current border router. The border or root router decapsulates and again encapsulates the packet to forward it to either the intermediate router or BS, which decapsulates the packet and finally delivers it to the MN (Fig. 3).

Cellular IP

Cellular IP [13, 14] proposes an alternative method to support local mobility (again, micro- and macro-mobility according to our definition) in a cellular network, which consists of interconnected cellular IP nodes. This protocol uses Mobile IP for wide-area mobility. It has many similarities with the host-based routing paradigm of HAWAII. In particular, Cellular IP is designed to support local mobility, say, between BSs in a cellular network (Fig. 4). Since MN addresses have no location significance inside a cellular IP network, the architecture uses the home IP address as a unique host identifier. When an MN enters a Cellular IP network, it communicates the local gateway's (GW's) address to its HA as the care-of address.

Nodes outside the Cellular IP network do not need any enhancements to communicate with nodes inside the network. When a CN sends packets to a roaming user, it uses the MN's home IP address. As in conventional Mobile IP, the HA intercepts the packets and sends the encapsulated packet to the



■ **Figure 4.** *Mobility using cellular IP.*

MN's current GW. The GW decapsulates the packet and forwards it to the MN's home address using a node-specific route. Thus, the nodes sending or receiving datagrams to/from the MN remain unaware of the node's location inside the Cellular IP network.

Wireless IP Network Architecture by TR45.6

Another framework for IP-based mobility management was recently developed by the Telecommunications Industry Association (TIA) Standards Subcommitte TR45.6 [15] to target 3G cellular wireless systems. The requirements have been set by the International Telecommunication Union (ITU) for IMT-2000. The framework uses Mobile IP with FAs for interdomain or global mobility. For intradomain or macro-mobility, the scheme proposes the use of dynamic HAs (DHAs), which reside in the serving network and are dynamically assigned by the visited authentication, authorization, and accounting (AAA) server. The DHA allows the roaming user to gain service with a local access service provider while avoiding unnecessarily long routing. The architecture defines a new node called a *packet data serving node* (PDSN) (which contains the FA), and uses VLR/home location register (HLR) (ANSI-41 or GSM-MAP) authentication and authorization information for the access network. The mobile node is identified by a network access identifier (NAI) [32] in the visiting or foreign network. An MN sends the registration message to the FA, which in turn interacts with an AAA server residing in that network or uses the broker network for authentication with the home network.

Limitations of Existing Protocols

In our opinion, the basic Mobile IP with its various enhancements as well as protocols like HAWAII and Cellular IP for mobility management in telecommunication environments have the following shortcomings:

- The basic Mobile IP has large handoff delay if the MN and HA or CN are separated by many hops in a wide area network. Location updates need to travel over the entire path from the MN to the HA/CN before the change in mobile location is effectively communicated and ongoing connections are restored. Data in transit will be lost until the handoff completes and a new route to the MN is established.
- In different versions of Mobile IPv4 (with and without route optimization) and in Mobile IPv6, location updates are always generated whenever the MN changes a subnet in the foreign network. Since subnet changes occur fairly rapidly, this approach results in frequent generation of location update messages. In situations with an extremely large population of MNs, the signaling load can become a significant portion of the traffic.
- Although the recent proposal on tunnel management [31] talks about regional registration when the distance between the visited and home networks of the MN is large, it does not specify an architecture that is directly applicable in telecommunication environments. Moreover, in this scheme, not only is the assignment of a GFA (a stabler globally valid care-of address) to a mobile performed by the FA, it is also suggested that the FA transparently append the GFA IP address information itself (as a registration extension) to the registration request message if the care-of address field is set to zero. We believe that practical implementation of such a mechanism would require the maintenance of valid security associations between all FAs and the HA, making the mobility management scheme significantly more complex. Finally, the idea of having the home network distribute the registration key associated with an MN to the corresponding GFA (to enable regional registrations in the visited domain) may weaken the strong security association paradigm between the HA and MN in conventional Mobile IP [7].
- Mobile IP schemes that specify the use of a collocated care-of address implicitly assume the availability of a pool of public addresses. As MNs become ubiquitous, the availability of such addresses may become an issue. This is particularly relevant for cellular environments since providers may be unwilling to spend resources to obtain chunks of the public address space. Furthermore, the use of public addresses by arbitrary MNs within the provider's domain may be restricted or prohibited due to security concerns, firewall restrictions, and so on.
- Since the current Mobile IP standard requires the mobile to change the care-of address (either FA or collocated) at every subnet transition, it is harder to reserve network resources on an end-to-end path between the CN and the mobile. For example, if Resource Reservation Protocol (RSVP) [33] is used to make reservations for quality of service (QoS) sensitive traffic, new reservations over the entire data path must be set up whenever the care-of address changes.
- The preceding limitations are largely avoided in HAWAII or Cellular IP by ensuring that the MN maintains a single care-of address while changing subnets or cells within a domain. However, this is achieved at the expense of requiring the establishment of source-specific routes within the administrative domain. Such a proposal does not appear to be very scalable since the state information and route lookup complexity in the routers will increase rapidly with an increased mobile population. The propagation of source-specific routes within a single domain can significantly increase signaling complexity.
- The Wireless IP network architecture (TR45.6) design uses existing standard protocols for mobility management and HLR/VLR for location update. Although this scheme offers

some flexibility in routing by assigning a DHA in the visitor network, it requires protocol upgrades at all CNs, which may limit the market acceptance of this architecture.

TeleMIP: Telecommunication Enhanced Mobile IP

Having identified some of the limitations of existing protocols and architectural proposals for supporting large-scale IP mobility and real-time packet communications in a commercial cellular environment, we are ready to present our architecture. Its essence is derived from the *registration-area*-based location management scheme currently employed in cellular networks. Such a scheme involves a combination of paging and location updates with a goal to minimize the overall cost by achieving an acceptable balance between these two kinds of traffic. Furthermore, we assume that BSs have layer 2 switching functionality similar to present-day cellular networks.

The proposed architecture, TeleMIP, is based on the observation that current IP mobility schemes have a subnet — and finer granularity of location resolution — and mostly no scoping for the transmission of location updates. Cellular IP [13], for example, proposes a base-station-level (layer 2) granularity similar to cellular networks. The current subnet-based FA scheme in Mobile IP, on the other hand, leads to a change in care-of addresses at every subnet transition. We propose a generalization of the FA concept by introducing a new node, the *mobility sgent* (MA), at network layer (layer 3) granularity, higher than that of a subnet, thus reducing the generation of global location updates. By limiting intradomain location updates to the MA, we further reduce the latency associated with intradomain mobility without resorting to source-specific routes. Finally, our two-level mobility management scheme allows the use of private addressing (and, if necessary, non-IP mobility management) within the provider's own domain.

Before presenting the architectural and operation details of our proposed protocol, let us define the various elements of our architecture and describe their roles in a TeleMIP-based mobility management solution.

Elements of TeleMIP

Most of the operational elements of our TeleMIP architecture have functionality similar to those specified in Mobile IP, with or without route optimization (as the case may be). For example, our definitions of HA, CN, home network (HN), foreign network (FN), and care-of address (CoA) are identical to the conventional Mobile IP definitions [7]. The TeleMIP architecture, however, requires some additional functionality in existing elements as well as an extra element, namely the MA:

- **MA**: An Internet host which is dynamically assigned by the network on the MN's visited network. It provides a more persistent CoA for a mobile host (MH) than currently provided by an FA. All incoming packets (and possibly outgoing ones) are routed via the MA, which thus acts as a proxy (point of attachment) for the MH in the FN. It may also have two interfaces depending on the network design; for example, if the subnet uses private address space, MA can act as a proxy/router with two interfaces.
- **MN**: A host that changes its point of attachment from one network or subnetwork to another. It may change its location without changing its IP address [7]. The MN in the TeleMIP architecture has to manage both its local and global (MA) care-of addresses. Outgoing packets can be tunnelled to the MA using the local care-of address as the outer source address; they are decapsulated at the MA and forwarded on to the global Internet.
- **FA**: An FA is present on an MN's visited subnet and provides configuration parameters to the MN. In general it assigns two addresses to the MN: an MA care-of address and an FA care-of address (this could be the FA's own address). The FA forwards the datagram to the MN. Also, an FA may serve as a local router for datagrams sent by registered MNs [7].
- **DHCP server** (or simply **server**): A host that returns configuration parameters to the MN [23]. In general, it assigns two addresses to the MN: the MA's address and subnet care-of address.

An Overview of the TeleMIP Architecture

Figure 5 shows our proposed intradomain mobility architecture. The FN is divided into several subnets depending on its geographical location. We assume that each subnet has at least one FA or DHCP server; these entities are functionally similar to the current FAs [7] or DHCP servers [23], respectively, with some modifications as discussed earlier. MAs are distributed throughout the provider's network domain and are primarily responsible for providing a globally reachable care-of address for registered MNs in that domain. Each FA or DHCP server must be associated with at least one MA in that domain. An MA is capable of handling several FAs or servers which may themselves be identified by a private addressing scheme unique to the specific domain. Whenever an MN registers in a new domain or region, it receives an MA's care-of address via the FA or server. This assignment can be performed by a load balancing algorithm. In such a scenario, MNs in a single subnet may be assigned to different MAs (e.g., using different hashing schemes). For example, MN 1 in subnet I is associated with MA 1, while MN 2 in the same subnet may be associated with MA 2.

In the TeleMIP architecture, an MN will be assigned two care-of addresses:

- A domain-specific care-of address from the public address space which is unchanged as long as the mobile stays within a specific domain or region. This is typically the address associated with the MA.
- A subnet-specific care-of address for roaming in a particular subnet. This address may have only local scope and can be either the care-of address of the FA or a locally valid collocated address. This address changes every time the mobile changes its foreign subnet.

When an MN enters a new domain or region, it will register the MA's care-of address with the HA during the initial location update process. As long as the mobile roams within this domain, all future correspondence from CNs will be directed toward this domain-specific address. The MN gets a new local care-of address every time it changes subnet; this address is obtained from the FA or DHCP server using conventional Mobile IP techniques. In general, the MA associated with a specific mobile will remain unchanged unless the association expires. The TeleMIP architecture requires communication between the MA and the associated FAs. This correspondence may indeed take place through proprietary or nonstandard protocols which are compatible with the existing telecommunications infrastructure.

The motion of the mobile between different subnets inside a domain will be transparent to the HA. In the conventional mode of TeleMIP operation, the HA will tunnel all packets received from CNs to the MA by using the domain-specific MH's care-of address. The MA will send packets directly to the MN's care-of address or through the currently associated FA. Whenever an MN changes subnets, it obtains a new local care-of address and subsequently informs the MA of this new local address binding. The MA is thus aware of the exact (subnet-level) location of the mobile and can consequently

Architecture	Location updates		Host-specfic routing updates
	Global (up to HN)	Local (within FN)	Local (within FN)
Mobile IP	P x N	----	----
HAWAII	P	P x N	P x K x N
Cellular IP	P	P x N	P x K x N
TeleMIP	P x (N/R)	P x N	----

Table 1. *A comparative chart for location updates.*

route the packet to the MN using the domain-specific routing protocol (without requiring source-specific routes). As long as the MN is under the control of a single MA, the MN does not transmit any location updates to the HA. This architecture thus ensures the localization of all intradomain mobility update messages within the domain. The scheme requires the MA to have a publicly accessible IP address associated with at least one interface. As already stated, the communication between the MA and the internal nodes within the domain (e.g., the FAs, DHCP servers, and MNs) can take place using private addressing (including possibly PSTN addresses) and routing schemes.

The advantages of our TeleMIP architecture should now be intuitively clear. By placing the global care-of address binding at a network granularity coarser than the subnet level as in the current Mobile IP architecture, we are able to significantly reduce the transmission of global updates at the expense of possible routing nonoptimality within the domain. Furthermore, since the frequency of subnet handoff within a small region is always larger than that of an interdomain handoff, we are able to considerably reduce the occurence of large-latency global location updates. Additionally, since local care-of addresses have no global visibility, this permits the use of private addressing schemes to handle intradomain mobility, thus enhancing the scalability of the mobility management scheme. A detailed quantitative comparison will be presented later.

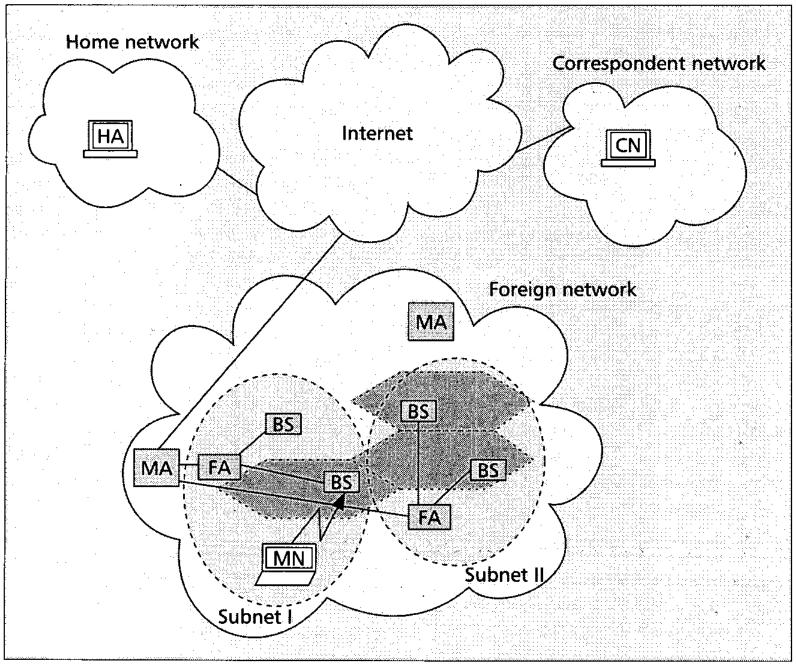


Figure 5. *Functional TeleMIP architecture.*

The Advantages of TeleMIP

The proposed TeleMIP architecture offers several advantages over other schemes:

- It aims to provide faster location updates and location tracking in IP-based telecommunication networks. By localizing the intradomain mobility update messages, this approach offers low latency and low handoff delay during registration. By appropriately placing the MAs, we can trade off between update latency and possibly nonoptimal routing within the domain.
- This approach provides flexibility to service providers by allowing the use of private address pools for local care-of addresses. Since obtaining IPv4 address in a commercial or regulatory environment may be expensive or not feasible, providers may use TeleMIP to avoid having to allocate public care-of addresses for each mobile (or even at each subnet).
- The TeleMIP architecture allows individual providers to perform their own trade-off regarding the number of subnets supported by an individual MA. Note that Cellular IP [13] provides BS-level (layer 2) granularity in a subnet: every time an MN changes a BS it updates the location. Current subnet-based Mobile IP architecture operates at subnet-level granularity by sending the location update to the HA. Such an approach may introduce additional latencies, and it is possible to have situations where the update latency is larger than the time between subnet transitions, thus leading to a complete failure of communication.
- Unlike the approach in [31], our solution is not merely a protocol but a more comprehensive architectural framework for supporting intradomain mobility in cellular wireless networks. The proposed mobility architecture requires that all registration information originate from the mobile itself, removing the need for security associations between FAs and the HA, and also allowing FA addresses to be configured from a private pool. By making the presence of an MA explicit to the MN (through the use of two different care-of addresses), we can remove the need to share any HA–MN related registration keys with the MAs. Moreover, the TeleMIP architecture proposes the use of distributed MAs and the assignment of MAs via some dynamic load balancing algorithm.
- It is well established that QoS-based protocols like RSVP do not work well if the destination address of a flow changes frequently. Therefore, it is always better to have a stable care-of address. Although our TeleMIP architecture does not provide end-to-end QoS (since the global reservation terminates at the MA and not at the MN itself), it does allow the setup of longer-term resources up to the MA. To our understanding, it is more important to support QoS guarantees over the public Internet, where congestion levels are less predictable, and policy decisions and engineering guidelines vary across service provider domains. Thus, the TeleMIP architecture is expected to be more conducive to stabler QoS guarantees in the more critical portion of the traffic path.

Performance Comparisons

Figure 6 shows an example network/domain with N subnetworks and M MAs. Each subnet has an FA or DHCP server. Each MA can handle R subnets. It is expected that R is fairly large; the performance gain from our suggested improvements will clearly be more compelling if R is large and a mobile spends a significant period of time within the domain controlled by a single MA. As a practical application in commercial environments, an MA could be expected to handle all subnets within a single city; a user would continue to be registered with a single MA as long as his/her mobility was confined to the same city. Therefore, $N/R \approx M$. It is also expected that $N \gg M$. Table 1 shows the number of location updates and corresponding routing entries generated by P mobiles as they visit the N subnets one by one (worst case) using different mobility protocols (including *TeleMIP*).

Here K represents the average number of intermediate routers where the host-specific routing entry must be updated (deleted or added) whenever a mobile changes subnets. The number of global updates represents the situation where there is only one domain root router and one gateway in the foreign domain and cellular network, respectively.

We also consider the following parameters for estimating various delays:

- $\Delta1$: time required for the registration message from MN to reach its *HA* (~200 ms)
- $\Delta2$: time required for the registration message from MN to reach the MA in the visiting domain (~10 ms)
- $\alpha1$: the rate at which the CN sends packets to the MN

Here $\Delta1$ (~200 ms) represents a typical transpacific delay (e.g., United States to Japan), whereas $\Delta2$ (~10 ms) is a typical MAN delay. Table 2 provides a comparison of TeleMIP with other protocols, using such metrics as the message *update latency*, *handoff delay*, and *packet loss* during message update.

We assume that the binding update delay for route optimization from HA to CN and MA to CN are the same as $\Delta1$ (~200 ms), and domain root routers in HAWAII and gateways in Cellular IP are placed at the same hierarchy as the MA in the visitor network, although it may not be true in a real scenario. While calculating different parameters with route optimization, binding acknowledgment is used to acknowledge receipt of a binding update message. For all performance metrics, the TeleMIP architecture shows considerable improvement over basic Mobile IP. For example, for voice applications, the update latency (~400 ms) causes a loss of 400 ms worth of voice samples, which may significantly degrade voice quality. TeleMIP, on the other hand, leads to a much lower (~10 ms) loss of voice samples, which is within a tolerable limit of interactive voice quality. Although HAWAII and Cellular IP show results similar to TeleMIP, we discussed earlier why these schemes, which assume that BSs have IP routing functionality, may require replacement of or upgrades to all existing/manufactured layer-2-capable BSs. This may not be feasible from a practical point of view. Moreover, BSs always require some time to propagate the path setup message or update the routing table entry; such functionality results in increased latency and delay (Table 1).

Conclusions

We survey several protocols, such as Mobile IP, HAWAII, and Cellular IP, and discuss their limitations in managing macro- (intradomain) mobility. Some of them have high latency during location update while changing subnets, thereby posing problems in supporting real-time service applications. We also present in detail an extension to the Mobile IP architecture, *TeleMIP*, for use in third-generation wireless networks. Our architecture introduces a new logical entity, the *mobility agent*, which offers a stable point of attachment to the mobile node. By localizing the scope of most location update messages (i.e., terminating at the MA) within an administrative domain or a geographical region, the proposed scheme supports fast handoff and real-time tracking within a domain. Among other advantages, this scheme reduces location update latency, the likelihood of losing binding update messages, and the extent of loss of inflight packets. Additionally, the TeleMIP approach permits the use of localized or private addresses, thus providing a more flexible addressing scheme, especially as the number of MNs continues to grow. By using load balancing schemes that allocate different mobile nodes to different MAs, TeleMIP promotes efficient use of existing network resources.

We have compared TeleMIP with other schemes regarding three performance metrics: message update latency, handoff delay, and packet loss during message update. Our results show that the TeleMIP architecture supports intradomain or macromobility more efficiently than other existing approaches. We are in the process of implementing a TeleMIP-based testbed for comparing its performance with that of MosquitoNet [17], SUN Microsystems [18], Cellular IP [14], University of Singapore [19], and Dynamics-HUT Mobile IP [20].

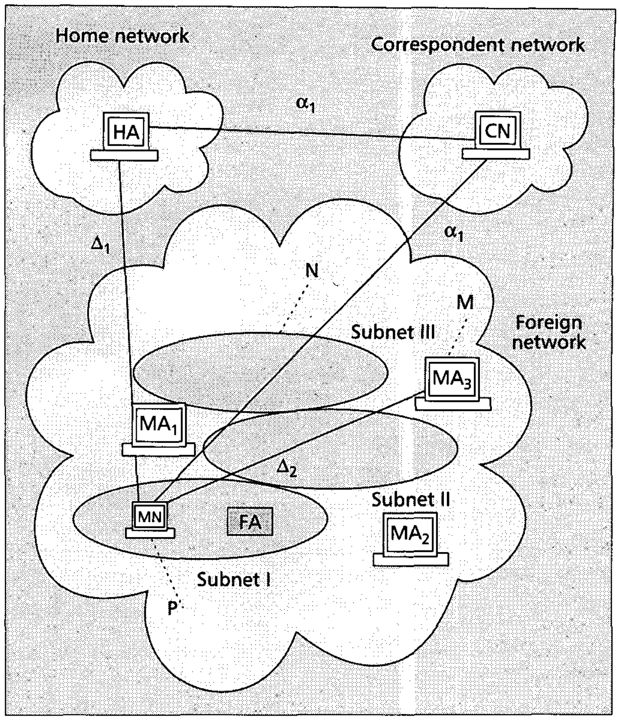


Figure 6. *An example foreign network with* N *subnetworks.*

Acknowledgments

The authors would like to thank the anonymous referees for their careful review and excellent comments which helped us improve the quality of the article. The work of Sajal K. Das was supported by Texas Advanced Research Program grant TARP-97-003594-013.

References

[1] J. Olov and G. Q. Maguire Jr., "The Effect of Using Co-Located Care-of-Addresses on Macro Handover Latency," Tech. rep., Dept. of Teleinformatics, Royal Inst. of Technology, Kista, Sweden, 1999.

Architecture	Parameters	
	Update latency/handoff delay	Packet loss
Mobile IP (without route optimization)	Δ_1 (~200ms)	$\alpha_1 \times \Delta_1$
Mobile IP (without route optimization)	$2\Delta_1$ (~400ms)	$\alpha_1 \times 2\Delta_1$
Mobile IP (IPv6)	Δ_1 (~200ms)	$\alpha_1 \times \Delta_1$
HAWAII (without Mobile IP route optimization)	$2\Delta_2$ (~20ms)	$\alpha_1 \times 2\Delta_2$
HAWAII (with Mobile IP route optimization)	$\Delta_1 + \Delta_2$(~210ms)	$\alpha_1(\Delta_1 + \Delta_2)$
Cellular IP (without Mobile IP route optimization)	Δ_2 (~10ms)	$\alpha_1 \times \Delta_2$
Cellular IP (with Mobile IP route optimization)	$\Delta_1 + \Delta_2$(~210ms)	$\alpha_1(\Delta_1 + \Delta_2)$
TeleMIP (without Mobile IP route optimization)	Δ_2 (~10ms)	$\alpha_1 \times \Delta_2$
TeleMIP (with Mobile IP route optimization)	$\Delta_1 + \Delta_2$(~210ms)	$\alpha_1(\Delta_1 + \Delta_2)$

■ **Table 2.** *A comparative delay chart with existing schemes.*

[2] J. S. M. Ho *et al.*, "Mobility Management in Current and Future Communication Networks," *IEEE Network*, vol. 12, Aug. 1998, pp. 39–49.

[3] J. S. M. Ho *et al.*, "Mobility Management in Next-Generation Wireless System," *Proc. IEEE*, vol. 87, Aug. 1999, pp. 1347–84.

[4] P. Rysavy, "General Packet Radio Service (GPRS)," *GSM Data Today Online J.*, Sept. 1998.

[5] E. Lycksell *et al.*, "IMT-2000 Standards: Network Aspects," *IEEE Pers. Commun.*, vol. 4, Aug. 1997, pp. 20–29.

[6] D. McFarlane *et al.*, "IMT-2000 Standards: Service Provider's Perspective," *IEEE Pers. Commun.*, vol. 4, Aug. 1997, pp. 8–13.

[7] C. Perkins, "IP Mobility Support," IETF RFC 2002, Oct. 1996.

[8] D. B. Johnson, C. Perkins, and A. Myles, "IHMP: A Mobile Host Protocol for the Internet," *Comp. Networks and ISDN Sys.*, vol. 27, no. 3, Dec. 1994, pp. 479–91.

[9] C. Perkins and D. B. Johnson, "Route Optimization in Mobile IP," Internet Draft, draft-ietf-mobileip-optim-07.txt, 1998; work in progress.

[10] C. Perkins and D. B. Johnson, "Mobility Support in IPv6," Internet draft, draft-ietf-mobileip-ipv6-07.txt, 1998; work in progress.

[11] S. Tripathi, P. Bhagwat, and C. Perkins, "Network Layer Mobility: An Architecture and Survey," *IEEE Pers. Commun.*, vol. 3, June 1996, pp. 54–64.

[12] T. La Porta, R. Ramjee, and L. Li, "IP Micro-Mobility Support Using HAWAII," Internet draft, draft-ietf-mobileip-hawaii-00.txt, June 1999; work in progress.

[13] A. G. Valko, "Cellular IP: A New Approach to Internet Host Mobility," *Comp. Commun. Rev.*, Jan. 1999, pp. 50–65.

[14] C-Y. Wan *et al.*, "Cellular IP," Internet draft, draft-valko-cellularip-01.txt, Oct. 1999; work in progress.

[15] T. Hiller, Ed., "Wireless IP Network Architecture Based on IETF Protocols," Ballot v. PN-4286, TIA/TR45, June 1999.

[16] E. Schooler *et al.*, "SIP: Session Initiation Protocol," IETF RFC 2543, Mar. 1999.

[17] MosquitoNet: "Mobile IP resources," Tech. rep., http://gunpowder.Stanford.EDU/mip/resource.html, 1998.

[18] SUN Microsystems:"Mobile IP resources," Tech. rep., http://playground.sun.com/vgupta/sunlabs-mobileip-new.tar.gz, 1999.

[19] Univ. of Singapore, "Mobile IP Implementation," Tech. rep., http://mip.ee. nus.edu.sg/v3.0beta, 1999.

[20] Helsinki University of Technology (HUT), "Dynamics- HUT Mobile IP," Tech. rep., http://www.cs.hut.fi/Research/Dynamics, 1999.

[21] E. Wedlund and H. Schulzrinne, "Mobility Support Using SIP," *Proc. 2nd ACM Int'l. Wksp. Wireless Mobile Multimedia*; S. K. Das and K. Basu, Eds., Aug. 1999, pp. 76–82.

[22] C. Castelluccia, X. Zhao, and M. Baker, "Flexible Network Support for Mobility," *ACM/IEEE, Proc. Mobicom '98*, Oct. 1998, pp. 145–56.

[23] R. Droms, "Dynamic Host Configuration Protocol," IETF RFC 2131, Mar. 1997.

[24] C. Perkins and D. Johnson, "Route Optimization in Mobile IP," Internet draft, draft-ietf-mobileip-optim-08.txt, Nov. 1997; work in progress.

[25] G. Montenegro and V. Gupta, "Sun's SKIP Firewall Traversal for Mobile IP," IETF RFC 2356, 1998.

[26] G. Montenegro, "Reverse Tunneling for Mobile IP", IETF RFC 2344, 1998.

[27] R. Caceres and V. N. Padmanabhan, "Fast and Scalable Handoffs for Wireless Internetworks," *Proc. MOBICOM '96*, ACM, Aug. 1996, pp. 76–82.

[28] R. Caceres and V. N. Padmanabhan, "Fast and Scalable Wireless Handoffs in Support of Mobile Internet Audio," *Mobile Networks and Apps.*, vol. 3, Jan. 1999, pp. 351–63.

[29] C. Perkins, "Mobile IP Local Registration with Hierarchical Foreign Agents," IETF Internet draft, Feb. 1996; work in progress.

[30] D. B. Johnson, "Hierarchical Foreign Agents and Regional Registration," Minutes of the Mobile IP Working Group Meeting, IETF, Mar. 1996.

[31] A. Jonsson, E. Gustafsson, and C. Perkins, "Mobile IP Regional Tunnel Management," Internet draft, draft-ietf-mobileip-reg-tunnel-01.txt, Aug. 1999; work in progress.

[32] C. Perkins and P. R. Calhoun, "Mobile IP Network Access Identifier Extension for IPv4," Internet draft, draft-ietf-mobileip-mn-nai-07.txt, July 1999; work in progress.

[33] S. Berson et al., "Resource Reservation Protocol (RSVP)," IETF RFC 2205, Sept. 1997.

Additional Reading

[1] C. Perkins, "IP Encapsulation within IP," IETF RFC 2003, 1996.

Biographies

SUBIR DAS [M] (subir@research.telcordia.com) received an integrated M.Tech. degree from the Institute of Radio Physics and Electronics, Calcutta University in 1990, and a Ph.D. degree in 1996 from Indian Institute of Technology, Kharagpur. Currently he is a research scientist in the Computer Networking Research Department, Telcordia Technologies (formerly Bellcore). During 1995–1996 he was a research associate and during 1997–1999 a faculty member in the Computer Engineering division, Electronics and Electrical Engineering Department, Indian Institute of Technology, Kharagpur. His current research interests include mobility management in 3G wireless access systems, wireless multimedia, security in next-generation networks, and auto-configurtion of ad hoc mobile networks.

ARCHAN MISRA received his M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland at College Park. Currently he is a research scientist in the Wireless OSS Department of Telcordia Technologies. His primary interests are in the design and provisioning of methodologies/architectures for supporting QoS guarantees on the Internet, and the development of mobility management solutions for IP and telecommunications networks. He has designed enhancements and developed several prototypes for Mobile IP-based seamless connectivity in commercial and military networks. His present primary focus is on the auto-configuration of ad hoc mobile networks and enhancing IP-based mobility management to support real-time/data traffic in wireless next-generation network architectures.

PRATHIMA AGRAWAL [F] is executive director of the Computer Networking Research Department and assistant vice president of the Internet Architecture Research Laboratory at Telcordia Technologies, Morristown, New Jersey. Prior to this she worked for 20 years in AT&T/Lucent Bell Laboratories in Murray Hill, New Jersey, where she was head of the Networked Computing Research Department. She received her B.E. and M.E. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, and her Ph.D. degree in electrical engineering from the University of Southern California. Her research interests are computer networks, mobile computing, and wireless communication systems. She has published over 150 papers and has received or applied for more than 40 U.S. patents. She is a member of the ACM. Presently, she chairs the IEEE Fellow Selection Committee.

SAJAL K. DAS received a B.Tech. degree in 1983 from Calcutta University, an M.S. degree in 1984 from the Indian Institute of Science at Bangalore, and a Ph.D. degree in 1988 from the University of Central Florida at Orlando, all in computer science. Currently he is full professor of computer science and engineering, and also founding director of the Center for Research in Wireless Mobility and Networking (CReWMaN) at the University of Texas at Arlington. From 1988 to 1999, he was on the faculty of computer science at the University of North Texas, Denton. He is a recipient of the Honor Professor Award from UNT in 1991 and 1997 for best teaching and scholarly research, and UNT's Developing Scholars Award in 1996 for outstanding research. His current research interests include resource management in wireless networks, mobile computing, QoS provisioning, wireless multimedia, distributed/parallel computing, network performance modeling and simulation. He has published over 150 research papers in these areas and directed several funded projects. He is a recipient of the ACM Mobicom '99 Best Paper Award.