CHAPTER 17

IP Micro-Mobility Management Using Host-Based Routing*

K. Daniel Wong, Hung-Yu Wei, Ashutosh Dutta, Kenneth Young, and Henning Schulzrinne

17.1 Introduction and Background

Global IP mobility solutions using Mobile IP or SIP are not optimized to handle micro-mobility management. For micro-mobility situations, low-latency handoffs are essential to reduce performance degradation. *Host-based routing* (HBR) schemes such as HAWAII, Cellular IP, and Micro-Mobility Protocol (MMP) are one of two main classes of IP micro-mobility management schemes; the other is hierarchical Mobile IP–derived schemes. This chapter discusses HBR schemes and examines their performance. Various simulation results and prototype system measurements demonstrate the superiority of HBR schemes over both MIP and hierarchical MIP-derived micro-mobility schemes in terms of fewer packets dropped per handoff for UDP traffic and better TCP throughput in various scenarios.

17.1.1 Chapter Overview

An overview of Mobile IP has been provided in Chapter 16. An overview of SIP for macro-mobility is included in Section 17.1.2. We do not claim that either Mobile IP or SIP is better for macro-mobility, but merely present them as alternatives. Both schemes are more suitable for macro-mobility than micro-mobility, though. Micro-mobility schemes are introduced in Section 17.1.3, although the introduction of schemes for micro-mobility based on HBR is deferred to Section 17.2 for more detailed coverage.

* This material is based upon work supported by DARPA under Contract MDA972-00-9-0009 (sub-contract RK3105 from BAE).

Following that discussion, performance-related issues are explored in Section 17.3. Section 17.4 contains selected performance results from our simulations and prototype test bed that illustrate the performance of HBR schemes. This is followed by conclusions in Section 17.5. Practical issues related to the integration of macro-mobility and micro-mobility protocols are beyond the scope of this chapter, but the reader is referred to [1].

17.1.2 The Application-Layer Macro-Mobility Management Alternative

A strength of Mobile IP, it being a network layer mobility solution, is that it is transparent to the applications above it. On the other hand, if the mobility solution were to be implemented at a higher layer (e.g., separately by each application), it might arguably be inefficient. However, this argument may not be as strong if a widely used application layer protocol were to be able to handle mobility.

Indeed, SIP [2] is rapidly gaining widespread acceptance (e.g., in IETF and 3GPP) as the signaling protocol of choice for Internet multimedia and telephony services. It fits into a possible future IP multimedia stack (Figure 17.1). SIP allows multiple participants to establish sessions consisting of multiple media streams. SIP components [i.e., user agents, servers (proxy and redirect), and registrars] provide an application layer mobility solution while interacting with other network protocols like DNS and DHCP. While SIP supports personal mobility (see Section 17.1.2.1) as part of its signaling mechanism, it can also be extended to provide support for terminal, service, and session mobility (Figure 17.2). Handoff, registration,

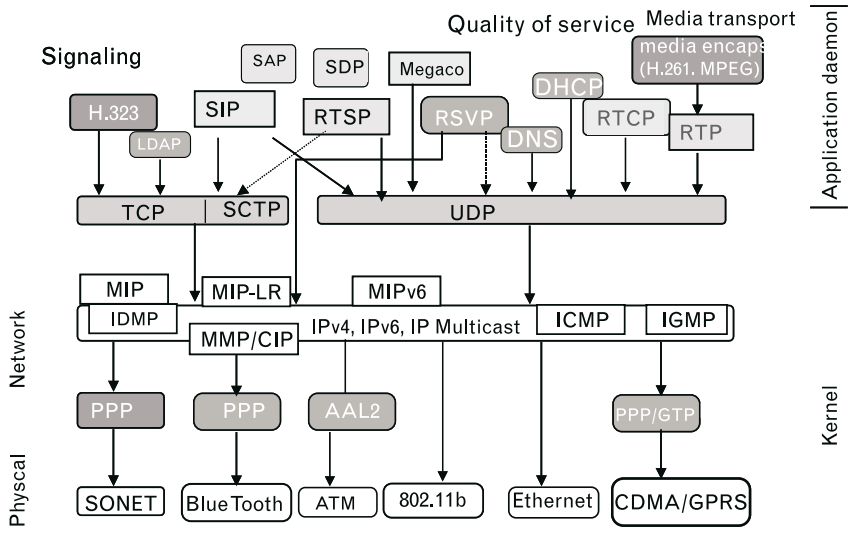


FIGURE 17.1
A possible future IP multimedia stack.

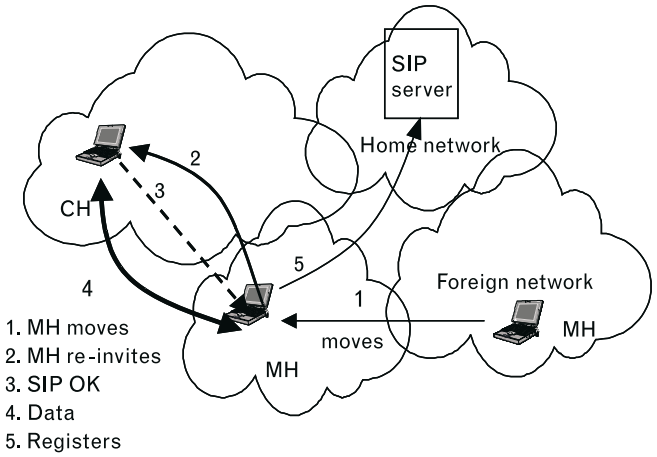


FIGURE 17.2
SIP terminal mobility illustrated.

configuration, dynamic address binding, and location management are key requirements for an SIP-based mobility scheme [3].

Mobility management in the wireless Internet may involve terminal, session, service, and/or personal mobility. Mobile IP and its derivatives, variations, and auxiliary schemes are network layer solutions that provide continuous media support when nodes move around, handling the terminal mobility problem. However, Mobile IP and related schemes by themselves do not provide means of device-independent personal, session, or service mobility. For delay-sensitive real-time applications, Mobile IP-based solutions suffer from limitations like triangular routing/registration, encapsulation overhead, and need for an HA in the home network. Mobile IP with MIP-RO route optimization alleviates the triangular routing problem but also tunnels binding updates to the HA. It requires changes in the operating system of the end hosts. MIPv6 is similar to SIP-based terminal mobility, updating the IP address on the correspondent host (CH) directly. However, it needs a 16-byte home address destination option.

Multimedia traffic is real-time or non-real-time, depending on delay and loss characteristics. Different transport mechanisms may be used for each type of traffic. Most real-time traffic should be carried over RTP/UDP, whereas non-real-time traffic has traditionally been carried over TCP. SIP-based terminal mobility provides subnet and domain handoff while a session is in progress. The SIP-based scheme provides a different approach for achieving terminal mobility by means of application layer signaling. This scheme does not rely on the mechanism of the underlying network components in the network core. Instead, proxy servers instituted by any third party service providers can provide mobility support.

When the MS moves from one subnet to another within the same administrative domain, SIP would support subnet handoff during the session as described below:

- The *mobile host* (MH) obtains a new temporary IP address through a protocol like DHCP.
- The MH reinvites the CH to its new temporary address. The identifier of the outbound proxy in the visited network is inserted in the SIP INVITE header.
- In case of domain handoff, a complete registration takes place.

A complete handoff procedure for a SIP session would consist of SIP signaling between the corresponding entities and actual media delivery. Delay associated with handoff would consist of several factors such as delay due to layer-2 detection, IP address acquisition by the mobile, activating the SIP signaling with the new address parameters, and actual delivery of media.

If the MH and CH are situated wide apart, then it may take some time for the reinvite to reach the CH. It has been proposed [4] that an RTP translator can be affiliated with a SIP proxy server that would intercept the traffic and would send the media to the current location of the mobile host. Thus, RTP translators reduce the end-to-end handoff delay (due to traversal of the INVITE request) to a one-way delay between the MH and the SIP proxy. In cases when both communicating hosts move during a session, each side would have to issue INVITE requests through their respective home proxy servers, where the MHs register their new location address after the movement.

While the RTP translator concept may reduce the micro-mobility problem somewhat, SIP does not in itself provide an optimized, targeted solution to the micro-mobility problem. Like Mobile IP, it is optimized for macro-mobility. Based on this brief examination of Mobile IP–based and SIP-based macro-mobility management, it can be deduced that a highly desirable property for a micro-mobility scheme is flexibility to work with a variety of macro-mobility schemes, and not just Mobile IP–based macro-mobility.

17.1.2.1 SIP Support for Other Types of Mobility

In addition to terminal mobility, SIP also supports other mobility concepts—namely, personal mobility, service mobility, and session mobility. Arguably, SIP offers a more unified macro-mobility management scheme than Mobile IP and its variations, which are more limited.

Personal mobility is the ability of users to originate and receive calls and access the subscribed network services on any terminal in any location in a transparent manner, and the ability of the network to identify end users as they move across administrative domains. SIP's URI scheme and registration mechanism are some of the main components used in providing personal mobility. A roaming subscriber is accessible independent of the device

the subscriber uses. Service mobility refers to the subscriber's ability to maintain ongoing sessions and obtain services in a transparent manner regardless of the subscriber's point of attachment. Session mobility allows a user to maintain a media session even while changing terminals such as transferring a session that began on a mobile device to a desktop PC after entering an office.

17.1.3 Micro-Mobility Management

The requirement that Mobile IP registration (or SIP reinvites) be performed every time an MH moves between subnets may cause high handoff latency. Various solutions have been proposed. The proposals generally implicitly or explicitly use a concept of micro-mobility regions where these regions comprise numerous subnets, and registrations with the HA are not necessary for movement of the MH within these regions. Registration with the HA is still necessary for movement between micro-mobility regions. Typically, Mobile IP handles the macro-mobility (mobility between micro-mobility regions), while a micro-mobility scheme handles micro-mobility (mobility within micro-mobility regions). Micro-mobility management schemes reduce the high handoff latency of Mobile IP by handling mobility within micro-mobility regions with low-latency local signaling.

17.1.3.1 Hierarchical Mobility Agent Schemes

Micro-mobility solutions using a hierarchy of mobility agents include Mobile IP *with regional registration* (MIP-RR) [5] and TeleMIP/*Intradomain Mobility Management Protocol* (IDMP) [6].

MIP-RR involves few modifications to Mobile IP. In a foreign network, the two level mobility hierarchy contains the upper-layer GFA and several lower-layer *regional foreign agents* (RFAs). All MHs under the GFA share the same COA.

Suppose an MH moves between subnets under a GFA with which it is already registered. As shown in Figure 17.3(b), the MH initiates its registration with FA2. Then the registration request is sent to GFA1. Since MN is already registered with GFA1, GFA1 does not initiate a home registration to HA, but just sends the registration reply to the MH through FA2. Since the HA does not need to be contacted in this scenario, MIP-RR reduces the handoff latency.

If the MH changes its GFA, it needs to register with its HA. As shown in Figure 17.3(a), the MH moves from FA3 to FA2, and its GFA is no longer GFA2. The MH sends a registration request to its new RFA, which is FA2, and then GFA1. Because GFA1 is a new GFA, it has to register with the HA. The HA sends the registration reply all the way through GFA1 and FA2 to the MH.

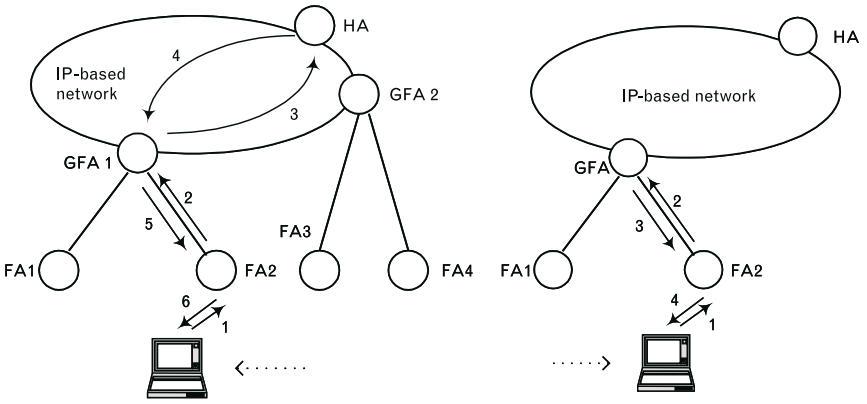


FIGURE 17.3
Mobile IP regional registration: (a) movement between regions; (b) movement within a region.

If the MH moves frequently within a GFA domain, it does not need to perform the time-consuming registration procedure with its HA. Therefore, the average handoff latency is reduced.

Like MIP-RR, TeleMIP is a hierarchical IP-based architecture that provides lower handoff latency and signaling overhead than Mobile IP. TeleMIP uses Mobile IP as a macro-mobility management protocol and can interwork with SIP-based mobility. IDMP is the micro-mobility protocol used with TeleMIP (Figure 17.4). IDMP uses multiple COAs that are taken care of by SAs and the MA at the subnet and domain level, respectively. SAs are FAs or DHCP servers at the subnet level that provide an MH with a

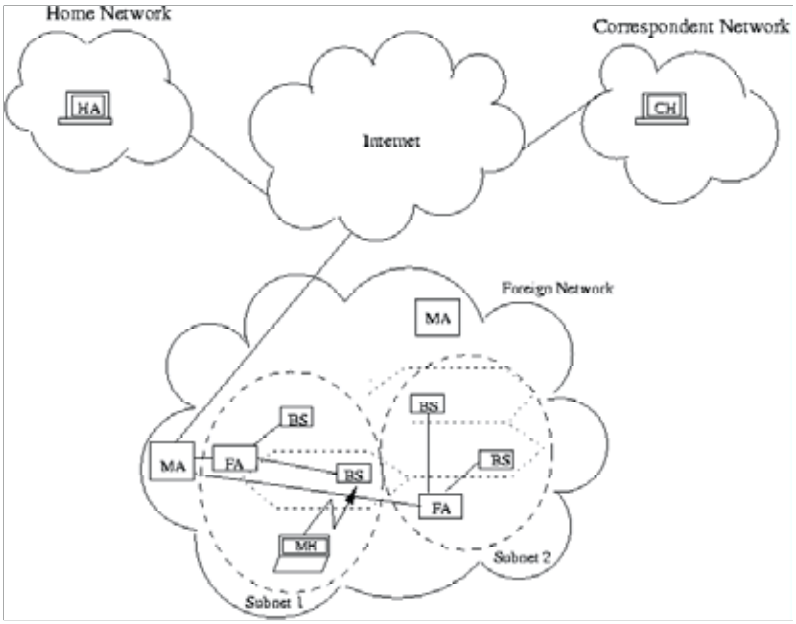


FIGURE 17.4
TeleMIP's IDMP.

locally scoped address, and they are analogous to MIP-RR RFAs. The locally scoped address provided by an SA identifies the MH's location within the domain. An IDMP MA is similar to a MIP-RR GFA and acts as a domain-wide point for packet redirection.

The serving MA provides an MH with a global COA that stays constant as the MH moves within the domain. Unlike MIP-RR, multiple MAs can be provisioned for load balancing and redundancy within the domain. All packets from the global Internet are tunneled to the MA (or one of the MAs, in the case of multiple MAs). The serving MA forwards packets to the MH using regular IP routing, with the local COA (colocated or FA) as the destination. It does this by un-encapsulating the packet and then performing a second encapsulation of the IP packet, with the local COA as the destination address.

On subsequent movement within the domain, the MH only obtains a new local COA. At that point, the MH needs only to update its MA with its new local COA. By limiting intradomain location updates to the MA, it reduces the latency associated with intradomain mobility. In addition, IDMP also provides the added advantage (over MIP-RR) of dynamic load balancing, within a domain.

17.2 HBR Overview

A class of micro-mobility management schemes is that which employs HBR, including Cellular IP [7], HAWAII [8], and MMP [9]. HBR schemes for micro-mobility could be considered a class of auxiliary schemes that deal with the handoff latency problem of Mobile IP. However, they have grown beyond being just a class of auxiliary schemes (e.g., MMP is designed to be usable with SIP mobility for real-time traffic, and with a Mobile IP variant for nonreal-time traffic). This flexibility is an advantage over micro-mobility schemes based on a hierarchical Mobile IP structure (e.g., MIP-RR or TeleMIP). A second major advantage of HBR schemes is that they offer the lowest latency networking rerouting solution for micro-mobility. This is because hierarchical Mobile IP–derived schemes like MIP-RR and IDMP only reduce the latency problem inherent in Mobile IP registration. The GFAs or MAs still need to cover a large area to be scalable and cost-effective. Location updates still need to reach them in these schemes. On the other hand, with HBR schemes, updates take an optimal path to the closest node that should handle the location/route update, namely the crossover node, as will be explained shortly. The performance results in Section 17.4 support these assertions with numerical results.

The distinctive characteristics of HBR schemes for micro-mobility are that (1) HBR is used within the micro-mobility regions, (2) very low-

latency handoffs are possible since the update message only needs to propagate to the crossover node, and (3) one or more special nodes (known as gateways or root nodes) are used as the demarcation point between each micro-mobility region and the rest of the Internet.

With HBR schemes, forwarding behavior is specified separately for each host. For example, nodes may route packets according to tables or caches indexed by unique host identifiers (e.g., their IP address). MMP, HAWAII, and Cellular IP are examples of HBR where the indexing is by host IP addresses. HBR differs from group-based routing schemes, where forwarding behavior is specified for groups of hosts. For example, for group-based routing, nodes may route packets according to tables or caches indexed by group identifiers (e.g., IP address prefix and netmask)—that is, packets with different destination addresses, but where the destination addresses match the prefix and netmask, will be routed in the same way.

A critical advantage of HBR schemes is that location management and routing can be integrated. With Mobile IP, location management is handled by registrations, while routing is overlaid on the existing IP-based network routing. The location management requires possibly long-latency registrations to a potentially distant HA whenever subnet boundaries are crossed, while the use of overlay routing over standard IP routing creates problems like triangular routes and encapsulation overhead. HBR, on the other hand, gives the power to update routes simultaneously and precisely with location management, precisely because the routes are host specific and do not affect routes to/from other MHs when changed. For the lowest latency handoffs, the intuition would be to update the routing information for an MH at the closest intersection with the old route (also known as the crossover node), whenever it performed a handoff. And this is indeed what happens with the HBR schemes for micro-mobility management.

Rather than present three separate descriptions of the three HBR schemes (MMP, HAWAII, Cellular IP), a generic, bare-bones HBR scheme will be introduced in Section 17.2.1, and then differences between MMP, CIP, and HAWAII will be discussed in Section 17.2.2. This approach brings out the essence of the HBR schemes before explaining the minor differences between them. A comparison with ad hoc routing protocols that use HBR-like routing follows in Section 17.2.3.

17.2.1 A Generic HBR Solution for IP Micro-Mobility

A generic HBR scheme is now presented to illustrate the essential workings of HBR in facilitating low-latency handoffs for IP micro-mobility. Actual protocols differ from this generic scheme in some aspects. For example, actual protocols use various enhancements to provide more seamless handoffs. There may be multiple nonoverlapping micro-mobility domains in a

given network. Movement between micro-mobility domains may be handled by a macro-mobility protocol like Mobile IP or SIP.

In the generic HBR scheme, each HBR micro-mobility domain has one root router that serves as the interface between the micro-mobility domain and the rest of the network. The other infrastructure nodes in the HBR micro-mobility domain are arranged in a strict inverse tree structure beneath the root router. In other words, every node is a child of one and only one other node (possibly the root router), and may be a parent of zero, one, or more other nodes. The inverse tree structure can be seen in Figure 17.5, which shows an abstraction of a generic HBR scheme for IP micro-mobility. Some of the nodes at the bottom of the hierarchy are BSs, with wireless interfaces, and whose coverage areas are called cells. Each infrastructure node other than the root router has one upstream interface and zero, one, or more downstream interfaces. The upstream interface is the interface towards the root router, while the downstream interface(s) is/are towards the BSs and MHs. How nodes know which are the upstream interfaces and which are the downstream interfaces is unspecified. However, one way this may happen is by listening to beacon messages sent periodically down by the root router (the interface through which the beacon arrives is recorded as the upstream interface and is used as the next hop for routing of any packet to the root router; the rest are considered downstream interfaces). Another way might be if the HBR scheme is implemented as an overlay over standard IP routers, as with HAWAII. Alternatively, they may be preconfigured.

When an MH first enters an HBR micro-mobility domain, the network would use access control mechanisms, such as authentication of the MH. Some form of macro-mobility signaling would be initiated, so that macro-mobility can be handled by the relevant protocol. The MH may use its home IP address, or it may obtain a temporary address for use in the HBR domain. The temporary address, if applicable, is called a COA, although it is

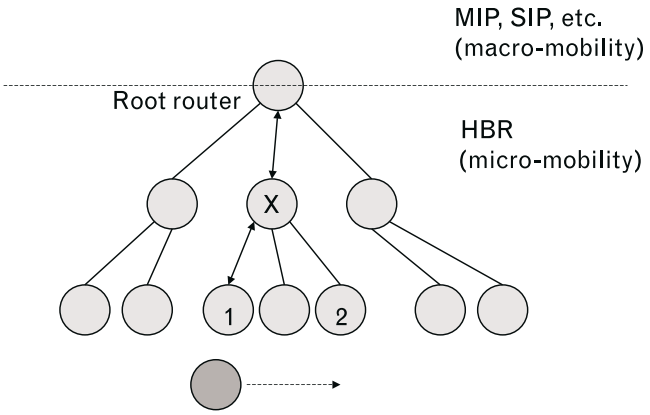


FIGURE 17.5
Abstraction of generic HBR scheme for IP micro-mobility management.

not necessarily exactly like a Mobile IP COA. It may be obtained through various means, and may be a globally routable address of the root router, or be colocated. In any case, the MH uses one IP address as long as it is in the same HBR domain, even when it moves from cell to cell.

Upstream routing of packets (from MH to gateway) is simple: once packets from a particular MH are admitted into the micro-mobility domain infrastructure, each node (BSs included, root router excluded) merely forwards upstream packets to the root router through their upstream interface. Therefore, the root router eventually gets upstream packets. It then routes them normally through the IP-based network, or it sends them back through its micro-mobility domain infrastructure, depending on the destination address.

Downstream routing of packets depends on routing caches maintained in each node. By definition of HBR, there are separate cache entries for each MH. These entries are set whenever upstream data passes through a node. The node reads the source IP address to identify the MH, and then binds it in the cache to the interface address (plus MAC address, if necessary) of the incoming packet, with the assumption that it is the right interface (and MAC address) to use for downstream packets destined to that MH. In order to facilitate simpler handoffs, the cache entries have a soft state and need to be periodically refreshed. Since it cannot be assumed that upstream data is always being transmitted, periodic control packets called route updates are transmitted by each MH to refresh the cache entries after a period of no upstream transmissions from that MH.

To explain how handoffs are handled, Figure 17.5 shows an MH (the unattached node at the bottom of the figure) moving from the cell covered by the BS labeled 1 (BS1) to the cell covered by the base station labeled 2 (BS2). The node marked "X" is the crossover node for this handoff. The crossover node is defined as the lowest node that is in the upstream of the paths from both BSs to the root router. In some cases, it may be the root router itself, but often the crossover node is below the root router. The MH initiates handoff by sending a route update through the new BS, BS2 in this case. As this route update propagates up to the root router, for each node below the crossover node, the routing cache may not have any binding for the MH, and the appropriate entry is added. At the crossover node, the critical switch occurs as the binding for the MH is updated to point towards the interface heading towards BS2. As the route update continues up to the root router from the crossover node, the routing cache entry at each node is updated as normal, as though no handoff had occurred.

HBR may not require any signaling with the old BS, BS1, nor any of the nodes between BS1 and the crossover node (some signaling occurs through these nodes in HAWAII, though, in one of the enhancements to provide seamless handoffs). Since the routing caches contain soft-state

entries, the entries will naturally time out and be removed without needing additional signaling. This is one of the reasons why soft-state routing cache entries are used. Another reason is that the MH does not need to perform any deregistration when it leaves the HBR domain, which it would need to do to update the routing caches if they had hard-state entries. Furthermore, if for any reason the MH loses connectivity or crashes, the hard-state entries would not be removed.

The generic HBR solution could also have some kind of paging scheme to provide paging gains over Mobile IP (see Section 17.3.1.1). One way this could be done would be by using paging caches, similar but parallel to the routing caches. The paging caches would have longer expiry times, so that they need not be refreshed as often as routing caches. When an MH is idle, it only needs to send paging updates occasionally.

17.2.2 Comparison Between HBR Schemes

In this section, specific differences between HBR schemes like HAWAII, Cellular IP, and MMP are discussed. Although each of these protocols has much in common with the generic HBR scheme just described, their design goals are not all the same. For HAWAII, the design goals [8] are to limit disruption to user traffic, enable efficient use of access network resources, enhance scalability by reducing updates to the HA, provide intrinsic support for QoS, and enhance reliability. For Cellular IP, the design principles [7] are to use universal building blocks as nodes, be a plug-and-play solution, be scalable, minimize the burden on the MH, and support passive connectivity. For MMP, the design principles are to limit disruption to user traffic, be robust, reliable, and survivable, enable efficient use of low-bandwidth access network resources, be a plug-and-play solution, be scalable, minimize the burden on the MH, support passive connectivity, and facilitate QoS support.

The MH in HAWAII is a Mobile IP client. The HAWAII domain looks like an FA to the MH, and its root router looks like an FA to the HA of the MH. On the other hand, the MH in Cellular IP needs to use Cellular IP signaling (e.g., route updates and the root router takes care of the Mobile IP signaling). As for MMP, the MH needs to use MMP signaling, and the root router acts on behalf of the MH to perform signaling for the macromobility scheme, whether it is Mobile IP or SIP. The advantage of having the MH be an ordinary Mobile IP client is that no changes are needed to MHs that already have Mobile IP client software. However, this choice is not made in Cellular IP, since that would go against the "universal building block" principle because then the BSs would have extra IP-level functionality in addition to Cellular IP node functionality (i.e., they would have to send updates to the root router on behalf of the MH). Furthermore, this

would reduce the ability to put together a plug-and-play wireless IP access network. This second reason applies to MMP as well. Additionally, however, MMP MHs cannot be Mobile IP MHs, since MMP is designed to work with several macro-mobility schemes, not just Mobile IP.

Although both Cellular IP and HAWAII are designed to work with Mobile IP as the macro-mobility management scheme, each works with a different mode of Mobile IP. In HAWAII, the MH acts as a Mobile IP MH in colocated COA mode, while in Cellular IP, the gateway acts as an enhanced FA that takes care of Mobile IP registration on behalf of the MH. Another major choice is whether the HBR scheme is implemented as an overlay over standard IP routers running in intradomain routing protocol (e.g., RIP, OSPF), as with HAWAII, or whether the HBR routing is the only routing that the HBR nodes are capable of, as with Cellular IP. The overlay approach allows more sophisticated seamless handoff techniques, as it allows HBR nodes to communicate with one another to forward buffered packets, and so on. It also may be able to rely on some of the reliability mechanisms of the underlying intradomain routing protocol, and may be easier to implement using existing equipment. However, it is not as lightweight and scalable over a range of wireless access environments as the nonoverlay approach, since full-fledged IP routers are used.

One of the design goals of MMP is to be robust, reliable, and survivable. Because of the efficiency in signaling and distribution of the routing information in the generic HBR scheme, only the nodes along the path between the serving BS and the root router know how to route to the MH. This could result in disruption in communications if a node or link fails, or if the root router fails. The problem is shared by Cellular IP, and to some extent, by HAWAII. With MMP, however, there is the option to use multiple root routers, and there is also the option for some or all the nodes to have more than one parent node. This provides for robustness and survivability, at the cost of a little more complexity and signaling traffic within the HBR domain.

Another difference between the schemes is in how handoffs are treated. The basic HBR handoff scheme is equivalent to the hard handoff scheme of MMP and Cellular IP. Although HBR schemes can achieve fast, low-latency handoffs by using only their very fast local updates, much faster than Mobile IP, packets could still be dropped. Therefore, various seamless handoff schemes have been proposed to improve performance even further. In HAWAII, there are four schemes for setting up the new path when handoffs occur. With the *multiple stream forwarding* (MSF) and *single stream forwarding* (SSF) schemes, the old BS receives the initial handoff message and signals with the new BS to set up a path with the new BS to forward packets there that have been buffered at the old BS. Using MSF could result in multiple out-of-order streams arriving at the BS through the new BS, as some earlier

packets being forwarded from the old BS to the new BS may arrive later than newer packets from the crossover node to the new BS for very short periods of time. SSF is more sophisticated, and results in a single stream of packets being forwarded to the new BS. The crossover node in this case needs to be informed by the old BS when it has cleared its buffers of packets for the MH, and only then would the crossover node switch packets for the MH over to the new BS. With the other two handoff schemes in HAWAII —*unicast nonforwarding* (UNF) and *multicast nonforwarding* (MNF) —it is the new BS that receives the initial handoff message. UNF, in the network, is like the basic hard handoff except that the old BS is informed through signaling from the new BS and it sends an acknowledgment back to the MH. The difference from simple hard handoff is that the MH is assumed to be able to communicate with both BSs during the handoff period, to reduce packet losses associated with hard handoff. As for MNF, it is like UNF except it uses a special dual-cast scheme (from crossover node to both BSs) for a short period of time from the time the crossover node receives the handoff message so the MH does not have to talk to two BSs simultaneously. On the other hand, the only seamless handoff scheme with Cellular IP, semi-soft handoff, is somewhat like MNF. The difference is that the MH switches back to the old BS after sending the initial handoff message, to reduce packet loss while the message is traveling to the crossover node.

There are also other differences between the actual protocols, such as paging schemes, but for more details, the reader is referred to the source documents (e.g., [10]).

17.2.3 Comparison with Ad Hoc Mobility Schemes

Another class of routing problems where various nonhierarchical, HBR-like schemes have been proposed for handling mobility is that of ad hoc routing in *mobile ad hoc networks* (MANET) [11]. In ad hoc networks, the nodes are not arranged in a fixed infrastructure, typically because they are mobile and are constantly changing positions with respect to one another. The distinction between the fixed infrastructure and mobile hosts may vanish, and every node may well be a mobile router. The lack of a fixed infrastructure makes the route discovery problem more difficult than in the case of the HBR schemes for micro-mobility management.

A variety of hierarchical routing protocols have been proposed, using concepts of ad hoc clusters of nodes based on factors like proximity and geographical location. However, such protocols may depend on assistance from the GPS or depend on the existence of a core of "backbone" nodes or use some heuristics for selecting cluster heads. Other ad hoc routing protocols are nonhierarchical but flat, which is closer to the routing within HBR micro-mobility domains. Among the flat ad hoc routing protocols, some,

like AODV routing, are reactive. These differ from the HBR schemes in that the routes are computed in an on-demand manner, as needed.

The flat, proactive ad hoc routing protocols such as DSDV routing may be closer to the HBR schemes. In DSDV, all nodes maintain a routing table that contains separate entries for all the possible destinations, which are periodically refreshed. Two differences between DSDV and HBR for micro-mobility are noticeable. While in DSDV, all nodes maintain a routing table for all the destinations; in HBR only the infrastructure nodes maintain these tables, and then only for MHs being served by one of their children or descendent nodes. The other difference is that the refresh problem is an order of magnitude more challenging in DSDV, since the nodes all are moving. Infrequent transmissions of full dumps are needed, where full dumps contain all routing information, and periodic transmissions of incremental packets are used to relay information on changes occurring since the last full dump. On the other hand, for HBR, route updates are always incremental and specific to mobile hosts, very quickly providing up-to-date information on MH location after movement occurs. Even the periodic refreshes are meant to be for optimizing network resource utilization more than to handle significant network topology changes.

Despite these advantages of HBR, ad hoc schemes must still be used in cases where the mobility situation demands it. However, whenever HBR can be used, it should be—for example, in less mobile or semi mobile networks (e.g., a tactical network where some "infrastructure" nodes move infrequently and remain on the same hierarchy even after moving)—since it also has additional advantages over ad hoc schemes. The root router provides a natural transition point between the micro-mobility region and the macro network. Moreover, MANETs run into a scalability problem with as few as 50 to 100 nodes because of all the updating and exchange of route information that goes on, whereas HBR domains can handle thousands of nodes.

An interesting and open area of research is where the crossover between an HBR network and a MANET might occur. For situations where there is a relatively stable infrastructure, HBR makes sense, whereas ad hoc routing protocols would need to be used in more mobile, fluid networks. A key question is how to qualify and quantify what is meant by a relatively stable infrastructure with core nodes, in which the network can take advantage of the core nodes to reduce signaling overhead. Ad hoc routing protocols like CEDAR that assume that core, high-bandwidth backbone nodes can be found, are closer to HBR in a sense. One could imagine a self-organizing protocol where nodes perform some self-discovery of the network and switch into an HBR mode or an ad hoc routing mode depending on certain conditions. It could periodically check if the conditions have changed, and switch modes if necessary.

17.3 Performance Issues

Performance is examined qualitatively in Section 17.3.1. Previously reported quantitative results are discussed in Section 17.3.2, as a prelude to discussing our performance results in Section 17.4.

17.3.1 A Qualitative Perspective

The major goal of micro-mobility schemes using HBR is providing fast, low-latency handoffs. This may result in fewer packets dropped during handoffs, leading to less disruption of UDP traffic and better TCP throughput performance.

While obtaining significant reductions in handoff latency compared to MIP is a major accomplishment of micro-mobility management schemes, other advantages have been claimed. These include:

- Reduction in signaling overhead through paging concepts;
- Reduction in packet header overhead in the low-bandwidth radio access network through not using encapsulation;
- Easier integration with QoS provisioning;
- Better use of scarce IP address resources by using fewer IP addresses than Mobile IP or hierarchical Mobile IP derivatives

These advantages will be discussed, qualitatively, in Sections 17.3.1.1 to 17.3.1.4. Other issues, such as scalability, will be discussed in Sections 17.3.1.5 to 17.3.1.6.

17.3.1.1 Paging Gains

The idea behind paging gains is that Mobile IP does not differentiate between active and idle MHs. It requires that MHs go through the registration process whenever MHs move between subnets, regardless of the activity level of the MH. There are two problems with this. First, the signaling overhead is high, even if the MH is idle (communications-wise) but moving around rapidly. Second, each of the registrations with movement between subnets would consume power from the MH's battery. The first of these problems is largely dealt with in an HBR domain even without paging, because macro-mobility signaling would not need to be invoked upon every handoff occurring. The second problem, however, can be dealt with using an idea (paging) borrowed from traditional wireless cellular networks.

Cellular mobile networks have long differentiated between active and idle states of a MH. When an MH is idle, it registers less often with the

network, the trade-off being that the network knows the MH's position with less precision and needs to page the MH to reach it when it needs to communicate with it. The less often the MH registers, the less precisely the network will know its position and the larger the paging area. The HBR schemes implement variants of the paging concept. However, this is not a fundamental flaw of Mobile IP, nor a fundamental advantage of HBR schemes. Moreover, a paging extension to Mobile IP has been proposed [12].

17.3.1.2 Reduction in Packet Header Overhead

Mobile IP adds at least 8 to 12 bytes per packet for minimal encapsulation, and more for alternative encapsulation schemes. With small packets, this can make a significant difference compared to a scheme without encapsulation overhead [13]. Since micro-mobility regions are often in wireless access networks where bandwidth efficiency may be at a premium, it is an advantage of HBR schemes that there is no encapsulation overhead. This advantage is over Mobile IP, and also over other non-HBR micro-mobility schemes that use encapsulation, like TeleMIP/IDMP and MIP-RR.

17.3.1.3 QoS

In the IntServ model for providing QoS, resource reservation protocols like RSVP are used to reserve network resources. The resource reservation, however, assumes that the endpoints have unchanging IP addresses. When an endpoint changes IP address (e.g., an MH obtains a new COA), the old reservations cannot be used, and new reservations need to be made. With HBR schemes, MHs keep the same IP address within an HBR domain, providing a more stable endpoint for RSVP than Mobile IP does.

17.3.1.4 Use of IP Addresses

There is a shortage of IP addresses in the IPv4 address space. Using Mobile IP for micro-mobility would require a pool of IP addresses to be set aside for use as COAs in every subnet. Since MHs with HBR can keep one IP address as they move within an HBR domain, a pool of COAs can be set aside for the entire HBR domain, if necessary, resulting in a more efficient use of IP addresses.

17.3.1.5 Scalability

HBR has a potential scalability problem in that the forwarding cache grows linearly with the number of hosts. To deal with the scalability problem, one solution is to use a group-based routing scheme. The Internet is an example

of a network with group-based routing. Furthermore, the groups are hierarchical, with smaller groups as subsets of larger groups, providing a very efficient and flexible way to specify routing behavior. At the minimum, a routing table may contain a default route that specifies how to route all packets.

The Internet uses a hierarchical routing scheme because HBR does not scale. An alternative way to deal with the scalability issue is to restrict the number of hosts involved, for example, to just the hosts roaming within a certain region. HBR works for micro-mobility because it is confined to a definite region, with a gateway between the micro-mobility region and the rest of the Internet.

17.3.1.6 Communications Between Two MHs in the Same HBR Domain

For cases where the CH is not in the same HBR domain as the MH, the routing within the HBR domain is optimal in both directions. Uplink packets go straight to the root router as they should, and downlink packets go straight to the correct BS, as they should. However, what happens when the CH is also an MH in the same HBR domain? The way that CIP, HAWAII, and MMP have been specified at this time, the route would be through the root router and down to the other BS. Even if the two MHs had a crossover node that was the direct parent of their respective BSs, packets would still go to the root router. The reason is that the HBR nodes along the path simply forward any uplink packets to the root router regardless of final destination.

Should this routing "inefficiency" be removed? A straightforward solution might be to check the destination address of every uplink packet with the contents of the routing cache to see if it should be forwarded down rather than to the root router. The problem, however, is that the uplink forwarding would become less efficient. It is unclear if the trade-off is worthwhile, given that the percentage of traffic from one MH to another in the same domain might be very low. Even if that percentage is not that low, the "inefficient" route through the root router is not a serious inefficiency because the added latency would not be very high and would be naturally bounded by the size of the HBR domain. Nevertheless, a solution has been proposed that introduces the concepts of optimizing Cellular IP node, proxy route-update packet and optimizing teardown packet [14].

17.3.2 Quantifying Performance

Attempts have also been made to quantify the performance of HBR schemes. Reference [7] describes an experimental prototype and measurements made thereupon, which show that TCP throughput decreases as the handoff rate increases. Measurements also show how semi soft handoff (also known as advanced binding handoff) performs better than hard handoffs,

experiencing less of a TCP throughput decrease as the handoff rate increases. Reference [8] provides performance results using a novel network simulator developed at Harvard University. It compares HAWAII with Mobile IP in terms of average number of dropped packets (UDP case) per handoff. TCP throughput of HAWAII is compared with that of Mobile IP as handoff frequency varies from 0.5 to 4 times per second.

While the first-order performance improvements are based on much reduced handoff latency (over Mobile IP) using HBR, second-order performance improvements may be available through various additional optimizations. For example, attempts at seamless (or almost seamless) handoffs, including semi-soft, MNF, UNF, MSF, SSF, can further reduce packet losses, at the cost of additional complexity. In this chapter, we focus more on the performance improvement related to reduced handoff latency over Mobile IP. Our simulation results extend, as well as complement, the performance results of [7], [8].

17.4 Performance Results

In this section, performance results (mostly from computer simulations using NS2, but also including some analytical and laboratory prototype results) are discussed. Section 17.4.1 introduces the simulation environment. Various results are discussed in Section 17.4.2, which is divided into two main types of simulations: UDP simulations and TCP simulations. Finally, Section 17.4.3 describes some measurement results from the laboratory prototyping.

17.4.1 Simulation Environment

The base simulation setup for our simulations of HBR schemes for micro-mobility management is illustrated in Figure 17.6. A simple wireless model is used that assumes perfect overlapping coverage, no propagation delay, and no transmission errors. Furthermore, handoffs are smooth and instantaneous at layer 2 and below. The link latency for the plain (straight-line) links in the micro-mobility domain are 2 ms, whereas the link latency of the dash-dot links in the Internet are 10 ms. The link bandwidths are 375 Kbps within the micro-mobility region, and 1.544 Mbps in the "Internet." Routing and paging cache entries need to be refreshed, and the route update and paging update intervals used are 3s and 60s, respectively. The size of each update packet was 100 bytes. Lightly loaded network conditions were simulated, with only one MH and one CH. TCP Tahoe was used as the transport protocol, with a flat application data rate of 200 Kbps. The application is assumed to not be delay sensitive (i.e., it is equally acceptable for the

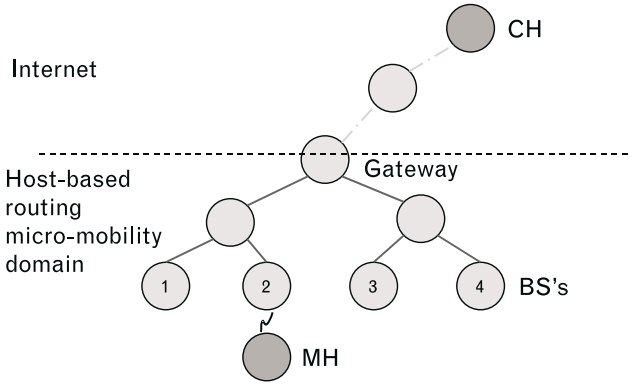


FIGURE 17.6
Base simulation setup for HBR micro-mobility simulations.

instantaneous throughput to fluctuate a lot as it is for the instantaneous throughput to be relatively constant, provided that the average throughput is the same). The handoff rate is once every 5 seconds on average, with exponentially distributed interhandoff intervals, and the handoffs are back and forth between the two base stations labeled BS1 and BS2. Each simulation was run at least as long as needed for 200 handoffs to occur.

Simulations were run comparing performance in terms of (1) average number of packets dropped per handoff for UDP traffic and (2) the TCP throughput. In different simulations, the effects of varying link latencies, handoff frequencies, application data rate, link bandwidths and other parameters were investigated. While TCP Tahoe was the default TCP used, TCP Reno was also simulated for comparison. Since there have been various seamless handoff schemes proposed, it has been decided that for this chapter only the basic HBR hard handoff and one representative seamless handoff scheme, the semi soft handoff scheme, be simulated. The reason for including basic HBR hard handoff is to bring out the performance gains of HBR schemes resulting simply from the low-latency route updates even without any auxiliary schemes for seamless handoffs. The reason for including one representative seamless handoff scheme is merely to confirm and illustrate that such schemes do indeed help further improve performance. In this, the choice of semi soft handoffs is somewhat arbitrary, partly because its performance is expected to be somewhat moderate compared with the other seamless handoff schemes.

The performance of HBR schemes has been compared with the performance of Mobile IP where the base simulation setup for Mobile IP simulations is shown in Figure 17.7. The wireless model, handoff model, and other parameters are almost identical to those for the HBR simulations, and the topology is similar to that in Figure 17.6, in order to allow for meaningful comparisons. Each BS now becomes a different subnet and has either an FA or an HA. The Mobile IP registration request and reply messages are each 100 bytes long. Minimal encapsulation is simulated, adding only 12

bytes to the unencapsulated packets. MIP periodic reregistrations are not simulated in this model because the intervals typically are long, in the order of many minutes, and so would hardly impact the simulation results. As for the dashed link, the link latency on that link was varied (from the original 2 ms to 10 ms to 100 ms), to simulate the real possibility that the HA is further away. It should be noted that the case where the dashed link has only 2 ms of latency should be useful to indicate the performance of either (1) MIP where the HA is very close to the FA, or (2) micro-mobility management by MIP-RR or TeleMIP/IDMP (without seamless handoff enhancements). In the case of MIP-RR or TeleMIP/IDMP, the HA in Figure 17.7 would be analogous to the GFA or the MA, in that it is very close to the FA (which would be the RFA or SA for MIP-RR or IDMP, respectively).

17.4.2 Simulation Results

17.4.2.1 CBR UDP Traffic

CBR UDP traffic was applied from CH to MH to investigate HBR performance compared with Mobile IP, in terms of number of packets dropped per handoff. The simulation environment was as described in Section 17.4.1. It could be expected that the results would be little impacted by varying the handoff rates, for reasonable handoff rates (not too large). Indeed, simulations verified this assumption. Recall that in the base case for both HBR and Mobile IP simulations (described in Section 17.4.1), the application data rate is 200 Kbps. For the base UDP simulations, this rate is accomplished by sending packets of 1,000 bytes every 40 ms. This can be described as a "heavy traffic" scenario, since the 375-Kbps links are more than 50% loaded. A "light traffic" scenario will also be investigated next, where the application data rate is 20 Kbps.

Some results for the simulations in the heavy traffic scenario are shown in Figure 17.8. As can be expected, increasing the link bandwidth would

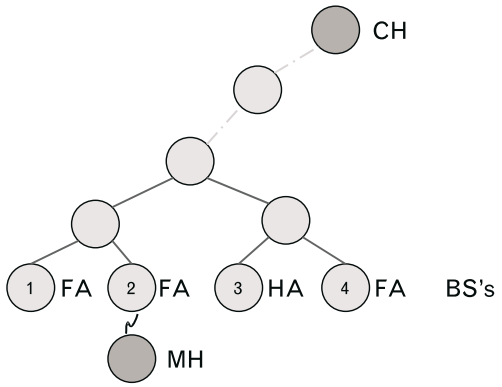


FIGURE 17.7
Base simulation setup for Mobile IP micro-mobility simulations.

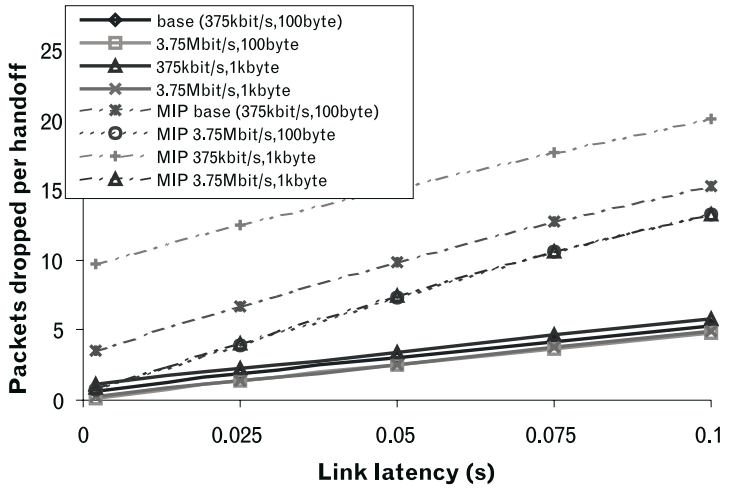


FIGURE 17.8
Packets dropped per handoff in a heavy-traffic scenario.

decrease the number of packets dropped per handoff, because the packets spend less time in transit. Similarly, increasing the size of the update packets would increase the number of packets dropped per handoff, as the update packets will take longer to arrive at their destination. These two effects were investigated by simulating micro-mobility domains with 10 times larger bandwidths, as well as cases of 10 times larger update packets. The plots marked "3.75 Mbps" have micro-mobility domains with link bandwidths of 3.75 Mbps, while the plots marked "1 kbyte" use large update packets 1 KB long. Curves whose labels are prefixed by "MIP" are those in which Mobile IP was run, and the rest use HBR by default (this statement applies to the rest of the performance results as well, not just this figure). The x-axis shows the link latency, which is the transmission latency of the straight-line links in the simulation setup. The *y*-axis shows the packets dropped per handoff.

Since this is a heavy traffic scenario, the number of packets dropped per handoff can be quite significant. However, it is the relative performance of HBR and Mobile IP that is of interest. With Mobile IP, the performance is worse (more packets dropped per handoff), even for this best-case Mobile IP scenario where the FA and HA are close together. Also noteworthy is the spread between the performance of the different cases when the link bandwidths are modified and/or the update packet size is modified. With HBR, the spread is very slight, from the best case (large link bandwidths and regular size update packets) to the worst case (regular size bandwidths and large update packets). With Mobile IP, the spread is much more pronounced, showing that it is much more sensitive to the settings of such variables.

Looking next at a light traffic scenario, two ways of reducing the traffic load are compared. First, reducing the size of the UDP packets 10-fold but keeping the rate of the packets at one every 40 ms can reduce traffic load

from 200 Kbps to 20 Kbps. Second, keeping the same packet size (1,000 bytes) but reducing the rate of the packets 10-fold can also reduce traffic load to 20 Kbps. As would be expected, the number of packets dropped per handoff in the first case would be significantly higher than in the second case. Indeed, this is the case, as illustrated in Figure 17.9. In this figure, "light" refers to the light traffic scenario, "sp" refers to short packets, and "lr" refers to low rate (the two ways to reduce the traffic load). It can be seen that for both HBR and Mobile IP, sp has more dropped packets per handoff. However, as in the heavy traffic scenario, the variation is greater for Mobile IP, in addition to the actual numbers being worse.

Having had a flavor of some simulation results, it is appropriate to address the issue of how generally the results can be interpreted. One question that arises is whether the results would be limited only to the unlikely case that an MH just moves back and forth between two BSs, rather than more realistic movement (e.g., randomly moving between all four BSs in the base simulation setup). This is a valid question, but the "to-and-fro" movement results are more generally useful because they can be extended to more general movement patterns according to the following methodology:

- Suppose the HBR domain is an n-tier domain, so the crossover node could be one level above the BSs, or up to n levels above the BSs.
- For each level from $i = 1$ to n, to-and-fro handoffs are simulated between any two BSs whose crossover node is i levels above the BSs. Let λ_i be the number of packets dropped per handoff.
- Given the HBR domain (or subregion within it) and mobility pattern, compute $E[h_i]$, the expected number of ith-tier handoffs, for each $i = 1$ to n, and define $h = \sum_{i-1}^{n} E[h_i]$.
- The overall expected number of packets dropped per handoff is then computed as the weighted average

$$\lambda = \sum_{i=1}^{n} \frac{E[h_i]}{h} \lambda_i \tag{17.1}$$

For example, for our base simulation setup, $n = 2$. For a handoff distribution that is uniform over the other three BSs, the crossover node would be two levels up for two of them and one level up for the other BS. Hence, the result λ_2 should be weighted by 2/3 and λ_1 by 1/3. It would be expected that the resulting value would be similar to what could be obtained by actually simulating handoffs under such conditions of random motion. Figure 17.10 shows the results. The curve labeled "1-up" is for to-and-from handoffs

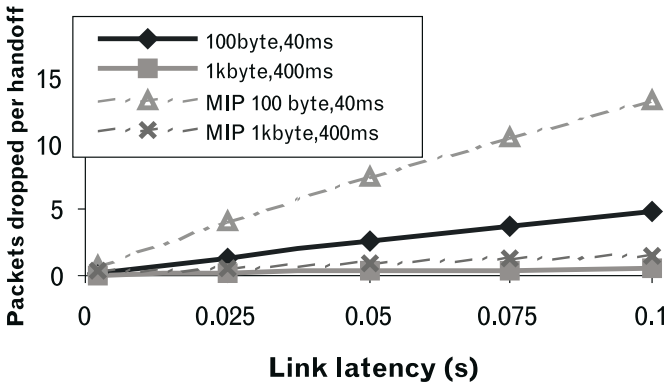


FIGURE 17.9
Packets dropped per handoff in a light-traffic scenario.

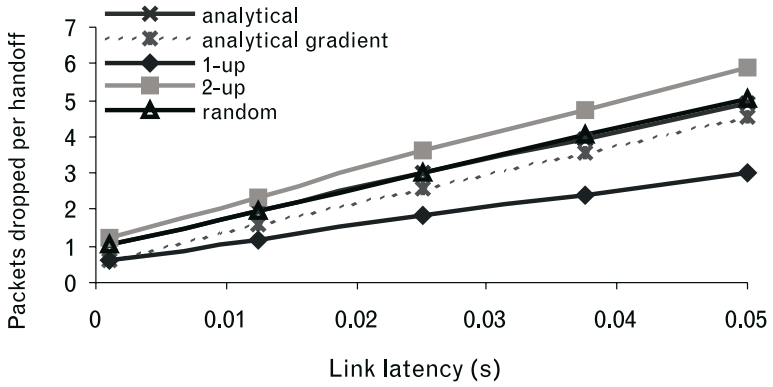


FIGURE 17.10
Comparing simulations of random handoffs with analytical model.

when the crossover node is one level above the BSs. The curve labeled "2-up" is similarly for to-and-fro handoffs when the crossover node is two levels above the BSs. The curve labeled "analytical" is where the performance in the more general random case is computed analytically according to (17.1). The curve labeled "random" is for the same case, but with results from actual simulations. It can be seen that "analytical" and "random" are almost the same curve, demonstrating that the analytical methodology works. One other curve can be seen in the figure, labeled "analytical gradient." This was obtained for the case where the 2-up simulation results were not used, but a 2-up scenario was emulated by simulating a 1-up scenario with double the link latency (up to 0.1 second instead of 0.05 second). The reason this underestimates the actual number of packets dropped per handoff for the random case is that it excludes the time for store-and-forward, and processing, at the intermediate nodes (just one such node in this case). Since this time is relatively constant, the offset of the resulting estimate from the real values is also roughly constant. Nevertheless, it at least provides the gradient of the correct curve and so is labeled "analytical gradient."

17.4.2.2 TCP Traffic

TCP throughput versus link latency is shown in Figure 1 7.11. "Base HBR" and "base MIP" are the results for the base simulations already described. "HBR semisoft" is for the case where the semi soft handoff scheme is used to further reduce handoff latencies. The figure shows the TCP throughput as link latency (of the straight-line links in Figure 17.6 and Figure 17.7) varies. HBR with semi-soft handoffs tolerates the most link latency and provides the highest throughput for any given link latency. Base Mobile IP shows a sharp deterioration in TCP throughput, which gets worse as the HA moves further away (not shown in this figure), which is more realistic.

The remaining curve in Figure 17.11 is labeled "flood MM." In this case, the root router floods the whole HBR domain with downstream packets, and upstream packets can arrive from any BS to the root router. This is not an HBR scheme, but is included to act as a performance bound. It is expected that no packets would be lost, since every BS is receiving the same stream, except for packets actually in the middle of being transmitted over the air when a handoff occurs. Notice that the throughput of flood MM starts to decrease at a link latency of just over 0.15 second. It may be conjecture that the reason is because that is where it runs into the so-called bandwidth-delay product bound. The TCP window size in the simulations is 20 KB, so the TCP "pipe" from sender to receiver can only hold that much data. Let link latency (in the HBR domain) be x seconds, and recall that the link latency of the "Internet" dash-dot links is 0.01 second. There are two of each type of link between the CH and MH. Therefore, the bandwidth-delay product from sender to receiver is (in kbits)

$$\beta = 2 \times 0.01 \times 1{,}544 + 2x \times 375 \tag{17.2}$$

In order to be within the window size constraint, it is necessary that $\beta < 160$, and so $x < 0.17.22$. It would be expected, however, that the

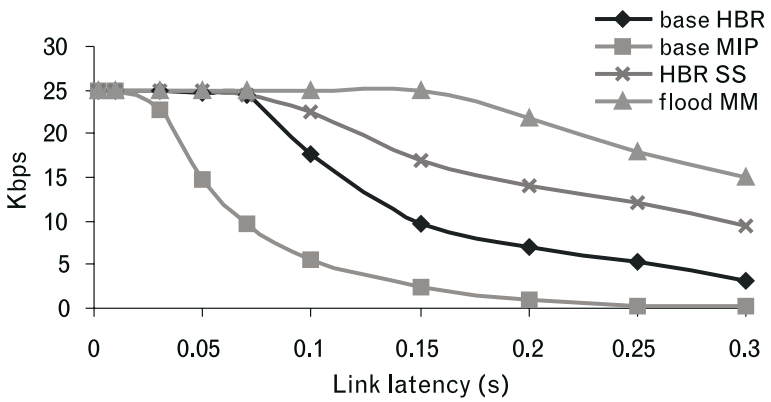


FIGURE 17.11
TCP throughput versus link latency.

throughput would start dropping for link latencies even a little less than 0.1722 second because the sender needs to allow time for acknowledgments from the receiver (ideally, therefore, it would need the bandwidth-delay product to be slightly less than the window size to allow room for it to keep sending without interruptions while waiting for the ACKs). Therefore, the performance of "flood MM" is reasonable and shows the bounds in performance for this scenario. It can be seen that HBR SS gets the closest to this bound. Furthermore, it can be expected that "flood MM" will not perform as well when there are multiple MHs, as flooding will affect other MHs the most, whereas HBR SS would still provide good results.

As in Section 17.4.2.1 for UDP, the results of the to-and-fro handoffs are also compared with that of actually simulating random handoffs, where the MH hands off from any BS with a uniform distribution to any of the other three. In Figure 17.12, the results are plotted for both HBR (with hard handoffs) and HBR SS (with semisoft handoffs). For "base HBR random" and "HBR SS random," the link latency is as given on the x-axis. For "base HBR to-and-fro" and "HBR SS to-and-fro," the actual link latencies used for the points on the curve are 5/3 the values on the *x*-axis.

This is because for the random handoff cases, the crossover node is expected to be two levels above the BSs two-thirds of the time and one level above the BSs one-third of the time; and through a similar reasoning process as for the UDP simulations, the 5/3 weighting factor for the to-and-fro simulations can be derived. Unlike the case of the UDP simulations, it is not expected that the curves will line up so well. The reason is that the number of packets dropped per handoff in the UDP simulations is linearly related to the handoff latencies, whereas the relationship for the TCP case is nonlinear. The results confirm this. Therefore, for TCP simulations, to-and-fro simulations may be useful to get an approximate understanding of performance for specific handoff situations like random handoffs, but actual simulations of the actual handoff situations are necessary to get specific performance results for specific scenarios.

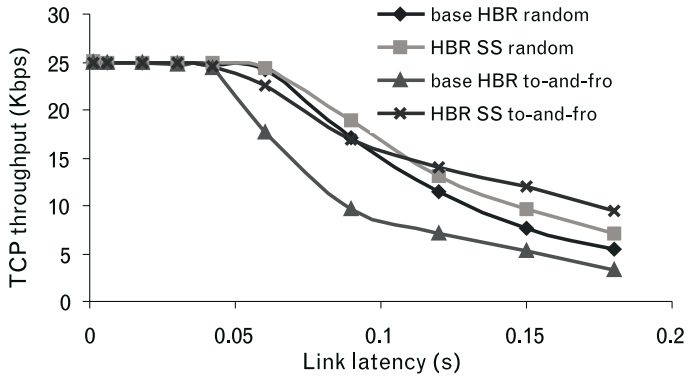
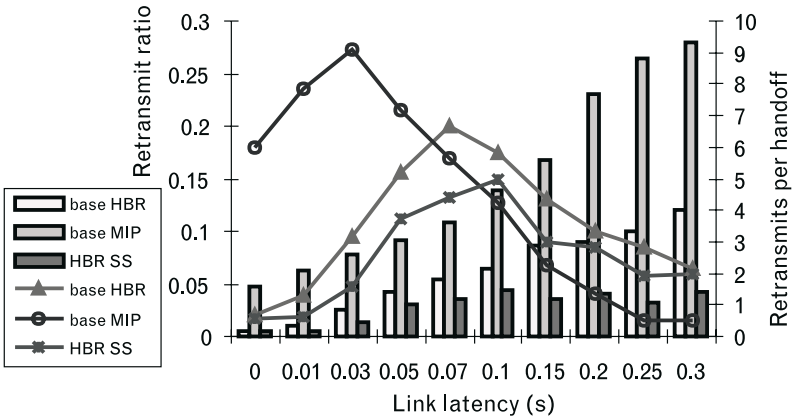


FIGURE 17.12
Comparing to-and-fro with random handoffs.

Although packets dropped per handoff is a useful performance indicator for UDP traffic, packet retransmission behavior is arguably more useful for TCP traffic. This is because throughput is more closely related to the packet transmission and retransmission behavior. Slight differences in number of packets dropped may result in timeouts occurring in one situation and not another, resulting in rather significant differences in packet retransmission behavior, and hence in throughput performance. The packet retransmission behavior for the different schemes is illustrated in Figure 17.13. For each of the three schemes (base HBR, base Mobile IP, and HBR SS), both the retransmit ratio and the number of retransmissions per handoff are plotted. The retransmit ratio refers to the ratio of retransmitted packets to total transmitted packets, and it is plotted as vertical bars with the values on the left *y*-axis. The values for the number of retransmissions per handoff, on the other hand, are plotted as regular curves, with the values on the right *y*-axis. The results for "flood MM" were also computed, but not plotted in the figure, because zero retransmissions occur throughout. Even the throughput decline for higher latencies is due to the bandwidth-delay product being constrained by the window size, not packet losses.

Looking at the retransmissions per handoff, it can be seen that all three curves follow the same basic pattern of increasing first, and then decreasing. The reason is that as the link latency increases, TCP is able to keep up with the increasing number of dropped packets by increasing the retransmissions and varying the instantaneous throughput so that can exceed 25 Kbps, allowing the average throughput to be still about 25 Kbps. There is a point, however, where TCP cannot "catch up" because of too many dropped packets, and the (average) throughput drops as a result. Since the average rate of packets is decreasing in this phase, the number of dropped packets per handoff, and retransmitted packets per handoff, also declines. It is interesting to note that Mobile IP has the worse performance in that the peak retransmission per handoff is the highest, and it occurs with the smallest link

FIGURE 17.13
Packet retransmission behavior.



latency before entering the declining throughput phase. Similarly, HBR SS performs the best in having the lowest peak occurring with the largest link latency. As for the retransmit ratios, these tend to increase as link latency increases. In the worst case, with Mobile IP, somewhere between a link latency of 0.2 second and 0.25 second, over one-fourth of the packets arriving at the MH are retransmitted packets!

In order to see how the TCP throughput is affected by the network latency between the HA and FA in Mobile IP, the latency on the dashed link in Figure 17.7 is varied from the base value of 2 ms to 100 ms to 500 ms. The resulting TCP throughput is shown in Figure 17.14, where the dashed link has a latency of 100 ms and 500 ms for Mobile IP 0.1 and MIP 0.5, respectively. These are not unreasonable values for Mobile IP, where HA and FA could be very far apart. The resulting degradation on performance is evident. The performance of "base HBR" is also included in the figure for reference. As expected, in this region where the link latency is 0.05s or less, HBR can achieve 25 Kbps throughput, unlike Mobile IP with the various latencies.

In the base simulations, the handoff rate is one handoff every 5 seconds on the average. To investigate the impact of different handoff rates, simulations were run (for HBR, HBR SS, and Mobile IP) with three average handoff intervals: 1 second, 5 seconds, and 10 seconds. The results are plotted in Figure 17.15. The solid lines show the performance of HBR. The dashed lines show the performance of HBR SS. The dotted lines show the performance of MIP. For each of these cases, the lines marked with crosses show the performance with handoff interval of 10 seconds, the lines with circles show the performance with handoff interval of 5 seconds, and the lines with triangles show the performance with handoff interval of 1 second. By looking at the three curves with crosses, it can be seen that the best performance is with the longer handoff intervals; whereas by looking at the three curves with triangles, it can be seen that the worst performance is with the shorter handoff intervals. This is expected because the more frequent the

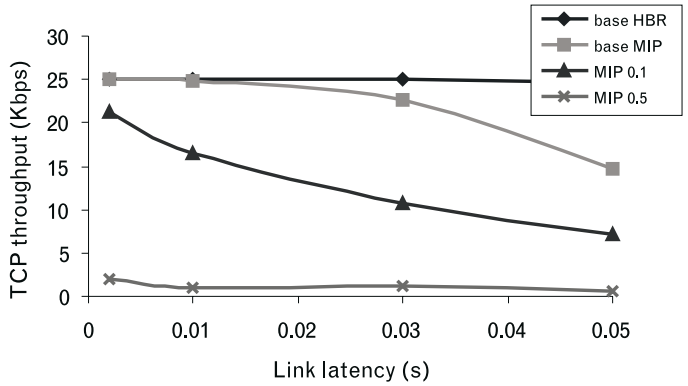


FIGURE 17.14
Performance degradation as latency between HA and FA increases.

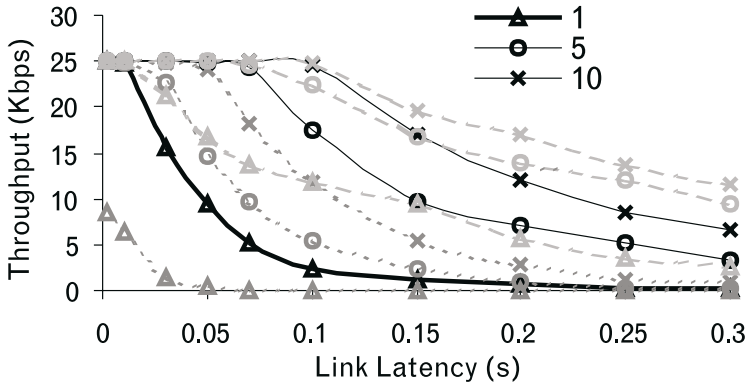


FIGURE 17.15
Effect of varying handoff rate.

handoffs, the more frequent the dropping of packets during handoff, resulting in more frequent transition of TCP into fast retransmit and slow start. It should also be noticed from the figure that for any given handoff rate, HBR SS performs best, followed by HBR, and then by Mobile IP.

Simulations were also run with TCP Reno. TCP Reno improves on the performance of TCP Tahoe when single packets are dropped, because of how it deals with congestion. In both cases, lost packets would result in duplicate ACKs being sent back to the transmitter, resulting in a retransmission of the lost packet (the fast retransmit algorithm). However, this is followed by slow start with TCP Tahoe, which can have a big impact on the TCP throughput. With TCP Reno instead, the fast recovery algorithm is used, going back to congestion avoidance instead of slow start. Despite performing better with single dropped packets, it has been previously found that TCP Reno performs poorly in general when multiple consecutive packets are dropped [15]. It was verified that this is the case for TCP Reno over HBR and Mobile IP as well.

Various other simulations were also run. These include cases where the basic simulation topology is modified, through vertical expansion (adding more layers of hierarchy) and horizontal expansion (adding more leaf nodes to each parent node). Also, cases of uneven link latencies have been explored. The results are similar to those with the basic simulation environment.

17.4.3 Results from Laboratory Prototype

The laboratory prototype used is shown in Figure 17.16. The purpose of the prototype was to confirm the simulation results with actual measurements performed on a prototype implementation. The nodes labeled "root node," "HBR BS1," and "HBR BS2" run the Linux (kernel 2.2.14) operating system, as do the MH and CH. The router is a standard Cisco router. The HBR prototype is based on the CIP version 1.1 software from Columbia

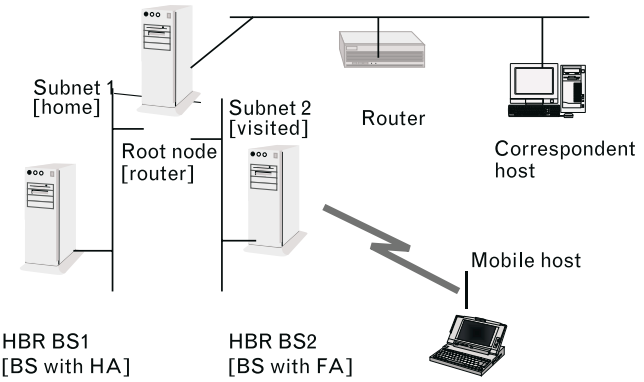


FIGURE 17.16
Laboratory prototype configuration.

University (http://comet.ctr.columbia.edu/cellularip), while the Mobile IP prototype is based on the Sun Laboratories Mobile IP prototype (https://playground.sun.com/pub/mobile-up/sunlabs).

If the HBR measurements are performed with one set of platforms and the Mobile IP measurements with another, then some irrelevant differences might creep in—for example, differences in processor speeds, cross-traffic, etc.—that would reduce the accuracy of the comparisons. In order to reduce irrelevant differences, both the HBR scheme and Mobile IP were therefore run on the same platforms and with the same hardware configuration and network connections. Switching back and forth between HBR and Mobile IP is a matter of typing a few commands to change a few interface configurations and routing table entries, and turn IP forwarding off (for HBR, the routing/paging caches are used instead) or on (for Mobile IP). Since the same hardware was used, both setups (HBR and Mobile IP) are shown on the same diagram (Figure 17.16). Where the functionality differs, the Mobile IP functionality is shown in square brackets beneath the HBR functionality (e.g., HBR BS1 becomes the BS with HA in the Mobile IP case).

TCP throughput between MH and CH was measured for both HBR and Mobile IP, using ttcp. Some of these results are shown in Figure 17.17. The values are averages over several measurements made at the ttcp receiving process, for CH to MH communications. The throughput of MH to CH traffic has also been measured and shows similar behavior. Throughput with Mobile IP for more than six handoffs per minute has not been included because the results become unstable in that region.

Measurements of TCP throughput were also made for traffic from the MH to the CH (Figure 17.18). The throughput degradations using Mobile IP with an increasing number of handoffs per minute are similar in this case as the previous. However, HBR performs better in this upstream direction. This is because even before the crossover node is aware of the handoffs, data packets following the handoff message are already taking the right path up to

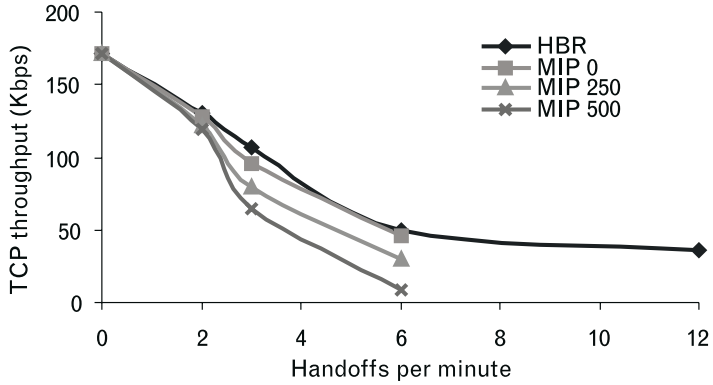


FIGURE 17.17
TCP throughput for data transfer from the CH to the MH.

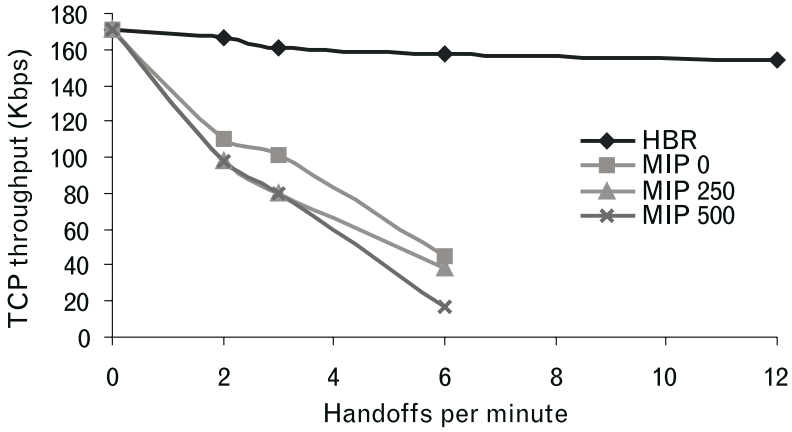


FIGURE 17.18
TCP throughput for data transfer from the MH to the CH.

the gateway (TCP acknowledgments may be lost during this time, though, accounting for the slight degradation in throughput as the handoff rate increases). Typically, the studies on HBR schemes focus on the downstream to the MH, because that is more critical for many applications (e.g., MH obtaining streaming video from a network server). It should be noted, however, that HBR schemes can perform even better in the upstream (this assertion is predicted on the assumption that data packets can immediately follow a route update packet). If security measures are in place that do not allow that, then upstream performance will be affected.

17.5 Conclusions and Future Directions

This chapter deals with HBR schemes, and in particular on how they are designed for reducing IP-level handoff latency to possibly orders of magnitude less than what may be experienced with pure macro-mobility schemes

based on Mobile IP and its variants, or SIP. This minimizes IP-level handoff disruptions, resulting in significantly fewer dropped packets and higher throughput. This chapter has provided an overview of the macro-mobility schemes and their shortcomings, especially where the shortcomings relate to handoff latency problems and the fact that they are not optimized to handle micro-mobility. An overview of two classes of micro-mobility schemes has been provided. The first class is the hierarchical Mobile IP-derived schemes like Mobile IP-RR and TeleMIP/IDMP. The second class is that of the HBR schemes. First, a generic HBR scheme is described that contains the essential features of the HBR schemes. Second, differences between actual HBR schemes like MMP, Cellular IP and HAWAII are explained. Some of the differences arise from differences in the design objectives, and the range of implementations of HBR schemes should provide network architects with sufficient flexibility to choose a micro-mobility solution that best fits a particular network. Third, comparisons are made between HBR domains and MANET. It was concluded that for certain types of very mobile networks, MANET ad hoc routing protocols must be used, but where a certain degree of fixed infrastructure exists that can support the HBR micro-mobility schemes, HBR would be preferred, because of reasons such as scalability and a more efficient distribution of routing information.

Next, performance issues of HBR micro-mobility schemes are discussed. In addition to providing the lowest latency handoffs of all the protocols discussed, the HBR schemes also (1) reduce signaling overhead through paging concepts, (2) improve bandwidth efficiency in the low-bandwidth radio access network by not using encapsulation in the radio access network, (3) facilitate QoS reservations by using an unchanging IP address while moving within the HBR domain, and (4) use IP address resources more efficiently than other schemes like Mobile IP with colocated COA. Results of simulations and laboratory prototype measurements are also reported. The goal of the performance studies reported in this chapter is to demonstrate the improvements of using HBR over Mobile IP for micro-mobility management. It is also explained how the Mobile IP results could be applied to hierarchical Mobile IP–derived schemes like MIP-RR, demonstrating that HBR performs better because of the lowest latency handoffs it provides. One implication of the simulation results is that using TCP over an IP micro-mobility management scheme magnifies the differences in handoff latencies between the schemes, because of the workings of the congestion control mechanisms like slow start.

Close to the top of the list of further simulations to do are simulations where there are multiple MHs in an HBR domain, where the traffic to/from other MHs would impact the performance of each MH. One expectation is that "flooding" micro-mobility management might not perform as well as it did in the results in this chapter (which is the best-case scenario for using flooding), because it would have the most impact on other

MHs, so it would not be practical. Further study is needed on transport protocols other than TCP, such as RTP and SCTP. Several modifications of TCP itself have been proposed for the problem of TCP interpreting errors and delays on wireless links as congestion [16]. The performance of such schemes over HBR micro-mobility schemes might be worth investigating.

This chapter has focused on the routing aspects of HBR schemes. However, it is very important to consider the implications of using HBR schemes together with schemes for QoS and security. Additionally, network management issues related to the HBR schemes should be investigated. HBR schemes may be more challenging to implement commercially than some other micro-mobility schemes, but they can be implemented slowly, in steps. Additional practical considerations are beyond the scope of this chapter. It is hoped, however, that the good performance of HBR schemes for IP micro-mobility management would serve as an incentive for further investigation in standards bodies like the IETF, and for engineers to work out the implementation issues.

References

[1] Wong, K. D., "Architecture Alternatives for Integrating Cellular IP and Mobile IP," *Proceedings of IEEE IPCCC 2002*, Phoenix, AZ, April 2002, pp. 197–204.

[2] Rpsenberg, J., et al., "SIP: Session Initiation Protocol," RFC 3261, IETF, June 2002.

[3] Dutta, A., et al., "Application Layer Mobility Management Scheme for Wireless Internet," *3G Wireless International Conference*, San Francisco, CA, May 2001.

[4] Schulzrinne, H., and E. Wedlund, "Application Layer Mobility Support Using SIP," *ACM Mobile Computing and Communications Review*, Vol. 4, No. 3, July 2000, pp. 47–57.

[5] Gustafsson, E., A. Jonsson, and C. Perkins, "Mobile IP Regional Registration," Internet Draft, IETF, March 2001, work in progress.

[6] Das, S., et al., "TeleMIP: Telecommunications Enhanced Mobile IP Architecture for Fast Intra-Domain Mobility," *IEEE Personal Communications Magazine*, Vol. 7, August 2000, pp. 50–58.

[7] Valko, A., "Design and Analysis of Cellular Mobile Data Networks," Ph.D. Dissertation, Technical University of Budapest, 1999.

[8] Ramjee, R., et al., "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks," available at http://www.bell-labs.com/user/ramjee/papers/hawaii.ps.gz.

[9] Dutta, A., et al., "Multilayered Mobility Management for Survivable Network," *Proceedings of MILCOM*, October 2001.

[10] Ramjee, R., T. La Porta, and L. Li, "Paging Support for IP Mobility Using HAWAII," Internet Draft, IETF, June 1999.

[11] Royer, E. M., and C.-K. Toh, "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," *IEEE Personal Communications*, April 1999, pp. 46–55.

[12] Zhang, X., et al., "P-MIP: Minimal Paging Extensions for Mobile IP," Internet Draft, IETF, July 2000, work in progress.

[13] Joa-Ng, M., and K. D. Wong, "IP Mobility Management for the ACN Platform," *Proceedings of MILCOM*, October 2000, pp. 465–469.

[14] Shelby, Z., et al., "Cellular IP Route Optimization," Internet Draft, IETF, June 2001, work in progress.

[15] Fall, K., and S. Floyd, "Simulation-Bbased Comparisons of Tahoe, Reno and SACK TCP," *Computer Communications Review*, Vol. 26, No. 3, July 1996, pp. 5–21.

[16] Balakrishnan, H., et al., "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links," *IEEE/ACM Transactions on Networking*, December 1997, pp. 756–769.