8

Wireless/Wireline Interworking

As the most pervasive wireline network, the Internet is the target wireline backbone network for the hybrid wireless/wireline network considered in this chapter. The Internet allocates a fixed address to each and every mobile user. All messages sent to a prescribed mobile user are first delivered to its permanent address located in the mobile's home network. The basic information transfer mechanism for the Internet is the Internet Protocol (IP), which resides in the network layer of the ISO (International Standards Organization) seven-layer reference model. As a protocol, IP is simple, but has no provision for traffic control. This means that IP can only offer best effort service. The current approach to forwarding messages from the mobile's fixed address to it in its new location is Mobile IP, introduced by IETF (Internet Engineering Task Force).

This chapter presents the salient features of the IP protocol, the operation of Mobile IP as an extension network to bridge information delivery to mobile users in their new locations, and TCP (transmission control protocol), in the transport layer, to oversee flow control for the IP-based network. We also present an overview of the Wireless Application Protocol (WAP), which has lighter overhead than TCP, and wireless ad hoc networks.

8.1 BACKGROUND

In wireless/wireline interworking, the front-end wireless segment provides a communications environment to support user roaming, while the backbone wireline segment extends the geographical

coverage. The Internet, which can provide global information delivery, is the most pervasive wireline network for use as the backbone network in wireless/wireline interworking. In this chapter, we consider the interworking of a wireless network with an IP-based network.

The Internet is a mesh connection of routers (nodes) by wirelines. The nodes at the edge of the Internet are called edge routers, while those in the network proper are called core routers. IP is used to route traffic among the Internet nodes; it is a simple and connectionless network layer protocol that provides a unique interface to diverse upper layer protocols. As a connectionless traffic transfer mechanism, IP handles the transfer of datagrams (e.g., packets). The main responsibility of IP is to address the end-hosts and to supervise the routing of datagrams to their destinations. The IP, which has no traffic control capability, is designed to provide *best effort* services only. To ensure network integrity, the Transmission Control Protocol (TCP) is introduced as a transport layer functionality to exercise flow control. TCP is a connection-oriented point-to-point transport layer protocol to provide end-to-end reliable and in-sequence data transfer over the best effort IP-based network. TCP uses a window mechanism to throttle the traffic at the edge router by shrinking the window size when necessary. A shrinkage of the window size at the edge router restricts the amount of traffic allowed to enter the network. Besides TCP, User Datagram Protocol (UDP) is also used to transfer datagrams in the transport layer. The difference between TCP and UDP is that UDP delivers packets from the source to the destination without any reliability or in-sequence guarantee.

A hybrid wireless/IP-based network supports both mobile and fixed hosts. From a mobility management perspective, it is more efficient to have a cluster of neighboring base stations, instead of a single base station, connected to the backbone network through a single attachment point. The cluster of base stations forms a subnetwork of the wireless segment. In the sequel, we will refer to these subnetworks as wireless networks. A pictorial view of this scenario is shown in Figure 8.1, where the wireless networks (home network and foreign network) are attached to the backbone IP-based network. The home network and the foreign network have different attachment

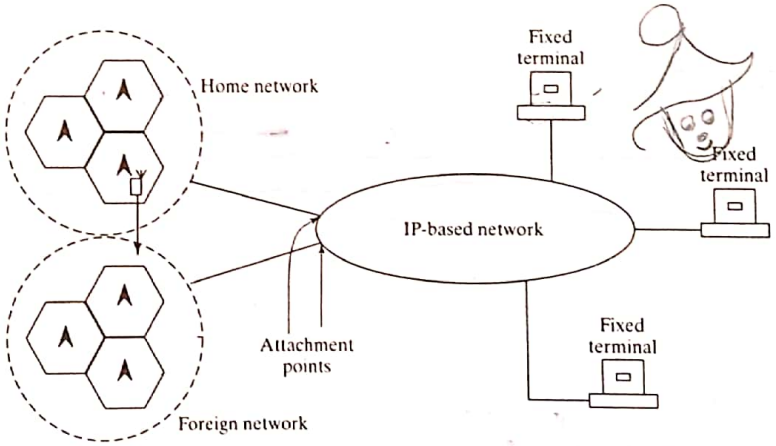

Figure 8.1 Hybrid wireless/IP-based network.

points. When a mobile signs on as a subscriber, it is allocated an IP address. The attachment point to the network where the mobile host initially resides houses the IP address. This network is the mobile's home network and the IP address is the mobile's home address. If the mobile moves to a foreign network without changing its IP address, it will be unable to receive information at the new site; if the mobile changes its IP address when it moves, it will have to terminate and restart any ongoing session.

In a wireless/Internet interworking environment, two of the most pressing problems are

a. the delivery of messages from the Internet to the mobile at its current location, and
b. traffic control to protect network integrity and to satisfy end-to-end quality of service (QoS) requirements.

The focus of this chapter is to address these two problems of wireless/Internet interworking. We first consider the problem of information delivery to the mobile in its current location in Section 8.2. We then describe the changes of IP version 6 from version 4, and the associated changes in the information delivery to mobiles in Section 8.3. The problem of traffic control to ensure end-to-end QoS satisfaction is studied in Section 8.4. End-to-end network performance is then evaluated in Section 8.5, taking into consideration the effect of traffic control. Wireless Application Protocol (WAP), with lighter overhead than TCP/IP, was introduced by the WAP Forum to facilitate Internet access from wireless terminals such as cellphones and portable terminals. The basic principle of WAP is discussed in Section 8.6. Section 8.7 gives an overview of mobile ad hoc networks.

The delivery of messages from the Internet server to the mobile user in its current location in a seamless manner is to be performed by a method referred to as Mobile IP.

8.2 MOBILE IP

When a mobile moves from its home network to a foreign network, messages from the Internet server for the mobile are still sent to the mobile's home address. A mechanism to allow the home network to forward the messages to the mobile in its new location is needed. The Internet Engineering Task Force (IETF) has proposed Mobile IP as an interface between the mobile's home network and the foreign network where the mobile currently resides [112, 15].

Mobile IP is a protocol that keeps track of the mobile's whereabouts and delivers Internet messages to the mobile at its current location. The operation of Mobile IP is enabled by the following functional entities:

Mobile node is a host or a router which can travel around the Internet while maintaining any ongoing communication session. In this text, the terms *mobile node* (MN), *mobile host* (MH) and *mobile station* (MS) are used interchangeably. The reason for this is that these terms are being used quite freely in the literature. A mobile node has a home address, which is a long-term IP address residing in its home network. When away from its home network, the mobile node is assigned a *care-of address* which reflects the mobile node's current point of attachment.

Correspondent node is a peer host with which a mobile node communicates.

Home address is an IP address that is assigned for an extended period of time to a mobile node. It remains unchanged regardless of where the node resides in the wireless segment.

Care-of address is the termination point of tunneling datagrams destined to a mobile node while it is away from home.

Collocated care-of address is an externally obtained local IP address temporarily assigned to an interface of the mobile node.

Home agent is a router with an interface on the mobile node's home network link, which the mobile node keeps informing of its current location, *care-of address*, as the mobile node moves from link to link. The home agent can intercept packets destined to the mobile node's home address and tunnel them to the mobile node's current location.

Foreign agent is a router with an interface on a mobile node's visiting network, which assists the mobile node in informing its home agent of its current care-of address.

Foreign agent care-of address is an IP address of a foreign agent, which has an interface on the foreign network being visited by a mobile node. A foreign agent care-of address can be shared by many mobile nodes simultaneously.

Home network is a network having a network prefix matching that of a mobile node's home address.

Foreign network is a network other than a mobile node's home network to which the mobile node is currently connected.

Virtual network is a network with no physical instantiation beyond its router. The router usually uses a conventional routing protocol to advertise reachability to the virtual network.

Link is a facility or medium over which nodes can communicate at the link layer.

Link-layer address is an address that identifies the physical endpoint of a link. Usually, the link-layer address is the interface's Medium Access Control (MAC) address.

Mobile node's home link is the link which has been assigned the same network-prefix as the network prefix of the mobile node's home address.

Mobile node's foreign link is the link that the mobile node is visiting, which has been assigned the same network prefix as the network prefix of the mobile node's care-of address.

Agent advertisement is the process in which foreign agents advertise their presence by using a special message.

Agent solicitation is the message sent by a mobile node to request agent advertisement.

Tunnel is the path followed by a datagram while it is encapsulated.

Binding entry is an entry in the home agent's routing table. Mobile IP maps the mobile node's home address into its current care-of address.

Messages from the Internet, destined for the mobile, are always sent to the mobile's permanent address in the mobile's home network. The Mobile IP interface is designed to deliver Internet messages from the mobile's home network to the mobile in its current location in a seamless manner. In Mobile IP, the routing of messages from the mobile's home network to the mobile in its current location is accomplished by allowing each mobile to have two IP addresses: a fixed *home address* for identification and a *care-of address* for routing. The home address remains unchanged regardless of where the mobile resides in the wireless segment. However, the care-of address changes at different access points.

In summary, Mobile IP uses an *agent* concept. The mobile has a *home agent* (HA) and a *foreign agent* (FA). The home agent maintains a database in which the mobile's home address

resides. When the mobile moves to a foreign network, it establishes an association with its foreign agent which, in turn, establishes an association with the mobile's home agent. That is, the mobile updates its registration with its home agent through the foreign agent. The registration updating procedure is similar to that described in Chapter 7. The operational features of Mobile IP are described in the next subsection.

8.2.1 Operation of Mobile IP

Figure 8.2 shows the functional relationships among the different entities in Mobile IP. The operation of Mobile IP is as follows. The home and foreign agents make themselves known by sending agent advertisement messages. After receiving an agent advertisement, the mobile determines whether it is in its home network or in a foreign network. The mobile basically works like any other node in its home network when it is at home. It routes packets using traditional IP routing protocols. When the mobile moves away from its home network, it obtains a care-of address on the foreign network by soliciting or listening for agent advertisements. The mobile node registers each new care-of address with its home agent, possibly by way of a foreign agent. Datagrams sent to the mobile node's home address are intercepted by its home agent, tunneled

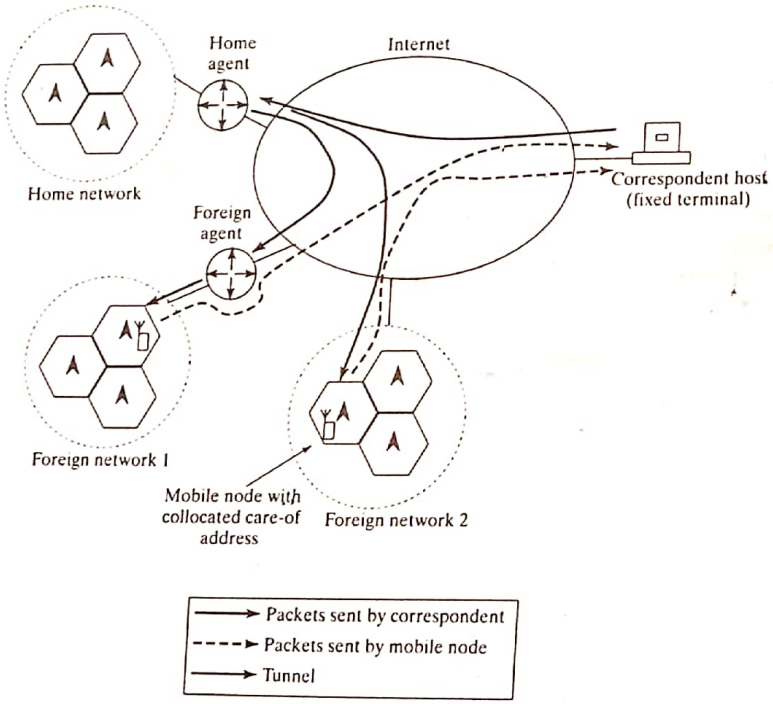

Figure 8.2 Entities in Mobile IP.

by its home agent to the care-of address, received at the tunnel endpoint (at either a foreign agent or the mobile node itself), and finally delivered to the mobile node. In the reverse direction, datagrams sent by the mobile node are generally delivered to their destination using standard IP routing mechanisms. The operation of Mobile IP is based on the cooperation of three major processes: agent discovery, registration, and tunneling (routing).

Agent Discovery is a process by which a mobile node determines its new attachment point or IP address as it moves from place to place within the wireless segment of the wireless/IP network. By agent discovery, a mobile node can (*a*) determine whether it is connected to its home link or foreign link, (*b*) detect whether it has changed its point of attachment, and (*c*) obtain a care-of address if it is connected to a foreign link. The mobile node identifies whether it is connected to the home or foreign link from agent advertisements sent periodically by agents (home, foreign or both) as multicasts or broadcasts to the link. In case a mobile node does not receive any agent advertisement, or it does not have the patience to wait for the next agent advertisement, the mobile node can send an agent solicitation to request an agent advertisement from the agent to which it is attached. When a mobile node is connected to its home link, it works exactly as a traditional node in a fixed place. When a mobile node detects that it has moved, it acquires a care-of address by reading it directly from an agent advertisement, or contacting Dynamic Host Configuration Protocol (DHCP), or using the manual configuration. Registration follows once the mobile node gets a new care-of address.

Registration is a process performed as a mobile node enters and remains in a foreign link. This process involves requesting services from a foreign agent and informing the home agent of a mobile node's new care-of address. Registration also involves reregistration upon expiration of a current registration and deregistration as the mobile node returns to its home link. Some of the characteristics of registration include having multiple, simultaneous care-of addresses and the ability to remove any number of them while retaining others. Registration consists of an exchange of two messages, a registration request and a registration reply, between the mobile node and its home agent, possibly involving a foreign agent as well, depending on the type of the mobile node's care-of address. While an agent discovery message is carried by the Internet Control Message Protocol (ICMP) payload portion, the registration message is carried by the User Datagram Protocol (UDP).

Tunneling (routing) is a process by which Mobile IP tunnels datagrams to the mobile node, whether it is or it is not away from its home network.

Example 8.1 Registration and Deregistration

When a mobile user moves to a visiting location, it has to register with its home agent. When the mobile returns to its home network, it also has to deregister with its home agent. Describe and illustrate with diagrams, using the concepts of care-of address and collocated care-of address, how registration and deregistration can be performed.

Solution

a. When the mobile host moves to a new foreign (serving) agent, it must register with its home agent using the foreign agent's care-of address. This registration process involves

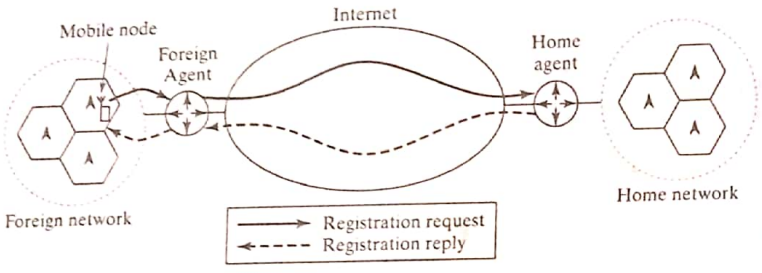

Figure 8.3 Registration with foreign agent's care-of address.

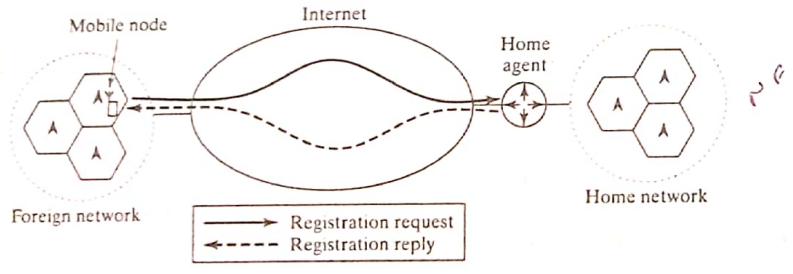

Figure 8.4 Registration with collocated care-of address.

the foreign agent. Figure 8.3, which shows both registration request and registration reply, illustrates the registration process with the foreign agent's care-of address.

b. The mobile host may register using a collocated care-of address. The collocated care-of address is an externally obtained local IP address temporarily assigned to an interface of the mobile host. As such, registration can be performed without the help of a foreign agent. Registration with a collocated care-of address is illustrated in Figure 8.4.

c. When the mobile host returns to its home network, it should deregister with its home address. This is done locally, without the involvement of the Internet. Figure 8.5 illustrates the deregistration process.

Registration can also serve as a means for a new mobile node to obtain the address of a home agent as it initially configures itself for Mobile IP. Registration in Mobile IP must be made secure so that fraudulent registrations can be detected and rejected. Otherwise, any malicious user on the Internet could disrupt communications between the home agent and the mobile node by the simple expediency of supplying a registration request containing a bogus care-of address.

Tunneling. Tunneling is a procedure in which the home agent encapsulates the message from the IP host for delivery to the mobile via its foreign agent. The encapsulation process involves

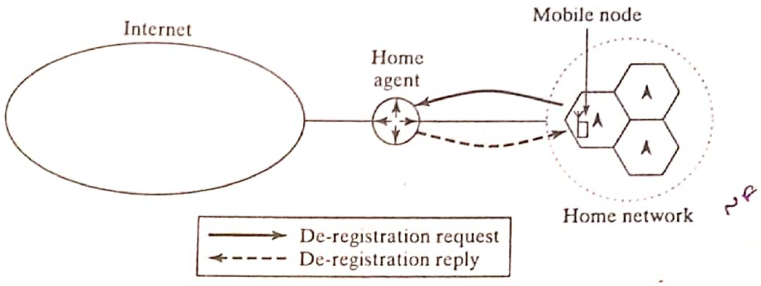

Figure 8.5 Deregistration.

shielding the inner IP header destination address (i.e., the mobile's home address) from intervening routers between the mobile's home network and its current location.

Mobile IP has proposed two routing approaches: *triangle routing* and *optimized routing*. The operational features and the merit and demerit points of triangle routing and optimized routing are as follows.

Triangle Routing A pictorial view of triangle routing is shown in Figure 8.6. The correspondent host (CH) is a fixed host connected to the Internet and communicates with the mobile host through the Internet. Datagram delivery between the correspondent host and the mobile node is performed using the following steps:

(1) A datagram from the correspondent for the mobile is sent to the mobile's home network using standard IP routing.

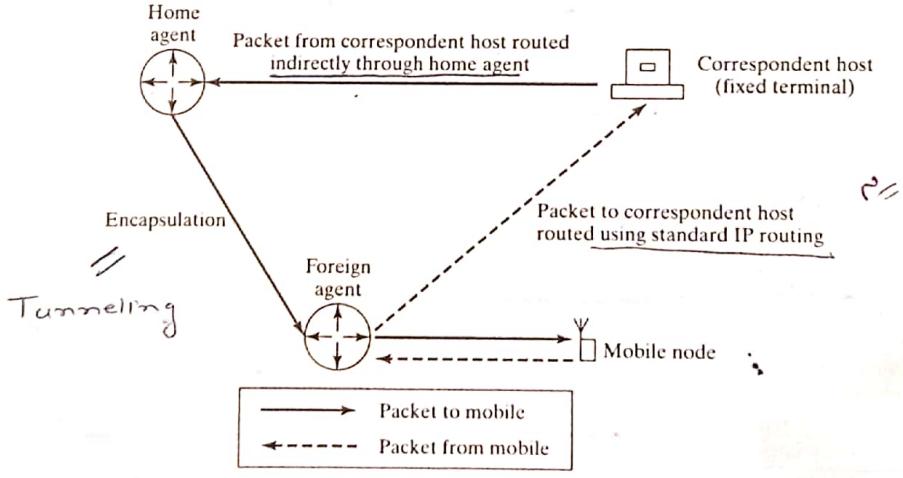

Figure 8.6 Triangle routing.

(2) Upon arrival at the home network, the datagram is intercepted by the home agent which, in turn, tunnels the datagram to the mobile's care-of address.
(3) At the foreign agent, the datagram is detunneled and delivered to the mobile.
(4) For datagrams sent by the mobile, standard IP routing is used to deliver each datagram to its destination. Note that the foreign agent is the mobile's default router.

The Mobile IP protocol with triangle routing is simple, the number of control messages to be exchanged is limited, and the address bindings are highly consistent since they are kept at one single point for a given host. One of the drawbacks of triangle routing is that the destination home agent is a fixed redirection point for exchanging every IP packet, even if a shorter route is available between source and destination. This can lead to unnecessarily large end-to-end packet delay. The other drawback is that the network links connecting a home agent to the network can easily be overloaded. Indeed, all session paths sharing the subnet field of their destination address converge into that subnet home agent, even if adjacent network links are idle.

Optimized Routing In optimized routing, the mobile host informs the correspondent host of its care-of address and has the packets tunneled directly to the mobile host, bypassing the home agent. The Mobile IP protocol with optimized routing allows every traffic source to cache and use binding copies. It supports a further update process by which a binding copy can be sent to the requiring nodes, which may keep it in their cache for immediate or future use. Local bindings enable most packets in a traffic session to be delivered by direct routing, with an apparent gain in terms of quality of service and scalability. In addition, a moving host can always inform its previous foreign agent about the new care-of address, so that packets tunneled to the old location (owing to out-of-date binding copy) can be forwarded to the current location. This should increase the overall quality of service in the case of high mobility. The disadvantage of the protocol is that it is quite complex, and the overhead incurred by message exchanges and processing (due to cache queries) can be critical. Also, cached bindings are possibly inconsistent since they are being kept in a distributed fashion. The main obstacle to implementing optimized routing resides in the security issues. The correspondent node must be informed of the mobile host's care-of address in order to tunnel data to the mobile host. In a hostile environment, an intruder can easily cut off all communications to the mobile host by sending a bogus registration if he/she knows the mobile's care-of address. Therefore, authentication/security measures have to be incorporated in the optimized routing.

Triangle routing is much simpler than optimized routing. In many cases, this is the preferred mode of routing datagrams from the correspondent host to the mobile host. However, as the mobile moves further and further away from its home network, the cost (delay) involved in the registration with the home agent can become prohibitively large. Methods to reduce registration costs are desirable. One approach is the incorporation of a *local anchor* to act as a *virtual home agent* such that the mobile host only needs to register with the close-by virtual home agent instead of the far away home agent.

The concept of local anchor as a means of reducing registration cost is introduced in Subsection 7.5.2. The same concept can be used in conjunction with the operation of Mobile IP.

8.2.2 Local Anchor for Mobile IP

In Chapter 7, we define a registration area within which the mobile does not have to register with its home location register (HLR). Here, we define an anchoring region within which the mobile only needs to register with the local anchor. The local anchor strategy can be described as follows:

a. Choose one agent as the focus of an anchoring region and name this agent as an anchor.
b. When the mobile moves within the anchoring region, it does not need to register with its home agent; instead, it registers with the anchor. That is, the local anchor acts as a virtual home agent.
c. When the mobile moves out of the anchoring region, it will register with its home agent and the new foreign agent will become the focus of the new anchoring region.

It is noted that the local anchor strategy only reduces registration costs, but has virtually no impact on delivery, since messages from the correspondent host are still sent to the mobile's home address, to be intercepted by the home agent and then tunneled to the mobile's care-of address. Thus, packets destined for the mobile will be forwarded from the home agent to the anchor agent first and, from there, to the foreign agent where the care-of address resides. This mode of packet forwarding is illustrated in Figure 8.7.

The anchor agent and the new foreign agent are two candidates that can decide whether the mobile should register with its home agent or not. The mobile does not have the knowledge of the network to make this decision. The decision making process can be based on static or dynamic information. The dynamic approach will have to use the mobile's past movement information and its current traffic information for decision making. The static approach makes use of information that is fixed for all time. For the purpose of establishing a new anchoring region, we will use the static approach, and use the distance from the old anchor agent to the new foreign agent as the criterion to decide whether or not to establish a new anchoring region.

The registration process in the local anchor scheme is shown in Figure 8.8. Depending on which agent (the home agent or the current anchor) the mobile should register with, the

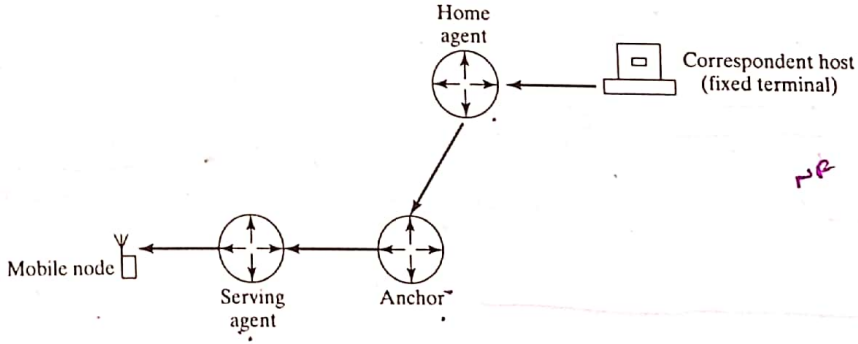

Figure 8.7 Packet forwarding in local anchor approach.

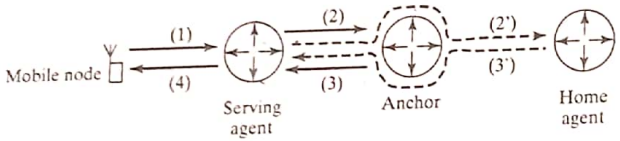

Figure 8.8 Registration in local anchor scheme.

registration process needs to consider two cases. The following example sets up the procedure for the registration process.

Example 8.2 Registration with Local Anchor

Using a static approach, construct the steps needed for the mobile to register, either with the anchor agent or the home agent. Assume that the new foreign agent takes responsibility for making the decision regarding which agent the mobile should register with.

Solution The following four steps constitute a procedure for the mobile to update its registration.

(1) The mobile sends the registration request indicating the current anchor agent and the home agent.
(2) There are two cases–
Case I: the new foreign agent decides that the mobile is still in its current anchoring region so it forwards the mobile's registration request to the anchor.
Case II: the new foreign agent decides that the mobile is out of its current anchoring region, so it forwards the mobile's registration request to the home agent.
(3) The anchor agent or the home agent sends the registration reply back to the foreign agent.
(4) The foreign agent returns an acknowledgment to the mobile and indicates who, the anchor or the home agent, sends this registration reply. In Case I, the mobile knows that it has not moved out of the current anchoring region and the anchor does not change. In Case II, the foreign agent becomes the focus of the new anchoring region and the mobile will update its anchor agent's IP address for later use.

8.2.3 Hierarchical Routing

For highly mobile users, the amount of registration traffic generated between the visited and the home networks can be quite large. In the preceding subsection, we have considered a local anchor approach to reduce the amount of registration with the home agent. The amount of registration between the home and the visited networks can also be reduced by using a hierarchical routing strategy. In hierarchical routing, a hierarchy of foreign agents is established in a tree structure, and multiple foreign agents are advertised in the agent advertisement. In this way, registrations can be localized to the foreign agent that is the lowest common ancestor of the care-of addresses at the two points of attachment of interest. To enable this, the mobile has to determine how high up the tree its new registration has to go, and then arrange for the transmission of the registration to each level of the hierarchy between itself and the closest common ancestor between its new and previous care-of addresses.

Example 8.3 Hierarchical Routing

Suppose that a mobile currently using the services of one foreign agent is migrating to use the services of a different foreign agent. If the foreign agents are hierarchically connected in a binary tree structure, a mobile moving from one foreign network to another foreign network may not involve a direct registration with its home agent.

a. Draw a binary tree, with a population of ten foreign agents, connected to the Internet as a subnetwork and, hence, to the mobile's home agent through the Internet.
b. Describe how the mobile moving from one serving foreign agent to another foreign agent will receive agent advertisements, and the situations under which the mobile does not have to register with its home agent.

Solution

a. The binary tree structure for a population of ten foreign agents is illustrated in Figure 8.9, where the mobile moves from location A to location B and then to location C. The attachment point to the Internet is FA_1. The home agent only directly "sees" foreign agent FA_1, while all the other foreign agents in the binary tree are not visible to the home agent.

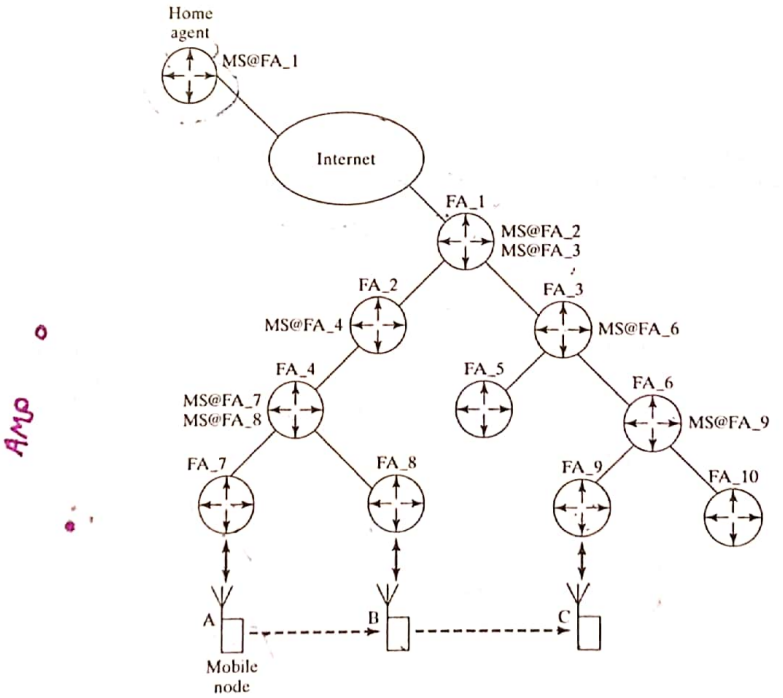

Figure 8.9 Hierarchical foreign agents.

b. As shown in Figure 8.9, the mobile is using the service of FA_7 while at location A. But it also receives agent advertisement from FA_4, FA_2 and FA_1. A registration is transmitted to each of these foreign agents as well as its home agent. Since an agent only "sees" its nearest neighbor in the hierarchy, the home agent believes that the mobile is located at the care-of address of FA_1; FA_1 believes that the mobile is located at FA_2, and so on, until foreign agent FA_7, which actually knows the whereabouts of the mobile. For illustration purposes, consider the following two scenarios.

   (1) When the mobile moves to FA_8 (at location B), it only has to cause the new registration to travel as far as FA_4.

   (2) When the mobile moves to foreign agent FA_9 (at location C), it receives advertisements which indicate the lineage of FA_9, FA_6, FA_3 and FA_1. By comparing the previous and the current lineages, the mobile determines that it has to cause the registration to travel up the hierarchical tree to foreign agent FA_1, but the registration still does not have to reach the home agent.

Note that the original datagram must be relayed to a number of intermediate nodes in the hierarchy. Each of the nodes is then charged with the responsibility of retunneling the datagram, if necessary, to the next lower level in the hierarchy.

8.3 INTERNET PROTOCOL (IP)

Until recently, Internet Protocol version 4 (IPv4) has been the protocol which supports Internet administration and operation. IPv4, with network entities such as mobile nodes, home agent, home address, foreign agent, care-of address, and the like, is the basis for the original development of Mobile IP. The basic IPv4 header, with a 32-bit source address and a 32-bit destination address, is shown in Figure 8.10. With only a 32-bit source address and a 32-bit destination address, the address allocation scheme in IPv4 is insufficient to support the rapidly growing Internet subscriber population. Since 1994, the IETF has been working on Internet Protocol version 6 (IPv6) to provide a remedy to the limitations inherent in IPv4, in terms of addressing, routing, mobility support, quality of service (QoS) provisioning, and so forth. The network entities of IPv6 are similar to those of IPv4, except that IPv6 does not have the concept of a foreign agent. With a 128-bit source address and a 128-bit destination address, IPv6 now supersedes IPv4. The basic IPv6 header is shown in Figure 8.11.

8.3.1 IPv6 versus IPv4

As the successor to IPv4, IPv6 can be installed as a normal software upgrade in Internet devices and is interoperable with IPv4. Also, IPv6 can run well on high-performance networks (e.g., ATM, fast Ethernet, and the like), and is efficient for low-bandwidth networks. The changes from IPv4 to IPv6 fall mainly into the following categories [77].

Expanded Routing Addressing Capabilities. With a 32-bit IP address in IPv4, which can hold up to $2^{32} - 1$, or over 4 billion hosts, one might think this address range is more than enough

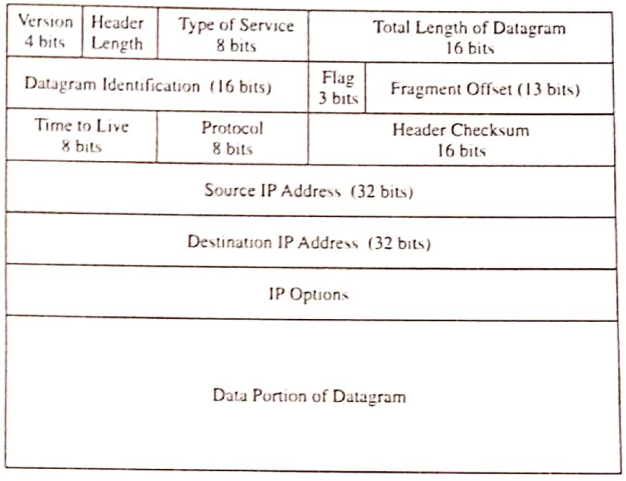

Figure 8.10 IPv4 header.

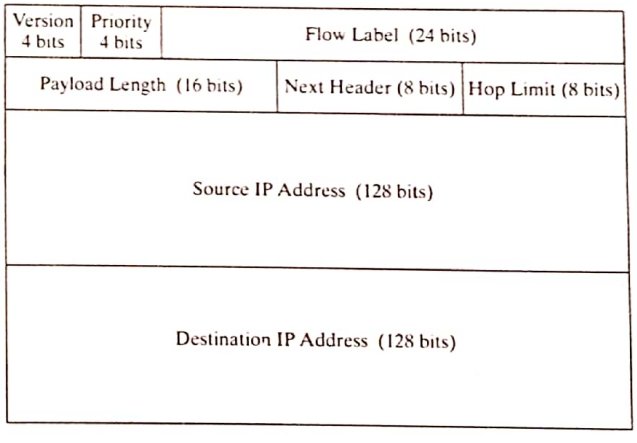

Figure 8.11 IPv6 header.

to support the addresses needed on the Internet. However, with this address space, one can easily run out of available addresses, as explained in the following:

a. Although the traditional two-level addressing scheme, with a network prefix and a host address, is convenient, there is a waste of address space. Once a network address is assigned to a particular network, a block of IP address is assigned to that network. Any improper assignment of network address will lead to a great waste of available IP addresses.

b. Many private networks, which are currently not connected to the Internet, are reusing IP addresses used by the public network or other private networks. They require many more IP addresses when connected to the Internet.
c. There may be devices other than the traditional hosts, possibly both wireless and wireline products, such as mobile telephones, wireless organizers, and the like, which need additional IP addresses to make themselves identifiable on the Internet.

The increase of IP address size from 32 bits to 128 bits in IPv6 allows the Internet to support more levels of addressing hierarchy, a much greater number of addressable nodes, and simpler autoconfiguration of addresses. The traditional two-level IP address structure, network address and host address, is modified. Unicast address, anycase address and multicast address are defined in IPv6. The meaning of unicast, multicast and anycast addresses are as follows:

Unicast address is simply a 128-bit network node address. Unicast address can be divided into two parts: a subnet prefix, which indicates the node's subnetwork; and an interface ID, which indicates the node's interface.

Multicast address in IPv6 replaces broadcast in IPv4. Multicast address is divided into two groups: the predefined groups, which are permanently assigned; and the transient groups, which are defined by specific organizations. The most common predefined multicast addresses are: All Nodes (all nodes connected), both routers and hosts; All Routers (not including hosts); and All Hosts (not including routers).

Anycast address is a new type of address defined to identify sets of nodes such that a packet sent to an anycast address is delivered to any one of the nodes assigned that address. Packets destined to a multicast address are sent to all nodes in that group, while packets destined to an anycast address are sent to only one node in that group. The use of anycast address in IPv6 allows nodes to control the path along which their traffic flows. The scalability of multicast routing is improved by adding a "scope" field to the multicast address.

There will be coexistence of both IPv6 addresses and IPv4 addresses and, also, it is impossible to replace all IPv4 routers with IPv6 routers. In IPv6, special addresses, called IPv4-compatible-IPv6 addresses, are introduced for assignment to those hosts and routers running IPv6, but which must route traffic across IPv4 networks. On the other hand, IPv4-mapped-IPv6 addresses are assigned to those hosts running IPv4.

Header Format Simplification. The IPv6 header has a fixed length of 40 octets. Some IPv4 header fields have been dropped, or made optional, to reduce the common-case processing cost of packet handling and to keep the bandwidth cost of the IPv6 header as low as possible, despite the increased size of the addresses. IPv6 extension headers, optional parts following the IPv6 header, are defined to carry additional information about the traffic being sent. Extension headers include

a. hop-by-hop option header, which defines special options that require hop-by-hop processing;
b. fragment header, which contains fragmentation and reassembly information;
c. destination options header, which contains optional information to be examined by the destination node;
d. routing header, which provides extended routing; and
e. authentication header, which provides packet integrity and authentication.

Improved Support for Options. Changes in the way that IP header options are encoded allow for more efficient forwarding, less stringent limits on the length of options, and greater flexibility for introducing new options in the future.

Priority. A 4-bit priority field is introduced for the source node to indicate the desired transmit and delivery priority of every packet relative to other packets from the same source. Traffic types are first classified as congestion-controlled traffic or non-congestion-controlled traffic; then one of eight levels of relative priority is assigned to each type of traffic.

Congestion-controlled traffic refers to traffic which can tolerate congestion, or delay. If network congestion happens, congestion-controlled traffic will be buffered or "backed off". A variable amount of packet delay or even out-of-order packet arrival is acceptable. IPv6 defines the following types of congestion-controlled traffic, in decreasing priority: Internet control traffic, interactive traffic, attended bulk transfer, unattended data transfer, filler traffic, and uncharacterized traffic.

Non-congestion-controlled traffic refers to that traffic which requires constant data rate, constant delivery rate (or at least relatively smooth data rate), or delivery delay. Examples of non-congestion-controlled traffic are real-time audio and real-time video.

Quality-of-Service Capabilities. A new capability is added to enable the labeling of packets belonging to particular traffic flows for which the sender requests special handling, such as voice or video. A flow is basically a series of packets originated by the source and having the same transmission requirements. No special transmission process is assigned to any particular flow label. A source must specify, or negotiate, what kind of special handling is requested before a flow is transmitted, possibly by means of other Internet control protocol. Flow labels are assigned pseudorandomly to ensure that there is no flow label reuse during the lifetime of that flow label.

Security Capabilities. IPv6 will support the five proposed security-related standards published by IETF. These security features, which are optional, are

RFC 1825 - Security Architecture for the Internet Protocol,
RFC 1826 - IP Authentication Header,
RFC 1827 - IP Encapsulating Security Payload (ESP),
RFC 1928 - IP Authentication Using Keyed MD5, and
RFC 1829 - The ESP DES-CBC Transform.

Two IP security mechanisms, security association and authentication, are combined to transmit IP packets that require both privacy and authentication. IPv6 includes the definition of extensions that provide support for authentication, data integrity, and confidentiality. This is included as a basic element of IPv6 and will be included in all implementations.

8.3.2 Mobile IPv6

Current activities by IETF on Mobile IP capture the salient features of IPv6. Mobile IPv6 uses the new and improved IPv6 *Routing Header*, along with the *Authentication Header*, and other pieces of IPv6 functionality to simplify routing to the mobile node and to perform route optimization

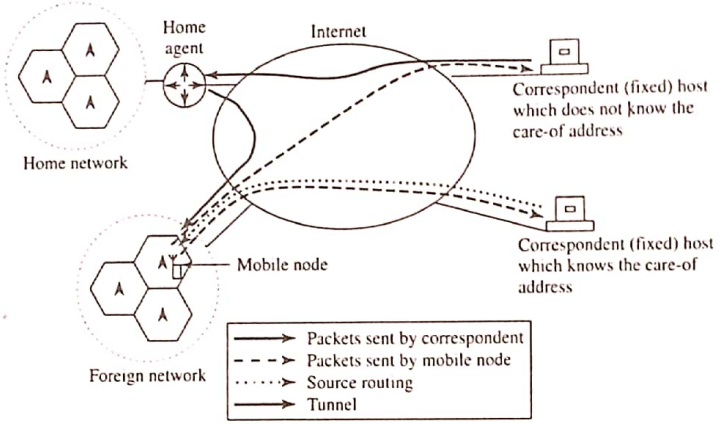

Figure 8.12 Mobile IPv6 operation.

in a secure manner. Mobile IPv6 has no foreign agent. The mobile node uses the *Address Autoconfiguration* procedure defined in IPv6 to acquire a collocated care-of address on a foreign link, and reports its care-of address to its home agent and selected correspondents. The operational procedure of Mobile IPv6 is depicted in Figure 8.12.

In Mobile IPv6, a correspondent node which knows the mobile's current care-of address can send packets directly to the mobile node by using an IPv6 Routing Header. Those correspondent nodes who do not have this information send packets without such a header. The packets are routed to the mobile node's home link, intercepted by the home agent, and tunneled to the mobile node's care-of address. When the mobile node returns to its home link, it notifies its home agent.

Mobile IPv6 has almost the same terminologies as Mobile IP (also called Mobile IPv4) discussed in Section 8.2, except for the absence of foreign agent. The concept of home agent, home link, care-of address and foreign link are roughly the same as in Mobile IPv4. Compared with Mobile IPv4, Mobile IPv6 has the following advantages:

a. The enormous address space in IPv6 allows very simple address autoconfiguration by means of Stateless Address Configuration (SAC). Because the mobile node can easily obtain a collocated care-of address by SAC, the foreign agent functionality is no longer needed. As a result, the foreign agent is eliminated from Mobile IPv6. This also implies that all Mobile IPv6 care-of addresses are collocated care-of addresses.
b. Mobile IPv6 uses the new IPv6 routing header to simplify routing to mobile nodes.
c. With the enhanced authentication header in IPv6 and the mandatory implementation of IP authentication header, Mobile IPv6 might adopt a wide scale of route optimization if a key management infrastructure becomes widely available on the Internet.
d. Mobile IPv6 uses both tunneling and source routing to deliver packets to mobile nodes. In the case of Mobile IPv4, tunneling is the only option.

8.4 TRANSMISSION CONTROL PROTOCOL (TCP)

8.4.1 Flow Control

As mentioned at the start of this chapter, the Internet Protocol only provides best effort services. Traffic control in the Internet uses the TCP in the transport layer to exercise flow control. TCP, which is designed to provide reliable end-to-end services, is a sliding window flow control mechanism. Control decisions are made based on control signals fed back from the receiver. In this sense, TCP is a reactive control method. The control signals fed back from the receiver are in the form of acknowledgment (ACK) packets. The main indicator of a problem of unsuccessful transmission through the IP network is congestion experienced in the routers along the path connecting the sender and the receiver. In conventional TCP schemes, designed primarily for exercising flow control in IP-based networks, control decisions to regulate the traffic allowed to enter the network are strictly based on network congestion. In a hybrid wireless/IP network, transmission errors incurred in the wireless propagation channel will also have to be taken into consideration in formulating the flow control policy.

As a reactive control scheme, the window size in TCP is regulated by a control signal fed back from the receiver to the transmitter. Packets arriving at the sender are stored in a buffer. The sender releases packets from the buffer and adjusts the size of the sliding window based on the acknowledgment fed back from the receiver. To facilitate flow control, the sender keeps a timer. If the sender fails to receive an acknowledgment for a packet from the receiver after a timeout interval, it retransmits that packet and exponentially backs off the timer. To determine an appropriate timeout value, TCP keeps track of the round-trip time for the data packets and the corresponding acknowledgment packets, and uses the accumulated knowledge to calculate the timeout value.

The sender transmits packets based on the size of the sliding window. At the start, the sender assumes a minimum window size and probes the network to determine the available network capacity. If the network is not in a congested state, so there is bandwidth to support an amount of traffic greater than that specified by the minimum window size, the window size is increased exponentially until the amount of traffic entering the network reaches a preset threshold. This exponential increase in the window size is referred to as a *slow-start*. When the threshold is reached, the sender probes the network continuously in a linear manner in an attempt to avoid network congestion.

If the sender fails to receive an ACK from the receiver for a particular packet after a timeout period, then the sender assumes the packet is lost inside the network, due to network congestion. The sender sets the threshold to one half of the last window size and throttles the window size to its initial value to avoid network congestion, and then retransmits the packet.

There are a number of ways the receiver can send acknowledgments. For example, the receiver may send an ACK only for each correctly received packet. In this case, the sender uses a timeout to retransmit the lost packets. An alternative is for the receiver to send duplicate acknowledgment packets for the last correctly received packet. Upon receiving a predetermined number of duplicate acknowledgment packets, the sender infers that there are missing packets in the receive buffer, and starts retransmitting the first unacknowledged packet, and then reduces the window size proportionately. A scenario in which TCP starts with slow-start and a preset

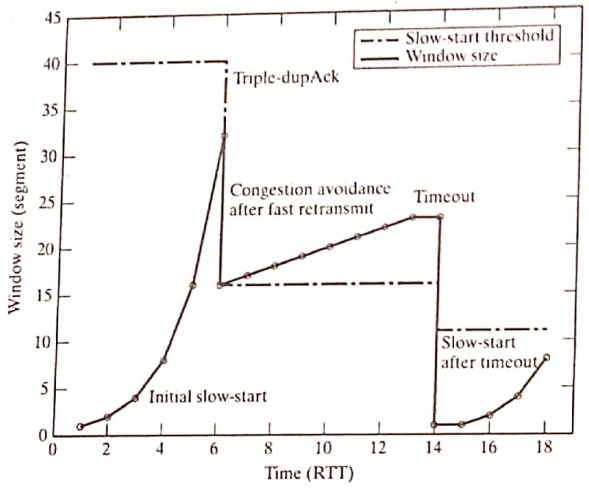

Figure 8.13 TCP window flow control strategy.

threshold value is shown in Figure 8.13, where the horizontal axis is in units of round-trip time (RTT) and the timeout interval is assumed to be one round-trip time. When the sender receives three duplicate acknowledgment packets (i.e., *Triple-DupAck* in Figure 8.13), it infers that the network is experiencing congestion. TCP then performs Fast Retransmit (see Subsection 8.4.1) and enters the congestion control phase by reducing the threshold value to one-half of the last window size.

8.4.2 Modified TCP

TCP is a connection-oriented transport layer protocol that is designed to provide reliable and in-sequence data delivery. However, if TCP is used without any modification in a hybrid wireless/IP network, a serious drop of throughput may occur. The reason is that a high bit error rate (BER), or disruption caused by poor wireless link quality, can corrupt packets, which may result in losing TCP data segments or acknowledgments. When acknowledgments do not arrive at the TCP sender within a prescribed interval of time, a timeout occurs. The sender retransmits the segment, exponentially backs off its retransmit timer for the next retransmission, and then reduces its window to one segment. Repeated errors will cause the window to remain small, resulting in a low throughput, especially on long links.

In order to ensure that the TCP connection to a mobile is efficient, it is necessary to prevent the sender from shrinking its congestion window when packets are lost, either due to bit errors or due to disconnection. When the mobile is reconnected, it should begin to receive data immediately. Several proposals have been reported in the literature for a modified or new TCP that is optimized for use over wireless links. These include Indirect Transmission Control Protocol (I-TCP), Berkeley Snoop Module, Fast Retransmit, and TCP for Mobile Cellular Network (M-TCP) [10, 20].

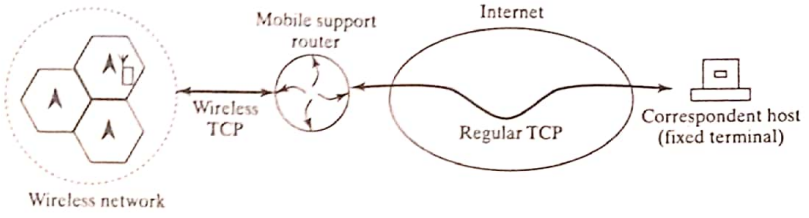

Figure 8.14 Indirect Transmission Control Protocol (I-TCP).

Indirect Transmission Control Protocol (I-TCP). In I-TCP, a connection between a mobile host and a fixed host is split into two separate connections at the base station—one between the mobile host and the base station or its mobile support router (MSR) over the wireless medium, and the other between the base station, or the MSR, and the fixed host over the wired network. The I-TCP scenario is shown in Figure 8.14. In this way, the special requirements of mobile hosts can be accommodated, which is backward compatible with the existing fixed network. All the specialized supports that are needed for mobile applications and for the low speed and unreliable wireless links can be built into the wireless side of the interaction while the fixed side is left unchanged. Data sent to the mobile host are received and acknowledged by the base station before being delivered to the mobile host. The wireless/wired link characteristics would be hidden from the transport layer, and only the wireless resources would be used for error control when the error is caused by the wireless link. With the I-TCP, the resulting benefits lie in that:

a. the flow control and congestion control functionalities on the wireless link are separated from those on the wired link;
b. a separate transport protocol for the wireless link can support notification of events such as disconnections, user movements and other features of the wireless link (e.g., changes in the available bandwidth) to the higher layers;
c. a partition of the connection into two distinct parts allows the base station to manage much of the communication overhead for a mobile host.

Throughputs are increased with I-TCP since the node, where the connection is split, may be one or two hops away from the mobile host's radio cell, and can adapt more quickly to the dynamic mobile environment because the round-trip time is shorter. However, I-TCP does not maintain end-to-end TCP semantics. This is because the TCP acknowledgments are not end-to-end; instead, there are separate acknowledgments for the wireless and the wired portions of the connection. One consequence is that the sender may believe that a segment is delivered correctly to the mobile host since the base station acknowledges it even if the mobile host is disconnected before receiving this segment. In other words, the sender does not know whether packets are actually received by the mobile host, and this may be a serious problem for many applications.

Berkeley Snoop Module. Snoop is another proposed solution for losses caused by a high BER. The Berkeley Snoop Module makes changes to the network layer software at the base station. It caches packets at the base station, inspects the TCP header of TCP data packets

and acknowledgments which pass through, and buffers copies of the data packets. Using the information from the headers, the snoop module detects lost packets (a packet is assumed lost when duplicate acknowledgments are received) and performs local retransmissions across the wireless link to alleviate problems caused by a high BER. The module also implements its own retransmission timer, similar to the TCP retransmission timeout, and performs selective retransmissions when an acknowledgment is not received within this interval. Routing protocol is also modified to enable low-latency handoff to occur with negligible data losses. Experiments have shown that the Berkeley Snoop Module achieves a throughput up to 20 times that of regular TCP, and handoff latencies over 10 times shorter than those of other mobile routing protocols. A drawback of the snoop module is that it does not perform as well in either the presence of lengthy disconnections or environments in which there are frequent handoffs. If the mobile host is disconnected for a lengthy period of time, the sender will automatically invoke congestion control because it does not receive acknowledgments for some segments. The snoop module will persistently generate packets and these packets will serve no purpose since the mobile is disconnected. If the mobile host moves into a new cell, the new base station starts up a copy of the snoop module on behalf of this mobile host. This snoop module begins with an empty cache and slowly builds up the cache, and the mobile host will see a poor TCP throughput. If the radio cell sizes are small, the performance degradation can be serious.

Fast Retransmit. Fast Retransmit is proposed to combat the effects of short disconnections on TCP throughput. During a handoff, since the mobile host cannot receive packets, unmodified TCP at the sender will think that a congestion has occurred and will begin congestion control (by reducing the window size and retransmitting) after a timeout. The timeout period may be long; even though the mobile host may have completed the handoff, it will have to wait for the full timeout period before it can begin receiving packets from the sender. Fast Retransmit forces the mobile host to retransmit, in triplicate, the last old acknowledgment as soon as it finishes a handoff. This forces the sender to reduce the congestion window to one-half and retransmit one segment immediately. Fast Retransmit does not split the TCP connection. However, if the mobile host were disconnected for a long time, the sender would already have invoked congestion control and shrunk its window to one segment. Similarly, if disconnections are frequent or the wireless links are poor, Fast Retransmit will do little to improve the throughput because the sender's congestion window will repeatedly get shrunk to half of its previous size.

TCP for Mobile Cellular Network (M-TCP). M-TCP works in a three-level hierarchy: mobile hosts, supervisor host, and the Internet, from the lowest to the highest by introducing the supervisor host (SH). The three-level M-TCP hierarchy is shown in Figure 8.15, where several base stations are controlled by one SH. The SHs are connected to the Internet and handle most routing and other mobile users' requirements. The advantages of this hierarchy are:

a. When a mobile host roams from one cell to another, the two base stations do not need to transfer any state information if they are controlled by the same SH.
b. The roaming mobile host remains within the domain of the same SH for long time periods because several base stations are controlled by the SH.

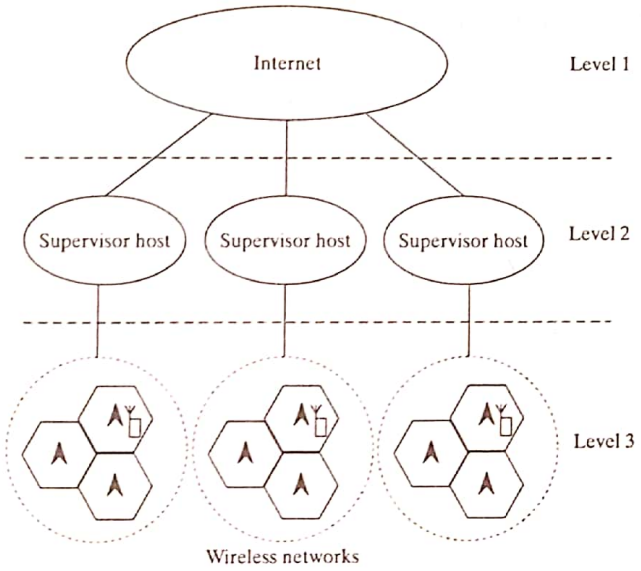

Figure 8.15 M-TCP hierarchy.

By introducing the SH, M-TCP maintains end-to-end TCP semantics while it delivers excellent performance when mobile hosts encounter disconnections. This is done by splitting the TCP connection at the SH. As packets arrive from a sender on the Internet, an acknowledgment is sent back and the SH deals with ensuring the completion of delivery. The drawback of this scheme is that it is fairly complex. Also, there will likely be a shortage of buffer space if a supervisor host services too many mobile hosts.

8.4.3 Modified UDP

User Datagram Protocol (UDP) is a datagram communication service built on top of IP. It adds multiplexing and error detection to the IP capabilities. In contrast to TCP, UDP does not use acknowledgments, and does not retransmit erroneous packets or control the flow. In wireless/IP interworking, a large percentage of packets will be lost by using UDP over wireless links. This is because UDP will continue to send packets even when transmission to a mobile host experiences signal fading. A simple concept would be to stop sending datagrams to a mobile host once it encounters fading. The goal of creating a new modified UDP (M-UDP) is to ensure that packets that have been lost are retransmitted. In the M-UDP protocol, the UDP connection is split in two at some host close to the mobile host. The host attempts to use any free bandwidth to retransmit packets lost during a fade, thus ensuring that the number of lost packets is kept small.

8.5 NETWORK PERFORMANCE

Communications networking operations are relatively complex. Exact analytical evaluation of network performance is often too difficult to tackle. Approximate analysis based on certain reasonable assumptions can yield valuable guidance for specifying system and traffic parameters. In this section, we evaluate the performance of the local anchor scheme described in Subsection 8.2.2, as an example. It is necessary to use some metrics as performance measures. Two such metrics are average handoff delay and average cost between two consecutive handoffs.

It is not easy to precisely model the Internet, since the route between any two hosts may change dynamically. But a network model is needed to facilitate performance analysis and evaluation.

8.5.1 Network Model

To simplify the analysis problem, assume that the route between the involved agents is fixed and, within the mobile's residence time, the delay on the corresponding route is also fixed. The delay of a path is proportional to the length of the path or the number of hops on the path. For simplicity, here we take the delay on a path as the cost of the path.

Consider the case in which the mobile receives information from a fixed terminal (the correspondent host). Assume that the correspondent host generates data packets destined for the mobile at a fixed rate λ. The network model for cost analysis is shown in Figure 8.16, where a, b, d, and f are the costs (delays) associated with the paths, respectively.

Assume a, b, and d are fixed, but f is a variable. In order to quantify f, assume that the relationship between the physical distance and the network distance can be modeled by a bifork tree, as shown in Figure 8.17. From this figure, it can be seen that, if any two agents are neighbors in

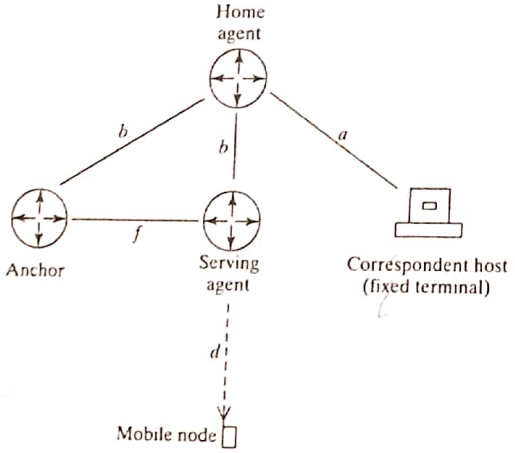

Figure 8.16 Network model for cost analysis.

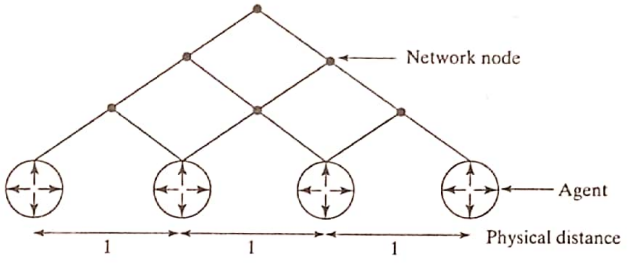

Figure 8.17 Network distance versus physical distance.

the physical location, they will have a distance (and therefore a delay) of two hops between them. Let d_p be the physical distance between two agents and assume each hop has a unity delay. Then $f = 2 \times d_p$. For example, if $d_p = 2$, there will be four hops between the two agents in the network.

8.5.2 Mobility Model 3

To capture the effect of the locality of the user movement on the average handoff delay, we introduce the following two-dimensional mobility model.

User movements in the network are modeled by boundary crossings between serving areas. The residence time of the mobile in each serving area (the interval between two consecutive handoffs) is assumed to be an exponentially distributed random variable with a constant mean value. Calls to the mobile are modeled as a Poisson arrival process with a constant rate. In addition, calls are assumed to be generated for randomly selected serving areas. The assumptions allow us to characterize the mobile's movement by a Markov state transition diagram. Furthermore, for convenience we consider rectangle radio cell clusters of equal size as the service areas. Each cluster has a unique attachment point to the Internet. As each cell cluster has four adjacent neighboring clusters, on each crossing of a cluster boundary, the mobile can move into one of four adjacent serving areas, with respect to the mobile's current serving area. The direction of each movement is modeled as a Markov process, as shown in Figure 8.18, where P_{back} is the probability that the mobile will move back to its previous serving area, and P_{same} is the probability that the next move will be in the same direction as the previous move. The probability of the mobile moving in any other direction is P_{other} which is equal to $(1 - P_{\text{same}} - P_{\text{back}})/2$. The probabilities P_{back} and P_{same} allow us to model various degrees of locality in the user's movements.

8.5.3 Handoff Delay with Local Anchor

Handoff delay is defined as the time interval from the instant when the mobile sends the registration request to the new foreign agent, to the instant when the mobile is allowed to send packets to, or receive packets from, the new foreign agent. A handoff will take place when necessary, whether or not a TCP connection exists. If there is no existing TCP connection when the handoff takes place, the handoff will finish when the mobile receives a handoff reply from the new foreign agent. In

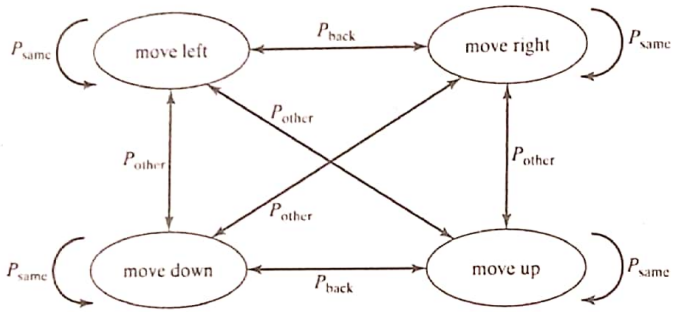

Figure 8.18 Two-dimensional Markov mobility model.

the operation of TCP in a hybrid wireless/IP environment, the air link needs to be considered separately. With M-TCP (see Subsection 8.4.1), the splitting occurs at the supervisor host. With the local anchor scheme in Mobile IP, the split can take place at the anchor agent [165]. This is the focal agent in the anchoring region which maintains the TCP connection state for the mobile host. That is, when the mobile moves within the anchoring region, its new serving agent does not keep track of the TCP connection state for the mobile; it therefore does not need to copy the mobile's TCP connection state from the old foreign agent. This will yield a reduction in handoff delay, in addition to the reduction in delay due to shorter registration path with the local anchor.

In the network model shown in Figure 8.16, if the mobile host has to register with its home agent, the registration request and reply have to go through paths with delays b and d, plus two processing times at the foreign agent and one processing time at the home agent. Let r denote the fixed processing time (cost) at each agent. Assume a split connection scenario, where the split is located at the focal agent of the anchoring region. We will refer to this transport layer connection as an anchor-based indirect TCP (I-TCP). Then, the registration delay, $t_{\text{I-TCP}}$, of this anchor-based I-TCP is given by

$$t_{\text{I-TCP}} = 2(b + d) + 3r. \tag{8.5.1}$$

After a successful registration, the new foreign agent sends a message to the old foreign agent and requests TCP state transfer. This will incur a transfer delay, t_{transfer}, of an amount

$$t_{\text{transfer}} = (n + 2)(f + r), \tag{8.5.2}$$

where n is the TCP buffer size, f is the delay between the anchor and the serving agent (see Figure 8.16 and Problem P8-11). Therefore, when the mobile host has to register with its home agent, the total handoff delay, $t_{\text{h\_I-TCP}}$, for the anchor-based I-TCP scheme is

$$t_{\text{h\_I-TCP}} = 2(b + d) + 3r + (n + 2)(f + r). \tag{8.5.3}$$

On the other hand, if the mobile roams within the anchoring region, it only needs to register with the anchor, as opposed to the home agent. In this case, the registration request and reply will

go through the paths with delays d and f, plus two processing times at the new foreign agent and one processing time at the anchor. Thus, the registration delay is

$$t_{\text{h\_anchor}} = 2d + 2f + 3r. \tag{8.5.4}$$

If the new foreign anchor decides that the mobile should register with its home agent, as opposed to the anchor, the handoff delay should be that given by Eq. (8.5.3).

Because of mobility, the mobile will always change its location in the anchor region. Once the mobile moves outside the anchoring region, the new foreign agent will become the focus of the new anchoring region. Thus, the value of f will change when the mobile moves.

Example 8.4 Variable Values of f

In a local anchor-based Mobile IP, the parameter f, the delay between the local anchor and the serving agent, is a variable. Why and how is f a random variable?

Solution The value of f depends on the radius of the anchoring region, in particular, the distance of other agents from the focal agent. An anchoring region with 25 agents is shown in Figure 8.19. The focal agent is located at the center. Any one of the other agents may be the serving agent (see Figure 8.16). Eight of the agents are located at a distance of 1 from the focus and 16 of the agents are located at a distance of 2 from the focus.

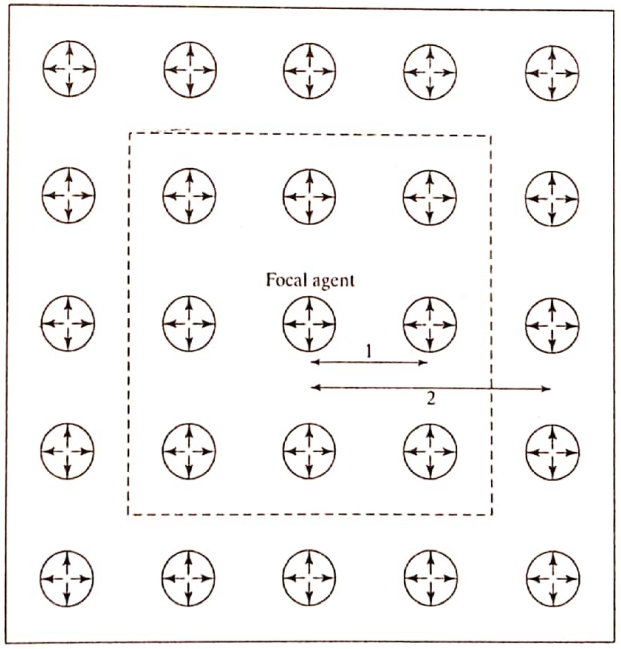

Figure 8.19 Anchoring region with 25 agents.

For the agents with distance 1 from the focus, $f = 2$; for the agents with distance 2 from the focus, $f = 4$. When the mobile moves outside the anchoring region, the value of f to be used to compute the TCP state transferring cost will be greater than 4 (e.g., $f = 6$).

In order to compute the average handoff delay, we need the knowledge of the probability distribution of the mobile staying at each service area. Let π_j be the probability that the mobile stays at service area j with or without the information of its previous area. The average handoff delay is then given by

$$t_{h\_anchor} = \sum_j \pi_j t_{h\_anchor}(j). \tag{8.5.5}$$

Example 8.5 Probability Distribution of Mobile's Location

Let $\pi = (\pi_1, \pi_2, \cdots, \pi_J)$ be the probability vector, based on the Mobility Model 3. Consider an anchoring region with 9 agents, as shown in Figure 8.20. Devise a method for calculating the limiting probability vector π.

Solution In Figure 8.20, the focal agent (anchor) is located at the center. Each of the other eight agents is at a distance of 1 from the focus. Consider only the situation of one step movement where a mobile has moved out of its previous area into a new area. Here we have four unique service areas, A, B, C, and the area outside the anchoring region. The mobile's movement and its location can be described by the following six states (i.e., $J = 6$):

S_1: the mobile is located in region A;
S_{21}: the mobile is located in region B and its last location is in region A;
S_{23}: the mobile is located in region B and its last location is in region C;
S_{24}: the mobile is located in region B and its last location is outside the anchoring region;

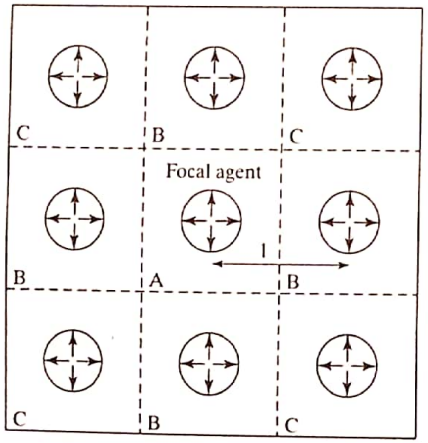

Figure 8.20 Anchoring region with nine agents.

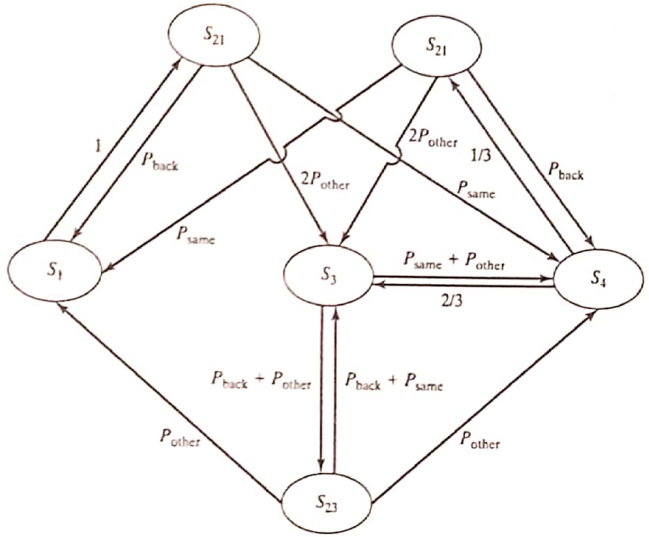

Figure 8.21 Markov chain for the anchoring region with nine agents.

S_3: the mobile is located in region C; and
S_4: the mobile moves out of the current anchoring region and needs to register with the home agent.

The above states characterize the Markov chain shown in Figure 8.21. By assigning S_1, S_{21}, S_{23}, S_{24}, S_3 and S_4 the positional values 1, 2, 3, 4, 5, and 6, respectively, we can construct the Markov transition matrix, M, for this Markov chain as

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ P_{back} & 0 & 0 & 0 & 2P_{other} & P_{same} \\ P_{other} & 0 & 0 & 0 & P_{back} + P_{same} & P_{other} \\ P_{same} & 0 & 0 & 0 & 2P_{other} & P_{back} \\ 0 & 0 & P_{back} + P_{other} & 0 & 0 & P_{same} + P_{other} \\ 0 & 0 & 0 & 1/3 & 2/3 & 0 \end{pmatrix}.$$

The limiting probability vector, π, can be readily computed from the relationship

$$\pi = \pi M.$$

The Markov chain and the corresponding transition matrix become significantly large when the radius of the anchoring region increases. An alternative approach is to simulate the mobile movements and compute the cost associated with the movements. In the limit, the simulation results should converge to the theoretical value quite well.

8.6 WIRELESS APPLICATION PROTOCOL (WAP)

The Transmission Control Protocol (TCP) has been designed to treat all data losses as being due to network congestion. Thus, conventional TCP erroneously considers data losses due to transmission errors and/or handoff disruptions as the result of network congestion. Taking all losses as being due to network congestion unnecessarily limits the network throughput. Newer versions of TCP (e.g., Tahoe, Reno, NewReno, SACK (Selective Acknowledgment), and other variants [7, 95]) attempt to make TCP aware of transmission errors as well as errors due to network congestion. While TCP/IP can be an effective flow/congestion control protocol for wireless/IP interworking, the large overhead associated with TCP/IP to take care of all sources of data loss is a huge drawback. This is perhaps the main reason behind the formation of the WAP (Wireless Application Protocol) Forum, a consortium of wireless equipment manufacturers and service providers. The WAP Forum introduces a set of WAP specifications for standardization, aiming to facilitate Internet access from wireless terminals such as cellphones, e-commerce portable terminals, and similar devices. The main advance of WAP over TCP is the light overhead that caters to application devices with limited computing power, low power consumption, and similar limitations. Although WAP-enabled wireless terminals and WAP-based Internet services are now available, WAP standards are still evolving.

As discussed in Section 8.4, TCP connections can be end-to-end or split connection (i.e., concatenation of two segments). WAP is based on a split connection principal, where the split is at the base station. The two segments are wireless (from the mobile terminals to the base station) and wireline (from the base station to the backbone network). The wireless segment employs protocols specifically designed for, and fine-tuned to, the wireless propagation channel.

The WAP suite addresses protocols from the transport layer up to the application layer. Specifically, the WAP protocol stack includes the application, session, transaction, and transport layers. To support the WAP protocol stack, the wireless segment of the WAP connection operates as a circuit-switching pipe in handling the transmissions by mobile/portable terminals to the base station. That is, with the WAP suite, the wireless segment is circuit-switched while the wireline segment, from the base station onward, is packet-switched. A request-response procedure is used to establish the WAP connection. The WAP architecture can thus be modeled as a *client-server* scenario, with the mobile terminal as the client and the Internet host as the *original server*. Client-server communication is handled by an interface or *gateway* located at the base station. The gateway interfaces the circuit-switching and the packet-switching modes by performing transmission format translation and conditioning of the data units suitable for forwarding to the client or the server. A functional block diagram of the WAP model is depicted in Figure 8.22.

8.6.1 Wireless Application Environment

Because the WAP standards are still evolving, documentation on the different layers of the WAP protocol stack consists mainly of WAP Forum documents [www.wapforum.org]. The application layer defines the Wireless Application Environment (WAE), which provides for interaction between WAP/Web applications and wireless devices containing a WAP microbrowser. WAE functions include: (*a*) Wireless Markup Language (WML), which accommodates the limitations of wireless devices with limited display capabilities, (*b*) WMLScript, which is WML's accompanying

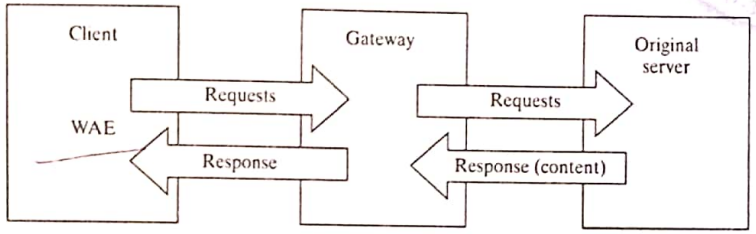

Figure 8.22 WAP programming model.

client-side scripting language that provides for additional intelligence and control over presentation, and (*c*) Wireless Telephony Application. In addition to supporting presentation services, similar to HTML (hypertext markup language), WML adds support for events and variables.

8.6.2 WAP Protocol Stack

The latest release is WAP 2.0 [WAP 2.0 Technical White Paper, Aug. 2001]. These protocols have been optimized for low bandwidth bearer networks with relatively long latency. In addition to the Wireless Application Environment (WAE), the other layers, together with their functionality, are as follows.

Wireless Session Protocol (WSP). WSP provides HTTP/1.1 functionality and incorporates new features, such as long-lived sessions and session suspend/resume. WSP provides the application layer with a consistent interface for two session services: a connection-oriented service that operates above the transaction layer protocol, and a connectionless service that operates above a secure or nonsecure datagram transport service. WSP is optimized for the low bandwidth and long latency inherent in most wireless transmission media. It enables the WAP client to negotiate, open, and maintain a session with the WAP gateway. If a connection closes prematurely, it can go into a sleeping mode and be wakened whenever the connection is reestablished.

Wireless Transaction Protocol (WTP). WTP is a light weight transaction-oriented protocol that is suitable for implementation in "thin" clients (e.g., handsets) and operates efficiently over wireless datagram networks. It is designed to reliably carry out transactions between the client and the server. WTP has the following salient features:

a. Improved reliability over datagram services—WTP relieves the upper layer from retransmissions and acknowledgments that are necessary when datagram services are used;
b. Improved efficiency over connection-oriented services—WTP has no explicit connection setup or tear down phases;
c. Advantages of using a message-oriented protocol, designed for services oriented toward transactions, such as "browsing".

Wireless Transport Layer Security (WTLS). The WTLS layer is designed to provide privacy, data integrity and authentication between two communicating applications. It provides the upper

sublayer of WAP with a secure transport service interface that preserves the transport service interface below it. Additionally, WTLS provides an interface for managing (e.g., creating and terminating) secure connections.

Wireless Datagram Protocol (WDP). WDP is a general datagram service which offers a consistent service to the upper layer protocols, and communicates transparently over one of the available underlying bearer services. This consistency is provided by a set of adaptations to specific aspects of these bearers, and provides a common interface with the upper layers to enable operations independent of the services of the wireless network.

8.6.3 WAP Gateway

The WAP gateway provides the interface between the client and the server. The protocol stacks for the client-gateway-server combination are illustrated in Figure 8.23. As can be observed from Figure 8.23, the WAP gateway has two protocol stacks, one for peer communications with the mobile hosts through the wireless propagation channel, and the other for peer communications with the Internet server through wirelines.

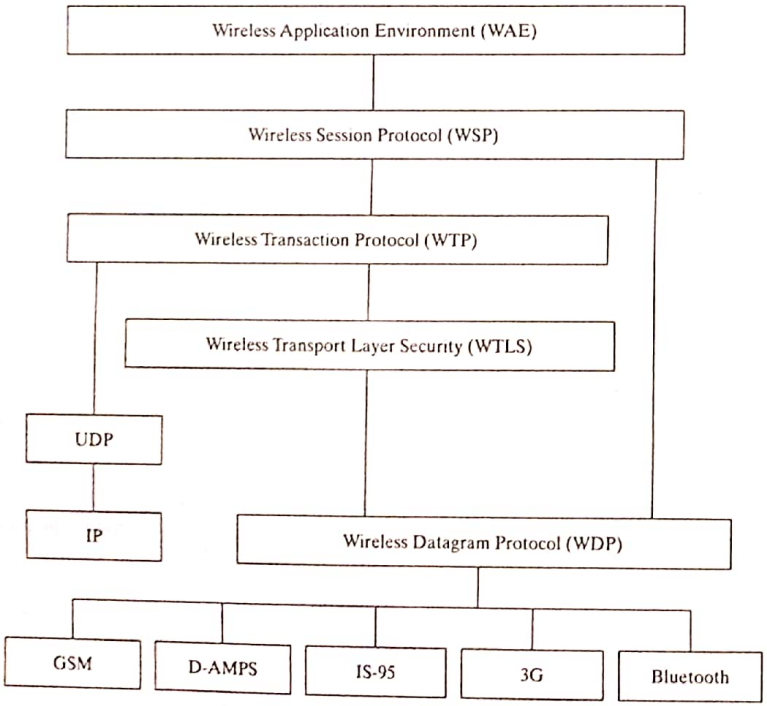

Figure 8.23 WAP protocol stack.

8.7 MOBILE AD HOC NETWORKS

There are two variations of mobile wireless communication networks. The first is known as mobile cellular networks, which is infrastructure-based, and the second is known as mobile ad hoc networks, which is infrastructureless. A mobile cellular network consists of an array of radio cells in which communications in each of the cells is handled by a base station. Thus, the base station is the fixed infrastructure which performs centralized administration. Mobile stations within the footprint of a base station directly communicate with that base station which, in turn, forwards (routes) traffic to designated destinations. Thus, among other tasks, the base station assumes the role of a *router*. A system with a fixed infrastructure is basically a two-hop system. Mobile cellular networks discussed so far in this text have been of the infrastructure-based variety. When a mobile moves outside the footprint of the currently serving base station, its connection with the destination is handed off to the base station with which the mobile station must now communicate.

Infrastructureless mobile networks are commonly known as *ad hoc networks*. Ad hoc networks have no fixed routers; all nodes are capable of movement and can be interconnected dynamically in an arbitrary manner. Nodes of ad hoc networks behave as routers that discover and maintain routes to other nodes in the network. A node in an ad hoc network directly communicates with other nodes in the network if they are within line-of-sight. However, there are also hidden (non-line-of-sight) nodes. Communication between a pair of hidden nodes needs to hop over one or more intermediate nodes. In this sense, ad hoc networks can be thought of as multihop networks.

The connectivity in ad hoc networks is much more complex than that in wireless networks with an infrastructure. This means *routing* in ad hoc networks is a more complex issue than in infrastructure-based networks. It follows that the existence of effective routing protocols is essential for effective and efficient operation of mobile ad hoc networks.

8.7.1 Ad Hoc Routing Protocols

Routing protocols for ad hoc networks can be categorized into two types: (*a*) table-driven and (*b*) source-initiated or demand-driven [17].

Table-Driven Routing Protocols. Table-driven routing protocols try to maintain consistent, up-to-date routing information from each node to every other node in the network. Each node is required to maintain one or more tables to store routing information. An example of table-driven protocols is Destination-Sequenced Distance-Vector (DSDV) Routing. DSDV is a table-driven algorithm based on the classical Bellman-Ford routing mechanism. The table maintained by each and every mobile node in the network contains all of the possible destinations and the number of hops to each destination in the network.

Source-Initiated on-Demand Routing. Source-initiated on-demand routing creates routes only when desired by the source node. When a node requires a *route* to a destination, it initiates a route discovery process within the network. This process is completed once a route is found or all possible route permutations have been explored. Once a route has been established, it is maintained by a route maintenance mechanism until either the destination becomes inaccessible along every path from the source or until the route is no longer needed. An example of on-demand

routing is Dynamic Source Routing (DSR) [69]. DSR is based on the concept of source routing. In source routing, mobile nodes are required to maintain route caches that contain the source routes of which the mobile is aware. Entries in the route cache are continually updated as new routes are learned. The source routing protocol consists of two phases: route discovery and route maintenance. When a mobile node has a packet to send to some destination, it first consults its route cache to determine whether it already has a route to the destination. If there exists a route to the destination, it will use that route to send the packet. On the other hand, if the node does not have a route to the destination, it initiates route discovery.

Route Discovery. The source node initiates route discovery by broadcasting a *route request* packet along all its outgoing links. The route request packet contains the address of the destination, the source node's own address, a unique identification number and a route record field. Each node receiving the packet checks whether it knows a route to the destination. If it does not, it adds its own address to the route record of the packet and then forwards the packet along its outgoing links. Route discovery by broadcasting is tantamount to *flooding*, which can consume a large amount of wireless resources. To limit the number of route requests sent on the outgoing links of a node, a mobile only forwards the route request if the request has not yet been seen by the mobile and if the mobile's address is not already in the route record.

A *route reply* is generated when the route request reaches either the destination node, or an intermediate node that contains in its route cache an unexpired route to the destination. By the time the packet reaches either the destination or such an intermediate node, it contains a route record with the sequence of hops taken. If the node generating the route reply is, in fact, the destination, it places the route record contained in the route request into the route reply. If the responding node is an intermediate node, it will append its cached route to the route record and then generate the route reply. To return the route reply, the responding node must have a route to the initiator. If the responding node does have a route to the initiator in its route cache, it may use that route. Otherwise, if symmetric links are supported, the node may reverse the route in the route record. If symmetric links are not supported, the node may initiate its own route discovery and piggy-back the route reply on the new route request.

Route Maintenance. Route maintenance is accomplished through the use of route error packets and acknowledgments. Route error packets are generated at a node when the data link layer encounters a fatal transmission problem. When a route error packet is received, the hop in error is removed from the node's route cache and all routes containing that hop are truncated at that point. In addition to route error messages, acknowledgments are used to verify the correct operation of the route links. Such acknowledgments include passive acknowledgments, in which a mobile is able to hear the next hop forwarding the packet along the route.

Applications. Ad hoc wireless networks have an important role to play in military applications. There are also commercial scenarios for ad hoc wireless networks, which include:

a. conference/meetings/lectures,
b. emergency services, and
c. law enforcement and similar services.

Approaches reported in the literature for ad hoc routing include paradigms that exploit such features as user demand, user location, power, and association parameters. Adaptivity and self-configuration are key features of these approaches. However, flexibility is also important. A flexible ad hoc routing protocol could responsively invoke table-driven and/or on-demand approaches based on situations and communication requirements. However, the toggling between these two approaches may not be trivial since the nodes involved must be in synchronism with the toggling. Coexistence of both approaches may also exist in spatially clustered ad hoc groups, with intracluster employing the table-driven approach and intercluster employing the demand-driven approach or vice versa.

8.7.2 Comments

In ad hoc networks, dynamic reconfiguration and establishment of routes are the most important features. Two common approaches in route discovery are table-driven and on-demand-driven. There are different schemes available for implementing each of these approaches. Routing protocols are schemes; as such, they are not based on systematic analytical formulation. Any scheme introduced needs to show that it performs well under different network topologies and movement patterns. Algorithmic determination and/or computer simulation are normally used to demonstrate the viability of the routing protocols.

For commercial scenarios such as conference, lectures, and the like, wider geographical coverage than that offered by ad hoc networking would be desirable. In these situations, a backbone network such as the Internet would be used to extend the geographical coverage. There are open challenging problems associated with the interworking of wireless ad hoc networks and IP-based networks.

SUMMARY

While the wireless communication network offers the flexibility for users to roam, the geographical coverage of a wireless system is nevertheless limited. A backbone network, either in the form of a wide area wireline network, or a global satellite network, is needed to provide global communications. The Internet is the most pervasive wireline network that has been enjoying universal acceptance. The interworking of a wireless front-end and an Internet backbone should provide an effective information transfer platform for supporting user roaming on a global basis.

Information transfer across the Internet is in the form of datagrams, and the Internet Protocol is a connectionless datagram transport mechanism that resides in the network layer of the protocol stack. By design, IP has no built-in traffic control capability, and only provides *best effort* services. A connection-oriented and window-based transport layer protocol, referred to as the Transmission Control Protocol, is used to exercise flow control over the Internet. TCP/IP is thus the adopted information transfer mechanism for the Internet. One of the fundamental limitations of the Internet is that each subscriber is only given one IP address, which resides in the subscriber's home network.

The interworking of a wireless segment with the IP-based backbone network requires an effective interface. Mobile IP, introduced by the Internet Engineering Task Force, is the peripheral network that provides the interface between the wireless segment and the IP segment. This chapter

has discussed the concepts of, and the entities that make up, Mobile IP. The key elements in Mobile IP are agents, both home agents and foreign agents, that provide the mobile user a care-of address, in addition to its home address. The home address is for *identification* while the care-of address, that changes at each access point, is for *routing*.

The concepts of Mobile IP, and the ramifications of TCP/IP within the context of wireless/IP interworking to provide global information delivery to roaming users, are described and discussed in this chapter.

Although TCP/IP is a viable flow control mechanism for the Internet, when modified forms of TCP are used to enforce flow control for an interworked wireless/IP network, the overhead associated with the application of TCP can be heavy. The WAP (Wireless Application Protocol) Forum has introduced WAP specifications, which provide end-to-end information transmission with QoS provisioning, with lighter overhead compared with TCP, in the wireless Internet access. An overview of WAP is included in this chapter.

Cellular wireless networks use a fixed infrastructure for central administration. In certain applications (e.g., in a military scenario or emergency situations), an infrastructureless network, commonly referred to as an ad hoc network, is preferred. The operational features of ad hoc networks are also described.

ENDNOTES

1. For an overview of IPv6, see the paper by Lee *et al.* [77]. For details of mobility support in IP networks, see the papers by Bhagwat, Perkins, and Tripathi [15], Campbell *et al.* [22], Das *et al.* [37], Perkins and Bhagwat [112], and Manzoni, Ghosal, and Serazzi [93], Ramjee *et al.* [126] [127], and Zhang and Mark [165].
2. For TCP and its performance enhancement, see the papers by Bakre and Badrinath [10], Barakat, Altman, and Dabbous [11], Brown and Singh [18], Caceres and Ifto [20], Ghani and Dixit [53].
3. For details of routing protocols in wireless ad hoc networks, see the papers by Broch, Johnson, and Maltz [17], Johnson and Maltz [69], Perkins and Bhagwat [111], Royer and Toh [131], and the special issue on wireless ad hoc networks of *IEEE Journal on Selected Areas in Communications* [57].
4. For QoS issues in wireless networking, see the papers by Chakrabarti and Mishra [25], Kim and Jamalipour [72], Sobrinho and Krishnakumar [143], Xylomenos and Polyzos [160].
5. For fundamentals of data communication networks, see the book by Bertsekas and Gallager [13]. For wireless network architectures, see the book by Lin and Chlamtac [87].

PROBLEMS

P8-1 The interworking of a wireless network with an IP-based network provides wide geographical coverage for information delivery to mobile users. Describe and discuss the problems created due to user roaming, and the mechanisms needed to ensure seamless information delivery to the mobile user at its current location.

P8-2 When a mobile migrates to a foreign network, certain actions must take place so that messages destined for the mobile in its current location can be delivered.

a. How does the mobile host determine its current location?
b. Describe the steps involved to facilitate delivery of messages from the correspondent host to the mobile host in its current location.
c. Why does Mobile IP need to use two IP addresses?
d. Why is tunneling needed when the home agent forwards messages to the mobile in its current location when it is away from its home network?

P8-3 Mobile IP is a peripheral network (or interface) that bridges a wireless network to an IP-based network for information delivery to a mobile user when it is away from its home network. There are three major processes that must work cooperatively in order for Mobile IP to operate effectively. Name the processes and describe the operational characteristics of each process.

P8-4 In Mobile IP, there are essentially two methods available for routing messages between a mobile host and an Internet correspondent host.

a. What are the two routing methods?
b. Describe and discuss the functions and the operation of the two routing methods.
c. Compare the advantages and disadvantages of the two routing schemes.
d. The simpler method is a popular choice. Under what condition would the more complex routing method be a more appropriate choice?

P8-5

a. Describe and discuss how techniques such as local anchor can reduce the network cost when the mobile user roams away from its home network.
b. If local anchor is the technique used to reduce registration cost, how do you choose the local anchor as the mobile migrates from one foreign network to another?

P8-6 In mobile Internet communications, traffic routing from a source host to a destination host can be performed using flat (one level) routing or hierarchical (multilevel) routing.

a. How does hierarchical routing reduce registration cost?
b. By arranging the foreign agents in a binary tree structure, one can reduce the need for the mobile to register with the home agent. If you have an array of 14 foreign agents, how should the foreign agents be populated in a binary tree to facilitate registration by the mobile as it moves from one leaf node to another leaf node?
c. In hierarchical routing, how are datagrams, which are destined for the mobile in its current location, tunneled to the mobile by the home agent?

P8-7 Internet Protocol version 4 (IPv4) is the basis for the original development of Mobile IP by the IETF (Internet Engineering Task Force). IPv4 has many salient features, but does have certain limitation in supporting Internet administration and operation. As a remedy to the limitation of IPv4, the IETF has introduced Internet Protocol version 6 (IPv6) as the answer.

a. What is the basic limitation of IPv4 that prevents it from being used as full deployment for Internet administration and operation? Explain.
b. Operationally, IPv6 is similar to IPv4, but there are basic differences. Discuss the basic differences that makes IPv6 more suitable for mobile Internet deployment.