SRI VENKATESHWARAA COLLEGE OF ENGINEERING & TECHNOLOGY
(Approved by AICTE, New Delhi & Affiliated to Pondicherry University, Puducherry.)
13-A, Villupuram – Pondy Main road, Ariyur, Puducherry – 605 102.
Phone: 0413-2644426, Fax: 2644424 / Website: www.svcetpondy.com

Department of Computer Science and Engineering

Subject Name: **MOBILE COMPUTING** Subject Code: **CS E84**

UNIT III

Mobile Networking: Virtual IP Protocols - Loose Source Routing Protocols - Mobile IP – CDPD – GPRS – UMTS - Security and Authentication – Quality of Service – Mobile Access to the World Wide Web.

---

2 Marks

1. What is mobile networking?

The mobile networking protocol should also be transparent to the hosts and routers which do not understand or support mobility. Thus, the mobility unaware routers should be able to route packets destined to a mobile host as normal IP data packets. Security is another important concern in internetworking. In mobile networking it is more so, since the mobile nodes will be visiting foreign networks, requesting services, and accessing data. Thus, it is important that the security of the visiting network is not breached due to the presence of a foreign node in its network.

2. Write Short notes on virtual IP protocols.

A mobile network is a virtual network with a virtual address space. A mapping is maintained between the physical or actual IP addresses and the Virtual IP addresses. This mapping is performed by the mobile host which obtains a *care of address* from the local network being visited using either the Dynamic Host Configuration Protocol (DHCP) or the BOOTP protocols or by any of the link layer protocols. Below, we describe two Virtual IP protocols.

3. What is loose source routing protocol?

This approach was proposed by David Johnson of CMU in 1993. It uses the Loose Source Route option available in the IPv4 for routing packet data. The option allows the source to specify the intermediate gateways in the IP packet. Thus, the source can control the route the IP packet takes. At each destination, the gateway picks up the next IP address from the IP packet, sets it as the destination, and advances a pointer stored in the IP packet header.

4. Define mobile internet protocol.

The Mobile Internet Protocol (Mobile IP) [88, 80, 19, 7, 59, 76] defines en-hancements to the Internet Protocol to allow routing of IP packets to mobile nodes in the internet. The IP version 4 assumes that the Internet Protocol of a node uniquely identifies the point of attachment of the node to the internet- work. Packets are routed based on the IP address.

5. What are the components of MIP?

The major architecture components of the mobile IP protocol are:
Mobile N o de (MN)
Home Agent (HA)
Foreign Agent (FA)

6. What is DHCP?

The mobile node obtains a temporary IP address on the foreign agent network to be used for forwarding. This can be done using the IETF Dynamic Host Configuration Protocol (DHCP).

7.Explain the routing path of a datagram.

The routing path of a datagram sent from a fixed host to a mobile node is as follows: (1) the datagram is sent from the fixed host to the home agent using standard IP routing; (2) the home agent encapsulates the received datagram inside another datagram and sends it to the foreign agent (IP-in-IP tunneling); (3) the encapsulated IP packet is received by the foreign agent, decapsu- lasted, and forwarded to the mobile node; (4) the mobile node replies by sending a datagram to the fixed host through the foreign agent.

8.Define cellular digital packet data.

CDPD is a connectionless multi-network protocol, proposed originally by the CDPD Forum (now called the WDF Forum). It is based on the early versions of Mobile-IP . The idea behind CDPD is to share unused channels in existing Advanced Mobile Phone Systems (AMPS) to provide up to 19.2 kbps data channel.

9.Explain GPRS.

GPRS is a GSM packet data service developed by the European Telecommunication Standards Institute (ETSI) as part of GSM phase 2+ developments. The goal of GPRS was to support data transfer rates higher than the 9.6 kbps achieved through GSM's circuit switching technology. Unlike Mobile-IP, GPRS is not restricted to IP packet data protocols, and offers connection to standard protocols (such as TCP/IP, X.25, and CLNP) as well as specialized data packet protocols.

10.Explain briefly about security and authentication.

In a mobile computing environment, it is desirable to protect information about the movements and activities of mobile users from onlookers. In addition to the basic security concerns in wire line systems (authentication, confidentiality, and key distribution), a new issue is the privacy and anonymity of the user's movement and identity. In fact, a typical situation arises when a mobile user registers in one domain (home domain) and appears in a different foreign do- main; the user must be authenticated and his solvency must be

confirmed. Usually during this process the user has to provide a non-ambiguous identity to his home domain and has to verify it. If no care is taken, this identity can be tapped on the air interface in a cellular environment or through the signaling protocols exchanged on the registered wired network.

11.What should be the quality of service in networking.

Mobile network protocols such as Mobile-IP and GPRS provide mobility trans- parency at the network layer level. This allows the higher layers of the protocol stack to be used unchanged. Unfortunately, there are ill consequences to this transparency that are mostly attributed to the constraints of the wireless and mobile environment. For example, transport layer protocols that rely heavily on timeout mechanisms for re-transmission, if used unchanged, will perform poorly under variable delays and limited bandwidth.

12.Explain about mobile access to the world.

More and more users are becoming increasingly dependent on information they obtain from the World Wide Web. Users are also demanding ubiquitous access, any time, anywhere, to the information they rely on. Several research efforts ex- plored the problems associated with wireless access to the Web. Most solutions used a Web proxy that enabled Web browsing applications to function over wireless links without imposing changes on browsers and servers. Web proxies are also used to prefetch and cache Web pages to the mobile client's machine, to compress and transform image pages for transmission over low-bandwidth links, and to support disconnected and asynchronous browsing operations.

13.What is wireless WWW?

A prototype consisting of commercially available PDAs and a wireless LAN has been used to provide a "proof of concept" for the Wireless World Wide Web (W4). A simplified version of Mosaic was ported to the PDA for the purpose of experimenting with response time performance and to sort out design choices. A PDA cache was used to improve the performance.

14.Explain the design of MOWSER..

The design is based on a mediator server that filters retrieved information according to the limitations of the mobile unit. Color, resolution, display mode, sound capability, and maximum file size are among the factors considered. The browser, called MOWSER, connects to two servers in the fixed network. The first is the preference server that maintains the user profile; the second is a proxy server that implements all the filtering indicated by the preference server.

15.What are the methods of web access optimizations?

- Web Express optimization methods are summarized below:
  C a c h i n g
  Differencing
  Protocol reduction
  Header reduction

16.What is dynamic URL?

The Mobisaic project at the University of Washington extends standard client browsers to support dynamic URLs and active documents. The Mosaic Web client and the URL syntax are modified so that when the user traverses a dynamic URL, the client resolves any references to dynamic information it may contain and sends the result back to the server. This is helpful in defining location-sensitive resources.

17. Draw the WAP protocol stack diagram.

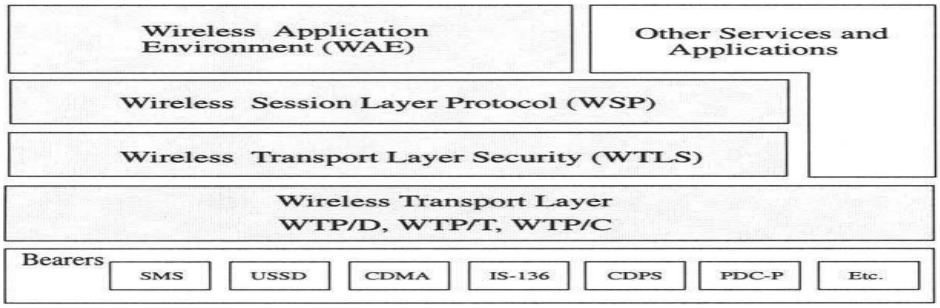


Figure 7.5 The WAP protocol stack

18.Explain the Loss profile approach.

A Loss Profile is proposed and is defined to be a description, provided by the application, of an "acceptable" manner in which data for its connection may be discarded. The loss profile is used in the event of bandwidth reduction at the wireless end of the connection. An elaborate example of a loss profile is given on viewer perception of a video clip under data loss. The loss profile is used by a specialized session layer which is transparent to the application.

19.What are the research efforts that address QoS concerns in the wireless and mobile environment.

- ✓ Optimizing TCP/IP for Mobile Networks
- ✓ QoS driven, full protocol stacks

20.Draw the diagram of GPRS.

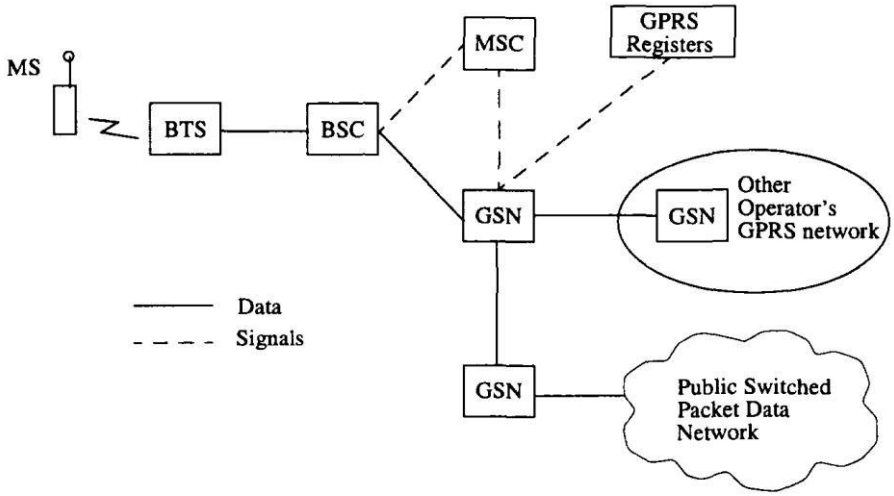


Figure 7.2 *General Packet Radio Service*

21.What is IETF?

The Internet Engineering Task Force (IETF) is a forum of working groups responsible for identifying operational problems and proposing solutions to these problems.

22. What is IRTF?

The Internet Research Task Force (IRTF) is a forum of working groups focusing on long-term research topics related to Internet protocols, applications, architecture, and technology.

11 Marks

1. EXPLAIN MOBILE NETWORKING

Internetworking mobile computers with the fixed-network raises the additional requirements of mobility transparency and mobility and location management. The mobility behavior of a node should be transparent to a peer node. A peer node should be able to communicate with a mobile node using some fixed IP address irrespective of the current point of attachment. The mobile net- working protocol should also be transparent to the hosts and routers which do not understand or support mobility. Thus, the mobility unaware routers should be able to route packets destined to a mobile host as normal IP data packets. Security is another important concern in internetworking.

In mo- bile networking it is more so, since the mobile nodes will be visiting foreign networks, requesting services, and accessing data. Thus, it is important that the security of the visiting network is not breached due to the presence of a foreign node in its network. Authentication of the mobile nodes and foreign networks is also important. Thus, at a minimum, mobile networking protocols should provide authentication and security features comparable to those found in fixed-network IP protocols such as IPv4 and IPv6.

In this section, various approaches and protocols for mobile internetworking

are examined, including:

- Early approaches: virtual IP mechanisms
- Loose source routing protocol
- The Mobile Internet Protocol (Mobile-IP) Cellular Digital Packet Data (CDPD)
- The General Packet Radio Service protocol (GPRS)

Emphasis is placed on protocol mechanisms, leaving out the details which can be obtained by following cited work and web resources.

1.1.1 Early Approaches: Virtual IP Protocols

In this approach, a mobile network is a virtual network with a virtual address space. A mapping is maintained between the physical or actual IP addresses and the Virtual IP addresses. This mapping is performed by the mobile host which obtains a *care of address* from the local network being visited using either the Dynamic Host Configuration Protocol (DHCP) [38] or the BOOTP protocols or by any of the link layer protocols. Below, we describe two Virtual IP protocols.

1.1.1.1 Sunshine And Postal

The earliest solution for managing mobile hosts was proposed by Sunshine and Postal in 1980. They proposed that the mobile hosts be assigned a virtual IP address which can be used to identify them. A mobile host in the foreign network is required to obtain a care-of-address, and to update its location in a mapping database. When a packet has to be routed

to the mobile host, its current location is looked up in the database and the packet is transmitted to that location.

1.1.1.2 The SONY Protocol

This protocol was proposed in 1992 by F.Teraoka et al. of Sony Laboratories. In this scheme, a mobile host has two IP addresses associated with it. A virtual address, which is immutable and by which it is known to the outside world, and a physical address, which is acquired from the local network. Two sub layers are introduced in the network layer and are used to map the physical address to the virtual address. The transport layer interfaces with the network layer through the virtual layer interface and addresses its packets to the virtual address of a mobile host. A cache called the Address Mapping Table (AMT) is used for fast address resolution. A copy of this cache is maintained at each host/router. The VIP (Virtual IP) is implemented as an IP option. A set of packet types is also defined for host communication.

On entering a foreign network, the mobile host obtains an IP address and informs its home network of its current location. The home network broadcasts this information so the AMT cache gets updated. A stationary host, when required to communicate with a mobile host, looks up its cache. If the mapping is available, the packet is transmitted in the normal fashion by appending the VIP header. If the cache entry is not available, the packet is addressed to the VIP address. A set of connection gateways are required for the co-existence of mobility aware and mobility unaware hosts on the network.

1.1.2 Loose Source Routing Protocol

This approach was proposed by David Johnson of CMU in 1993. It uses the Loose Source Route option available in the IPv4 for routing packet data. The option allows the source to specify the intermediate gateways in the IP packet. Thus, the source can control the route the IP packet takes. At each destination, the gateway picks up the next IP address from the IP packet, sets it as the destination, and advances a pointer stored in the IP packet header.

The home network maintains a database of all mobile hosts native to its net- work. When a mobile host changes location, it informs its home network of its new location. When an IP packet destined to the mobile host arrives at the home network, the packet is forwarded to the mobile host at the current location address, and the corresponding source host is informed of the current location of the mobile host. The corresponding host can use this information to cache the location, thus avoiding communication with the home network until the mobile host changes its location again. Source route set up is done by the corresponding host.

Loose Source Routing is an IP option which can be used for address translation. LSR is also used to implement mobility in IP networks.

Loose source routing uses a source routing option in IP to record the set of routers a packet must visit. The destination of the packet is replaced with the next router the packet must visit. By setting the forwarding agent (FA) to one of the routers that the packet must visit, LSR is equivalent to tunneling. If the corresponding node stores the LSR options and reverses it, it is equivalent to the functionality in mobile IPv6.

The name loose source routing comes from the fact that only part of the path is set in advance. This is in contrast with strict source routing, in which every step of the route is decided in advance when the packet is sent.

1.1.3 The Mobile Internet Protocol (Mobile-IP)

The Mobile Internet Protocol (Mobile IP) [88, 80, 19, 7, 59, 76] defines en-hancements to the Internet Protocol to allow routing of IP packets to mobile nodes in the internet. The IP version 4 assumes that the Internet Protocol of a node uniquely identifies the point of attachment of the node to the internet- work. Packets are routed based on the IP address. In a mobile environment, the point of attachment of the mobile node will be different from time to time, and the mobile nodes could be attached to different networks. For IPv4 to work correctly in the mobile environment, the mobile node will either have to be assigned a new IP address every time it changes its point of attachment, or the host specific routing information has to be supplied throughout the network. Both of these alternatives result in scalability and connection management problems. The mobile IP protocol describes a mechanism which allows nodes to change their point of attachment on the Internet.

The major architecture components of the mobile IP protocol are:

- Mobile Node (MN): is a host or a router that changes its point of attachment to the network from one sub network to another. The MN is known throughout the network by an IP address assigned to it in the home net- work. The mobile node can communicate from any location as long as the link layer connectivity to the internetwork is established.
- Home Agent (HA): is a mobile-IP capable router on the mobile node's home network. The HA maintains the location information for the mobile node. It also acts as the tunneling agent for packets destined to the mobile node. The HA manages the registration and authorization information of all the mobile modes belonging to its network.
- Foreign Agent (FA): is a mobile-IP capable router that the mobile node has visited. After attaching to the foreign network, the mobile node is required to register itself with the FA. The FA detunnels and routes the packets destined to the mobile node. The FA may also act as a default router for mobile nodes registered with it.

The mobile IP protocol can be summarized as follows:

1. The mobility agents (HA and FA) in the network broadcast their avail-ability through agent advertisement packets.

2. The mobile node, after connecting to a network, receives information about the mobility agents through the agent advertisement broadcasts. Alternatively, the mobile nodes can solicit the agent information if no broadcasts have been received.

3. The mobile node determines the network it is attached to. If it is connected to the home network, it operates without mobility services. If it is returning back to the home network, the mobile node deregisters itself with the HA and operates without mobility services.

4. If the mobile node is attached to a foreign network, a care-of-address is obtained from the FA.

5. The mobile node operating from a foreign network registers itself with its home agent. The foreign agent then acts as a relay in this registration process.

6. When the mobile node is away from its home network, datagrams destined to the mobile node are intercepted by the home agent, which then tunnels these datagrams to the mobile node's care-of-address. The tunneled pack-ets destined to the mobile node are detunneled either by the foreign agent or by the mobile node itself. In the latter case, the mobile node obtains a temporary IP address on the foreign agent network to be used for forwarding. This can be done using the IETF Dynamic Host Configuration Protocol (DHCP).

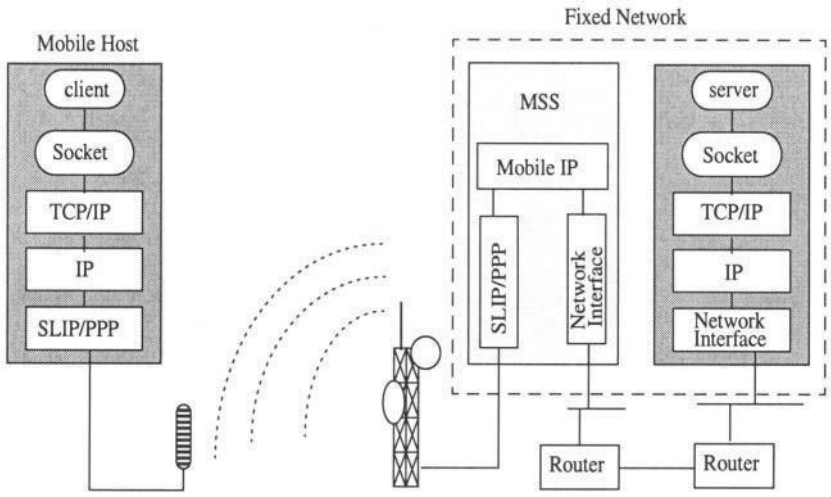


Figure 7.1 Mobile Internet architecture using Mobile-IP

7. The datagrams originating from the mobile node are routed in the normal fashion. The foreign agent may act as a default router in this case.

The routing path of a datagram sent from a fixed host to a mobile node is as follows: (1) the datagram is sent from the fixed host to the home agent using standard IP routing; (2) the home agent encapsulates the received datagram inside another datagram and sends it to the foreign agent (IP-in-IP tunneling); (3) the encapsulated IP packet is received by the foreign agent, DE capsulated, and forwarded to the mobile node; (4) the mobile node replies by sending a datagram to the fixed host through the foreign agent. The Mobile IP protocol stack on the fixed network and on the mobile unit is depicted in Figure 7.1.

The Mobile Host Protocol, known as Mobile-IP, is an evolving standard being developed by the IETF Working Group on IP Routing for Wireless/Mobile Hosts. Standards for both IPv4 and IPv6 have been proposed and are being reviewed for enhancements in scalability and performance. In particular, the triangular routing between the mobile node, the home agent, and the foreign agent (that must be performed every time the mobile node switches over to another communication cell) is a bottleneck that is being removed in IPv6 [81]. Packets addressed to
the mobile node's home address are transparently routed to its care-of address. The optimized protocol enables IPv6 nodes to cache the binding of a mobile node's home address with its care-of address, and to then send any packets destined for the mobile node directly to it at this care-of address.

The Mosquito Net project at Stanford aimed at relaxing the requirement of foreign agent availability. Mosquito Net follows the IETF specification of Mobile-IP to support host mobility, but does not require FA support in foreign networks visited by the mobile node.

More details on achieved and ongoing efforts in Mobile IP and its routing optimization can be found in [62, 60, 14, 11, 88, 80].

1.1.3.1 Support for Ad-Hoc Mobility

An ad-hoc mobile network is a collection of wireless mobile nodes forming a temporary network without the aid of any established infrastructure or centralized administration. Examples of ad-hoc networks include wireless portable devices of a group of collaborator, such as an emergency team in a disaster area. No routing is needed between ad-hoc nodes which are within transmission range of each other's. Otherwise, additional nodes must be used to form a sequence of hops from the source to the destination. Routing algorithms in the ad-hoc environment are therefore a necessary support for this mode of mobile connection.

Traditional routing algorithms used in wire line networks use distance vector or link state routing algorithms, which rely on periodically broadcasting routing advertisements by each router node. The distance vector algorithm [55] broad- casts its view of the distance from a router node to each host. The link state routing algorithm broadcasts its view of the adjacent network links. Neither algorithms is suitable for the ad-hoc environment because periodic broadcasts will drain battery power quickly.

Research in ad-hoc routing is dedicated to finding algorithms that avoid the needless battery consumption and the inefficient use of the wireless bandwidth. Dynamic source routing is one such algorithms due to Johnson and Maltz. It allows for route discovery, route maintenance, and the use of route caches. To discover an available route, a source node sends out a route request packet indicating the source, the target nodes, and a request identifier. When a mobile node receives a route request packet, it checks a list of recently processed requests. If a request is found for the same source and request id, the request is dropped and no further action is taken. Otherwise, the address of the node servicing the request is added to the route request packet before the packet is re-broadcasted. However, if the address of the node servicing the request is identical to the target node address, the requested route is discovered, and a reply is sent to the source node.

Due to unpredictable node mobility, cached routes may become incorrect. Route maintenance is therefore necessary in this environment. This is achieved by requiring nodes routing packets to acknowledge successful forwarding and to send error messages to the
source node if a route ceases to exist. Active monitoring such as MAC -level acknowledgements, as well as passive monitoring (listening to nearby broadcast, in a promiscuous mode), can be used in route maintenance.

Other recent ad-hoc routing protocols that can be found in the literature include the on-demand distance vector routing, the Location-Aided Routing (LAR) algorithm, and the Distance Routing Effect Algorithm.

1.1.4 Cellular Digital Packet Data (CDPD)

CDPD is a connectionless multi-network protocol, proposed originally by the CDPD Forum (now called the WDF Forum). It is based on the early versions of Mobile-IP . The idea behind CDPD is to share unused channels in existing Advanced Mobile Phone Systems (AMPS) to provide up to 19.2 kbps data channel.

Even though CDPD and Mobile-IP are similar, their terminologies are different. CDPD follows the OSI model terminology. For example, the mobile node is called a Mobile End-System (M-ES); the home and foreign agents are called Mobile Home and Mobile Serving Functions (MHF and SF respectively) and re- side in a mobile data intermediate system (MD-IS). A Mobile Database Station (MDBS) is also defined which deals with the air link communications and acts as a data link layer relay between the M-ES and the serving MD-IS . Two pro-tools, the Mobile Node Registration Protocol (MNRP) and the Mobile Node Location Protocol (MNLP), are responsible for registration of the M-ES with its home MD-IS and the proper routing of packets destined for the M-ES.

The main resemblance between CDPD and Mobile-IP is in the triangular rout- ing approach between the mobile node and the home and foreign agents. The main differences can be summarized as follows:

- The user's IP address must be assigned by the CDPD service provider. Mobile IP makes no such assumptions.
- Mobile IP allows the mobile node to also be a foreign agent. Combining the M-ES and the Serving MD-IS was not considered and is not practical in CDPD.
- CDPD's mobility tunneling is based on CLNP. Mobile IP's mobility tunneling is based on the IP-in-IP protocol, which is IP-based.
- Mobile IP operates completely above the data link layer. CDPD mobility, on the other hand, is mostly above the data link layer.
- Since the infrastructure of the CDPD network is closed there are less secrudity considerations for CDPD.

While the standardization process of Mobile IP has been progressing rather slowly, CDPD has been deployed for a few years now, and is receiving the support of major AMPS carriers. However, due to its lack of openness, the future of CDPD deployment and/or acceptance can only be guessed.

What is CDPD?

Cellular Digital Packet Data (CDPD) is a technique used for transmitting small chunks of data, commonly referred to as packets, over the cellular network in a reliable manner. It allows users to send and receive data from anywhere in the cellular coverage area at any time, quickly and efficiently. CDPD technology provides extensive, high speed (data can be sent over the Airlink at a rate of 19.2 kilobits per second), high capacity, cost effective data services to mobile users. With this technology, both voice and data can be transmitted over existing cellular channels.

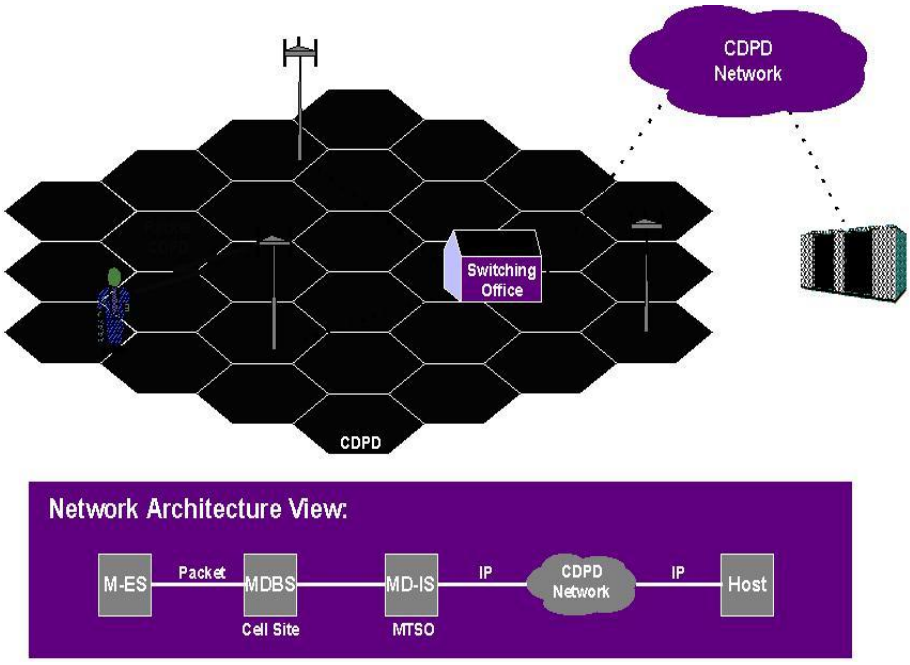
What defines the CDPD Network

By building CDPD as an overlay to the existing cellular infrastructure, and using the same frequencies as cellular voice, carriers are able to minimize the capital expenditures required to offer the service while offering the same coverage area (footprint) their customer base has grown accustomed to. In comparison, it costs approximately $1 million to build out a new cellular cellsite and only about $50,000 to build the CDPD overlay to an existing site.

The CDPD overlay network is made up of a combination of key components that operate together to provision the overall service. These components are described below:

- The Mobile End System (M-ES) which is defined as any mobile computing device which is equipped with a CDPD modem (e.g. a PC). Unlike voice cellular phones, the decision to initiate a transfer, or hand-off from one cell to another cell is under the control of the CDPD M-ES itself, as it is the M-ES which is responsible for monitoring the received signal strength of the cellular channels being used.
- The Fixed End System (F-ES) which is defined as a stationary computing device, such as a host computer or an on-line information service.
- The Mobile Data Intermediate System (MD-IS) which is a stationary network component with similar responsibilities to the cellular voice switch. It is responsible for keeping track of the M-ES's location and routing data packets to and from the CDPD network and M-ES appropriately. It has been referred to as the "brain" of the network, because of its functionality. Not only is it responsible for ensuring that an M-ES is valid to log on to the network, but it also stores information on the M-ES's last known location, traffic statistics and billing information.
- The Mobile Data Base Station (MDBS) is primarily responsible for RF channel management. It is located at the voice cell sites and is responsible to instruct the M-ES to "hop" to new channels for continued communication in the event voice communication (which is the priority traffic) is detected. It also handles the leg work for the M-ES in locating new channels when a hand off is required between cell sites.
- The Intermediate System (IS) is made up of (off the shelf) routers which are CDPD compatible with the primary responsibility for relaying the data packets.

The way these componenets interact with each other can be seen from the graphic depiction below:



How does CDPD work?

To effectively integrate voice and data traffic on the cellular system without degrading the level of service provided to the voice customer, the CDPD network implements a technique called channel hopping. The way this works is that when a CDPD mobile data unit desires to initiate data transmission, it will check for availability of a cellular channel. Once an available channel is located, the data link is established.

As long as the assigned cellular channel is not needed for voice communications, the mobile data unit can continue to transmit data packet bursts on it. However, if a cellular voice customer initiates voice communication, it will take priority over the data transmission. At such time, the mobile data unit will be advised by the Mobile Data Base Station (which is the CDPD serving entity in the cell and constantly checks for potential voice communication on the channel) to "hop" to another available channel. In the event that there are no other available channels, then data transmission will be temporarily discontinued. It is important to note that these channel hops are completely transparent to the mobile data user. As far as the user can see, there is only one data stream being used to complete the entire transmission.

1.1.5 The GSM General Packet Radio Service (GPRS)

GPRS is a GSM packet data service developed by the European Telecommunication Standards Institute (ETSI) as part of GSM phase 2+ developments. The goal of GPRS was to support data transfer rates higher than the 9.6 kbps achieved through GSM's circuit switching technology. Unlike Mobile-IP, GPRS is not restricted to IP packet data protocols, and offers connection to standard protocols (such as TCP/IP, X.25, and CLNP) as well as specialized data packet protocols. Mobile-IP, however, influenced the design of Mobility management in GPRS.

Figure 7.2 shows the architecture of a GSM system that uses GPRS. In addition to the Base Transceiver Station (BTS), Base Station Controller (BSC), and the Mobile Switching Center (MSC), a new logical network node called the GPRS support node (GSN) was introduced in order to create an end-to-end packet transfer mode. Physically, the GSN can be integrated with the mobile switching center (MSC), or it can be a separate network element based on the architecture of data network routers. GSN is a mobility router that provides connection and interoperability with various data networks, mobility management with the GPRS registers, and delivery of data packets to MSs , independently of their locations.

One GSN is designated the Gateway GSN (GGSN) and acts as a logical inter- face to external packet data networks. The GGSN is similar to the home agent in Mobile-IP . It updates the location directory of the mobile station (MS) using routing information supplied by the Serving GSN node (SGSN). The latter is similar to the foreign agent in Mobile-IP. GGSN also routes the external data network protocol packet encapsulated over the GPRS backbone to the SGSN currently serving the MS. It also de-capsulate and forwards external data net- work packets to the appropriate data network and handles the billing of data traffic.

The SGSN is responsible for the delivery of packets to the mobile stations within its service area. The main functions of the SGSN are to detect new GPRS MSs in its service area, handle the process of registering the new MSs along with the GPRS registers, send/receive data packets to/from the GPRS MS, and keep a record of the location of MSs inside of its service area. The GPRS register acts as a database from which the SGSNs can

ask whether a new MS in its area is allowed to join the GPRS network. For the coordination of circuit and packet switched services, an association between the GSM MSC and the GSN is created. This association is used to keep routing and location area information up-to-date in both entities.

1.1.6 Security and Authentication Issues in Mobile Networks

In a mobile computing environment, it is desirable to protect information about the movements and activities of mobile users from onlookers. In addition to the basic security concerns in wire line systems (authentication, confidentiality, and key distribution), a new issue is the privacy and anonymity of the user's movement and identity. In fact, a typical situation arises when a mobile user registers in one domain (home domain) and appears in a different foreign do- main; the user must be authenticated and his solvency must be confirmed. Usually during this process the user has to provide a non-ambiguous identity to his home domain and has to verify it. If no care is taken, this identity can be tapped on the air interface in a cellular environment or through the signaling protocols exchanged on the registered wired network.

In CDPD , all the mobility management, as well as security-related activity, are concentrated in the Massage-Data Intermediate System (MD-IS) . Each MD-

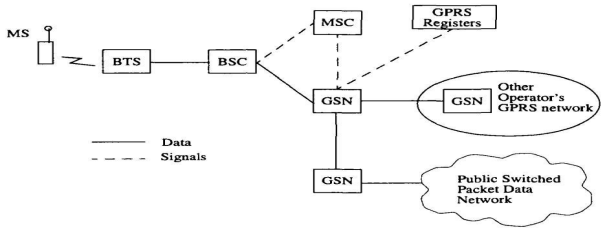


Figure 7.2 *General Packet Radio Service*

IS controls an area covered by a number of base stations. Upon arrival to a new area, the mobile unit engages in a *Diffie-Hellamn* key exchange protocol with the local MD-IS. As a result, both parties obtain a shared secret key. Subsequently, the mobile unit encrypts its real identity (Network Equipment Identifier) and transmits it to the local MD-IS. This approach allows the local MD-IS to discover the real identity of the mobile unit. Unfortunately, the key exchanging protocol itself is not secure. This means that an active attacker masquerading as the local domain authority can engage in the key exchange protocol with the mobile unit and obtain a shared key.

2. EXPLAIN QUALITY OF SERVICE IN MOBILE NETWORKS

Mobile network protocols such as Mobile-IP and GPRS provide mobility transparency at the network layer level. This allows the higher layers of the protocol stack to be used unchanged. Unfortunately, there are ill consequences to this transparency that are mostly attributed to the constraints of the wireless and mobile environment. For example, transport layer protocols that rely heavily on timeout mechanisms for re-transmission, if used unchanged, will perform poorly under variable delays and limited bandwidth.

This is especially true for applications that require continuous-media streams. Existing session protocols are not of much use under frequent disconnections and reconnections of the same mobile computation. Similarly, existing presentation layer protocols are inappropriate to use unchanged, in the wireless and mobile environment. For example, a user with a limited

display and limited battery PDA will not be able to browse the Web unless the presentation of the downloaded data is changed to suite her PDA's capabilities. Regardless of which particular upper layer in the protocol stack suffers the consequences of transparency, the effect on the end-user will always be felt as unacceptable fluctuations in the perceived QoS.

In this section, we describe the following three research efforts that address QoS concerns in the wireless and mobile environment.

- *Optimizing TCP/IP for Mobile Networks.* Transport or network layer solutions to get TCP/IP to work despite the fluctuations in the underlying network QoS. Solutions are not application-sensitive and do not address an overlay of heterogeneous networks.

- *QoS driven, high-level communication protocols.* Session and/or application layer protocols directly addressing QoS parameters. Solutions are sensitive to applications, but do not address network heterogeneity issues.

- *QoS driven, full protocol stacks.* All layers are aware of either QoS or the limitations introduced by mobility and by the wireless networks. Two research efforts will be discussed including, the BARWAN project and the Wireless Application Protocol (WAP) standard.

2.2.1 Optimizing TCP/IP for Mobile Networks

Since mobile users will need connection-oriented communication to obtain re- mote services, they will have to use transport protocols developed for the fixed network. Unfortunately, such protocols like TCP perform poorly when used un- modified in the mobile network. For example, TCP acknowledgment timeout is in the range of tens of milliseconds. A mobile unit crossing cell boundaries blanks out during a hand-off procedure that could last up to 1,000 milliseconds. This leads to sender timeouts and repeated re-transmissions.

Another source of re-transmission is the high error rate inherent in the wireless transmission characteristics. Another problem that can lead to performance degradation un- der standard TCP is bandwidth allocation under unpredictable mobility. An unpredicted number of mobile users can move into the same cell, thus competing on sharing the limited wireless link. Under this scenario, it is difficult to build applications or services that provide performance guarantees or quality of service. A few approaches have been proposed to optimize and extend the standard TCP protocol so that it can be used efficiently under a mobile network protocol such as Mobile IP.

2.2.1.1 Yavatkar et al

Yavatkar et al proposed an approach whereby the communication path be- tween the mobile end and the fixed end is split into two separate connections: one over the wireless link and another over the wired links. The connection over the wireless link may either use regular TCP or a specialized transport protocol optimized for better performance. The splitting of a connection is transparent to an application and no changes are necessary to protocol software on the stationary hosts. A new session layer protocol called Mobile Host Protocol (MHP) is introduced atop standard TCP. MHP compensates for wireless link characteristics and for host migration.

It is located at both the base station and the mobile host. An advantage of this approach is that performance degradation in TCP is limited to a "short" connection over the wireless hop, while traffic over the "long" connection over the wired network can be protected from the impact of erratic behavior over the wireless link. A second alternative is proposed in the same work which is

similar to the MHP alternative except that MHP uses a specialized protocol instead of TCP over the wireless hop. The specialized protocol differs from standard TCP in that the former uses selective acknowledgement by the receiver, in which a bitmask is used to indicate all missing segments of the connection stream. This way, the recovery of all losses can be performed via a single round trip message, resulting in a better throughput performance.

Another approach similar to Yavatkar's is the I-TCP protocol (Indirect Trans- port Layer Protocol), which also splits the communication path between the mobile host and the fixed network host into two connections; the first be- tween the mobile host and the base station, over the wireless link, using the I-TCP protocol; and the second between the base station and the fixed network host using standard TCP.

3.2.1.2 Balakrishnan et al

Balakrishnan et al took a slightly different approach to improve the performance of TCP in the mobile network. They focused on the re-transmission behavior of TCP due to hand-off. They redesigned the network layer so that it caches packets at the base stations. Retransmission can therefore be performed locally between the base station and the mobile unit. The gain is that the erratic transmission characteristics of the wireless link are dealt with in isolation of the rest of the fixed network. Experimental evaluation showed a throughput increase of up to 20 times over standard TCP. Their results are based on the Lucent/NCR Waveland network.

2.2.1.3 Caceres et al

Similar research by Caceres and Iftode addressed the problem of communication pauses due to hand-off. They observed that such pauses are interpreted by standard TCP (Tahoe in their experiment) as packet losses due to congestion, which consequently causes retransmissions that get further timed out during the hand-off. They proposed using the fast re-transmission option available in TCP-Tahoe immediately after hand-off is completed. Their experimental verification shows clear smoothening of TCP performance during hand-off.

2.2.2 QoS Driven, High-Level Communication Protocols

Optimizing the behavior and performance of transport protocols is not sufficient to maintain the QoS required by applications. For example, most Web browsers use multiple TCP connections to access a multimedia page. While this parallelism achieves speedup in the fixed network, it is slow and inappropriate in the wireless and mobile environment. In addition to transport optimizations, what was found needed are application-aware (or application-specific) mechanisms to monitor, request, and maintain QoS from the application or user point of view. This section describes high-level, above-transport protocols that understands application QoS requirements and resource limitations.

2.2.2.1 The Loss Profile Approach

Seal and Singh considered the problem of unpredictable mobility and its effect on the degradation of the wireless communication performance. They addressed the case where the aggregate bandwidth required by all mobile units in an overloaded cell exceeds the cell's available bandwidth. Their mechanism is simple and relies on policies and measures for discarding parts of the data of the mobile users. Instead of discarding data in an arbitrary manner, guidelines are proposed to avoid discarding critical portions of the data. A Loss Profile is proposed and is defined to be a

description, provided by the application, of an "acceptable" manner in which data for its connection may be discarded. The loss profile is used in the event of bandwidth reduction at the wireless end of the connection. An elaborate example of a loss profile is given on viewer perception of a video clip under data loss. The loss profile is used by a specialized session layer which is transparent to the application.

2.2.2.2 QEX: The QoS Driven Remote Execution Protocol

In the problem of fluctuations in the quality of service (QoS) in a federation of heterogeneous networks is addressed. The work describes a design of a distributed system platform that supports the development of adaptable services. The design allows services to tolerate the heterogeneity of the environment by dynamically adapting to changes in the available communication QoS.

The implementation of the distributed system is based on APM Ltd.'s ANSAware software suite, which is based on the ANSA architecture that has had some influence on the ISO Reference Model for Open Distributed Processing (RM-ODP). The purpose of this effort is to propose extensions to emerging distributed systems standards in order to support mobile services. The basic ANSAware platform is extended to support operation in the mobile environ- ment by introducing the notion of explicit bindings, which is a QoS-aware RPC protocol for objects called QEX.

Explicit bindings allow application programmers to specify QoS constraints on bindings between objects, and to detect violations of these constraints at run time. To support explicit bindings, a new remote procedure call protocol has been developed for ANSAware. The new RPC is able to maintain QoS information on the underlying communications infrastructure and to adapt to changes in the perceived QoS. Moreover, it is able, via explicit bindings, to pass on relevant QoS information to interested applications. This allows the applications themselves to adapt to changes in the QoS. Binding parameters include specification of parameters such as the desired throughput, latency, and jitter associated with the binding. Clients are returned a binding control interface as a result of an explicit bind operation. To control the QoS of the flow once the binding has been established, the control interface includes a pair of operations *setQoS()* and *getQoS()*. These operations take as arguments a set of QoS parameters which can then be passed by the stream binding to the underlying transport protocol. A call-back mechanism is also provided to inform client objects of QoS degradations reported by the underlying transport service.

The work is being put to test using an adaptive collaborative mobile application designed to support field engineers in the U.K. power distribution industry.

2.2.3 QoS Driven, Full Protocol Stacks

Future mobile services will be built upon federations of heterogeneous networks maintained and administered by different providers. The mobility of users will force an application to migrate along overlays of networks that vary in their bandwidth, latency, range, and transmission characteristics. Unless the application adapts to variations in the network overlay, the application performance is bound to suffer. A network overlay can include a cellular network, a personal communication system (PCS), a wireless LAN, an Internet connection, and/or a satellite communication loop, among other networks. In addition to the het- erogeneity of networks, the heterogeneity of the mobile platforms imposes a great impediment to mobile application portability. Unless applications adapt to the capabilities and limitations of the mobile computer with respect to the type and media of communicated data, applications will remain

proprietary to the specific mobile computer platforms they were originally designed for. This section describes a research project that proposes a full stack solution as an overlay network stack atop a heterogeneous collection of wireless subnets. This section also describes an ongoing standardization effort called WAP that aims at proposing a specification of a full ISO/OSI-like network stack that is wireless and mobile aware.

2.2.3.1 BARWAN: The Wireless Overlay Network Architecture

The BARWAN project at the University of California at Berkeley developed an architecture that supports applications' graceful adaptation to the available bandwidth and latency of the wireless network. The architecture assumes an overlay of various wireless networks ranging from regional-area, wide-area, metropolitan-area, campus-area, in-building, and in-room wireless networks. A testbed of wireless overlay network management that supports media-intensive applications has been used to demonstrate the adaptability features of BARWAN. The testbed that has been developed in the San Francisco Bay Area includes the participation of over six local carriers including Nextel and Metricom. The testbed integrates the participants' networks and allows full coverage of the greater Bay Area. The BARWAN architecture is gateway- centric, meaning it provides gateway connections from the mobile host to each participating wireless networks. Medical imaging applications have been developed to drive the testbed.

The layered architecture of BARWAN is shown in Figure 7.3. It shows all layers designed for wireless overlay network integration and for providing application support. The lowest layer is the wireless overlay subnets, which are the carrier networks including data link interface, and possibly carrier network rout- ing. The details of this layer depends on the specific subnets being integrated. Next is a layer called the Overlay Network Management Layer which includes network and transport functionalities including location tracking, QoS-based hand-off management, other QoS services, and connection-oriented transport mechanisms. The next higher up layer is the Session Management Layer which provides a "transactional" transport (called message-oriented interface).

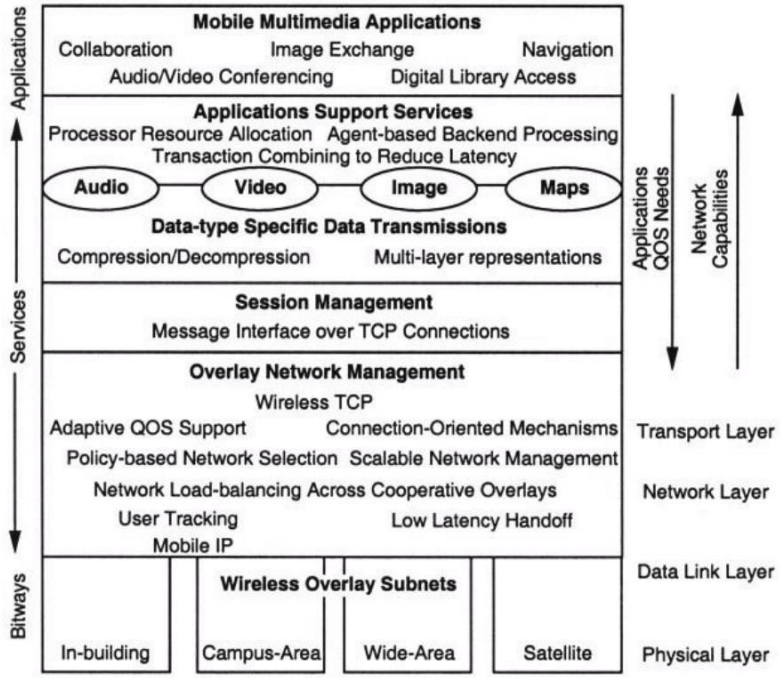


Figure 7.3 The BARWAN conceptual layered architecture

The layer attempts to optimize transport connections related to the same application by session sharing whenever possible. On top of the session layer is the Application Support Services including support for various data types and continuous media such as audio and video. Finally, the mobile multimedia application is on top of the stack. Figure 7.3 also shows how the quality of services needs pass down the layers from applications towards the network management layers, while information about network capabilities propagates up the layers.

2.2.3.2 The Wireless Application Protocol (WAP)

In June 1997, Ericsson, Nokia, Motorola, and Phone.Com (previously Unwired Planet) formed a consortium for the standardization of an open middleware architecture for wireless application. The objective was to create the specification of a wireless application environment and a wireless ISO/OSI-like protocol stack. The goal was to provide the needed interoperability to connect different portable devices, via heterogeneous wireless networks, into the internet and corporate intranets. The focus was to bring the internet content and advanced services to digital cellular phones and other hand-held devices such as smart communicators and PDAs.

In January 1998, the consortium created a nonprofit company named the WAP Forum with the mission of enabling: (1) interoperability across heterogeneous portable devices, wireless networks, and internet contents, and (2) portability of third party wireless software and applications across different portable devices that are WAP-compliant. Currently, the WAP Forum is creating a set of specifications for the Wireless Application Environment and for each layer in the WAP protocol stack.

The architectural infrastructure of WAP is depicted in Figure 3.4 and consists of: (1) hand-held devices ranging from digital cellular phones, to smart communicators such as the Nokia 9000, to palmtop computers. Only devices that will be WAP-compliant (implement the WAP stack and wireless application environment) are part of the WAP infrastructure, (2) Wap-compliant wireless networks, which are carrier networks augmented with the WAP stack on top of the

air link interfaces, (3) WAP-compliant internet information providers such as Web servers, that must conform to levels of presentations of information suitable to the capabilities of the hand-held device requesting the information, and (4) WAP-compliant TeleVAS providers.

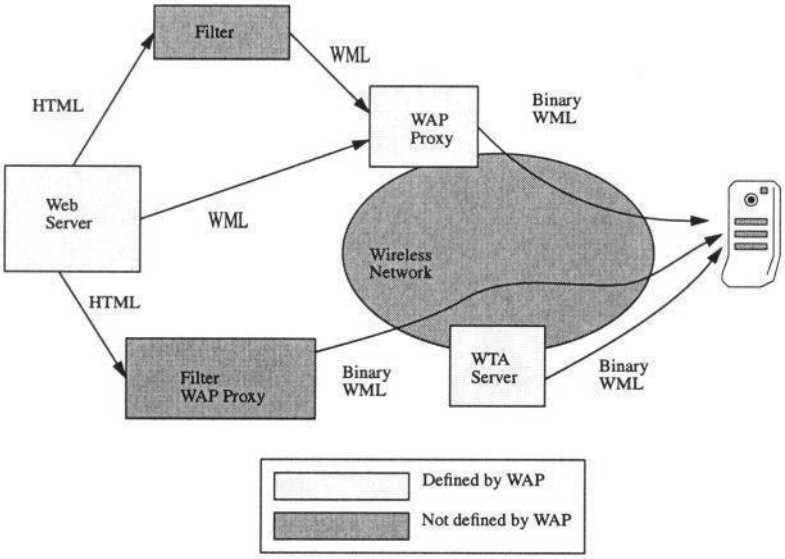


Figure 7.4 The WAP architectural infrastructure

In the heart of the WAP standard is the WAP protocol stack shown in Fig- ure 7.5. The stack is similar to the ISO/OSI stack and consists of a lowest layer containing air link interfaces such as GSM's GPRS, CDPD, D-AMPS, among others. This lowest layer corresponds to both the physical and the data link layers combined in the OSI stack. On top of the air link is the transport layer, in which datagram and connection-oriented streams are supported. Also, *transactional* connections are supported to enable electronic commerce applications. This layer corresponds to both the network and the transport layers of the OSI stack combined. On top of the transport, WAP dedicates a layer for security. This includes encryption, authentication, and capabilities. On top of security is the session layer which is responsible for enabling multi-tasking on the hand-held device. This is because multiple connections can be maintained as multiple sessions managed by the session layer. The session layer, which is the most elaborate layer, also contains critical QoS features including:

- exception mechanisms to allow applications to register interest in QoS related network events and parameter thresholds. This allows the application to be mobility-aware, by using QoS API to program how to adapt to changes in the environment.

- mechanisms for capability and content negotiation. This will enable the WAP stack itself to partner through its pieces (on the fixed network, on the wireless network gateways, and on the hand-held device) to perceive and adapt to the mobility and the changes in network characteristics. When certain information content is being delivered, the WAP stack negotiates with the device the capability to receive and display the contents. The negotiation decides for the feasibility of the transfer and for the level of filtering that might be needed to deliver the the information while maintaining QoS.

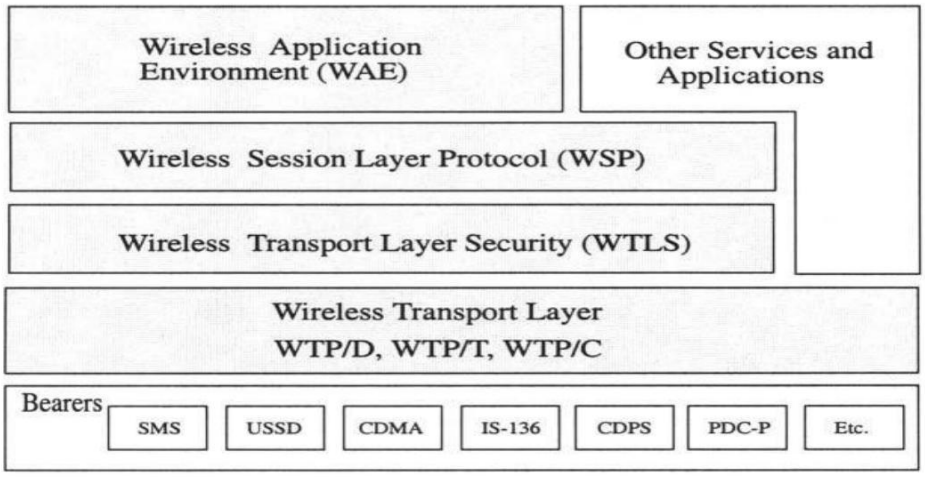


Figure 7.5 The WAP protocol stack

The first capability provides applications with the environment awareness needed to initiate QoS adaptations. The second capability, on the other hand, provides the system with automated awareness mechanisms not only of the environment, but also of the device capabilities and the characteristics of the information content.

Another standardization effort similar to WAP is the Mobile Network Computer Reference Profile (MNCRF), which is based on the NCRF standard developed by the Open Group. The first draft of the standard has been released addressing the unique requirements of mobile network computing. Details of this initiative are available in a white paper and a reference specification document.

3. EXPLAIN MOBILE ACCESS TO THE WORLD WIDE WEB

More and more users are becoming increasingly dependent on information they obtain from the World Wide Web. Users are also demanding ubiquitous access, anytime, anywhere, to the information they rely on. Several research efforts explored the problems associated with wireless access to the Web. Most solutions used a Web proxy that enabled Web browsing applications to function over wireless links without imposing changes on browsers and servers. Web proxies are also used to prefect and cache Web pages to the mobile client's machine, to compress and transform image pages for transmission over low-bandwidth links, and to support disconnected and asynchronous browsing operations.

3.3.1 The Wireless WWW (W4)

In a prototype consisting of commercially available PDAs and a wireless LAN has been used to provide a "proof of concept" for the Wireless World Wide Web (W4). A simplified version of Mosaic was ported to the PDA for the purpose of experimenting with response time performance and to sort out design choices. A PDA cache was used to improve the performance.

3.3.2 Dynamic Documents

The concept of dynamic documents was introduced in as an approach to extending and customizing the WWW for mobile computing platforms. Dynamic documents are programs executed on a mobile platform to generate a document; they are implemented as Tcl scripts as part of the browser client.

A modified version of the NCSA Mosaic browser was used to run the dynamic documents it retrieves through a modified Tcl interpreter. The interpreter is designed to execute only commands that do not violate safety. By using dynamic documents, an adaptive e-mail browser that employs application-specific caching and prefetching is built. Both the browser and the displayed e-mail messages are dynamically customized to the mobile computing environment in which they run.

Dynamic documents can solve the problem of limited resources in the mobile host. For example, the Tcl script could be a filter that reduces an incoming image so that it fits the screen size or resolution. Unfortunately, dynamic documents being placed at the client side are not wireless-media sensitive. This is because filtering occurs after all transmitted information is received by the client. Although caching and prefetching can alleviate some of the communication overhead, excess data (that would be reduced by the dynamic document) is, however, communicated, leading to inefficient utilization of the wireless bandwidth.

3.3.3 Dynamic URLs

The Mobisaic project at the University of Washington extends standard client browsers to support dynamic URLs and active documents. The Mosaic Web client and the URL syntax are modified so that when the user traverses a dynamic URL, the client resolves any references to dynamic information it may contain and sends the result back to the server. This is helpful in defining location-sensitive resources. Active documents are Web pages that notify the client browser when dynamic information changes. This feature also supports location-sensitive information by keeping the mobile client aware of service relocation or of services offered by a mobile server.

3.3.4 Mobile Browser (MOWSER)

In [65], a design of a mobile-aware Web browser is discussed. The design is based on a mediator server that filters retrieved information according to the limitations of the mobile unit. Color, resolution, display mode, sound capability, and maximum file size are among the factors considered. The browser, called MOWSER, connects to two servers in the fixed network. The first is the preference server that maintains the user profile; the second is a proxy server that implements all the filtering indicated by the preference server.

MOWSER assumes that the user is aware of the mobile unit limitations, which in a way sacrifices transparency. Similar to the dynamic document approach, MOWSER does not directly consider the limitations of the wireless media (although the maximum file size indirectly preserves the limited bandwidth).

3.3.5 Web Express

Web Express uses the proxy approach to intercept and control communications over the wireless link for the purposes of reducing traffic volume and optimizing the communication protocol to reduce latency. Two components are inserted into the data path between the Web

client and the Web server: (1) the Client Side Intercept (CSI) process that runs in the client mobile device and (2) the Server Side Intercept (SSI) process that runs within the wired and fixed network (see Figure 7.6). The CSI intercepts HTTP requests and, together with the SSI, performs optimizations to reduce bandwidth consumption and transmission latency over the wireless link. From the viewpoint of the browser, the CSI appears as a local Web proxy that is co-resident with the Web browser.

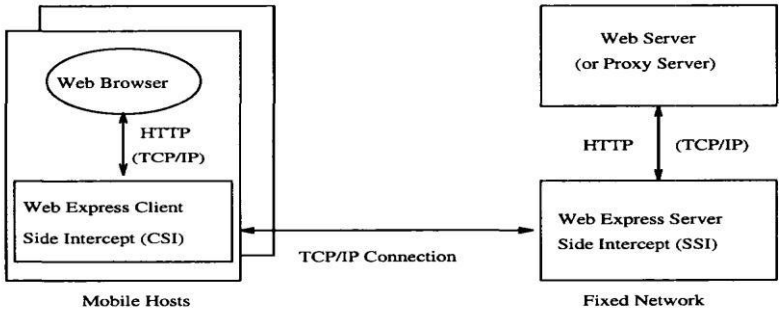


Figure 7.6 WebExpress proxy intercept model

On the mobile host, the CSI communicates with the Web browser over a local TCP connection (using the TCP/IP "loopback" feature) via the HTTP protocol. Therefore, no external communication occurs over the TCP/IP connection between the browser and the CSI. No changes to the browser are required other than specifying the (local) IP address of the CSI as the browser's proxy address. The CSI communicates with an SSI process over a TCP connection using a reduced version of the HTTP protocol. The SSI reconstitutes the HTML data stream and for- wards it to the designated CSI Web server (or proxy server). Likewise, for responses returned by a Web server (or a proxy server), the CSI reconstitutes an HTML data stream received from the SSI and sends it to the Web browser over the local TCP connection as though it came directly from the Web server.

The proxy approach implemented in Web Express offers the transparency ad- vantage to both Web browsers and Web servers (or proxy servers) and, there- fore, can be employed with any Web browser. The CSI/SSI protocols facilitate highly effective data reduction and protocol optimization without limiting any of the Web browser functionality or interoperability. Web Express optimization methods are summarized below:

- *Caching:* Both the CSI and SSI cache graphics and HTML objects. If the URL specifies an object that has been stored in the CSI's cache, it is returned immediately as the response. The caching functions guarantee cache integrity within a client-specified time interval. The SSI cache is populated by responses from the requested Web servers. If a requested URL received from a CSI is cached in the SSI, it is returned as the response to the request.

- *Differencing:* CSI requests might result in responses that normally vary for multiple requests to the same URL (e.g., a stock quote server). The concept of differencing is to cache a common base object on both the CSI and SSI. When a response is received, the SSI computes the difference between the base object and the response and then sends the difference to the CSI. The CSI then merges the difference with its base form to create the browser response. This same technique is used to determine the difference between HTML documents.

- *Protocol reduction:* Each CSI connects to its SSI with a single TCP/IP connection. All requests are routed over this connection to avoid the costly connection establishment overhead. Requests and responses are multiplexed over the connection.

- *Header reduction:* The HTTP protocol is stateless, requiring that each request contain the browser's capabilities. For a given browser, this information is the same for all requests. When the CSI establishes a connection with its SSI, it sends its capabilities only on the first request. This information is maintained by the SSI for the duration of the connection. The SSI includes the capabilities as part of the HTTP request that it forwards to the target server (in the wire line network).