provides system specific information such as the access parameters for the selected cell and the information related to the surrounding cells to support cell selection and location registration procedures in the mobiles.

7.4.2 GPRS Radio Interface

To support high speed packet data access, the TDMA structure in GSM is extended and new logical channels are introduced to support control signaling as well as packet traffic flow over the GPRS radio interface. In GPRS, mobile station can receive multiple downlink time slots as well as transmit on multiple uplink time slots. The burst structure and frame structure remain to be the same as GSM except that the multiframe has been extended to 52 frames. This is elaborated in Figure 7.10 The 52-multiframe consists of 52 consecutive assigned time slots with the same number (e.g., TS 2 as in Figure 7.10 Of the 52 time slots, 48 are used for sending the actual GPRS data (radio blocks). Of the remaining four, the two time slots marked 'X' (numbers 25 and 51) are used for neighbor cell identification, similar to the IDLE frame in the 26 multiframe in the GSM speech case. The mobile searches the SCH burst that holds the BSIC and timing information for the neighbor. During the 'X' time slot, the serving cell does not receive or transmit anything. Finally, the two time slots marked 'P' (numbers 12 and 38) form the *packet timing advance control channel* (PTCCH) channel used for timing advance regulation. We shall elaborate the new logical channels for the GPRS service in this section.

The physical channel dedicated to packet data traffic is called a *packet data channel* (PDCH). Different packet data logical channels can occur on the same PDCH. Similarly, there are two types of logical channels for packet data, namely the *packet data traffic channels* (PDTCH) and the *packet control channels* (PCCHs). The PDTCH is the logical channel setup to carry user dedicated information for packet switched sessions. The mapping of physical channel (PDCH) to PDTCH is illustrated in Figure 7.10 A user may simultaneously occupy multiple PDTCHs (multiple time slots) for high speed access. On the other hand, one PDTCH may be shared by multiple users for low data rate packet access.

Several new packet control channels are introduced to support GPRS packet data access. The packet control channels can be divided into the *packet common control channels* (PCCCHs), which are the logical control channels shared by all mobiles in the cell, and the *packet dedicated control channels* (PDCCHs), which are the logical control channels dedicated to mobile users. For PCCCHs, we have the *packet random access channel* (PRACH), the *packet paging channel*(PPCH), the *packet access grant channel* (PAGCH), the *packet notification channel* (PNCH) as well as the *packet broadcast control channel* (PBCCH). They are elaborated below.

Packet Random Access Channel (PRACH): The PRACH is the only uplink PCCCH shared by all mobiles for initial access or transfer of bursty packet data. For example, Figure 7.11 illustrates the handshaking between the GPRS mo-

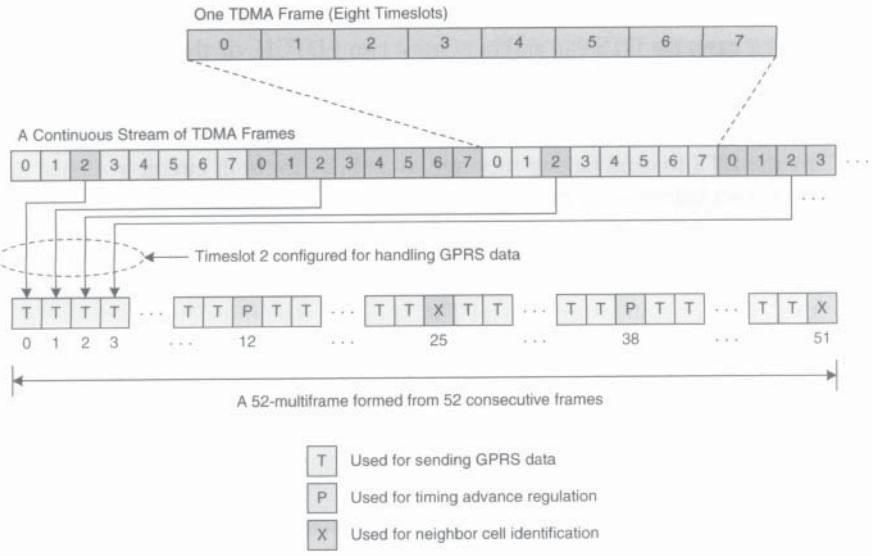


Figure 7.10 Illustration of GPRS frame structure and the mapping of physical channels to packet data traffic channels (PDTCH).

bile and the BTS for uplink packet data access. The GPRS transmits a packet channel request to the BTS via the PRACH. Alternatively, the GPRS mobile may use the RACH to send the access request.

Packet Access Grant Channel (PAGCH): This is the common logical control channel used by the BTS in the packet data transfer establishment phase to assign uplink packet channel (either PACCH [to be introduced later] or PDTCH [for delivering a small number of uplink packets]) in response of the PRACH access message sent by the mobile. For example, in Figure 7.11, the BTS may respond to a mobile packet access request using a *one-phase procedure* or a *two-phase procedure*. In the one-phase procedure, the BTS assigns an uplink packet data channel for a number of radio blocks to be transferred. One of more PDTCH may be assigned depending on the requested resource indicated in the access message in PRACH. The network informs the mobile of this assignment through the PAGCH. In the two-phase procedure, the network needs more information for resource reservation and it indicates the need for two-phase procedure in the PAGCH. Specifically, the BTS setup the uplink resource of PACCH (to be introduced later) and informs the mobile via the PAGCH. The mobile and the network exchange packet resource request and packet resource assignment signaling messages on the PACCH to complete the resource negotiation.

On the other hand, the mobile may use RACH to send the access request and in that case, the BTS can assign at most two PDTCH via the downlink AGCH for uplink packet data transfer.

Packet Paging Channel (PPCH): This is a common logical control channel used by the BTS to send paging message to mobiles for both circuit-switched and packet-switched services.

Packet Notification Channel (PNCH): This is a common logical control channel used by the BTS to send a *point-to-multipoint multicast* (PTM-M) notification to a group of mobiles prior to a PTM-M packet data transfer.

Packet Broadcast Control Channel (PBCCH): This common logical control channel is used by the BTS to broadcast system information specific for packet data. If PBCCH is not allocated in the physical channel, the packet data specific system information is broadcast on the existing GSM BCCH.

The packet dedicated control logical channels include *packet associated control channel* (PACCH) and the *packet timing advance control channel* (PTCCH). They are elaborated below.

Packet Associated Control Channel (PACCH): This is a dedicated control logical channel (dedicated to mobile) to convey mobile specific signaling messages such as power control, resource assignment and reassignment. The PACCH shares resources with PDTCHs. For example, in Figure 7.11, the packet resource request message (mobile to BTS) and the packet resource assignment message (BTS to mobile) are delivered in the PACCH. An mobile currently engaged in packet data transfer may also be paged for circuit switched services on PACCH.

Packet Timing Advance Control Channel (PTCCH): This is a full duplex dedicated control logical channel. The mobile uses the uplink PTCCH to transmit a random access burst. With this information, the BSS estimates the timing advance requirement. The timing advance information is delivered to the mobile using the downlink PTCCH.

7.5 CORE NETWORK INTERFACE AND SERVICES

The core network of the GSM/GPRS infrastructure supports the basic call switching functions as well as mobility management of mobiles. As illustrated in Figure 7.2, the GPRS/GSM core network adopts an overlay approach. The circuit-switched domain (MSC/IWF) interfaces with the PSTN via PCM-64 in the payload path and SS7 in the signaling path. The packet-switched domain (SGSN/GGSN) interfaces with the public internet via standard IP protocol [99]. The VLR/HLR distributed database

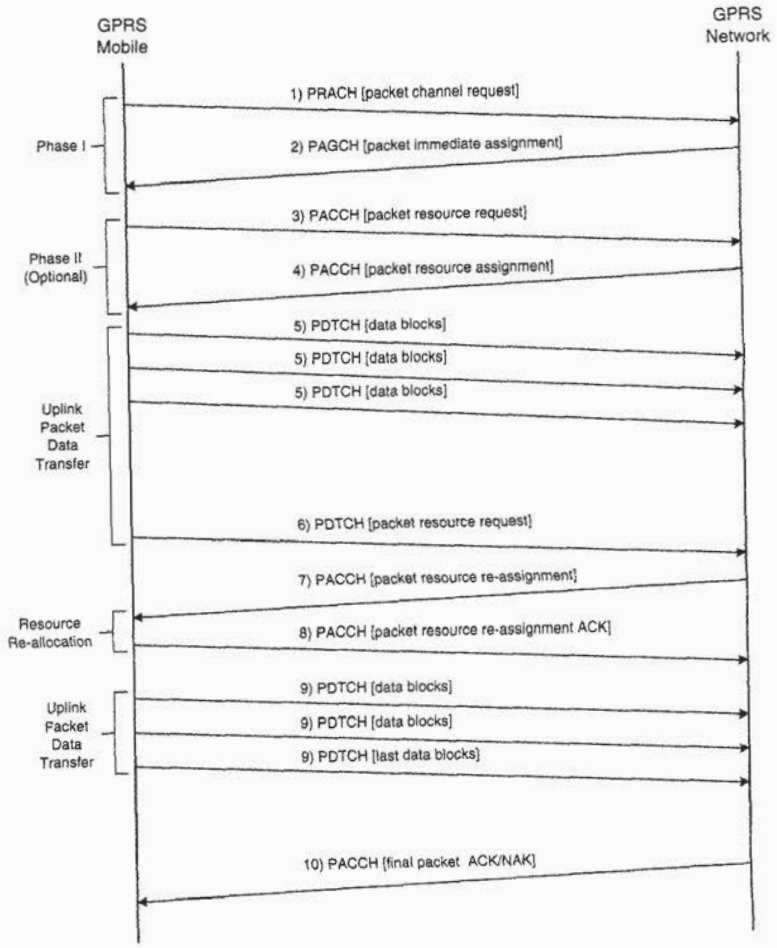


Figure 7.11 Illustration of GPRS uplink packet data access.

supports both the circuit-switched and the packet-switched core networks. We shall elaborate the circuit-switched and the packet-switched domains of the core network in the following subsections.

7.5.1 Circuit-Switched Domain—GSM Core Network

The critical components in the GSM core network includes the MSC, GMSC, IWF, VLR and HLR. The MSC and GMSC are directly involved in the call switching for circuit-switched voice connection. The IWF is involved for circuit-switched data connection. In both cases, the MSC/IWF, VLR and HLR are involved in the control signaling during call setup, call termination as well as mobility management. We shall elaborate the signaling procedures of these services with examples below.

Figure 7.12 illustrates the protocol architecture of the *control plane* of the circuit-switched domain. The BSC connects to the MSC through the A-interface. The MSC connects to the other MSC, GMSC, VLR and HLR through the GSM-MAP protocol, which is derived from the *Transaction Capabilities Application Part* (TCAP) of the SS7 framework. The GSM networks track the location of mobiles so that incoming calls can be routed to the target subscriber. To exercise the location tracking, a cluster of base stations is grouped to form a *location area* (LA). Mobile has to perform a *location update procedure* whenever it crosses a location area. The location update procedure is defined in the mobility management (MM) layer at the mobile. The corresponding peer of the MM layer is at the MSC. Above the MM layer is the *call control layer* (CC) which is responsible for the ISDN-like call setup and call termination procedures. The corresponding peer of the CC at the core network is the MSC.

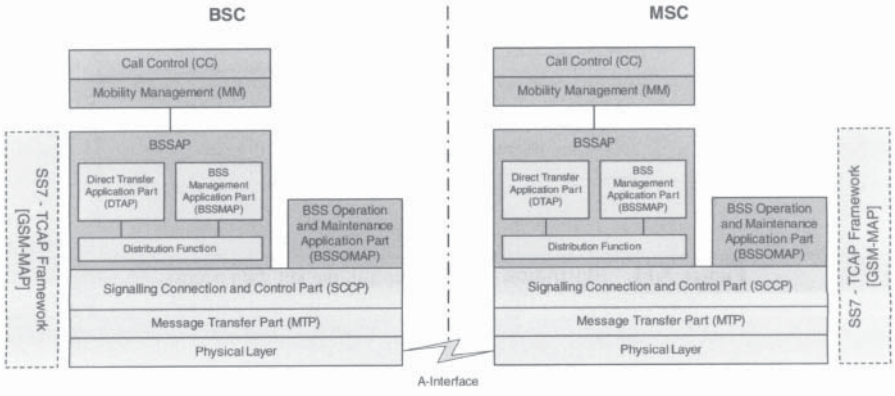


Figure 7.12 Protocol architecture of the control plane of the circuit-switched domain in GSM system.

The mobility management, call setup and call termination, as well as international roaming are all supported by the distributed database in the *visiting location register* (VLR) and *home location register* (HLR) [217]. One MSC is associated with one VLR in the core network and the entire core network is associated with one HLR. The structures of the VLR and HLR are elaborated below.

Home Location Register (HLR): The HLR maintains a permanent database record for every mobile in the GSM system. The search key used in HLR to identify a mobile record is the *IMSI*. A mobile record in the HLR includes the MSC number, the subscription profile as well as the address of the last VLR visited.

Visiting Location Register (VLR): The VLR maintains a dynamic database record for the mobiles under its location area. The search key used in VLR to identify a mobile record is the *TMSI*, which is a temporarily address generated by the VLR. The dynamic record in the VLR includes the IMSI, the address of the HLR, the address of the MSC currently serving the mobile (MSRN) and the current location area identity (LAI).

We shall elaborate the mobility management, call setup, call termination and international roaming procedures in the following subsections.

7.5.1.1 Mobility Management: Location Update Procedure When a GSM network receives an incoming call, the GMSC has to decide where to route the call to and page the target mobile for the incoming call. When the network has no idea of the current location of the target mobile, all the BTS have to send paging message to look for the target mobile and hence, the paging overhead is enormous for each incoming call. This is obviously impractical. To reduce the paging overhead, the GSM core network is designed to keep track of the location of the mobile down to a *location area* (LA) resolution. A LA, identified by LAI, consists of a cluster of BTSs. Whenever the mobile (in idle state) crosses the LA boundary, the mobile has to initiate location update procedure. The size of the LA is a tradeoff of the uplink signaling overhead (location update overhead) and downlink paging overhead during idle state.

There are three cases for location update, namely the *inter-LAI movement within the same MSC*, the *inter-MSC movement* as well as the *inter-VLR movement*. Figure 7.13 illustrates the scenario of the first case. Say the mobile moves from LAI-1 to LAI-2 and both LAIs belong to the same MSC. The location update procedure is summarized into three steps below.

Step 1 (MS to MSC): The mobile station sends a *Location Update Request Message*, which is a signaling message generated at the MM layer of the mobile. The message includes the previous LAI, current LAI, previous MSC, previous VLR address and TMSI. The mobile (in steady state) identifies itself by the TMSI. However, if the VLR does not have record for the mobile (e.g., first time

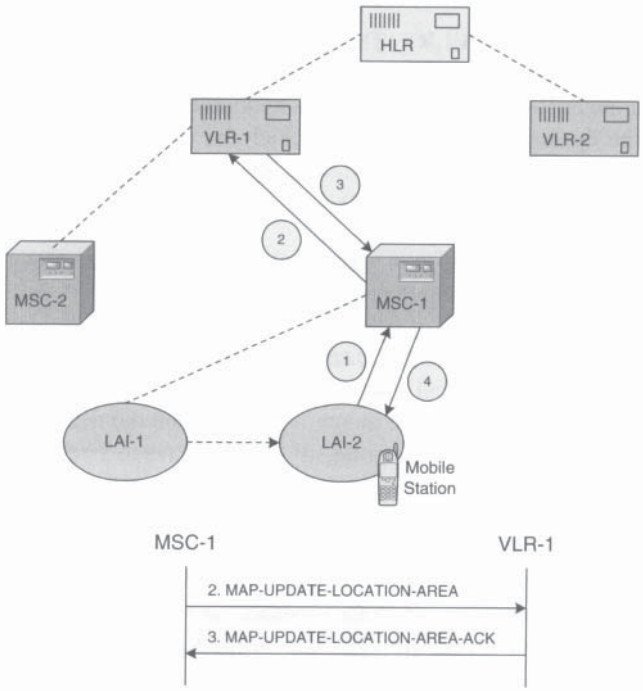


Figure 7.13 Case 1: Inter-LAI movement within the same MSC.

registration), the mobile will be asked to send the IMSI. The location update request message terminates at the MSC.

Step 2 (MSC to VLR): The MSC, upon receiving the location update message, forward the location update request to the VLR based on GSM-MAP dialogue. The location update transaction into the VLR includes the MSC address, TMSI, previous LAI and the current LAI. The VLR locates the mobile record based on the TMSI and update the LAI field.

Step 3 (Completion): Since the location update does not involve new MSC, the VLR sends an ACK to the MSC, completing the location update dialogue.

Figure 7.14 illustrates the scenario of the second case. Say the mobile moves from LAI-1 to LAI-2 which involves different MSCs. The location update procedure is summarized into four steps below.

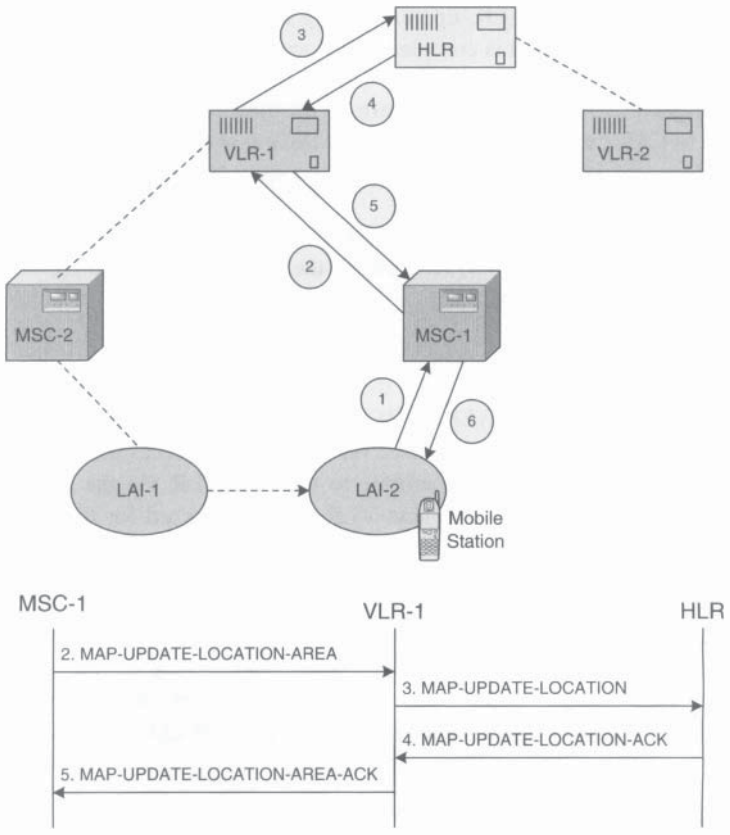


Figure 7.14 Case 2: Inter-LAI movement involving a change of MSC.

Step 1 (MS to MSC to VLR): The mobile station sends a location update message to the MSC with TMSI, previous LAI, current LAI and previous MSC address. The MSC forward the location update request to the VLR based on GSM-MAP dialogue.

Step 2 (VLR Processing): The VLR locates the mobile record based on TMSI as the key. VLR notices that LAI-1 and LAI-2 belongs to two different MSCs. Hence, the VLR updates the LAI and MSC address fields in the mobile record. In addition, it derives the HLR address from the IMSI (in VLR record) and sends a *MAP-UPDATE-LOCATION* to the HLR (containing IMSI, address of the new MSC as well as the address of the VLR).

Step 3 (HLR Processing): The HLR identifies the mobile record based on the IMSI and updates the MSC address in the record. HLR replies with an ACK to the VLR to complete the GSM-MAP dialogue.

Step 4 (Completion): The VLR, upon receiving the ACK from HLR, replies with an ACK to the MSC. This completes the location update procedure.

Finally, Figure 7.15 illustrates the scenario of the third case. Say the mobile moves from LAI-1 to LAI-2 which involves different MSCs as well as different VLRs. The location update procedure is summarized into four steps below.

Step 1 (MS to MSC to VLR): Similarly, the mobile sends a location update message to the MSC and the MSC forward the location update to the new VLR based on GSM-MAP dialogue.

Step 2 (VLR Processing): The new VLR does not have the record of the mobile based on the TMSI supplied from the location update request. Therefore, the new VLR sends a *MAP-SEND-IDENTIFICATION* message to the old VLR with the TMSI. The old VLR, based on the TMSI, retrieves the mobile record and delivers the IMSI of the mobile to the new VLR via the *MAP-SEND-IDENTIFICATION-ACK*. The new VLR creates a record for the mobile and generates a new TMSI for the record.

Step 3 (HLR Processing): The new VLR sends a registration message to the HLR (*MAP-LOCATION-UPDATE*) with IMSI, address of the new MSC as well as the address of the new VLR. The HLR locates the mobile record based on the IMSI and update the VLR address as well as the MSC address fields in the record. It completes the MAP dialogue with the MAP-LOCATION-UPDATE-ACK to the new VLR.

Step 4 (Completion): The new VLR sends the generated TMSI to the mobile via the new MSC. At the same time, the new VLR deletes the mobile record in the old VLR.

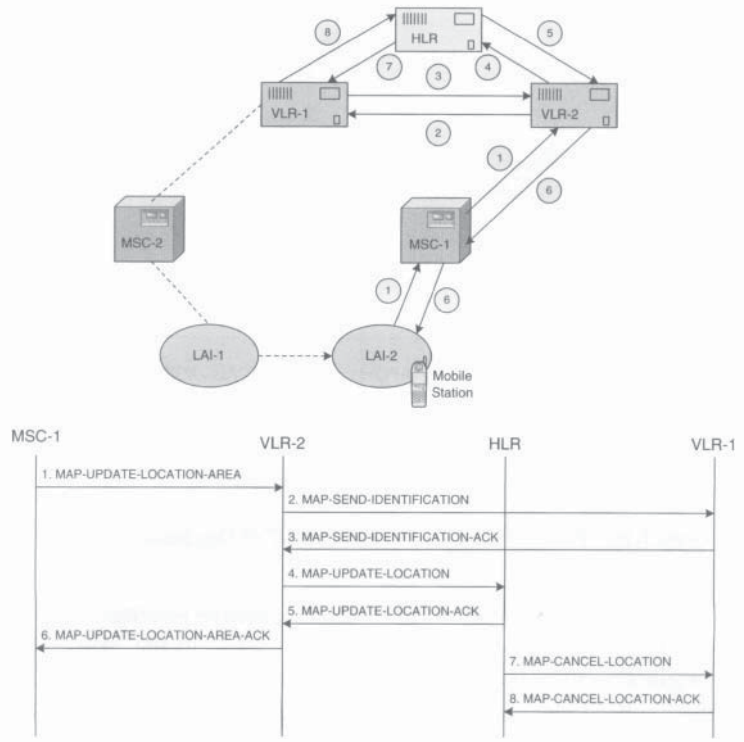


Figure 7.15 Case 3: Inter-LAI movement involving a change of MSC and VLR.

7.5.1.2 Call Control: Call Origination and Termination Based on the location update procedures, the call setup procedures are described below. Figure 7.16 illustrates the call origination procedure at the core network level. In contrast to the call setup procedure described in earlier section, we shall focus on the end-to-end procedure involving the core network and the mobile.

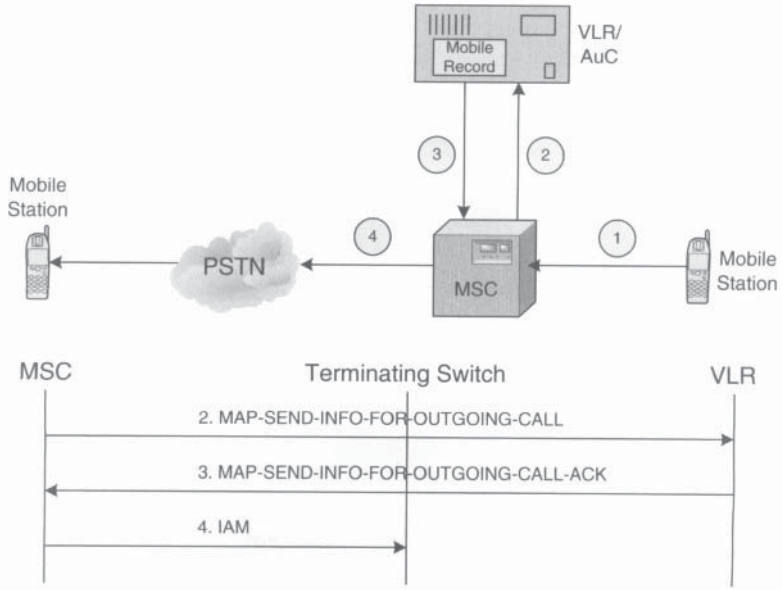


Figure 7.16 Core network procedure for the Call Origination in GSM.

Step 1 (MS to MSC): The mobile sends a *call origination message*, which is a signaling message at the *call control protocol layer*, to the MSC. Hence, the call origination message is transparent to the BTS and the BSC. In the call origination message, the mobile identifies itself using TMSI obtained during registration procedure.

Step 2 (MSC to VLR): The MSC forwards the call origination request to the VLR to locate the mobile record using *MAP-SEND-INFO-FOR-OUTGOING-CALL* GSM-MAP dialogue.

Step 3 (VLR Processing): The VLR retrieves the mobile record based on the TMSI. The VLR checks if the mobile is allowed to make outgoing calls and replies the MSC with *MAP-SEND-INFO-FOR-OUTGOING-CALL-ACK*, completing the GSM-MAP dialogue.

Step 4 (Call Setup): If the result is positive, the MSC shall setup voice trunks according to the standard PSTN procedures as well as the radio resource of the

BTS to carry the voice call. This triggers a number of radio interface handshaking to setup the voice channels as illustrated in Figure 7.9.

Figure 7.17 illustrates the end-to-end call termination procedure between the mobile and the core network. The detail procedure is outlined in the steps below.

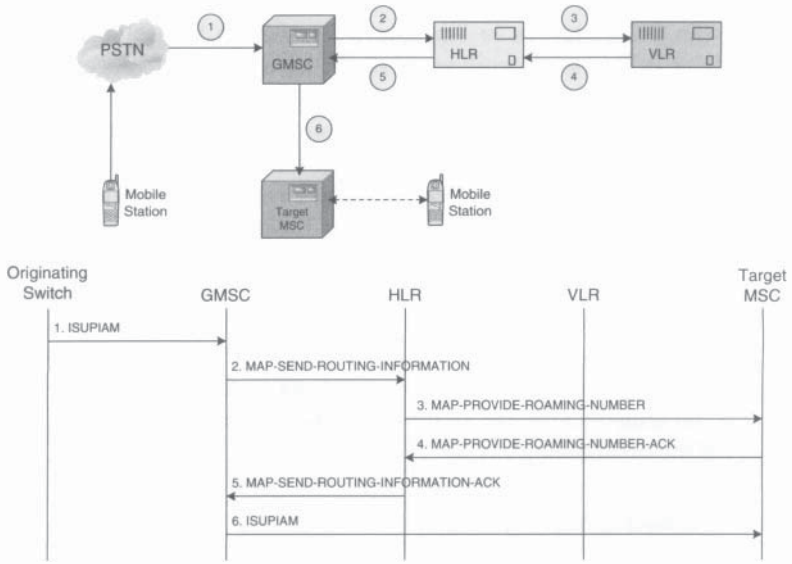


Figure 7.17 Core network procedure for the Call Termination in GSM.

Step 1: The PSTN phone initiates a fixed-line to mobile call and the call is routed to the GMSC from the PSTN via SS7 - IAM.

Step 2: The GMSC queries the HLR via *MAP-SEND-ROUTING-INFORMATION*, which contains the MSISDN (called number).

Step 3: The HLR identifies the mobile record based on the MSISDN and sends the *MAP-PROVIDE-ROAMING-NUMBER* to the target VLR to obtain the MSRN.

Step 4: The VLR creates the MSRN using the MSC address stored in the mobile record. The MSRN is sent back to the GMSC via the HLR, completing the GSM-MAP dialogue.

Step 5: The target MSC address can be derived from the MSRN at the GMSC. Hence, the voice trunk is setup from the GMSC to the target MSC using SS7 IAM. The mobile is paged from all the BTSs in the current LAI. Upon the mobile receives the page message and responds with page response message, the handshaking procedure over the radio interface defined in Figure 7.9 continues.

7.5.1.3 International Roaming One of the important features supported in GSM network is the *international roaming*. In fact, international roaming can be regarded as a special case of call termination described in the above subsection, with an exception that the called mobile may be a visiting mobile. The international roaming is supported by the VLR/HLR distributed database architecture. Suppose a GSM mobile user (John) from Taiwan roams to Singapore. When the GSM mobile registers in the GSM network in Singapore, the VLR in Singapore network will update the HLR in Taiwan network based on the location update procedure in Case 3. Suppose a person in Taiwan calls John. The call will be processed according to the call termination procedure described in Figure 7.17 and hence, two local voice trucks (within Taiwan and within Singapore) and an international voice trunk (between Taiwan and Singapore) will be resulted. The caller is charged for a local GSM call but John is charged for an international call from Taiwan to Singapore. On the other hand, suppose a person from Hong Kong calls John. Following the call termination procedure in Figure 7.17, this will result in two international segments, from HK to Taiwan and from Taiwan to Singapore. Suppose now, a caller in Singapore calls John (who is roaming in Singapore), two international segments (Singapore to Taiwan and Taiwan to Singapore) will be resulted although both parties are in Singapore. This undesirable call setup is called the *Tromboming effect*. Figure 7.18 illustrates the call processing of the tromboming effect.

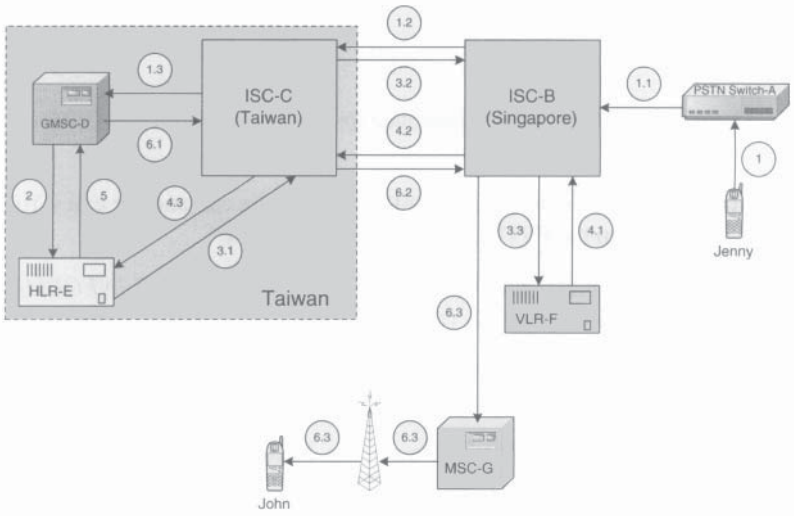


Figure 7.18 Illustration of Tromboming Effect in GSM International Roaming. A user in Singapore (Jenny) calls the roamer (John).

Step 1: Say Jenny calls John's number (Taiwan's GSM network). This is interpreted as an international call and hence, the SS7-IAM will be routed to an international

switching center (ISC-B). The call will be routed to ISC-C in Taiwan, which will be routed to the GMSC-D in Taiwan's GSM network (John's home network).

Step 2: The GMSC-D in Taiwan queries the HLR-E in Taiwan for the MSRN. The HLR-E realizes that John is roaming in Singapore and queries the VLR-F in Singapore based on GSM-MAP dialogue. The MSRN is returned to GMSC-D.

Step 3: Based on the MSRN, the GMSC-D sets up the voice trunk to the MSC-G in the Singapore network, connecting to John. Hence, the voice path involves two international trunks, namely from Singapore to Taiwan (ISC-B to ISC-C) and from Taiwan to Singapore (GMSC-D to MSC-G), resulting in tromboning.

Obviously, the tromboning effect is highly undesirable. The key step is to stop the call being routed to the ISC-C as otherwise, the international segments cannot be eliminated in the voice path. We shall discuss several possible solutions below.

Roamer Location Cache (RLC): Figure 7.19 illustrates the architecture of the RLC-based solution. For instance, a RLC is colocated at the ISC-B in the visited system. The RLC acts as a HLR proxy for the visited system by intercepting the GSM-MAP message during the location update procedure. Specifically, during the location update procedure, the mobile (John) registers to the VLR in Singapore network. The VLR sends the *MAP-UPDATE-LOCATION* to the roamer's HLR (Taiwan). Since the HLR is in Taiwan, the message will be routed to ISC-B. The RLC in ISC-B duplicates the MAP message and creates a record with IMSI and VLR/MSC address for the mobile. RLC still does not have the MSISDN for the mobile. The RLC requests this information from the HLR (Taiwan) using "MAP-RESTORE-DATA". The MSISDN is returned from the HLR (Taiwan) to the RLC through the "MAP-INSERT-SUBSCRIBER-DATA".

During call processing, Jenny dials John's mobile number as before. Similarly, the call is routed to ISC-B. The ISC-B looks at the prefix of the called number and recognizes it as a potential roamer. Before routing the call to the ISC-C in Taiwan, the ISC-B searches the local RLC using the called MSISDN. If no entry is found, the call is routed to ISC-C as usual. If an entry is found for John's number, the call is routed to the MSC (Singapore). In this case, the RLC is acting as the HLR proxy. Hence, the international voice trunks are avoided in the speech path of the call. The advantage of the RLC solution is user transparency as the caller (Jenny) does not need to be aware of the special arrangement. On the other hand, the disadvantage of the RLC solution is the need to modify the ISC-B to understand GSM-MAP. Hence, agreements between the GSM operators and the international carriers are needed.

Roamer Location Cache within GSM (RLC-GSM): Figure 7.20 illustrates the architecture of the RLC solution within the GSM network. Unlike the previous solution, the RLC is now located within the GSM network so that no modification of the ISC-B is needed. Instead, the RLC is colocated at the operator-owned

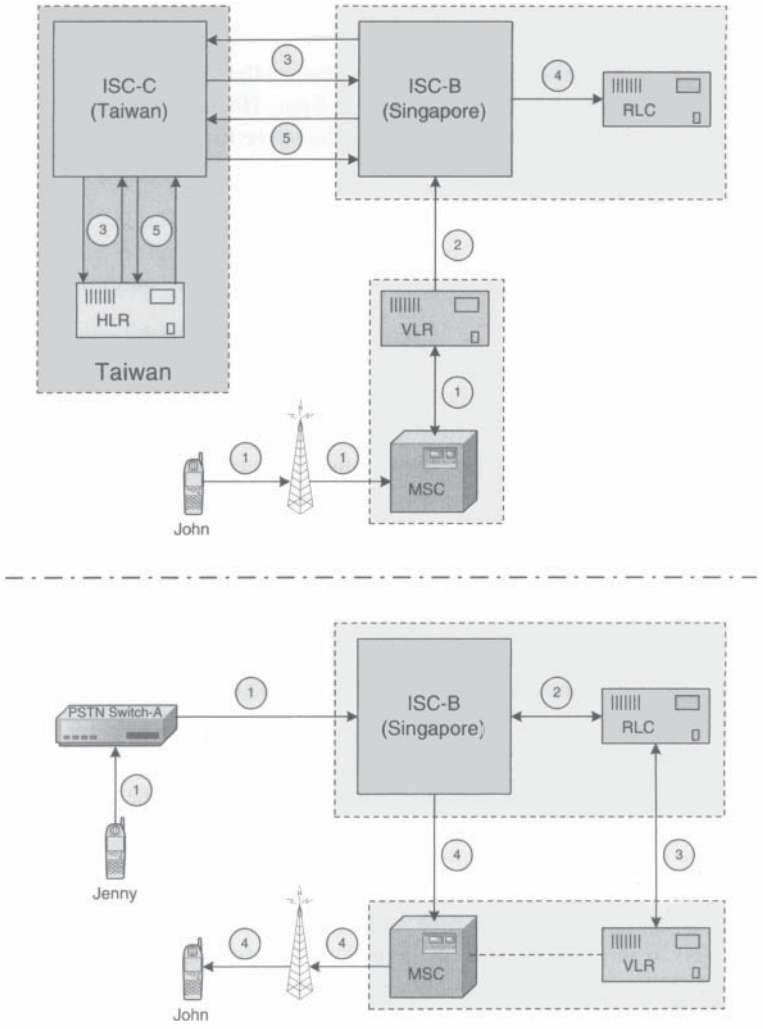


Figure 7.19 Illustration of the Roamer Location Cache (RLC) in GSM system.

switch (switch D). During the location update procedure, the VLR recognizes the location update is from a roaming user. Hence, the VLR sends the "MAP-UPDATE-LOCATION" to the RLC. The RLC creates a record for the mobile with IMSI, VLR and MSC address. The RLC relays the "MAP-UPDATE-LOCATION" to the HLR in Taiwan. After the location update procedure is completed, the RLC obtains the MSISDN similar to the RLC-based solution.

During call processing, Jenny calls a special access number (which is routed to switch D). After connecting to switch D, Jenny is asked for John's number. Switch D will search the RLC for John's record using the MSISDN. If there is no such entry, the call will be routed to ISC-B as usual. However, if an entry is found in RLC, the call is routed to the GMSC locally in Singapore network. Similar to the above solution, the RLC is acting as the HLR proxy. The advantage of this solution is that only modification within the GSM network is needed. However, extra modification at the VLR is needed and the dialing procedure for user (Jenny) is different from the regular procedure. Hence, the solution is not transparent to the general users.

Extractor: Figure 7.21 illustrates the roaming solution using *extractor*. The extractor-based solution is similar to the RLC-based solution except that no modification is needed in the GSM network. Instead, a new element called *extractor* is introduced into the GSM network. The extractor monitors (but does not modify) the GSM-MAP message between the VLR and the foreign HLR. It will capture the GSM-MAP message when there is location update dialogue between the VLR and the foreign HLR. RLC obtains the MSISDN similar to the previous solutions and acts as the HLR proxy during call processing. The call processing operation is exactly the same as the RLC within GSM solution. The advantage of the extractor-based solution is that it is transparent to the VLR and no modification of VLR is needed. On the other hand, the solution is not transparent to the end users.

International Roamer Access Code (IRAC): In fact, the key to avoid the tromboning effect is to route the call into the visited GSM network before it reaches the ISC. The GSM operators of the visited network reserves an IRAC. In order to make a call to John, Jenny dials the NDC1 + IRAC + John's number (MSISDN). Based on NDC1, the PSTN routes the call to the GMSC of the visited network. From IRAC, the GSM network recognizes that it is an international roaming call. From the MSISDN, the GMSC search for John's record in the HLR (Taiwan) to obtain the MSRN for normal call processing. The GMSC of the visited network will route the call to the local MSC and this results in local trunks to carry the voice call. The advantage of this approach is that no modification of the GSM network or ISC is needed. However, the solution is not transparent to the end users as special IRAC dialing rule is needed.

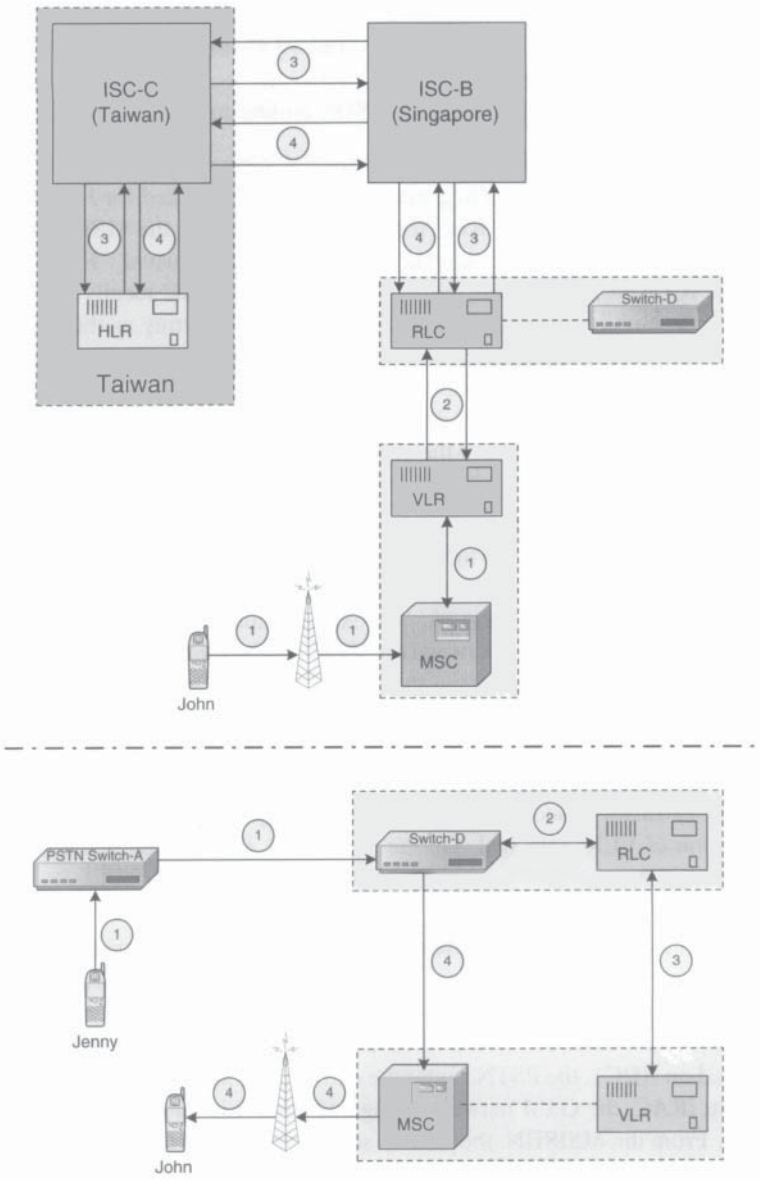


Figure 7.20 Illustration of the Roamer Location Cache (RLC) within GSM system.

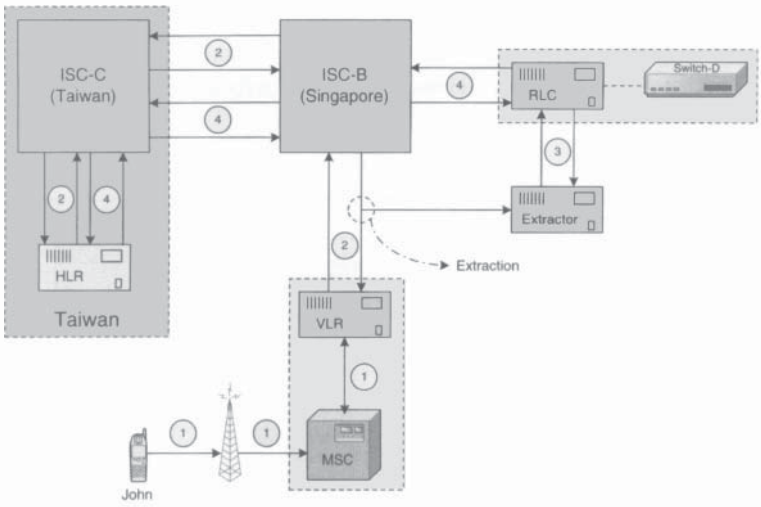


Figure 7.21 Illustration of the Extractor-based solution for International Roaming in GSM system.

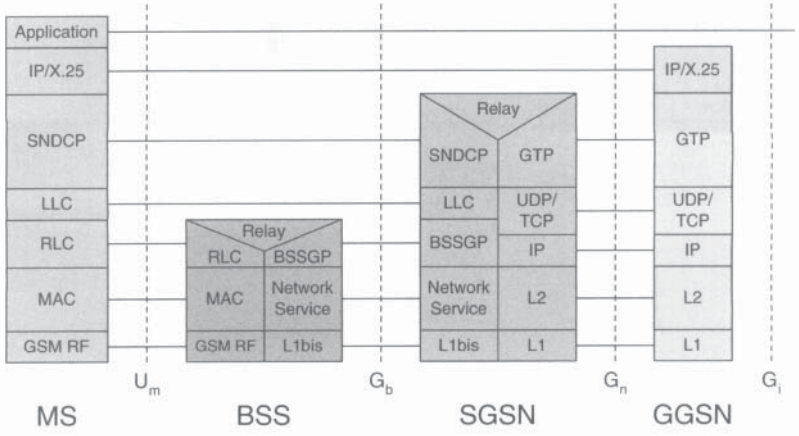
7.5.2 Packet-Switched Domain—GPRS Core Network

Unlike the GSM core network, the GPRS core network (SGSN and GGSN) is packet-switched (IP) based. Figure 7.22(a) and 7.22(b) illustrate the protocol architecture of the user plane and control plane for GPRS system. The user plane illustrates the protocol stacks involved in delivering the user application across the GPRS network. The control plane illustrates the protocol stacks involved in signaling across the GPRS networks to establish packet-switched connections. Both the user plane and control plane are supported by the LLC/RLC layers. The RLC layer offers a reliable logical channel between the mobile and the BSC based on ARQ and retransmission. and BSC. The LLC layer defines a logical link control layer protocol between the mobile and the SGSN. The LLC spans from the mobile to the SGSN and is intended for use with both acknowledged and unacknowledged data transfer. Encryption is also provided in the LLC layer for secured data transmission.

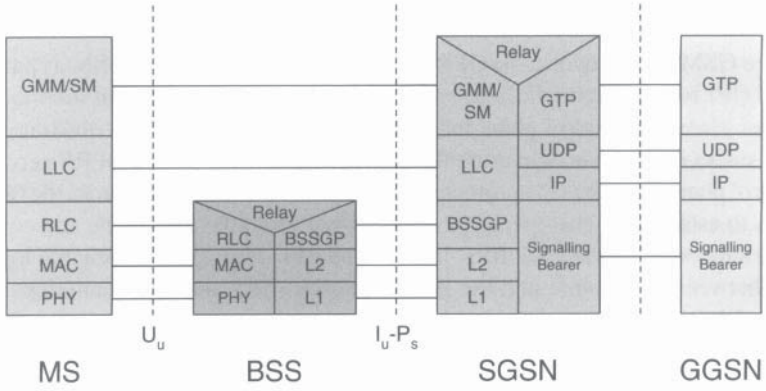
On the user plane, the application data (such as TCP/IP) is delivered to the SGSN over the LLC logical link. The relay function in SGSN delivers the application data to the GGSN via the GPRS Tunneling Protocol (GTP).

On top of the LLC layer, there are two important protocol layers in the control plane. The main function of the GPRS Mobility Management (GMM) is to support mobility management functionality of a mobile such as GPRS attach, GPRS detach, security, routing area update, location update. The main function of the Session Management (GMM/SM) protocol supports mobility management functionality of a mobile such as GPRS attach, GPRS detach, security, routing area update, location

update. The main function of the SM layer is to support PDP context handling of the user terminal. SM comprises of procedures for the PDP context activation, deactivation, and modification. These procedures will be elaborated later in this section.



(a) user plane



(b) control plane

Figure 7.22 Protocol architecture of the user plane and control plane of the packet-switched domain in GPRS system.

Figure 7.23 illustrates the typical flow of events for a GPRS mobile. The GPRS mobile first physically attaches to the strongest cell and monitors the system broadcast messages and paging message. After that, the GPRS mobile registers to either the CS domain (IMSI attach) or PS domain (GPRS attach) to establish a mobility management

logical connection. Say the GPRS mobile attaches to the PS domain. On top of the GMM connection, the GPRS mobile establishes a SM connection with the GGSN. Application data packets will be delivered via the logical channel established. During this state, the GPRS mobile may have to location area / routing area update as well as SGSN relocation. Finally, the GPRS mobile release the logical connection in the mobility management (GPRS detach or IMSI detach).

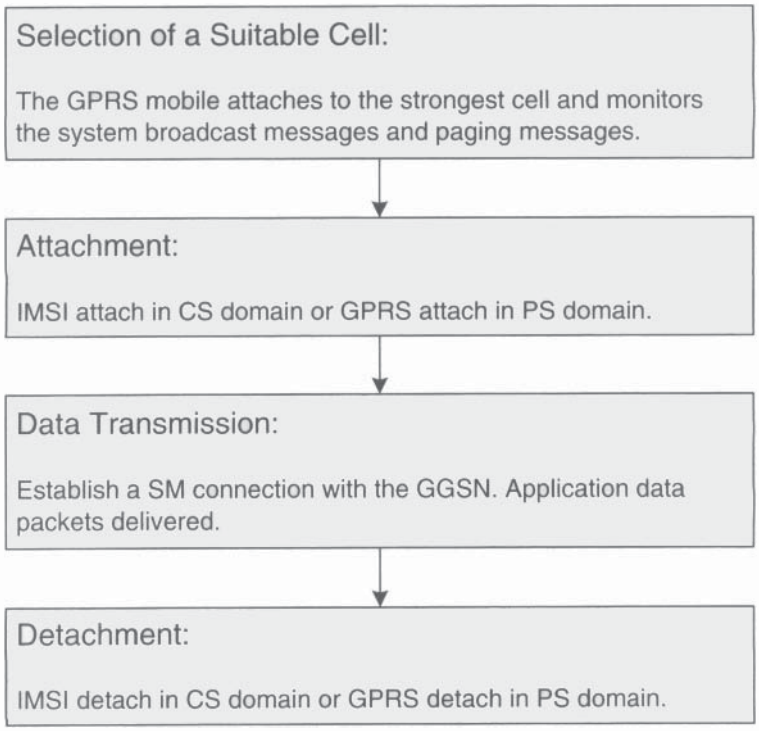


Figure 7.23 Illustration of the events involved during a packet data session in GPRS system.

7.5.2.1 GPRS Mobility Management In this section, we shall elaborate on the signaling procedures related to the GPRS mobility management functions. Specifically, we shall discuss the GPRS attach and detach procedures. Before a logical channel for the packet data flow can be established, the GPRS mobile has to establish a GMM connection with the SGSN. The signaling procedure for the GMM connection establishment refers to *GPRS attach procedure*. Similarly, the signaling procedure to release the GMM connection refers to *GPRS detach procedure*. Figure 7.24 illustrates the scope of the GMM connection in a GPRS network. In other words, the GPRS attach/detach refers to the registration/cancellation of the mobile to the SGSN in the packet-switched (PS) domain. On the other hand, the IMSI attach/detach refers to

the registration/cancellation of the mobile to the MSC/VLR in the circuit-switched (CS) domain.

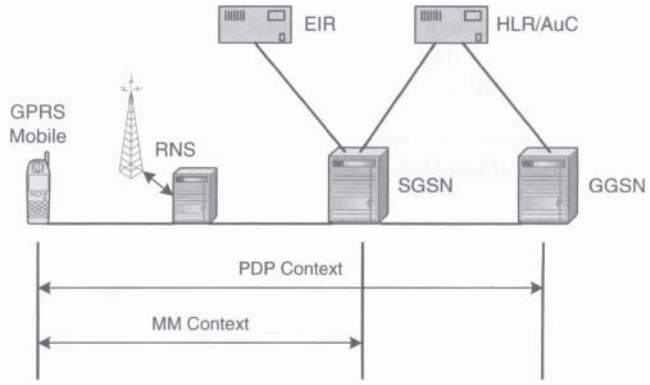


Figure 7.24 Illustration of the MM context and PDP context in GPRS systems.

When a GMM connection is established (GPRS attach), a mobility management context (MM Context) will be created for the mobile. The MM context at the SGSN includes the IMSI, TMSI, MSISDN, MM state, Routing Area Identity (RAI) and cell identity, the address of the VLR currently serving the MS, IP address of the new SGSN where buffered packets should be forwarded,.etc. The MM context is maintained throughout the entire GMM connection.

Figure 7.25 illustrates the signaling flow of a typical GPRS attach procedure. The GPRS attach is initiated by the GPRS mobile and is divided into three phases. In the first phase, a *GPRS Attach Request* message (GMM layer message) will be delivered to the SGSN from the mobile via the logical channel provided by the LLC layer. In the second phase, the SGSN authenticates the identity of the GPRS mobile user and the equipment. The data exchanged will be ciphered. During the third phase, the SGSN address will be registered in the mobile record in HLR (similar to the location update procedure in GSM network described in the previous section). Finally, the SGSN/VLR will generate a temporarily ID (P-TMSI) and notify the mobile with the *GPRS Attach Accept* message via the LLC connection.

Figure 7.26 illustrates the signaling flow of GPRS detach. The GPRS detach can be initiated by the mobile or the SGSN. The PDP context between the mobile and the GGSN will be released and the buffered user data in the SGSN will be deleted. The LLC connection will be released and the corresponding mobile record in the HLR will be purged.

7.5.2.2 GPRS Session Management The GPRS session management (SM) establishes, modifies and releases packet data sessions with well-defined QoS between the mobile and the GGSN in GPRS system. Before the packet data session can start, the mobile and the GGSN has to establish a logical connection in the SM layer.

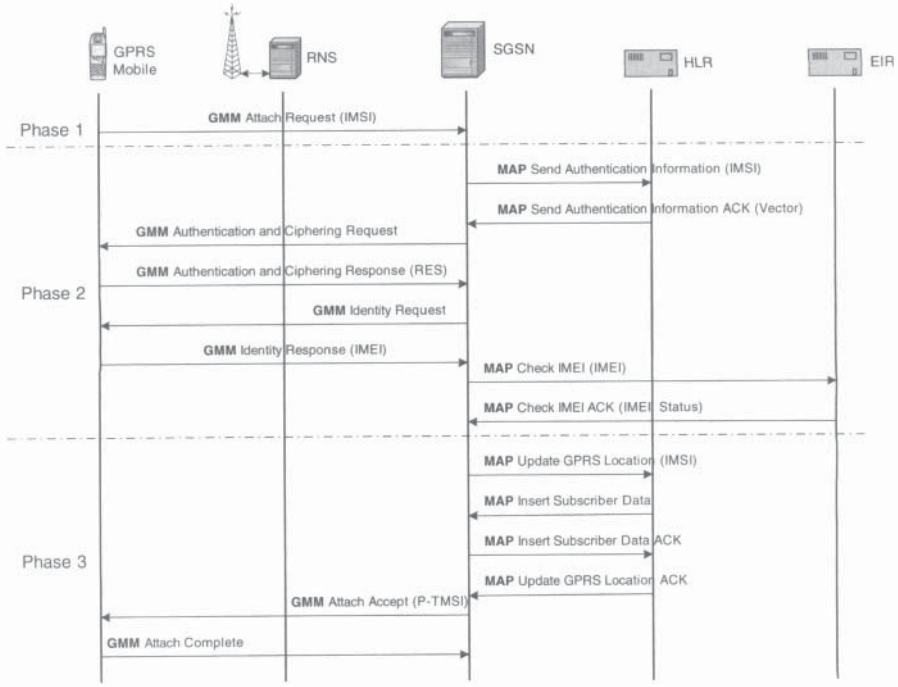


Figure 7.25 Illustration of the signaling flow in GPRS attach.

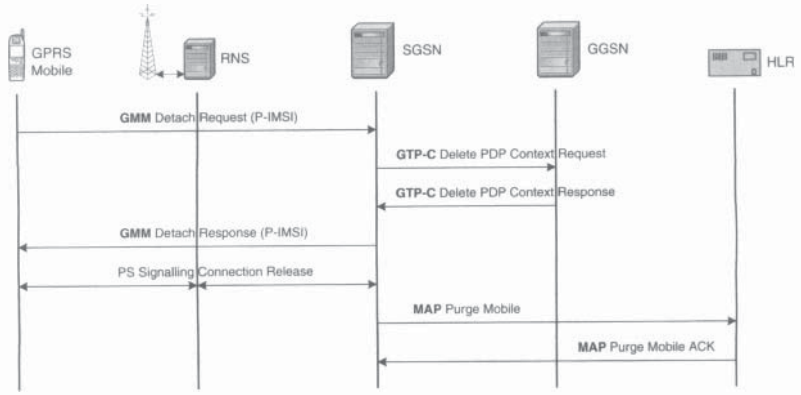


Figure 7.26 Illustration of the signaling flow in GPRS detach.

When the MM connection is established, a packet data protocol (PDP) context, which contains the important QoS attributes about the packet data session, will be created. PDP is a generic name for the user application protocol, which can be X.25 or IP. A mobile may have multiple PDP contexts simultaneously with each PDP context associated with a SM protocol entity. Figure 7.27 illustrates the information contained in the PDP context of the mobile, SGSN and GGSN.

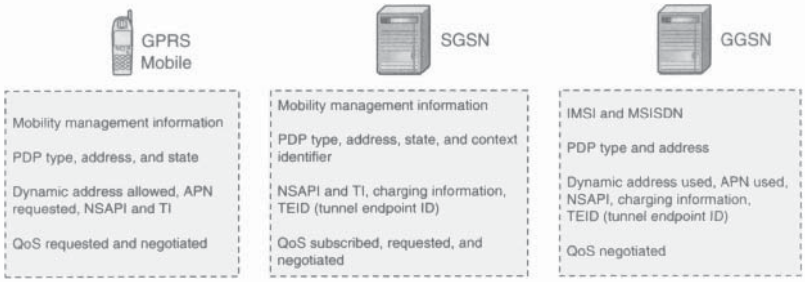


Figure 7.27 Illustration of the PDP context stored at the mobile, SGSN and GGSN.

There are three types of SM procedures defined for GPRS, namely the *PDP Context Activation*, *PDP Context Modification* and *PDP Context Deactivation*. All these procedures requires the GMM connection being established. Note that the PDP context can be activated by the mobile or the network. Figure 7.28 illustrates the signaling flow in a PDP context activation scenario (activated by the mobile). The mobile sends a *PDP Context Activation Request* to the SGSN over the GMM connection. The request contains the PDP type, PDP address (static/dynamic) as well as QoS requirement. The SGSN forwards the PDP activation request to the GGSN where a GTP tunnel is created between the PDP context in SGSN and the PDP context in GGSN. The GGSN obtains or activate the PDP address (e.g., IP address) from the external data network (e.g., ISP). If the GSGN reply (3.b) is positive, the SGSN activates the PDP context and is ready to forward / relay packets between the mobile and GGSN. In addition, the LLC connection (between SGSN and BSC) and the RLC connection (radio access bearer) between the BSC and the mobile will be established. Finally, the SGSN delivers the PDP Context Activation Accept message to the mobile, completing the PDP context activation procedure. From this point onward, the PDP address in the GPRS mobile is visible to the external network and PDP data can flow between the mobile and external network (e.g., internet). The GPRS mobile is said to be *online*. Note that due to the packet switched nature of the connection, the GPRS mobile can be online but at the same time, no dedicated radio resource is allocated at the BTS to support the logical connection.

Similarly, PDP context deactivation can be initiated by the mobile or the network. Figure 7.28 illustrates the signaling flow for mobile-initiated PDP context deactivation. The mobile sends the *Deactivate PDP context request* (with a transaction

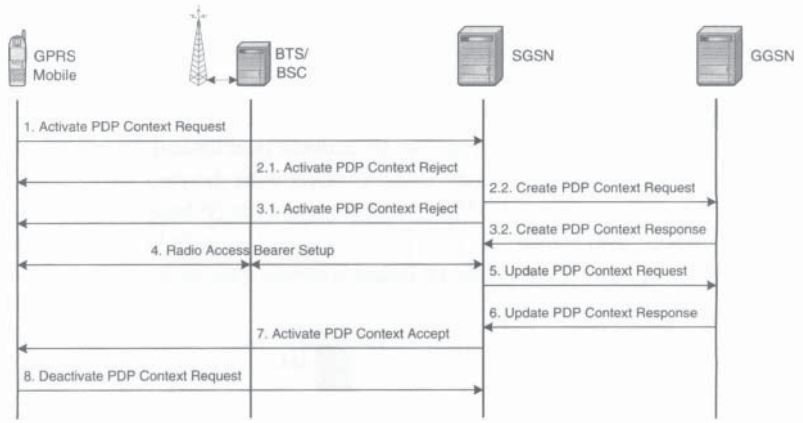


Figure 7.28 Illustration of the signaling flow in PDP context activation (activated by mobile) of GRPS network.

identifier TI to identify which PDP context to deactivate) to the SGSN via the LLC connection. The SGSN and GGSN exchange the signaling message to deactivate the PDP context. The GGSN deletes the PDP context and reclaims the dynamic PDP address into the system pool (if any). After all the concerned PDP contexts are released, the SGSN returns the accept message to the mobile. Note that if the mobile indicates "Tear Down Indicator" in the deactivation request message, all the PDP contexts of the current PDP address will be released.

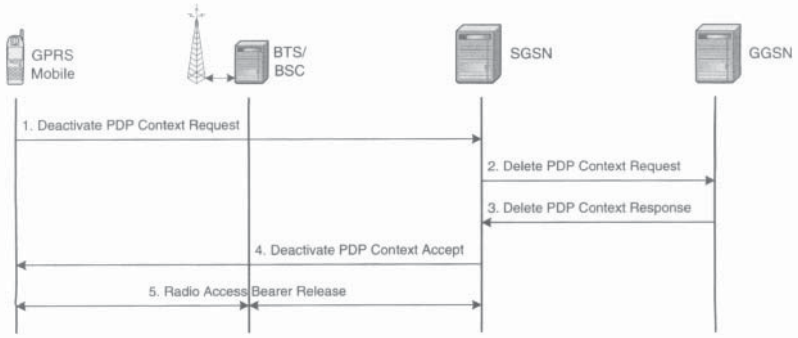


Figure 7.29 Illustration of the signaling flow in PDP context deactivation (initiated by mobile) of GRPS network.

Figure 7.30 illustrates the end-to-end wireless internet access based on GPRS network. The RLC layer between the GPRS/UMTS mobile and the BSC/RNC offers a logical channel (with ARQ protection). Above the RLC, there are user plane and

control plane. The user plane consists of regular TCP/IP stack which terminates into the SGSN/GGSN. The control plane consists of RRC (UMTS), GMM and SM (GPRS/UMTS). The SM is responsible for handshaking with GGSN on the end-to-end QoS negotiation as well as IP address activation (static/dynamic IP). After the SM layer established the PDP context, the IP address is activated and visible to the external internet. The TCP layer establishes a socket with the peer TCP layer in the web server. Hence, the GPRS/UMTS network offers an IP-based wireless bearer (between the mobile and the GGSN). Applications such as web browsing and VPN can be run transparently on top of the IP-based wireless data bearer.

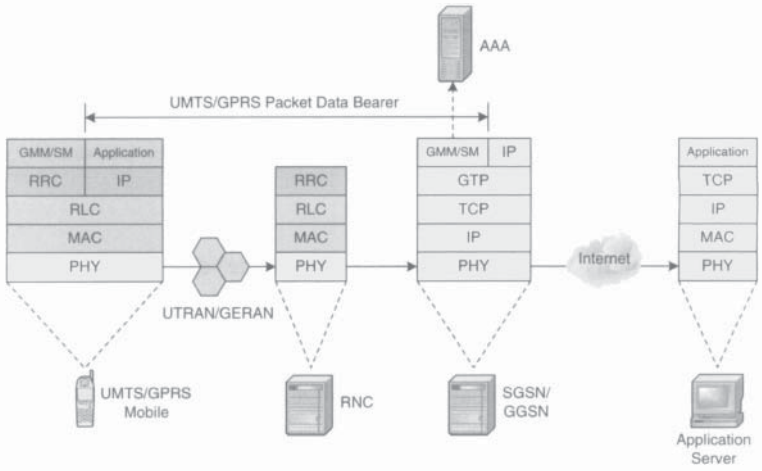


Figure 7.30 End-to-end protocol architecture of wireless internet access for GPRS/UMTS system.

7.6 SUMMARY

In this chapter, we have reviewed the system architecture of GSM and GPRS network. We started from the radio interface of the GSM and GPRS systems. Both radio interfaces are designed based on FDMA/TDMA where the spectrum is partitioned into 200kHz carrier by FDMA. Each 200kHz carrier is further partitioned into 8 time slots based on TDMA. In GSM system, each mobile can occupy one time slot only. The physical time slots are grouped into various logical channels. There are two types of logical channels defined in GSM systems, namely the *traffic channels* (TCH) and the *control channels* (CCH). TCHs are intended to carry user information (e.g., speech packets or data packets) while CCHs are used to carry control or signaling information. We have elaborated on the super-frame structure of the GSM system as well as the specific types and usage of the control logical channels and traffic logical

channels. In GPRS system, each mobile is allowed to be allocated multiple downlink time slots to boost up the physical bit rate for high speed wireless data applications. Similarly, we have elaborated on the various types of new logical channels to support the GPRS operation.

Next, we have focused on the core network design. The GSM core network is designed based on circuit-switched architecture and therefore interfaces directly to the PSTN (which is also circuit-switched) via PCM-64 (payload path) and SS7 (signaling path). To support packet switched data service, the GPRS core network overlays another layer of packet-switched based core network (SGSN/GGSN). Hence, this new layer of GPRS core network interfaces naturally to the external public data network such as internet. In the GSM system, the protocol architecture for the control plane as well as the mobility management, call setup, and international roaming procedures are introduced. Examples are given to illustrate the signaling flow and interactions between the MSC, VLR and HLR. In the GPRS core network, we have elaborated the protocol architecture in both the user plane and control plane. We have elaborated the GPRS mobility management procedures (GPRS attach/detach) as well as the session management procedures (PDP context activation/modification/deactivation). A GPRS mobile can be online (or visible to the external network) only after GPRS attach and PDP context activation. Finally, an example is given to illustrate the protocol architecture of the end-to-end IP-based bearer offered by GPRS network.

PROBLEMS

7.1 To upgrade from GPRS to EDGE, the operator does not need to change any base station hardware [True/False].

7.2 Which of the following entities are involved in the transport of traffic payload information of circuit switched data connection in GSM system?

- Base Transceiver Station (BTS)
- Interworking Function (IWF)
- Visiting Location Register (VLR)
- Home Location Register (HLR)

7.3 Which of the following is involved in the packet data traffic flow of GPRS system?

- Interworking Function (IWF).
- Mobile Switching Center (MSC).
- GGSN
- SGSN

7.4 To avoid tromboning effect in international roaming, one has to avoid routing the call setup to the GMSC of the home network whenever possible [True/False].

7.5 Location update is used to balance out the paging signaling loading [True/False].

CHAPTER 8

WIDEBAND CDMA AND BEYOND

One half of the world must sweat and groan that the other half may dream.
—H. W. Longfellow: Hyperion, I, 1839

8.1 INTRODUCTION

Engineers and practitioners have always been trying hard to utilize the elegant results from wireless communication research to produce high quality systems that can meet the ever increasing demands from consumers. Third generation (3G) mobile communication systems are the most recent prominent examples. Specifically, in 3G systems, channel-adaptive technologies are used so as to further boost the bandwidth efficiency of the wireless spectra. As detailed in this chapter, high performance channel-adaptive scheduling techniques are implemented in 3G systems to enable high quality services to the heterogeneous demands of various diverse mobile applications such as video phone, multimedia messaging and traditional voice services. We shall focus on the Universal Mobile Telecommunications System (UMTS) as illustration.

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

UMTS, being a wideband CDMA (WCDMA) standard, can manifest in two different versions in pairing uplink and downlink: FDD (frequency division duplex) and TDD (time division duplex). In the FDD mode (the paired spectrum mode due to the need of two 5 MHz wide frequency bands for uplink and downlink.), each physical channel comprises of a unique code. In the TDD mode (the unpaired specturm mode due to the need of only one 5 MHz wide frequency band), each physical channel comprises of a unique code and unique time-slots in a frame. Notice that the chip rate in UMTS is 3.84 Mcps and the frame duration is 10 ms. The 10ms frame is further divided into 15 slots. Thus, we have a total of 2,560 chips per slot, leading to a symbol rate of 2,560 symbols per slot. With a spreading factor ranging from 4 to 256 for an FDD uplink and from 4 to 512 for an FDD downlink, the channel symbol rate ranges from 7,500 symbols per second to 960 Ksymbols per second. On the other hand, the spreading factor in TDD mode ranges from 1 to 16 and thus, the channel symbol rate ranges from 240 Ksymbols per second to 3.84 Msymbols per second.

UMTS can support both circuit switched connections (e.g., for conventional voice services) and packet switched connections. Indeed, packet switched connections are envisioned as the major driving force in the 3G market because such connections enable on-demand variable rate application services (e.g., multimedia messaging, video phone and location based services, etc.) and is well-suited for bursty sources.

The scheduling actions in the early UMTS standard—Rel 99—are located at the network (RNC). As will be elaborated in this chapter, the packet data scheduling in UMTS—Rel 99 is based on a *macroscopic time scale* where the resources allocated to a *data bearer* are dynamically adjusted based on the source buffer status. For example, if there are ten packet switched data users in the cell, there will be ten packet switched data bearers setup in which the radio resource assigned to these ten data bearers will be dynamically adaptive to the individual buffer status. On one hand when the buffer has plenty of packets, high data rate dedicated physical traffic channels will be setup for the data bearer. On the other hand, when the buffer is empty, the dedicated physical traffic channels associated with the data bearer will be released. As discussed in Chapter 3, this approach does not exploit the multi-user selection diversity and therefore, the bandwidth efficiency is low. Nevertheless, the performance of the packet data services in the UMTS Rel 99 is already much more superior than that in GPRS or EDGE systems due to the faster data rates in the dedicated physical channels.

To exploit the advantages of channel adaptation and multi-user selection diversity, the High-Speed Downlink Packet Access (HSDPA) scheme is incorporated in a more recent version of the UMTS standard—Rel 5. The HSDPA system adopts a *microscopic scheduling* approach on the radio resource (time-slots/codes) for its serving active users over very short duty cycles (up to 2ms). To facilitate the microscopic scheduling, the scheduling algorithm resides at the base station instead of the RNC to minimize the potential delay in the execution of scheduling. The base station obtains the instantaneous channel quality estimates from the mobiles (UEs) and selects one mobile (UE) to transmit on the high data rate shared traffic channel at the current

time-slot based on a scheduling algorithm (which takes into consideration of system throughput, QoS requirements, fairness, or a combination of these). This is fundamentally different from the Rel 99 approach. For example, if there are ten packet switched data users in the cell, there will be one high data rate traffic channel shared dynamically between these ten users in the HSDPA systems. In fact, it is shown that the microscopic scheduling approach in the HDSPA systems is theoretically optimal due to the multi-user selection diversity.

In Section 8.2, we first briefly introduce the key features of the UMTS architecture. For the fine details and complete specifications, the reader is referred to the following excellent texts and research articles: [192, 24, 141, 151, 96, 343, 125, 260]. In Section 8.3, we shall discuss the design of the *macroscopic scheduling* mechanism in the UMTS Rel 99. In Section 8.4, we shall focus on the *microscopic scheduling* mechanism in the HDSPA systems.

8.2 UMTS ARCHITECTURE

A high level view of the UMTS architecture is shown in Figure 8.1 We can see that the UMTS architecture is divided into three main components: User Equipment (UE), UTRAN (UMTS Terrestial Radio Access Network), and Core Network (CN). Following the GSM architecture, UE is a mobile device consisting of a Subscriber Identification Module (SIM) and a wireless transceiver with the appropriate air interface equipment (WCDMA in this case). UTRAN is responsible for handling all radio-related functionalities. CN, adopted from the GSM architecture for easy migration of technology, is responsible for switching/routing calls/data connections to/from external networks such as ISDN, PSTN, and the Internet.

In the UTRAN, the radio resource management functions such as the overload control, admission control, code allocation, outer-loop power control, soft handover and dynamic bearer reconfigurations are primarily residing in the RNC. The Node B is responsible for the physical layer processing over the air such as channel coding, interleaving, rate adaptation and spreading. Manifested as a base station (BS), a Node B also handles inner loop power control and diversity combining for softer handover between different sectors of the Node B. In Rel 5, it is also responsible for "fast scheduling", as detailed in Section 8.4. Each RNC is designated to handle multiple Node Bs and there can be multiple RNCs in a UTRAN. The Node B and RNC are connected using fixed line transmission facilities such as ATM and the connection is defined as the Iub interface. The RNC may also be inter-connected with each other and the connection is defined as the Iur interface. Besides offering higher physical bit rate over the air, the UMTS system is differentiated from the 2.5G systems (such as GPRS) in the Quality of Service (QoS) dimension as well. For instance, UMTS offers four QoS classes over the wireless interface and the QoS differentiation is enforced at the RNC as well.

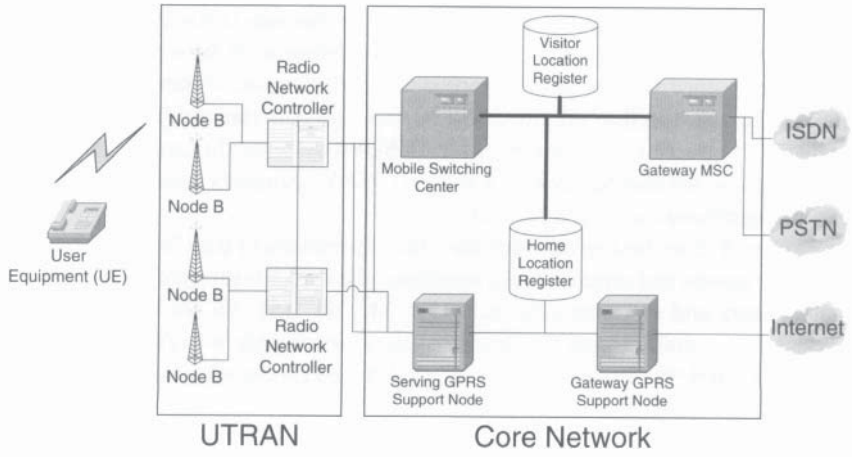


Figure 8.1 A high level view of the UMTS architecture, composed of the User Equipment (UE), Radio Network Controller (RNC), Node B, Home Location Register (HLR), Visitor Location Register (VLR), Mobile Switching Center (MSC), Gateway MSC, Serving GPRS Support Node (SGSN), and Gateway GPRS Support Node (GGSN).

In the CN, the primary role is to support the switching of circuit switched connections as well as packet switched connections. Unlike the GSM or IS-95, UMTS offers both circuit switched services and packet switched services. However, there is an implicit challenge in the design of the core network in UMTS. On one hand, the target public network of the circuit switched service is essentially the Public Switched Telephone Network (PSTN), which is based on circuit switching architecture. On the other hand, the target public network of the packet switched service is essentially the Internet, which is based on the packet switching architecture. Hence, a fundamental issue on the core network design of UMTS system is whether it should be circuit switched or packet switched. In Rel 99, an *overlay approach* is adopted. Basically, the core network has two layers, namely the circuit switched layer and the packet switched layer. In the circuit switched layer, it consists of Mobile Switching Center (MSC), which is the core element for mobility management, authentication, and switching of users' calls. The CN is connected to the PSTN via the Gateway MSC (GMSC) over the PCM-64 payload interface and SS7 signaling interface respectively. In the packet switched layer, it consists of Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). Both the SGSN and GGSN are IP routers with additional functionality such as mobility management, authentication and data encryption for packet switched services. The CN is connected to the public internet via the GGSN which routes the IP packets between the UMTS CN and the internet. In addition to the switching elements in the CN, there are distributed database to support user mobility and supplementary services such as call forwarding. They are

the Home Location Register (HLR) and the Visiting Location Register (VLR). HLR is a database for storing the master copy of a user's service profile which includes the user's service specifications and dynamic status information such as call forwarding details. VLR is a database for storing a visiting user's service profile.

UMTS is an evolving standard. The above-mentioned architecture is the basic structure as specified in UMTS standard Rel 99. More recent releases are Rel 4 and Rel 5 (as such, Rel 99 is retrospectively called Rel 3). The major addition in Rel 4 is the specification of using ATM for QoS (quality-of-service) control inside the CN. On the other hand, in Rel 5, an all IP based network is specified to be used inside the CN. In other words, a single IP-based core network is proposed in the UMTS-Rel 5 to interface with both the PSTN and public Internet. While the interface to the public internet is straight-forward, the interface of the circuit switched voice service to the PSTN is done via VoIP gateway. In addition, an IP Multimedia Sub-system (IMS) is added to interface with the GGSN in the CN. In the IMS, there are three components, namely the Media Resource Function (MRF), Call Session Control Function (CSCF), and Media Gateway Control Function (MGCF). MRF is responsible for controlling media stream resources and the multiplexing of different media streams. CSCF is the gateway interfacing the GGSN and the IMS. CSCF is also used as a firewall separating the CN from other operators' network. MCCF is for handling protocol conversions. We shall not elaborate further on the core network operation as this is beyond the scope of this chapter.

8.2.1 Radio Interface

The radio interface protocol architecture is shown in Figure 8.2 The protocol stack can be partitioned into two vertical planes, namely the control plane and the user plane. The control plane is responsible for the signaling, coordination and control functions in call setup (both circuit switched and packet switched), call release, radio resource management, session management as well as mobility management. The user plane is responsible for the user applications such as voice, video streaming as well as TCP/IP. In practice, the protocol layers above the physical layer are usually implemented in firmware and the protocol layers in both planes co-exist as concurrent threads or processes in the mobile phone.

The upper protocol layers in both the user plane and the control plane are all supported by a common suite of lower layers, namely the Radio Link Control (RLC) layer, the Medium Access Control (MAC) layer as well as the Physical layer (PHY). The RLC layer is responsible for offering *logical channels* to the upper layer applications for both planes. For example, packet retransmission, segmentation and re-assembly are the primary services offered by the RLC layer. The MAC layer is responsible for the multiplexing between the *transport channels* (offered by the physical layer) and the logical channels (offered to the RLC layer). The PHY layer is responsible for offering various *transport channels* to the MAC layer based on various physical channels. It is responsible for the modulation, coding, spreading, power control and

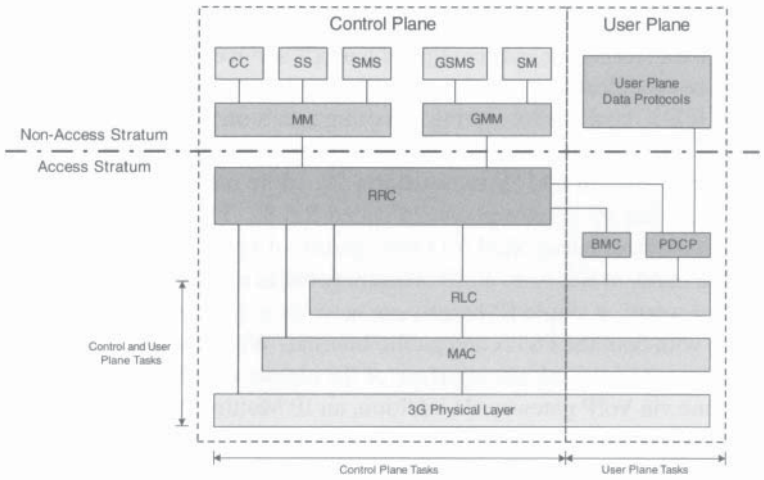


Figure 8.2 The radio interface protocol stack in UMTS.

the multiplexing of packets from various transport channels into physical channels. Note that the RLC layer and the MAC layer terminate at the RNC while the PHY layer terminates at the Node B.

In the user plane, user specific applications such as TCP/IP are interfaced to the logical channels offered by the RLC layer via the Packet Data Convergent Protocol (PDCP), which is mainly responsible for IP header compression. The PDCP layer terminates at the RNC and therefore, the IP header is re-generated at the RNC and delivered to the SGSN as a regular IP packet as illustrated in Figure 8.3 The IP layer in the user specific application terminates at the GGSN (transparent to the UTRAN) and the TCP layer is transparent to the UMTS network and the internet.

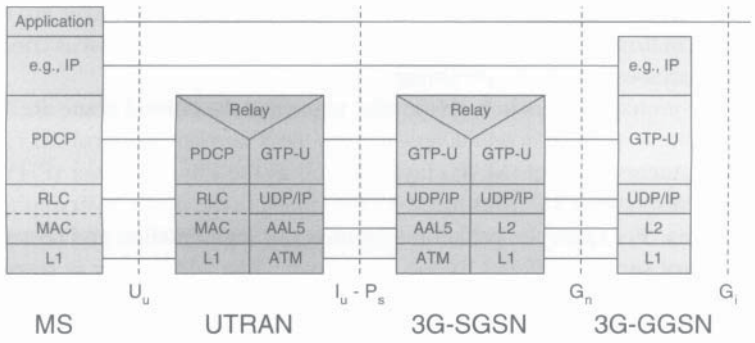


Figure 8.3 UMTS user plane protocol architecture.

In the control plane, there are various upper layer signalling protocols, namely the Radio Resource Manangement (RRC), the *Mobility Management* (MM) as well as the Session Management (SM). The RRC layer, which terminates at the RNC, is responsible for the radio dependent resource management such as outer-loop power control, soft handoff, admission control, overload control and dynamic bearer reconfiguration. The MM and SM layer terminate at the core network and are responsible for the location area / routing area update as well as QoS management of the established sessions. This is illustrated in Figure 8.4 We shall elaborate the functions of the PHY layer, MAC layer, RLC layer and RRC layer in the following sections.

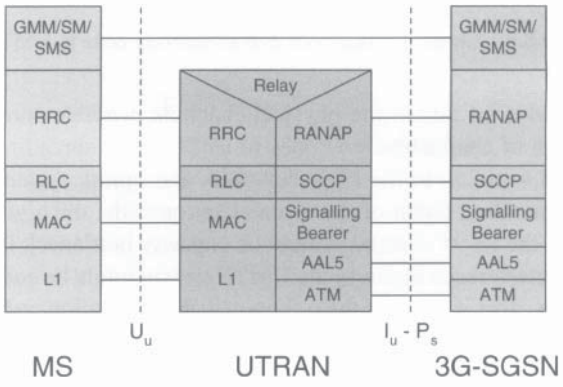


Figure 8.4 UMTS control plane protocol architecture.

8.2.2 PHY Layer

The PHY layer of UMTS is based on wideband CDMA technology. Specifically, the user separation is done based on hybrid *deterministic CDMA*[1] and *random CDMA* approaches. In the downlink, each physical channel is spread using a *channelization code* and a *scrambling code* as illustrated in Figure 8.5 The channelization code is used for the actual spreading operation (bandwidth expansion) and is targeted for the separation of intra-cell users whereas the scrambling code is used for the suppression of inter-cell interference. The downlink scrambling code is unique per cell. In the uplink, the physical channels from the mobiles are spread by both the channelization code and the scrambling code. However, the channelization code, which is used for the actual spreading operation (bandwidth expansion), is targeted to separate different physical channels from the same mobile whereas the scrambling code is used to suppress interference between different users (intracell as well as intercell).

[1] Deterministic CDMA refers to the case where the spreading codes between different users are completely orthogonal.

Unlike the downlink case, each mobile is assigned a unique scrambling code (by the RNC).

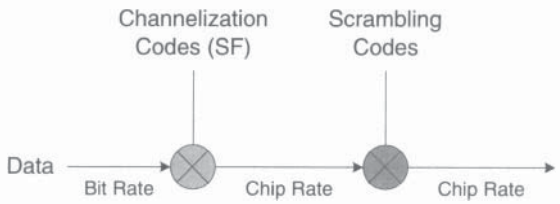


Figure 8.5 Channelization code and scrambling code in UMTS.

To support various bit rate in the physical channels, variable spreading factor is adopt in the design of channelization codes in UMTS. The spreading factor ranges from 4 to 512 and 4 to 256 in the FDD downlink and uplink systems respectively. The smaller the spreading factor of the channelization code, the higher the physical bit rate is. Unlike the IS-95 systems where the capacity bottleneck is usually in the uplink direction (interference limited), the UMTS system might be code limited in the downlink direction. This is because the orthogonal channelization codes are designed based on the *Orthogonal Variable Spreading Factor* (OVSF) codes as illustrated in Figure 8.6 On one hand, if all the users are low data rate users or voice users using channelization codes with spreading factor of 128, the downlink can support at most 128 orthogonal channels. However, if there is one high data rate user using a physical channel with spreading factor of 4 ($c_{4,1}$), all the other codes (of higher spreading factor) derived from the code $c_{4,1}$ cannot be used in order to maintain mutual orthogonality between code channels. Hence, a single high data rate user consumes $1/4$ of the voice capacity in the cell. If a UMTS cell has a few high data rate users, the number of voice channels will be quite limited due to the shortage of orthogonal codes. In other words, the system may be *code-limited*. This situation is aggregated by the way that packet switched data connections are serviced in UMTS Rel 99. For instance, if there are ten packet switched data users, there might be ten dedicated physical channels set up. One solution is to deploy the secondary scrambling code in the base station at the expense of increased mutual interference between code channels in the two scrambling code domains. In UMTS Rel 5 (HSDPA), the situation will be eased because only one high speed shared physical channel will be setup to service the various packet data users.

In UMTS systems, channels can be organized into three layers, namely the *physical channels*, the *transport channels* and the *logical channels*. A physical channel is characterized by a channelization code and a scrambling code based on WCDMA technology. A transport channel defines how and with which type of characteristics the data is transferred in the physical layer and it is the interface between the MAC layer and the PHY layer. A logical channel is a virtual connection for disseminating a

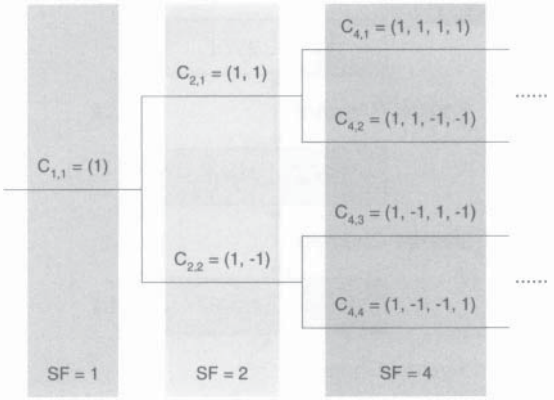


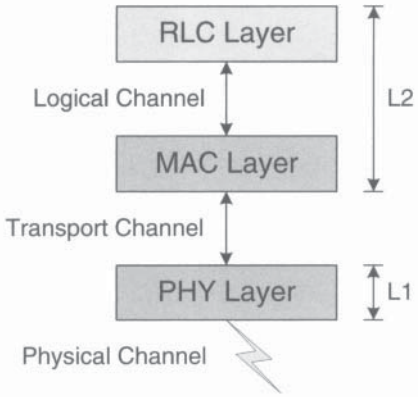
Figure 8.6 Orthogonal Variable Spreading Factor (OVSF) codes.

certain type of information (which can be control information from the control plane or data applications from the user plane) and it is the interface between the MAC layer and the RLC layer. Figure 8.7 shows the relationships among the three layers of channels in UMTS.

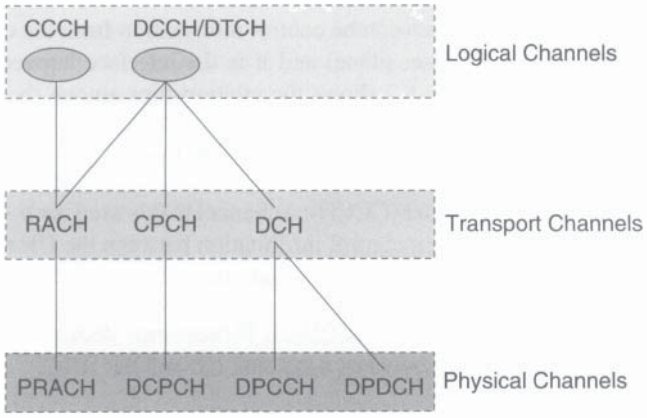
The UMTS logical channels are briefly described below:

- **Common Control Channel** (CCCH): a channel that is used for both the downlink and uplink for carrying control information between the UE and the RNC before any dedicated control channel is set up;
- **Dedicated Control Channel** (DCCH): a bidirectional dedicated channel carrying control information between a specific UE and the RNC;
- **Dedicated Traffic Channel** (DTCH): a bidirectional channel between a specific UE and the CN carrying traffic data (user plane);
- **Common Traffic Channel** (CTCH): a downlink channel carrying traffic data to a designated group of UEs;
- **Broadcast Control Channel** (BCCH): a common downlink channel carrying all the necessary system or cell specific information, such as the random access codes or time-slots, to all mobiles in the cell;
- **Paging Control Channel** (PCCH): a common downlink channel carrying paging or notification messages from the network to a specific UE, i.e., when the network wants to initiate communication with the terminal;

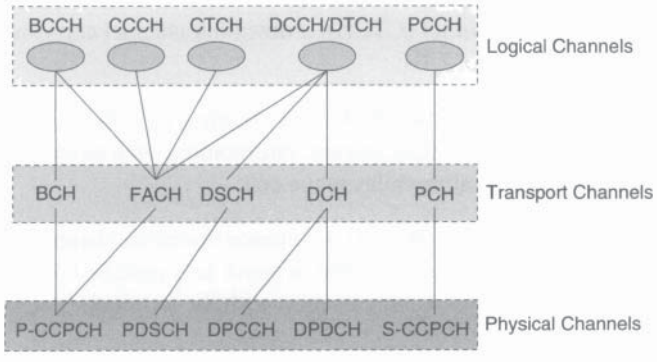
These logical channels are multiplexed or encapsulated by the transport channels in the MAC layer. Transport channels can be classified into three categories, namely



(a) three layers of channels



(b) uplink channels



(c) downlink channels

Figure 8.7 Three layers of channels in UMTS.

the *common transport channels*, the *dedicated transport channels* and the *shared transport channels*. Common transport channels are shared by all users in the cell and are sometimes referred to *overhead channels* because they usually do not contribute to the transmission of user data[2]. Shared transport channels are shared by a designated group of users only. Usually, there is no closed loop power control or soft handoff in both the common transport channels and the shared transport channels. Dedicated transport channels are set up for a specific user only and they support closed loop power control and soft handoff. Note that a dedicated logical channel is not neccessarily mapped to a dedicated transport channel in the MAC layer. This is further elaborated in Section 8.3. A list of transport channels are elaborated in the following.

- **Broadcast Channel** (BCH): a downlink common transport channel supporting the BCCH on the downlink;
- **Paging Channel** (PCH): a downlink common transport channel supporting the PCCH on the downlink;
- **Forward Access Channel** (FACH): a downlink common transport channel supporting several logical channels such as the CCCH, CTCH, DCCH, and DTCH. The FACH channel can be used to carry signaling information or a small amount of user data to a UE within a cell. A cell can have several FACH channels but the primary one must have low data rate in order to be received by all users. In-band signaling is used in the FACH to indicated for which user the data is intended;
- **Random Access Channel** (RACH): an uplink common transport channel supporting the CCCH and DCCH to carry control information or supporting DTCH to carry small amount of bursty data;
- **Common Packet Channel** (CPCH): an uplink common transport channel supporting the DCCH to carry control information or DTCH to carry small amount of bursty data;
- **Downlink Shared Channel** (DSCH): a downlink shared transport channel supporting DCCH or DTCH;
- **Dedicated Channel** (DCH): a dedicated downlink and uplink transport channel supporting DCCH and DTCH between the network and a specific UE;

The data carried by the transport channels are delivered by the MAC layer to the physical layer in the basic units of *transport blocks* once every *Transmission Time Interval* (TTI) (10ms, 20ms, 40ms or 80ms). The transport format of each transport block is identified by the Transport Format Indicator (TFI), which is used in the inter-layer communication between the MAC layer and the PHY layer. Several transport

[2]In fact, RACH and FACH can be used to carry short burst of user data.

channels can be multiplexed together into a single Code Composite Transport Channel (CCTrCh) in the physical layer as illustrated in Figure 8.8 Cyclic redundancy check (CRC) and channel encoding are performed on a per-transport channel basis. The encoded bit stream from a transport channel is zero-padded, interleaved (over the TTI) and partitioned into a number of radio frames. For example, a transport channel with TTI = 40ms is partitioned into four radio frames after channel encoding and CRC. Various transport channels, each may have different TTIs, channel coding rates as well as data size, are multiplexed together to form a single CCTrCh based on a *rate matching algorithm* on a 10ms frame-by-frame basis. The CCTrCh is further mapped into a number of physical channels and transmit over the air using CDMA.

In fact, UMTS protocol architecture allows QoS specification per transport channel because each transport channel can have different channel encoding, TTI, block size, FER target as well as rate matching attributes[3]. This is a unique feature that cannot be found in other competing standards such as 3G1X. Similar to the transport channels, there are also three types of physical channels, namely the *common channels*, the *dedicated channels* and the *shared channels*. They are elaborated below.

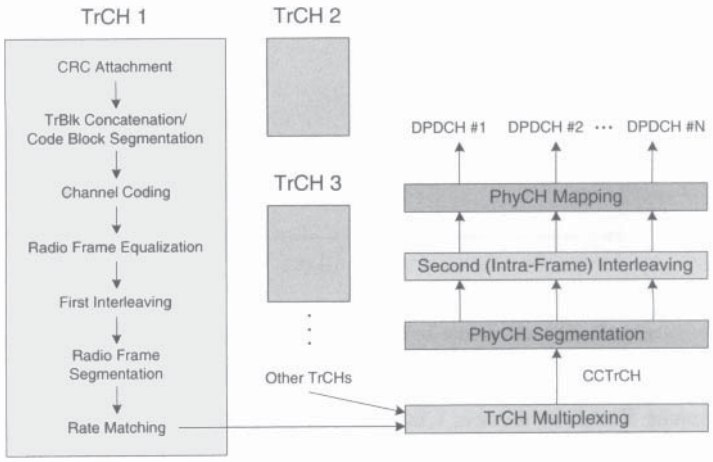
- **Common Pilot Channel (CPICH):** This is a downlink common channel with a constant spreading factor of 256 scrambled by the *cell-specific* primary scrambling code. This is used as a common reference for channel estimation and timing acquisition at the mobile.

- **Synchronization Channel (SCH):** This is a common physical channel used for cell searching by the mobiles. There are two sub-channels, namely the primary SCH and secondary SCH as illustrated in Figure 8.9 Both the P-SCH and S-SCH are sent only during the first 256 chips of each time-slot and they are transmitted in parallel. Please refer to [4, 5] for the details of the cell search operation.

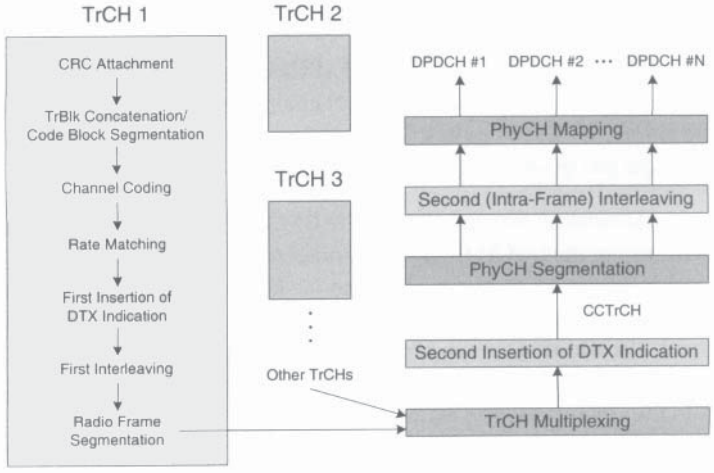
- **Primary Common Control Physical Channel** (P-CCPCH): This is a common physical channel supporting the BCH transport channel, carrying cell broadcast information. It has a fixed data rate of 30 kbps and a fixed spreading factor of 256. Since the P-CCPCH is time-multiplexed with the SCH, the first 256 chips in every time-slot is not transmitted and hence, the effective data rate is reduced to 27kbps. There is no closed-loop power control in this channel and it needs to be demodulated by all terminals in the cell.

- **Secondary Common Control Physical Channel** (S-CCPCH): This is a common physical channel used to support the PCH and FACH (transport channels) in the downlink. A fixed spreading factor of 256 is used and variable data rates can be supported by Discontinuous Transmission (DTX) and rate matching. Similar to the P-CCPCH, there is no closed-loop power control in S-CCPCH.

[3] Rate matching attribute is a parameter defined per transport channel indicating the priority of bit puncturing or repetition in the CCTrCh multiplexing.



(a) uplink CCTrCh multiplexing



(b) downlink CCTrCh multiplexing

Figure 8.8 Multiplexing of several transport channels into a single CCTrCh in the physical layer.

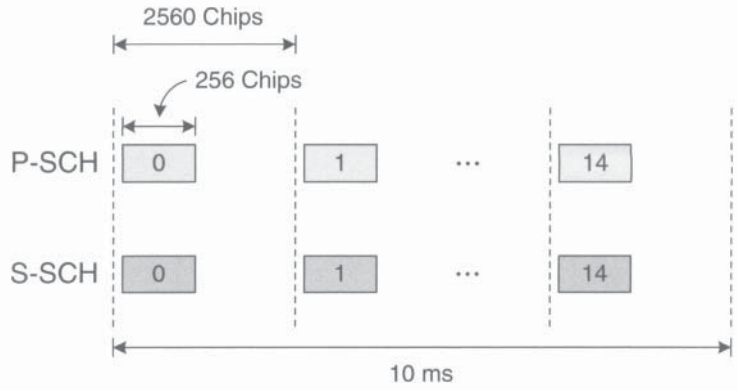


Figure 8.9 Structure of P-SCH and S-SCH for cell search.

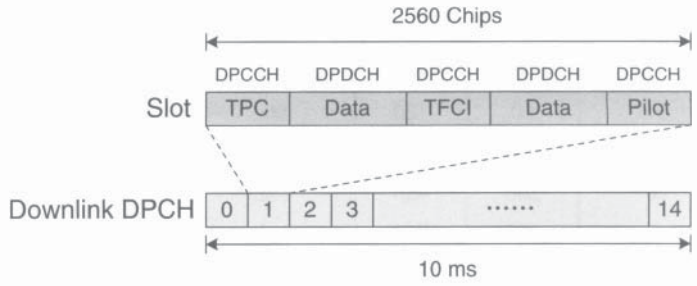
- **Physical Random Access Channel** (PRACH): This is a common physical channel supporting the RACH transport channel. It is a contention based uplink channel which can be used to carry control information or small amount of user data from the mobile to the network. Similarly, no closed-loop power control is supported.

- **Physical Common Packet Channel** (PCPCH): This is a common physical channel supporting the CPCH transport channel for carrying bursty data traffic in the uplink. However, unlike the PRACH, channel can be reserved for several frames and fast power control is supported.

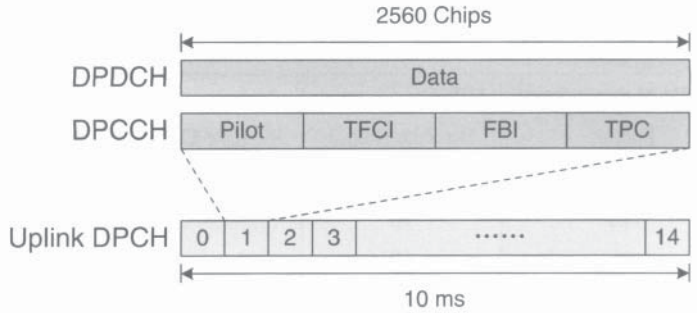
- **Physical Downlink Shared Channel** (PDSCH): This is a shared physical channel supporting the DSCH transport channel on the downlink. It is used to carry control information or bursty data and is shared by several designated users. It supports the use of fast power control but there is no soft handoff for this channel. In addition, it supports a variable spreading factor on a frame-by-frame basis so that the bit rate is adaptive to the physical channel condition. The PDSCH must always be associated with a downlink DPCH to indicate which terminal should decode the information in the PDSCH and the associated spreading code. Since the information of the PDSCH is associated with a downlink DPCH, the PDSCH frame is not started before three time-slots after the end of the associated downlink DPCH.

- **Dedicated Physical Channel** (DPCH): This is a dedicated physical channel in both uplink and downlink used to support the DCH between the RNC and a specific mobile terminal. There are two sub-channels that constitute the DPCH, namely the *Dedicated Physical Data Channel* (DPDCH) and the **Dedicated Physical Control Channel** (DPCCH). The DPDCH carries the data part of the

DCH whereas the DPCCH carries the physical layer control information such as the fast power control commands, dedicated pilots as well as the Transport Format Combination Indicator (TFCI)[4]. In the downlink, the DPDCH and the DPCCH are time multiplexed and use normal QPSK modulation as illustrated in Figure 8.10(a). The spreading factor of the DPCH does not vary on a frame-by-frame basis and the data rate is varied by the rate matching operation by bit puncturing, bit repetition or discontinuous transmission (DTX). In the uplink, the DPDCH and the DPCCH are I-Q multiplexed with different channelization codes c_D and c_C respectively as illustrated in figure 8.10(b). The DPCCH always has a fixed spreading factor of 256 whereas the DPDCH has a variable spreading factor on a frame-by-frame basis (4–256).



(a) downlink DPCH



(b) uplink DPCH

Figure 8.10 Multiplexing of DPDCH and DPCCH in downlink and uplink.

Table 8.1 illustrates the set of possible spreading factors and the associated bit rates of DPCH in the downlink and uplink.

[4]TFCI is used at the receiving side to demultiplex data from a single CCTrCh into a number of transport channels on a frame-by-frame basis.

Table 8.1 Spreading factors and bit rates in DPCH.

(a) downlink DPCH

Spreading Factor	Channel Symbol Rate (kbps)	Channel Bit Rate (kbps)	DPDCH Channel Bit Rate Range (kbps)	Maximum User Data Rate with 1/2-Rate Coding (Approx., kbps)
512	7.5	15	3–6	1–3
256	15	30	12–24	6–12
128	30	60	42–51	20–24
64	60	120	90	45
32	120	240	210	105
16	240	480	432	215
8	480	960	912	456
4	960	1920	1872	936
4, with 3 parallel codes	2880	5760	5616	2.3M

(b) uplink DPCH

DPDCH Spreading Factor	DPDCH Channel Bit Rate (kbps)	Maximum User Data Rate with 1/2-Rate Coding (Approx., kbps)
256	15	7.5
128	30	15
64	60	30
32	120	60
16	240	120
8	480	240
4	960	480
4, with 6 parallel codes	5740	2.3M

Table 8.2 summarizes the mapping of transport channels to physical channels in UMTS systems.

Table 8.2 A summary of transport channel mappings in UMTS.

Transport Channel	Physical Channel
(UL/DL) Dedicated channel DCH	Dedicated physical data channel DPDCH Dedicated physical control channel DPCCH
(UL) Random access channel RACH	Physical random access channel PRACH
(UL) Common packet channel CPCH	Physical common packet channel PCPCH
(DL) Broadcast channel BCH	Primary common control physical channel P-CCPCH
(DL) Forward access channel FACH (DL) Paging channel PCH	Secondary common control physical channel S-CCPCH
(DL) Downlink shared channel DSCH	Physical downlink shared channel PDSCH
Signaling physical channels	Synchronization channel SCH Common pilot channel CPICH Acquisition indication channel AICH Paging indication channel PICH CPCH status indication channel CSICH Collision detection/Channel assignment indicator channel CD/CA-ICH

8.2.3 MAC Layer

CPCH The MAC layer in the UMTS protocol stack is a very important component in that it serves many purposes:

- Mapping logical channels to transport channels;
- Choosing suitable transport format for each transport channel based on the instantaneous source data rate;
- Managing priority in a set of data traffic flows within each UE;
- Managing priority among different UEs using dynamic scheduling techniques;
- Identifying specific UEs on the common transport channels;

- Marshaling of data PDUs between transport block sets and the underlying physical layer dedicated channels;
- Measuring and monitoring of traffic volume;
- Ciphering for transparent mode RLC;
- Choosing access service class for RACH and CPCH sessions.

As shown in Figure 8.11, the MAC layer can be divided into four major components: MAC-d, MAC-c/sh, MAC-b and MAC-hs. MAC-d is responsible for handling dedicated channels and is residing at the Serving RNC (S-RNC)[5]. MAC-c/sh is responsible for handling the common and shared channels. MAC-b is responsible for handling the broadcast channel. Both MAC-c and MAC-b are residing at the Controlling RNC (C-RNC)[6]. MAC-hs is a high-speed extension of the MAC layer in HSDPA and is responsible for the high speed DSCH transport channel. Unlike the other MAC components, the MAC-hs is residing at the Node B to facilitate fast scheduling.

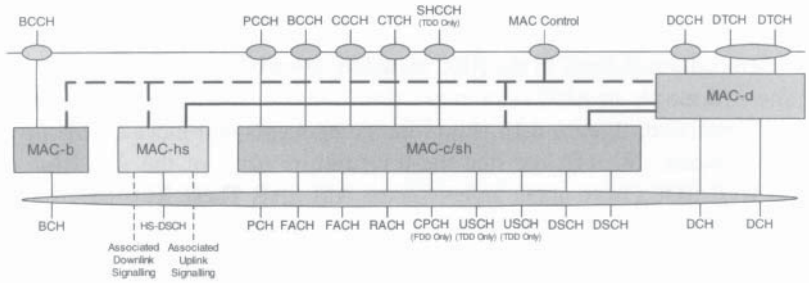
8.2.4 RLC Layer

The major role of the RLC layer is to provide segmentation and retransmission functions for the logical channels offered by the MAC layer and is residing at the S-RNC. There are three different operation modes, namely the transparent mode (Tr), unacknowledged mode (UM), and acknowledged mode (AM). The architecture of the RLC layer is shown in Figure 8.12 We can see that Tr and UM modes are characterized as unidirectional and AM mode is featured as bidirectional.

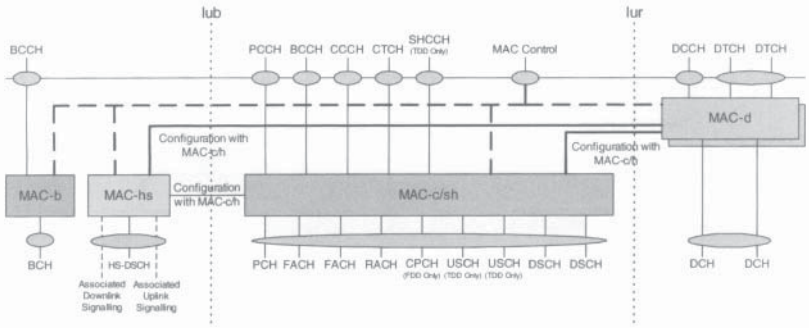
The Tr mode offers basic Service Data Unit (SDU) transfer, segmentation and reassembly and SDU discard and it does not add any protocol overhead to higher layer SDUs. Specifically, corrupted SDUs are either discarded or marked as erroneous. Segmentation and re-assembly is an optional feature in the Tr mode. If segmentation is not configured in the RLC layer, the SDU from the upper layers is used as the Protocol Data Unit (PDU) to the MAC layer. On the other hand, if segmentation is configured, the packet length of the SDU received from the upper layers must be an integer multiple of the PDU length so that segmentation is done without any RLC overhead. This mode is useful for real-time applications, such as streaming or conversational applications, where low-overhead logical channel is needed.

The UM mode offers concatenation, padding, ciphering and sequence number checking in addition to the services offered by the Tr mode. It does not offer retransmission function and as such, data delivery is not guaranteed. Consequently, corrupted data PDUs are either marked or discarded depending on the system configuration. Unlike the Tr mode, RLC protocol overhead is added in the UM mode

[5]The S-RNC is the anchor point with respect to a UE when the call is setup. As the mobile moves to other Node Bs, the S-RNC will not be changed unless the S-RNC relocation procedure has been executed. With respect to a UE, the S-RNC is unique.
[6]Each Node B has a unique RNC and this is the C-RNC.



(a) MAC layer components in each UE



(b) MAC layer components in UTRAN

Figure 8.11 Three components of the MAC layer in the UE and the UTRAN.

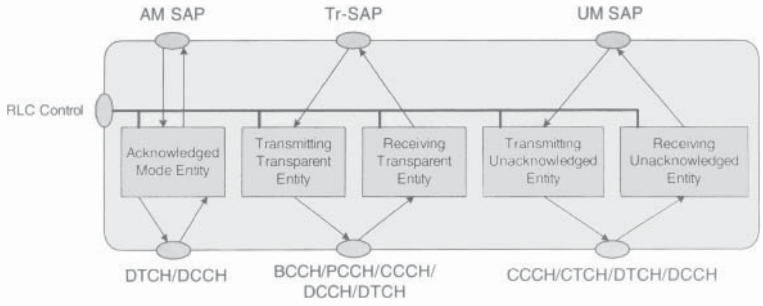


Figure 8.12 The architecture of the RLC layer.

to facilitate concatenation, padding, ciphering and sequence number checking. For instance, the UM mode is used for certain RRC signaling procedures whereas only a unidirectional logical channel is required and there are acknowledgment and retransmission procedures defined in the RRC signaling already.

In the AM mode, an ARQ (Automatic Repeat Request) mechanism is employed for error correction. Figure 8.13 shows the components and mechanisms involved in the AM mode. We can see that the transmitting party of the AM-RLC entity receives RLC SDUs from upper layers via the AM-SAP. These RLC SDUs are then segmented (in case of a large SDU) into fixed length AMD PDUs. Alternatively, the SDUs may be concatenated into an AMD PDU if the SDUs are smaller than the AMD PDU length. The AMD PDUs buffered in the retransmission buffer are either deleted or retransmitted. The MUX multiplexes the AMD PDUs from the Retransmission buffer that need to be retransmitted. Ciphering is performed (if configured) followed by sending the encrypted PDUs to either one or two DCCH or DTCH logical channels.

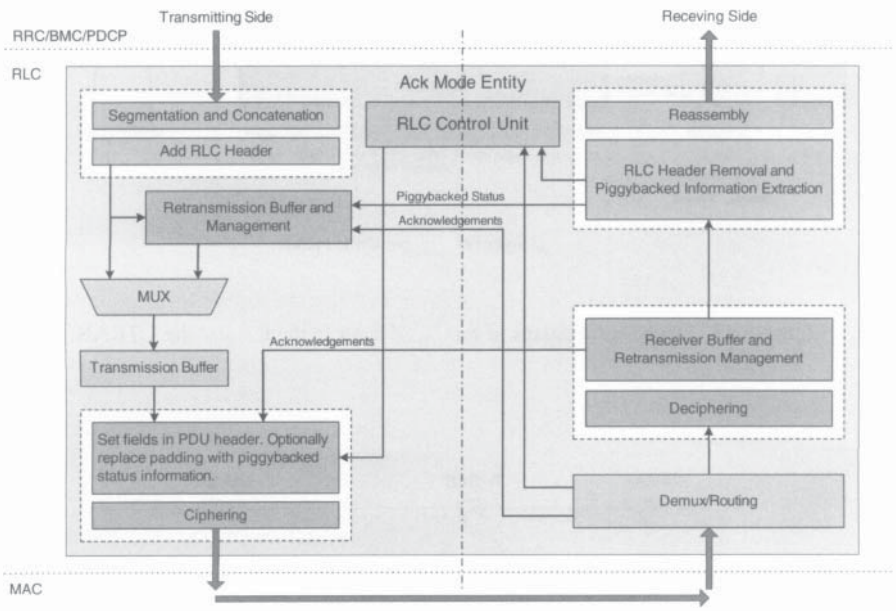


Figure 8.13 Illustration of the components and mechanisms involved in the RLC-Acknowledged Mode.

On the other hand, the receiving party of the AM-RLC entity receives AMD and control PDUs via the specific logical channels. Afterward, the AMD PDUs are routed to the Deciphering unit followed by delivery to the receiving buffer. Reassembly is performed in that the AMD PDUs are accumulated in the buffer until a complete RLC

SDU is received. The receiver then acknowledges successful reception (or requests for retransmission of missing AMD PDUs).

In summary, the RLC layer provides the following services:

- Segmentation and reassembly of higher-layer PDUs into/from smaller RLC payload units (RLC-Tr, RLC-UM, RLC-AM);
- Assembling by concatenation of RLC SDUs (RLC-UM, RLC-AM);
- Padding (RLC-UM, RLC-AM);
- Transfer of user data (RLC-Tr, RLC-UM, RLC-AM);
- Error correction (RLC-AM);
- In-order delivery of higher-layer PDUs (RLC-AM);
- Duplicate detection (RLC-AM);
- Flow control (RLC-AM);
- Sequence number check (RLC-UM);
- Protocol error-detection and recovery (RLC-AM);
- Ciphering (RLC-UM, RLC-AM);

8.2.5 RRC Layer

RRC layer is a very important layer in the control plane responsible for the radio resource management. It resides at the S-RNC and using RRC messages, the set-up, modification, and tear-down of layers 2 and 1 protocol entities can be accomplished. Mobility management functions in the UTRAN level such as measurements, handovers and cell updates with respect to a specific UE are also under the control of RRC layer. The RRC architecture is shown in Figure 8.14 We can see that there are four major functional entities:

- The Dedicated Control Function Entity (DCFE) is responsible for handling all functions and signaling for one particular UE. It relies mostly on AM mode RLC.
- The Paging and Notification control Function Entity (PNFE) is responsible for handling the paging of idel mode UEs. It mainly employs the PCCH logical channel.
- The Broadcast Control Function Entity (BCFE) is responsible for handling the system information broadcasting. It utilizes either BCCH or FACH logical channels.

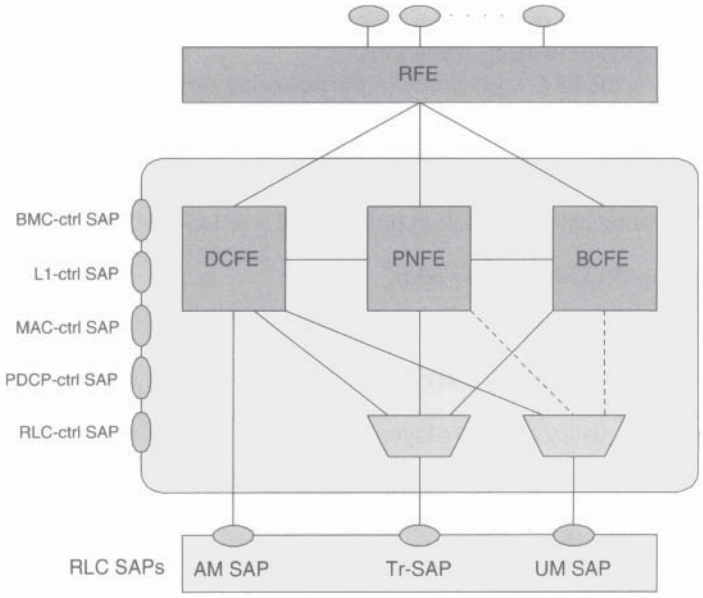


Figure 8.14 Protocol architecture of the RRC layer.

- The Routing Function Entity (RFE) is responsible for the routing of higher layer messages.

RRC interacts closely with each layer in the UMTS protocol stack so as to provide control information, and in turn receive measurement feedback from these layers. This is illustrated in Figure 8.15 Specifically, RRC provides the following functions:

- Cell Broadcast Service (CBS) control;
- Initial cell selection and cell re-selection;
- Paging;
- Broadcast of information;
- Establishment, maintenance, and release of an RRC connection between a UE and the UTRAN;
- Assignment, reconfiguration, and release of radio resources for the RRC connection;
- Control of requested QoS;
- UE measurement reporting and control of the reporting;

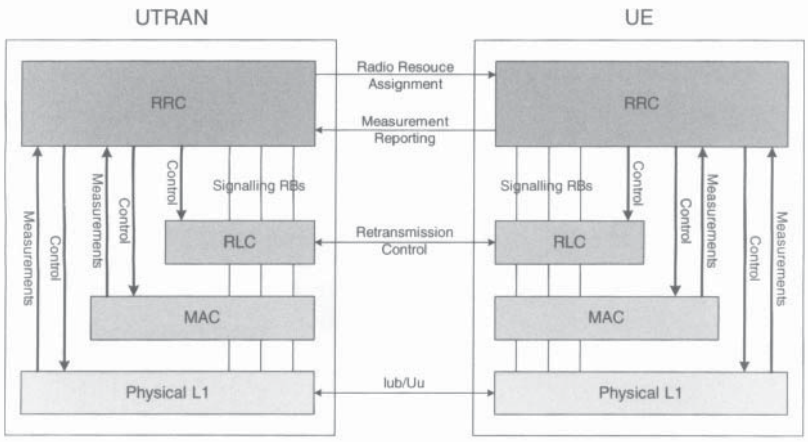


Figure 8.15 Interactions between the RRC layer and other protocol components.

- RRC message integrity protection;
- Arbitration of radio resources on uplink DCH;
- Slow dynamic channel allocation (DCA) in the TDD mode;
- Timing advance in the TDD mode;
- RRC connection mobility functions;
- Outer loop power control;
- Control of ciphering functions.

Figure 8.16 shows the transition mechanism of a UE's RRC states. After the UE is powered on, it chooses a suitable cell with the strongest pilot signal strength for service through monitoring the common downlink pilot channel. After the UE acquires the pilot channel through cell search procedure, the UE is able to receive system information and cell broadcast messages from the BCH channel. At this moment, the RRC layer at the UE is at the idle state. The UE stays in the idle mode until it sends a request to the UTRAN to establish a dedicated control channel (DCCH) with the RRC peer at the S-RNC. When this happens, the RRC layer is moved from the idle state to the *connected state*. Note that setting up a DCCH does not necessarily imply the setting up of dedicated transport channels and dedicated physical channels.

When a logical connection (DCCH) is set up between the UE and the S-RNC, both the RRC peers in the UE and the S-RNC will be at the connected state. However, there are four sub-states in the connected state, namely the *Cell-DCH* state, the *Cell-FACH* state, the *Cell-PCH* state as well as the *URA-PCH* state. In the *Cell-DCH* state, a

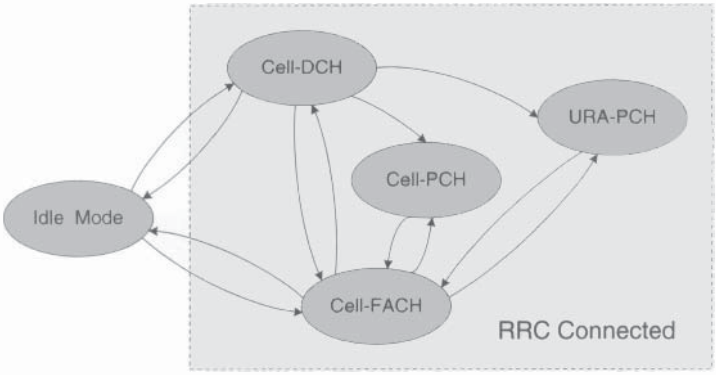


Figure 8.16 State transition diagram of RRC layer.

dedicated physical channel is allocated to the UE for the logical DCCH and DTCH (or radio bearer). In addition, the S-RNC recognizes the location of the UE down to the cell level. The UE carries out measurements and transmits measurement reports according to downlink measurement control information from the RNC. This state can be entered directly from the idle state during call setup or from the *Cell-FACH* state (as triggered by traffic volume change).

In the *Cell-FACH* state, the UE is not allocated any dedicated physical channel. The DCCH and DTCH logical channels are supported by the RACH, CPCH and FACH transport channels. In this state, the UE is also capable of listening to the BCH for acquiring system information as well as performing cell re-selections. Similar to the *Cell-DCH* state, the S-RNC recognizes the location of the UE down to the cell level. Hence, after a re-selection process is performed, the UE needs to send a *cell update* message to notify the S-RNC about its new location. This state can be entered directly from the idle state or from the *Cell-DCH* state (as triggered by traffic volume change).

In the *Cell-PCH* state, the UE is still recognized on a cell level but it can only be reached via the PCH. The major feature of this state is that the battery consumption is less than that in the *Cell-FACH* state because monitoring of the paging channel includes a discontinuous reception (DRX) capability. The UE also obtains system information on the BCH. If the UE performs a cell re-selection, it has to move to the *Cell-FACH* state first so as to carry out the *cell update* procedure.

Finally, in the *URA-PCH* state, it is very similar to the *Cell-PCH* state except that the UE needs to execute the *cell update* procedure only if the UE detects a change in the *UTRAN Registration Area* (URA). The UE obtains the URA of the current cell from the BCH and a cluster of cells will share the same URA. This arrangement is to reduce the uplink signaling loading to the system due to *cell updating*. The UE in this state is basically in a dormant mode and the DCCH cannot be used in this state.

The UE has to move to *Cell-FACH* state to re-activate the DCCH for any potential activities.

We shall illustrate the important role of the RRC states in the dynamic bearer reconfiguration of packet switched connections in the next section.

8.3 PACKET SWITCHED CONNECTIONS IN UMTS (REL 99)

As variable rate bursty data services are the major new incentives for using 3G wireless systems, great efforts have been exerted on designing efficient algorithms to enable such services. However, provisioning of packet data services efficiently is much more challenging than that of the constant bit rate voice services because of the inherent heterogeneity in the users' requirements and their associated resource conditions (e.g., signal-to-noise ratio, channel quality, data urgency, etc.).

Specifically, there are four basic types of traffic classes supported by UMTS:

- **Conversational Class**: real-time symmetric traffic, mostly voice services;
- **Streaming Class**: real-time asymmetric traffic, usually from the network to the user, e.g., video streaming;
- **Interactive Class**: non-real-time packet traffic, high integrity, e.g., Web surfing;
- **Background Class**: non-real-time packet traffic, delay insensitive, e.g., emails.

Judicious radio resource management is needed to satisfy users' QoS requirements while maximizing the system's performance and utilization. In UMTS Rel 99, the scheduler is residing at the S-RNC and is responsible for the dynamic allocation of the radio resources, in terms of time durations, codes, and power levels, in the radio bearers set up for the packet data users over a macroscopic time scale. As will be discussed below, the scheduler needs to decide on what types of channels to use for different UEs in a dynamic manner according to the system loading and the user buffer levels. In the followings, we shall focus on the packet switched data connection setup and dynamic reconfiguration incorporated in the Rel 99 standard of UMTS.

8.3.1 Radio Bearer for Packet Switched Users

In the UMTS systems, the *radio bearer* is defined as a logical channel between the UE and the RNC to carry data payload (from the user plane) or signaling data (from the control plane). In other words, the radio bearer is the logical channel offered by the RLC layer to the upper layers. If the radio bearer is used to carry signaling data, it is called the *signaling radio bearer*. Otherwise, it is called the *data radio bearer*.

While the data radio bearer is a logical channel for the user plane, it can be mapped into three types of transport channels, namely the *common transport channels* (CPCH,

RACH, FACH), the *dedicated transport channels* (DCH) and the *shared transport channels* (DSCH). Common transport channels can be used to carry user packet data. In a typical deployment scenario, there are usually only one or just a handful of RACH and FACH channels per sector. A distinctive merit of using common transport channels for user data is that the set-up time is minimal because these channels are already setup. However, there is no closed loop power control as well as soft handoff. Instead, the cell re-selection mechanism is employed when the UE moves to the next cell. Hence, this is most suitable for transmitting small and bursty user packets.

Dedicated transport channels are bidirectional in nature and support fast power control and soft handoff. Radio link level performance is thus greatly enhanced. However, the set-up time for dedicated channels is usually quite long and this introduces setup overhead. In addition, the dedicated transport channels will consume more code space in the downlink. Hence, it is most suitable for transmitting large volume of user packets.

The shared transport channel (DSCH) operates in a time-division scheduling manner in that a single orthogonal code is shared among many users in the time domain. Compared with the dedicated transport channels, the DSCH can save the code consumption in the downlink as well as reduce the associated setup time. Hence, it is suitable for transmitting large volume or bursty user packets. In fact, in HSDPA, a high speed version of the DSCH (HS-DSCH) is created to support fast scheduling of packet data users. A short-coming of the DSCH is that it does not support soft handoff.

Table 8.3 summarizes the possible mappings of various transport channels for supporting data radio bearer.

8.3.2 Setup of Packet Switched Connection

In order to set up a packet switched connection for a data user, the UE has to first establish a DCCH (either based on DCH or FACH/RACH). This is initiated by the UE sending a *RRC connection request* message to the S-RNC via the RACH channel (which is mapped to the CCCH). The corresponding RRC state change can be from the idle state to the Cell-DCH state or from the idle state to the Cell-FACH state. Once the RRC connection is successfully set up, the next step is to establish a data radio bearer. Specifically, a *radio bearer setup* message is sent from the S-RNC to the UE over the DCCH (already established RRC connection). Figure 8.17 illustrates an example of packet switched data connection setup with low traffic volume. The RRC-state of the UE is changed to the *Cell-FACH* after the DCCH is set up. Next, the data radio bearer (RB1) is also setup based on the FACH transport channel.

Unlike the circuit switched connection, the resource requirement of a packet switched connection has to be dynamically adjusted because of the bursty nature of the source. This is called the *dynamic bearer reconfiguration* or scheduling over macroscopic time scale and the decision is made by the RRC layer at the S-RNC. To facilitate the S-RNC to perform dynamic bearer reconfiguration, the RRC layer

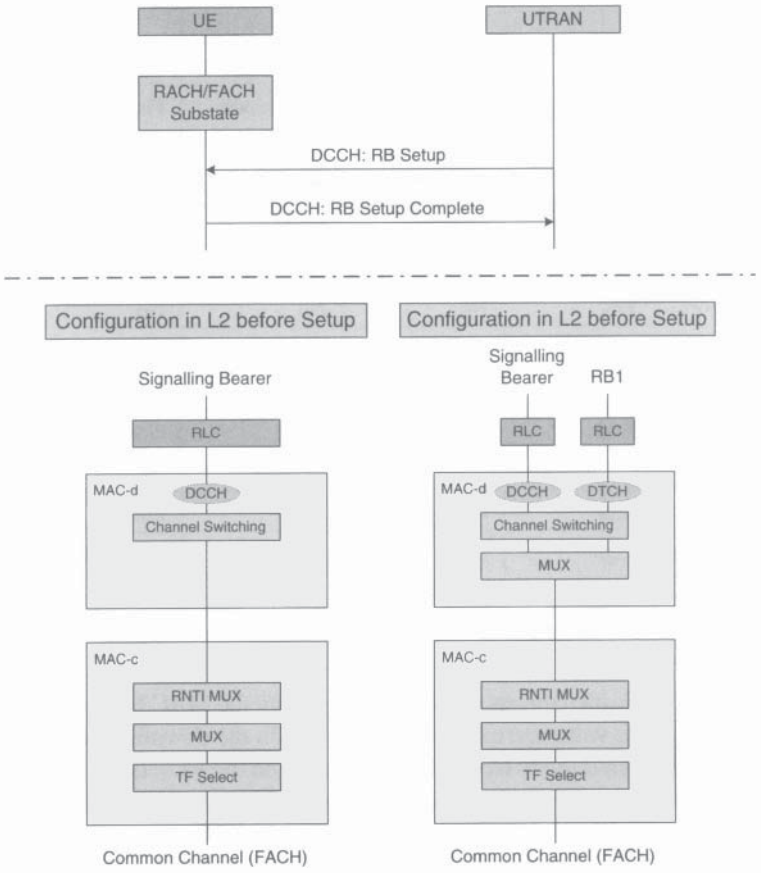


Figure 8.17 An example of packet switched data connection setup in UMTS with low traffic volume.

Table 8.3 A summary of transport channel mappings to support data radio bearer in UMTS.

TrCh	DCH	RACH	FACH	CPCH	DSCH
TrCH Type	Dedicated	Common	Common	Common	Shared
Applicable UE State	CELL_DCH	CELL_FACH	CELL_FACH	CELL_FACH	CELL_FACH
Direction	Both	Uplink	Downlink	Uplink	Downlink
Code Usage	Accordingly to max. bit rate	Fixed code allocations in a cell	Fixed code allocations in a cell	Fixed code allocations in a cell	Fixed code allocations in a cell
Power Control	Fast closed-loop	Open-loop	Open-loop	Fast closed-loop	Fast closed-loop
SHO Support	Yes	No	No	No	No
Target Data Traffic Volume	Medium or high	Small	Small	Small or medium	Medium or high
Suitability for Bursty Data	Poor	Good	Good	Good	Good
Setup Time	High	Low	Low	Low	Low
Relative Radio Performance	High	Low	Low	Medium	Medium or high

has to monitor the instantaneous system loading (from the MAC layer) as well as the instantaneous traffic volume (from the RLC layer). In the downlink, these measurement can be obtained directly from the MAC layer and the RLC layer of the S-RNC. In the uplink, the traffic volume measurement is performed at the UE RLC layer per instruction of the S-RNC[7]. This is illustrated in Figure 8.18

Suppose the downlink traffic volume increases such that the buffer level at the RLC rises. The S-RNC may trigger a *transport channel re-configuration* as illustrated in Figure 8.19 Dedicated transport channels (DCH1 and DCH2) are setup to support the existing DCCH and the radio bearer (RB1) in the reconfiguration in response to the rise in the downlink traffic volume. In this case, the RRC state changes from the *Cell-FACH* state to the *Cell-DCH* state because dedicated transport channels are setup after reconfiguration.

[7]To trigger uplink traffic volume measurement, the S-RNC can send a *traffic volume measurement* request to the UE and the UE responses by sending a *measurement report* to the S-RNC. The traffic volume measurement can also be configured as periodic or event driven.

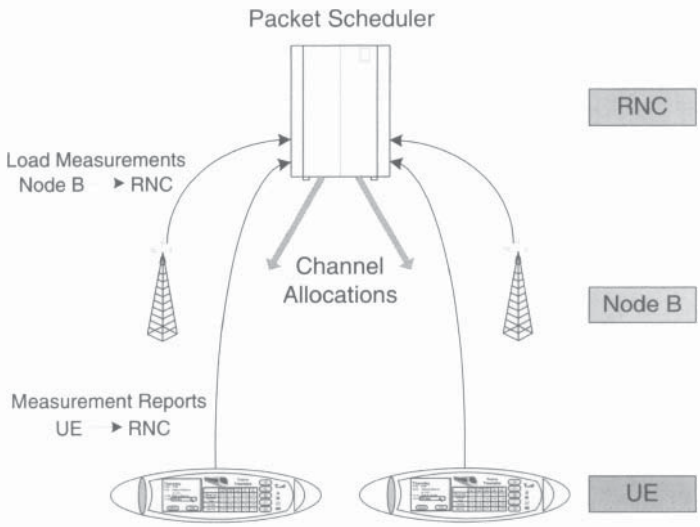


Figure 8.18 A schematic diagram illustrating the scheduling mechanism in UMTS Rel 99.

On the other hand, the dedicated transport channels may be released and replaced by the common transport channels to support the DCCH and RB1 in the example when the traffic volume drops. This transition is usually triggered by the *inactivity timer* which times the idle period when the buffer in the RLC layer is empty. When this happens, the RRC state will change from the *Cell-DCH* state back to *Cell-FACH* state. Figure 8.20 summarizes the RRC state changes in the lifetime of a packet switched data connection.

8.4 PACKET SCHEDULING IN HSDPA (REL 5)

The HSDPA concept can be seen as a continue evolution of the DSCH time division scheduling and a new transport channel, namely the *high speed DSCH* (HS-DSCH) is defined for more efficient scheduling of high bit rate packet data users. In addition, fast scheduling is made possible by introducing a new MAC layer called the *MAC-hs* residing at the Node B. The fast scheduling (over 2ms intervals) together with the HS-DSCH transport channel allow the HSDPA system to exploit multi-user selection diversity over the microscopic fading channels and therefore achieving higher peak data rate, spectral efficiency as well as QoS control for bursty and downlink asymmetric packet data users. In the following, we shall give an overview of HSDPA.

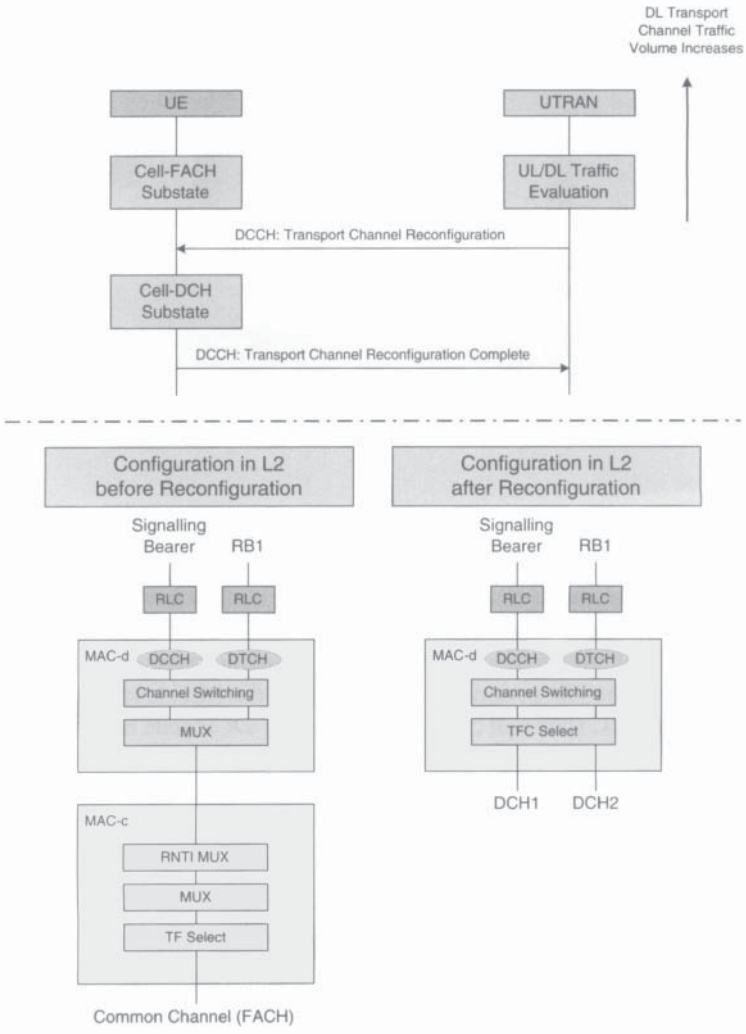


Figure 8.19 An example of transport channel re-configuration due to an increase of downlink traffic volume.

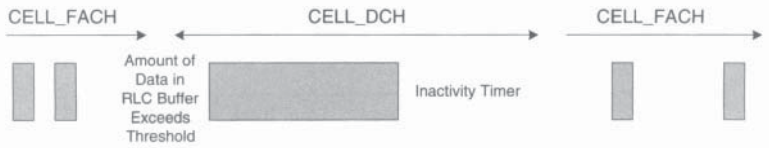


Figure 8.20 An example of RRC state changes in the lifetime of a packet switched data connection.

8.4.1 Key Enabling Technologies in HSDPA

There are two key enabling features in HSDPA, namely the adaptive modulation and channel coding, hybrid ARQ in the physical layer (HS-DSCH) as well as the fast scheduling at the MAC-hs layer in Node B. They are elaborated as follows.

- **Adaptive Modulation, Channel Coding and Multi-code transmissions:**

  On the HS-DSCH channel, two fundamental CDMA features, namely the variable spreading factor and the fast power control[8], have been deactivated and replaced by *short packet size*, *adaptive modulation and coding* (AMC), *multi-code operation* and *fast hybrid ARQ*. While being more complicated, the replacement of fast power control with AMC yields a power efficiency gain due to the elimination of the inherit power control overheads. In the HS-DSCH, the spreading factor is fixed to be 16 which gives good data rate resolution with reasonable complexity. In order to increase the efficiency of AMC to exploit the microscopic channel fading, the packet duration is reduced to 2ms.

  The means of adaptation in HDSPA are the channel coding rate, modulation level, the number of multi-code employed as well as the transmit power allocated per code. To be able to exploit the microscopic fading, the AMC and the multi-code of the HSDPA must be able to cover a wide dynamic range. In the modulation design, the HSDPA incorporated 16QAM in addition to the QPSK modulation in order to increase the peak data rate. The combination of rate $3/4$ encoding and 16QAM gives a peak bit rate of 712 kbps per code channel (SF=16). On the other hand, higher robustness is available with rate $1/4$ and QPSK modulation giving a peak bit rate of 119 kbps per code channel. Given very good channel condition, a single user can simultaneously receive up to 16 code channels with an aggregate bit rate of 10.8Mbps. Table 8.4 illustrates various combinations of channel encoding and modulation level as well as the corresponding data rate in HDSPA.

[8]Moreover, due to the existence of buffer for the bursty data sources and a relaxed delay constraint relative to the circuit switched voice users, the optimal transmission strategy should be *water-filling* instead of *equalizing* the fading channel. Hence, the removal of fast power control does not incur any fundamental performance penalty in information theoretical sense.

Table 8.4 Various data rates available in DS-DSCH of HSDPA using different modulation, channel coding and multi-code allocations.

TFRC	Data Rate (1 code)	Data Rate (5 codes)	Data Rate (15 codes)
QPSK, Rate 1/4	120 kbps	600 kbps	1.8 Mbps
QPSK, Rate 1/2	240 kbps	1.2 Mbps	3.6 Mbps
QPSK, Rate 3/4	360 kbps	1.8 Mbps	5.3 Mbps
16QAM, Rate 1/2	477 kbps	2.4 Mbps	7.2 Mbps
16QAM, Rate 3/4	712 kbps	3.6 Mbps	10.8 Mbps

The dynamic range of the AMC for single code and multi-code systems is illustrated in Figure 8.21 The curve includes the gain from fast hybrid ARQ based on chase combining which significantly improves the throughput at low E_s/η_0. The dynamic range of E_s/η_0 variation is usually over 20dB due to the Rayleigh fading. Observe that the multi-code operation allows a smaller granularity of the AMC and resulting in a smoother throughput curve. Furthermore, the multi-code operation allows the instantaneous throughput to vary over a wider dynamic range of E_s/η_0 (around 32 dB).

- **Fast Hybrid ARQ:**

  The hybrid ARQ (HARQ) protocol adopted by the HSDPA system is a stop-and-wait protocol. The transmitter persistently transmits the current packet until it has been successfully received by the UE (indicated through the acknowledgement (ACK)). In order to avoid resource idling (and therefore a waste) while waiting for the ACK, a maximum of 8 parallel HARQ processes may be setup for the UE so that different HARQ processes transmit in different TTI (2ms each). The control of HARQ is located at Node B over the new MAC-hs layer so that the storage of the unacknowledged data packets and the subsequent scheduling of retransmissions does not involve the RNC. Hence, Iub signaling (between Node B and RNC) is avoided and the resulting retransmission delay in HSDPA is much lower than the conventional retransmission in the RNC (RLC layer). Typical retransmission delay in HSDPA is around 8 – 12ms.

  The HSDPA concept supports two types of retransmission strategies, namely the *chase combining* (CC) and the *incremental redundancy* (IR). The basic idea of CC is to transmit an identical version of the erroneously detected data packet and then for the decoder to combine the soft information of the received copies weighted by the SNR prior to decoding. In the IR scheme, additional redundancy information is incrementally transmitted if the decoding fails in the first attempt. In general, the IR scheme requires a lower SNR relative to

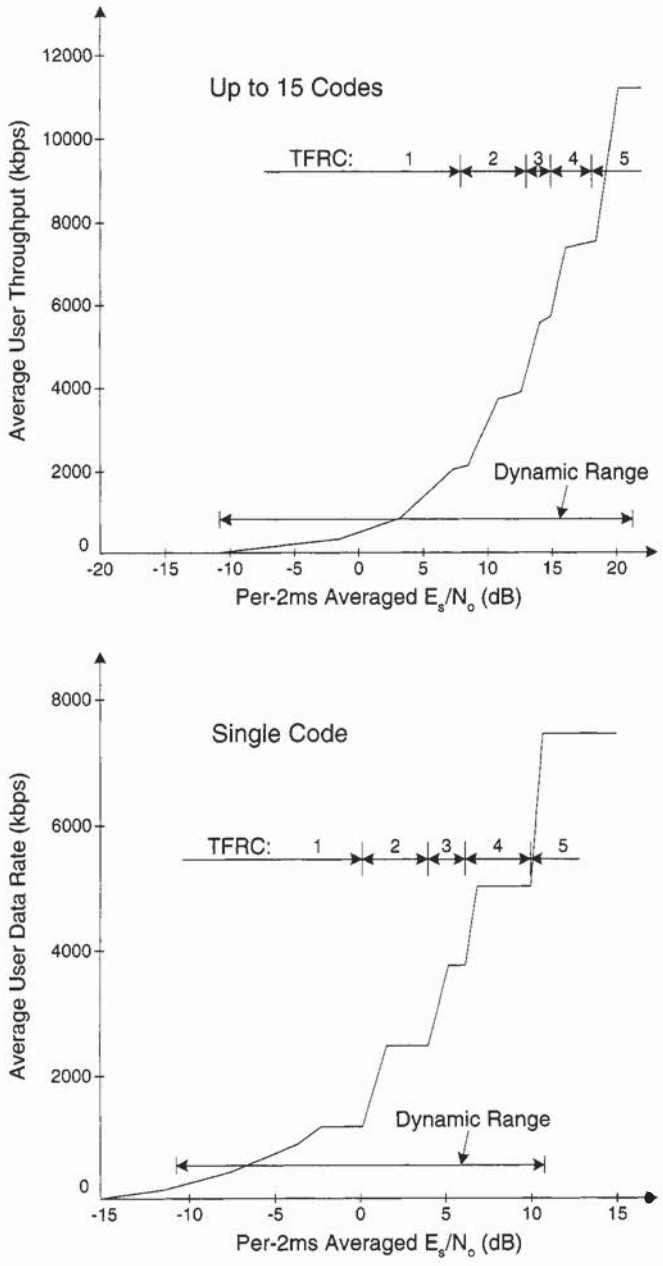


Figure 8.21 Dynamic range of AMC for single code and multi code HSDPA systems.

the CC scheme at the same target BLER but the disadvantage of the IR scheme is a higher memory requirement in the UE. Hence, the possibility of using IR scheme is defined by the UE capability class.

- **Fast Scheduling:** While the AMC and multi-code transmission in HS-DSCH allow a smooth link level adaptation over a large dynamic range of SNR variation, fast scheduling is introduced in the MAC-hs residing in the Node B to facilitate multi-user selection diversity over the Rayleigh fading channels. The scheduling decisions as well as the packet buffers are located at the Node B and no Iub signaling (between Node B and RNC) is needed in HSDPA. Hence, the HS-DSCH can be assigned to one packet data user in 2ms TTI resolution. This is shorter than the coherence time of most pedestrian users in the system and therefore, the microscopic fading remains quasi-static within a packet transmission.

  In HSDPA, there are three new channels introduced in the physical layer as outlined below.

  - **High-Speed Downlink Shared Channel (HS-DSCH)**: This is the channel carrying user data over the downlink and thus, it is the major resource for the scheduling algorithm to manipulate.
  - **High-Speed Shared Control Channel (HS-SCCH)**: This is the channel carrying the physical layer control information over the downlink for several key purposes including decoding, combining, and retransmission.
  - **High-Speed Dedicated Physical Control Channel (HS-DPCCH)**: This is an uplink channel carrying important control information such as the ARQ acknowledgments (ACK) and the downlink channel quality indicator (CQI).

  Similar to the DSCH in UMTS Rel 99, there is always a low data rate downlink DCH associated with the HS-DSCH. While the HS-DSCH is shared by a number of UEs, the associated downlink DCH is dedicated per UE. Besides the downlink DCH, there is also an uplink HS-DPCCH (per UE) associated with the HS-DSCH in HSDPA. Figure 8.22 illustrates the multi-user fast scheduling concept in HSDPA. The Node B tracks the channel quality of each active packet data users in the downlink direction by monitoring the transmit power of the downlink DCH associated with the HS-DSCH. In addition, the UE can also be requested to report a CQI periodically over the uplink HS-DPCCH. The CQI is an indication of the adaptation mode and number of codes currently supported by the UE based on the current channel condition. The feedback cycle of the CQI can be set as a network parameter adjustable in 2ms steps. To facilitate HARQ, the UE is also required to send an ACK/NAK response on the HS-DPCCH.

In Node B, the CSI of a UE is estimated based on the power level of the downlink DCH, the ACK/NACK ratio as well as the CQI. Depending on the scheduling algorithm, the Node B then schedules data transmission to UE on the HS-DSCH. Prior to sending data to the HS-DSCH, the Node B sends a detail message to the active packet data users via the associated HS-SCCH. This message describes the employed AMC mode, the multi-code set as well as the HARQ process of the HS-DSCH transmission and it is transmitted 2 slots in advance of the HS-DSCH as illustrated in Figure 8.23

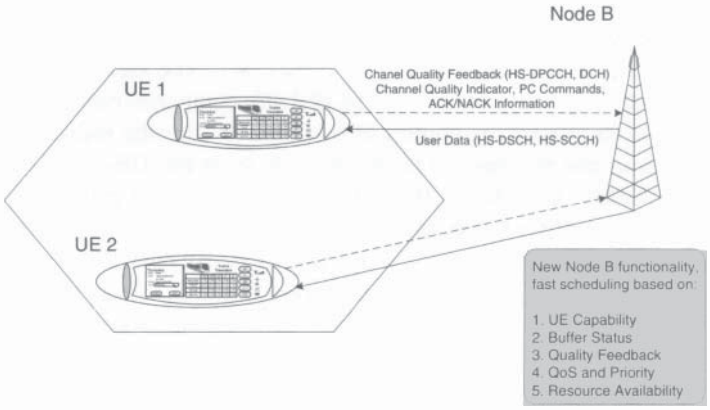


Figure 8.22 A schematic diagram illustrating the scheduling mechanism (HSDPA) in the UMTS Rel 5.

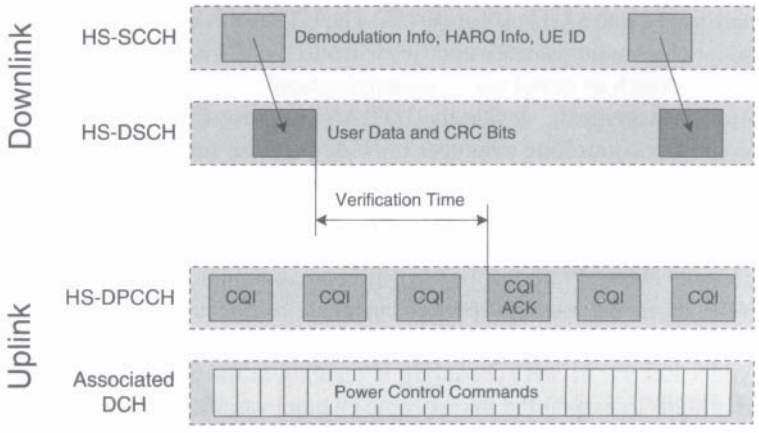


Figure 8.23 Inter-channel operation of HSDPA scheduling.

8.4.2 Continued Evolution

HSDPA provides a significant cell capacity gain for packet data traffic in UMTS and is thus an important part of the continuous 3G evolution. Since the HSDPA systems offers improved code efficiency and dynamic range in the user data rate, it can utilize the improvements in the detector performance such as channel equalizations, multi-user and multi-code interference cancellations as well as advanced MIMO techniques. In addition, the introduction of fast scheduling in HSDPA paves the way for cross-layer scheduling design incorporating the time, frequency and spatial dimensions on a multi-user perspective.

Naturally, the "high speed" idea can be extended to the uplink also. That is the so-called HSUPA (High Speed Uplink Packet Access). The idea is that except for scheduling, all the actions are reversed in HSUPA compared with HSDPA. That is, scheduling is still under the control of the Node B. But the measurements and combining actions are also moved to the Node B from the UEs. However, there is one major obstacle in realizing HSUPA—the power control problem. Over the downlink, the power source is centralized at the Node B. Thus, the power level can be dynamically adjusted solely at the Node B's discretion. By contrast, over the uplink, there are numerous heterogeneous power sources—the UEs. Thus, the interference induced is not easily controllable at the receiver located at the Node B. Indeed, uplink power control cannot be abandoned as in the downlink due to the near-far problem. Much more research is needed in this area.

8.5 SUMMARY

In this chapter, we have elaborated the application packet scheduling using UMTS as an example. We have reviewed a general architecture of UMTS. A UMTS network can be partitioned into a UTRAN and a CN. The UTRAN is handling radio dependent functions (such as radio resource control) whereas the CN is handling radio independent functions (such as switching of connections and packets, mobility management and session management). Inside the UTRAN, we have the RNC acting as a base station controller controlling a number of Node Bs (base stations).

In the UTRAN, channels are partitioned into three layers, namely the logical channels (or radio bearer), transport channels and the physical channels. The logical channels form a common interface for both the user plane and control plane protocols. Various logical channels are mapped to different combinations of transport channels in the MAC layer. Various transport channels are further multiplexed into different combinations of physical channels in the PHY layer.

In UMTS Rel 99, packet scheduling is residing at the S-RNC (RRC layer) and therefore, it fails to exploit the microscopic channel variation in the fading channels due to the inherit delay in the Iub signaling (over 500ms). Packet data users can be served on DCH based on code scheduling or DSCH based on time scheduling. In the former case, the physical radio resource assigned to the active UE are dynamically

adjusted based on radio bearer reconfiguration as triggered by traffic volume changes in the RRC layer. It suffers from a larger channel setup overhead and is suitable only for large volume of bursty data. In the latter case, a number of packet data users are time sharing the DSCH and it incurs a lower setup overhead and delay. In both cases, the peak data rate of the packet user is around 2Mbps.

In UMTS Rel 5, HSDPA is introduced to significantly improve the spectral efficiency of packet data users. There are three enabling technologies behind HSDPA, namely the AMC, the fast HARQ and the fast scheduling. Three new channels (HS-DSCH, HS-SCCH, HS-DPCCH) are created to facilitate an adaptive link with large dynamic range over the supported bit rates. In addition, a new MAC layer (MAC-hs) is introduced at the Node B and fast scheduling can be done at the Node B without going through the Iub signaling. This substantially reduces the scheduling latency and the scheduling is based on time division over the shared HS-DSCH. The scheduling period is 2ms and this is sufficient to exploit the microscopic fading for pesdestrian users. As a result of the fast scheduling and the adaptive physical layer, multi-user selection diversity can be effectively exploited and the HSDPA could deliver a much higher spectral efficiency relative to UMTS Rel 99.

PROBLEMS

8.1 UMTS system allows QoS differentiations in the air interface [True/False].

8.2 In UMTS Rel 99, the system capacity of high bit rate packet data users mixed with regular voice users will be limited by the number of orthogonal codes [True/False].

8.3 The outerloop power control is done at the Node-B of UMTS systems [True/False].

8.4 In UMTS, one way to increase the physical bit rate is to assign multiple codes to the same user [True/False].

8.5 In UMTS Rel-99, the RNC is responsible for dynamically controlling the physical resources allocated to the data bearer [True/False].

8.6 Since the RLC layer in UMTS mobile has packet retransmission capability in case a packet is corrupted, there is no need for the TCP layer to do any TCP packet retransmission because all the potential error packets are taken care by the lower layer (RLC layer) in the protocol stack [True/False].

This Page Intentionally Left Blank

PART III

SHORT-RANGE WIRELESS TECHNOLOGIES

This Page Intentionally Left Blank

CHAPTER 9

IEEE 802.11X WLAN STANDARDS

The ether of space is nothing but nominative of the verb to undulate.

—Author unidentified

9.1 INTRODUCTION

Internet becomes truly pervasive when low-cost wireless technologies are widely available. Specifically, short-range wireless communications have by and large enabled people to get network access anytime anywhere. In this chapter we talk about one of the two most important short-range wireless technologies.

Interoperability:

Wireless LAN (WLAN) has been in widespread deployment for several years and the first IEEE 802.11 standard was introduced in 1997. The main reason for developing this standard is to provide interoperability in two dimensions: interoperability between wireless LANs and existing wired LANs, for example, IEEE 802.3 Ethernet, and also interoperability between wireless devices from different vendors.

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

In order to facilitate the cooperation between wired and wireless LANs, the IEEE 802.11 architecture has been developed with the concept of Basic Service Set (BSS) and how an IEEE 802.11 device communicates with other networking technologies through the Access Point (AP) and Distribution System (DS). By employing the DS concept, a WLAN not only can interconnect with other WLANs' BSS, but also can integrate with existing wired LANs. The IEEE 802.11 architecture is discussed in detail in Section 9.3 below.

On the other hand, before the IEEE 802.11 standard was created, different vendors developed their own wireless LAN products based on their own ideas, and this lead to a serious problem of incompatibility between wireless devices from different vendors. In order to facilitate the interoperability issue, the IEEE 802.11 working group has been set up to provide generally acceptable standards and specifications on Medium Access Control (MAC) mechanisms and Physical (PHY) links for wireless devices from different vendors, so as to share the wireless medium in a fair manner. These issues are discussed in detail in Section 9.4 and Section 9.5.

Performance:

In view of the demand for different performance requirements from the wireless LANs, different task groups have been developed including the famous ones like A, B, G, E. The standard provided by Task Group A, that is IEEE 802.11a, can provide theoretical capacity up to 54 Mbps with 24 orthogonal channels. This is a tempting capacity comparable to wired networks. However, it is not as popular as that of IEEE 802.11b. The main reasons are discussed in detail in Section 9.5. On the other hand, IEEE 802.11a/b/g has been developed to focus on data exchange in the wireless medium. However, only limited QoS requirements have been considered. In order to solve this problem, another task group IEEE 802.11e has been developed to highlight the issue of multimedia delivery in wireless LANs, based on existing IEEE 802.11a/b/g physical layers. This topic is discussed in Section 9.6.

9.2 DESIGN GOALS

There are several objectives in designing IEEE 802.11, which are summarized in Table 9.1 below.

Besides the objectives listed in Table 9.1, there may be some other goals which are under development, such as providing a larger throughput capacity, improving the coverage area and providing wireline comparable functions, etc. Indeed, as user applications' demands keep increasing, future design goals of WLAN extensions would be even more ambitious.

9.3 IEEE 802 ARCHITECTURE

The development of WLAN is assisted by the IEEE 802.11 working group which is responsible for providing specifications and standards on WLAN. The IEEE 802.11

Table 9.1 The design goals of WLAN.

Objective	Description
Seamless Wireline LAN Extension	Wireline LAN has been developed and used for quite a long time with many innovative inventions involved. People prefer new technology development to at least exploit some of those innovative ideas and provide backward compatibility. In order to do so, the design of IEEE 802.11 architecture has put into consideration of interoperability between different BSSs and wireline networks. At the same time, only minimum amount of changes are applied to WLAN's MAC and PHY layers, with upper layers unaffected.
Wireless and Mobile Capabilities	WLAN provides the stations with the capability to connect to an existing network through the wireless medium. This in return provides users with great convenience, as users can conveniently move with their wireless devices without the consideration of the limitation and location of the corresponding wired access point.
Low Cost	There are some considerations in achieving low cost: • The frequency bands used by IEEE 802.11x are all by and large unlicensed bands, so they can be used without charge as long as frequency band usage and power regulation has been followed, and the issue of collision with other wireless technologies is addressed in the design. • The popularity of WLAN devices in return leads to low cost mass production. Nowadays, WLAN devices can be obtained with reasonable and generally affordable price. • The interoperability with wireline LAN is achievable with the help of an access point to form BSS. No new device is required to integrate with existing infrastructure except an access point.

architecture has been suggested in the MAC/PHY specification. There are several components involved and some of the important components are listed in Table 9.2 [323, 6].

The basic building block of the architecture is Basic Service Set (BSS). A BSS may be composed of a single or multiple station(s). Since the wireless medium within a BBS is shared by all the stations, it is necessary for them to use the same MAC protocol and to be controlled by a single coordination function. There are two main types of coordination functions suggested by the IEEE 802.11 working group—Distributed Coordination Function (DCF) and Point Coordination Function (PCF). These functions are discussed in detail in Section 9.4.

Table 9.2 The components of the IEEE 802.11 architecture.

Architecture Component	Description
Station (STA)	A device that has implemented standard IEEE 802.11 Media Access Control (MAC) and Physical (PHY) layers.
Access Point (AP)	A device that has implemented station's functionalities and provided an interface for other associated wireless stations to access a Distribution System.
Basic Service Set (BSS)	A group of stations managed by a single coordination function.
Distribution System (DS)	A system which connects multiple Basic Service Sets and/or integrates with other networking technologies, for example, IEEE 802.x LANs to form an Extended Service Set.
Portal	A logical location where the Media Access Control Sevice Data Unit (MSDU) enters the Distribution System (DS).
Extended Service Set (ESS)	A group of one or more BSSs and integrated LANs that shown to be a single BBS to Link Logical Control (LLC) Layer of any station inside any associated BBS.
MAC Protocol Data Unit (MPDU)	The data unit exchanged by two peer MAC users using the service of the PHY layer.
MAC Service Data Unit (MSDU)	The information delivered by MAC protocol data unit.

Sometimes it is preferable to have a larger wireless network. However, due to physical limitation of the wireless medium and limitation on PHY standard, every station can only transmit data within a certain range. This causes a limitation on coverage area to a BSS. In view of this, the IEEE 802.11 working group suggested the concept of Distribution System (DS). DS is a backbone system in which interconnection between multiple BSSs are facilitated by connecting to it through Access Points (AP). An Access Point is a station which communicates with other stations within the same BBS via the wireless medium. At the same time, it communicates with other entities within the DS. In fact, though the IEEE 802.11 working group has defined the concept of DS, it has not limited the implementation of DS, so that DS can be a switch, another wireless or wired network.

In order to allow other networking technologies' network, for example, Ethernet, to integrate with the IEEE 802.11 architecture, a portal is used to connect to DS. Similar to AP, a portal can be a switch, bridge or hub which acts as an interface between the network and DS.

The Extended Service Set (ESS) consists of DS and BSSs and/or other networking technologies' networks. By definition, ESS refers to the integration of components

listed Table 9.1 and appears to be a single BSS to Logical Link Control (LLC) Layer of devices within the network. In other words, it appears to be a single LAN. With this concept, a station's movement from one BSS to another BSS is possible to be transparent to LLC.

The concepts discussed above are illustrated in Figure 9.1. Here, each station is connected to a single BSS. While BSSs are connected to DS through AP, at the same time, an IEEE 802.x LAN is integrated with BSSs by connecting to DS through a portal. This is one of the simplest configurations.

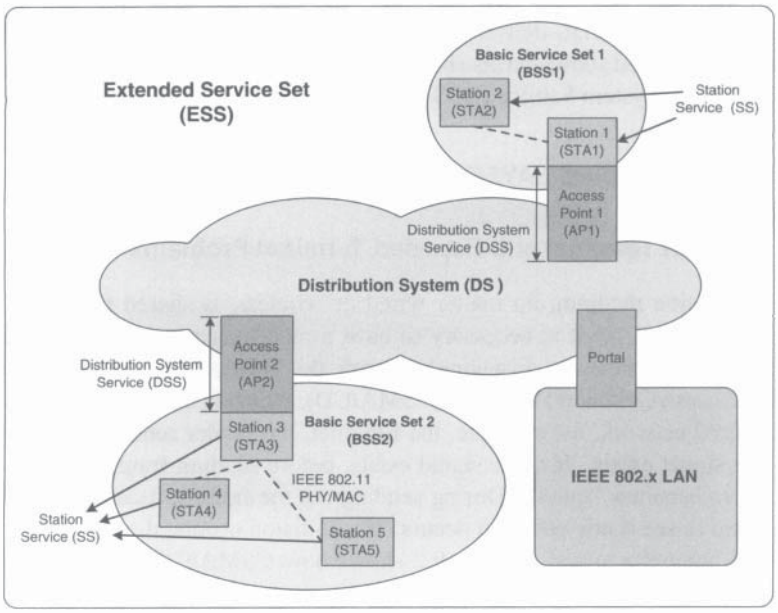


Figure 9.1 IEEE 802.11 architecture.

IEEE 802.11 Service

IEEE 802.11 defines nine services [323, 6]. Six of them are used to support MSDU movement between different stations within ESS and three of them are used for controlling the network access and privacy.

The nine services are:

1. Authentication
2. Association
3. Deauthentication
4. Deassociation

5. Distribution
6. Integration
7. Privacy
8. Reassociation
9. MSDU delivery

Obviously, from the list above, authentication, deauthentication and privacy are three services used to control network access and confidentiality. All other services are used to support MSDU delivery.

The services listed above can also be divided into two groups—Station Service (SS) and Distribution System Service (DSS). Their functions are summarized in Table 9.3

9.4 IEEE 802.11 MAC LAYER

9.4.1 Hidden Terminal and Exposed Terminal Problems

The transmission medium, no matter wired or wireless, is shared by all the users within the network, so it is necessary to have a mechanism to avoid simultaneous sending of data by the users. In a wired network, this is usually done by Carrier Sense Multiple Access/Collision Detection (CSMA/CD), which is briefly reviewed below.

In a wired network, for example, the Ethernet, the sender senses the medium to see if any signal exists. If some signal exists, before sending frames, it waits until the medium becomes "quiet". During sending out the frames, it continues auditing the medium to see if any collision occurs. If a collision occurs, the sender stops and send out a jamming signal. Figure 9.2 shows how CSMA/CD works in Ethernet. We can see that Computers A and C sent a frame to Computer B simultaneously since they both found the medium is free. Then a collision occurs near Computer B. The collision can then be detected by all the computers inside the network. The two computers then wait a certain amount of time before trying to send out their frame again.

As can be seen in Figure 9.2, the collision is detected by the sender instead of receiver. However, we are only interested in whether the receiver can receive a "clean" frame. In a wired network, this is not a problem at all as the signal level of a "clean" frame is similar everywhere within the network including the sender side. However, this is not the case in a wireless network. Figure 9.3 shows the problem of CSMA/CD in a wireless environment. Here, Stations A and C are within the transmission range of each other, but they are near the end of each other's transmission range. Stations A and C are sending a frame to Station B simultaneously. However, since they are sending a frame at that moment, the signal level of their own frame is much higher than that of their opponent. In this case, it is very difficult for them to detect whether there is a collision.

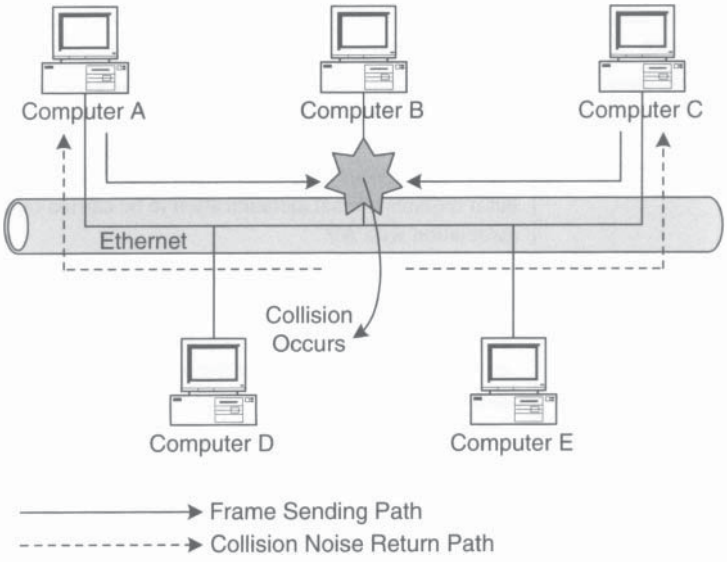


Figure 9.2 Wired network CSMA/CD illustration.

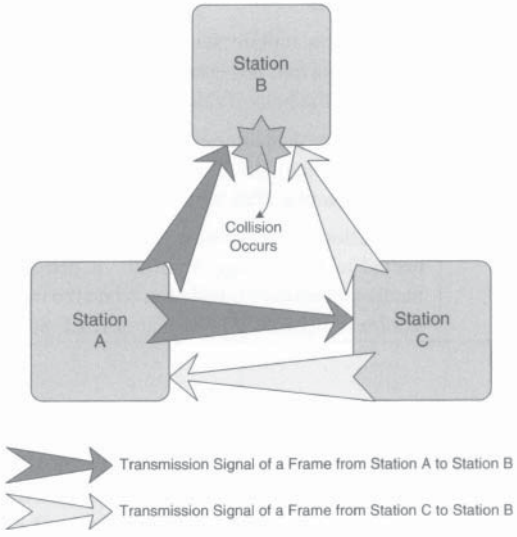


Figure 9.3 Wireless network CSMA/CD problem.

Table 9.3 IEEE 802.11 services.

Type	Service	Description
SS	Authentication	Unlike wired networks, the transmission medium used in wireless networks is open and shared. This service is required for stations to establish the identities of their communication peers within BSS. IEEE 802.11 does not fix the scheme used, but it requires mutual authentication to be carried out before association with AP.
	Deauthentication	This service is activated whenever any existing authentication is about to be terminated.
	Privacy	With reasons similar to those of authentication, this service is provided for preventing the message to be read by recipients other than the intended one.
	MSDU delivery	This service is used to deliver the data unit from one station to another.
DSS	Association	Before a station is allowed to send and receive frames, it has to associate with certain AP. In this case, either station and AP can determine the address and identity of each other. AP can then supply the associated station information to DS, so DS can route frames within ESS.
	Deassociation	Either a station or AP within the same BSS sends out a notification to notify others about the termination of existing association.
	Reassociation	Whenever a mobile station moves out of the transmission range of current AP, established association is then transferable with the help of this service.
	Distribution	This is the primary service used by stations to exchange frames through DS. That is, the frames are routed from one BSS to another BSS within ESS.
	Integration	Integration service allows the IEEE 802.x network to be connected to DS through a portal. It performs any necessary media and address translation when the messages are sent by station in IEEE 802.11 to that of IEEE 802.x or vice versa.

Other than the problem described previously, there are two main problems in the wireless environment—the hidden terminal and exposed terminal problems [313] that make schemes like CSMA/CD in a wired network fail and/or inefficient.

First of all, let us focus on the following scenario. In Figure 9.4, the transmission ranges of Stations A and C only cover up to Station B and so their signal cannot reach each other. This causes the so-called hidden terminal problem. Specifically, the problem is caused by the fact that Station A does not realize the existence of

station C and vice versa. In this case, when Station A senses the environment, it finds out that the environment is free, so it sends out a frame to station B. At this moment, Station C also wants to send out a frame to station B. However, as it cannot hear the transmission from station A, it also thinks that the environment is free and sends out its frame. This leads to a collision at Station B. However, the collision is not realized by both Stations A and C, and so they continue sending their frame until they finish sending.

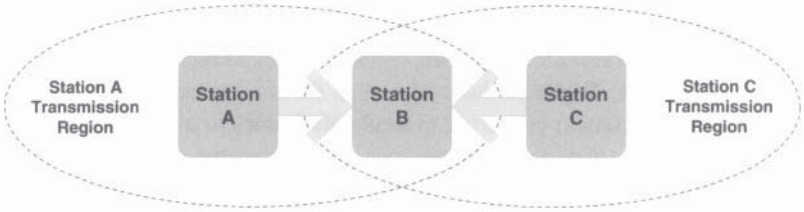


Figure 9.4 The hidden terminal problem.

Another issue is the exposed terminal problem. In order to explain this problem, let us focus on the scenario in Figure 9.5. We can see that Station B's transmission range covers both Stations A and C, but not Station D, while the transmission range of Station C covers both Stations B and D, but not Station A. In this case, Station B senses the environment and finds out that it is free, so it sends out the frame to Station A. However, at this moment, Station C also wants to send a frame to Station D. It senses that Station B is sending out a frame, so in order to avoid collision, it waits until Station B finishes sending. Indeed, we only care about the frame at receiver side instead of sender side. Thus, Station C, in fact, can also send out its frame simultaneously, without waiting for Station B to finish sending.

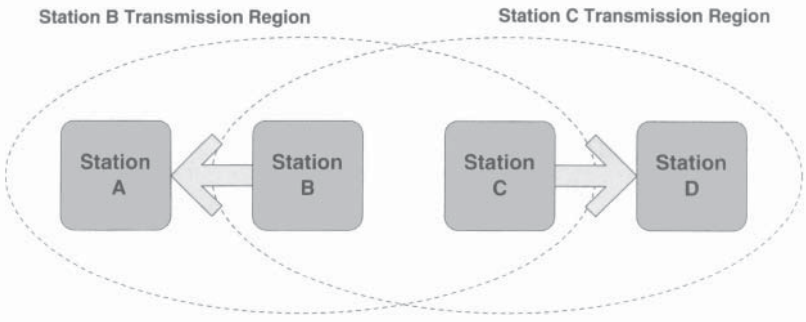


Figure 9.5 The exposed terminal problem.

9.4.2 Four-Frame Sequence

As explained in the previous section, there are hidden terminal problem and exposed terminal problem in using carrier sense mechanism in wireless environment. In order to solve the hidden terminal problem, a four-frame sequence is introduced. The sequence consists of Request-to-Send (RTS) frame, Clear-to-Send (CTS) frame, Data frame and Acknowledgement (ACK) frame.

Before sending out data frames, the sender sends out RTS after it finds that the medium is free. If the target recipient receives RTS, it sends out CTS in reply if it has not received any RTS/CTS from other stations before. When the sender receives CTS, it knows that the environment around the receiver is free to receive data frames, so it sends out data frames. While stations other than the sender receive RTS/CTS , they expect that there is a sender preparing to send data frames to a recipient within their transmission range, so they wait until the receiver receives the data frames and sends out ACK which is used as a signal of finishing sending, or until Network Allocation Vector (NAV) is deactivated. NAV is a flag which is used to indicate stations that the medium is busy, so that the stations activate it when they receive RTS/CTS. During the period when NAV is activated, the stations ignore other RTS/CTS and do not send out their own frames.

After explaining the four-frame sequence, the details on how this sequence can be used to solve the hidden terminal problem is explained in Figure 9.6. Here, Stations A and C are out of the transmission range of each other, i.e., Stations A and C are hidden from each other. Only Station B can receive signal from both stations. The whole sequence is explained as follows:

1. Station A senses that the medium is free and sends out RTS to Station B.

2. Station B receives RTS and sends out CTS as reply to Station A. Station C also receives CTS, so it knows that Station A is going to send data to Station B and so it sets the NAV (CTS) value.

3. Station A starts sending out data frames.

4. Station C wants to send data, but it has received CTS from Station B, so it waits until receiving ACK.

5. Station B receives data frames and sends out ACK, both Stations A and C receive it.

6. Station C sends out RTS now, as it knows that Station A has finished sending.

7. The process of RTS-CTS-DATA-ACK sequence repeats once again.

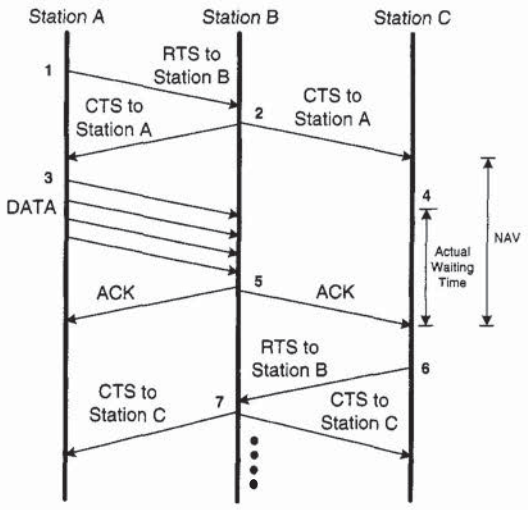


Figure 9.6 RTS/CTS solution to the hidden terminal problem.

9.4.3 DCF and PCF

To introduce the concepts of DCF (distributed coordination function) and PCF (point coordination function), let us consider the MAC architecture of IEEE 802.11. Figure 9.7 shows that there are two medium access coordination mechanisms in IEEE 802.11 MAC. They are DCF and PCF [323, 6]. DCF only provides contention service. By using DCF, stations are required to compete for the wireless medium. However, as shown in Figure 9.7, PCF is built on top of DCF, it uses the services of DCF in order to provide the functions of contention-free service for upper layers.

As discussed in the previous section, collision detection is difficult to be implemented in the wireless medium. In DCF, Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) is used instead. By using this algorithm, first of all, a station senses the medium if it wants to send out an MAC frame. It sends out the frame only if it discovers that the medium is free; otherwise, it waits until the medium becomes free again.

Even though CSMA/CA algorithm is used in DCF, there is still a high probability that multiple stations sense that the medium is free and send out their frame simultaneously, especially at the moment immediately after a sender finishes sending a frame and the medium becomes free again. In order to avoid this, various types of Interframe Space (IFS) are introduced together with a random backoff mechanism. IFS is the time between two frames, with variable length depending on the types. Different priority is then given to different services or stations. Generally speaking, the shorter the length of IFS, the higher the chance that a station can control the medium. Table 9.4 gives the description of various IFS.

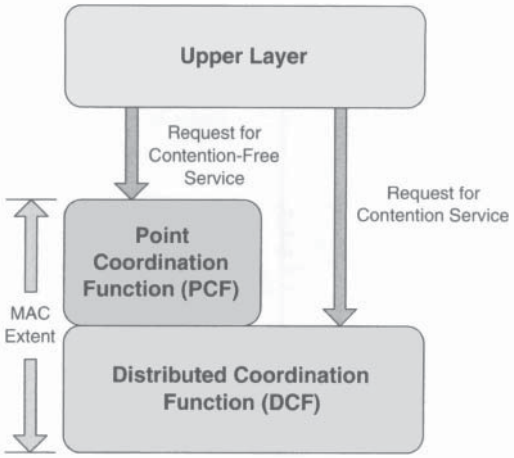


Figure 9.7 IEEE 802.11 MAC architecture.

Table 9.4 The various types of IFS.

Type	Description
Short IFS (SIFS)	This is the shortest length of IFS. It is the amount of time waited by stations in reply of RTS, CTS and data frame. Due to its short duration, using it essentially prevents other stations waiting for the medium access from using the medium, so the whole sequence of data frame transmission can be finished.
Point Coordination Function IFS (PIFS)	This is the second shortest length of IFS. In PCF, a centralized controller continuously polls the stations for traffics during contention free period. PIFS is the time waited before the start of contention free period and also the time wait for the response of polling.
Distributed Coordination Function IFS (DIFS)	This is the longest length of IFS. It is the minimum amount of time waited by stations after the medium becomes free in DCF.

Other than using various types of IFS, in order to avoid collisions, random backoff mechanism is also introduced in DCF. When the stations wait for DIFS after the medium becomes free, they enter backoff time. They continue waiting for a certain amount of time based on the slot randomly chosen. Initially, there are eight slots to choose from. However, if a collision occurs, the number of slots is then doubled (exponentially, and up to 256 slots) after each retransmission. This method is called binary exponential backoff. As long as one of the stations successfully sends out the frame during backoff time, the remaining waiting time slots chosen by all other stations

are kept so that their backoff waiting process can be resumed in next contention window.

Figure 9.8 demonstrates a basic access method of DCF, illustrating the operation of various IFS types and random backoff mechanism.

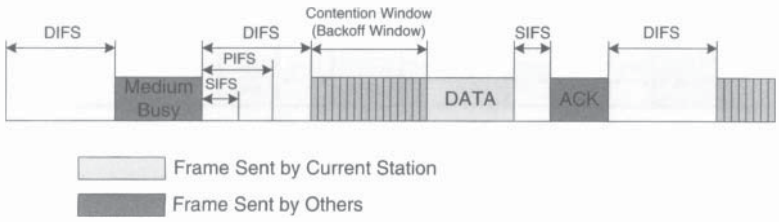


Figure 9.8 DCF basic access method.

Since DCF is contention-based in nature, it is only suitable for asynchronous traffic which does not mandate any QoS requirements. However, in order to support time-sensitive traffics, the scheme must provide some contention-free period dedicated for that kind of traffics. This leads to the concept of PCF. As indicated in the MAC architecture shown in Figure 9.7, PCF is built on top of DCF. It is possible for it to support both asynchronous and time-sensitive traffics simultaneously with the use of alternating contention-free period and contention period structure. This is indicated in Figure 9.9.

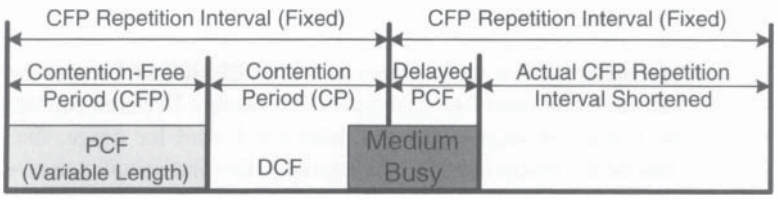


Figure 9.9 Alternating contention-free period (CFP) and contention period (CP).

As shown in Figure 9.9, there is a repeated fixed period of time for a CFP (contention-free period) to occur. This repeated fixed period of time is divided into two subperiods, they are CFP and CP. During CFP, the central point coordinator (usually the AP) polls each of the stations within the BSS. The polling structure of CFP is shown in Figure 9.10. We can see that before entering CFP, the point coordinator waits for PIFS and accesses the medium afterwards and this indicates the beginning of CFP. Afterwards, the point coordinator polls Station 1 as shown in the figure. If Station 1 has frames to send, it replies with a frame containing the DATA and ACK after waiting for SIFS. Then the point coordinator polls the next station, Station 2, as

illustrated in Figure 9.10. Since Station 2 does not have any frame to send and thus does not reply to the polling messages. The coordinator then waits for PIFS and poll the next stations or send out CF-END frame which indicates the end of CFP.

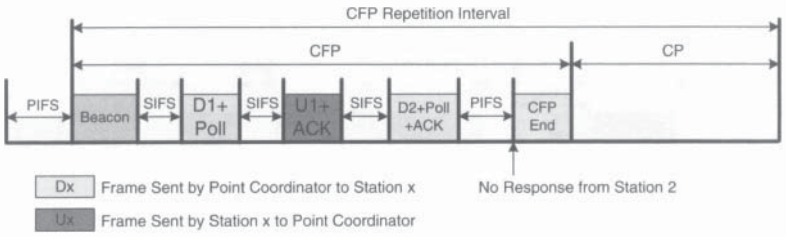


Figure 9.10 CFP structure.

After CFP, it comes to CP. During this period, each station acts just like that in DCF. They compete for the medium with the help of IFS and random backoff as described previously. However, as indicated in Figure 9.9, there may be chances that the medium is busy during the start of CFP. In this case, the point coordinator needs to wait until the medium becomes free and starts the CFP afterwards. In this case, the actual CFP repeat period for this time is shortened to meet the original schedule.

9.5 IEEE 802.11 PHYSICAL LAYERS

9.5.1 Physical Link Features

The transmission capabilities of the physical link of IEEE 802.11 keep on improving. With the help of techniques like Orthogonal Frequency Division Multiplexing (OFDM) and the release of more and more unlicensed band for usage, the bandwidth capacity has been greatly increased. Originally, the first specification was released in 1997, in which three physical link types were proposed—Direct Sequence Spread Spectrum (DSSS), Frequency Hopping Spread Spectrum (FHSS), and infrared [323, 128].

DSSS is a technology which divides the 2.4 GHz ISM frequency band into channels with spacing of 5 MHz in center frequency. The center frequency of channel 1 is located at 2.412GHz and so on. Each channel then takes up 22 MHz bandwidth. There are altogether 14 channels defined. However, due to spectrum regulation in different countries, there are at most 13 channels available. The system implementing DSSS divides a data bit into 11 chips according to Barker Sequence as shown in Figure 9.11. By doing so, the original narrow band signal is spread into a relatively wide band signal. The receiver side receiving the wideband signal can use the same sequence to decode the original narrow band signal due to good autocorrelation property of Barker Sequence.

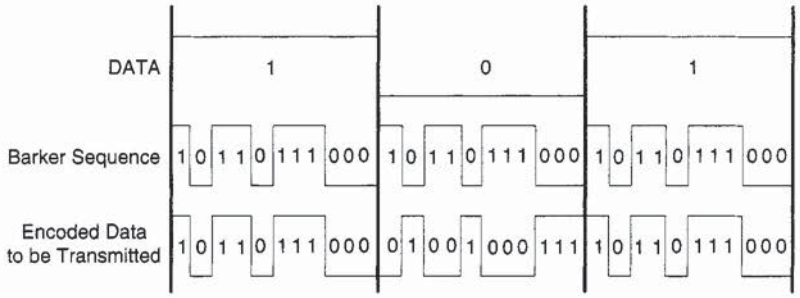


Figure 9.11 DSSS mechanism.

In the original IEEE 802.11, only two data rates are provided, they are 1 Mbps and 2 Mbps. They are using DBPSK and DQPSK as modulation scheme respectively. DBPSK is able to encode one bit in one symbol, while DQPSK can encode two bits in one symbol.

FHSS is a technology which divides the 2.4 GHz ISM frequency band into frequency slots of 1 MHz each. The center frequency of a frequency slot in return determines the channel number of the slot. Channel 1 corresponds to the central frequency of 2.401 GHz and so on. The total number of channels depends on countries, ranging from 23 channels in Japan up to 78 channels in the United States and most European countries. By dividing the frequency band into channels, the stations implementing FHSS keep on changing channels where they do not stay more than 0.4 second in each hop according to US regulations. Which channels are to be hopped next is determined by some predefined hopping sequence set with regulation on minimum hopping distance depending on countries. Figure 9.12 demonstrates a sample FHSS system with two stations, without consideration of hopping distance.

As mentioned before, the two data rates defined in IEEE 802.11 are 1 Mbps and 2 Mbps. In the FHSS case, with 1 MHz occupied frequency and two-level GFSK, 1 Mbps data rate can be obtained. By using two-level GFSK, only one bit is encoded into one symbol, so the data rate supported is 1 Mbps. Similar to the case of 1 Mbps, for the date rate of 2 Mbps, four-level GFSK modulation is used, with two bits encoded into one symbol.

Infrared can be divided into two main categories—line of sight infrared and diffused infrared. IEEE 802.11 defines diffused infrared in which a device is not required to be strictly within line of sight of each other. The infrared light is reflected by walls, ceiling, etc. IEEE 802.11 infrared defines two data rates—1 Mbps and 2 Mbps. For 1 Mbps, four data bits are mapped to a predefined group of 16-bit symbol. This modulation scheme is called 16 Pulse Position Modulation (PPM). For the case of 2 Mbps, two data bits are mapped to a predefined group of four bits symbol and it is called 4 PPM.

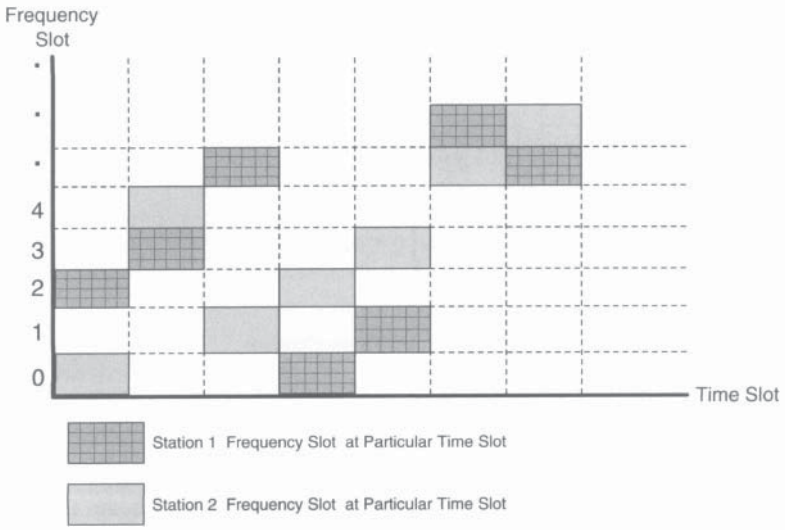


Figure 9.12 Orthogonal FHSS hopping sequence without consideration of hopping distance.

Orthogonal Frequency Division Multiplexing (OFDM) is not defined in original IEEE 802.11 PHY. However, it is proved to be effective in wireless medium usage and so it is included in IEEE 802.11a/g PHY. It is similar to Discrete Multi Tone (DMT) used in the Digital Subscriber Line (DSL). The frequency channels in the frequency band are divided into channels of 5 MHz spacing in center frequency, while OFDM occupies four channels with a total of 20 MHz. Within this 20 MHz, it is further divided into 52 smaller frequency slots with 0.3125 MHz apart and are called subcarriers. Within these subcarriers, only 48 are used to transmit data, while the remaining four are used for control. In this case, data bits from a single source is possible to be sent out through all of the subcarriers in parallel and more advanced modulation techniques like QAM, which can code more bits into a symbol, can be applied.

9.5.2 IEEE 802.11a

IEEE 802.11a specifies the usage of WLAN in 5GHz unlicensed band—Unlicensed National Information Infrastructure (UNII) [323, 7]. This band provides much larger amount of bandwidth for usage and is less crowded than that of 2.4 GHz ISM band. With the usage of OFDM, there are 13 (original) + 11 (currently new) orthogonal channels for usage according to United States Regulation. That is, 24 channels can be operated within the same region without interfering with each other. The original channels occupy the frequency from 5.15–5.35 GHz and 5.725–5.825 GHz, while the

new channels occupy the frequency from 5.470–5.725 GHz according to new FCC rules in November 2003.

As we have discussed earlier, OFDM and wide bandwidth channels allow more complicated wireline modulation techniques to be used. They support modulations like BPSK, QPSK, 16-QAM, 64-QAM. This in return provides a higher throughput capacity. With the help of OFDM and different modulation schemes supported, data rates ranging from 6 Mbps to 54 Mbps are supported

9.5.3 IEEE 802.11b

IEEE 802.11b was proposed in 1999 together with IEEE 802.11a [323, 8]. It uses the 2.4 GHz ISM band. There are several design objectives of IEEE 802.11b that make it more popular than IEEE 802.11a. One of the main reasons is that IEEE 802.11b is using the 2.4 GHz ISM band which is the same as that defined in IEEE 802.11. Together with the support of low data rate DSSS (1–2 Mbps), backward compatibility is provided. This facilitates the interoperability of IEEE 802.11b and existing WLAN products on the market. At the same time, with relatively simpler modulation techniques and mass production, IEEE 802.11b devices are relatively cheaper.

IEEE 802.11b uses the 2.4 GHz ISM band and has three orthogonal channels. However, in order to provide data rates larger than 2 Mbps, another chip spreading method is used. In original DSSS, 11 bits Barker Sequence is used to spread the data bit into 11-bit transmitted symbol, which is then recovered at the receiver side. Consequently, 1M symbols are transmitted in a second. However, with the usage of Complementary Code Keying (CCK), an eight-bit sequence is used to encode four bits or eight bits data. In this case, spreading effect can be achieved together with multi-bits data encoding that results in a higher data rate. Thus, by using CCK, 1.375M symbols are transmitted in a second which results in a data rate of 5.5 Mbps with the use of 4-bit encoded symbol and even up to 11 Mbps with the use of 8-bit encoded symbol.

9.5.4 IEEE 802.11g

Although IEEE 802.11b is more popular than IEEE 802.11a, there is a gap in throughput capacity between these two standards. People enjoy better interoperability brought by IEEE 802.11b on one hand, while they would like to have a larger throughput capacity on the other hand. In order to get the best of both worlds, IEEE 802.11g is created to provide a higher throughput capacity in the 2.4 GHz ISM band [9].

To get a higher throughput capacity, OFDM is adopted in IEEE 802.11g. However, to provide backward compatibility to IEEE 802.11b devices, DSSS is also implemented. In order to use OFDM with the compatibility of IEEE 802.11b, another mode of operation is introduced in IEEE 802.11g, that is DSSS-OFDM mode. In using this mode of operation, the headers and permeable are sent by using DSSS, so

that other IEEE 802.11b devices notice about the frame sent by IEEE 802.11g, while the data are sent by OFDM to achieve the highest possible data rate. With different mode of operations, IEEE 802.11g can support data rates ranging from 1 Mbps up to 54 Mbps with three coexisting orthogonal channels.

9.5.5 IEEE 802.11ag Dual Band

As indicated in the previous section, the main aim of IEEE 802.11g is to provide a throughput capacity higher than IEEE 802.11b by applying OFDM to the 2.4 GHz medium [71]. However, due to several limitations, using solely an IEEE 802.11g network may not be a good solution. It is because:

- IEEE 802.11g has three orthogonal channels only.
- The 2.4 GHz ISM band is crowded with many devices including microwave oven, cordless phone, Bluetooth, etc. Significant interferences are expected.
- IEEE 802.11g network's performance is severely affected in the presence of IEEE 802.11b stations due to the backward compatibility issue.

On the other hand, using solely an IEEE 802.11a network may not be a good solution either. Because:

- IEEE 802.11a does not provide backward compatibility to other IEEE 802.11x networks.
- Generally speaking, IEEE 802.11a devices are more expensive than IEEE 802.11b or IEEE 802.11g devices.
- IEEE 802.11a devices are less popular than that IEEE 802.11b.

Since both IEEE 802.11g and IEEE 802.11a have their own limitations and advantages, it is possible to combine both devices in order to get the best of both worlds. In view of this, dual-band IEEE 802.11/ag devices are implemented and are expected to dominate the WLAN market in near future. Generally speaking, since IEEE 802.11a has more orthogonal channels and aggregated bandwidth, and it is using 5 GHz UNII band which is less crowded, it is more suitable to be used to support QoS-based applications which generally require larger bandwidth and are more sensitive to delays. On the other hand, since IEEE 802.11g devices are more popular and are integrated into many devices which do not require or do not have strict QoS requirements, it is most suitable for some "best effort" applications, for example, normal data networking, some relatively low quality audio and video streaming etc.

Since WLAN becomes more and more popular, and the users of both QoS-based and "best effort" applications are expected to increase due to the integration of WLAN to new electronic devices, IEEE 802.11ag clearly becomes one of the best solutions to improve the performance of WLAN in future.

9.6 IEEE 802.11E FOR QOS PROVISIONING

The MAC and PHY of various IEEE 802.11 standards have been discussed in previous sections. In IEEE 802.11 MAC, only PCF has addressed some QoS-related issues. However, it does not provide a very clear picture on what and how QoS to be supported in WLAN. At the same time, there are two severe problems in PCF that make QoS support less effective. The two problems are unpredictable Beacon delay due to busy medium in CP and the unknown transmission time of the polled stations in CFP. In view of the problems, IEEE 802.11e standard is being developed based on the PHY of IEEE 802.11a/b/g. In the IEEE 802.11e draft standard, two main types of traffic classification are addressed—Traffic Category and Traffic Stream. These two traffic classification methods are described in Figure 9.13 [126].

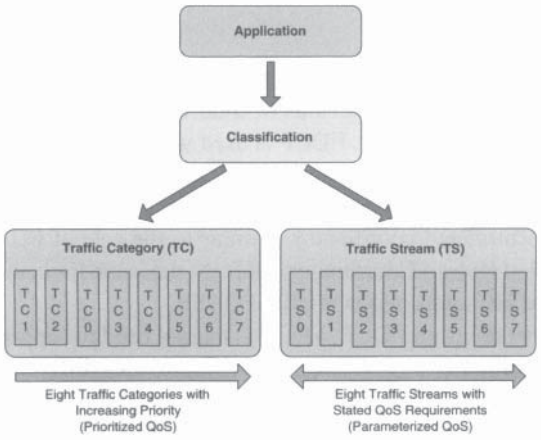


Figure 9.13 IEEE 802.11e traffic classification types.

In Traffic Category (TC), traffic flows are categorized into eight distinct classes with different priorities according to their traffic types predefined, from TC0–TC7. As shown in Table 9.5, a traffic priority is assigned based on the urgency to use the medium, the lowest priority is assigned to TC1 which refers to background traffic, while the highest priority is assigned to TC7 which refers to network control traffic. It is similar to the Differentiated Service suggested in wired networks.

For the case of Traffic Stream (TS), it is similar to that of Integrated Service in wired networks. In each station, eight traffic streams can be supported. Each stream is required to be characterized by the Traffic Specification (TSPEC). In TSPEC, information characterizing the traffic stream, including the packet size, arrival rate, maximum delay and delay variance should be specified.

In order to support these two traffic classifications, two new MAC modes of operation have been introduced in IEEE 802.11e. They are Enhanced Distributed Coordi-

Table 9.5 IEEE 802.11e traffic categories.

Traffic Category	Type	Priority
TC1	Background traffics	1 (lowest)
TC2	Spare traffics	2
TC0	Best effort data traffics	3
TC3	Excellent data traffics	4
TC4	Controlled load data traffics	5
TC5	Multimedia traffics with delay less than 100 msec	6
TC6	Multimedia traffics with delay less than 10 msec	7
TC7	Network control traffics	8 (highest)

nation Function (EDCF) and Hybrid Coordination Function (HCF). Similar to legacy IEEE 802.11 PCF, there are two periods of time within the CFP repetition interval. Similar to the role of DCF in PCF, EDCF is used within CP only and HCF is used in CFP. Unlike that of PCF, it is possible for HCF to be used within CP. A station that has implemented this function is called enhanced station. It may be possible for this station to act as centralized coordinator (similar to the role of PC in PCF) and this coordinator is called Hybrid Coordinator (HC). At the same time, the BSS with IEEE 802.11e supported HC and stations is called QBSS.

First of all, let us focus on EDCF. EDCF is mainly used to support traffic category (TC) that has been discussed previously [126]. It provides differentiated services for different priority queues waiting for accessing the shared wireless medium. It can support up to a maximum of eight priority queues. For differentiation purpose, first of all, the packets are classified into different queues, then the queues are prioritized by setting different Arbitration IFS (AIFS) values with AIFS length at least being equal to DIFS. The main reason for this is to provide support for non-QoS supported stations, so they can still compete for the wireless medium using the original DCF method. As shown in Figure 9.14, the higher the priority, the lower the AIFS value is set to that queue.

Similar to DCF, EDCF is using CSMA/CA. That is, before sending out a frame, the station is required to sense the medium to make sure that it is free, then each queue within the station waits for AIFS and make sure the medium is free. Then the queues can get into the contention window period similar to that of DCF. However, in EDCF, each queue can be associated with a persistence factor (PF) which controls the contention window size after each failure transmission of frame from that specific queue. In case that multiple backoff counters reach zero simultaneously (virtual collision), the medium access right is given to the highest priority queue, all other queues within the station undergo backoff again. The queue successfully gets the right to transmit within the station may have collision with other stations, however

if collision does not occur, it can transmit their frames up to a maximum duration predefined in Beacon and this is called Transmission Opportunity (TXOP).

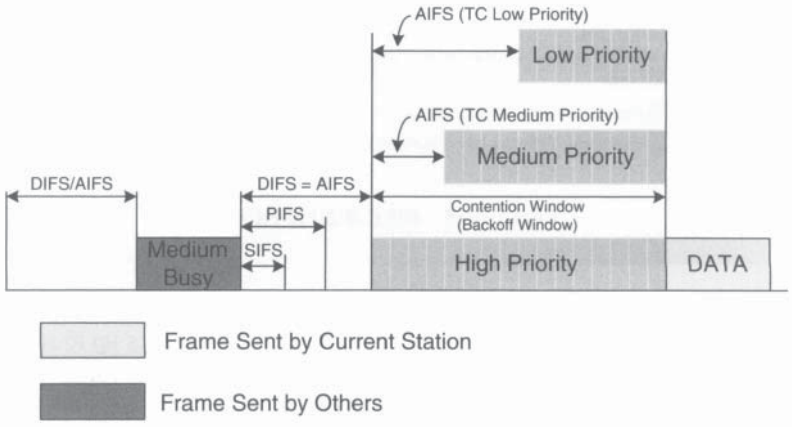


Figure 9.14 IEEE 802.11e EDCF.

HCF is mainly targeted for Traffic Stream (TS) [126] as we described earlier. The concept of HCF is very similar to PCF in the sense that there is a centralized coordinator which controls the access of medium by polling the stations. However, unlike that of PCF, stations are required to signal the HC with the QoS specifications (TSPEC) which state the requirements of the traffic stream of applications. In this case, HC grants certain TXOP to specific stations at certain period of time during CP or CFP.

As shown in Figure 9.15, during the CFP, the station polled by HC is allowed to transmit the frame within the TXOP specified in the CF-POLL frame. The CFP is terminated by the time specified in Beacon or specifically by a CF-END frame. During the CP, the QoS-enabled stations compete for the medium using EDCF as stated previously. However, it is also possible for HC to poll stations specially during CP. Since HC only waits for PIFS, while all other stations wait for AIFS (at least DIFS), HC has the priority access to the medium.

By using HCF and EDCF, the problems of delayed start of CFP and unknown duration of polled stations transmission can be solved, since the transmission duration is controlled by TXOP, QoS CF-Poll frame and Beacon frame. By doing so, QoS can be supported in WLAN environment.

9.7 ADVANCED DEVELOPMENTS

As we have discussed in Section 9.5, various Physical Link technologies have been proposed and eventually various IEEE 802.11 standards have been developed to suit

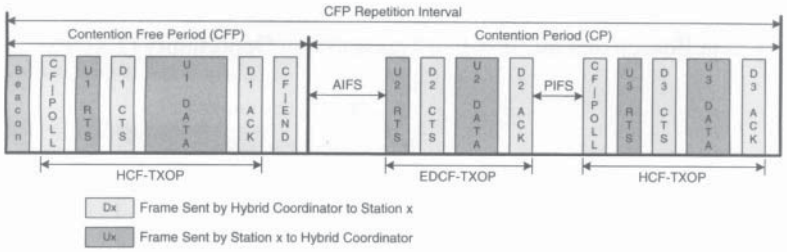


Figure 9.15 IEEE 802.11e HCF.

the needs of different users. However, users constantly request for better performances including higher throughput and larger coverage area. This comes up to some non-standardized ideas for increasing the throughput and extending the support range of standardized IEEE 802.11.

In order to increase the currently supported throughput capacity of the IEEE 802.11g, some WLAN vendors suggested and implemented some solutions like bursting, combined frames, multiple channels support etc. and these ideas are elaborated in detail as follows.

First of all, let us start with bursting. If a station knows that it has multiple frames to send out, after it takes control of the medium (after DIFS and random backoff), it continues controlling the medium by only waiting for SIFS which is the shortest possible IFS defined in IEEE 802.11 after each successive data frames are sent. In this case, a burst of frame can be sent out consecutively without the overhead of relatively large DIFS and contention windows. The whole sequence of actions is explained in Figure 9.16 [73].

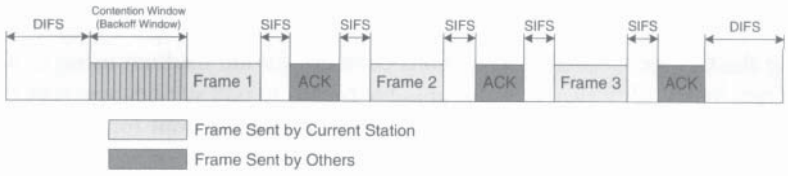


Figure 9.16 Bursting of MAC frames.

Other than bursting, another approach suggested in increasing the throughput is the combined frames approach. The scenario, in fact, is very similar to the case of bursting. When a station knows that it has multiple frames to send out, it can then make use of the opportunity that it has successfully taken control of the medium to send out a combined frames up to a maximum possible size. In this case, frames of different sizes are sent out instead. This approach is illustrated in Figure 9.17.

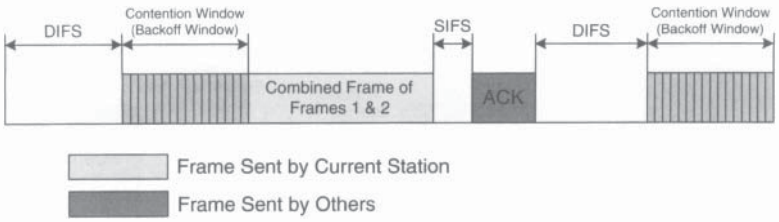


Figure 9.17 Combined MAC frames.

Furthermore, as IEEE 802.11b/g has three orthogonal channels, while IEEE 802.11a has 24 orthogonal channels, it is possible for an AP to choose more than one channel to transmit data if it finds that the bandwidth available is not enough to cope with the demands. In this case, the stations can make use of channels which are not used by other APs to increase the bandwidth significantly.

With the use of above idea, a network with more than 54 Mbps of data rate can be achieved. Other than higher throughput, users also expect longer supported range, larger coverage area and more reliable transmission of WLAN. In order to meet the demands of users, some vendors suggest the idea of enhancing the received sensitivity based on standard IEEE 802.11 . As shown in Figure 9.18 [72], when the station is within the original transmission range of IEEE 802.11, AP communicates with it using standard IEEE 802.11a/g with the transmission rate of 54 Mbps to 6 Mbps. However, once the station moves out of the original range, communication is carried out with smaller additional transmission rates from 3M to 0.25M. It is because a smaller transmission rate, in fact, has lower requirements in receive sensitivity and accuracy by the use of some simple modulations. In other words, signals with lower receive sensitivity can still be recognized and successfully decoded. Together with the improved MAC and PHY to support low Signal to Noise Ratio (SNR), the coverage area can be greatly increased.

Although the idea suggested results in a higher throughput and extended transmission range, there are some problems. The most serious problem is that the suggested idea is non-standardized. In other words, vendors can implement their own design and this leads to a serious problem in interoperability between WLANs from different vendors. At the same time, although the vendors suggesting these idea insist that they have backward compatibility to third party standard IEEE 802.11 design, their idea, in fact, leads to other problems that may break the design goals of IEEE 802.11. Just like the case of bursting and combined frames, these two methods lead to a higher throughput to one station, however, other stations implementing standard IEEE 802.11 may be required to wait for longer time before they can access the medium and this leads to the problem of fairness.

Finally, a recent development in further enhancing the capability of WLAN is to use the multiple-input-multiple-output (MIMO) physical layer technology. In simple

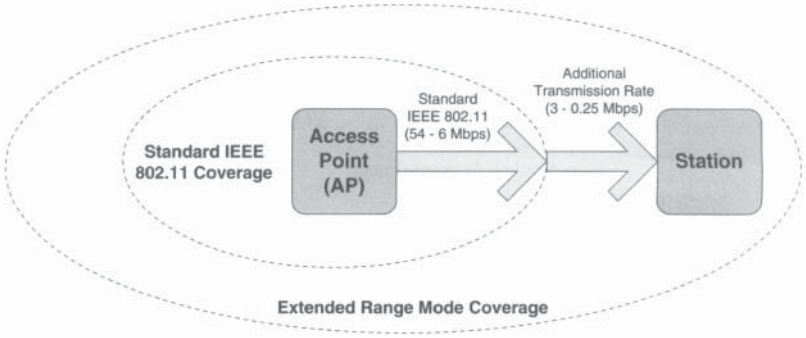


Figure 9.18 WLAN extended range.

terms, using multiple (say n) transmit antennae and multiple receive antennae, the data rate can be increased by n times. Indeed, it is estimated that the raw data rate of the MIMO based WLAN (called IEEE 802.11n) will be over 200Mbps.

9.8 PRACTICAL ILLUSTRATION: HOME NETWORK

WLAN is widely used in many places, for example, at home and at company, to provide networking facilities to the location where putting additional wires is impossible. Recently, there is a trend that WLAN is integrated into shopping malls and coffee shops to create hot spots for users to have temporary access to the Internet. In this section, we focus on a foreseen usage scenario of wireless network in near future—wireless home network [264].

In a home environment shown in Figure 9.19, coax cable is used to communicate with outside world. After the cable gets into the home, it is further split into three cables by cable splitter to support different locations within the home. There are three types of stations existing in the home environment. The first one is the station directly supported by the cable, while the second one is the station which communicates with the repeater which acts as a bridge between the cable and wireless medium, so as to extend the coverage of wireless network by transmitting the signal back to an access point. The third one is the station which communicates with an access point which implements both wireless station and coax station functionalities.

As shown in Figure 9.19, audio and video sources such as television programmes, movies, songs etc. is temporarily stored in the Set-Top-Box (STB) inside the home and transmitted through the coax cable. When Station 1 requested a certain audio or video programme, the programme is then moved to the thin Set-Top-Box and sent to Station 1 through the wireless medium. Since IEEE 802.11a/g has a bandwidth capacity of up to 54 Mbps, together with IEEE 802.11e QoS proposal, it is sufficient in supporting multimedia distribution. At the same time, when Station 2 wants to surf

the Internet, it can send the request to the repeater which then forwards the request to the access point which is connected with a cable modem for Internet access. In this case, solely coax cable and IEEE 802.11x wireless access technologies are sufficient in providing entertainment like VOD and Internet access to all users in all locations at home.

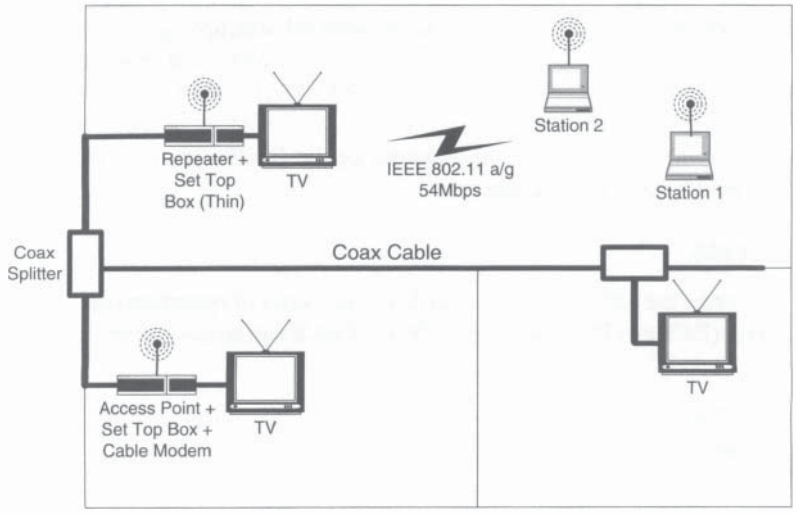


Figure 9.19 Wireless home network scenario.

In this scenario, solely cable may not provide the coverage area comparable with that of combination of both cable and wireless technologies, while solely wireless technologies may not provide enough bandwidth and coverage area, as some of the locations may be blocked by thick walls which cause great signal attenuation. Combination of both technologies can provide a highly satisfactory solution and it is expected to be more and more popular in near future.

9.9 SUMMARY

In this chapter, various IEEE 802.11x standards have been discussed. Within the standards, various MAC modes of operation, from the original DCF and PCF to the currently proposed QoS-supported HCF and EDCF have been discussed in detail. Moreover, various PHY link types including DSSS, FHSS, infrared and OFDM have also been mentioned together with their integration and various usages in IEEE 802.11x standards.

Since WLAN brings us many conveniences, more and more people and electronic devices vendors choose WLAN as a medium of communication. These lead to the constant demand for increasing the bandwidth and coverage area. In view of this,

vendors have tried to develop their own improved add-on technologies based on the standard IEEE 802.11x in order to achieve the above-mentioned goals and to provide the backward compatibility to those third party devices. Besides, they have also started thinking about the possibility of combination and implementation of various standards on a single device.

Same as many people, the authors also expect the potential usages of WLAN in various environment. Just like the home network example given in Section 9.8, with the bandwidth comparable to that of wired networks and large coverage area supported without the need of extra wires, IEEE 802.11x technologies can be used in home in combination of coax to provide the features of home theater and Internet access everywhere. Innovative applications of the IEEE 802.11x technologies are only bounded by our imagination.

PROBLEMS

9.1 Describe the differences between different modes of operation in IEEE 802.11x MAC layer (DCF and PCF) and how HCF and EDCF can be used to solve the problem of PCF in supporting QoS in wireless network.

9.2 Describe the methods used by vendors to increase the throughput capacity and the coverage area of WLAN, and how these actions in return affect the development of WLAN.

9.3 Briefly describe OFDM. Describe the differences between IEEE 802.11a and IEEE 802.11g and explain why IEEE 802.11ag may be a good solution in solving the problem of lacking bandwidth capacity.

CHAPTER 10

BLUETOOTH WPAN

No one is born a master.

—Italian Proverb

10.1 INTRODUCTION

While IEEE 802.11x WLAN technologies discussed in the previous chapter are already of a low cost, they are still considered not viable for many low-end commodity devices such as keyboard, mouse, etc. Indeed, to enable "cable replacement" for such commodity devices, WLAN technologies are too heavy-weight. In this chapter, we discuss about a low cost but effective short-range wireless technology.

Interoperability:

Bluetooth was originally developed by Ericsson in 1994 and its development is now managed by Bluetooth Special Interest Group (SIG). Bluetooth was developed in view of the great necessity in short-range cable replacement solutions, while existing solutions are insufficient in satisfying the needs of users. During the development

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau

Copyright © 2007 John Wiley & Sons, Inc.

of Bluetooth, special care has been taken on interoperability in two dimensions—interoperability with existing "cable replacement" interfaces and interoperability with other Bluetooth devices.

In order to facilitate fast pace development of Bluetooth, Bluetooth SIG decided to build up the protocols in protocol stacks from scratch that are necessary for the functions of Bluetooth only. Examples include the Core Protocols which include both Bluetooth PHY and MAC. On the other hand, some other protocols in Bluetooth are modified and/or adopted from existing solutions. There are quite a number of such protocols. For example, OBEX is an existing protocol developed for object exchange in infrared, and WAP is also an existing protocol developed for carrying Internet materials for portable devices like mobile phone. These existing protocols have been adopted into Bluetooth. Consequently, applications originally designed for other cable replacement interfaces can be used in Bluetooth also. This not only increases the development pace of Bluetooth, but also provides interoperability with current existing cable replacement solutions. Together with Bluetooth protocol architecture, these issues are further discussed in Section 10.3.

On the other hand, Bluetooth also provides interoperability with other Bluetooth devices by standardizing the protocol stacks and profiles. Profiles, in fact, record the protocol requirements necessary to be implemented for supporting specific usage model. By standardizing usage model with profiles, any products claimed to support specific usage models and profiles should implement the same protocol stacks under the same requirements. In this manner, interoperability between different Bluetooth devices can be ensured. Together with Bluetooth profiles, these issues are further discussed in Section 10.3.

Performance:

Bluetooth, similar to IEEE 802.11 b/g, uses the 2.4 GHz ISM band. It provides a theoretical capacity up to 1 Mbps due to the design of its PHY and MAC. This is covered in detail in Section 10.4. On the other hand, Bluetooth supports both asynchronous traffics and synchronous traffics with the help of two physical channels—ACL and SCO. With the help of SCO, some time slots can be reserved for time sensitive applications like audio, which provides certain capability of QoS. This is further discussed in Section 10.4. Moreover, some current suggestions on the improvement of QoS support in Bluetooth are discussed in Section 10.5. Bluetooth can provide a theoretical capacity up to 1 Mbps. However, in actual scenarios, this theoretical capacity cannot be achieved. Together with some currently released analytical and simulated results, this issue is further elaborated in Section 10.6.

10.2 DESIGN GOALS

IEEE 802.11 has been developed for years. There was a great controversy on the necessity of the development of a new wireless technology in complementary to the

existing one. Bluetooth has been developed with several objectives that defy the controversy. These objectives can be summarized in Table 10.1.

Table 10.1 Bluetooth design goals.

Objective	Description
Low Cost	Bluetooth is a one-chip solution with the goal of US$5 per chip in the long run due to the expected mass production. With this low cost, vendors would be able to integrate Bluetooth into their products without great increase in their production costs and users would be able to afford the products.
Cable Replacement	The development objective of Bluetooth is not aggressive, it mainly targets at functioning as a short-range cable replacement for devices like headset of mobile phone, PDA, etc. Together with its low power consumption and low cost, more and more users use Bluetooth devices as a replacement for lengthy, inconvenient cables.
Ad Hoc Networking	In the market, many devices like remote control, PDA, mobile phone, etc. provide infrared as wireless point-to-point communication technologies. This has been considered appropriate because most of the connections between these devices are only set up temporarily (i.e., ad hoc in manner). In view of this, with the interoperability brought by Profiles specification, different products can communicate with each other and set up a temporarily connection as long as they have implemented the same profiles.

Currently, more and more usage of Bluetooth has been discovered and implemented. New design objectives keep emerging to support new usage models. For example, there is a proposal of increasing the theoretical capacity of Bluetooth from the original 1 Mbps to 2 Mbps or more.

10.3 BLUETOOTH PROTOCOL STACK

The development of Bluetooth technologies and Bluetooth protocol specifications are carried out by Bluetooth Special Interest Group (SIG). Originally, Bluetooth specifications are divided into two parts—Core specification and Profiles specification. The Core specification covers a layered protocol stacks architecture, in which protocols including the Bluetooth Physical layer up to Bluetooth specific Logical Link Control layer are defined. At the same time, some implementation details including maximum and minimum power dissipation, and the maximum supporting range, are also included. On the other hand, Profile Specification mainly focuses on specific usage models of Bluetooth. The specification v1.1 covers 13 Profiles and more extended profiles are released currently. In each profile, the specification defines clearly on what Bluetooth Protocol Stacks are required to implement and what other actions

could be implemented as an options. By doing so, devices from different vendors which claim to support specific profiles would be able to interoperate [1, 2, 323].

First of all, let us consider the Core specification in some more details. In Core specification, a layered Bluetooth Protocol Stack are defined as shown in Figure 10.1. We can see that within the protocol stack, four group of protocols are involved. They are Core Protocols, Cable Replacement Protocol, Telephony Control Protocol and Adopted Protocols. Their components and corresponding functions are summarized from Table 10.2 to Table 10.5.

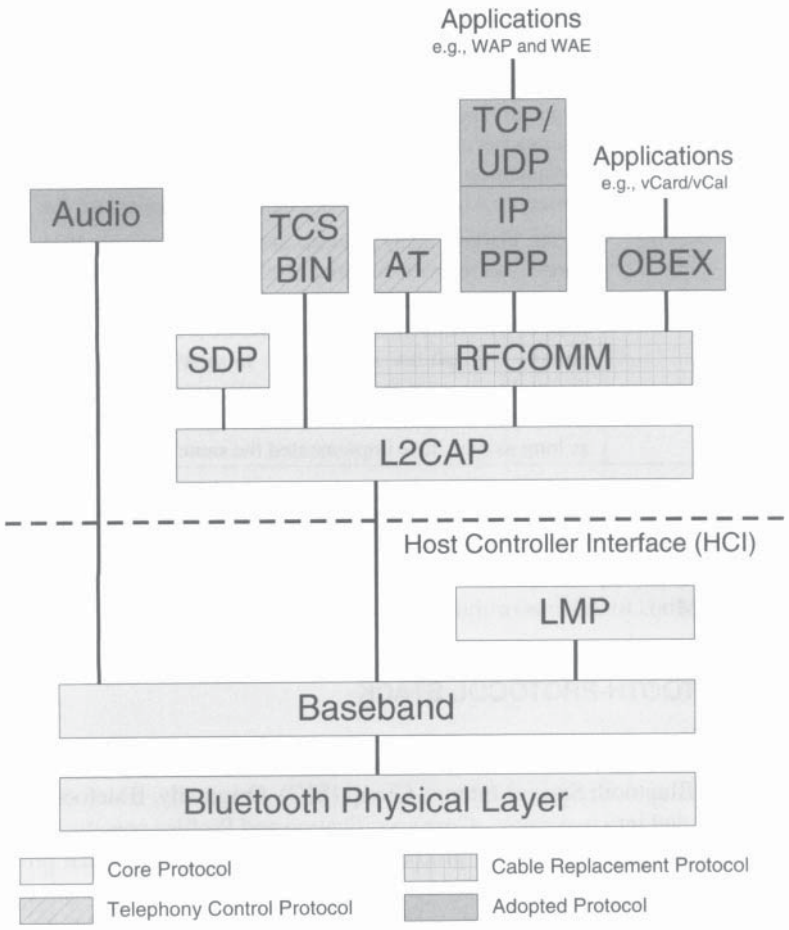


Figure 10.1 Bluetooth protocol stack.

After discussing the protocol stack in Core specification, We now proceed to describe the Profiles specification. Figure 10.2 shows Profiles specification v1.1

Table 10.2 Members of Core Protocols and their functions.

Member	Description
Bluetooth Physical Layer	This layer specifies the details about the radio characteristics of Bluetooth, including the frequency, frequency hopping usage, power and modulation.
Baseband	This layer specifies connection establishment in piconet including the physical link establishment, clock synchronization, etc.
Link Manager Protocol (LMP)	This layer manages the link setup and ongoing link management. On the other hand, it is also responsible for security issues including authentication and encryption.
Logical Link Control and Adaptation Protocol (L2CAP)	This layer acts as an interface between the upper protocol layer and baseband which multiplex traffic flows that are connectionless and connection-oriented. It also provides packet segmentation and reassembly for better management of the link usages.
Service Discovery Protocol (SDP)	This layer is mainly responsible for discovering other Bluetooth devices in nearby physical locations, and their corresponding characteristics including the services provided and some basic information, for example, the profiles, the link types and the frame types supported, etc.

Table 10.3 Member of Cable Replacement Protocol and its function.

Member	Description
Radio Frequency Communications Protocol (RFCOMM)	This layer is used to emulate a serial port over L2CAP layer based on European Telecommunications Standards Institute (ETSI)'s Technical Standard (TS) 07.10 that is originally designed for Global System for Mobile (GSM) communications. This protocol emulates RS232 serial port (or commonly named as COM port)'s control and data signal over L2CAP protocol. By doing so, upper layer protocols generally used for wired networks (for example, PPP over modem) can be used here.

Table 10.4 Member of Telephony Control Protocol and its function.

Member	Description
Telephony Control Specification Binary (TCS BIN)	This layer is a bit-oriented protocol defined for call control signaling which is used to build up a data and speech call between Bluetooth devices based on the ITU-T Q.931. At the same time, mobility management procedures are also provided to maintain group of Bluetooth devices.

Table 10.5 Members of Adopted Protocol and their functions.

Member	Description
Point-to-point Protocol (PPP)	This is a protocol defined by Internet Engineering Task Force (IETF) for carrying IP datagrams traffic over serial point to point link (for example, RS232 link).
Internet Protocol (IP)	Connectionless network protocol commonly used in Internet.
Transmission Control Protocol (TCP) and User Datagram Protocol (UDP)	TCP and UDP are respectively connection-oriented and connection-less transport layer protocol used in Internet.
Object Exchange Protocol (OBEX)	This protocol is originally developed by Infrared Data Association (IrDA) to support object exchange between devices. This protocol is subsequently adopted into Bluetooth for object exchange between devices.
Wireless Application Protocol (WAP) and Wireless Application Environment (WAE)	A protocol which provides sending and receiving Internet materials on small wireless devices like mobile phone. This is discussed in detail in Chapter 15.

and their corresponding relationship. As we have discussed earlier in this section, Bluetooth Profiles specification v1.1 has included 13 profiles and there are constantly some more profiles released as extended profiles due to the discovery of more usage models of Bluetooth. Examples include Personal Area Networking (PAN) for building up a network, Human Interface Device Profile (HID) for Bluetooth keyboard, mouse or gamepad, Basic Printing Profile (BPP) for Bluetooth enabled printer, etc.

In fact, the main function of profiles is to define which specific components of protocol stacks should be implemented in supporting a certain usage model. Consequently, interoperability is provided between products from different vendors that claim to support specific profiles. The most basic one is Generic Access Profile which specifies most core protocols. As a result, most profiles depend on Generic Access Profile, as shown in Figure 10.2. On top of it are Telephony Cordless Binary Based Profiles which include TCS BIN layer, and Serial Port Profile which includes RFCOMM layer. With the help of RFCOMM, many other profiles can be implemented, such as General Object Exchange Profile which includes OBEX layer, LAN Access Profile (LAP)[1] which includes PPP layer and optional TCP/IP layer, Dial-Up Networking Profile which includes AT commands layer for modem control and optional PPP layer, etc. By building profiles on top of others, adopted/additional functions can be implemented easily and this allows Bluetooth to interoperate with programs using protocols originally designed for other technologies. OBEX originally designed for infrared for object exchange is one of the best examples.

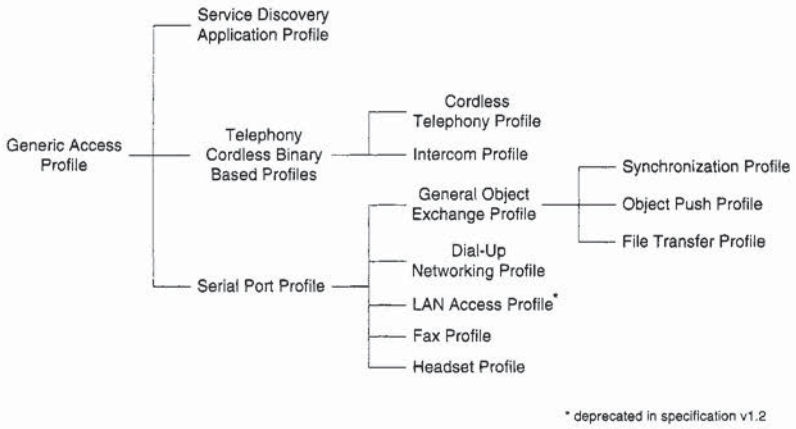


Figure 10.2 Bluetooth profiles specification v1.1.

[1]LAP is deprecated in v1.2 [2] due to the release of PAN, which is specially designed for supporting networking applications. Some profiles, for example, WAP, which rely on LAP are also moved above PAN in the new release.

In fact, due to the development of more and more usage models on Bluetooth, more and more profiles are expected and developed. As a result, Bluetooth protocol stacks keep on expanding with some newly developed layers specifically for Bluetooth, for example Bluetooth Network Encapsulation Protocol (BNEP) in PAN profile, and some newly adopted layers.

10.4 BLUETOOTH PHYSICAL AND MAC LAYERS

10.4.1 Bluetooth Physical Layer

Bluetooth uses the 2.4 GHz ISM frequency band, similar to that of IEEE 802.11 discussed in Chapter 9. This frequency band is unlicensed and can be used by devices as long as specific frequency and power regulations are followed. Accordingly, this frequency band is crowded and so any technology using this band should have mechanism in withstanding collisions with other technologies using the same frequency band.

In view of this, Bluetooth adopted Frequency Hopping Spread Spectrum (FHSS) technique described in Section 9.5 in its Physical (PHY) layer with 79 channels[2] and bandwidth of 1 MHz each. The frequency band details are described in Table 10.6 [1, 2, 53, 323]. Within this frequency band, Bluetooth has a relatively high hopping rate (1600 hops per second) so as to get a better resistance to interference from devices using the same band. Together with GFSK as the modulation scheme, it can provide up to a maximum throughput capacity of 1 Mbps.

Table 10.6 The frequency band usage of Bluetooth.

Frequency Band Range	2.400 - 2.4835 GHz
RF Channel	$f = 2.402$ GHz + n MHz, $n = 0, ..., 78$
Lower Guard Band	2 MHz
Upper Guard Band	3.5 MHz

Other than frequency bands and channels requirements, radio specification in Bluetooth also specifies three power classes which in return determine the transmission range of Bluetooth devices. The classes and their corresponding power output are summarized in Table 10.7.

10.4.2 Bluetooth MAC Layer

In this section, we describe Bluetooth MAC layer in detail [1, 2, 53, 323]. Bluetooth Baseband, in fact, plays a crucial role in the Bluetooth MAC layer, as it supports the functions of MAC and helps in frequency hopping, clock synchronization, etc.

[2]In specification v1.1, both 79 and 23 hopping sequence channels are defined, but the 23 hopping sequence channels are deprecated in specification v1.2.

Table 10.7 The three power classes of Bluetooth.

Class	Maximum Power Output	Minimum Power Output	Remarks
1	100 mW (20 dBm)	1 mW (0 dBm)	Longest transmission range with power control over +4 dBm. Power control is optionally below +4 dBm.
2	2.5 mW (4 dBm)	0.25 mW (-6 dBm)	Optional power control.
3	1 mW (0 dBm)	N/A	Optional power control.

The linkage between Bluetooth devices is based on a master-slave relationship, in which one device becomes the master and other devices become slaves. A master can control not more than seven active slaves. This structure is called *piconet* and is further discussed in Section 10.5. Within a piconet, the master device and slave devices communicate in time division duplex (TDD) manner in which the data transmissions use up the whole frequency slot (1 MHz) with downlink (from the master to slaves) and uplink (from slaves to master) transmission divided into different time slots. This is illustrated in Figure 10.3. Similar to the case of PCF in IEEE 802.11, the master has the function similar to that of point coordinator in the sense that the master polls each slave by sending them a downlink frame. If a slave has backlogged data to send, it then replies with a uplink frame in next time slot. Unlike that of PCF, each time slot in Bluetooth is fixed to be $625\mu s$ which corresponds to 1600 hops in a second, while PCF does not have the concept of time slot instead. As indicated in Figure 10.3, within each time slot, the frame is transmitted using a specific frequency $f(k)$. That is, the master sends the frame using frequency $f(k)$ in the first time slot, then the slave replies with the frame using another frequency $f(k+1)$ in the next time slot.

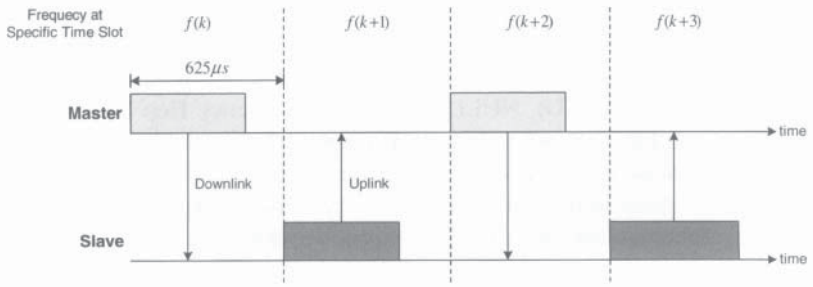


Figure 10.3 Time division duplex in Bluetooth MAC.

The frequency that is used in next time slot is specific to each piconet and this is called Frequency Hopping Sequence. This sequence is derived from the master BD_ADDR and clock. BD_ADDR is extended from IEEE 802 architecture, and be-

cause of this, BD_ADDR is unique for each Bluetooth device. Even though frequency hopping sequence is used, there are still chances that the frequency used by different piconets within the same time slot is the same and this leads to collisions. However, with the use of frequency hopping techniques, the chances of colliding within specific time slot becomes smaller and consecutive collisions become unlikely. With the use of error detection and correction techniques like Forward Error Correction (FEC), which is elaborated below, the effect of collision can be practically handled.

Figure 10.3 shows alternating transmission of frame making use of 1-slot time only. However, it is possible for either side to send out frames which take up multi-slots. There are three types of frames—1, 3, 5 slots frames and they are shown in Figure 10.4. Within the transmission of the multi-slot frames, the frequency is unchanged. However, after the transmission of the frame, the next frequency used is the frequency that should be used after multiple time slots, as shown in the Figure 10.4.

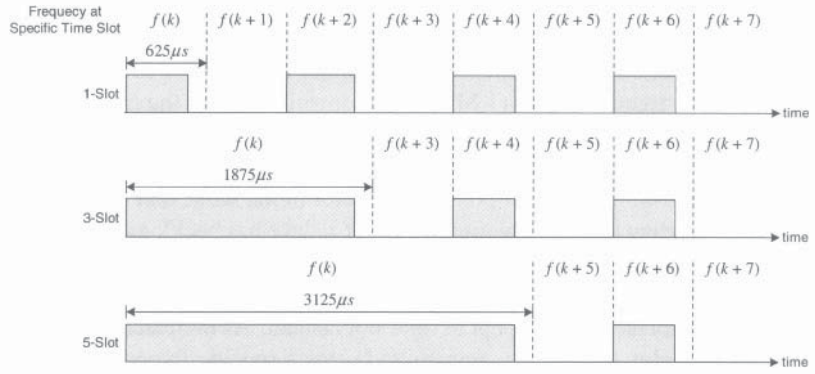


Figure 10.4 Multi-slot frames in Bluetooth.

There are several types of packets defined in Bluetooth. They are control packets, Asynchronous Connectionless (ACL) packets, Synchronous Connection-Oriented (SCO) packet, and extended Synchronous Connection-Oriented (eSCO). Control packets include Identity (ID), NULL, POLL and Frequency Hop Synchronization (FHS) packets. Table 10.8 shows ACL packet types, whereas Table 10.9 shows SCO and eSCO types of packets. Different types of packets are used for supporting different physical links defined in Bluetooth, which are based on different error concealment techniques. There are two main types of physical links that can be established between Bluetooth devices—ACL link and SCO link. The characteristics of these two links are discussed in detail below.

ACL refers to the point-to-multipoint connections between a master and its associated slaves within a piconet. The master exchanges packets with its slaves in a round-robin fashion (although other ordering methods may also be possible), only the slaves with ACL connection addressed by the master is allowed to transmit at that specific time slots. ACL connection is still allowed even though a SCO connec-

Table 10.8 Bluetooth ACL packet types.

Type	Slot	Max Symmetric Rate (kbps)	Max Asymmetric Rate (kbps)	Max Payload (Bytes)	Description
DM1	1	108.8	108.8 | 108.8	17	With CRC and 2/3 FEC.
DH1	1	172.8	172.8 | 172.8	27	With CRC, but no FEC.
DM3	3	258.1	387.2 | 54.4	121	With CRC and 2/3 FEC.
DH3	3	390.4	585.6 | 86.4	183	With CRC, but no FEC.
DM5	5	286.7	477.8 | 36.3	224	With CRC and 2/3 FEC.
DH5	5	433.9	723.2 | 57.6	339	With CRC, but no FEC.
AUX1	1	185.6	185.6 | 185.6	29	No CRC and FEC.

Table 10.9 Bluetooth SCO and eSCO packet types.

Type	Slot	Max Symmetric Rate (kbps)	Max Payload (Bytes)	Description
HV1	1	64.0	10	No CRC, but 1/3 FEC.
HV2	1	64.0	20	No CRC, but 2/3 FEC.
HV3	1	64.0	30	No CRC and FEC.
DV	1	64.0 + 57.6 (Data)	19 (10 + 9 (Data))	No CRC and FEC (voice); with CRC and 2/3 FEC (data).
EV3	1	96	30	With CRC, but no FEC.
EV4	3	182	120	With CRC and 2/3 FEC.
EV5	3	288	180	With CRC and FEC.

tion is established between two Bluetooth devices and there must only be one ACL connection between two Bluetooth devices.

ACL is mainly used to support asynchronous type of traffic, for example, file data, and as such, no reservation is allowed (i.e., packet switching connection). As shown in Table 10.8, there are two main types of packets—DMx and DHx, where x is a number corresponding to the number of slots occupied by that packet type. The main difference between these two types of packets is that 2/3 FEC with (15,10) shortened Hamming code is used in DMx packet type. That is the reason why the maximum throughput capacity achieved by using DMx traffic is around 2/3 that of DHx traffic. Apart from FEC, Automatic Retransmission Request (ARQ) is used in ACL traffics to give the packet further protection.

SCO refers to the point-to-point connection between a master and a single slave with a fixed bandwidth allocated to support this connection. It is mainly used to support synchronous type of traffic, for example, voice (circuit switching connection). Due to this reason, the master reserves a pair of time slots at regular interval for the corresponding slave. Within the reserved time slots, the slave can send out a packet in next time slot no matter whether the packet from master is addressed to it. There are maximum of three simultaneous SCO connections.

Other than DM1, there are four basic types of packets allowed in SCO. They are HVx and DV. As shown in Table 10.9, the main difference between various HVx packets is on the types of FEC used: HV1 uses 1/3 FEC, HV2 uses 2/3 FEC and HV3 does not use FEC. They all provide a fixed transmission rate of 64 kbps which is a standard bandwidth requirement for 8-bit digitized voice. While for the case of DV, both voice and data are be encoded into it with the voice portion providing up to 64 kbps and data portion providing up to 57.6 kbps. For this type of packets, ARQ is generally not required. However, it is adopted in data portion of DV. On the other hand, EVx is the packet types defined in specification v1.2 in eSCO [2]. eSCO is also considered SCO type because it also reserves slots between a master and specific slave. However, the main difference between them is that eSCO allows retransmission in all its packet and allows both symmetric and asymmetric with up to three slots to be used.

10.5 PICONETS AND SCATTERNETS

10.5.1 Overview

There are three topologies defined in the specifications—point-to-point connection, single point-to-multipoint connection (piconet) and multiple point-to-multipoint connection (scatternet). Their basic structures are shown in Figure 10.5 [1, 2, 53]. In the figure, point-to-point connection is shown on the right single connection between a master and a slave. As shown in the figure, there are three different piconets. The masters of Piconet 1 and Piconet 2 connect to three slaves, respectively, while the master of Piconet 3 connects to two slaves. Together, these three piconets are connected together to form a single scatternet. As can be seen in Figure 10.5, there are two methods of linking two piconets together. The two methods are shown in the connection between Piconets 1 and 2, which makes use of master-slave bridge, and the connection between Piconet 1 and 3, which makes use of slave-slave bridge.

10.5.2 Piconet Formation

Before the scatternet formation details are discussed, let us focus on the piconet formation details. In order to set up a piconet, there are two main procedures required—inquiry and paging [1, 2, 53].

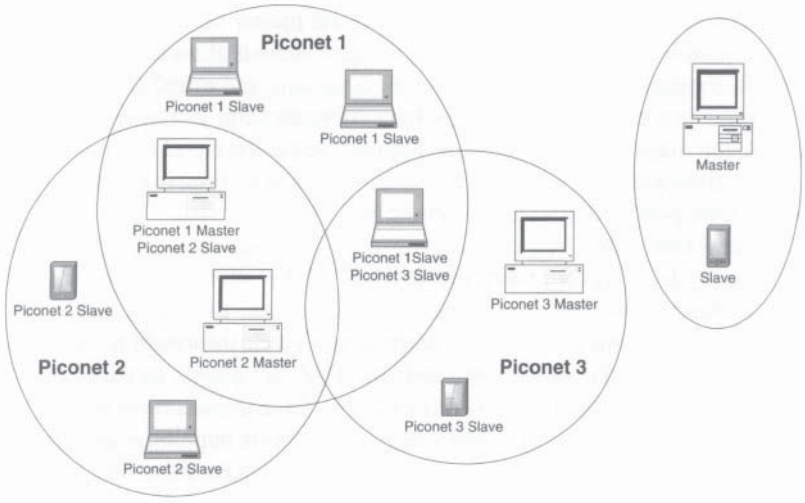


Figure 10.5 Bluetooth topologies.

Inquiry procedure is mainly used by a potential master device to find other surrounding devices which are willing to form a piconet. Initially, all the devices are either in Standby or Connection state as shown in Figure 10.6. Firstly, the potential master starts by entering the inquiry state and sends out Identity (ID) packet with inquiry access code (IAC). The potential master makes use of common frequency channels (known as wake up carriers) to send out the ID packet. The wake up frequency channels include 32 out of total 79 Bluetooth channels. The IAC ID packet is sent out over each of these wake up channels.

On the other hand, a potential slave enters the inquiry scan state periodically to look for IAC ID packets. In receiving IAC ID packet, the potential slave enters inquiry response substate and generate a Frequency Hop Synchronization (FHS) packet in response which includes the slave's BD_ADDR, clock and other information. However, there may be chances that multiple potential slaves receive IAC ID packet simultaneously. In order to avoid collision in replying, potential slaves follow the random backoff mechanism stated in Section 9.4 before sending out the response. After response has been sent, the potential slaves then enter page scan state and wait for paging from the potential master. If no paging is received after a specific amount of time, the slaves switch back to inquiry scan state due to a possible collision of FHS packets in the previous state. On the other hand, the potential master may not want to reply the FHS packets from slaves as soon as possible. It is because it would like to wait for as many replies as possible from the surrounding. As such, it stays in inquiry state for some time. The duration of this procedure may last for 10.24s in order to provide enough time to get replies from all surrounding potential slaves.

After receiving enough responses, the potential master then enters page state as shown in Figure 10.6. The main purpose of this procedure is allow the potential master to establish connection with potential slaves in piconet and establish a synchronous hopping sequence between all devices for communication in the piconet. At the beginning, the master calculates a page hopping sequence based on specific slave's BD_ADDR and clock information. This specific frequency hopping can then be used to page specific potential slave in this procedure. The master sends an ID packet with device address code (DAC) which is the lower part of a slave's address using specific page frequency hopping. In receiving DAC ID packet, the slave echoes the same DAC ID packet to show that it has successfully received. The master then responds with a FHS packet indicating its own address and clock information. In receiving this packet, the potential slave then sends the DAC ID packet to show that it has received the packet successfully. At this time, the slave transits from page response substate to connection state and it then uses the connection hopping sequence derived from master's clock and address to communicate later. On the other hand, the master continues to page other potential slaves and enters connection state after all paging operations have been done.

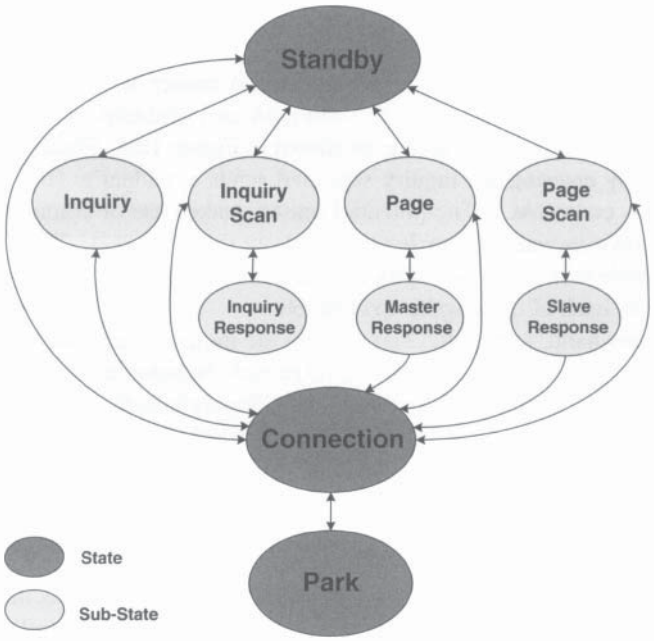


Figure 10.6 Bluetooth transition states.

At this point, a piconet is formed with a master and multiple slaves in connection state with a synchronized clock and hopping sequence. However, there are several

modes that a slave can participate in the piconet. The modes are shown in Figure 10.7, while their functions are summarized in Table 10.10

Table 10.10 Bluetooth connection state modes.

Mode/State	Description
Active	The slave gets a 3-bit AM_ADDR code and actively participate in the activities of piconet. Since there are only 3-bit AM_ADDR, there are at most seven active, sniff and hold mode slaves in piconet.
Sniff	The slave also gets AM_ADDR, but only interests in the messages in specified time-slot to which master has been informed before joining the piconet. Sniff mode slave can enter reduced power mode in-between wakeup time.
Hold	The slave also gets AM_ADDR. In this mode of operation, ACL packet exchange is not supported, but SCO packet exchange can continue. The slave can entered reduced power mode in specific piconet. This mode is mainly used by bridge node in which actively participate in more than one piconet simultaneously is impossible, so the activities in particular piconet can be temporarily held and the node can switched between different piconets by using this mode and still be considered as connected member.
Park[a]	In this state, a slave is given a 8-bit parking member address (PM_ADDR) and is given up its own AM_ADDR. In this case, it can not actively participate in the piconet, but it is still considered as a part of it. With this state, the slaves can freely get into lowered power mode and the piconet is allowed to have more than seven members with 8-bit PM_ADDR in this state.

[a] Originally Park state is also considered as a mode in Connection State, it becomes a new state in specification v1.2 with similar functionalities as specification v1.1.

10.5.3 Scatternet Formation

As discussed in the previous section, a piconet cannot have more than seven active members. In order to increase the size of the network and also increase the coverage area, Bluetooth devices can form a scatternet, which is constructed by connecting multiple piconets together. Since there is still no conclusion on how scatternets should be formed, in this section, a number of methods currently suggested are discussed [53].

In scatternet formation, two main categories can be concluded—centralized approach and distributed approach. In the centralized approach, a leader is elected from all the Bluetooth devices within the transmission range of each other. Then the leader makes decision on the final topology of the scatternet. The final structure may be

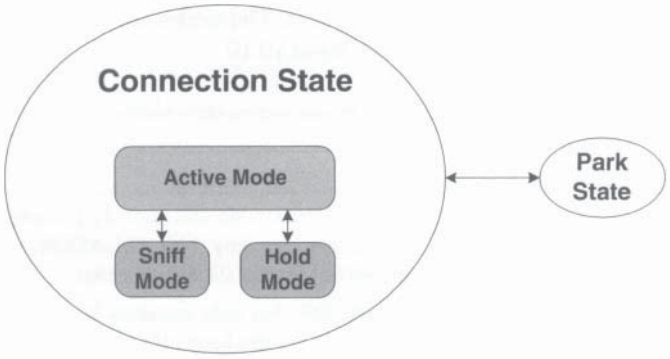


Figure 10.7 Bluetooth connection states.

a tree, a star, a ring, etc. In the distributed approach, discovery of neighbors, connections and other topology control actions are done locally. Consequently, different number of localized clusters are formed. These localized clusters are then combined to form a scatternet.

First of all, some centralized scatternet formation approaches are briefly discussed:

1. **Bluetooth Topology Construction Protocol (BTCP).** It is a three-phase scatternet formation protocol. The first phase involves an election of a leader by Voting process, in which each device is assumed to have one VOTE at the beginning and alternately performs inquiry and inquiry scan. Once connected, the one with a larger VOTE wins and get the VOTE of loser. The winner continues inquiring for more nodes, while the loser enters page scan state. At the end of first phase, a centralized controller with global information is elected. The second phase involves the topology construction decision making by the leader. The decisions on which nodes are designated as the master, slave and bridge are forwarded to specific devices. Then the actual connection is made in the third phase. The final topology may be a mesh, tree, etc.

2. **Bluetree Scatternet Formation.** This formation protocol is similar to BTCP where a leader is elected at the beginning to become a root, so called Blueroot. This root with complete information then initiates a spanning tree formation process with itself being the root of the spanning tree. As a result, a spanning tree structure is formed.

We now consider some semi-centralized or distributed scatternet formation approaches:

1. **Bluenet Scatternet Formation.** Bluenet Scatternet Formation is a three-phase scatternet formation protocol. In the first phase of this protocol, similar to centralized scatternet formation algorithms, alternating inquiry and inquiry scan

are done by each independent node to acquire the neighborhood information and exchange information with others. Afterwards, piconets are formed by several nodes nearby to form small clusters across a distributed area. In the second phase, any remaining independent nodes in phase one should enter page scan mode to wait for paging. In the last phase the slaves in a piconet then actively search for other slaves in other piconets to form a scatternet. The final topology may be a tree or some other sparsely connected structure.

2. **BlueConstellation Scatternet Formation.** This algorithm is another three-phase protocol for scatternet formation. Similar to the above protocols, the first phase is to exchange information among neighbouring nodes. In the second phase, Bluestars (piconets) is formed by electing the master with the largest weight of clustered nodes around. In the final phase, the Bluestars are connected to form BlueConstellation with the property that each master is within three hops' distance of each other. By doing so, the maximum number of hops of a node in one piconet to another node in another piconet can be determined and fixed.

3. **Randomized Scatternet Formation.** In this scatternet formation algorithm, all the nodes are the leaders at the beginning. Each leader randomly decides a probability within 1/3 to 2/3 on whether it performs a SEEK operation, which is an inquiry action, or stay on performing a SCAN operation, which is an inquiry scan action. When a match occurs, the two components are merged together with one of the leader retired, which will then restructure to maintain two constraints predefined (i.e., each leader has either no slave or one unshared slave, and each leader should have no more than k slaves). After a number of iterative leader retiring operations, there is only one leader left and finally a scatternet is formed.

In the above two categories of scatternet formation, there are several points in common. Firstly, an initial setup and maintenance is required to maintain the structure. Secondly, the topology formed is closely related to how a packet can be routed from one node in one piconet to another node in another piconet. However, most of the scatternet formation algorithms stated above do not consider the second point. The following scatternet formation algorithms have taken this issue into account and provide feature of "self-routing":

1. **Scatternet Formation Algorithm based on Search Tree.** This algorithm is very similar to the Bluetree Scatternet Formation. However, it is different in that the root is chosen based on the BD_ADDR with the node having the largest BD_ADDR to be the root. Then it makes decision on the tree topology construction based on a tree-sort like procedure. In this procedure, an invariant in a node A of the tree should be maintained with the property: $max(c_i) < min(c_j)$ for all $i < j$, where c_i and c_j are children of node A, while $max(c_i)$ and $min(c_j)$ are the largest BD_ADDR and smallest BD_ADDR in c_i and c_j,

respectively. By using this algorithm, a sorted tree structure will be formed, so that routing in such structure can be easily done.

2. **BlueRing Scatternet Formation.** Similar to most of the algorithms, the assumption that all nodes should be within the transmission range of each other is maintained. After the information exchange step, the one with the largest BD_ADDR is elected as the leader. The leader then assigns some other nodes in the network to be masters to form a ring with one slave destined as slave-slave bridge in between. With this structure, routing can be done by passing through ring of masters.

Some of the self-routing scatternet formation algorithms have been discussed above. However, the efforts of forming a scatternet may not be worthwhile in the case where there are low traffic and frequent topology changes. Due to these reasons, some researchers suggested the usage of traditional ad hoc routing techniques in carrying packets from a node to another without the necessity of forming a structure beforehand. There are several such routing algorithms suggested currently in Bluetooth, for example, Routing Vector Method, on-demand scatternet routing, zone routing protocol, which are adopted from traditional ad hoc routing protocols like Ad Hoc On Demand (AODV), Dynamic Source Routing (DSR) and Zone Routing, which are discussed in detail in Chapter 19.

10.5.4 QoS Consideration in Piconet and Scatternet

In the previous section, piconet and scatternet formation approaches have been discussed in detail. However, in the piconet and scatternet formation, only limited QoS is considered [53].

In the case of piconet, the master polls its slaves in simple round robin fashion. However, this method may not be an efficient one since some slaves may have more traffics than the others. Bluetooth uses TDD in MAC, so in order to improve QoS, the possible methods include using better scheduling algorithms in TDD and/or improve the Segmentation and Reassembly process (SAR) in the frame. In view of this, suggestions on improvement have been proposed and some of them are summarized below:

1. **Piconet Queue-State-Dependent Scheduling.** An example is head-of-line priority policy (HOL-PP). In HOL-PP, the master-slave communications are classified into three classes. The first class is the one with 100% utilization of channels (1-1,1-3,3-1). The second class is the one with 75% utilization of channels (3-0, 0-3). The third class is the one with only 50% of utilization (1-0, 0-1). The allocation of time-slots and polling frequency for specific master-slave pair are then be adjusted according to the classification.

2. **Piconet SAR and Scheduling.** Examples are SAR-BF and SAR-OSU. SAR-BF aims at optimizing the channels utilization by segmenting the upper layer

packet into the slot frames that best utilize the channels. On the other hand, SAR-OSU aims at reducing the overall delays. Consequently, it may prefer the usage of the slot frame with size larger than the upper layer packet size, so as to send out the frames as soon as possible.

On the other hand, focus has been given on scatternet formation. However, little has been suggested on the QoS issue of scatternets. In a scatternet, it relies on the bridge nodes to relay packet from one piconet to another. As such, in order to consider the issue of QoS in scatternets, better time scheduling has to be done on the bridge nodes. Some of the suggestions on improving the performance of scatternets are summarized in the following:

1. **Scatternet Nodes Randomized Rendezvous Scheduling.** The bridge nodes in a scatternet have neighbors (links) from different piconets, however, they cannot meet with more than one of them at the same time. In this case, Pseudo Random Coordinated Scatternet Scheduling (PCSS) has been suggested with the idea of randomly choosing the rendezvous time by using the master's clock and slaves' addresses.

2. **Locally Coordinated Scheduling (LCS) in Scatternets.** This approach also addresses the rendezvous issue of bridge nodes. However, instead of randomly choosing a point, this algorithm calculates the start time of next rendezvous by observing the data rate trend. At the same time, it also finds out the duration of next rendezvous by using the queue size and past transmission history, so that the duration is about right to transmit all the backlogged traffics.

In summary, some of the QoS considerations in piconets and scatternets have been suggested together with the general ideas on how to improve QoS. However, there is no consensus on which one is the best and so this topic is still under development.

10.6 PERFORMANCE ISSUES

As it has been discussed previously, Bluetooth is using the 2.4 GHz ISM frequency band, which is crowded with devices of other technologies. However, even though this issue is neglected, multiple Bluetooth piconets formation with overlapping coverage area also brings up great interferences since they are using the same frequency band independently. Unfortunately, this case happens frequently nowadays when Bluetooth becomes more and more popular with users forming their own Bluetooth personal area network (PAN) using multiple electronics devices such as PC, laptop, PDA, mobile phone, etc., which are generally not shared with other users. In this section, we describe a study on the performance issue [74].

In order to determine the performance of Bluetooth in a more realistic case, several issues are required to be mentioned beforehand. Firstly, the power received by a receiver depends on the distance between the transmitter and receiver, the transmitted

power and path loss exponent used, etc. Secondly, the occurrence of interferences depends on the ratio between power received from the intended transmitter and all other surrounding transmitters. As long as the ratio is higher than certain threshold value, the received signal is still valid. This in turn determines the probability of a successful transmission of a packet in a piconet.

With the above information and the assumption that seven piconets are located within the region where a large portion of each piconet is interfered by the others, the following analytical and/or simulated results are shown:

1. The analytical results show that the average throughput of Bluetooth with interference is better than that of FDMA in which the whole bandwidth has to be divided into seven non-overlapping frequency bands for seven independent piconets.

2. Bluetooth performance varies with the number of obstacles and interferences with a given receive threshold value, so different average throughput can be expected in different environment.

3. The shorter the distance between the transmitter and receiver within the piconets, the higher the probability that it will receive packets correctly, which in turn increases the throughput.

4. By using the model for simulation, it has been shown that for the case in ACL DH1, DH3, DH5 packets, the simulated performance is 3% to 4% lower than the ideal case when interference is neglected, however, the performance drops by 14% to 30% when interference occurs.

From the results, it can be concluded that Bluetooth, in fact, has relatively good performance when compared to some access technologies like FDMA. However, the average throughput is varied greatly according to the environment with different obstacles and interferences.

10.7 PRACTICAL ILLUSTRATION: SENSOR NETWORK

Bluetooth has become a publicly accepted standard for short-range communication and has been included in many electronic devices as a complementary communication technology to infrared. More and more devices are integrated with Bluetooth technology, for example, PDA, laptop, mobile phone, printer, etc. In this section, we present a practical security system in which Bluetooth is used as a communication technology for assisting in building up a sensor network, which has attracted much researchers' attention.

There are three types of nodes involved in the system—sensor nodes, relay nodes and control node. A sensor node may be an embedded system which has the functions of detecting movement from outside environment and controlling the Bluetooth

module to transmit the movement information out. In fact, relay nodes and control node are also sensor nodes that carry extra functions. Relay nodes are used to route information from the sensor nodes, while the control node acts as an interface between the sensor network and local security processing system. The nodes in the Bluetooth sensor network build up a tree-based hierarchical structure with the control node as a root of the tree and relay nodes as intermediate nodes, while sensor nodes as leaves. By using this simple configuration, simple routing can be achieved by constantly pushing the information to the root. Because of this, intermediate nodes failure recovery and route maintenance can be done easily.

The whole scenario can be summarized in Figure 10.8 [66], and the details are discussed as below:

1. Sensor Node A detects movement.
2. Sensor Node A converts the motion event into digital data and controls the Bluetooth module to transmit the corresponding information out.
3. Sensor Node B receives the data from Sensor Node A and helps in routing the data to next hop (Sensor Node C) towards the control node. So in this case, Sensor Node B acts as a relay node.
4. Similar to Sensor Node B, Sensor Node C also acts as a relay node and forwards the information to Control Node D through Bluetooth.
5. Sensor Node D is a control node which connects to a local security processing system (e.g., a workstation). In receiving the movement information from Sensor Node A routed through Sensor Nodes B and C, security guards located inside the security room monitoring the local security processing system receive a warning message about the intrusion into the building.
6. The intrusion information may also be routed to remote/centralized security processing system through Internet. In this case, multiple geographically locations can be monitored and actions can be made in a centralized manner.

In the above case, a Bluetooth sensor network is used in a security system. In fact, this configuration can also be used in many scenarios. A good example is remote site rescue—a temporary rescue center is set up outside the rescue site and Bluetooth sensor nodes are distributed in the rescue site to provide information inside to the control node outside with a local processing system, which will then be able to carry the information to the centralized rescue office in remote location.

10.8 SUMMARY

In this chapter, Bluetooth has been discussed in detail. First of all, we discuss the design goals of Bluetooth—a short-range cable replacement with low cost and low power

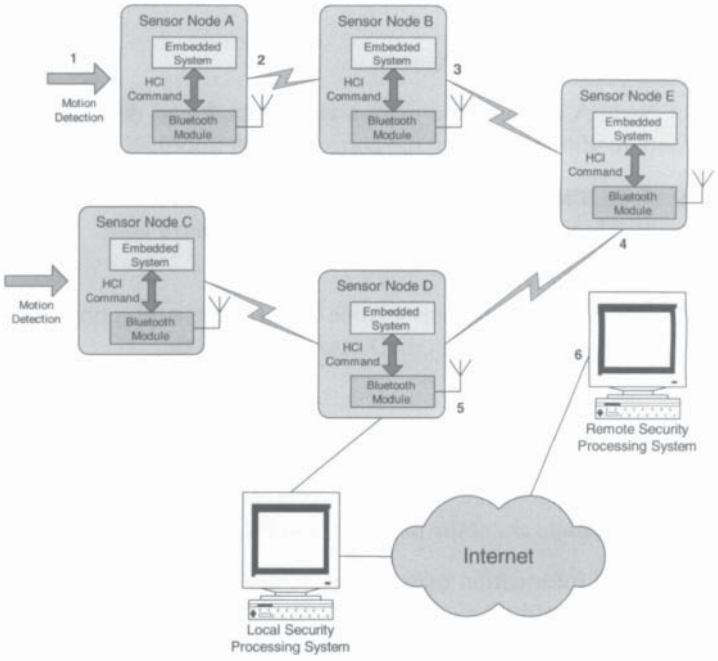


Figure 10.8 Bluetooth sensor network security system.

consumption. Then Bluetooth's protocol stack and profiles which aim at providing interoperability between devices from different vendors have been gone through in detail. Afterwards, Bluetooth PHY and MAC are discussed.

With the structural constraints of Bluetooth, Bluetooth single point-to-multipoint (piconets) and multiple point-to-multipoint (scatternets) have been discussed in detail with the focus on their formation, together with the possibility of QoS support. After that, we come to a performance study on Bluetooth MAC throughput which shows that Bluetooth, in fact, gives a good performance, but its performance may be unstable subject to environment conditions. Lastly a simple Bluetooth sensor network security system scenario is shown to demonstrate the practical usage of Bluetooth.

All in all, as for its original target, Bluetooth has been successful in providing a short-range ($\tilde{1}$0m) cable replacement solution. People start thinking about other potential usage models on Bluetooth, which brings out more and more extended profiles recently. It is expected that Bluetooth's future development will also follow IEEE 802.11x in the sense that they will be capable of carrying a larger bandwidth capacity and have a longer transmission range. However, with the development of UWB (Ultra Wideband), which provides bandwidths comparable to wired networks and also targets on WPAN, the future of Bluetooth is difficult to be predicted at this moment.

PROBLEMS

10.1 Bluetooth uses the same frequency band and provides similar functionalities as Wireless LAN, e.g., IEEE 802.11b/g. Briefly describe the design goals of each technology and explain why the products of these two technologies can co-exist in the market.

10.2 Briefly describe the differences between Wireless LAN MAC and Bluetooth MAC, and how they can be used to support QoS applications.

10.3 What are the differences between piconets and scatternets? Briefly describe different methods of formation of scatternets.

This Page Intentionally Left Blank

CHAPTER 11

COEXISTENCE ISSUES

The art of life lies in a constant readjustment to our surroundings.

—Kakuzo Okakura

11.1 INTRODUCTION

Low cost short range wireless communications bring us great convenience. But such convenience could be haunted by the ubiquity of such devices. In this chapter, we discuss about the coexistence problem involving Bluetooth and WLAN devices.

Interoperability:

As we have learnt from previous chapters, ISM band is mainly provided for the usage of short range wireless communications. Ideally, devices using this band should be interoperable with each other. As detailed in Section 11.2, the major issue is to withstand interference from other devices. However, the spread spectrum techniques adopted by current popular technologies, e.g., Bluetooth that uses FHSS and WLAN that uses DSSS, are designed based on specific conditions without regard to the

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

existence of devices using different short-range technologies. Indeed, these different ISM band short-range technologies operating in close proximity result in a serious packet collision problem. Section 11.3 will focus on this issue.

The approaches suggested in handling coexistence issues are described in Section 11.4. These solutions can be divided into two main categories: collaborative mechanisms and non-collaborative mechanisms. The collaborative solutions suggested provide interoperability among main interference technologies by colocating multiple (usually two) air interface solutions within the same physical unit. On the other hand, two main methods (AFH and MDMS) are suggested in the non-collaborative category. AFH approach involves changes in frequency hopping sequence. As a result, interoperability can only be supported among the devices that use AFH. On the other hand, the MDMS approach is based on skipping the usage of certain frequencies in an adaptive manner. Consequently, this approach only requires changes made on the master devices. Indeed, all other slaves devices can still interoperate without any modifications. AFH approach is discussed in detail in Section 11.4.2.3, Section 11.5 and Section 11.6. MDMS approach is discussed in detail in Section 11.7.

Performance:

From the simulation results reported in some studies, two main AFH approaches—TG2-AFH and ISOAFH—outperform traditional pseudo-random frequency hopping technique used in handling interference issue. Specifically, ISOAFH works well in resource and power limited environment. However, this is not the case for TG2-AFH which does not take into account resource and power constraints. On the other hand, in a performance study that involved TG2-AFH, ISOAFH and ISOMDMS, it is found that ISOAFH greatly outperforms other two mechanisms in handling interferences in all scenarios. Interestingly, ISOMDMS is able to produce better performance than TG2-AFH in some cases. This indicates that AFH may not be the only suitable technology in combating interference. We will further elaborate on these issues in Section 11.8.

11.2 THE ISM BAND SPECTRUM

In order to facilitate short range transmission technologies development, government has provided some unlicensed spectra specifically for this purpose. Unlicensed does not mean free from any regulations. In fact, the government has regulated the usage of these spectrum by specifying the maximum power transmitted, frequency band range, etc. The most notable example is ISM bands shown in Figure 11.1 [61].

ISM bands are mainly provided for wireless communication devices for commodity purposes. The devices using these bands do not need to pay license fee. Nevertheless, they must follow the power regulation and the usage of spread spectrum techniques in order to provide acceptable interference to others and survive from interference of others. Currently, there are three ISM bands. They are located at 900

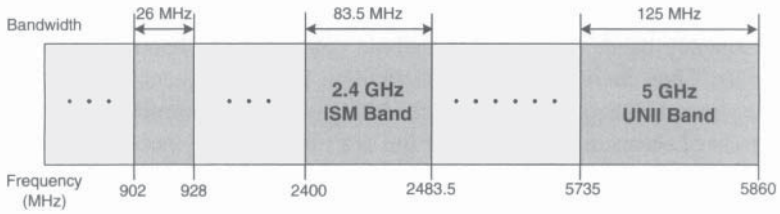


Figure 11.1 ISM bands in United States.

MHz, 2.4 GHz, and 5 GHz frequency regions as shown in Figure 11.1. The most notable one is 2.4 GHz ISM band which has been discussed briefly in Chapter 9 and Chapter 10. Indeed, 2.4 GHz ISM band is the frequency band used by many of the devices nowadays, including IEEE 802.11b/g, Bluetooth, Microwave Oven, cordless phone, etc. On the other hand, 5 GHz ISM band is less crowded currently, mainly used by IEEE 802.11a, HiperLAN2, etc.

The low cost and open usage of the ISM band inevitably leads to severe interference between different products integrated with different technologies using the same frequency band. This is further elaborated in Section 11.3 below.

11.3 PACKET COLLISION

In the previous section, several unlicensed bands have been introduced. Because any device can use unlicensed ISM bands free of charge as long as regulations defined by the government of specific countries have been followed, large number of products produced using these frequency bands lead to severe interference and packet collision. In this section, the general concept of packet collision is further elaborated [61].

The 2.4 GHz ISM band is one of the most commonly used unlicensed band nowadays. It has been shared by two popular kinds of devices—Bluetooth and IEEE 802.11b/g WLANs. As a result, these two types of device are two main sources of collision. First of all, let us briefly review some general characteristics of these two technologies here.

- IEEE 802.11b is a kind of short range wireless technology using 2.4 GHz frequency band. It divides the whole 2.4 GHz ISM band into channels with center frequency spacing of 5 MHz. Each channel takes up 22 MHz bandwidth in total. Under United States FCC regulations, there are 11 channels available in which only three of them can be used simultaneously. It uses Direct Sequence Spread Spectrum (DSSS) in physical layer. Consequently, its frequency band occupancy is relatively static and wide. Further details can be found in Section 9.5.

- Bluetooth is another kind of short range wireless technology using 2.4 GHz frequency band. It divides the whole 2.4 GHz ISM band in channels of 1 MHz. Thus, there are 79 channels in total. It uses Frequency Hopping Spread Spectrum (FHSS) in physical layer. A single piconet is formed by a master and a group of slaves in which they use the same hopping sequence to communicate. The frequencies used in the TDD time slots are predefined by this hopping sequence and specific to that piconet. Since master-slave communication is carried out using different frequency every time (hopping), its frequency band is relatively narrow and dynamic. The readers are referred to Section 10.4 for further details.

Generally speaking, packet transmission collision occurs only when both technologies use overlapping frequency in overlapping period of time. This concept illustrated in Figure 11.2. As can be seen, both Bluetooth and WLAN IEEE 802.11b devices are located within transmission range of others. We can also see that there are two packets collision occurred, namely the second and third WLAN packet transmissions, corresponding to third and seventh Bluetooth packet transmissions.

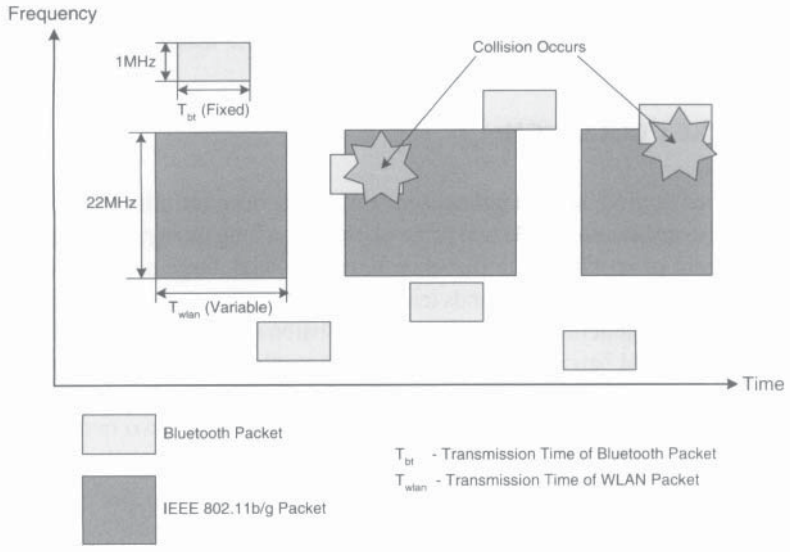


Figure 11.2 Packet collision example.

It should be noted that the severity of this collision, in fact, depends on many factors including the separation of devices, signal strength etc. Figure 11.2 shows that Bluetooth packets hop across different frequencies over different time slots. As a result, packet collision is only temporary. By using this technique, narrowband and fast-changing interferences can be coped with successfully. However, collision

occurs frequently as WLAN frequency band is relatively static and wide. On the other hand, WLAN uses DSSS to cope with collision. As long as the signal to interference ratio (SIR) is lower than certain threshold value, the interference problem can be ignored. This technique is best for handling background noise. However, this may not be a good solution for narrowband and fast-changing interferences.

As a result, both Bluetooth and IEEE 802.11b WLAN use their own methods in solving the problem of interference. However, the methods adopted may not be an effective solution as these schemes are not targeted at the main interference source. Furthermore, with increasing popularity of both technologies nowadays, the interference problem becomes more and more severe. Together with serving as a reference for less crowded 5 GHz ISM band which is expected to be frequently used sooner or later, many possible solutions are proposed and they are introduced in Section 11.4.

11.4 POSSIBLE SOLUTIONS

In Section 11.3, the concept of packet collision has been introduced. Specifically, packet collision is assumed to occur only when more than one short-range wireless technologies use overlapping frequency during the same period of time. In view of this, we can easily think of two possible methods to cope with this problem, namely, obtaining frequency orthogonality, and/or time orthogonality among different kinds of technologies.

There are two main groups of solutions using the methods suggested above. They are collaborative mechanisms and non-collaborative mechanisms. The classification of coexistence mechanisms is shown in Figure 11.3 [61]. Most of the mechanisms suggested involve modifications in Bluetooth WPAN side instead of IEEE 802.11b WLAN side.

11.4.1 Collaborative Mechanisms

Collaborative mechanism mainly requires the cooperation of the affected wireless technologies to minimize the mutual interferences. It can be further divided into collocated and non-collocated [61].

11.4.1.1 Collocated Mechanism Collocated mechanism refers to the cooperation of the affected wireless technologies which are implemented into a single physical unit. Since they are implemented into the same physical unit, within such a short separation distance, the power transmitted by either Bluetooth and IEEE 802.11 b/g WLAN will cause interference to each other once they occupy adjacent frequency in overlapping period of time. Moreover, the transmission power by either side may cause interference even if non-overlapping frequency is used. Consequently, only time orthogonality method can be adopted to prevent interference. Time division

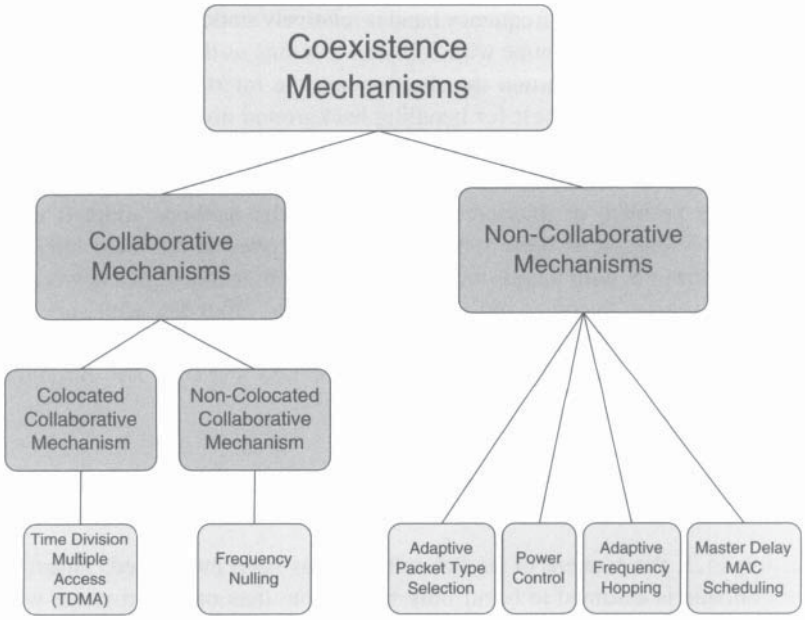


Figure 11.3 Various coexistence mechanisms.

multiple access (TDMA) technique is generally applied in this case. It is a kind of MAC layer solution which solve the problem by alternatively interleaving the transmission of Bluetooth and WLAN packets. In this case, either kind of packets will be transmitted at certain period of time and so coexistence can be achieved. The general concept of TDMA of multiple packets is shown in Figure 11.4.

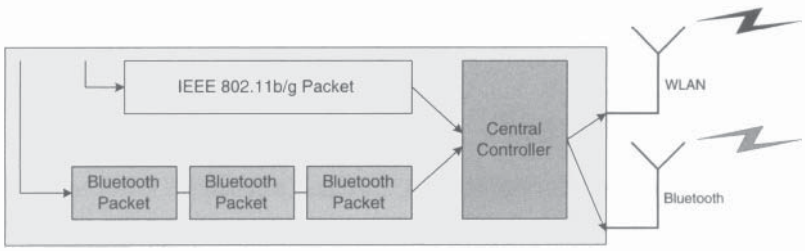


Figure 11.4 Collocated TDMA system conceptual diagram.

Some of the proposed TDMA approaches in collocated mechanism are summarized here.

- **Alternating Wireless Medium Access (AWMA) scheme.** This scheme simply uses a time division approach which divides a IEEE 802.11b Beacon interval into two sub-interval. They are Bluetooth sub-interval and IEEE 802.11 sub-interval. By doing so, alternating exclusive periods are allocated to both technologies to transmit their own packets.
- **MEHTA scheme.** This scheme requires the MAC layer information of both technologies. Whenever Bluetooth or IEEE 802.11b would like to transmit a packet, the transmission request is forwarded to a central controller (MEHTA) for approval. They can transmit their packets only when permission is granted. Permission is granted based on certain criteria including signal strength of the technology, the differences between the center frequencies of both technologies used, etc.

11.4.1.2 Non-collocated Meachanism By definition, non-collocated mechanism refers to the mechanism that both technologies are implemented into different physical units, but they are required to cooperate to achieve coexistence. This approach is not commonly used. A notable example is Frequency Nulling.

This mechanism requires the IEEE 802.11b device to be integrated with a Bluetooth receiver which is mainly used to snoop Bluetooth packets around. By doing so, the frequency hopping sequence of Bluetooth piconet around is then known. In this case, a 1 MHz wide frequency null can be inserted into the appropriate region of 22 MHz of IEEE 802.11 channel whenever a collision on frequency domain is expected.

11.4.2 Non-Collaborative Mechanisms

As indicated in Figure 11.3 [61], there are four main types of solutions in this category. They are adaptive packet selection, power control, adaptive frequency hopping, and master delay MAC scheduling. Non-collaborative mechanism refers to the coexistence mechanism which does not involve any information exchange between possibly interference sources (i.e., the nearby WPAN and WLAN). Since information exchange is not required, it is more applicable in real life environment. Indeed, non-collaborative mechanism is generally considered as a more important approach. Because explicit information exchange is not carried out, these approaches generally involve two stages—channel classification and adaptive control actions. This is further elaborated below.

11.4.2.1 adaptive packet selection mechanism Firstly, the adaptive packet selection mechanism is briefly described. In Section 10.4, different types of Bluetooth MAC layer frames are introduced. Specifically, they are frames occupying one slot, three slots and five slots and frames with different error detection and correction mechanisms—no FEC or 2/3 FEC. DH5 frame, which has no FEC and occupies five slots, can carry the largest amount of information with the least overhead. However, DH5 frame is more vulnerable to interferences as it occupies the channels with longer

time and is not equipped with any error correction mechanism. On the other hand, DM1 frame carries the least information with largest overhead.

In view of this, an idea is that Bluetooth should select appropriate packet types for transmission at specific channels. At the beginning, a Bluetooth device has to store statistics on channels that encounter interference. Adaptive control mechanisms can then be employed. Some example control mechanisms are presented below.

- A longer frame (with more slots) is used in good channel which will then cover the slots that are originally assigned to bad channels immediately after the good channels.
- The successful packet transmission probability is dynamically calculated. Shorter frame is used if the successful transmission probability is low. This mechanism is known as IBLUES.

11.4.2.2 Power Control mechanism In this mechanism, transmission power of Bluetooth can be limited at minimum level. At this level, only required performance and data integrity are maintained. By controlling the transmission power, all the interference sources nearby will experience reduced level of interference.

Thus, mechanism has been proposed for the channel detection process. The transmission power is maintained at minimum level with a margin. Accordingly, when data lost caused by propagation loss is detected, the transmission power is increased. By using such heuristic power adjustment, maximum data rate can be achieved by using minimum amount of power.

11.4.2.3 Adaptive Frequency Hopping Adaptive Frequency Hopping (AFH) is a non-collaborative mechanism generally accepted as a solution for solving coexistence problem in ISM band. Indeed, AFH has already been adopted in Bluetooth specification v1.2. Originally the frequency hopping sequence is generated randomly and the devices within the piconet use that generated sequence to communicate with each other, without regard to the fact that some channels in the sequence may be overlapping with other interference sources. However, in the case of Adaptive Frequency Hopping, whenever the frequency hops generated are occupied by some interference sources, the affected hops are swapped to some other frequency hops that are non-overlapping with others. The general principle of AFH is illustrated in Figure 11.5. Here, channel 10 of Bluetooth is selected to be used for communication within the piconet. However, this channel is also occupied by a interference source nearby. As a result, channel 10 is swapped to channel 40, which is non-overlapping with others during that time slot.

In general, AFH involves several processes. They are illustrated in Figure 11.6 and their corresponding functions are summarized below:

1. Device identification is mainly used to make sure that the devices involved support AFH mode. In other words, it is mainly added for backward compatibility.

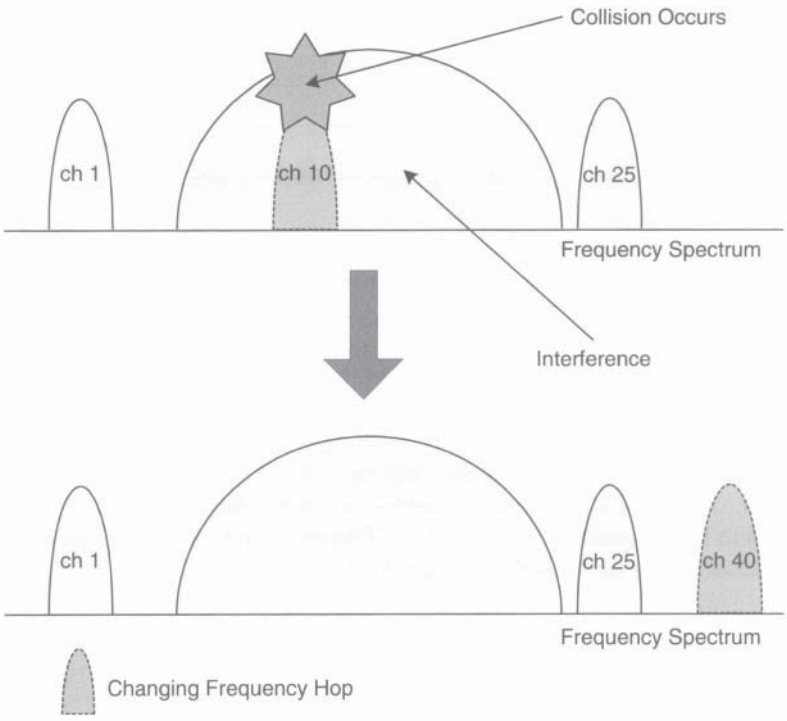


Figure 11.5 Adaptive Frequency Hopping: a conceptual illustration.

2. Channel Classification is mainly used to classify the quality of the channels.
3. Classification Information Exchange is a process for exchanging measurement information between master and slaves within the piconet.
4. Adaptive Frequency Hopping mechanism is the actual process which make changes to the frequency hops generated dynamically based on the information of the channels.

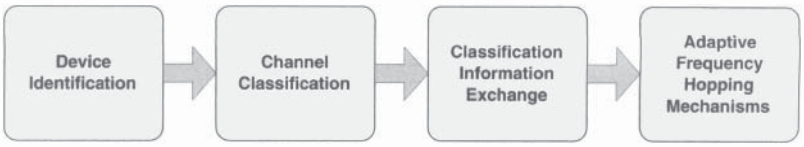


Figure 11.6 Adaptive Frequency Hopping processes.

Due to its demonstrated efficacy and practicality, AFH is generally accepted as a promising solution in avoiding packet collision. Indeed, AFH is adopted by IEEE 802.15 TG2 as a mechanism to tackle the issue. Variants of the AFH mechanism are further discussed in Section 11.5. In Section 11.6, we also describe an idea which further improves AFH solution suggested in IEEE 802.15 TG2.

11.4.2.4 Master Delay MAC Scheduling Master Delay MAC Scheduling (MDMS) is another non-collaborative mechanism in solving the problem of packet collision. The main idea behind is simple—whenever the frequency hop generated is found to be occupied by other sources. Transmission at that time slot is skipped. This idea is further elaborated in Section 11.7.

11.5 IEEE 802.15 TG2

The interference between WLAN (e.g., IEEE 802.11 b/g) and WPAN (e.g., Bluetooth) has caught significant attention. Bluetooth SIG and IEEE 802.15 then organized a task group to work on this issue specifically. The task group is known as coexistence Task Group 2 (TG2)[1]. TG2 is mainly responsible for quantifying the mutual interference between WLAN and WPAN. It is also responsible for developing a set of practices to facilitate combating the coexistence issue.

Specifically, IEEE 802.15 TG2 recommended AFH mechanism to be implemented in Bluetooth to avoid packet collision between WLAN and WPAN. This involves the appending an adaptive frequency hops selection scheme after the original Bluetooth frequency hops selection scheme as shown in Figure 11.7. Originally, the frequency

[1] The task group is now in hibernation until further notice.

hopping sequence is generated randomly with the use of parts of master BT_ADDR and its clock information. With the addition of new scheme, the sequence generated by the original one is then passed through a frequency remapping module which aims at modifying the hops inside the sequence that are classified as "bad" channels and maintaining the pseudo-random nature inside the sequence. The frequency remapping function makes use of the information from sequence generation module which stores the information of channels and decides when and which frequencies are to be used. After passing through remapping module, an adaptive frequency hopping sequence is formed.

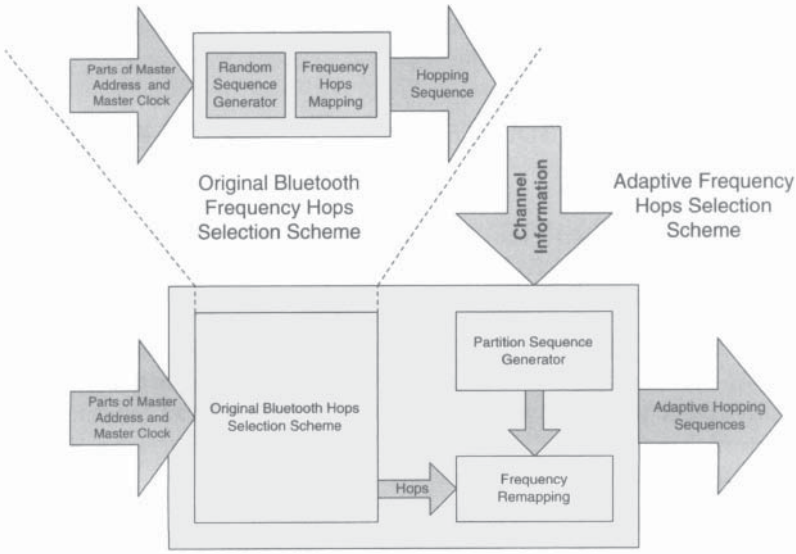


Figure 11.7 Structure of TG2 adopted AFH.

The two additional modules in adaptive frequency hops selection scheme performs two main functions in most of the non-collaborative mechanisms—channel classification and adaptive control actions.

In the channel classification process, TG2 mandatorily requires both the master and slaves to keep track of the channels conditions based on some channel assessment methods such as Bit Error Rate (BER), Packet Error Rate (PER), Packet Loss Rate (PLR) and Received Signal Strength Indicator (RSSI), etc. Generally speaking, it is common to use BER, PER and PLR for channel condition assessment.

After the measurements, the slaves then transmits the results back to master through an LMP message. Channel status is generated on channel-by-channel basis. After gathering the information from slaves, a list of "good" and "bad" channels are generated, each of which is then used to modify the frequency hops in adaptive control actions.

Let us consider the adaptive control actions on ACL links. In channel classification process, channels have been classified into "good" channels set and "bad" channels set. The frequency hops generated by using original Bluetooth frequency hops selection scheme are then compared against the elements in these two lists. Specifically, the frequency hop classified as a "bad" channel is substituted with a "good" channel by the remapping module. However, there is a regulation on the minimum number of frequency hops $N_{\min}$. According to FCC regulation in United States, $N_{\min}$ is 15. Thus, if the number of available "good" channels is larger than $N_{\min}$, some "bad" channels will be modified to one of the "good" channels in this process (Mode L operation); otherwise "bad" channel will continue to be used in order to comply with the regulation (Mode H operation). After remapping, the final sequence generated is called an adaptive frequency hopping sequence.

11.6 INTERFERENCE SOURCE ORIENTED ADAPTIVE FREQUENCY HOPPING

In Section 11.5, IEEE 802.15 TG2 adopted AFH mechanism in handling coexistence issue between WLAN and WPAN has been discussed in detail. However, this method still has room for improvement due to the following reasons.

1. TG2 adopted AFH mechanism assesses the channels on a channel-by-channel basis. However, it does not take into account the observation that if the interference source is WLAN, then the adjacent channels of "bad" channels actually have high probability of subjecting to interference also. This is because WLAN uses a wide frequency band statically.

2. Since channel assessment in TG2 AFH is done on a channel-by-channel basis, this generally results in a relatively longer response time from the presence of interference to the detection of interference and the response of appropriate actions. In other words, there is still room for improvement in channel classification mechanism.

According to above explanations, the improvements of TG2 AFH are usually focused on its deficiency in channel classification process. A recent attempt is an approach called Interference Source Oriented AFH (ISOAFH) [61]. This AFH mechanism mainly focuses on improving the channel classification process by identifying the interference source and analyzing its corresponding characteristics. Firstly, Table 11.1 gives a detailed comparison between IEEE 802.11b channel frequency allocation and that of Bluetooth.

From Table 11.1, we can see that if the interference source is IEEE 802.11b WLAN which makes use of channel 4, then the corresponding Bluetooth channels (channels 14–36) will be highly affected. With this observation, ISOAFH tries to locate the interference source (WLAN) channel(s) and then prevents Bluetooth frequency hops selection scheme to select the group of frequencies affected.

Table 11.1 Comparison of IEEE 802.11b WLAN and Bluetooth WPAN channel frequency allocation.

IEEE 802.11b channel Number (US)	Occupied frequency range (MHz)	Bluetooth channel Number
channel 1	2401 — 2423	channel 0 — 21
channel 2	2406 — 2428	channel 4 — 26
channel 3	2411 — 2433	channel 9 — 31
channel 4	2416 — 2438	channel 14 — 36
channel 5	2421 — 2443	channel 19 — 41
channel 6	2426 — 2448	channel 24 — 46
channel 7	2431 — 2453	channel 29 — 51
channel 8	2436 — 2458	channel 34 — 56
channel 9	2441 — 2463	channel 39 — 61
channel 10	2446 — 2468	channel 44 — 66
channel 11	2451 — 2473	channel 49 — 71

The channel classification process in ISOAFH first groups the Bluetooth channels according to IEEE 802.11b WLAN channels listed in Table 11.1. This results in 11 groups of Bluetooth channels. Instead of keeping track of every Bluetooth channel status, only the aggregate PER of these 11 groups is stored. These 11 groups include some overlapping Bluetooth channels, e.g., Bluetooth channel 10 is located in Group 1, 2 and 3. If a packet error is detected in Bluetooth channel 10, then Group 1, 2 and 3 PER will also increase. By doing so, the group corresponding to the interference source will have the highest PER and so the interference source channel can also be located. In this case, "good" groups set and "bad" groups set are formed.

In the associated Adaptive Control Actions process, the master informs its slaves in the piconet about the channel groups status by using modes of revocation that are numbered according to groups. In other words, if group 6 is marked as "bad" group, mode of revocation number 6 will be sent to each slave. In receiving the LMP message from master with revocation number, all slaves remap frequency hops in group 6 to other channels in "good" groups.

A critical system parameter, $N_{\min}$, which indicates the minimum number of hops involved in a sequence has been briefly discussed in Section 11.5. According to FCC regulation in United States, the number should be at least 15. On the other hand, the maximum number of orthogonal channels in IEEE 802.11b WLAN is three which occupies channel 1, 6 and 11 according to US regulations. With the interferences of these three 802.11b channels, the remaining unaffected Bluetooth channels include

channels 21–23, 47–49, and 73–78. Consequently, this minimum requirement can always be fulfilled.

11.7 INTERFERENCE SOURCE ORIENTED MASTER DELAY MAC SCHEDULING

We have discussed various AFH approaches in handling coexistence issue. AFH is an aggressive approach in that it aims at searching for another "good" channel to replace the "bad" channel. However, this approach has certain limitations. The main limitation is backward compatibility. In using AFH, it is necessary to make changes to all devices, including master and slaves. However, thousands or even millions of Bluetooth devices which does not support AFH have already been produced and are being used in the market. In view of this, MDMS approach is considered as a more practicable solution [61].

Let us first review the principles of MDMS. Specifically, MDMS is a non-collaborative coexistence mechanism which can also be divided into two main steps—channel classification and adaptive control actions. In channel classification process, it uses the method adopted by normal non-collaborative mechanism (same as the one shown in TG2-AFH). In adaptive control actions process, the master first obtains the "good" channels set and "bad" channels set. Afterwards, the master will check the sets against the existing hopping sequence in pair as master has to guarantee both uplink and downlink transmission in ACL connection. Whenever a "bad" channel is encountered, the master simply delays the transmission of the packet to specific slave until next "good" channel hop. The whole process is illustrated in Figure 11.8.

Since the whole process does not include any modification of the original hopping sequence and the master has the complete control on the usage of time slots, changes on slaves are unnecessary. Only master has to be changed with simple firmware upgrade to support MDMS. As a result, backward compatibility is supported by using this approach.

In view of the deficiency in TG2-AFH approach, the channel classification process in MDMS is also enhanced. The modification procedures are exactly the same as that described in Section 11.6. By using this method, the interference sources characteristics will be considered by grouping channels into 11 groups. The readers are referred to Section 11.6 for the details of the classification procedures. After classification of "good" groups of channels and "bad" groups of channels, the master applies the same techniques discussed above to skip the "bad" channels.

11.8 PERFORMANCE ISSUES

Non-collaborative approaches are more popular in tackling the problems of coexistence issues as they can be used readily in realistic application scenarios. Recently, a performance study of two main AFH approaches and traditional pseudo-random

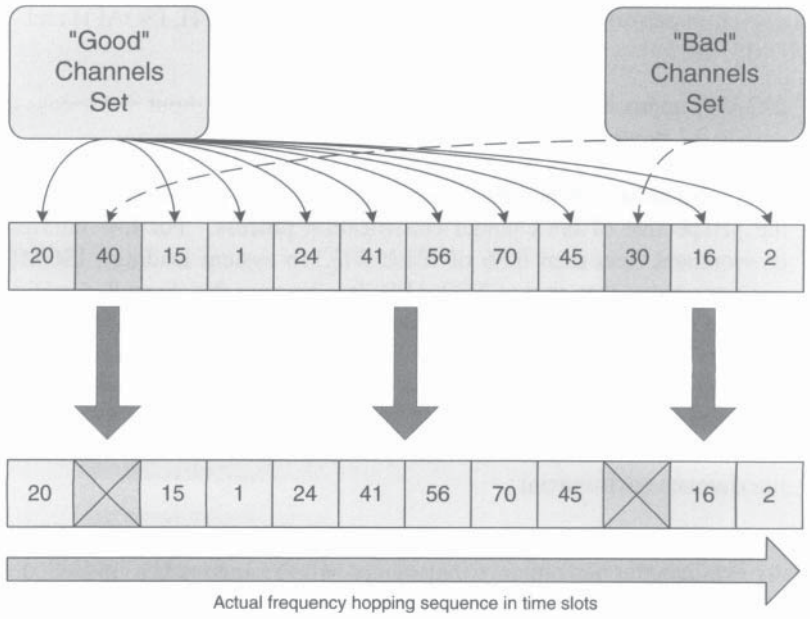


Figure 11.8 An illustration of the MDMS mechanism.

frequency hopping approach under limited resource and power environment has been carried out [61]. The study involved the three main non-collaborative approaches: TG2-AFH, ISOAFH and ISOMDMS.

Conclusions derived from the performance study [61] are summarized below.

- Both TG2-AFH and ISOAFH perform better than the traditional Pseudo-Random Frequency Hopping in the present of interference sources. The results are especially significant if the interference source is IEEE 802.11b WLAN. The main reason behind is that both AFH approaches obtain extra information in selecting frequencies for transmission.
- The performance of TG2-AFH mechanism is severely deteriorated in the presence of power and resource constraints while ISOAFH is less sensitive to these constraints. The main reason behind is that less information is concerned and stored in the case of ISOAFH (11 groups) than that of TG2-AFH (79 channels).
- AFH based techniques exhibit similar performance compared to that of pseudo-random frequency hopping techniques if the interference source is pure Bluetooth (i.e., pure "inter-piconet" interference) with varying number of piconets. It is because the narrow and fast-changing characteristics of interfering piconets affect the accuracy of channel and this greatly affects the ability of AFH based techniques to classify "good" and "bad" channels.

Moreover, in performance comparisons involving TG2-AFH, ISOAFH and ISOMDMS, the following observations were made.

- ISOAFH approach performs the best in terms of throughput and access delay among the three approaches considered.
- AFH mechanisms do not always perform better than MDMS mechanism from the perspective of the channel classification process. For low interference environment (less than 60% of IEEE 802.11b system loading), ISOMDMS performs better than that of TG2-AFH. It is because the channel classification mechanism of ISOMDMS allows it to response faster than that of TG2-AFH by skipping "bad" frequency channels in case of interference. However, when the interference becomes severe, ISOMDMS starts consecutively skipping "bad" frequency channels while TG2-AFH will swap "bad" frequency channels into "good" one. As a result, TG2-AFH performs better than ISOMDMS in high interference environment.

From the summary listed above, an overall conclusion is that AFH approaches generally obtain better performance than that of MDMS approaches. Indeed, despite AFH's implementation problems in existing devices, AFH should be the approach adopted for solving the problem of interference in long run.

11.9 PRACTICAL ILLUSTRATION: BLUETOOTH AND WI-FI COLOCATE TRANSMISSIONS

It is very common nowadays for owners to deploy IEEE 802.11b/g WLAN access point or Bluetooth access point in their coffee shops, shopping malls, etc., to act as internet access hotspot. Indeed, such facilities add values to the core services (e.g., coffee) purchased by the clients. However, due to interference problems, simultaneous deployment of both kinds of access points inevitably results in poor performance. Recently, a practical performance study [365] modeling a scenario using a collocated collaborative network access point has been carried out.

As discussed above, packet collision between Bluetooth and IEEE 802.11 b/g is a serious problem due to the popularity of these two technologies nowadays. Some access point manufacturers observe that it is actually possible to implement both Bluetooth and IEEE 802.11 b/g access point into a single device as shown in Figure 11.9. Here, both Bluetooth and IEEE 802.11 b/g WLAN client devices are served by the same network access point, which is equipped with dual air interface—Bluetooth and WLAN interfaces. As a result, Internet access service can be provided to both Bluetooth and WLAN clients.

In order to support dual air interfaces which are competing for frequency resources in nature, a collocated collaborative design and implementation detail are shown in Figure 11.10. This prototype system [365] is a PC that contains dual air interfaces

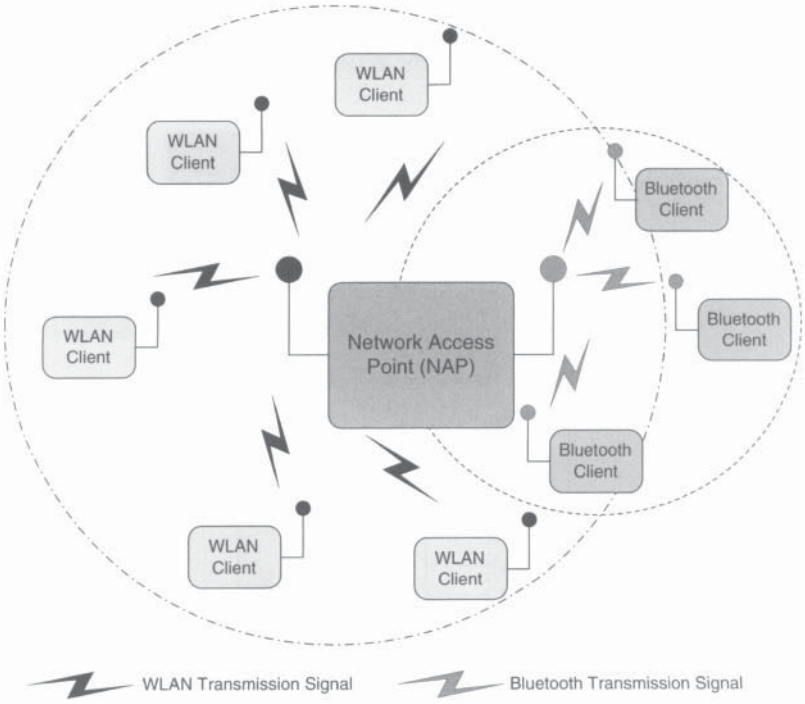


Figure 11.9 Dual air interface network access point scenario.

and operates in Linux environment. With the use of Linux operating system, each air interface will be associated with a software interface, i.e., bnep0 for Bluetooth and wlan0 for WLAN. This software interface can be used to control the packet traffics by assigning some rules in ingress, forward and egress points in this interface. As shown in the figure, the packets from both bnep0 and wlan0 are directed to a virtual interface imq0. Within imq0, the packets are queued up and packet scheduling algorithms can be applied to it. After applying the scheduling algorithm, the packet scheduled can then be transmitted by using the corresponding interface.

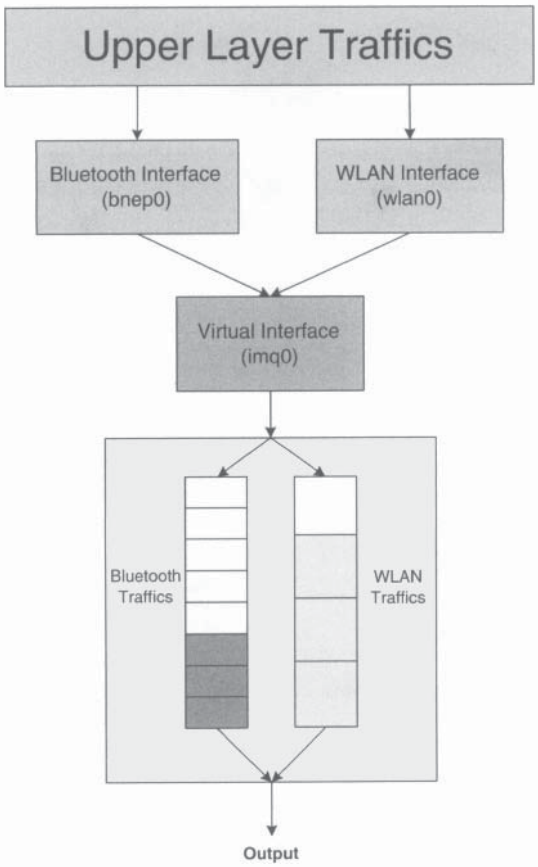


Figure 11.10 Bluetooth-WLAN network access point design and implementation.

Other than some common scheduling algorithms, the administrator of Bluetooth-WLAN NAP may also implement their own scheduling algorthms based on certain criteria such as Bandwidth Balancing, Round Robin Balancing, Interference Consideration, etc. The NAP keeps track of moving averages of the instantaneous packet

transmission rates of Bluetooth and WLAN in order to classify the conditions of the corresponding wireless channels. Specifically, there are three cases: increasing, decreasing and steady. For instance, a persistence decreasing of the transmission rate gives a strong indication of interference. In this case, the virtual interface will skip dequeuing of a packet from either interfaces. By doing so, further interference can be avoided.

In this section, we have described a Bluetooth-WLAN NAP scenario which can be used to support both Bluetooth and WLAN clients simultaneously. However, we have to emphasize that this is only one possible implementation of collocated collaborative coexistence mechanism. Indeed, vendors keep on developing their own standards and many products have been sold in the market already. Because of this, standardization becomes a pressing issue for interoperability among these different NAPs.

11.10 SUMMARY

In this chapter, we first describe in detail the most commonly used ISM bands. Then, packet collision problem in ISM bands has been discussed in detail, especially for the case of Bluetooth and IEEE 802.11b/g WLAN in 2.4 GHz ISM band. Since collision is mainly a result of overlapping frequency band usage in overlapping period of time, obtaining orthogonality in either frequency domain or time domain can effectively solve the problem.

Indeed, there are many proposed solutions in tackling the coexistence problem. Specifically, in view of the severity of the problem, IEEE 802.15 has setup TG2 to focus on this issue and proposed some recommended solutions. AFH based solution has been suggested and proposed by TG2 and this solution is also recently incorporated in Bluetooth specification v1.2.

However, researchers have also found htat the solution proposed by IEEE 802.15 TG2 may not be a good solution since it does not exploit the knowledge about the type of interference source. Due to this reason, some researchers have proposed modified versions of AFH, one of which is known as ISOAFH. Generally, these improved versions include identification of interference source and channel classification specifically designed for identifying the channel occupied by interference source.

Although AFH approaches generally obtain good performance in tackling interference problem, all the devices involved are required to cooperate in order to carry out the same AFH mechanism. In other words, legacy devices are unable to enjoy the advantage brought about by AFH mechanisms. In view of this, MDMS mechanism which focus on time domain based orthogonality has also been proposed. The main advantage of this mechanism is that only the master device is required to make changes by updating the firmware, and the slaves devices are unaffected. A modified version called ISOMDMS which includes channel classification modifications

has also been presented in this chapter. Finally, a practical scenario of collocated collaborative network access point has been covered and discussed in detail.

PROBLEMS

11.1 Briefly describe the original Bluetooth frequency hops selection scheme and how adaptive frequency hops selection scheme can be extended from it to solve the problem of coexistence interference.

11.2 Briefly describe the differences between ISOAFH and TG2-AFH. In particular, explain why ISOAFH could obtain better performance than TG2-AFH.

11.3 Briefly describe the differences between AFH and MDMS. Suggest some conditions that MDMS may obtain better performance.

CHAPTER 12

COMPETING TECHNOLOGIES

It takes all sorts to make a world.

—English Proverb, traced by Smith to 1620

12.1 INTRODUCTION

As short-range wireless communication is a lucrative market, there are many other similar technologies competing with Bluetooth and IEEE 802.11x WLAN. In this chapter we focus on three such competing technologies: IrDA, HomeRF, and HiperLAN.

Interoperability:

Interoperability is a main problem in the competing technologies for short-range wireless communications. One major competing technology that dominated the short-range realm is infrared. For using infrared as physical layer, there are standards from IEEE 802.11 and IrDA. Of the two technologies, IrDA is the one more commonly used. IrDA actually comprises several standards: IrDA-data, IrDA-control and the

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

most recently proposed AIr (Advanced Infrared). For data communications purpose, IrDA-data is the one that is widely used. Multiple standards in irfrared greatly hinder its interoperability. From an implementation perspective, different manufacturers tend to have their considerations on the transmission range due to reasons like potential hazards posed to the human eye due to reception of infrared with higher power. As a result, some manufacturers implement IrDA with a transmission range shorter than the suggested one meter. This issue also greatly hinders the interoperability. Details about the infrared technologies are further discussed in Section 12.2.

HomeRF is another major competing technology in the short-range wireless domain. HomeRF is a major competitor for IEEE 802.11 WLAN as they both use similar frequency band and provide similar services. Most notably, HomeRF adopts the idea of IFS in IEEE 802.11 WLAN but its IFS scheme is incompatible with that of IEEE 802.11 WLAN. This hinders the wide-spread acceptance of the HomeRF technology. Indeed, with greater support from manufacturers and better specifications with multiple task groups, IEEE 802.11 is expected to overshadow HomeRF. HomeRF becomes fading out in the market. The details of HomeRF are further discussed in Section 12.3.

Last but not the least, HiperLAN/2 standards developed in Europe becomes a main competitor for IEEE 802.11a. HiperLAN/2 is widely touted to have a better performance than that of IEEE 802.11a. One special feature is that it includes the consideration of integration with other wired network. However, these idea greatly increase the complexity of implementing the HiperLAN/2 technologies. Consequently, this leads to serious interoperability and implementation problems due to complex frame structure and tight time constraint. More about HiperLAN/2 is further discussed in Section 12.4.

Performance:

IrDA offers wireless communication range for at least 1 m with variable throughput capacity up to 16 Mbps by using different kinds of modulation schemes to support different transmission rates. Indeed, IrDA's throughput capacity is much higher than that achievable by its current main competitor—Bluetooth. The details about different modulation schemes used in order to attain these different transmission rates are elaborated in Section 12.2.4.

On the other hand, HomeRF achieves the wireless communication range for at least 50 m with variable throughput capacity up to 800 Kbps for voice communication, 1.6 Mbps for voice and data communication, and 10 Mbps for pure data communication. We provide a detailed discussion about this in Section 12.3.4. Specifically, we discuss about the different MAC schemes used for different traffic flows.

Finally, HiperLAN/2 achieves wireless communication range for around 150 m with variable throughput capacity similar to that of IEEE 802.11a. The throughput ranges from 6 Mbps up to 54 Mbps. This is further elaborated in Section 12.4.4.

12.2 IRDA

12.2.1 Design Goal

Infrared has been developed for more than ten years. Similar to the case of Bluetooth, it was developed in view of the great market demand for short range cable replacement solutions for portable devices. As a result, it is designed based on similar objectives as that of Bluetooth—Low Cost, Cable Replacement and Ad Hoc connectivity. The readers are referred to Section 10.2 for the design goals of Bluetooth. The design goals of infrared are summarized in Table 12.1 below.

Table 12.1 Design goals of infrared.

Low Cost	Infrared transmitter and receiver can be produced with the cost as low as US$2—3 due to its simple hardware architecture (one LED for transmitter and one photodiode for receiver). As a result, many portable devices can integrate this solution with negligible cost.
Cable Replacement	One of the main objectives of infrared is to provide short range cable replacement solution to portable devices—PDA, mobile phone, portable computers, etc. As a wide range of portable devices are already equipped with IrDA for communication, its objective as cable replacement solution is considered as successful.
Ad Hoc Connectivity	Infrared's design expects the users to to set up temporary connections only. As a result, only some simple applications like exchange of business card, exchange of picture, etc. are suggested in its original design.
Secure	Infrared does not need to adapt highly complex security measure since the communication between devices has limited range and requires line-of-sight.

12.2.2 Transmission techniques

There are three basic transmission techniques using infrared light—directed beam infrared, omnidirectional infrared and diffused infrared [323].

Directed beam infrared's transmission signal is focused. It can be used to create a point-to-point link between two infrared devices. Its range of transmission depends on many factors such as transmission power and degree of focusing. One of the example usages of this type of transmission technique is to build a token ring LAN as shown in Figure 12.1. Here, each IR transceiver is used to support a computer or a group of computers connected together by hubs. As a result, a token ring network is built.

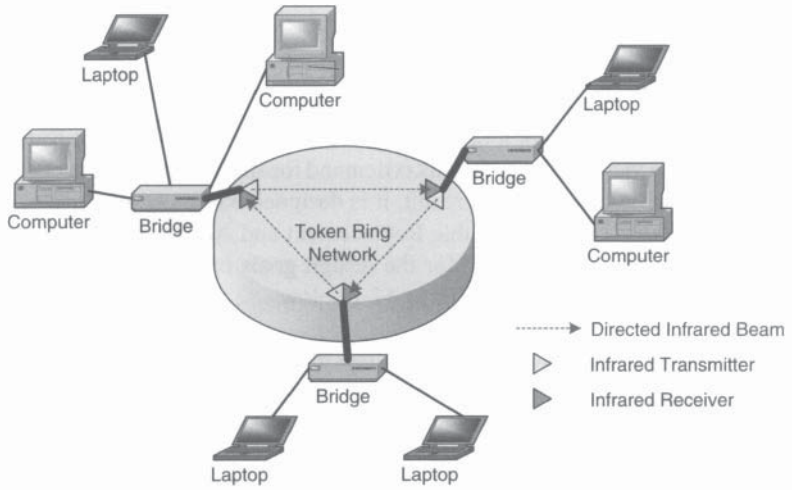


Figure 12.1 Directed beam infrared example.

Omnidirectional infrared generally involves a single access point located at the ceiling so as to provide coverage area large enough to cover the whole room. The access point is connected to an IR transceiver in which transmission signal is radiated omnidirectionally, so that receivers of stations within the coverage area are possible to receive the signals. On the other hand, the transmitters of other stations transmit directed signal pointed to the receiver of access point. The whole configuration is illustrated in Figure 12.2.

Figure 12.3 illustrates the operation of diffused infrared. Here, the transceiver attached to each station focuses at a point which is possible to reflect the IR signal. For example, the lightly painted ceiling in the room is such a target point. By doing so, the directed beam is reflected omnidirectionally by the ceiling and all the receivers within the area can receive the signal.

In the market nowadays, most devices are equipped with IrDA compatible infrared transceiver for data communications. As a result, most IR devices employ directed beam infrared type with point to point link transmission only. On the other hand, IEEE 802.11 PHY specifications also include infrared as a possible physical medium. Specifically, diffused infrared is specified. The readers are referred to Section 9.5 for more details about IEEE 802.11 infrared.

12.2.3 IrDA Protocol Stack

IrDA (Infrared Data Association), founded in June 1993, aims at providing an open standard for short range wireless infrared communication. IrDA was established in view of the fact that there are many manufacturers implementing their own standards

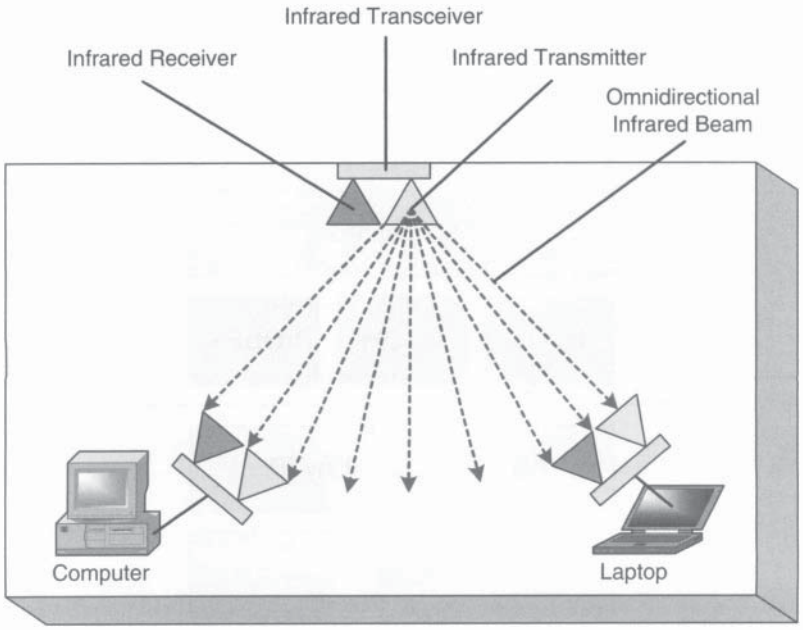


Figure 12.2 Omnidirectional infrared example.

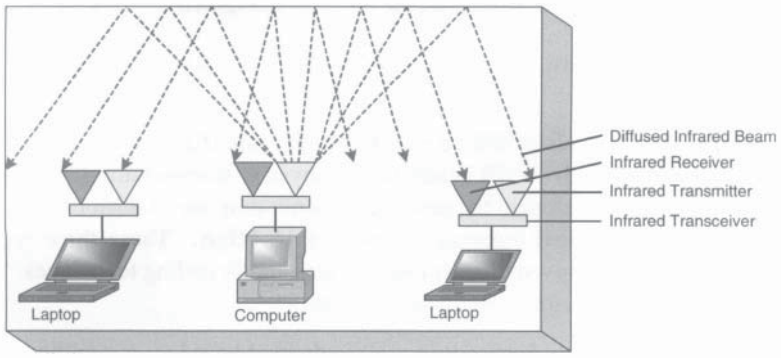


Figure 12.3 Diffused infrared example.

on infrared technology at that time. This leads to incompatibility among different infrared devices from different products. Indeed, even for IrDA standards on infrared, there are more than one standard, including IrDA-Data, IrDA-Control and AIr. In this section and the next, IrDA-Data communication protocol stack, its corresponding physical layer, and datalink layer are discussed in detail.

A simplified IrDA protocol stack is shown in Figure 12.4. Details about each protocol components are summarized below [349].

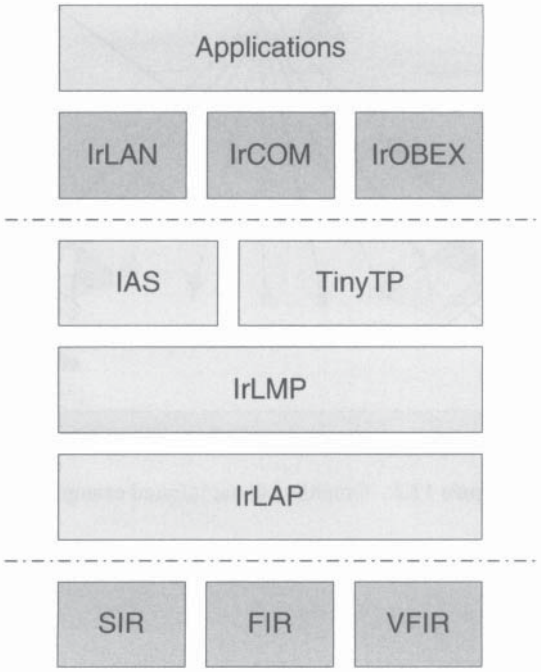


Figure 12.4 Simplified IrDA protocol stack.

- Physical layer involves SIR (Serial Infrared), FIR (Fast Infrared) and VFIR (Very Fast Infrared)[1]. SIR refers to the physical transmission rate less than 1.152 Mbps, FIR refers to the physical transmission rate of 4 Mbps, and VFIR refers to the physical transmission rate of 16 Mbps. These three types of physical layers involved different modulation and/or coding techniques. These will be further elaborated in the next section.
- IrLAP (Inrared Link Access Protocol) is defined based on a modified HDLC protocol. In general, it is responsible for device discovery, negotiation and

[1] SIR, FIR, VFIR are obsolete terms in specifying Infrared physical layer. However, it is used in this chapter for presentation purpose.

establishment of a physical connection. IrLAP allows establishment of a single reliable connection between exactly two devices with the description on its establishment and termination including the supported connection rates of both side. In IrLAP connections, there are one primary device and multiple secondary devices. Despite that in theory it is possible to have more than one IrLAP connections in a device (point to multi-point connection), practically most devices only support one single IrLAP connection (point to point connection).

- IrLMP (Infrared Link Management Protocol) is mainly used to multiplex multiple infrared applications/services into a IrLAP connection. As a result, multiple applications listening for incoming connections without interfering with others are allowed.
- TinyTP (Tiny Transport Protocol) is mainly used to provide two basic functions—flow control, segmentation and reassembly. Specifically, flow control can be provided on virtual service channels similar to that of TCP. On the other hand, a large packet can be divided into smaller packets which will then be reassembled at the receiver side.
- IAS (Information Access Service) is mainly used by the user application to advertise their services and query the services provided by the opposing hosts. Its function is similar to that of SDP in Bluetooth protocol stack. At the beginning, only the IAS service is present in the service record. When an application starts up, it registers itself to IAS. Then, a specific selector—LSAP (Link Service Access Point) Selector is given to record this application so as to be accessed by other devices.
- IrCOMM (Infrared Communication)'s functions are similar to that of RFCOMM described in Bluetooth protocol stack in Section 10.3. This protocol is mainly used to emulate RS-232 serial port via infrared. As a result, legacy serial port applications can be used between IrDA devices.
- IrOBEX (Infrared Object Exchange Protocol) is a simple protocol in which PUT and GET commands generally used in HTTP protocol, are defined for object exchange. By using this protocol, binary data object can be transferred between devices supporting this protocol.

12.2.4 IrDA Physical layer and Data Link layer

IrDA operates at around 850–950nm wavelength. As it has been discussed in previous section, IrDA involves multiple physical layer transmission rates with different modulation and/or coding techniques. Common to all transmission rates, IrDA standard requires the devices to work at least 1 meter with 15 to 30 degree half angle deflection (in cone shape) for both transmitter and receiver under daylight environment. There are three main types of physical layers—SIR, FIR and VFIR [3, 349].

SIR refers to the transmission rate up to 1.152 Mbps. It is further divided into lower rate with throughput up to 115.2 kbps and higher rate up to 1.152 Mbps. In lower transmission rate type, pulse modulation—RZI (Return to Zero Inverted), with $3/16$ length of the original duration of a bit (i.e., $3/16$ mark-to-space ratio) or a fixed length of 1.63 μs pulse is used. The data format of this type of transmission is inherited from UART frame with a start bit and stop bit between data bits. With the help of start bit, asynchronous transmission word can be used.

On the other hand, for high-speed transmissions, the pulse modulation used is RZI with $1/4$ mark-to-space ratio. The packet format is also a little bit different from that of lower rate in that two start words are transmitted at the beginning, followed by address word to indicate the target device, data, 16-bit CRC and a stop word. This type of transmission is synchronous in nature.

Figure 12.5 illustrates the pulse modulation schemes that is mentioned above. Here, RZI is compared with NRZ (Non-Return to Zero), which is analogous to data bits. NRZ is a pulse modulation scheme which uses a pulse period to represent data bit 1 and non-pulse period to represent data bit 0. At the same time, during the bit pulse period, the pulse does not return to zero position. RZI uses the opposite concept that a pulse represents data bit 0 and the pulse will return zero position during the bit period as shown in the figure.

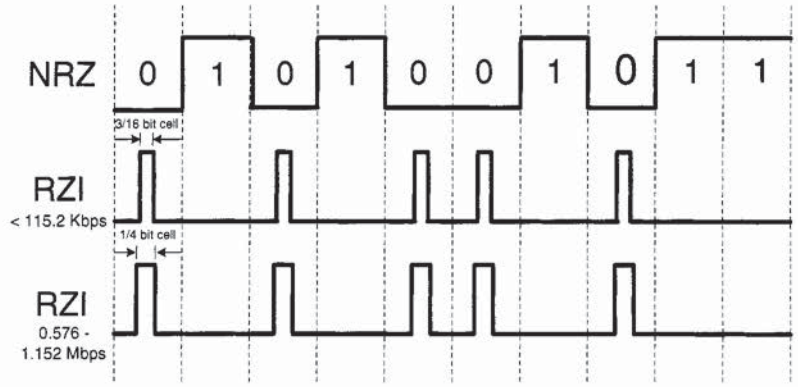


Figure 12.5 Pulse modulation of different SIR schemes.

Now let us consider FIR, which represents the infrared transmission rate of 4 Mbps. In order to achieve this transmission rate, 4 PPM modulation with $1/4$ mark-to-space ratio is used. Figure 12.6 illustrates the idea of 4 PPM. As can be seen, the original data bits are paired up. Two bits of data are encoded into one of the four position pulses as shown in the figure. That is, $00 \rightarrow 1000$, $01 \rightarrow 0100$, $10 \rightarrow 0010$ and $11 \rightarrow 0001$. In case of data bits pair 00, the LED flashes at the first position (a chip period) and then there will be nothing for the three remaining position. Unlike the

mechanism in SIR, FIR mechanism requires only one flash to represent two bits. This leads to two time faster speed for data transmission.

The packet format of FIR is a little bit different than that of SIR. In a packet, there is a permeable field existed at the beginning for receiver to establish phase lock. Then a start flag and information field are followed. Stop flag is the last. However, unlike that of SIR, FIR packet uses 32-bit CRC.

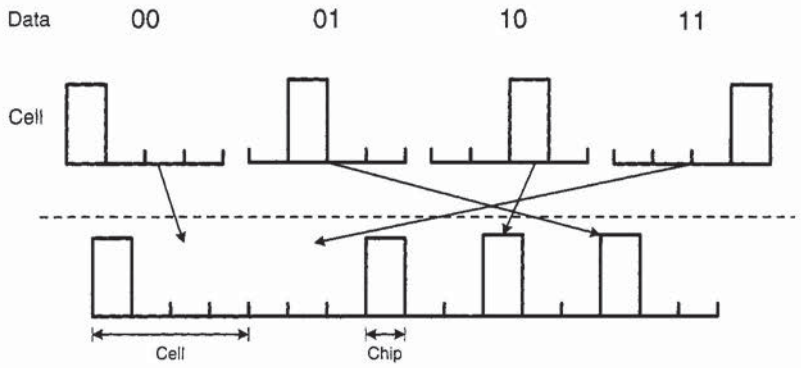


Figure 12.6 Four pulse position modulation scheme.

Finally, VFIR refers to the infrared transmission rate of 16 Mbps. It incorporates a new modulation code HHH (1, 13). The specified data rate is achieved by a low duty cycle, rate 2/3 and (1, 13) run-length limited (RLL) code. With (1, 13) RLL code, it guarantees that there is at least one space between each pulse and at most 13 spaces between each pulse. The chip clock is 24 Mchips/s. With the rate of 2/3, the IR transmission speed achieved is 16 Mbps.

The data encoding techniques of HHH(1, 13) are quite complicated. We provide a brief overview here. Firstly, there are 3-bit representing present state inside and 6-bit representing internal inputs. The state is first initialized to (1, 0, 0) and the internal inputs bits depend on initial input data bits pair. Depending on the input data bits pair, internal input bits and present state, next state is then determined with 3-bits output for transmission every time.

Let us move on to discuss the data link layer of infrared. As discussed earlier, the data link layer functionality of infrared can be regarded as provided by IrLAP protocol, which has the functions of device discovery and maintaining physical connection. Accordingly, it has two modes of operation—contention mode and connection mode. Contention mode is used in device discovery in which the transmission speed is fixed at 9600 bps. In this mode of operation, the device senses the media for at least 500 ms before sending out any query. This mode is mainly used to discover the devices within the visible region associated with their nicknames and MAC address. A MAC address in infrared is a 32-bit address randomly chosen by the device. If the device senses that there is an address conflict, an address conflict resolution procedures is

carried out in this mode of operation; otherwise, a connection will be established with thetarget device with a negotiated rate. IrLAP then changes to connection mode. Communication can then be carried out between two infrared devices with that rate.

12.3 HOMERF

12.3.1 Design Goal

The main objective of HomeRF is to provide an interoperable wireless networking environment with both local and remote (Internet and/or PSTN) contents including voice, data and, audio and/or video streaming, etc. In other words, it targets at integrating all the communication services at home with both wired and wireless environment by using HomeRF.

The development of HomeRF was coordinated by a working group[2] formed by companies including Siemens, Xilinx, AT&T, Motorola, etc.

12.3.2 Structure of HomeRF

There are two main types of network structure in HomeRF—Ad Hoc Networks and Managed Networks [126]. Before these two structures are discussed in detail, three types of node involved are introduced below. It should be noted that it is possible for a device to implement more than one service. As a result, some devices may become nodes of multiple types.

- A-node (Asynchronous node). This type of node is mainly used to transfer asynchronous data services including file transfer. A number of machines are regarded as this type of node such as PCs, PDAs, laptop computers, etc.
- S-node (Streaming node). This type of node is mainly used to receive or transfer some streaming data (asynchronous in nature) from or to other devices. Since this type of node requires a moderately strict QoS support, it should receive priority service than normal A-node traffic. Devices having streaming of audio and/or video applications such as audio headset, PC streaming of radio programme, etc., are regarded as this type.
- I-node (Isochronous node). This type of node is mainly used to support isochronous data service with strict QoS requirements such as a telephone call. A cordless phone is a typical example.

The idea of Ad Hoc Network topology in HomeRF is very similar to that of IEEE 802.11 WLAN. In this structure, there is no centralized control on the medium access. Peer communication is allowed by competing for the shared wireless medium. This

[2]This workgroup was disbanded in January 2003.

type of network only allows A-nodes. Figure 12.7 illustrates an example of this kind of network.

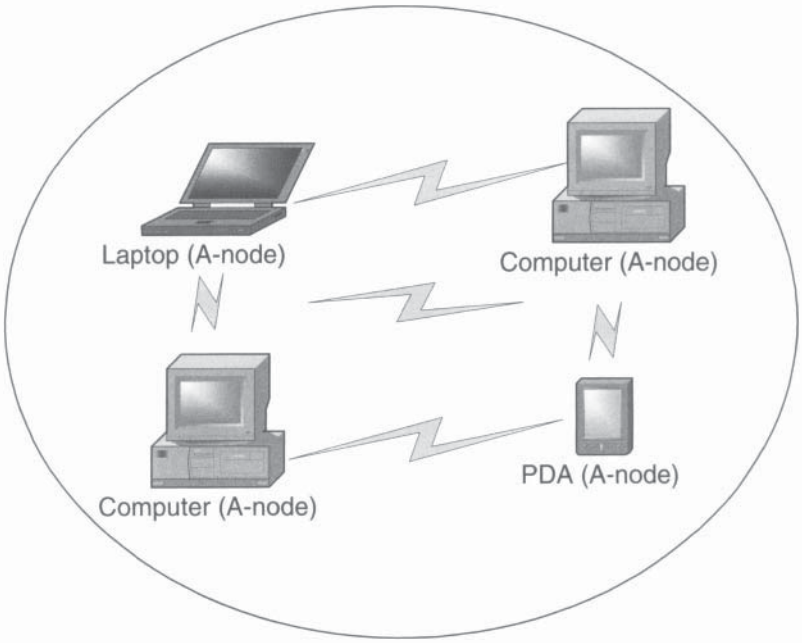


Figure 12.7 HomeRF ad hoc network example.

On the other hand, for a managed network, there will be a centralized control point, which manages the resources of the whole network by allowing A-nodes to exchange data among themselves, other wired networks, or the Internet. At the same time, it manages the S-nodes by assigning priority channel access to them through session setup process. It also manages the I-nodes by connecting them to PSTN telephone network and allocating dedicated time and bandwidth to maintain the conversation with specified QoS requirements. Figure 12.8 illustrates a managed network example.

12.3.3 Protocol Stack

Similar to all other technologies that have been mentioned, e.g., Bluetooth, IrDA, etc., HomeRF also has its own specialized protocol stack as shown in Figure 12.9. It supports three types of traffics: Asynchronous, Streaming and Isosynchronous [126]. In order to support different requirements of different traffic types, different MAC types are used. CSMA/CA is used for normal asynchronous traffic, CSMA/CA with priority access is used for streaming traffic as this such traffic requires a certain degree of QoS supports. Last but not the least, TDMA is used for isochronous traffic in order

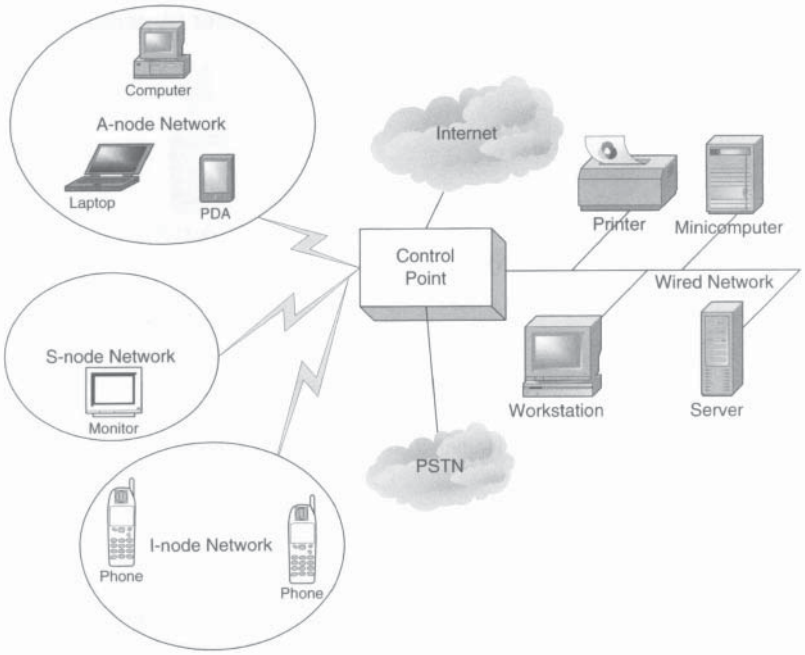


Figure 12.8 HomeRF managed network example.

to provide dedicated time and bandwidth for voice traffic. HomeRF MAC is further discussed in next section.

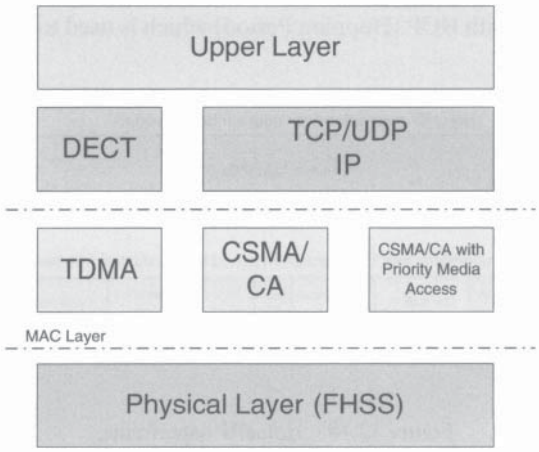


Figure 12.9 HomeRF protocol stack.

From Figure 12.9, we can see that TCP or UDP is used over IP for asynchronous and streaming traffic. However, the DECT (Digital Enhanced Cordless Telecommunications) which is a cordless phone standards in Europe is used for isochronous traffic. Due to the usage of DECT, the I-nodes in the network will be able to connect to normal PSTN network with the normal call setup.

12.3.4 HomeRF PHY and MAC

HomeRF operates at 2.4 GHz frequency band [126]. It chooses FHSS which has been discussed in detail in Bluetooth PHY in Section 10.4. Since it aims at supporting both data and voice communications, the channels division mechanism also depends on the communication types. For data and voice communication, frequency bands are divided into 75 channels with 1 MHz center frequency spacing. This results in 1.6 Mbps throughput rate. On the other hand, for pure data communication, 5 and 10 Mbps throughput rates are supported by dividing the frequency band into 15 channels with center spacing of 5 MHz each.

Meanwhile, certain adaptation mechanism is used in HomeRF to detect whether interference occurs. If interference is detected, it examines the hopping sequence so as to swap those hops within the interference region. The main objective of this mechanism is to prevent two consecutive hops from affecting by the interference source.

We can see that the PHY of HomeRF is very similar to that of Bluetooth. Now, let us consider the HomeRF MAC. Figure 12.10 illustrates the superframe of HomeRF

MAC. We can see that here are two main types of structures. The first one includes only contention period, while the second one includes both contention period and contention free period. A superframe lasts for a fixed period of 20 ms. Both types of superframe start with HOP (Hopping Period) which is used to change the hopping frequency used.

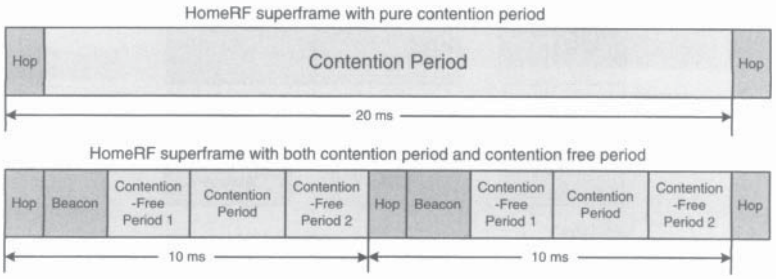


Figure 12.10 HomeRF superframe.

The contention period in both types of superframe are mainly used to transfer asynchronous traffics and streaming traffics. The standards define that the channel access method used in this period is CSMA/CA with priority access and time reservation. As a result, this period is mainly used by A-node and S-node for transferring data or streaming voice. The idea of contention resolution is very similar to that of DCF in IEEE 802.11 with the use of IFS and random backoff mechanism. Figure 12.11 illustrates the contention period with both asynchronous and streaming traffics. There are at most eight simultaneously streaming traffics exists. The CP assigns smaller stream slot number to the traffic which requires stricter QoS requirements. Recall IFS explanation in Section 9.4.3, the shorter the IFS, the higher the chance that the traffic can get control of the medium. In this case, the CF assigns the stricter QoS requirements traffic with shorter IFS (DIFS). As shown in Figure 12.11, the DIFS[3] of these eight streaming sessions is shorter than that of asynchronous traffics. Thus, priority access is given to these eight streaming traffics.

Now let us consider the contention free period. In order to provide a contention free period for isochronous traffics which are sensitive to delay and jitters, the CP has to provide a dedicated time slot and bandwidth. In view of this, HomeRF uses TDMA for this kind of traffic. Similar to that of streaming traffics, it allows eight simultaneous isochronous streams. Figure 12.12 illustrates an example of isochronous transmission in contention free period. We can see that there are two frames with each 10 ms. Within each frame, there are two contention free periods. The second contention period is mainly used to transmit the isochronous traffics as shown. On the other hand, the first contention period in the frame immediately after the previous

[3]DIFS in HomeRF is a little bit different from that of IEEE 802.11. The DIFS in HomeRF includes the contention windows time, while IEEE 802.11 does not include contention window time.

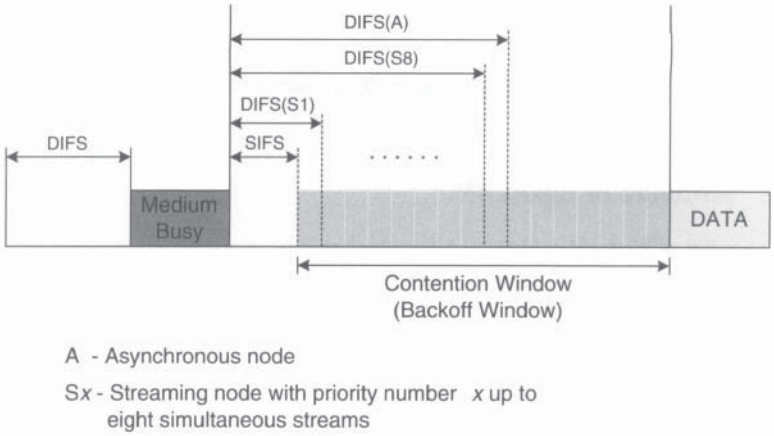


Figure 12.11 CSMA/CA with priority access.

frame is mainly used for retransmission of the lost traffics in previous contention free period. In other words, this contention free period may not occur if there is no loss in previous transmission. This approach supports high quality voice transmissions such as DECT, which requires strict delay and jitter bounds.

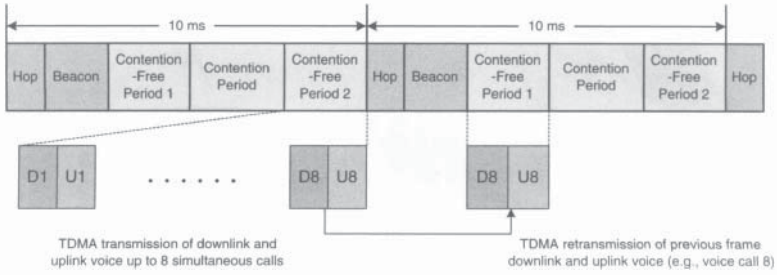


Figure 12.12 TDMA of HomeRF contention free period.

12.4 HIPERLAN

12.4.1 Design Goal

The main objective of HiperLAN is to provide a high performance WLAN in Europe which supports access to external network architecture such as Ethernet, ATM network, 3G network, etc. On the other hand, it also targets at both traditional data traffic together with multimedia traffics with certain QoS requirements.

In order to facilitate the development of HiperLAN[4], ETSI (European Telecommunications Standards Institute) Project BRAN (Broadband Radio Access Networks) was developed. ESTI works together with more than 50 companies to come up with the standards.

12.4.2 Structure of HiperLAN

Similar to that of HomeRF, HiperLAN is designed to support two main structures—Business Network and Residential Network [126, 284]. Let us first focus on the business network, which involves an access point acting as centralized control of the medium. All the traffics are routed through the access point. Figure 12.13 illustrates an example of business network model. This is also known as centralized mode of operation in HiperLAN.

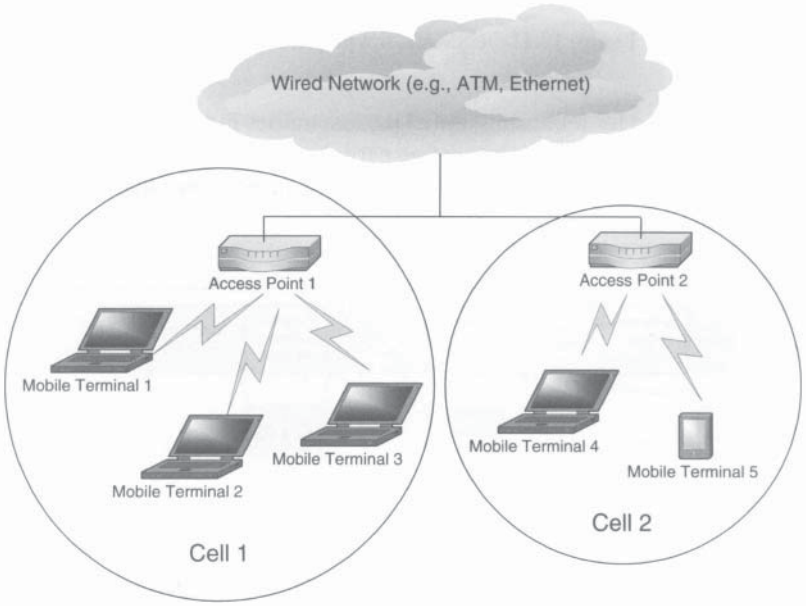


Figure 12.13 HiperLAN business network example.

On the other hand, there is another structure known as residential network. It is, in fact, an Ad Hoc network model. However, in this configuration, a CC (central controller) is needed to coordinate the medium usage among all the devices within the subnet. The CC can be dynamically selected from any Mobile Terminal. This is also known as direct mode of operation. Unlike centralized mode of operation, traffics

[4]Current version of HiperLAN is HiperLAN version 2 or HiperLAN/2. Throughout this chapter, HiperLAN refers to HiperLAN/2.

are sent in a peer-to-peer manner among mobile terminals. Figure 12.14 illustrates an example of this kind of structure.

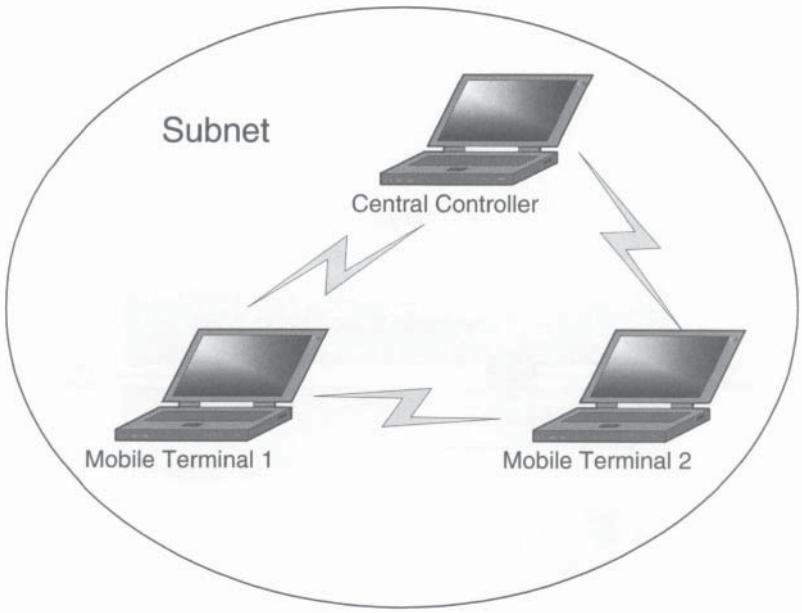


Figure 12.14 HiperLAN residential network example.

12.4.3 Protocol Stack

HiperLAN/2's protocol stack is shown in Figure 12.15 [108, 126, 284]. Here, it can be seen that HiperLAN/2 protocol stack actually composes of three layers—CL (Convergence Layer), DLC (Data Link Control) Layer including MAC and RLC, etc. and PHY (Physical) layer. The functionalities of PHY and DLC layer are elaborated in detail in the next section.

The convergence layer is a major component that enables HiperLAN/2 to provide interface between wireless mobile terminal and wired network or core network such as Ethernet, ATM, 3G network, etc. Convergence layer acts as an interface which performs packet transformation from fixed or variable size packet from upper layer to fixed size packet of proper priority class transmission in HiperLAN/2 network and vice versa. Essentially, convergence layer performs the functions of segmentation and reassembly (SAR) for packets.

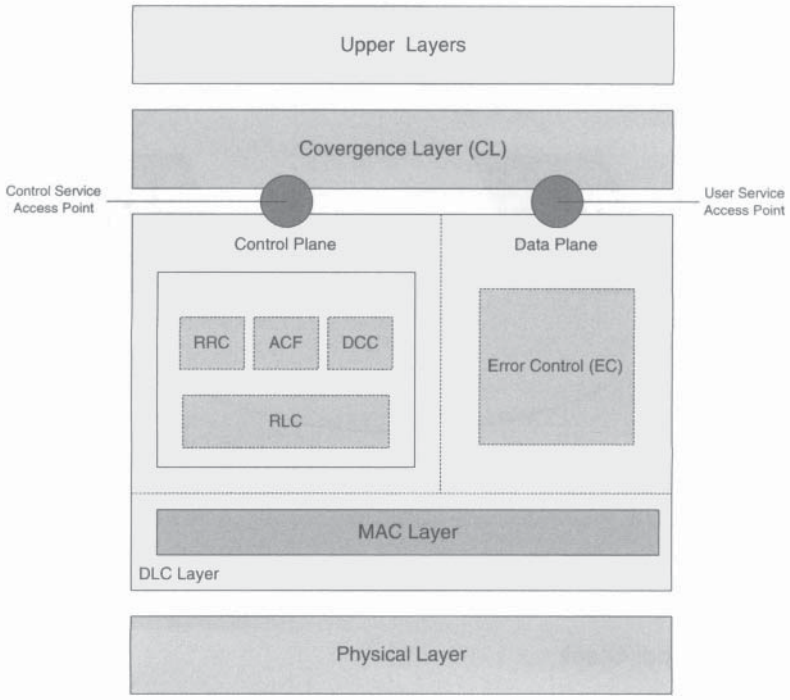


Figure 12.15 HiperLAN/2 protocol stack.

12.4.4 HiperLAN/2 PHY and DLC

In this section, we discuss about the detailed features of HiperLAN/2 PHY and DLC layer [108, 126, 284]. Let us consider the PHY layer first. The PHY layer in HiperLAN/2 is very similar to that of IEEE 802.11a WLAN. It operates at 5 GHz ISM band with OFDM. Since OFDM is adopted, it is required to operate into channels of 20 MHz center frequency spacing. Modulation schemes like BPSK, QPSK, 16-QAM and 64-QAM are also adopted into OFDM in order to provide transmission rate of 6 Mbps up to theoretical maximum of 54 Mbps.

Unlike that of IEEE 802.11a, dynamic frequency selection (DFS) scheme is adopted into HiperLAN/2. This scheme allows AP to choose the frequency based on the consideration of interference source. At the same time, P/CC can adjust their transmission rates by changing the modulation scheme used to adopt the channel conditions.

DLC in HiperLAN/2 is quite complicated. Indeed, the DLC structure is one of the main features of HiperLAN/2 that differentiates it from other technologies. As shown earlier in Figure 12.15, in the protocol stack, there are three main DLC components: RLC (Radio Link Control), EC (Error Control), and MAC layer. RLC and EC functions are summarized below.

- RLC implements the control plane of DLC and is mainly used to manage the use of wireless radio networks and exchange the control data between nodes. Multiple instances are present in AP with each of them represents a MT. However, only one instance exists in a MT. It consists of three main functions as shown in Figure 12.15.

  - DCC (DLC connection control) function is mainly used to manage the establishment, maintenance and termination of a connection between MT and AP. The establishment is managed by the set-up procedures which is usually initiated by a MT with the QoS parameters provided if necessary. Connection release procedures are also provided for termination of the connection.

  - ACF (Association Control Function) is mainly used to provide any necessary functions in association including, authentication, encryption, association, dissociation, link capabilities information exchanges, etc. As it has been suggested in Section 12.4.2, the wireless medium is under centralized control by AP/CC no matter in direct mode of operation or centralized mode of operation. As a result, association is a necessary function for every MT accessing HiperLAN/2 network.

  - RRC (Radio Resource Control) includes four main functions including dynamic frequency selection (DFS), handover, station availability check, power saving and control. These functions are necessary for efficient and effective use of the shared wireless medium.

- EC (Error Control) is implemented into the data plane of DLC for detecting and correcting (if possible) the error of the data frames sent and received. There are three modes of operation.
  - Acknowledged mode uses retransmission techniques to recover the unacknowledged packets. The Automatic Retransmission Request (ARQ) scheme used is selective repeat with variable window size. By doing so, reliable transmission is provided.
  - Repetition mode transmits arbitrary number of times of the same packets in consecutive order. It is unnecessary to inform the transmitter with acknowledgement. By multiple transmission of the same packet, reliable transmission can be provided with high probability.
  - Unacknowledged mode does not use any mechanism to retransmit the packets in case of error. As a result, the transmission is unreliable.

MAC layer [108, 126, 284] in HiperLAN/2 is quite complicated compared to that of IEEE 802.11 WLAN. It adopts Time Division Duplex (TDD) with Time Division Multiple Access mechanisms (TDMA). Thus, different time slots are assigned for downlink, direct link and uplink. The wireless medium is controlled by AP/CC. Each frame has a duration of 2 ms. Figure 12.16 illustrates an example of HiperLAN/2 MAC frame. As can be seen, some of the transport channel components of the MAC frame have also been shown. These includes Broadcast Channel (BCH), Frame Channel (FCH), Access Feedback Channel (ACH), Random Access Channel (RCH) and also the location of the three phrases—downlink, directlink and uplink phrases in the frame.

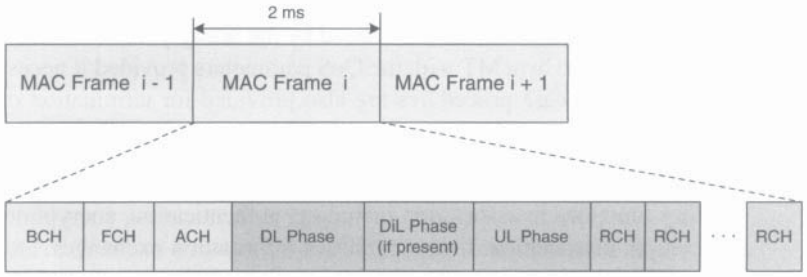


Figure 12.16 HiperLAN/2 MAC frame example.

In the MAC frame construction process, transport channel represents the message format in carrying the information from the logical channel. On the other hand, logical channel is mainly used to represent the types of message that a particular channel carry. In order words, the message (either control message or data message) is first mapped to an appropriate logical channel which will then be mapped to different transport

channels in the MAC frame for transmission. There are altogether ten logical channels and six transport channels. Their corresponding mappings in downlink (from the AP/CC to MT), direct link (peer to peer transmission between MTs) and uplink (from MT to AP/CC) are shown in Figure 12.17.

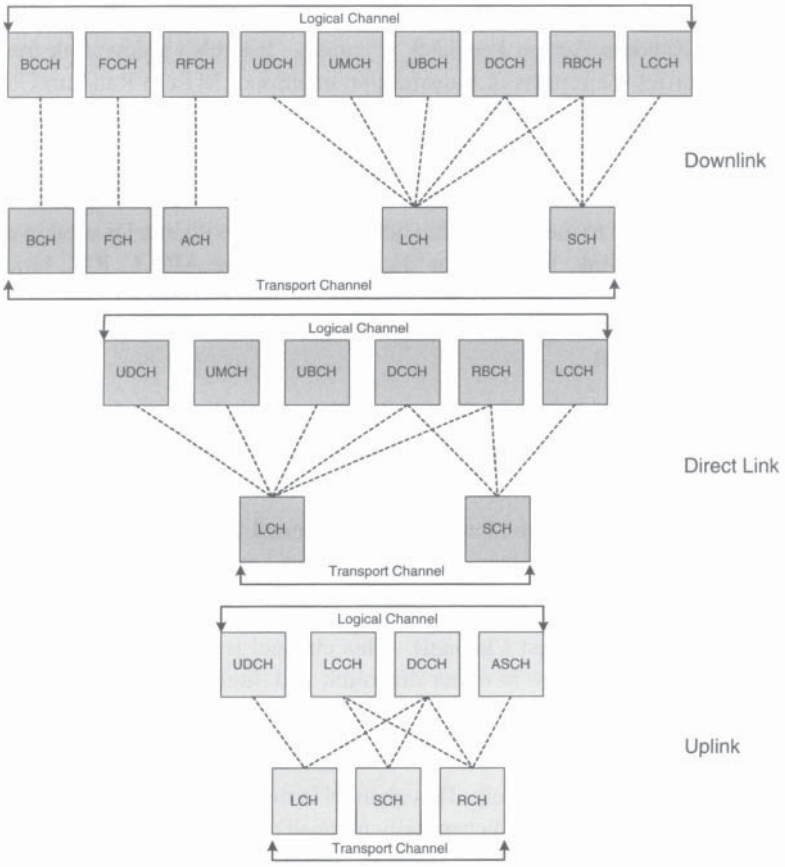


Figure 12.17 HiperLAN/2 logical and transport channels mapping in downlink, direct link and uplink phrases.

The corresponding functions of these altogether 16 channels are summarized below. First of all, let us start with the 10 logical channels.

- BCCH (Broadcast Control Channel). It carries downlink broadcast control message related to the whole radio network to all nodes in the network. It is mapped to Broadcast Channel (BCH) in transport channel.

- FCCH (Frame Control Channel). This is also a kind of downlink message to the MTs in the network. This channel carries the information related to the frame structure. This channel is also known as resource grants. It carries information about how many LCHs and SCHs a particular MT can get for transmitting and receiving data. This channel is mapped to Frame Channel (FCH) in transport channel.

- RFCH (Random Access Feedback Channel). It is also a downlink message which carries the reports of transmission attempt of MTs in Random Channel (RCH) in previous MAC frame. This channel is mapped to Access Feedback channel (ACH) in transport channel.

- RBCH (RLC Broadcast Channel). This channel is used to broadcast control message related to the whole radio network cell. This channel is used in downlink and direct link. Information like, MAC ID of the AP/CC, RLC broadcast message, some encryption related information and convergence layer ID information can use this channel. This channel is mapped to LCH and/or SCH in transport channel.

- DCCH (Dedicated Control Channel). This channel is mainly used to carry RLC messages in either downlink, uplink and direct link in LCH, SCH or RCH transport channel.

- UBCH (User Broadcast Channel). This channel is mainly used to carry the user broadcast message from the Convergence Layer (CL) in either downlink and direct link. It is mapped into LCH transport channel.

- UMCH (User Multicast Channel). This channel is mainly used to carry the user multicast message in either downlink and direct link. It is also mapped into LCH transport channel. Together with UBCH, this channel can either use repetition or unacknowledged mode for error control.

- UDCH (User Data Channel). This channel is mainly used to carry user data in all the links mentioned. It is generally mapped to LCH transport channel. This channel is granted for a connection described in FCCH.

- LCCH (Link Control Channel). This channel is mainly used to carry error control information such as ARQ results in all the link types mentioned. However, it is specifically used to carry the resource request (RR) message in uplink direction. As a result, this channel is mapped into SCH or RCH transport channel.

- ASCH (Association Control Channel). This channel is mainly used by MT to transmit association or handover message to an AP/CC in case it is not associated to any. This message is mainly transport through RCH transport channel.

The six transport channels are introduced below.

- BCH (Broadcast Channel). This channel is mainly used to carry information related to the whole radio cell including information like AP ID, power transmitted, etc. It transmits message in downlink direction.

- FCH (Frame Channel). This channel is mainly used to carry information related to the MAC frame structure. In other words, it transmits information on the allocation of the resources. Similar to BCH, this channel is also transmitted in downlink direction.

- ACH (Access Feedback channel). This channel is mainly used to carry information and results related to the access attempts in RCH of previous frame. Similar to the previous two channels, this one is also transmitted in downlink direction.

- LCH (Long Transport Channel). This channel is mainly used to carry user data and control information. This channel can be used in both direction.

- SCH (Short Transport Channel). This channel is mainly used to carry relatively short control information. This channel can be used in both direction.

- RCH (Random Access Channel). This channel is mainly used to give MTs which are not allocated for any SCH an opportunity to transmit their control information to AP/CC. This information can be resource request (RR), association message, etc. Consequently, this channel is mainly used in uplink direction.

12.4.5 QoS consideration in HiperLAN/2

The complicate DLC layer of HiperLAN/2 has been mentioned in previous section. In fact, this complicated structure allows HiperLAN/2 to support per-flow QoS [126, 284]. HiperLAN/2 adopts a centralized mode of operation in its design with an AP/CC assigned to perform the functions of allocating resources. Bandwidth allocation is performed by signaling: resource request and resource grants. Firstly, MTs send out the resource request indicating the characteristics of the traffic flows to AP/CC. AP/CC will then grant MTs with the numbers of packets which is allowed to send based on bandwidth resources availability, traffic characteristics, etc., through resource grant message. This centralized mode of operation effectively supports per-flow QoS in HiperLAN/2 with the help of signaling and admission control.

12.5 PRACTICAL ILLUSTRATIONS—IMPLEMENTATION OF HIPERLAN

DLC layer in HiperLAN/2 allows it to achieve high performance with QoS considerations. However, in the market, IEEE 802.11a products are widely accepted. One of the main reason is that HiperLAN/2 design is highly complicated due to the reason of complex DLC and restricted time constraint. In this section, we describe a practical implementation of HiperLAN/2 on a FPGA board to demonstrate the difficulties in the implementation of HiperLAN/2 [337].

Figure 12.18 illustrates HiperLAN/2 MAC protocol together with the implementation idea and division of software and hardware module. Such a division is necessary, as the restricted time constraint does not allow all the protocol processing to be done in software. In particular, hardware speed is needed for some of the time critical functions like parsing of broadcast message, CRC error checking, etc. In other words, hardware module is adopted to reduce the time of performing some time critical tasks.

As shown in Figure 12.18, the implementation of DLC involves a software scheduler for management of the available resources. The scheduler on AP involves generation of allocation map of a frame with information from traffic table which stores information related to QoS agreement with connected MTs. The execution time of the resource grants in the frame is used to sort the mappings. Afterwards, broadcast control frames are formed and stored. All these will be done by software parts in AP. On the other hand, the scheduler in MT will be required to map the resource grants carried in the broadcast control message in the traffic table. After the scheduler has finished its jobs, the software part generates a memory address list which contains the memory addresses of the connection list that is going to be sent out in current frame. These are the main functions done by the software modules in AP and MT HiperLAN/2 implementation.

As shown in Figure 12.18, the remaining parts of DLC and PHY layer are implemented in hardware logics in a FPGA board. The host controller interface is implemented by PCI controller which acts as a communication interface between the software module and hardware module. The most difficult part of hardware module design and implementation involves the implementation of frame processor in AP and MT. The frame processor needs to carry out many functions including the management of the air interface, the generation of acknowledgement map for the RCH field of previous frame, terminals synchronization, etc.

Since the time for executing certain events need to be precise, time descriptor which contains information about when and what events needed to be carried out are proposed. The generation of this description is done by parsing the broadcast message which involves information about the whole frame. The slot number suggested inside the time descriptor is used to match that of MAC frame slot. In order to have a precise matching which results in synchronization of AP and MT, a 40 MHz clock with slots exactly equal to the MAC frame slot is used after BCH is sensed.

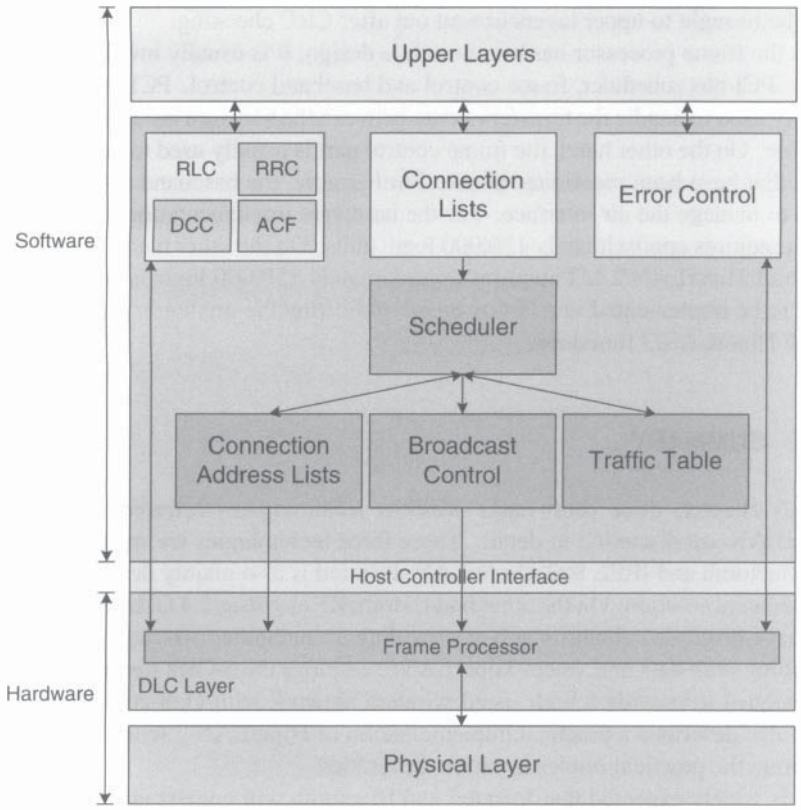


Figure 12.18 HiperLAN/2 MAC protocol architecture.

Moreover, ACH field in the current frame which is used to gives acknowledgment to the RCH in previous frame. As the RCH requires very fast response time, it is also implemented in frame processor. This is mainly done by handling the CRC checking of the received packets in the frame processor. As a result, an error occurs in RCH results in a CRC error which is detected by hardware module. The hardware module can then produce appropriate acknowledgment mappings accordingly. On the other hand, there is latency introduced by PCI bus and baseband modules. Thus, different queues are used to store the process data for either uplink or downlink. The data can then be brought to upper layers or sent out after CRC checking.

In the frame processor hardware module design, it is usually involves three main parts: PCI bus scheduler, frame control and baseband control. PCI bus scheduler is mainly used to handle the transfer of data between the software module and hardware module. On the other hand, the frame control part is usually used to handle the functions that have been mentioned above. Furthermore, the baseband control is mainly used to manage the air interface. For the hardware implementation of HiperLAN/2 AP, it requires approximately 170,000 logic units. On the other hand, the implementation of HiperLAN/2 MT requires approximately 150,000 logic units. These parts have to be implemented in a FPGA board, indicating the implementation difficulties of the HiperLAN/2 functions.

12.6 SUMMARY

In this chapter, three short-range wireless technologies—Infrared, HomeRF and HiperLAN are discussed in detail. These three technologies are major competitors for Bluetooth and IEEE 802.11x WLAN. Infrared is also mainly designed as a cable replacement solution. On the other hand, HomeRF also uses 2.4 GHz frequency band. We have discussed about its aim at providing an integrated wireless home network including both data and voice. HiperLAN/2 employs the 5 GHz frequency band and is designed to provide a high speed wireless network with QoS consideration. We have also described a practical implementation of HiperLAN/2 with FPGA board to illustrate the practical implementation difficulties.

It is widely expected that Infrared and Bluetooth will coexist in the market. It is because they are competing yet complementary in nature. The main advantages of IrDA are: it is easy to use, it comes with low implementation cost, it does not need complicated security. On the other hand, HomeRF, which occupies 2.4 GHz band, has very similar MAC layer with that of IEEE 802.11. Since this standard does not really provide many advantages over that of IEEE 802.11 and they are competing in many aspects, it seems that HomeRF does not have a competitive edge. Indeed, HomeRF working group was disbanded in January 2003. For the case of HiperLAN/2, its initial ideas of supporting multiple network architectures and incorporating QoS consideration are good. However, as illustrated in the practical illustration section, it requires both hardware and software support for precise time synchronization,

framing support in order to build such a complex system. As a result, it is believed that IEEE 802.11 would drive out HiperLAN/2 due to its distributed approach and relatively less complicated design.

PROBLEMS

12.1 Briefly discuss the similarities and the differences between IrDA and Bluetooth in their protocol stack and some current usage models. Explain why these two technologies simultaneously exists in market with both competing and complementary effects.

12.2 Briefly discuss the PHY and MAC of HomeRF. Explain why this technology fades out in the market.

12.3 HiperLAN/2 is expected to be a competing technology of IEEE 802.11a WLAN. Briefly discuss the advantages and disadvantages of HiperLAN. Explain why people expect that IEEE 802.11 would drive out HiperLAN/2.

This Page Intentionally Left Blank

PART IV

PROTOCOL ADAPTATIONS FOR WIRELESS NETWORKING

This Page Intentionally Left Blank

CHAPTER 13

MOBILE IP

Home is home, be it never so homely.

—English Proverb, traced by Apperson to c. 1300

13.1 INTRODUCTION

The Internet extends its reach into the wireless realm with the help of various protocol adaptation mechanisms. For one thing, the location and addressing of hosts are among the most important issues. In this chapter, we discuss about a mature mobility management adaptation framework for making wired IP interoperable with its wireless counterpart.

Interoperability:

Due to the advancement of wireless technology, there is a proliferation of mobile devices which are equipped with wireless interface such as WLAN or Bluetooth. Because these devices are mobile, it is very likely that the users may carry them from one Internet access point (typically wireless access point) to another Internet access

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

point. Thus, it is beneficial to keep the connection when they move across different domains, as if they were connected to a wireline connection all the time. Mobile IP is designed to fulfill this goal. In Mobile IP, a mobile host uses some advertisement mechanisms to discover its home agents and foreign agents. When it moves from its original home network to a foreign network, it gets a care-of address from the foreign agent and registers the address to it home agent. After that, all packets destined to the mobile host will be intercepted by the home agent and then are sent through a tunnel to foreign agent for delivery to the mobile host. This makes the mobile host get connected wherever it goes. Interoperability is manifested here as a transparent connection migration with the help of different agents across multiple networks.

Performance:

Using Mobile IP, when a mobile host roams across different domains, it can get uninterrupted Internet services. However, this comes with a cost—there is overhead induced in the discovery and registration process. For example, if the mobility of the device is high, the registration process may cause excessive signalling traffic and a long service delay may be incurred. On the other hand, when an Internet host sends packets to the mobile host, since it is not aware of the location of the mobile host, all packets will be sent to the home agent first, regardless of whether the mobile host is close to it. All these request for efficient schemes for enhancing the performance of Mobile IP. Indeed, from a performance perspective, much work is still being done in optimizing the connection migration process.

13.2 ADVERTISEMENT MECHANISMS

Advertisement mechanisms comprise several protocols and procedures for a mobile host to discover its home agent when it is at its home network and foreign agent when it is away from home. Thus, effectively, an agent discovery process is carried out. Due to its similarity to the router advertisement process defined in Internet Control Message Protocol (ICMP), the agent discovery process in Mobile IP makes use of ICMP router advertisement messages by attaching special extensions to the standard ICMP messages.

As shown in Figure 13.1, the home agent and the foreign agent periodically broadcast advertisement messages. Specifically, at the home network, there might also be a foreign agent for serving other visiting mobile hosts (similarly, there could be a home agent at the foreign network). Thus, when a mobile host receives an agent advertisement message, it must compare the network portion of its home network address with the network portion of the advertising agent's IP address. When these portions match, the mobile host knows that it is at the home network, and it can receive IP datagrams with forwarding by the home agent to the foreign agent; otherwise, the mobile host is at the foreign network, and it needs to get a care-of address so as to receive IP datagrams forwarded by the home agent.

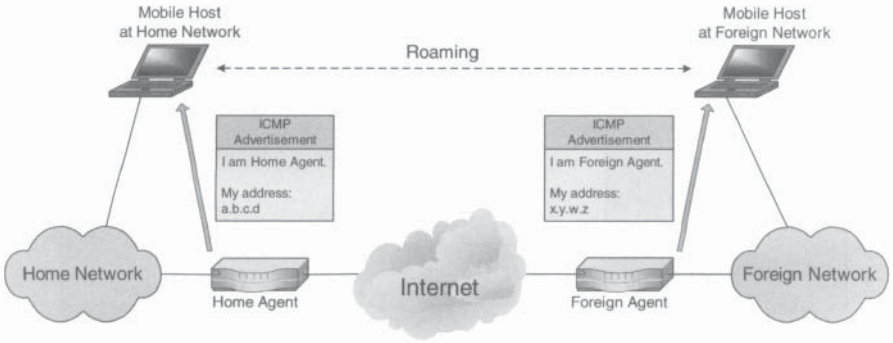


Figure 13.1 Agent advertisement.

From the illustration shown in Figure 13.1, we can see that the agent advertising process provides a mechanism for the mobile host to determine its point of attachment. When the mobile host is roaming, handoff from one network to another one occurs at the physical layer, without notification to the IP network layer. Consequently, the agent advertising process must be a continuous process, where the home and foreign agents need to periodically broadcast advertisement messages.

As mentioned before, the mobile host needs to get a care-of address when it discovers that it is at a foreign network. Now, let us see how it acquires such an address by utilizing the advertisement message received from the foreign agent. Figure 13.2 shows the ICMP router advertisement with the agent advertisement extension, and Table 13.1 summarizes the description of the various fields in the agent advertisement extension.

In Figure 13.2, it also shows an optional prefix-length extension succeeding the agent advertisement extension. The objective of this extension is to indicate the number of bits in the router's address that define the network number, allowing the mobile host to compare the network portion of its home IP address with that of the router's IP address. The value of the type field of the extension is 19. The prefix length field is the number of leading bits that define the network number of the corresponding router address listed in the ICMP router advertisement portion. Finally, the length field is the number of such prefix lengths.

From the agent advertisement message received, if the mobile host recognizes that it is on a foreign network, it will acquire a care-of address advertised in the agent advertisement message. After that, it will carry out the registration process to tell its home agent its care-of address. But before we go on to discuss the registration process, there are few points that are worth discussing:

Agent Solicitation: If a mobile host needs the information of the agents on its network immediately, it is allowed to send an ICMP router solicitation mes-

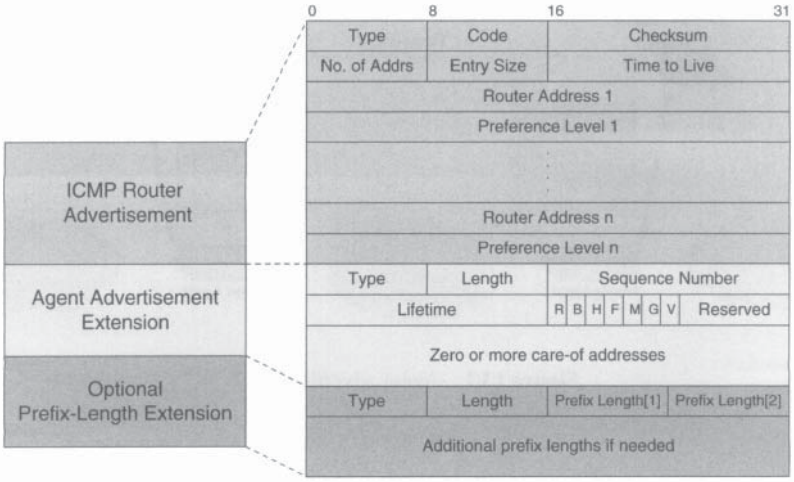


Figure 13.2 Agent advertisement message.

sage so as to elicit an agent advertisement message. Any agent receiving the solicitation message will issue an agent advertisement message.

Collocated Care-Of Address: Besides obtaining a care-of address form its foreign agent, a mobile host can alternatively acquire a care-of address from other independent services such as DHCP. This situation happens in some cases where there is no foreign agent available in the visiting network or all the foreign agents are busy. Specifically, the mobile host can obtain an IP address for its network interface via DHCP, and then use the address as the care-of address, which is known as a collocated care-of address.

Movement Detection: As mentioned earlier, by comparing the network portion of its home address and that of the advertising agent's address, a mobile host can determine whether it is on the home network or it has moved to a foreign network. In the other aspect, the mobile host can detect movement by using the lifetime field of the agent advertisement message. Specifically, when an agent advertisement message is received from an agent, the mobile host records the lifetime as a timer. Later on, if the timer expires before it can receive another advertisement message from the same agent, it can assume that contact has been lost with the agent.

Table 13.1 Description of the fields in the agent advertisement extension.

Field Name	Description
Type	Value 16, indicating that this is an agent advertisement.
Length	The length of the extension, which depends on how many care-of addresses are being advertised.
Sequence Number	The count of agent advertisement messages sent since the agent was initialized.
Lifetime	The maximum lifetime, in seconds, that this agent will accept a registration request from a mobile host.
R	Registration with this foreign agent (or another foreign agent on this network) is required. Mobile nodes that have already acquired a care-of address from this foreign agent must re-register.
B	The foreign agent is busy and will not accept any more registration. This bit is set to prevent existing customers from thinking that the foreign agent had crashed, and moving away unnecessarily.
H	The agent is a home agent.
F	The agent is a foreign agent. Note that bits F and H are not mutually exclusive, and that B cannot be set unless F is also set.
M	This agent supports minimal encapsulation (RFC 2004).
G	This agent supports GRE encapsulation (RFC 1701).
V	This agent supports Van Jacobson header compression (RFC 1144).
Care-Of Address	The care-of address(es) supported by this agent. There must be at least one care-of address if the bit is set.

13.3 REGISTRATION

After a mobile host recognizes that it is on a foreign network and then acquires a care-of address, it needs to inform its home agent about its care-of address to request the home agent to forward its IP datagrams to the care-of address. In addition, it also needs to tell the home agent how long it wants to use the care-of address and indicate special features that are available from the foreign agent. All these are done by carrying out a registration process. Two types of messages are defined in the registration process, they are registration request message and registration reply message, both are sent using UDP. Figure 13.3 depicts the registration process, and the details are described as follows:

1. The mobile host sends the registration request message to the foreign agent to ask it to pass the message to the home agent.

2. The foreign agent relays the request message to the home agent.

3. The home agent decides whether to accept or reject the request from the mobile host, and sends the decision as a registration reply message to the foreign agent.
4. The foreign agent passes the reply message to the mobile host.

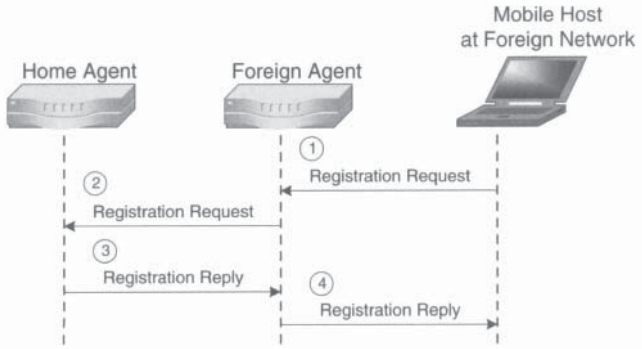


Figure 13.3 The registration process.

Note that during the registration process the foreign agent is considered a passive agent which agrees to pass the mobile host's request to the home agent and then pass back the reply from the home agent to the mobile host. On the other hand, if the mobile node is using a collocated care-of address, it will directly register with its home agent, without the intervention of a foreign agent.

Now, let us investigate the two types of registration messages, which are carried in UDP segments. The packet formats of the registration request message and the registration reply message are shown respectively in Figures 13.4 and 13.5, while Tables 13.2 and 13.3 describe the fields of the corresponding messages.

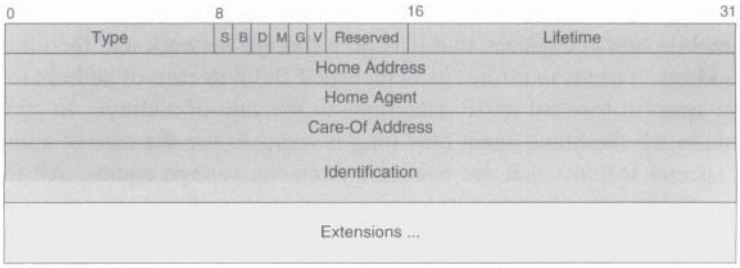


Figure 13.4 Mobile IP registration request message.

Security Measures in the Registration Procedure

The mobile host relies on the registration messages to inform the home agent of its care-of address so as to receive IP datagrams forwarded by the home agent. This

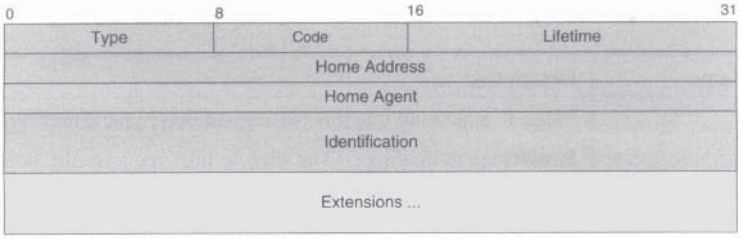


Figure 13.5 Mobile IP registration reply message.

makes it critical to to detect and reject fraudulent registration messages to prevent a malicious user from disrupting the communication between the home agent and the mobile host in the following cases:

1. A malicious user replays old registration request messages to the home agent to cut existing connection between the home agent and the mobile host.

2. A malicious user pretends to be a foreign user to send a registration request message (with the malicious use's IP address being the care-of address) to the home agent in order to intercept traffic destined to the mobile host.

In Mobile IP, the measures applied to protest against the above interruptions are based on the use of the identification field in the registration request and reply messages, and the use of authentication extensions. The identification field is basically a timestamp or newly generated random number (a nonce, a **n**umber used **once**), which makes each registration message different. Specifically, the home agent and the mobile host have to agree on reasonable values for the timestamp or nonce chosen, and resynchronization should be allowed by using the registration reply code 133 as shown in Table 13.4 By using the identification field, a registration reply can be matched to a registration request, and therefore suspicious registration messages (e.g., replayed messages) can be rejected.

The authentication extensions, which are appended to the registration request and reply messages, are used to authenticate the sender and receiver in the message sending between the home agent and the mobile host. Specifically, they are used to protect the fields of a registration message, particularly the important identification field for fighting against replay attacks. In the registration process, despite that the foreign agent acts as a relay to pass messages between the home agent and the mobile host, authentication extensions can also be used for authentication between the home agent and the foreign agent, and between the foreign agent and the mobile host. Thus, there are totally three types of authentication extensions defined in Mobile IP.

Figure 13.6 describes the format of an authentication extension. The type field indicates the type of authentication extension, while the length field indicates the size of the remainder of the authentication extension. The security parameter index (SPI)

Table 13.2 Description of the fields in the registration request messages.

Field Name	Description
Type	Value 1, indicating that this is a registration request message.
S	Simultaneous bindings. The mobile host requests the home agent to retain its previous binding, i.e., the association of the its home address with its care-of address. Using simultaneous bindings, the home agent will forward a datagram to each of the care-of addresses that are registered. In effect, multiple copies of a datagram will be sent to the mobile host to increase reliability in situations like wireless handoff.
B	Broadcast datagrams. It is used to tell the home agent to encapsulate broadcast datagrams from the home network to care-of address.
D	Decapsulation by mobile host. The mobile host is using a collocated care-of address and thus will decapsulate by itself the tunnelled datagrams.
M	It tells the home agent to use minimal encapsulation (RFC 2004).
G	It tells the home agent to use GRE encapsulation (RFC 1701).
V	It tells the home agent to use Van Jacobson header compression (RFC 1144).
Lifetime	The number of seconds that the mobile host requests the home agent to honor the registration. A value of zero is used for de-registration.
Home Address	The home address of the mobile host. This is for telling the home agent to receive datagrams destined to this home address and then forward to the care-of address.
Home Agent	The address of the mobile host's home agent. This tells the foreign agent to which this registration request message should send to.
Care-Of Address	The care-of address of the mobile host. It tells the home agent to which the datagrams of the mobile host should forward.
Identification	A 64-bit number used for replay protection, discussed below.
Extensions	Additional extensions that are required, such as the authentication extension, discussed below.

is an index which identifies a security context between the sender and the receiver. Specifically, it selects the authentication algorithm, mode, and secret (a shared key or public/prviate key pair) used to compute the authenticator in the authenticator field. The authenticator is a code generated by the sender to authenticate a registration message, and then used by the receiver to check that the message has not been altered.

In Mobile IP, the default authentication algorithm uses keyed-MD5 to in prefix+suffix mode to produce a one-way 128-bit message digest as the authenticator.

Table 13.3 Description of the fields in the registration reply messages.

Field Name	Description
Type	Value 3, indicating that this is a registration reply message.
Code	The result of registration request, detailed in Table 13.4
Lifetime	If the registration is accepted, it is the number of seconds that the registration will be honored by the home agent; a value of zero indicates that the mobile host has be de-registered.
Home Address	The home address of the mobile host. This is for telling the home agent to receive datagrams destined to this home address and then forward to the care-of address.
Home Agent	The address of the mobile host's home agent.
Identification	A 64-bit number used for replay protection. It is used for matching a registration reply to a registration request, discussed below.
Extensions	Additional extensions that are required, such as the authentication extension, discussed below.

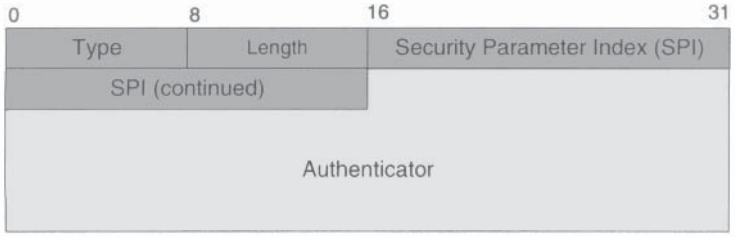


Figure 13.6 Mobile IP authentication extension.

The detailed operations are depicted in Figure 13.7. As shown in the figure, the 128-bit authenticator is computed over the shared key, and then the protected fields of the registration request/reply message, and finally the shared key again. With the authenticator inserted, the registration message is transmitted. After the receiver receives the registration message, the receiver computes its own 128-bit authenticator based on the shared key and the protected fields of the received registration message. Then a comparison is done between the computed authenticator and the authenticator in the received registration message. If the two are equal, the receiver can conclude that the registration message has not been altered.

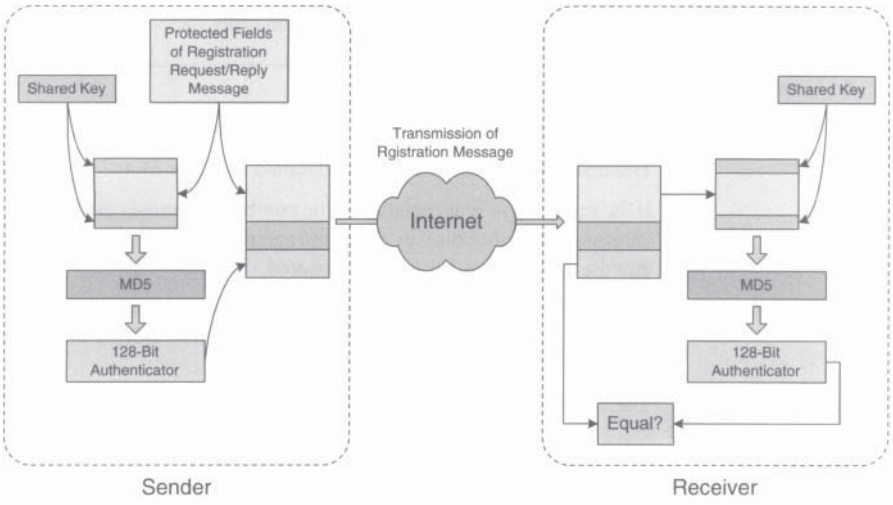


Figure 13.7 The keyed-MD5 authentication algorithm in Mobile IP.

13.4 TUNNELING APPROACHES

After the mobile host registers its care-of address at the home agent, the home agent intercepts IP datagrams destined to the mobile host and then tunnel the IP datagrams to the foreign agent and then to the mobile host. Specifically, two main processes are involved—interception of IP datagrams by the home agent and transmission of IP datagrams to the mobile host through tunneling.

There are two types of traffic destined to the mobile host. One type of traffic is originated from a node in the mobile host's home network. The other type is originated from outside the home network and is transmitted across the Internet (or from other topology of routers and links) and going through the home agent, which serves as a gateway for the home network. In the case of traffic going through the home agent, the interception of IP datagrams by home agent can be accomplished directly without any additional effort. However, if the traffic is generated from within the home networking, the home agent needs to perform a gratuitous ARP (address resolution protocol, a protocol for the translation between network-layer addresses and link-layer addresses) broadcasting beforehand, telling the other nodes in the home network to send IP datagrams destined to the mobile host to the home agent (at the link level).

The reason for performing a gratuitous ARP broadcasting by the home agent is that, while the mobile host is at home, the other nodes on the same network are likely to have ARP cache entries for the mobile host, which will become stale when the mobile host is away from home. Thus, it needs to invalidate those stale ARP cache

entries in order for the mobile host to correctly receive its IP datagrams. For the same reason, when the mobile host returns home, it needs perform a gratuitous ARP broadcasting so that its home address is associated to its own link-layer address by the other nodes within the home network.

After a mobile host's IP datagrams are captured, the datagrams are tunneled to the foreign agent for delivery to the mobile host. The tunneling can be done by one of the following encapsulation algorithms:

1. **IP-within-IP Encapsulation.** Defined in RFC 2003. It is the simplest approach and must always be supported (default algorithm).
2. **Minimal Encapsulation.** Defined in RFC 2004. It involves fewer fields in an IP datagram.
3. **Generic Routing Encapsulation (GRE).** Defined in RFC 1701. It is a generic encapsulation algorithm developed before the development of Mobile IP.

In the IP-within-IP encapsulation approach, the entire IP datagram sent by the Internet host is inserted in a new IP datagram as the payload. The header in the original IP datagram is unchanged except that the TTL field is decreased by one. On the other hand, for the header in the new IP datagram, the protocol field is given a value of 4, indicating the presence of the encapsulated IP datagram. In addition, the IP address of the home agent is used as source address while the care-of address of the mobile host is used as the destination address; the values of other fields are selected from the header of the original IP datagram. At the exit of the tunnel, the original IP datagram is restored and delivered to the mobile host.

Minimal encapsulation can be used if the home agent, mobile host and foreign agent all agree to use. It results in lower overhead when compared to the IP-within-IP encapsulation approach. However, it does not support fragmentation to deal with tunnels with smaller path maximum transmission units (MTUs). Figure 13.8 shows the operation of minimal encapsulation. As can be seen in the figure, a value of 55 in the protocol field indicates the presence of minimal encapsulation. Furthermore, a new header called minimal encapsulation header is inserted between the header and payload of the original IP datagram, while the fields in the original header are modified accordingly to reflect the change, such as the length, checksum, source and destination addresses, etc. The fields in the minimal encapsulation header are detailed in Table 13.5.

At the exit of the tunnel, the fields in the minimal encapsulation header are restored to the original IP header and then the minimal encapsulation header is removed. The length and checksum fields are re-calculated to reflect the change.

13.5 ROUTE OPTIMIZATIONS

From the above-mentioned discussion, we know that IP datagrams from an Internet host and destined to a mobile host need to be routed to its home agent first and then

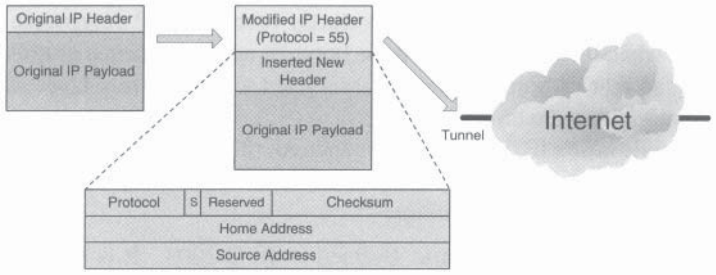


Figure 13.8 Minimal encapsulation.

tunneled to the foreign agent of the mobile host. On the other hand, IP datagrams from the mobile host can be directly routed to the Internet host. This kind of asymmetric routing, the so-called triangular routing [276] as illustrated in Figure 13.9, is far from optimal. The situation is even worse when the Internet host and the mobile host are close to each other, but datagrams from the Internet host to the mobile host are routed to and back from the distant home agent first.

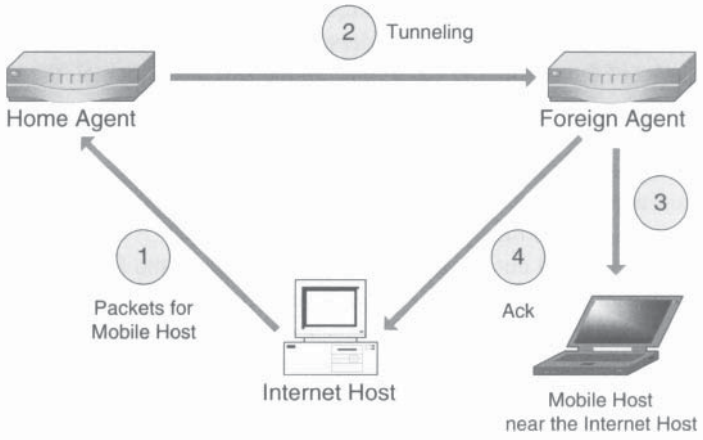


Figure 13.9 Triangular routing.

To solve the triangular routing problem, a route optimization protocol has been introduced. Basically, the protocol defines several messages so as to inform the correspondent host of an up-to-date mobility binding of the mobile host (i.e., the association of the mobile host's home address with its care-of address). With this binding, the correspondent host can directly send encapsulated datagrams to the care-of address of the mobile host. However, to achieve all these, it also means that changes

are required in the correspondent cost so that it can understand the rout optimization protocol.

There are two main types of message in the route optimization protocol, namely binding request and binding update messages. A corresponding node can issue a binding request message to the home agent to query the current care-of address of the mobile host. The home agent will then reply with a binding update message. Besides sending binding update messages upon request, the home agent can also send binding update messages to corresponding nodes when it detects that it needs to tunnel datagrams received from the corresponding nodes.

Similar to the security issues in the registration process, a malicious user may send a forged binding update message to the correspondent host so as to eavesdrop datagrams of the mobile host. To secure the the binding update, similar measures employed in the registration process can be applied, such as the use of the identification field and authentications.

Another problem associated with route optimization is the handoff problem. When a mobile host transits from a network to another, it needs to find a new foreign agent and undergo the registration process to tell the home agent its new care-of address. Subsequently the home agent tell the correspondent host the new binding of the mobile host. However, between the time the mobile host leaves the previous foreign network and the time the correspondent host is notified of the new binding, a considerable amount of datagrams might have been sent to the previous foreign agent. The previous foreign agent will forward the datagrams back to the home agent to send to the mobile host, after it discovers that the mobile host has left. The effect of delayed transfer of datagrams will then be magnified in upper layers like TCP, which has adopted some traffic control mechanisms.

In order to solve the above handoff problem, once the mobile host finds a new foreign agent, it can ask the new foreign agent to send a binding update message to the previous foreign agent. Then before the registration process is completed, the previous foreign agent can directly send the mobile host's datagrams to the new foreign agent, without the involvement of the home agent. However, there is authentication issue here. Specifically, we need a way to persuade the foreign agent that the binding update by the new foreign agent is sent on behalf of the mobile host. There are many ways in achieving this kind of authentication and it is left as a problem at the end of this chapter.

In addition to route optimization, there are some other techniques that can improve the performance in Mobile IP, for example:

1. **Simultaneous bindings.** A mobile can perform multiple registrations at the same to register more than one care-of address. Doing so, the mobile host will receive multiple copies of a datagram, which helps to increase the reliability of data transfer via the error-prone wireless media.

2. **QoS management.** To provide differentiated services to a mobile host, at the home network, the mobile host can negotiate a service level specification (SLS)

with the bandwidth broker (BB) to get a certain level of service [12]. When the mobile host moves away from home to a foreign network, in order to maintain differentiated services to the mobile host, the SLS used in the home network can be transformed and transmitted to the foreign network via the BBs in both networks.

13.6 PRACTICAL ILLUSTRATIONS—HIERARCHICAL MOBILITY MANAGEMENT

In Mobile IP applications, most of the time, the mobile host roams across multiple subnets within a single network of domain, that is, the user's mobility is local to a single administrative domain. A typical example is that while you are walking around in the campus with a PDA equipped with a WLAN adapter, the campus network provides you with uninterrupted Internet access using Mobile IP, wherever you go (e.g., library, cafe, bookstore, etc.).

With reference to the above application scenario, since each time a mobile host comes to a new subnet it gets a new care-of address, when the mobility of the mobile host increases, it needs to register its new care-of addresses more often with its home agent. As a result, with frequent location update, the traffic induced in the registration process can be excessive. The situation is even worse when the population of mobile users is large in the home network. It is because the home agent is likely to be overwhelmed by the large volume of registration messages from the mobile users with high mobility, which in turn causes the registration process to take a longer time to finish. Subsequently, the performance of data transfer in the mobile most suffers.

To solve the above problems, intuitively, we need to make the home agent unaware of the mobile host's mobility within the foreign domain. Specifically, we need to make the home agent see a single unchanged care-of address so as to reduce registration messages. To accomplish this, we can adopt a hierarchy of foreign agents [230].

As illustrated in Figure 13.10, originally the mobile host enters the foreign domain at $subnet_1$ and gets a care-of address CoA_1 from the foreign agent FA_1 in that subnet. Then the mobile host (MH) registers CoA_1 with the home agent (HA). Later on, the mobile host moves to $subnet_2$ and gets a new care-of address CoA_2 from the foreign agent FA_2. Instead of registering the new care-of address CoA_2 with the home agent, the mobile host asks FA_2 to inform FA_1 of its new care-of address. This way, the home agent continues to tunnel the mobile host's IP datagrams to CoA_1 and then FA_1 re-tunnels the datagrams to FA_2 for delivery to the mobile host. However, it should be emphasized that when the level of hierarchy increases, the latency of datagram transfer from FA_1 to the mobile host will also increase. Whenever a threshold level of hierarchy is reached (the threshold can be dynamically adjusted based on the mobility of the mobile host and the traffic load), the mobile host needs to register its current care-of address with the home agent to set up a new foreign agent hierarchy.

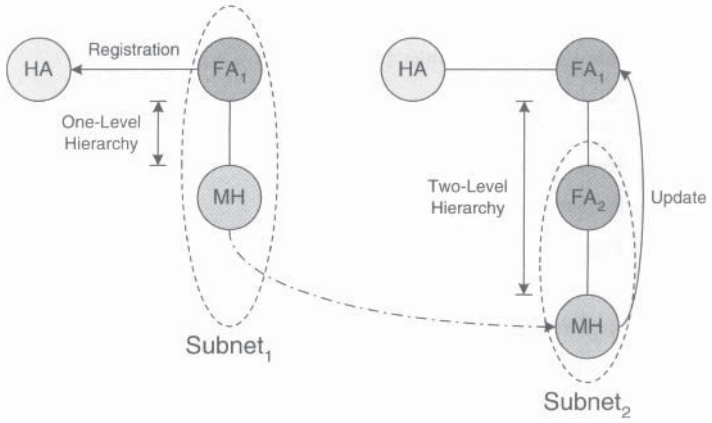


Figure 13.10 Dynamic hierarchical mobility management.

13.7 SUMMARY

In this chapter, we have discussed the three key design features of Mobile IP, namely advertisement mechanisms, registration process and tunneling approaches. The triangular routing problem is discussed and the corresponding route optimization approaches are presented. Finally, as a practical illustrative example, we talk about a hierarchical mobility management approach for reducing signaling traffic in the registration process.

PROBLEMS

13.1 How can a mobile host know whether it is at the home network or foreign network?

13.2 What is the triangular routing problem? How to solve it?

13.3 What will be the problem when a mobile host roams across different domains frequently? Describe a scheme to mitigate the problem.

Table 13.4 Code values for a registration reply message.

Code Value	Description
	Registration Successful
0	Registration accepted.
1	Registration accepted but simultaneous mobility bindings not supported.
	Registration Rejected by the Home Agent
128	Reason unspecified.
129	Administratively prohibited.
130	Insufficient resources.
131	Mobile host failed authentication.
132	Foreign agent failed authentication.
133	Registration identification mismatch.
134	Poorly formed request.
135	Too many simultaneous mobility bindings.
136	Unknown home agent address.
	Registration Rejected by the Foreign Agent
64	Reason unspecified.
65	Administratively prohibited.
66	Insufficient resources.
67	Mobile host failed authentication.
68	Home agent failed authentication.
69	Requested lifetime too long.
70	Poorly formed request.
71	Poorly formed reply.
72	Requested encapsulation unavailable.
73	Requested Van Jacobson header compression unavailable.
80	Home network unreachable (ICMP error received).
81	Home agent host unreachable (ICMP error received).
82	Home agent port unreachable (ICMP error received).
83	Home agent unreachable (other ICMP error received).

Table 13.5 Description of the fields in the inserted header of minimal encapsulation.

Field Name	Description
Protocol	It is copied from the protocol field of the original IP header for identifying the protocol type of the original IP payload.
S	It is 1 if original source address is present; otherwise, it is 0.
Checksum	Checksum of the header.
Home Address	It is copied from the destination address in the original IP header.
Source Address	It is present when the value of S is 1, and is copied from the source address of the original IP header.

This Page Intentionally Left Blank

CHAPTER 14

IPV6

Riches are often abused, never refused.

—Danish Proverb

14.1 INTRODUCTION

Driven by the shortage of IP addresses, an improved version of the Internet Protocol has been in pressing need for many years.

Interoperability:

IPv6 is designed, at the very beginning, to interoperate with IPv4 because a "flag day" for the transition from IPv4 networks to IPv6 networks is simply infeasible. Therefore, IPv4 networks and IPv6 networks will coexist for a certain period of time. Consequently, some transition mechanisms are needed. Currently, the main three types of transition mechanism are dual-stack, tunneling, and translation. On the other hand, a number of advanced features are incorporated in IPv6 in order to increase efficiency when different network nodes interoperate with each other. Features like

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau
Copyright © 2007 John Wiley & Sons, Inc.

simplified header format and the use of extension headers provide enhanced routing efficiency when compared to IPv4. Mobile IPv6, which takes experience from Mobile IPv4 and utilizes advanced features in IPv6, provides mobile hosts with enhanced mobility service so as to facilitate efficient roaming across different networks.

Performance:

Since IPv6 was released as a draft standard in 1995, network device manufacturers and operating system providers have started their efforts to support IPv6. Since then, IPv6 implementations have matured, and the deployment of IPv6 networks worldwide has been transiting from experimental networks to operational networks. Because of this transition trend, it becomes a concern to investigate the performance of IPv6 on user applications.

Zeadally *et al.* have conducted an empirical performance comparison of IPv4 and IPv6 protocol stack implementations of commodity operating systems [369]. They investigated performance metrics such as throughput (TCP and UDP), round-trip time (RTT), socket creation time, and TCP connection time. The experiment was conducted on two identical workstations connected by a point-to-point link. Their performance results obtained from a Red Hat Linux operating system indicate that there is no significant difference on the performance between IPv4 and IPv6. Specifically, in the throughput experiments (TCP and UDP), there is notable decrease in IPv6 implementation for messages smaller than 512 bytes, which is due to the overhead of the larger header size of IPv6. On the other hand, when compared to IPv4, the round-trip latencies for IPv6 are lower. As for TCP connection time, IPv6 results in a slightly increased connection time due to the increased IPv6 header size, while there is more or less no difference in socket creation time.

On the other hand, Wang *et al.* have conducted a performance comparison of IPv6 and IPv4 in networks of a larger scale [347]. They collected packets from IPv6/IPv4 dual-stack Web servers located in 44 countries to study performance metrics like connectivity, packet loss rate, and RTT. Their measurement results indicate that IPv6 connections tend to have lower RTTs than their IPv4 counterparts. However, at the same time, IPv6 connections suffer higher packet loss rate. Finally, it is found that tunneling does not show notable performance degradation in IPv6.

14.2 DESIGN GOALS

The Internet Protocol version 4 (IPv4) was developed in the early 1970s to facilitate information exchange between educational and government entities in the United States. At the beginning, it was intended for use in a small sized and closed system of limited number of access points, and therefore the original IP protocol is rather simple and the developers did not foresee requirements such as security. However, as more and more machines are interconnected, it was realized that the features provided by the original IP protocol is insufficient to meet the needs in a large network. The situation is even worse when the Internet became world wide public and the emergence of World

Wide Web (WWW) in the early 1990s. Since then, a number of Internet applications have been developed and the problems of IPv4 have become more apparent. In order to mitigate the shortcomings of IPv4 found in practical use, many additional protocols and standards have been developed to work collaboratively with IPv4, such as Internet Control Message Protocol (ICMP) and Network Address Translation (NAT). This makes the current IPv4 protocol very complex.

Since the proliferation of web pages, emails, and instant messengers, the number of Internet users has been increased unprecedentedly. Besides, people are becoming more dependent on data and information services provided by the Internet. This trend continues to grow with the advancement of mobile communication technology. Consequently, there has been an explosive growth in IP-capable devices, since it is advantageous to utilize existing Internet infrastructure. Handheld game consoles, such as Sony Playstation Portable (PSP) and Nintendo DS (NDS), are equipped with Wi-Fi devices and can access to the Internet. Many PDAs and mobile phones are also Wi-Fi-enabled. In additional, IP-phones and IPTV are very common nowadays. Thus, IP-enabled devices are penetrating into every aspect of our life. With this explosive growth in IP-capable devices, one apparent and most serious problem of IPv4 is the shortage of IP addresses. To mitigate this address space problem, NAT has been suggested to map public IP addresses with private addresses such that a few number of IPv4 addresses are used to enable a high number of users with private addresses to connect to the Internet. However, NAT prevents end-to-end communication as it modifies end node addresses during address mapping. Consequently, some applications are not able to work with NAT, such as peer-to-peer applications. Therefore, inevitably, the problem of running out IPv4 addresses need to be solved radically.

In the early 1990s, the Internet Engineering Task Force (IETF) started to develop a successor protocol to IPv4. The Internet Protocol Next Generation (IPng) working group was started in 1993 to investigate different proposals and recommendations. In 1995, the RFC 1883 (Internet Protocol, Version 6 (IPv6) Specification) [82] was published, which was then obsoleted by RFC 2460 [83] in 1998.

Although solving the address space problem in IPv4 is the main driving force for the introduction of IPv6, IPv6 is not merely about extending the address space of IPv4. Based on the experience that we have from the development and usage of IPv4, IPv6 is also designed to augment many other aspects of IPv4 and many additional features are integrated in IPv6. Specifically, proven functions that make IPv4 successful have been retained and modified to extend the scalability and flexibility, limitations of IPv4 have been removed, additional functions that are originally designed to work collaboratively with IPv4 are now integrated in IPv6 to enhance the efficiency and security, and finally new features are added to ease network management, to provide mobility support to mobile nodes, etc.

The following gives an overview of the main features of IPv6:

Expanded Address Space: The length of IP address field is extended from 32 bits to 128 bits. In practice, it supports virtually infinite number of IP addresses.

In addition, the address space can be structured in a hierarchical way to allow optimized global routing of packets.

Simplified Header Format: The header of IPv6 has a fixed length of 40 bytes. Since 32 bytes are used for source and destination addresses, only 8 bytes are left for incorporating other information, and some fields of the IPv4 header have been either eliminated or made optional. This simplified header format of IPv6 allows for faster processing of IP packets.

Extension Headers: In IPv4, optional fields are integrated in the base header. However, in IPv6, optional fields are put in extension headers after the based header. Extension headers are added only when needed and may not be processed by intermediate nodes between the source and destination so as to allow fast processing of packets.

Autoconfiguration: IPv6 supports stateless and stateful autoconfiguration of IP addresses to reduce the network management costs. In the stateful mode, the operation is similar to the use of DHCP for IPv4, where a node requests for an IP address dynamically from a DHCPv6 server. On the other hand, the stateless mode allows a mobile node to be connected to a network in an easy and efficient way. For example, the mobile node can autoconfigure for a global IP address by using its MAC identifier, in addition to getting a network prefix from an IPv6 router.

Security: IPsec is mandatory in IPv6. The Authentication Header (AH) and the Encapsulating Security Payload Header (ESP) are included as extension headers.

Mobility Support: Mobility support provided by IPv6 removes some of the limitations found in IPv4 implementations. Mobile IPv6 aims at providing mobile users with an efficient way to maintain connections when they move between networks. For example, route optimization is supported in Mobile IPv6 and foreign agents are no longer needed.

Quality of Service (QoS) Support: To support QoS in IPv4, various additional protocols have be developed to work with IPv4. On the contrary, IPv6 integrates QoS support by using flow label. Specifically, packets belonging to different traffic flows can be labeled such that the sender can request special handling of the packets.

Unlike the switch from using Network Control Protocol (NCP) to IP in one day in 1983, there will be no such flag day for IPv6. Thus, another key feature of IPv6 is that IPv6 is designed to be interoperable with IPv4. Specifically, IPv4 and IPv6 coexisting schemes have been developed, such as dual IP layer approach and tunneling. This allows incremental upgrade from IPv4 to IPv6, since an IPv4 host can be upgraded anytime independently of other hosts and routers.

IPv6 Addressing

As IPv6 addresses are four times the length of 32-bit IPv4 addresses, the dotted decimal notation as adopted in IPv4 is inappropriate for IPv6 addresses. The IPv6 addressing architecture is defined in RFC 3513 [149].

Figure 14.1 shows an example IPv6 address taken from [81]. As can be seen, an IPv6 address typically consists of eight blocks of four hexadecimal digits (i.e., each block contains 16 bits), and each block is separated by a colon. In each block, if there are leading zeroes, then the zeroes can be omitted. Moreover, if the entire 16-bit block is zero, then the block can be reduced to two consecutive colons. If there are consecutive blocks of zeroes, the blocks can also be reduced to two consecutive colons. However, to avoid confusion, this kind of double-colon notation can only be used once in an IPv6 address.

Similar to IPv4 addresses, an IPv6 address can be divided into two portions, the network portion and the host portion. These two portions can be explicitly defined by using a prefix length (i.e., the number of bits used to represent the network portion), as illustrated in Figure 14.1

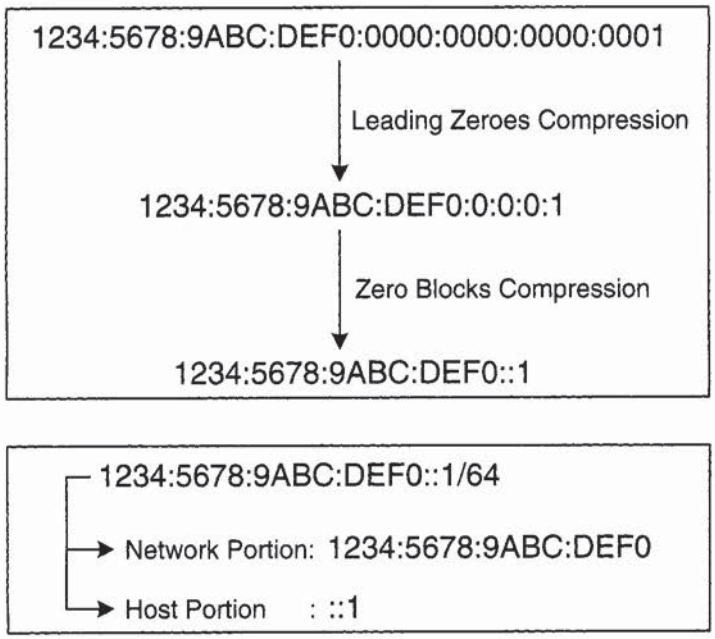


Figure 14.1 IPv6 addressing example.

In IPv4, there are three classes of address, namely unicast, multicast, and broadcast. In IPv6, broadcast is considered a special case of multicast, and a new class of address

called anycast is introduced. Anycast addresses can be used when we want to assign a single global unicast address to more than one host.

IPv6 Header Format

Figure 14.2 shows a comparison between the IPv4 header format and the IPv6 header format. Table 14.1 describes he fields in the IPv6 header. By comparing the header of IPv4 with that of IPv6, we can find that several fields in the IPv4 header have been removed, meaning that some operations are revised in IPv6:

IHL (Internet Header Length): As the IPv6 header has a fixed length of 40 bytes, a length indicator is no longer needed.

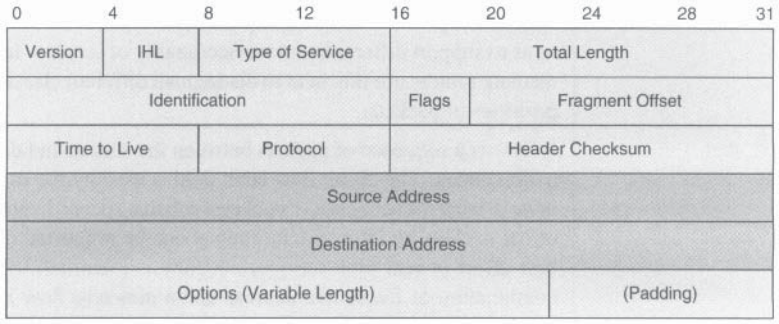
Identification, Flags, Fragment Offset: These fields have been moved to an extension header, called the Fragment Header. In IPv6, fragmentation and reassembly can only be performed at the source node and destination node, respectively. This means that intermediate routers are not allowed to fragment and reassemble packets during routing of packets. During packet transmission, the Fragment Header is used to send a packet of size larger than the path MTU (maximum transmission unit) to a destination node, and the source node will divide the original packet into several smaller packets. If a router receives a packet that is too large to be forwarded, the packet will be discarded and a "Packet Too Big" ICMPv6 error message will be sent to the source node to inform the use of smaller packet size. Since fragmentation and reassembly are time-consumption operations, removing them at intermediate routers can help speed up the routing task significantly.

Header Checksum: Since other layers in the Internet protocol stack also perform error checking (e.g., TCP and UDP in the transport layer), error checking at the network layer seems to be redundant and can be removed completely. Another reason is that point-to-point connections nowadays have very low error rate. Again, removing error checking aims at allowing fast packet processing in intermediate routers.

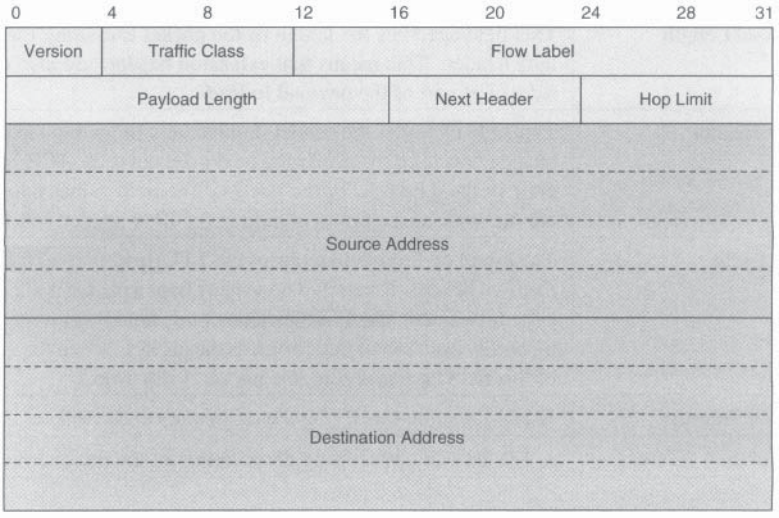
Options: The Options field is no longer a part of the IPv6 header and has been moved to extension headers. The removal of this variable-length field results in the fixed 40-byte IPv6 header.

Transition Mechanisms

As the number of Internet-connected machines is significantly large, a "flag day" for upgrading from IPv4 to IPv6 is practically infeasible. Also, because of the upgrading cost, many systems may not upgrade from IPv4 to IPv6 in the immediate future. Despite that IPv6-capable systems can be made backward compatible with IPv4 systems, existing IPv4 systems are merely not able to handle IPv6 packets. Therefore, inevitably, IPv4 and IPv6 coexisting mechanisms are needed to allow the gradual transition from IPv4 networks to IPv6 networks, and more importantly to



(a) IPv4 header format



(b) IPv6 header format

Figure 14.2 Header formats of IPv4 and IPv6.

Table 14.1 Description of the fields in the IPv6 header.

Field Name	Description
Version	Value 6, indicating that this is an IPv6 header.
Traffic Class	The usage of this field is similar to the TOS (type of service) field in IPv4. It is used to specify different types of IP packets so as to support differentiated service/quality of service. Intermediate routers use this field to distinguish different classes or priorities of packets.
Flow Label	A flow is a sequence of packets between the source and destination nodes. The 20-bit flow label field is used by the source node to label the sequence of packets such that special handling of the packets by intermediate routers can be requested (e.g., best effort or real-time service). A flow is identified by the combination of the source address and a non-zero flow label value. Hosts or routers that do not understand the flow label field should set the field to zero, and a packet with the flow label field set to zero does not belong to any flow.
Payload Length	This field specifies the length of the packet excluding the 40-byte header. This means that extension headers are also considered as part of the payload in IPv6.
Next Header	This field identifies the header immediately following the IPv6 header. It is used to specify the next protocol (e.g., TCP, UDP, etc.) in the data field of the packet. The field values adopted are the same as the protocol field in the IPv4 header.
Hop Limit	The 8-bit hop limit field replaces the TTL (time to live) field in the IPv4 header. It specifies how many hops a packet is allowed to be forwarded, and thus the value of this field is decremented by one by each router that forwards the packet. When the value of this field becomes zero, the packet is discarded.
Source Address	A 128-bit value specifying the originator of the packet.
Destination Address	A 128-bit value specifying the recipient of the packet.

allow the interoperability between the two networks. Existing transition mechanisms can be classified into three categories, dual-stack, tunneling, and translation.

The simplest way of allowing coexistence of both IPv4 and IPv6 protocol stacks is the dual-stack approach. In this approach, the network interface of a node is configured with an IPv4 address as well as an IPv6 address. Therefore, the node can run both protocols in parallel, and hence the node is called a IPv6/IPv4 node. Depending on whether the incoming packet is an IPv4 or IPv6 packet, appropriate protocol stack is used to handle the packet. Similarly, the IPv6/IPv4 node needs to

determine whether the next node supports IPv6 or IPv4 only. This can be achieved by using DNS, which can return the IPv6 and/or IPv4 address(es) of the next node.

In the dual-stack approach, if both the source and destination nodes are IPv6-capable but there is intermediate node that can speak only IPv4, then it might turn out that the source and destination nodes are communicating using IPv4 packets finally. As Figure 14.3 [194] illustrates, node A is IPv6-capable and wants to send a packet to another IPv6-capable node, node E. Since the next node of node A, i.e., node B, is an IPv6/IPv4 node, node A can send an IPv6 packet to node B. Afterwards, node B needs to forward the IPv6 packet to node C. However, since node C can only understand IPv4 protocol, node B needs to convert the IPv6 packet to an IPv4 packet, by performing appropriate address mapping and dropping of some IPv6-specific fields. Consequently, when node D receives the IPv4 packet from node C and constructs an IPv6 packet for transmission to node E, some fields in the original IPv6 packet sent from node A are lost.

While the dual-stack approach is simple and flexible, it has the disadvantage of consuming more processor power and memory. The reason is that the IPv6/IPv4 node needs to keep all type of tables (e.g., routing table) twice for each protocol. Nevertheless, the dual-stack approach, as we will see later, forms the basis for other transition mechanisms.

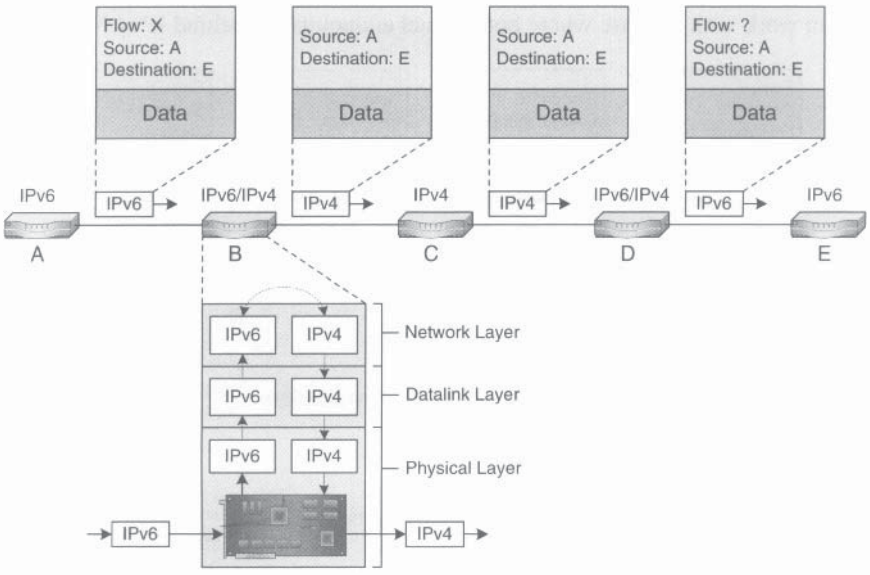


Figure 14.3 Dual-stack approach [194].

As mentioned earlier, in the dual-stack approach, if there is intermediate node that can handle only IPv4 packets, then some IPv6-specific fields will be lost dur-

ing packet transmission between two IPv6-capable nodes. Tunneling approach is another transition mechanism that can be used to solve this problem. Specifically, by using tunneling, information/data from one protocol (e.g., an IPv6 packet) can be encapsulated inside the packet of another protocol (e.g., an IPv4 packet). Effectively, the latter acts as a tunnel for transmission of original information/data in the former without any augmentation.

Figure 14.4 [194] illustrates the idea of the tunneling approach. At node B, it discovers that node C can only handle IPv4 packets. Then node B encapsulates the IPv6 packet received from node A in an IPv4 packet and send the IPv4 packet to node C, who is situated inside an IPv4 tunnel. At this point, we say that node B is the tunnel entry point, and the IPv4 packet will be transmitted via node C to the tunnel exit point, i.e., node D. At node D, the received IPv4 packet will be decapsulated, and the original IPv6 packet sent from node A will be delivered to node E, without any modification of the IPv6-specific fields.

In the tunneling approach, it is required that the tunnel endpoint (e.g., the tunnel exit point at node D in Figure 14.4) should have a public IPv4 address. However, if the tunnel endpoint is located behind one or more IPv4 NATs, other mechanism is needed. Teredo is designed to allow IPv6 packets to traverse IPv4 NATs by using IPv4-based User Datagram Protocol (UDP) messages [267]. Teredo consists of three main components—Teredo server, Teredo client, and Teredo relay. Teredo, however, does not work with the case where both tunnel endpoints are behind IPv4 NATs.

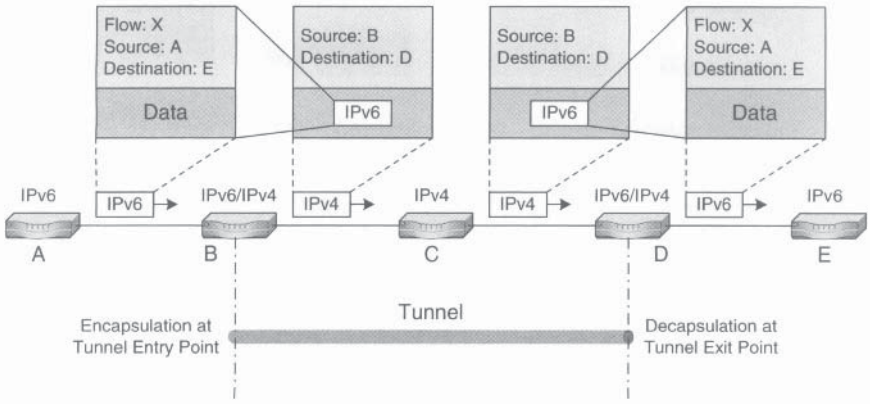


Figure 14.4 Tunneling approach [194].

The discussion above focuses on providing connectivity between IPv6-capable nodes via an IPv4 network. In order to provide connectivity between an IPv6 node and an IPv4 node, translation mechanisms can be applied, for example, the Network Address Translation-Protocol Translation (NAT-PT). NAP-PT is applied in a router that supports both IPv4 and IPv6. NAT-PT uses a pool of IPv4 addresses for dynamic assignment to IPv6 nodes when a session is initiated between an IPv6 node and an

IPv4 node. Then the NAP-PT router performs addressing mapping in the IPv4 and IPv6 headers. In addition, if an application program embeds IP addresses within the data of a packet, then address translation within the data should also be performed in order to match the address mapping in the headers. To perform this kind of application level translation, an Application Level Gateway (ALG) can be employed within the NAT-PT router.

Figure 14.5 illustrates how NAT-PT works. In the example, node A, whose IPv6 address is 1234:5678:9ABC::DEF0, wants to communicate with node B, whose IPv4 address is 238.12.234.56. Node A sends a packet destinated for node B with a destination address 2228:38::238.12.234.56. The prefix 2228:38::/96 is advertised by the NAT-PT router such that all packets sent from an IPv6 node will be routed through the NAT-PT router. When the NAT-PT router receives the IPv6 packet from node A, corresponding mapping of IP address and port number is carried out as depicted in Figure 14.5 For inbound traffic, the mapping is performed in a similar way.

There are several disadvantages of using NAT-PT. First of all, end-to-end security cannot be provided, which is a common pitfall of using any form of NAT. Secondly, there might be information lost when translating from one protocol to another protocol. Finally, the NAT-PT router is the performance bottleneck, since all packets from the same session need to go through the same NAT-PT router.

14.3 MOBILITY SUPPORT

As a result of the advanced mobile communication technology, mobile devices such as smart phones are penetrating into our lives. By using a mobile device, we can access Internet services anytime and anywhere. In addition, to leverage existing Internet infrastructures, IP services are commonly implemented on mobile devices. For mobile devices, mobility is a key concern. Specifically, it is advantageous if existing connections can be maintained when a mobile host roams across different wireless networks. In Chapter 13, we have discussed how Mobile IPv4 supports mobility to mobile hosts.

Using Mobile IP, when a mobile host is attached to a new network, its network prefix changes, meaning that a new care-of address (CoA) is assigned to it. Because of this reason, when a considerable number of mobile hosts are attached to a foreign network (due to the proliferation of mobile devices), limitation of address space in IPv4 may become an issue in Mobile IPv4. On the other hand, Mobile IPv4 has other limitations such as lack of support of route optimization. In view of these problems, Mobile IPv6 is defined in RFC 3775 [172]. Since Mobile IPv6 is designed based on the experience from Mobile IPv4 and advanced features of IPv6, the mobility service provided by Mobile IPv6 is considered as one of the killer applications in IPv6 [250].

When compared with Mobile IPv4, the key changes in Mobile IPv6 are as follows:

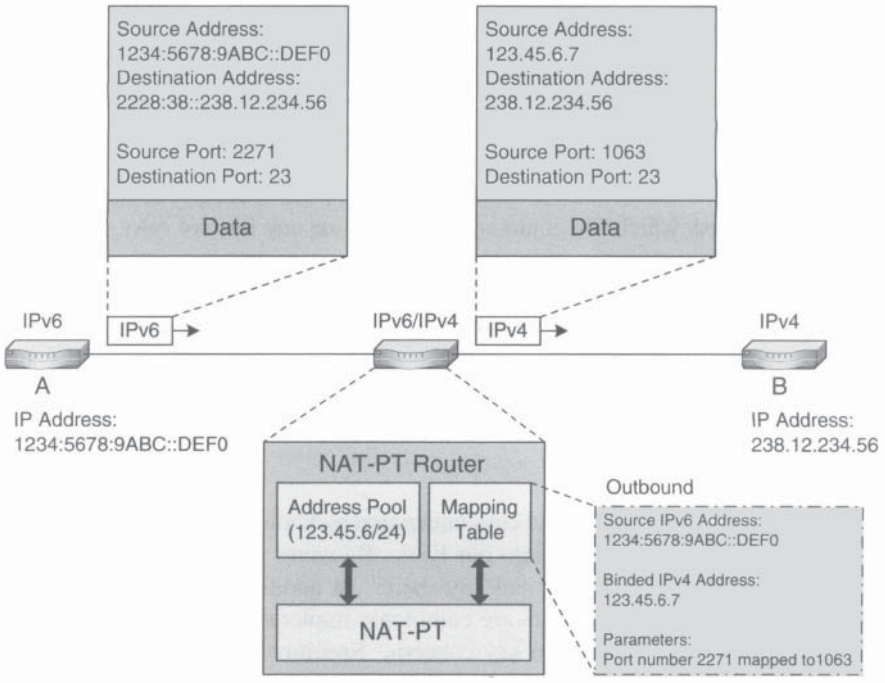


Figure 14.5 Translation approach—Network Address Translation-Protocol Translation (NAT-PT).

Foreign Agents: In Mobile IPv6, foreign agents are not necessary. Mobile hosts can obtain a CoA by using stateless or stateful autoconfiguration. In the stateful approach, a CoA can obtained from a DHCPv6 server. In the stateless approach, a mobile host can use the network prefix advertised by the router of the foreign network to form the upper 64 bits of the CoA, and to use its MAC address to form the lower 64 bits of the CoA.

Route Optimization: Route optimization is supported in Mobile IPv6. In route optimization, it is required that the mobile host registers its CoA with the correspondent host. As for the security measure required for this correspondent registration, a return routability procedure is defined for authenticating the mobile host to the correspondent host. Then, to send packets from the correspondent host to the mobile host, a special routing header (type 2) is used by the correspondent host.

Use of Neighbor Discovery Protocol: In Mobile IPv4, ARP (address resolution protocol) is used. However, the IPv6 neighbor discovery protocol is used in Mobile IPv6.

Home Agents Discovery: Mobile IPv6 supports multiple home agents. A mechanism called Dynamic Home Agent Address Discovery is defined in Mobile IPv6 in order to enable a mobile host to learn about the reconfiguration of the home network (e.g., a change of IP address of the home agent).

In Mobile IPv6, an extension header called Mobility Header (MH) is defined. This extension header has a next header value 135. Using the header, 11 mobility messages are defined. Mobility messages are used by the mobile host, correspondent host and home agent to carry out the return routability procedure, binding update procedure, and fast handover.

14.4 HOME AGENTS DISCOVERY

When a mobile host moves to a new foreign network, it needs to send a binding update message to its home agent. However, the mobile host needs to determine the address of its home agent first. Dynamic Home Agent Address Discovery is a mechanism in Mobile IPv6 for serving this purpose. It involves the use of ICMPv6 messages, a home agents list and anycast addresses. This mechanism is useful for a mobile host to find new home agents dynamically when its original home agent goes down.

As illustrated in Figure 14.6, when the mobile host needs to find the address of its home agent, firstly it sends an ICMPv6 Home Agent Address Discovery Request message to the home agent anycast address. Then in the home network, the home agents who are configured with the home agent anycast address will respond with a Home Agent Address Discovery Reply message. In the reply message, it consists of a Home Address field storing one or more home agent addresses. Finally, on getting

a home agent address, the mobile host can send a binding update message to its home agent.

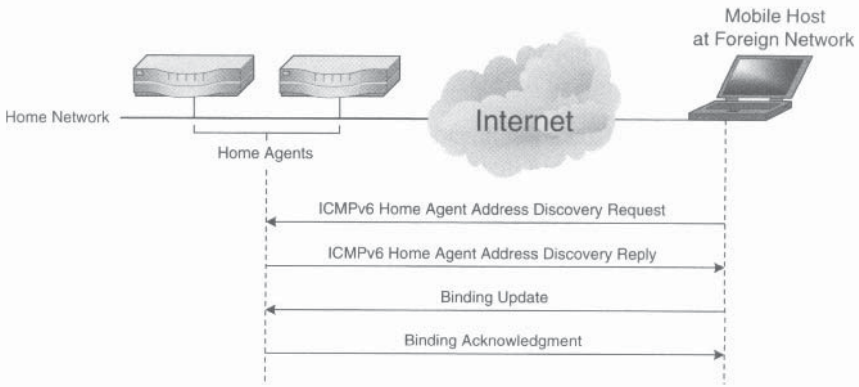


Figure 14.6 Dynamic Home Agent Address Discovery.

In order to reply a mobile host's Home Agent Address Discovery Request messages, each home agent is required to maintain a home agents list, which contains all the home agents on the home network link. Each home agent maintains the list by monitoring router advertisements on the network, since a router will advertise itself as a home agent by setting the H-bit in the router advertisement packet. Specifically, the list contains the following information:

- Link-local addresses of the home agents on the home network link.
- Remaining lifetimes of the entries for the home agents. The entry for the home agent whose lifetime has expired must be deleted.
- One or more global IPv6 unicast addresses for the home agents.
- The preference value for each home agent. This value can be found in router advertisements. A higher value means a higher preference. This preference value is used for sorting home agents when replying Home Agent Address Discovery Request messages.

14.5 PRACTICAL ILLUSTRATIONS—IPV6 BASED VTHD NETWORK

Since it is inevitable to use IPv6 in the long term, governments or organizations will benefit from earlier introduction of IPv6 services to their existing IPv4 networks. With this consideration in mind, in the VTHD project carried out in France, which is partially funded by the French government, IPv6 services have been introduced in the very high broadband IP/wavelength-division multiplexing (WDM) network (VTHD

network) for new-generation Internet applications [11]. In VTHD, both IPv4 and IPv6 services are supported, and it has a public IPv6 subnet, which is connected to the global IPv6 network via Open Transit (France Telecom's international IP service for ISPs).

Figure 14.7 [11] shows the VTHD architecture. The VTHD backbone network has 10 transit routers interconnected by 2.5 Gb/s optical channels, with 12 additional edge routers connected to the backbone for aggregating network traffic from campuses or organizations. The VTHD network is assigned with a /42 prefix, namely 2001:688:1F80::/42, which allows 64 /48 prefixes ranging from 2001:688:1F80::/48 to 2001:688:1FBF::/48. Out of the 10 transit routers in the backbone network, seven of them are connected with external sites. Therefore, address space 2001:688:1F80::/42 can be divided into eight blocks, each having eight /48 prefixes, as illustrated in Figure 14.7 (the first block, 2001:688:1F80—2001:688:1F87, is used by the backbone). In each block, only the first quartet of the eight /48 prefixes is allocated while the second quartet is reserved for use in the case of network increase. When a partner node is connected to a transit router node, it will obtain a /48 prefix from the first quartet of the node. If all the /48 prefixes in the first quartet have been used, a /48 prefix can be obtained in the second quartet.

As the VTHD network supports both IPv4 and IPv6 services, some transition mechanisms are implemented in the VTHD network. Specifically, the dual-stack approach is mainly implemented in the IPv6 backbone nodes. In addition to the dual-stack approach, VTHD also implements some tunneling transition mechanisms, namely IPv6-in-IPv4 tunneling and 6PE tunneling [266]. In 6PE tunneling, IPv6 islands are interconnected over a Multi-Protocol Label Switching (MPLS)-enabled IPv4 cloud. In this approach, IPv6 packets are received on IPv6 Provider Edge routers (6PE), which are IPv6/IPv4 dual-stack routers. Afterwards, appropriate encapsulation and decapsulation are performed at the 6PE routers in order to transmit the IPv6 packets.

In order to test the performance of the VTHD network, a testbed as depicted in Figure 14.8 was set up [11]. In the test console, TCL scripts were executed so as to control some parameters like frame loss rate on the test devices. Afterwards, throughput results were measured. As reported in [11], the test was aimed at studying the performance of the three transition mechanisms as mentioned above. Performance results indicated that the dual-stack approach and 6PE tunneling achieved nearly wire-speed performance, and that the IPv6-in-IPv4 tunneling approach had some limitations. Besides investigating the performance of the three transition mechanisms, the performance of IPsec was also studied in the VTHD project. It is reported that IPsec over IPv6 achieved the same performance as IPsec over IPv4 [11].

In the VTHD project, different network devices such as routers and test devices, were intentionally acquired from different manufacturers so as to study the interoperability issues. Through the experiments conducted in the project, it is asserted that network equipment nowadays is IPv6 ready for the transition from IPv4 networks to IPv6 networks.

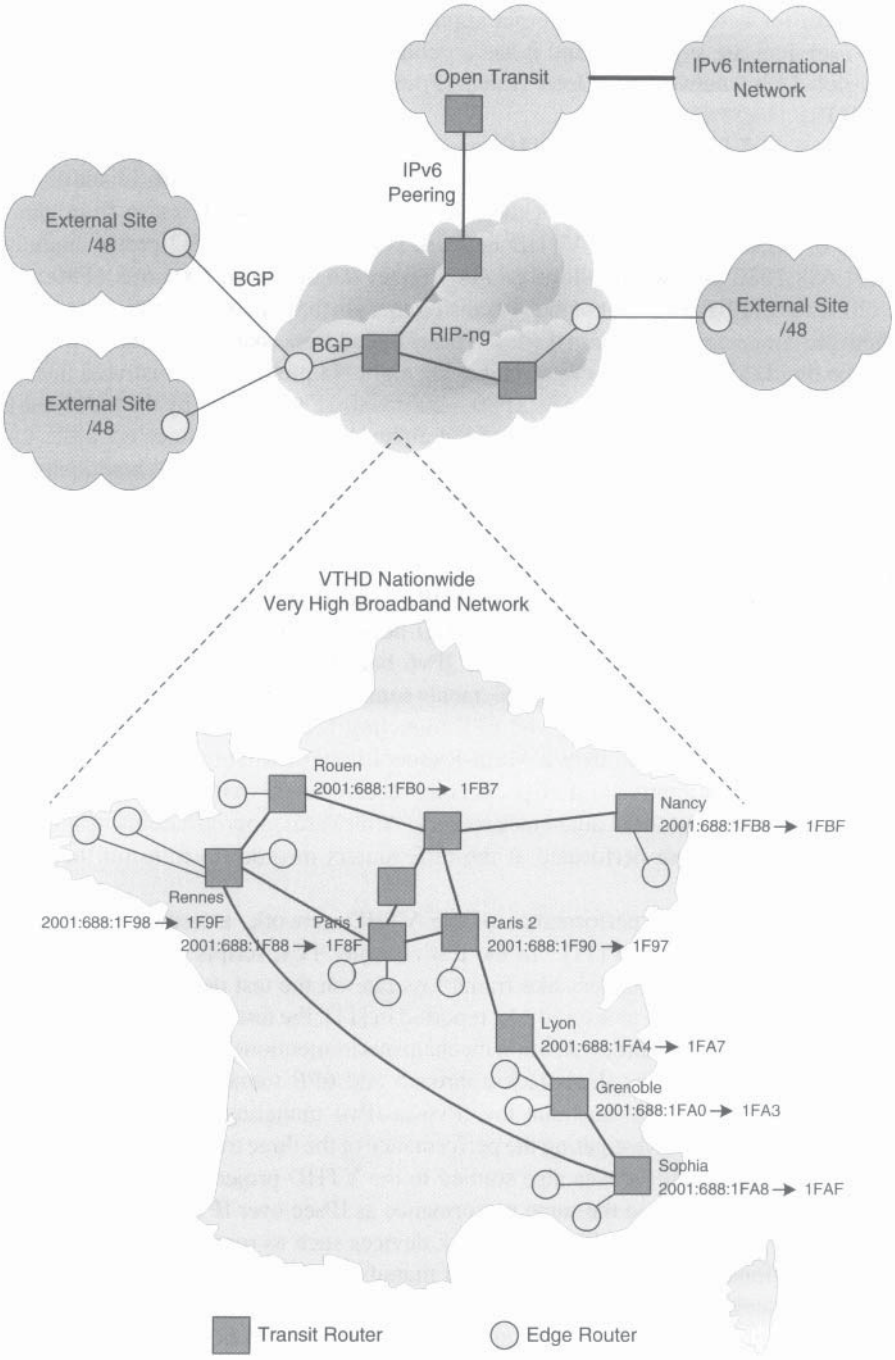


Figure 14.7 VTHD architecture [11].

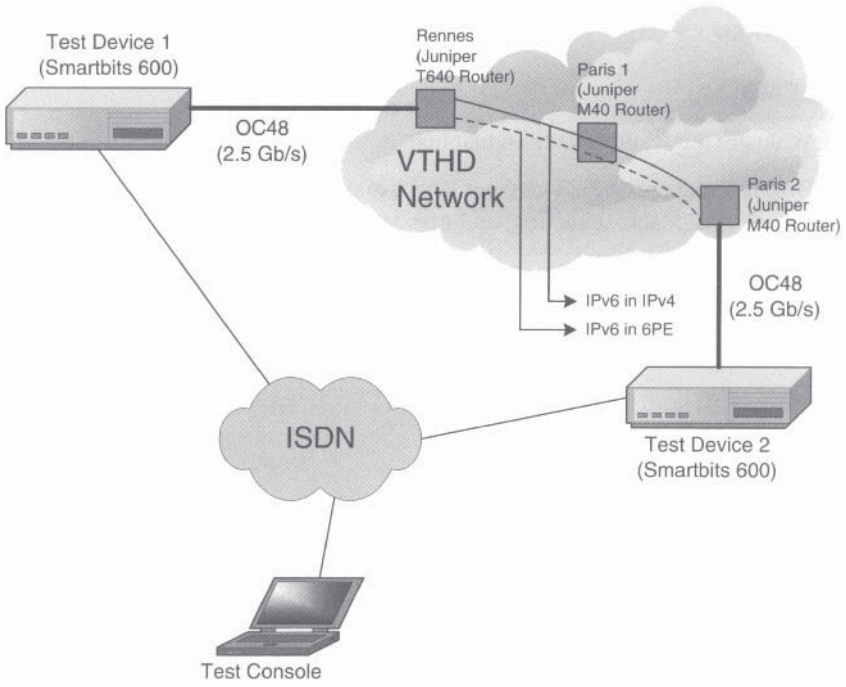


Figure 14.8 VTHD testbed setup [11].

14.6 SUMMARY

In this chapter, we have discussed the key features and design goals of IPv6. Specifically, the addressing architecture, header format, coexisting mechanisms with IPv4, and mobility support (Mobile IPv6) have been introduced. For practical illustrations, we have seen how IPv6 services were deployed in the VTHD network in France. Despite that IPv6-ready hardware (e.g., network equipment) and software (e.g., operating systems) components have been matured and provide optimized performance, IPv6 networks are still being deployed gradually. Nevertheless, IPv6 is inevitable in the long term, and governments or organizations will benefit from earlier introduction of IPv6 services.

PROBLEMS

14.1 What are the features of IPv6 that help to increase packet routing efficiency?

14.2 Briefly compare the advantages and disadvantages of the three transition mechanisms—dual-stack, tunneling, and translation.

14.3 What are the key differences between Mobile IPv4 and Mobile IPv6?

CHAPTER 15

WIRELESS APPLICATION PROTOCOL (WAP)

Translation is at best an echo.

—George Borrow; Lavengro, XXV, 1851

15.1 INTRODUCTION

As the TCP/IP protocol suite was not designed for a wireless environment where bandwidth resource is expensive (consider a cellular network), modifications and enhancements are necessary for adopting it in a wireless network.

Interoperability: WAP aims at bridging the gap between the mobile world and the Internet. It is designed to bring diverse Internet content and other data services to mobile users across different wireless technologies so that mobile users can reach the Internet to have different information services anytime and anywhere. In order to enable mobile devices to access ordinary web servers, WAP tries to utilize existing Internet protocols and standards as much as possible, such as XML, HTML, HTTP, and TLS. The adapted protocols and standards are optimized for deployment in a

Wireless Internet and Mobile Computing. By Yu-Kwong Ricky Kwok and Vincent K. N. Lau

Copyright © 2007 John Wiley & Sons, Inc.

wireless environment where mobile devices are limited in capabilities. To access ordinary web servers, WAP-enabled mobile devices rely on a WAP gateway to provide protocol conversion between the WWW protocol stack and the WAP protocol stack.

Performance: The WAP protocol stack is a lightweight protocol stack that is designed to address the limitations of wireless devices and the wireless network. It is designed with two main goals—to minimize bandwidth requirement and to maximize the number of supported network types (e.g., 9.6 Kbps in GSM). Each layer of the protocol stack is designed to be scalable and efficient. For example, in Wireless Transaction Protocol (WTP), there is no explicit connection setup or teardown, reducing much amount of overhead when compared to TCP. To reduce transmission time, WAP uses binary-coded WML (wireless markup language) pages. In addition, WAP specifies a caching model and user agent profile (UAProf) for efficient delivery of device-specific content.

15.2 WAP SERVICE MODEL

As the Internet provides a wealth of information (and the information provided continues to expand every day), people are becoming more and more reliable on the Internet. Being able to access the Internet, they can retrieve useful information such as maps, weather forecast, etc. They can also handle routine activities electronically, such as e-banking, e-commerce, e-checkin, etc. The Internet has been helping people to make their lives easier. Because of this reason, it would be very convenient if we can reach the Internet to have different information services anytime and anywhere. This is possible with the advancement of wireless technologies such as GSM, CDMA, and TDMA.

Because of the above-mentioned pressing need for providing ubiquitous Internet services (i.e., wireless data services) for mobile users, at the beginning, different service providers developed different proprietary solutions/protocols. The different solutions/protocols developed are characterized by the limitations of mobile phones or mobile terminals and the network that connect them:

Limited Processor Power and Memory: Being portable, mobile devices are usually not equipped with powerful CPU and large memory capacity. Thus, they are not supposed to execute computation-intensive tasks.

Limited Battery Life: Despite the battery technology is evolving, it is still not able to cope with the needs of users (e.g., a mobile device is typically equipped with many functionalities such as camera and music player, which all consume significant power). Thus, the battery life time of mobile devices is still a concern.

Limited Input and Output Facilities: Mobile devices are typically equipped with input devices of limited capabilities, which are not as convenient as keyboards

and mouse-pointing devices found in desktop computers and laptops. On the other hand, the display of mobile devices is small, which does not allow substantial information to be displayed on the screen each time.

Low Bandwidth: The bandwidth of wireless networks is relatively low (e.g., 9.6 Kbps in GSM), which constrains the amount of information to be delivered to a mobile user.

High Latency: A relatively high latency is resulted due to the cumulative effects of hardware and software factors when a message goes through a wireless network. The message needs to go through various elements of the wireless network, which consist of various physical network devices and radio transmitters. On the other hand, the message also needs to go through various protocol layers.

Unpredictable Availability and Stability: Wireless networks are much less reliable when compared to wired networks. Mobile users might be frequently disconnected from a wireless network due to channel fading, lost radio coverage, etc., when they roam from place to place.

The above features vary from mobile device to mobile device and from network to network. Moreover, different mobile users have different needs. All these make the proprietary solutions/protocols developed for providing Internet services incompatible to each other. In order to solve the incompatibility issue and to develop a universal and open standard, Wireless Application Protocol (WAP) Forum was founded in 1997 by Ericsson, Motorola, Nokia, and Phone.com. WAP 1.1 was published in 1999 and WAP 2.0 was published in 2001. In 2002, the WAP Forum consolidated into the Open Mobile Alliance (OMA) and the specification work from WAP continues within OMA [262].

WAP aims at bridging the gap between the mobile world and the Internet, by bringing diverse Internet content and other data services or value-added services to mobile users across different wireless technologies. It specifies an application environment and a set of communication protocols so that different manufacturers, network providers, content providers, and developers can collaborate to provide mobile devices with the ability to have technology-independent access to the Internet and telephony services. WAP is designed to meet the following requirements:

Interoperability: Any WAP-enabled mobile device is able to communicate with any network supporting WAP services.

Scalability: Protocols and services should scale with number of customers. It allows the personalization of services in accordance with user needs.

Efficiency: Provide QoS according to the characteristics of wireless networks.

Reliability: Provide a consistent and predictable platform for service deployment.

Security: Ensure user data integrity and provide protection to devices and services.

WAP tries to utilize existing Internet protocols and standards as much as possible (e.g., XML, HTML, HTTP, TLS). Those adapted protocols and standards are optimized for deployment in a wireless environment where mobile devices are limited in capabilities, and the wireless network bears the properties of low bandwidth, high latency and unpredictable availability and stability. The following summarizes the key elements of WAP which specify how to improve existing services and how to develop new applications and value-added services in a quick and interoperable manner:

Programming Model: The WAP programming model is similar to the conventional WWW programming model. Since manufacturers, service providers, and developers are familiar with the conventional WWW programming model, it allows a quick introduction of WAP services.

Markup Language: WAP uses wireless markup language (WML), which is a subset of the extensible markup language (XML). WML is designed to fit small mobile devices, while still being able to fulfill the same functionality of HTML.

Scripting Language: WMLScript, which is similar to JavaScript, is used to to define script-type programs for incorporation of enhanced functionalities to WAP-based services.

Small Browser Specification: It specifies a browser suitable for mobile devices having limited processor power, memory, input capability, and small display.

Optimized Protocol Stack: The lightweight protocol stack is designed to address the limitations of wireless devices and the wireless network. It minimizes the bandwidth required and hence maximizes the number of types of wireless networks that are able to support WAP-based services.

Wireless Telephony Application Interface (WTAI): WTAI specifies an application framework for providing telephony services. It enables a WAP-based mobile device to make calls by calling special WMLScript or accessing special URLs.

When a user goes mobile, his concern about retrieving information from the Internet is different from that when he is at home or office. For example, when he is out of town, he might want to keep in touch with his family or colleagues in a quick and stable way through emails, and have a fast and convenient way to keep an eye on the stock market. On viewing emails and browsing the Internet about the stock market, he does need to have fancy images or a mass of unrelated information, which are simply a waste of the valuable bandwidth and incur a high transmission latency. Therefore, WAP is designed to bring Internet services to mobile users in a fast, accurate, and convenient way.

Wireless Markup Language (WML)

Web pages created using HTML are intended to be displayed in machines with larger screens, such as desktop computers and laptops. In addition, today's HTML pages are often created with fancy images or videos, which require relatively more powerful processors for viewing. Thus, HTML pages are not suitable for use in mobile devices with limited processor power and screen. Moreover, such kind of HTML pages are large in size, which are not suitable for delivery in wireless networks of low bandwidth and high latency. In order to address these challenges in providing web-based information services to mobile devices, WML is designed to describe data and the format that data should be presented.

Like HTML, WML is a tagged language. WML adopts a deck and card metaphor. Each WML document is made up of multiple cards, and cards are grouped into a deck. Similar to an HTML page, a deck is the unit of transmission and is identified by a URL. In order to save bandwidth in a wireless network, WML pages can be encoded in a binary format before transmission. When navigating with a WML browser, a user interacts with cards, and moves forward and back through the deck. Figure 15.1 illustrates an example of a deck with two cards.

```
<wml>
  <card id="Card1" title="First Card">
    <p>
    Hello World!
    </p>
  </card>

  <card id="Card2" title="Second Card">
    <p>
    WAP is fun!
    </p>
  </card>
</wml>
```

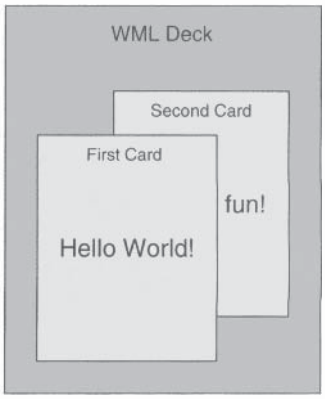


Figure 15.1 A WML deck with two cards.

WMLScript

WMLScript is a scripting langauge which complements WML. WMLScript is designed to work well in a wireless environment, like WML. Scripts written in WMLScript can be compiled into bytecode for efficient transmission over wireless networks. Moreover, the WMLScript bytecode interpreter is compact in size, which allows efficient execution of scripts will less memory and processor requirements.

The key functionalities of using WMLScript are described as follows:

User Interaction and Input Validation: In a WML card, if the user is requested to input some information and then send to a server, then it can save bandwidth and latency if the WML browser can call some WMLScript to verify if the