LECTURE NOTES

ON

13A05801 MOBILE COMPUTING

IV YEAR B.TECH II SEMESTER CSE

SYLLABUS

UNIT-I:Wireless LANS and PANS: Introduction, Fundamentals of WLANS, IEEE 802.11 Standards, HIPERLAN Standard, Bluetooth, Home RF.

Wireless Internet: Wireless Internet, Mobile IP, TCP in Wireless Domain, WAP, Optimizing Web over Wireless.

UNIT-II:

AD HOC Wireless Networks: Introduction, Issues in Ad Hoc Wireless Networks, AD Hoc Wireless Internet.MAC Protocols for Ad Hoc Wireless Networks: Introduction, Issues in Designing a MAC protocol for Ad Hoc Wireless Networks, Design goals of a MAC Protocol for Ad Hoc Wireless Networks, Classifications of MAC Protocols, Contention - Based Protocols, Contention - Based Protocols with reservation Mechanisms, Contention – Based MAC Protocols with Scheduling Mechanisms, MAC Protocols that use Directional Antennas, Other MAC Protocols.

UNIT -III:

Routing Protocols: Introduction, Issues in Designing a Routing Protocol for Ad Hoc Wireless Networks, Classification of Routing Protocols, Table –Driven Routing Protocols, On – Demand Routing Protocols, Hybrid Routing Protocols, Routing Protocols with Efficient Flooding Mechanisms, Hierarchical Routing Protocols, Power – Aware Routing Protocols.

Transport Layer and Security Protocols: Introduction, Issues in Designing a Transport Layer Protocol for Ad Hoc Wireless Networks, Design Goals of a Transport Layer Protocol for Ad Hoc Wireless Networks, Classification of Transport Layer Solutions, TCP Over Ad Hoc Wireless Networks, Other Transport Layer Protocol for Ad Hoc Wireless Networks, Security in Ad Hoc Wireless Networks, Network Security Requirements, Issues and Challenges in Security Provisioning, Network Security Attacks, Key Management, Secure Routing in Ad Hoc Wireless Networks.

UNIT –IV:

Quality of Service: Introduction, Issues and Challenges in Providing QoS in Ad Hoc Wireless Networks, Classification of QoS Solutions, MAC Layer Solutions, Network Layer Solutions, QoS Frameworks for Ad Hoc Wireless Networks.

Energy Management: Introduction, Need for Energy Management in Ad Hoc Wireless Networks, Classification of Ad Hoc Wireless Networks, Battery Management Schemes, Transmission Power Management Schemes, System Power Management Schemes.

UNIT –V:

Wireless Sensor Networks: Introduction, Sensor Network Architecture, Data Dissemination, Data Gathering, MAC Protocols for Sensor Networks, Location Discovery, Quality of a Sensor Network, Evolving Standards, Other Issues.

TEXT BOOKS:

1. Ad Hoc Wireless Networks: Architectures and Protocols - C. Siva Ram Murthy and B.S.Manoj, PHI, 2004.

2. Wireless Ad- hoc and Sensor Networks: Protocols, Performance and Control – Jagannathan Sarangapani, CRC Press

REFERENCE BOOKS:

1. Ad hoc Mobile Wireless Networks – Subir Kumar sarkar, T G Basvaraju, C Puttamadappa, Auerbach Publications,2012.

2. Wireless Sensor Networks - C. S. Raghavendra, Krishna M. Sivalingam, 2004, Springer.

3. Ad- Hoc Mobile Wireless Networks: Protocols & Systems, C.K. Toh , Pearson Education.

UNIT-1: WIRELESS LANS AND PANS

1.1 INTRODUCTION

The field of computer networks has grown significantly in the last three decades. An interesting usage of computer networks is in offices and educational institutions, where tens (sometimes hundreds) of personal computers (PCs) are interconnected, to share resources (*e.g.,* printers) and exchange information, using a high-bandwidth communication medium (such as the Ethernet). These privately-owned networks are known as local area networks (LANs) which come under the category of small-scale networks (networks within a single building or campus with a size of a few kilometers). To do away with the wiring associated with the interconnection of PCs in LANs, researchers have explored the possible usage of radio waves and infrared light for interconnection [1]. This has resulted in the emergence of wireless LANs (WLANs), where wireless transmission is used at the physical layer of the network. Wireless personal area networks (WPANs) are the next step down from WLANs, covering smaller areas with low power transmission, for networking of portable and mobile computing devices such as PCs, personal digital assistants (PDAs), which are essentially very small computers designed to consume as little power as possible so as to increase the lifetime of their batteries, cell phones, printers, speakers, microphones, and other consumer electronics. This chapter highlights the issues involved in the design of WLANs and PANs. It consists of the following sections:

1. Fundamentals of WLANs: The technical issues in WLANs must be understood in order to appreciate the difference between wired networks and wireless networks. The use of WLANs and their design goals are then studied. The types of WLANs, their components, and their basic functionalities are also brought out in this section.

2. IEEE 802.11 Standard: This section introduces a prominent standard in WLANs, the IEEE 802.11 standard. The medium access control (MAC) layer and the physical layer mechanisms are explained here. This section also covers some of the optional functionalities, such as security and quality of service (QoS).

3. HIPERLAN Standard: This section describes another WLAN standard,

HIPER-LAN standard, which is a European standard based on radio access.

4. Bluetooth: This section deals with the Bluetooth standard, which enables personal devices to communicate with each other in the absence of infrastructure.

5. HomeRF: This section discusses the issues in home networking (HomeRF

standard) and finally illustrates the technical differences between Bluetooth, HomeRF, and other technologies such as infrared [portable devices that use the infrared interface of the Infrared Data Association (IrDA) for transmission], which are the current technological alternatives in the PAN area.

1.2 FUNDAMENTALS OF WLANS

This section deals with the fundamental principles, concepts, and requirements of WLANs. This section also brings out WLAN types, their components, and some of their functionalities. In what follows, the terms "node," "station," and "terminal" are used interchangeably. While both portable terminals and mobile terminals can move from one place to another, portable terminals are accessed only when they are stationary. Mobile terminals (MTs), on the other hand, are more powerful, and can be accessed when they are in motion. WLANs aim to support truly mobile work stations.

Technical Issues

Here the technical issues that are encountered in the design and engineering of WLANs are discussed. In particular, the differences between wireless and wired networks, the use of WLANs, and the design goals for WLANs are studied.

Differences Between Wireless and Wired Transmission

• Address is not equivalent to physical location: In a wireless network, address refers to a particular station and this station need not be stationary. Therefore, address may not always refer to a particular geographical location.

• Dynamic topology and restricted connectivity: The mobile nodes may often go out of reach of each other. This means that network connectivity is partial at times.

• Medium boundaries are not well-defined: The exact reach of wireless signals cannot be determined accurately. It depends on various factors such as signal strength and noise levels. This means that the precise boundaries of the medium cannot be determined easily.

• Error-prone medium: Transmissions by a node in the wireless channel are affected by simultaneous transmissions by neighboring nodes that are located within the direct transmission range of the transmitting node. This means that the error rates are significantly higher in the wireless medium. Typical bit error rates (fractions of bits that are received in error) are of the order of 10^{-4} in a wireless channel as against 10^{-9} in fiber optic cables. The above four factors imply that we need to build a reliable network on top of an inherently unreliable channel. This is realized in practice by having reliable protocols at the MAC layer, which hide the unreliability that is present in the physical layer.

Use of WLANs

Wireless computer networks are capable of offering versatile functionalities. WLANs are very flexible and can be configured in a variety of topologies based on the application. Some possible uses of WLANs are mentioned below.

• Users would be able to surf the Internet, check e-mail, and receive Instant Messages on the move.

• In areas affected by earthquakes or other such disasters, no suitable infrastructure may be available on the site. WLANs are handy in such locations to set up networks on the fly.

• There are many historic buildings where there has been a need to set up computer networks. In such places, wiring may not be permitted or the building design may not be conducive to efficient wiring. WLANs are very good solutions in such places.

Design Goals :

The following are some of the goals which have to be achieved while designing

WLANs:

• Operational simplicity: Design of wireless LANs must incorporate features to enable a mobile user to quickly set up and access network services in a simple and efficient manner.

• Power-efficient operation: The power-constrained nature of mobile computing devices such as laptops and PDAs necessitates the important requirement of WLANs operating with minimal power consumption. Therefore, the design of WLAN must incorporate power-saving features and use appropriate technologies and protocols to achieve this.

• License-free operation: One of the major factors that affects the cost of wireless access is the license fee for the spectrum in which a particular wireless access technology operates. Low

cost of access is an important aspect for popularizing a WLAN technology. Hence the design of WLAN should consider the parts of the frequency spectrum (*e.g.,* ISM band) for its operation which do not require an explicit licensing.

• Tolerance to interference: The proliferation of different wireless networking technologies both for civilian and military applications and the use of the microwave frequency spectrum for non-communication purposes

(*e.g.,* microwave ovens) have led to a significant increase in the interference

level across the radio spectrum. The WLAN design should account for this and take appropriate measures by way of selecting technologies and protocols to operate in the presence of interference.

• Global usability: The design of the WLAN, the choice of technology, and the selection of the operating frequency spectrum should take into account the prevailing spectrum restrictions in countries across the world. This ensures the acceptability of the technology across the world.

• Security: The inherent broadcast nature of wireless medium adds to the requirement of security features to be included in the design of WLAN technology.

• Safety requirements: The design of WLAN technology should follow the safety requirements that can be classified into the following:

(i) interference to medical and other instrumentation devices and (ii) increased power level of transmitters that can lead to health hazards. A well-designed WLAN should follow the power emission restrictions that are applicable in the given frequency spectrum.

• Quality of service requirements : Quality of service (QoS) refers to the provisioning of designated levels of performance for multimediaia traffic. The design of WLAN should take into consideration the possibility of supporting a wide variety of traffic, including multimedia ia traffic.

• Compatibility with other technologies and applications: The interoperability among the different LANs (wired or wireless) is important for efficient communication between hosts operating with

different LAN technologies. In addition to this, interoperability with existing WAN protocols such as TCP/IP of the Internet is essential to provide a seamless communication across the WANs.

Network Architecture:

This section lists the types of WLANs, the components of a typical WLAN, and the services offered by a WLAN.

Infrastructure Based Versus Ad Hoc LANs:

WLANs can be broadly classified into two types, infrastructure networks and ad hoc LANs, based on the underlying architecture. Infrastructure networks contain special nodes called *access points* (APs), which are connected via existing networks. APs are special in the sense that they can interact with wireless nodes as well as with the existing wired network. The other wireless nodes, also known as mobile stations (STAs), communicate via APs. The APs also act as bridges with other networks. Ad hoc LANs do not need any fixed infrastructure. These networks can be set up on the fly at any place. Nodes communicate directly with each other or forward messages through other nodes that are directly accessible.

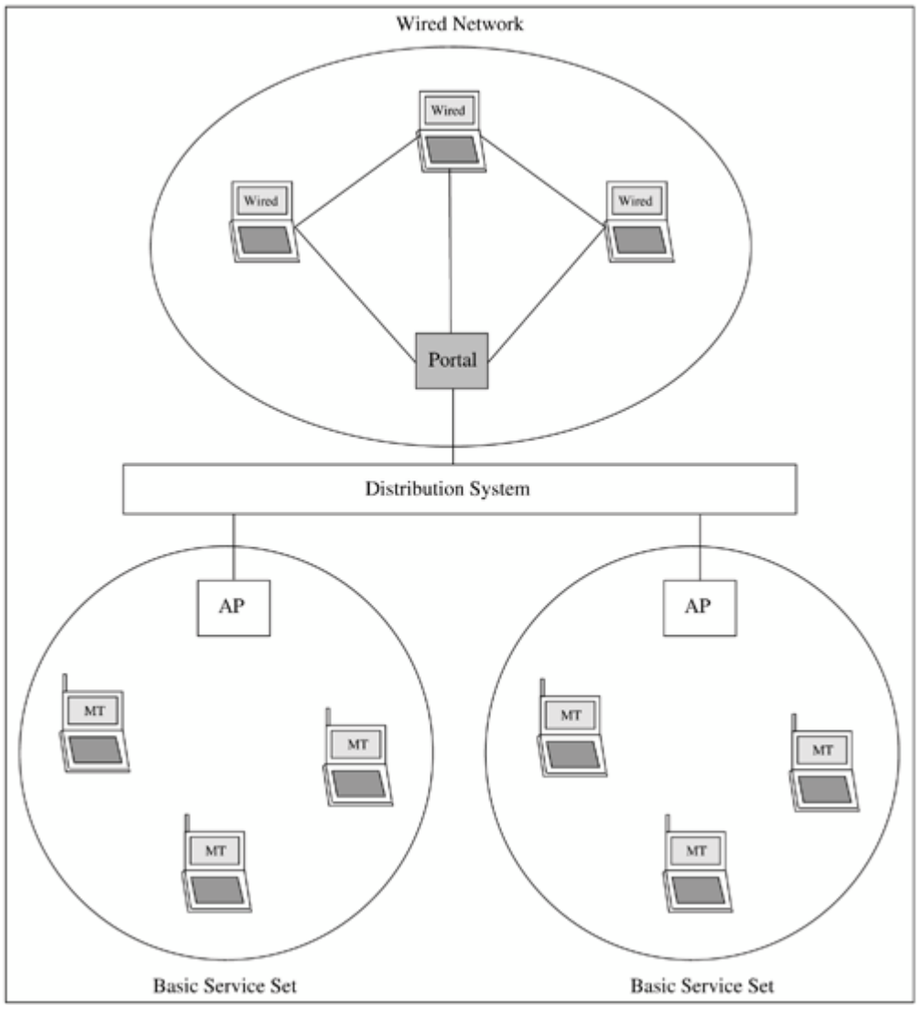
Components in a Typical IEEE 802.11 Network :

IEEE 802.11 is the most popular WLAN standard that defines the specification for the physical and MAC layers. The success of this standard can be understood from the fact that

the revenue from the products based on this standard touched $730 million in the second quarter of the year 2003. The principles and mechanisms followed in this standard are explained later. In what follows, the basic components in a typical IEEE 802.11 WLAN [2] are listed. The set of stations that can remain in contact (*i.e.*, are associated) with a given AP is called a basic service set (BSS). The coverage area of an AP within which member stations (STAs or MTs) may remain in communication is called the basic service area (BSA). The stations that are a part of a BSS need to be located within the BSA of the corresponding AP. A BSS is the basic building block of the network. BSSs are connected by means of a distribution system (DS) to form an extended network. DS refers to an existing network infrastructure. The implementation of the DS is not specified by the IEEE 802.11 standard. The services of the DS, however, are specified rigidly. This gives a lot of flexibility in the design of the DS. The APs are connected by means of the DS.

Portals are logical points through which non-IEEE 802.11 packets (wired LAN packets) enter the system. They are necessary for integrating wireless networks with the existing wired networks. Just as an AP interacts with the DS as well as the wireless nodes, the portal interacts with the wired network as well as with the DS. The BSSs, DS, and the portals together with the stations they connect constitute the extended service set (ESS). An ad hoc LAN has only one BSS. Therefore, ad hoc LANs are also known as independent basic service sets (IBSSs). It may be noted that the ESS and IBSS appear identical to the logical link control (LLC). Figure below gives a schematic picture of what a typical ESS looks like.

Figure Extended Service Set.



Services Offered by a Typical IEEE 802.11 Network:

The services offered by a typical IEEE 802.11 network can be broadly divided into two categories: AP services and STA services. The following are the AP services, which are provided by the DS:

• Association: The identity of an STA and its address should be known to the AP before the STA can transmit or receive frames on the WLAN. This is done during association, and the information is used by the AP to facilitate routing of frames.

•Re association : The established association is transferred from one AP to another using re association. This allows STAs to move from one BSS to another.

• Disassociation: When an existing association is terminated, a notification is issued by the STA or the AP. This is called disassociation, and is done when nodes leave the BSS or when nodes shut down.

• Distribution: Distribution takes care of routing frames. If the destination is in the same BSS, the frame is transmitted directly to the destination, otherwise the frame is sent via the DS.

• Integration: To send frames through non-IEEE 802.11 networks, which may have different addressing schemes or frame formats, the integration service is invoked.

The following are the STA services, which are provided by every station, including APs:

• Authentication: Authentication is done in order to establish the identity of stations to each other. The authentication schemes range from relatively insecure handshaking to public-key encryption schemes.

• De authentication: De authentication is invoked to terminate existing authentication.

• Privacy: The contents of messages may be encrypted (say, by using

the WEP algorithm, which is explained later) to prevent eavesdroppers from reading the messages.

• Data delivery: IEEE 802.11 naturally provides a way to transmit and receive data. However, like Ethernet, the transmission is not guaranteed to be completely reliable.

1.3 IEEE 802.11 STANDARD

After the fundamental issues in WLANs are clearly understood, the reader is in a position to appreciate the *de facto* standards for WLANs. IEEE 802.11 is a prominent standard for WLANs, which is adopted by many vendors of WLAN products. A later version of this standard is the IEEE 802.11b, commercially known as *Wi-Fi* (wireless fidelity). The IEEE 802.11 standard, which deals with the physical and MAC layers in WLANs, was brought out in 1997. This standard is explained in this section.

It may be observed that IEEE 802.11 was the first WLAN standard that faced the challenge of organizing a systematic approach for defining a standard for wireless wideband local access (small-scale networks capable of transmitting data at high rates). As mentioned earlier, in contrast to other LAN standards, wireless standards need to have provisions to support mobility of nodes.

The IEEE802.11 working group had to examine connection management, link reliability management, and power management — none of which was a concern for other standards in IEEE 802. In addition, provision for security had to be introduced. For all these reasons and because of several competing proposals, it took nearly ten years for the development of IEEE 802.11, which was much longer compared to the time taken for the development of other

802 standards for the wired media. Once the overall picture and the ideas became clear, it took only a reasonable duration of time to develop the IEEE 802.11a and IEEE 802.11b enhancements. Under the IEEE 802.11 standard, MTs can operate in two modes: (i) *infrastructure mode,* in which MTs can communicate with one or more APs which are connected to a WLAN, and (ii) *ad hoc mode,* in which MTs can communicate directly with each other without using an AP.

Physical Layer

IEEE 802.11 supports three options for the medium to be used at the physical level — one is based on infrared and the other two are based on radio transmission. The physical layer is subdivided conceptually into two parts — physical medium dependent sub layer (PMD) and physical layer convergence protocol (PLCP). PMD handles encoding, decoding, and modulation of signals and thus deals with the idiosyncrasies of the particular medium. The PLCP abstracts the functionality that the physical layer has to offer to the MAC layer. PLCP offers a service access point (SAP) that is independent of the transmission technology, and a clear channel assessment (CCA) carrier sense signal to the MAC layer. The SAP abstracts the channel which can offer up to 1 or 2 Mbps data transmission bandwidth. The CCA is used by the MAC layer to implement the CSMA/CA mechanism. The three choices for the physical layer in the original 802.11 standard are as follows: (i) frequency hopping spread spectrum (FHSS) operating in the license-free 2.4 GHz industrial, scientific, and medical (ISM) band, at data rates of 1 Mbps [using 2- level Gaussian frequency shift keying (GFSK) modulation scheme] and 2 Mbps (using 4-level GFSK); (ii) direct sequence spread spectrum (DSSS) operating in the 2.4 GHz ISM band, at data rates of 1 Mbps [using differential binary phase shift keying (DBPSK) modulation scheme] and 2 Mbps [using differential quadrature phase shift keying (DQPSK)]; (iii) infrared operating at wavelengths in 850-950 nm range, at data rates of 1 Mbps and 2 Mbps using pulse position modulation (PPM) scheme.

Carrier Sensing Mechanisms:

In IEEE 802.3, sensing the channel is very simple. The receiver reads the peak voltage on the cable and compares it against a threshold. In contrast, the mechanism employed in IEEE 802.11 is relatively more complex. It is performed either physically or virtually. As mentioned earlier, the physical layer sensing is through the clear channel assessment (CCA) signal provided by the PLCP in the physical layer of the IEEE 802.11. The CCA is generated based on sensing of the air interface either by sensing the detected bits in the air or by checking the received signal strength (RSS) of the carrier against a threshold. Decisions based on the detected bits are made somewhat more slowly, but they are more reliable. Decisions based on the RSS can potentially create a false alarm caused by measuring the level of interference.

Basic MAC Layer Mechanisms

This section describes the MAC layer as specified by the IEEE 802.11 standard. The primary function of this layer is to arbitrate and statistically multiplex the transmission requests of various wireless stations that are operating in an area. This assumes importance because wireless transmissions are inherently broadcast in nature and contentions to access the shared channel need to be resolved prudently in order to avoid collisions, or at least to reduce the number of collisions. The MAC layer also supports many auxiliary functionalities such as offering support for roaming, authentication, and taking care of power conservation.

The basic services supported are the mandatory asynchronous data service and an optional real-time service. The asynchronous data service is supported for unicast packets as well as for multicast packets. The real-time service is supported only in infrastructure-based networks where APs control access to the shared medium.

Distributed Foundation Wireless Medium Access Control (DFWMAC)

The primary access method of IEEE 802.11 is by means of a distributed coordination function (DCF). This mandatory basic function is based on a version of carrier sense with multiple access and collision avoidance (CSMA/CA). To avoid the hidden terminal problem (which is explained later), an optional RTS-CTS mechanism is implemented. There is a second method called the point coordination function (PCF) that is implemented to provide real-time services. When the PCF is in operation, the AP controls medium access and avoids

simultaneous transmissions by the nodes.

Inter-Frame Spacing (IFS) :

Inter-frame spacing refers to the time interval between the transmission of two successive frames by any station. There are four types of IFS: SIFS, PIFS, DIFS, and EIFS, in order from shortest to longest. They denote priority levels of access to the medium. Shorter IFS denotes a higher priority to access the medium, because the wait time to access the medium is lower. The exact values of the IFS are obtained from the attributes specified in the physical layer management information base (PHYMIB) and are independent of the station bit rate.

• **Short inter-frame spacing** (**SIFS**) is the shortest of all the IFSs and denotes highest priority to access the medium. It is defined for short control messages such as acknowledgments for data packets and polling responses. The transmission of any packet should begin only after the channel is sensed to be idle for a minimum time period of at least SIFS.

• **PCF inter-frame spacing (PIFS)** is the waiting time whose value lies between SIFS and DIFS. This is used for real-time services.

• **DCF inter-frame spacing (DIFS)** is used by stations that are operating under the DCF mode to transmit packets. This is for asynchronous data transfer within the contention period.

• **Extended inter-frame spacing (EIFS)** is the longest of all the IFSs and denotes the east priority to access the medium. EIFS is used for resynchronization whenever physical layer detects incorrect MAC frame reception.

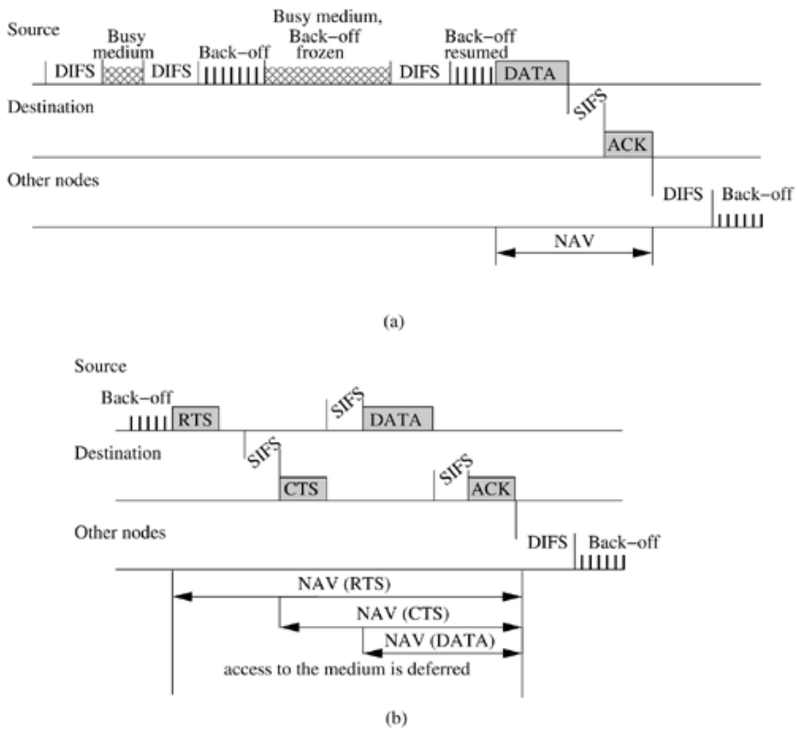
CSMA/CA Mechanism

Carrier sense with multiple access and collision avoidance (CSMA/CA) is the MAC layer mechanism used by IEEE 802.11 WLANs. Carrier sense with multi le access and collision detection (CSMA/CD) is a well-studied technique in IEEE 802.x wired LANs. This technique cannot be used in the context of WLANs effectively because the error rate in WLANs is much higher and allowing collisions will lead to a drastic reduction in throughput. Moreover, detecting collisions in the wireless medium is not always possible. The technique adopted here is therefore one of collision avoidance.

The Medium Access Mechanism:

The basic channel access mechanism of IEEE 802.11 is shown in Figure If the medium is sensed to be idle for a duration of DIFS, the node accesses the medium for transmission. Thus the channel access delay at very light loads is equal to the DIFS. If the medium is busy, the node *backs off,* in which the station defers channel access by a random amount of time chosen within a *contention window*(*CW*). The value of *CW* can vary between CW_{min} and CW_{max}. The time intervals are all integral multiples of slot times, which are chosen judiciously using propagation delay, delay in the transmitter, and other physical layer dependent parameters. As soon as the back-off counter reaches zero and expires, the station can access the medium. During the back-off process, if a node detects a busy channel, it freezes the back-off counter and the process is resumed once the channel becomes idle for a period of DIFS. Each station executes the back-off procedure at least once between every successive transmission.

IEEE 802.11 DCF and RTS-CTS mechanism.



In the scheme discussed so far, each station has the same chances for transmitting data next time, independent of the overall waiting time for transmission. Such a system is clearly unfair. Ideally, one would like to give stations that wait longer a higher priority service in order to ensure that they are not starved. The back-off timer incorporated into the above mechanism tries to make it fair. Longer waiting stations, instead of choosing another random interval from the contention window, wait only for a residual amount of time that is specified by the back-off timer.

Contention Window Size

The size of the Contention Window (CW) is another important parameter. If the CW is small in size, then the random values will be close together and there is a high probability of packet collision. On the other hand, if the size of CW is very large, there will be some unnecessary delay because of large back-off values. Ideally, one would like the system to adapt to the current number of stations that are contending for channel access. To effect this, the truncated binary exponential back-off technique is used here, which is similar to the technique used in IEEE 802.3. The initial contention window is set to a random value between (0, CWmin) and each time a collision occurs, the CW doubles its size up to a maximum of CWmax. So at high load, the CW size is high and therefore the resolution power of the system is high. At low loads, small CW ensures low access delay. The specified values of CW min and CWmax for different physical layer specifications are given in Below Table

Table IEEE 802.11 parameters

Parameter	**802.11 (FHSS)**	**802.11 (DSSS)**	**802.11 (IR)**	**802.11b**	**802.11a**
t_{slot}	50 μsec	20 μsec	8 μsec	20 μsec	9 μsec
SIFS	28 μsec	10 μsec	10 μsec	10 μsec	16 μsec
PIFS	SIFS $+t_{slot}$				
DIFS	SIFS$+(2 \times t_{slot})$				
Operating Frequency	2.4 GHz	2.4 GHz	850-950 nm	2.4 GHz	5 GHz
Maximum Data Rate	2 Mbps	2 Mbps	2 Mbps	11 Mbps	54 Mbps
CWmin	15	31	63	31	15
CWmax	1,023	1,023	1,023	1,023	1,023

Acknowledgments

Acknowledgments (ACKs) must be sent for data packets in order to ensure their correct delivery. For uncast packets, the receiver accesses the medium after waiting for a SIFS and sends an ACK. Other stations have to wait for DIFS plus their back off time. This reduces the probability of a collision. Thus higher priority is given for sending an ACK for the previously received data packet than for starting a new data packet transmission. ACK ensures the correct reception of the MAC layer frame by using cyclic redundancy checksum (CRC) technique. If no ACK is received by the sender, then a retransmission takes place. The number of retransmissions is limited, and failure is reported to the higher layer after the retransmission count exceeds this limit.

RTS-CTS Mechanism

The *hidden terminal problem* is a major problem that is observed in wireless networks. This is a classic example of problems arising due to incomplete topology information in wireless networks that was mentioned initially. It also highlights the non-transitive nature of wireless transmission. In some situations, one node can receive from two other nodes, which cannot hear each other. In such cases, the receiver may be bombarded by both the senders, resulting in collisions and reduced throughput. But the senders, unaware of this, may get the impression that the receiver can clearly listen to them without interference from anyone else. This is called the hidden terminal problem. To alleviate this problem, the RTS-CTS mechanism has been devised as shown in Figure

How RTS-CTS Works

The sender sends a request to send (RTS) packet to the receiver. The packet includes the receiver of the next data packet to be transmitted and the expected duration of the whole data transmission. This packet is received by all stations that can hear the sender. Every station that receives this packet will set its *network allocation vector* (NAV) accordingly. The NAV of a station specifies the earliest time when the station is permitted to attempt transmission. After waiting for SIFS, the intended receiver of the data packet answers with a clear to send (CTS) packet if it is ready to accept the data packet. The CTS packet contains the duration field, and all stations receiving the CTS packet also set their NAVs. These stations are within the transmission range of the receiver. The set of stations receiving the CTS packet may be different from the set of stations that received the RTS packet, which indicates the presence of some hidden terminals. Once the RTS packet has been sent and CTS packet has been received successfully, all nodes within receiving distance from the sender and from the receiver are informed that the medium is reserved for one sender exclusively. The sender then starts data packet transmission after waiting for SIFS. The receiver, after receiving the packet, waits for another SIFS and sends the ACK. As soon as the transmission is over, the NAV in each node marks the medium as free (unless the node has meanwhile heard some other RTS/CTS) and the process can repeat again. The RTS packet is like any other packet and collisions can occur only at the beginning when RTS or CTS is being sent. Once the RTS and CTS packets are transmitted successfully, nodes that listen to the RTS or the CTS refrain from causing collision to the ensuing data transmission, because of their NAVs which will be set. The usage of RTS-CTS dialog before data packet transmission is a form of *virtual carrier sensing.*

Overhead Involved in RTS-CTS

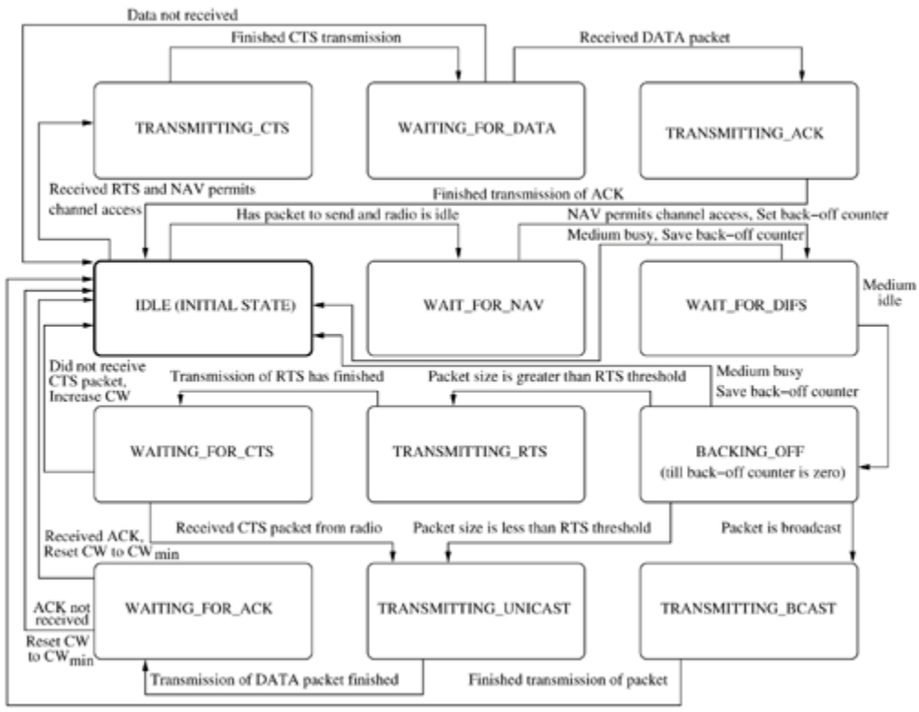
It can be observed that the above mechanism is akin to reserving the medium prior to a particular data transfer sequence in order to avoid collisions during this transfer. But transmission of RTS-CTS can result in non-negligible overhead. Therefore, the RTS-CTS mechanism is used judiciously.

An RTS threshold is used to determine whether to start the RTS- CTS mechanism or not. Typically, if the frame size is more than the RTS threshold, the RTS-CTS mechanism is activated and a four-way handshake (*i.e.,* RTS-CTS-DATA-ACK) follows. If the frame size is below the RTS threshold, the nodes resort to a two-way handshake (DATA-ACK).

MAC as a State Machine

Figure diagrammatically shows what has been discussed so far. It models the MAC layer as a finite state-machine, and shows the permissible transitions. It must be noted that the state-machine is simplistic and is given only to ease the understanding of the fundamental mechanisms at the MAC layer. The functioning of the finite state-machine is explained in what follows.

Figure MAC state transition diagram.



If a node has a packet to send and is in the IDLE state, it goes into the WAIT_FOR_NAV state. After the on-going transmissions (if any) in the neighborhood are over, the node goes to the WAIT_FOR_DIFS state. After waiting for DIFS amount of time, if the medium continues to be idle, the station enters the BACKING_OFF state. Otherwise, the station sets its back-off counter (if the counter value is zero) and goes back to the IDLE state. During back-off, if the node senses a busy channel, the node saves the back-off counter and goes back to the IDLE state. Otherwise, it goes into one of three states. If the packet type is broadcast, the node enters the TRANSMITTING_BCAST state where it transmits the broadcast packet. If the packet type is uncast and the packet size is less than the RTS threshold, the node enters the TRANSMITTING_UNICAST

state and starts transmitting data. If the packet size is greater than the RTS threshold, the node enters the TRANSMITTING_RTS state and starts transmitting the RTS packet. After the RTS transmission is over, the node enters the WAITING_FOR_CTS state. If the CTS packet is not received within a specified time, the node times out and goes back to the IDLE state, and increases the *CW* value exponentially up to a maximum of CW_{max}. If

the CTS packet is received, the node enters the TRANSMITTING_UNICAST state and starts transmitting data. After the uncast packet is transmitted, the node enters the WAITING_FOR_ACK state. When the node receives the ACK, it goes back to the IDLE state and reduces the CW value to CW_{min}.

If a node receives an RTS packet when in IDLE state and if the NAV of the node indicates that no other on-going transmissions exist, the node enters the TRANSMITTING_CTS state and starts transmitting the CTS packet. After the CTS packet is transmitted, the node enters the WAITING_FOR_DATA state and waits for the data packet from the sender. On receiving the data packet, the node enters the TRANSMITTING_ACK state and starts transmitting the ACK for the data packet. When the ACK has been transmitted, the node goes back to the IDLE state. If the data packet is not received, the receiver returns to the IDLE state.

Fragmentation

Bit error rates in the wireless medium are much higher than in other media. The bit

error rate in fiber optics is only about 10^{-9}, whereas in wireless, it is as large as 10^{-4}. One way of decreasing the frame error rate is by using shorter frames. IEEE 802.11 specifies a fragmentation mode where user data packets are split into several smaller parts transparent to the user. This will lead to shorter frames, and frame error will result in retransmission of a shorter frame. The RTS and CTS messages carry duration values for the current fragment and estimated time for the next fragment. The medium gets reserved for the successive frames until the last fragment is sent.

The length of each fragment is the same for all the fragments except the last fragment. The fragments contain information to allow the complete MAC protocol data unit (MPDU, informally referred to as packet) to be reassembled from the fragments that constitute it. The frame type, sender address, destination address, sequence control field, and indicator for more fragments to come are all present in the fragment header. The destination constructs the complete packet by reassembling the fragments in the order of the sequence number field. The receiving station ensures that all duplicate fragments are discarded and only one copy of each fragment is integrated. Acknowledgments for the duplicates may, however, be sent.

Other MAC Layer Functionalities

There are several other functionalities that the MAC layer provides in IEEE 802.11 WLANs. The functionalities described in this section are the point coordination function (PCF) which is used for QoS guarantees, timing synchronization, power management, and support for roaming.

Point Coordination Function

The objective of the point coordination function (PCF) is to provide guarantees on the maximum access delay, minimum transmission bandwidth, and other QoS parameters. Unlike the DCF, where the medium contention is resolved in a distributed manner, the PCF works by effecting a centralized contention resolution scheme, and is applicable only in networks where an AP polls the nodes in its BSS. A point coordinator (PC) at the AP splits the access time into super frame periods. The super frame period consists of alternating contention free periods (CFPs) and contention periods (CPs). The PC will determine which station has the right to transmit at any point of time. The PCF is essentially a polled service with the PC playing the role of the polling master. The operation of the PCF may require additional coordination to perform efficient operation in cases where multiple PCs are operating simultaneously such that their transmission ranges overlap. The IFS used by the PCF is smaller than the IFS of the frames transmitted by the DCF. This means that point-coordinated traffic will have higher priority access to the medium if DCF and PCF are concurrently in action. The PC controls frame transmissions so that contentions are eliminated over a limited period of time, that is, the CFP.

Synchronization

Synchronization of clocks of all the wireless stations is an important function to be performed by the MAC layer. Each node has an internal clock, and clocks are all synchronized by a timing synchronization function (TSF). Synchronized clocks are required for power management, PCF coordination, and frequency hopping spread spectrum (FHSS) hopping sequence synchronization. Without synchronization, clocks of the various wireless nodes in the network may not have a consistent view of the global time. Within a BSS, quasi periodic beacon frames are transmitted by the AP, that is, one beacon frame is sent every target beacon transmission time (TBTT) and the transmission of a beacon is deferred if the medium is busy. A beacon contains a time-stamp that is used by the node to adjust its clock. The beacon also contains some management information for power optimization and roaming. Not all beacons need to be heard for achieving synchronization.

Power Management:

Usage of power cords restricts the mobility that wireless nodes can potentially offer. The usage of battery-operated devices calls for power management because battery power is expensive. Stations that are always ready to receive data consume more power (the receiver current may be as high as 100 mA). The transceiver must be switched off whenever carrier

sensing is not needed. But this has to be done in a manner that is transparent to the existing protocols. It is for this reason that power management is an important functionality in the MAC layer. Therefore, two states of the station are defined: sleep and awake. The sleep state refers to the state where the transceiver cannot receive or send wireless signals. Longer periods in the sleep state mean that the average throughput will be low. On the other hand, shorter periods in the sleep state consume a lot of battery power and are likely to reduce battery life. If a sender wants to communicate with a sleeping station, it has to buffer the data it wishes to send. It will have to wait until the sleeping station wakes up, and then send the data. Sleeping stations wake up periodically, when senders can announce the destinations of their buffered data frames. If any node is a destination, then that node has to stay awake until the corresponding transmission takes place.

Roaming:

Each AP may have a range of up to a few hundred meters where its transmission will be heard well. The user may, however, walk around so that he goes from the BSS of one AP to the BSS of another AP. Roaming refers to providing uninterrupted service when the user walks around with a wireless station. When the station realizes that the quality of the current link is poor, it starts scanning for another AP. This scanning can be done in two ways: active scanning and passive scanning. Active scanning refers to sending a probe on each channel and waiting for a response. Passive scanning refers to listening into the medium to find other networks. The information necessary for joining the new BSS can be obtained from the beacon and probe frames.

Other Issues

Improvements in the IEEE 802.11 standard have been proposed to support higher data rates for voice and video traffic. Also, QoS provisioning and security issues have been addressed in extended versions of the standard. These will be discussed in the remainder of this section.

Newer Standards

The original standards for IEEE 802.11 came out in 1997 and promised a data rate of 1-2 Mbps in the license-free 2.4 GHz ISM band . Since then, several improvements in technology have called for newer and better standards that offer higher data rates. This has manifested in the form of IEEE802.11a and IEEE 802.11b standards, both of which came out in 1999. IEEE 802.11b, an extension of IEEE 802.11 DSSS scheme, defines operation. 2.4GHz ISM band at data rates of 5.5 Mbps and 11 Mbps, and is trademarked commercially by the Wireless Ethernet Compatibility Alliance (WECA) as Wi- Fi. It achieves high data rates due to the use of complimentary code keying (CCK). IEEE 802.11a operates in the 5 GHz band (unlicensed national information infrastructure band), and uses orthogonal frequency division multiplexing (OFDM) at the physical layer. IEEE 802.11a supports data rates up to 54 Mbps and is the fast Ethernet analogue to IEEE 802.11b. Other IEEE 802.11 (c, d, and h) task groups are working on special regulatory and networking issues. IEEE 802.11e deals with the requirements of time- sensitive applications such as voice and video. IEEE802. 11f deals with inter- AP communication to handle roaming. IEEE 802.11g aims at providing the high speed of IEEE 802.11a in the ISM band. IEEE 802.11i deals with advanced encryption standards to support better privacy.

QoS for Voice and Video Packets

In order to offer QoS, delay-sensitive packets (such as voice and video packets) are to be given a higher priority to get ahead of less time-critical (*e.g.,* file transfer) traffic. Several mechanisms have been proposed to offer weighted priority. Hybrid coordination function (HCF) can be used where the AP polls the stations in a weighted way in order to offer QoS. Extended DCF is another mechanism which has been proposed where the higher priority stations will choose the random back-off interval from a smaller CW. Performance of WLANs where voice and data services are integrated.

Wired Equivalent Privacy

Security is a very important issue in the design of WLANs. In order to provide a modest level of physical security, the wired equivalent privacy (WEP) mechanism was devised. The

name WEP implies that this mechanism is aimed at providing the level of privacy that is equivalent to that of a wired LAN. Data integrity, access control, and confidentiality are the three aims of WEP. It assumes the existence of an external key management service that distributes the key sequence used by the sender. This mechanism relies on the fact that the secret key cannot be determined by brute force. However, WEP has been proven to be vulnerable if more sophisticated mechanisms are used to crack the key. It uses the pseudo-random number key generated by RSA RC4 algorithm which has been efficiently implemented in hardware as well as in software. This mechanism makes use of the fact that if we take the plain text, XOR (bit-by-bit exclusive OR) it with a pseudo-random key sequence, and then XOR the result with the same key sequence, we get back the plain text.

1.4 HIPERLAN STANDARD

The European counterparts to the IEEE 802.11 standards are the high- performance radio LAN(HIPER LAN) standards defined by the European Telecommunications Standards Institute (ETSI). It is to be noted that while the IEEE 802.11 standards can use either radio access or infrared access, the HIPERLAN standards are based on radio access only. The standards have been defined as part of the ETSI broadband radio access networks (BRAN) project. In general, broadband systems are those in which user data rates are greater than 2 Mbps (and can go up to 100s of Mbps). Four standards have been defined for wireless networks by the ETSI.

• HIPERLAN/1 is a wireless radio LAN (RLAN) without a wired infrastructure, based on one-to-one and one-to-many broadcasts. It can be used as an extension to a wired infrastructure, thus making it suited to both ad hoc and infrastructure- based networks. It employs the 5.15 GHz and the 17.1 GHz frequency bands and provides a maximum data rate of 23.5 Mbps.

• The HIPERLAN/2 standard intends to provide short-range (up to 200 m) wireless access to Internet protocol (IP), asynchronous transfer mode (ATM[1]), and other infrastructure-based networks –and, more importantly, to integrate WLANs into cellular systems. It employs the 5 GHz frequency band and offers a wide range of data rates from 6 Mbps to 54 Mbps. HIPERLAN/2 has been designed to meet the requirements of future wireless multimediaia services.

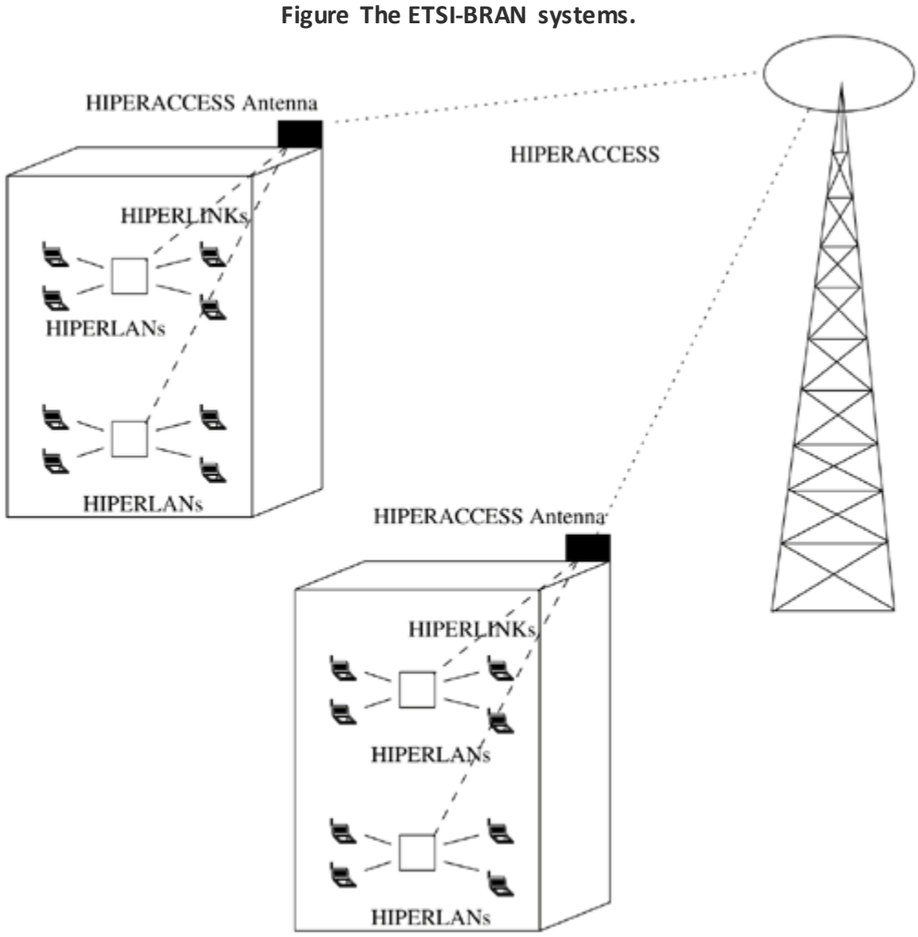
-ATM networks are connection-oriented and require a connection to set up prior to transfer of information from a source to a destination. All information to be transmitted — voice, data, image, and video — is first fragmented into small, fixed-size packets known as cells. These cells are then switched and routed using packet switching principles.

• HIPERACCESS (originally called HIPERLAN/3) covers "the last mile" to the customer; it enables establishment of outdoor high-speed radio access networks, providing fixed radio connections to customer premises.

HIPERACCESS provides a data rate of 25 Mbps. It can be used to connect HIPERLAN/2 deployments that are located far apart (up to 5 Km away). It offers point-to- multipoint communication n.

• The HIPERLINK (originally called HIPERLAN/4) standard provides high speed radio links for point-to-point static interconnections. This is used to connect different HIPERLAN access points or HIPERACCESS networks with high-speed links over short distances of up to 150 m. For example, the HIPERLINK can be employed to provide links between different rooms or floors within a large building. HIPERLINK operates on the 17 GHz frequency range. Figure shows a typical deployment of the ETSI standards. The standards excluding HIPERLAN/1 are grouped under the BRAN project. The scope of the BRAN has been to standardize the radio access network and the functions that serve as the interface to the infrastructural networks.

Figure The ETSI-BRAN systems.



HIPERLAN/1

HIPERLAN/1 is a RLAN standard that was introduced by the ETSI in 1995. The standard allows nodes to be deployed either in a pre-arranged or in an ad hoc fashion. Apart from supporting node mobility, HIPERLAN/1 provides forwarding mechanisms (multi-hop routing). Thus, coverage is not limited to just the neighboring nodes. Using a clever framing scheme as explained later in this section, HIPERLAN/1 provides a data rate of around 23.5 Mbps without utilizing much power, thus having the capability to support multimedia ia data and asynchronous data effectively. This data rate is significantly higher than that provided by IEEE 802.11. The HIPERLAN/1 protocol stack is restricted to the two lower-most layers in the OSI reference model: the data link layer (DLL) and the physical layer. The DLL is further divided into the medium access control (MAC) sub layer and the channel access control (CAC) sub layer. The sections that follow describe the standard.

The Physical Layer

The tasks of the physical layer are modulation and demodulation of a radio carrier with a bit stream, forward error-correction mechanisms, signal strength measurement, and synchronization between the sender and the receiver. The standard uses the CCA scheme (similar to IEEE 802.11) to sense whether the channel is idle or busy.

The MAC Sub layer

The HIPERLAN/1 MAC (HM) sublayer is responsible for processing the packets from the higher layers and scheduling the packets according to the QoS requests from the higher layers specified by the HM QoS parameters.

The MAC sub layer is also responsible for forwarding mechanisms, power conservation schemes, and communication confidentiality through encryption– decryption mechanisms. Because of the absence of an infrastructure, the forwarding mechanism is needed to allow the physical extension of HIPERLAN/1 to go beyond the radio range of a single station. Topology-related data are exchanged between the nodes periodically with the help of special packets, for the purpose of forwarding. In order to guarantee a time-bound service, the HM protocol data unit (HM- PDU) selected for channel access has to reflect the user priority and the residual lifetime of the packet (the time remaining for the packet to expire). The MAC layer computes the channel access priority for each of the PDUs following a mapping from the MAC priority to the channel access mechanism (CAM) priority. One among those PDUs which has the highest CAM priority and the least residual time will be selected for access to the channel.

The CAC Sub layer

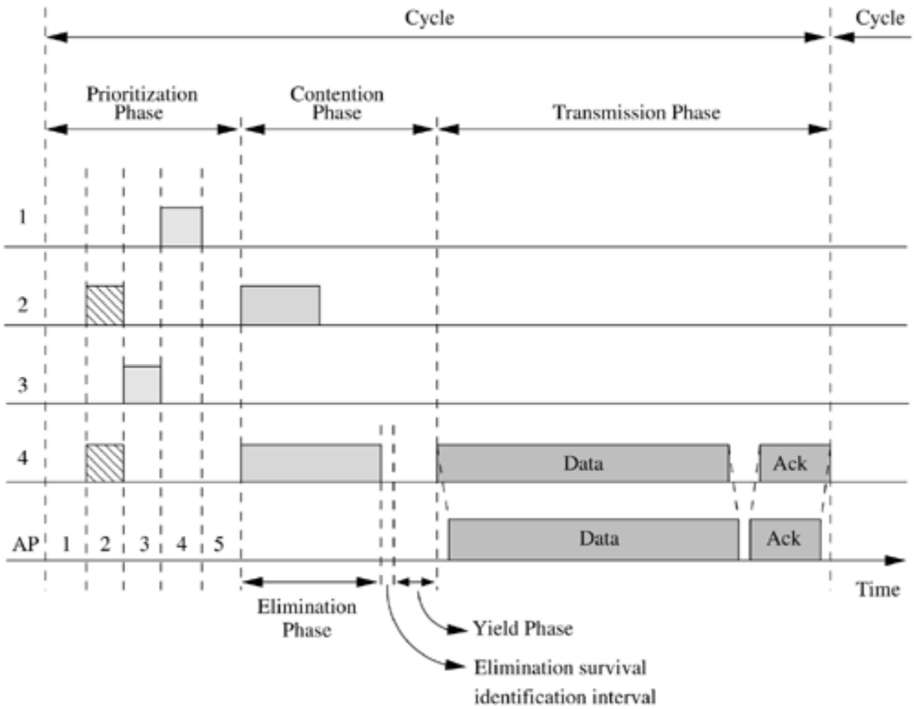
The CAC sub layer offers a connectionless data service to the MAC sub layer. The MAC layer uses this service to specify a priority (called the CAM priority) which is the QoS parameter for the CAC layer. This is crucial in the resolution of contention in the CAM.

EY-NPMA

After a packet with an associated CAM priority has been chosen in the CAC sub layer for transmission,

the next phase is to compete with packets of other nodes for channel access. The channel access mechanism is a dynamic, listen-and-then-talk protocol that is very similar to the CSMA/CA used in802.11 and is called the elimination yield non-preemptive multiple access (EY- NPMA) mechanism. Figures shows the operation of the EY-NPMA mechanism in which the nodes 1, 2, 3, and 4 have packets to be sent to the AP. The CAM priority for nodes 2 and 4 is higher with priority 2 followed by node 3 with priority 3, and node 1 with the least priority of 4. The prioritization phase will have *k* slots where *k* (can vary from 1 to 5 with *k − 1* having higher priority than *k*) refers to the number of priority levels.

The operation of EY-NPMA.



The entire process of channel access occurs in the form of channel access cycles. A synchronization interval occurs after the end of every such cycle. This access cycle is comprised of three phases: prioritization, contention, and transmissin.

1. **Prioritization:** This phase culls out nodes with packets of the highest CAM priority and lets them participate in the next phase. The prioritization phase consists of two events, namely, priority detection and priority assertion. During the priority detection period, a node listens to the channel for a number of time slots proportional to the CAM priority assigned to the packet that the node wants to send. In Figure the nodes 2 and 4 wait for one slot and assert their priority in the second slot as they hold packets with higher priority, and nodes 3 and 1 wait for slots equal to their priority level. By listening to the channel, nodes 3 and 1 detect the existence of other nodes with higher priority and hence leave the prioritization phase. If a low-priority node has succeeded in waiting up to this slot, it enters the priority assertion period during which it sends a burst, signaling its selection to the next stage. In this process, the node(s) with the highest CAM priority will finish the prioritization phase first and hence will be selected for the next phase.

2. **Contention:** This phase is to eliminate as many nodes as possible, in order to minimize the collision rate during transmission. This phase extends to a maximum of 13 slots, each of the same width as that of the slots in the prioritization phase. In this phase, the nodes that transmitted a burst in the previous phase, resolve access to the channel by contention. This phase consists of two sub-phases, namely, the elimination phase and the yield phase. Nodes in this phase (nodes 2 and 4 in Figure) get to transmit a burst for a geometrically distributed number of time slots [the probability of a node's transmission extending to a slot length of *k* slots (where $k < 12$ slots) is 0.5^{k+1}] which is then followed by a sensing period of 1 slot. During this period, if a node detects another node's burst, it stops the contention process (node 2 in Figure). This period during which each contending node will have to listen to the channel for a slot duration is called the elimination survival identification interval. If the channel is sensed idle during this interval, the node reaches the yield phase. This period is also called elimination survival verification. This ensures that the node(s) which sent the elimination burst for the maximum number of slots will be chosen for the next phase. The next phase is the yield phase which complements the eliminationn phase; it involves each node listening to the channel for a number of time slots (up to a maximum of 15 slots, each with duration $\frac{1}{4}^{th}$ of the slot duration in the prioritization phase). This is in fact similar to the back-off state in which the probability of backing off for *k* slots is 0.1×0.9^k. If the channel is sensed to be idle during these slots, the node is said to be eligib le for transmission. The node that waits for the shorter number of slots initiates transmission and other nodes defer their access to the next cycle to begin the process afresh.

3. **Transmission:** This is the final stage in the channel access where the transmission of the selected packet takes place. During this phase, the successful delivery of a data packet is acknowledged with an ACK packet. The performance of EY-NPMA protocol suffers from major factors such as packet length,

number of nodes, and the presence of hidden terminals. The efficiency of this access scheme varies from 8% to 83% with variation of packet sizes from 50 bytes to 2 Kbytes. The above-described channel access takes place during what is known as the channel synchronization condition. The other two conditions during which channel access can take place are (a) the channel free condition, when the node senses the channel free for some amount of time and then gains access, and (b) the hidden terminal condition, when a node is eliminated from contention, but still does not sense any data transmission, indicating the presence of a hidden node.

Power Conservation Issues

The HIPERLAN/1 standard has suggested power conservation schemes at both the MAC and the physical layers.

At the MAC level, the standard suggests awake/sleep modes similar to the DFWMAC in IEEE 802.11. Two roles defined for the nodes are the p-savers (nodes that want to implement the function) and the p-supporters (neighbors to the p-saver that are deputized to aid the latter's power conservation). The p- saver can receive packets only at predetermined time intervals and is active only during those intervals, in the process saving power. At the physical level, a framing scheme has been adopted to conserve power. The physical burst is divided into high bit rate (HBR) and low bit rate (LBR) bursts. The difference between the two bursts lies in the keying mechanisms employed for them – the HBR burst is based on Gaussian minimum shift keying (GMSK) that yields a higher bit rate, but consumes more power than frequency shift keying (FSK) used for the LBR bursts. The LBR burst contains the destination address of the frame and precedes the HBR burst. Any node receiving a packet, first reads the LBR burst. The node will read the HBR burst only if it is the destination for that frame. Otherwise, the burst is simply ignored, thereby saving the power needed to read the HBR burst.

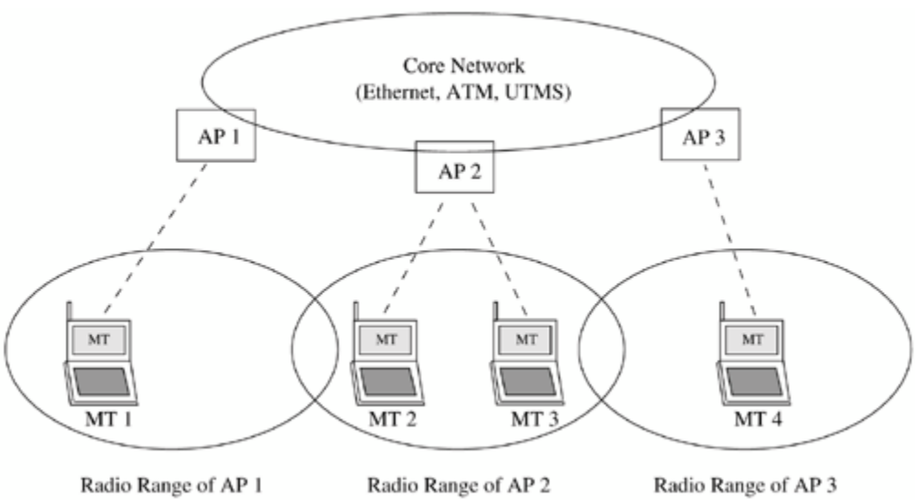
Failure of HIPERLAN/1

In spite of the high data rate that it promised, HIPERLAN/1 standard has always been considered unsuccessful. This is because IEEE Ethernet had been prevalent and hence, for its wireless counterpart too, everybody turned toward IEEE, which came out with its IEEE 802.11 standard. As a result, hardly any manufacturer adopted the HIPERLAN/1 standard for product development. However, the standard is still studied for the stability it provides and for the fact that many of the principles followed have been adopted in the other standards. For further details on the standard.

HIPERLAN/2

As seen earlier, the IEEE 802.11 standard offers data rates of 1 Mbps while the newer standard IEEE802.11a offers rates up to 54 Mbps. However, there was a necessity to support QoS, handoff (the process of transferring an MT from one channel/AP to another), and data integrity in order to satisfy the requirements of wireless LANs. This demand was the motivation behind the emergence of HIPERLAN/2. The standard has become very popular owing to the significant support it has received from cellular manufacturers such as Nokia and Ericsson. The HIPERLAN/2 tries to integrate WLANs into the next-generation cellular systems. It aims at converging IP and ATM type services at a high data rate of 54 Mbps for indoor and outdoor applications. The HIPERLAN/2, an ATM compatible WLAN, is a connection-oriented system, which uses fixed size packets and enables QoS applications easy to implement. The HIPERLAN/2 network has a typical topology as shown in Figure. The figure shows MTs being centrally controlled by the APs which are in turn connected to the core network (infrastructure-based network). It is to be noted that, unlike the IEEE standards, the core network for HIPERLAN/2 is not just restricted to Ethernet. Also, the AP used in HIPERLAN/2 consists of one or many transceivers called access point transceivers (APTs) which are controlled by a single access point controller (APC).

A typical deployment of HIPERLAN/2.



There are two modes of communication in a HIPERLAN/2 network, which are described by the following two environments:

- **Business environment:** The ad hoc architecture of HIPERLAN/1 has been extended to support a centralized mode of communication using APs. This topology corresponds to business environments. Accordingly, each AP serves a number of MTs.
- **Home environment:** The home environment enables a direct mode of communication between the MTs. This corresponds to an ad hoc architecture that can be operated in a plug-and-play manner. The direct mode of communication is, however, managed by a central control entity elected from among the nodes called the central controller (CC). There are several features of HIPERLAN/2 that have attracted many a cellular manufacturer. These features are part of the discussion on the protocol stack of HIPERLAN/2 below. The HIPERLAN/2 protocol stack consists of the physical layer, convergence layer (CL), and the data link control (DLC) layer.

The Physical Layer

The physical layer is responsible for the conversion of the PDU train from the DLC layer to physical bursts that are suitable for radio transmission. HIPERLAN/2, like IEEE 802.11a, uses OFDM for transmission. The HIPERLAN/2 allows bit rates from 6 Mbps to 54 Mbps using a scheme called link adaptation. This scheme allows the selection of a suitable modulation method for the required bit rate. This scheme is unique to HIPERLAN/2 and is not available in the IEEE standards and HIPERLAN/1. More details on the physical layer.

The CL is the topmost layer in the HIPERLAN/2 protocol stack is the CL. The functions of the layer are to adapt the requirements of the different higher layers of the core network with the services provided by the lower layers of HIPERLAN/2, and to convert the higher layer packets into ones of fixed size that can be used by the lower layers. A CL is defined for every type of core network supported. In short, this layer is responsible for the network-independent feature of HIPERLAN/2.

The CL is classified into two types, namely, the packet-based CL and the cell- based CL. The packet-based CL processes variable-length packets (such as IEEE 802.3, IP, and IEEE 1394). The cell-based CL processes fixed- sized ATM cells. The CL has two sub layers, namely, the common part (CP) and the service-specific convergence sub layer (SSCS). The CP is independent of the core network. It allows parallel segmentation and reassembly of packets. The CP comprises of two sub layers, namely, the common part convergence sub layer (CPCS) and the segmentation and reassembly (SAR) sub layer. The CPCS processes the packets from the higher layer and adds padding and additional information, so as to be segmented in the SAR. For further information on the CP. The SSCS consists of functions that are specific to the core network. For example, the Ethernet SSCS has been standardized for Ethernet core networks. The SSCS adapts the different data formats to the HIPERLAN/2 DLC format. It is also responsible for mapping the QoS requests of the higher layers to the QoS parameters of HIPERLAN/2 such as data rate, delay, and jitter.

The DLC Layer

The DLC layer constitutes the logical link between the AP and the MTs. This ensures a connection-oriented communication in a HIPERLAN/2 network, in contrast to the connectionless service offered by the IEEE standards.

The DLC layer is organized into three functional units, namely, the radio link control (RLC) sub layer on the control plane, the error control (EC) sub layer on the user plane, and the MAC sub layer. The following discussion describes the features of the DLC layer.

The RLC Sub layer

The RLC sub layer takes care of most of the control procedures on the DLC layer. The tapes of the RLC

can be summarized as follows.

• **Association control function (ACF):** The ACF handles the registration and the authentication functions of an MT with an AP within a radio cell. Only after the ACF procedure has been carried out can the MT ever communicate with the AP.

• **DLC user connection control (DCC):** The DCC function is used to control DLC user connections. It can set up new connections, modify existing connections, and terminate connections.

• **Radio resource control (RRC):** The RRC is responsible for the surveillance and efficient utilization of the availablele frequency resources. It performs the following tasks:

Dynamic frequency selection: This function is not availablele in IEEE 802.11, IEEE 802.11a, IEEE802.11b, and HIPERLAN/1, and is thus unique to HIPERLAN/2. It allows the AP to select a channel (frequency) for communication with the MTs depending on the interference in each channel, thereby aiding in the efficient utilization of the available frequencies.

Handoff: HIPERLAN/2 supports three types of handoff, namely, sector handoff (moving to another sector of the same antenna of an APT), radio handoff (handoff between two APTs under the same APC), and network handoff (handoff between two APs in the same network).

Power saving: Power-saving schemes much similar to those in HIPERLAN/1 and IEEE 802.11 have been implemented.

Error Control (EC)

Selective repeat (where only the specific damaged or lost frame is retransmitted) protocol is used for controlling the errors across the medium. To support QoS for stringent and delay-critical applications, a discard mechanism can be provided by specifying a maximum delay.

The MAC Sub layer

The MAC protocol is used for access to the medium, resulting in the transmission of data through that channel. However, unlike the IEEE standards and the HIPERLAN/1 in which channel access is made by sensing it,the MAC protocol follows a dynamic time division multiple le access/time division depleting (TDMA/TDD) scheme with centralized control. The protocol supports both AP-MT uncast and multicast transfer, and at the same time MT-MT peer-to-peer communication. The centralized AP scheduling provides QoS support and collision-free transmission. The MAC protocol provides a connection-oriented communication between the AP and the MT (or between MTs).

Security Issues

Elaborate security mechanisms exist in the HIPERLAN/2 system. The encryption procedure is optional and can be selected by the MT during association. Two strong encryption algorithms are offered, namely, the data encryption standard (DES) and the triple-DES algorithms.

1.4 BLUETOOTH

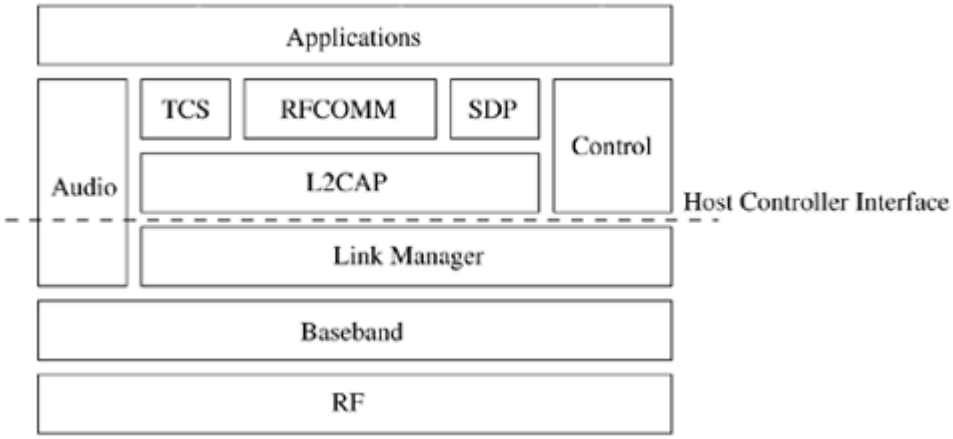
WLAN technology enables device connectivity to infrastructure-based services through a wireless carrier provider. However, the need for personal devices to communicate wirelessly with one another, without an established infrastructure, has led to the emergence of personal area networks (PANs). The first attempt to define a standard for PANs dates back to Ericsson's Bluetooth project in 1994 to enable communication between mobile phones using low- power and low-cost radio interfaces. In May 1998, several companies such as Intel, IBM, Nokia, and Toshiba joined Ericsson to form the Bluetooth Special Interest Group (SIG) whose aim was to develop a *de facto* standard for PANs. Recently, IEEE has approved a Bluetooth-based standard (IEEE 802.15.1) for wireless

personal area networks (WPANs). The standard covers only the MAC and the physical layers while the Bluetooth specification details the whole protocol stack. Bluetooth employs radio frequency (RF) technology for communication. It makes use of frequency modulation to generate radio waves in the ISM band. The project was named after Danish King Harald Blatand (A.D. 940-981) (who was known as Bluetooth due to his fondness for blueberries), who unified the Scandinavians by introducing Christianity. Low power consumption of Bluetooth technology and an offered range of up to ten meters has paved the way for several usage models. One can have an interactive conference by establishing an ad hoc network of laptops. Cordless computer, instant postcard [sending digital photographs instantly (a camera is cordlessly connected to a mobile phone)], and three-in-one phone [the same phone functions as an intercom (at the office, no telephone charge), cordless phone (at home, a fixed-line charge), and mobile phone (on the move, a cellular charge)] are other indicative usage models.

Bluetooth Specifications

The Bluetooth specification consists of two parts: core and profiles. The core provides a common data link and physical layer to application protocols, and maximizes reusability of existing higher layer protocols. The profiles specifications classify Bluetooth applications into thirteen types. The protocol stack of Bluetooth performs the functions of locating devices, connecting other devices, and exchanging data. It is logically partitioned into three layers, namely, the transport protocol group, the middleware protocol group, and the application group. The transport protocol group consists of the radio layer, baseband layer, link manager layer, logical link control and adaptation layer, and the host controller interface. The middleware protocol group comprises of RFCOMM, SDP, and IrDA (IrOBEX and IrMC). The application group consists of applications (profiles) using Bluetooth wireless links, such as the modem dialer and the Web-browsing client. The following sections discuss the concepts involved in the design of transport protocols in Bluetooth communications, and also provide an overview of the middleware and application layer protocols. Figure sows the protocol stack of Bluetooth.

Bluetooth protocol stack.



Transport Protocol Group

This group is composed of the protocols designed to allow Bluetooth devices to locate

each other and to create, configure, and manage the wireless links. Design of various protocols and techniques used in Bluetooth communications has been done with the target of low power consumption and ease of operation. This shall become evident in the design choice of FHSS and the master–slave architecture. The following sections study the various protocols in this group, their purpose, their modes of operation, and other specifications.

Radio (Physical) Layer

The radio part of the specification deals with the characteristics of the transceivers and design specifications such as frequency accuracy, channel interference, and modulation characteristics. The Bluetooth system operates in the globally available ISM frequency band and the frequency modulation is GFSK. It supports 64 Kbps voice channels and asynchronous data channels with a peak rate of 1 Mbps. The data channels are either asymmetric (in one direction) or symmetric (in both directions). The Bluetooth transceiver is a FHSS system operating over a set of *m* channels each of width 1 MHz. In most of the countries, the value of *m* is 79. Frequency hopping is used and hops are made at a rapid rate across the possible 79 hops in the band, starting at 2.4GHz and stopping at 2.480 GHz. The choice of frequency hopping has been made to provide protection against interference. The Bluetooth air interface is based on a nominal antenna power of 0 dBm (1 mW) with extensions for operating at up to 20 dBm (100 mW) worldwide. The nominal link range is from 10 centimeters to 10 meters, but can be extended to more than 100 meters by increasing the transmit power (using the 20 dBm option). It should be noted here that a WLAN cannot use an antenna power of less than 0 dBm (1 mW) and hence an 802.11 solution might not be apt for power-constrained devices.

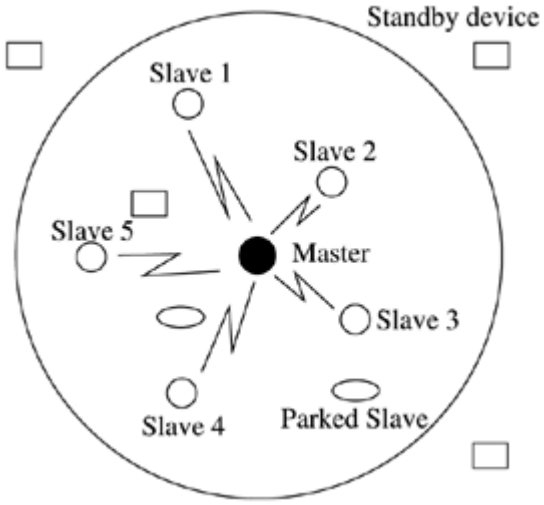
Baseband Layer:

The key functions of this layer are frequency hop selection, connection creation, and medium access control. Bluetooth communication takes place by ad hoc creation of a network called a *piconet.* The address and the clock associated with each Bluetooth device are the two fundamental elements governing the formation of a piconet. Every device is assigned a single 48-bit address which is similar to the addresses of IEEE 802.xx LANdevices. The address field is partitioned into three parts and the lower address part (LAP) is used in several baseband operations such as pionet identification, error checking, and security checks. The remaining two parts are proprietary addresses of the manufacturing organization. LAP is assigned internally by each organization. Every device also has a 28-bit clock (called the *native clock*) that ticks 3,200 times per second or once every 312.5 μs. It should be noted that this is twice the normal hopping rate of 1,600 hops per second.

Piconet

The initiator for the formation of the network assumes the role of the *master* (of the piconet). All the other members are termed as *slaves* of the piconet. A piconet can have up to seven active slaves at any instant. For the purpose of identification, each active slave of the piconet is assigned a locally unique active member address AM_ADDR. Other devices could also be part of the piconet by being in the parked mode (explained later). A Bluetooth device not associated with any piconet is said to be in standby mode. Figure shows a piconet with several devices.

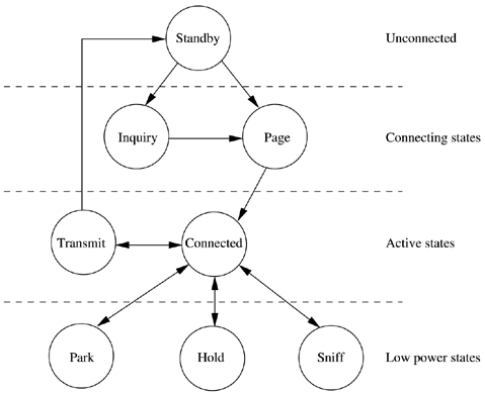
A typical piconet.



Operational States

Below Figure shows the state diagram of Bluetooth communications. Initially, all the devices would be in the standby mode. Then some device (called the master) could begin the inquiry and get to know the nearby devices and, if needed, join them into its piconet. After the inquiry, the device could formally be joined by paging, which is a packet-exchange process between the master and a prospective slave to inform the slave of the master's clock. If the device was already inquired, the master could get into the page state bypassing the inquiry state. Once the device finishes getting paged, it enters the connected state. This state has three power-conserving sub-states – hold, sniff, and park (described later in this section). A device in the connected state can participate in the data transmission.

Figure Operational states.



Frequency Hopping Sequences

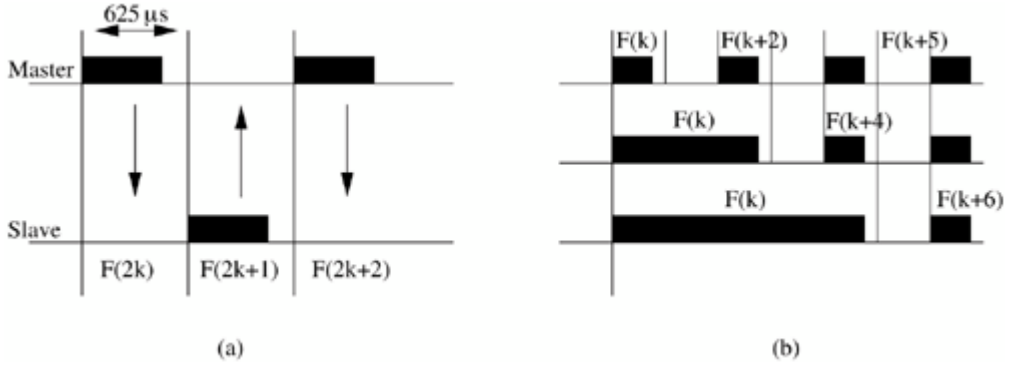
It is evident (in any wireless communication) that the sender and the receiver should

use the same frequency for communication to take place. A frequency selection module (FSM) is present in each device to select the next frequency to be used under various circumstances. In the connected state, the clock and the address of the device (master) completely determine the hopping sequence. Different combination of inputs (clock, address) are used depending on the operational state. During the inquiry operation, the address input to FSM is a common inquiry address. This common address is needed because at the time of inquiry no device has information about the hopping sequence being followed. The address of the paged device is fed as input to the FSM for the paging state.

Communication Channel

The channel is divided into time slots, each 625 μs in length. The time slots are numbered according to the Bluetooth clock of the piconet master. A time division duplex (TDD) scheme is used where master and slave alternately transmit. The master starts its transmission in even-numbered time slots only, and the slave starts its transmission in odd-numbered time slots only. This is clearly illustrated in Figure(a). The packet start shall be aligned with the slot start. A Bluetooth device would determine slot parity by looking at the least significantant bit (LSB) in the bit representation of its clock. If LSB is set to 1, it is the possible transmission slot for the slave. A slave in normal circumstances is allowed to transmit only if in the preceding slot it has received a packet from the master. A slave should know the master's clock and address to determine the next frequency (from the FSM). This information is exchanged during paging.

Figure Transmission of packets over a channel.



Packet-Based Communication

Bluetooth uses packet-based communication where the data to be transmitted is fragmented into packets. Only a single packet can be transmitted in each slot. A typical packet used in these communications has three components: access code, header, and payload. The main component of the access code is the address of the piconet master. All packets exchanged on the channel are identified by the master's identity. The packet will be accepted by the recipient only if the access code matches the access code corresponding to the piconet master. This also helps in resolving conflicts in the case where two piconets are operating currently on the same frequency. A slave receiving two packets in the same slot can identify its packet by examining the access code.

The packet header contains many fields such as a three-bit active slave address, a one-bit ACK/NACK for ARQ scheme [Automatic Repeat re Quest — anytime an error is detected, a negative acknowledgment (NACK) is returned and the specified frames are retransmitted], a four-bit packet type to distinguish payload types, and an eight-bit header error check code to detect errors in the header. Depending on the payload size, one, three, or five slots may be used for the packet transmission. The hop frequency which is used for the first slot is used for the remainder of the packet. While transmitting packets in multiple slots, itis important that the frequencies used in the following time slots are those that are assigned to those slots, and that they do not follow the frequency sequence that should have normally applied. This is illustrated in Figure 2.10 (b). When a device uses five slots for packet transmission, the next packet transmission is allowed in F(k+6) and not in F(k+2). Also note that the receiving time slot becomes F(k+5) as opposed to F(k+1). On this slotted channel, both synchronous and asynchronous links are supported. Between a master and a slave there is a single asynchronous connectionless link (ACL) supported. This is the default link that would exist once a link is established between a master and a slave. Whenever a master would like to communicate, it would, and then the slave would respond. Optionally, a piconet may also support synchronous connection oriented (SCO) links. SCO link is symmetric between master and slave with reserved bandwidth and regular periodic exchange of data in the form of reserved slots. These links are essential and useful for high-priority and time-bound information such as audio and video.

Inquiry State

As shown in Below Figure a device which is initially in the standby state enters the inquiry state. As its name suggests, the sole purpose of this state is to collect information about other Bluetooth devices in its vicinity. This information includes the Bluetooth address and the clock value, as these form the crux of the communication between the devices. This state is classified into three sub- states: inquiry, inquiry scan, and inquiry response. A potential master sends an inquiry packet in the inquiry state on the inquiry hop sequence of frequencies. This sequence is determined by feeding a common address as one of the inputs to the FSM. A device (slave) that wants to be discovered will periodically enter the inquiry scan state and listen for these inquiry packets. When an inquiry message is received in the inquiry scan state, a response packet called the frequency hopping sequence (FHS) containing the responding device address must be sent. Devices respond after a random jitter to reduce the chances of collisions.

Page State

A device enters this state to invite other devices to join its piconet. A device could invite only the devices known to itself. So normally the inquiry operation would precede this state. This state also is classified into three sub-states: page, page scan, and page response.

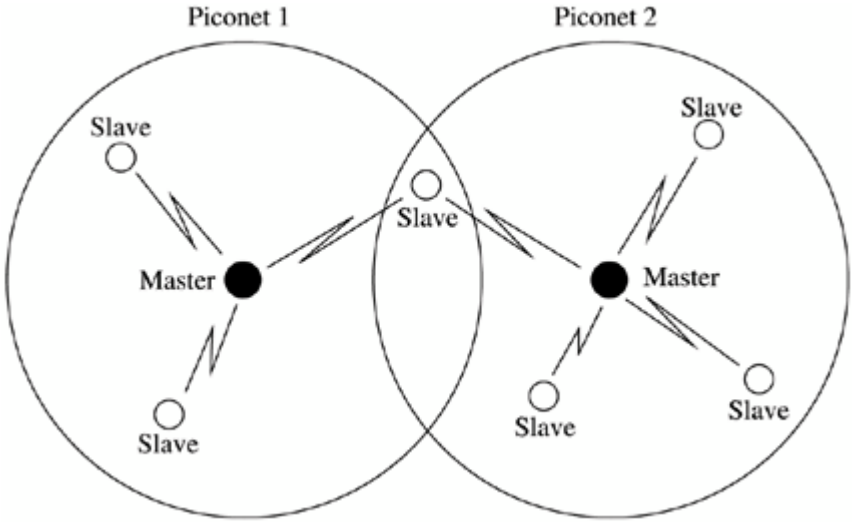
In the page mode, the master estimates the slave's clock based on the information received during the inquiry state, to determine where in the hop sequence the slave might be listening in the page scan mode. In order to account for inaccuracies in estimation, the master also transmits the page message through frequencies immediately preceding and succeeding the estimated one. On receiving the page message, the slave enters the slave

page response sub- state. It sends back a page response consisting of its ID packet which contains its device access code (DAC). Finally, the master (after receiving the response from a slave) enters the page response state and informs the slave about its clock and address so that the slave can go ahead and participate in the piconet. The slave now calculates an offset to synchronize with the master clock, and uses that to determine the hopping sequence for communication in the piconet.

Scatter nets and Issues

Piconets may overlap both spatially and temporally, that is, many piconets could operate in the same area at the same time. Each piconet is characterized by a unique master and hence the piconets hop independently, each with its own channel hopping sequence as determined by the respective master. In addition, the packets carried on the channels are preceded by different channel access codes as determined by the addresses of the master devices. As more piconets are added, the probability of collisions increases, and a degradation in performance results, as is common in FHSS systems. In this scenario, a device can participate in two or more overlaying piconets by the process of time sharing. To participate on the proper channel, it should use the associated master device address and proper clock offset. A Bluetooth unit can act as a slave in several piconets, but as a master in only a single piconet. A group of piconets in which connections exist between different piconets is called a*scatternet* (*Figure*).

Figure A typical scatter net.



When a device changes its role and takes part in different piconets, it is bound to lead to a situation in which some slots remain unused (for synchronization). This implies that complete utilization of the available bandwidth is not achieved. An interesting proposition at this juncture would be to unite the timings of the whole of the scatternet . But this may lead to an increase in the probability of packets colliding. Another important issue is the timing that a device would be missing by participating in more than one piconet. A master that is missing from a piconet (by momentarily becoming a slave in another piconet) may miss

polling slaves and must ensure that it does not miss beacons from its slaves. Similarly, a slave (by becoming a master or slave in another piconet) that is missing from a piconet could appear to its master to have gone out of range or to be connected through a poor-quality link.

Link Manager Protocol

Link manager protocol (LMP) is responsible for setting and maintaining the properties of the Bluetooth link. Currently, the major functionality of this layer is power management and security management. It also provides minimal QoS support by allowing control over parameters such as delay and delay jitter. Normally, a paging device is the default master of the piconet, but, depending on the usage scenario, the roles of the master and a slave could be switched and this is coordinated by exchange of LMP packets.

Power Management

The Bluetooth units can be in several modes of operation during the connection state, namely, active mode, sniff mode, hold mode, and park mode. These
modes are now described.

• **Active mode:** In this mode, the Bluetooth unit actively participates in the piconet. Various optimizations are provided to save power. For instance, if the master informs the slave when it will be addressed, the slave may sleep until then. The active slaves are polled by the master for transmissions.

• **Sniff mode:** This is a low-power mode in which the listening activity of the slave is reduced. TheLMP in the master issues a command to the slave to enter the sniff mode, giving it a sniff interval, and the slave listens for transmissions only at these fixed intervals.

• **Hold mode:** In this mode, the slave temporarily does not support ACL packets on the channel (possible SCO links will still be supported). In this mode, capacity is made available for performing other functions such as scanning, paging, inquiring, or attending another piconet.

• **Park mode:** This is a very low-power mode. The slave gives up its active member address and is given an eight-bit parked member address. The slave, however, stays synchronized to the channel. Any messages to be sent to a parked member are sent over the broadcast channel characterized by an active member address of all zeros. Apart from saving power, the park mode helps the master to have more than seven slaves (limited by the three-bit active member address space) in the piconet.

Bluetooth Security

In Bluetooth communications, devices may be authenticated and links may be encrypted. The authentication of devices is carried out by means of a challenge- response mechanism which is based on a commonly shared secret link key generated through a user-provided personal identification number (PIN). The authentication starts with the transmission of an LMP challenge packet and ends with the verification of result returned by the claimant. Optionally, the link between them could also be encrypted.

Logical Link Control and Adaptation Protocol (L2CAP)

This is the protocol with which most applications would interact unless a host controller is used. L2CAP supports protocol multiplexing to give the abstraction to each of the several applications running in the higher layers as if it alone is being run. Since the data packets

defined by the baseband protocol are limited in size, L2CAP also segments large packets from higher layers such as RFCOMM orSDP into multiple smaller packets prior to their transmission over the channel. Similarly, multiple received baseband packets may be reassembled into a single larger L2CAP packet. This protocol provides QoS on certain parameters such as peak bandwidth, latency, and delay variation when the link is established between two Bluetooth units.

Host Controller Interface

This is the optional interface layer, provided between the higher (above LMP) and lower layers of the Bluetooth protocol stack, for accessing the Bluetooth hardware capabilities. Whenever the higher layers are implemented on the motherboard of a host device, this layer is needed. Such an approach could prove beneficial as the spare capacity of the host device (say, a personal computer) could be utilized. The specification defines details such as the different packet types as seen by this layer. Command packets that are used by the host to control the device, event packets that are used by the device to inform the host of the changes, and data packets come under this category.

Middleware Protocol Group

The basic functionality of the middleware protocol group is to present to the application layers a standard interface that may be used for communicating across the transport layer, that is, the applications need not know the transport layer's complexities, they can just use the application programming interfaces (APIs) or higher level functions provided by the middleware protocols. This group consists of the RFCOMM layer, service discovery protocol (SDP), IrDA interoperability protocols, telephony control specification (TCS), and audio. The RFCOMM layer presents a virtual serial port to applications using the serial interface. Any application which is using the serial port can work seamlessly on Bluetooth devices. RFCOMM uses an L2CAP connection to establish a link between two devices. In the case of Bluetooth devices, there is no device which will be static and hence services offered by the other devices have to be discovered. This is achieved by using the service discovery protocol (SDP) of the Bluetooth protocol stack. Service discovery makes the device self-configured without manual intervention. The IrDA interoperability protocol is not for communication between Bluetooth devices and Infrared devices. It is only for the existing IrDA applications to work on Bluetooth devices without any changes. The main protocols in the IrDA set are Ir OBEX (IrDA object exchange) for exchanging objects between two devices and IrMC (infrared mobile communications) for synchronization. Audio is the distinguishing part of Bluetooth. Audio is given the highest priority and is directly carried over the baseband at 64 Kbps so that a very good quality of voice is provided. Another important point to note here is that audio is actually not a layer of the protocol stack, but only a specific packet format that can be transmitted directly over the SCO links of the baseband layer. Telephony control is implemented using the telephony control specification – binary (TCS-BIN) protocol. TCS defines three major functional areas: call control, group management, and connectionless TCS. Call control is used to set up calls which can be subsequently used to carry voice and data traffic. TCS operates in both point-to-point and point-to-multipoint configurations. One of the main concepts of TCS is that of the wireless user group (WUG). Group management enables multiple telephone extensions, call forwarding, and group calls. For example, consider multiple handsets and a single base set.

When a call comes in to the base set, all the multiple handsets can receive this call. In a similar fashion, calls can also be forwarded. The functionalities of TCS include *configuration distribution* and *fast intermember access*.Configuration distribution is the mechanism used to find the information about the other members in a group. Fast inter member access is a method for two slaves to create a new piconet. A WUG member uses the information from the configuration distribution and determines another member which it wants to contact. Then it sends the device's information to the master, which forwards it to this device. The contacted device then responds with its device address and clock information and places itself in a page scan state. Then the master contacts the device initiating the communication. This device now pages the contacted device and forms a new piconet. This explains how a new piconet is formed between two slaves with the help of the master. In all the above cases, a connection-oriented channel is established. To exchange simple information such as adjusting volume or signaling information, establishing such a channel is overkill and hence connectionless TCS has been provided for having a connectionless channel.

Bluetooth Profiles

These profiles have been developed to promote interoperability among the many implementations of the Bluetooth protocol stack. Each Bluetooth profile specification has been defined to provide a clear and transparent standard that can be used to implement a specific user end function. Two Bluetooth devices can achieve a common functionality only if both devices support identical profiles. For example, a cellular phone and a headset both have to support the Bluetooth headset profile for the headset to work with the phone. The Bluetooth profiles spring up from the usage models. In all, 13 profiles have been listed and these can be broadly classified into the following four categories:

1. **Generic profiles:** The Generic access profile, which is not really an application, provides a way to establish and maintain secure links between the master and the slaves. The service discovery profile enables users to access SDP to find out which applications (Bluetooth services) are supported by a specific device.
2. **Telephony profiles:** The cordless telephony profile is designed for three- in-one phones. The Intercom profile supports two-way voice communication between two Bluetooth devices within range of each other. The Headset profile specifies how Bluetooth can provide a wireless connection to a headset (with earphones/microphones) for use with a computer or a mobile phone.
3. **Networking profiles:** The LAN Access profile enables Bluetooth devices to either connect to a LAN through APs or form a small wireless LAN among themselves. The dial-up networking profile is designed to provide dial-up connections via Bluetooth-enabled mobile phones. The FAX profile, very similar to the dial-up networking profile, enables computers to send and receive faxes via a Bluetooth-enabled mobile phone.
4. **Serial and object exchange profiles:** The serial port profile emulates a serial line (RS232 and USB serial ports) for (legacy) applications that require a serial line. The other profiles, generic object exchange, object push, file transfer, and synchronization, are for exchanging objects between two wireless devices.

Bluetooth is the first wireless technology which has actually tried to attempt to make all the

household consumer electronics devices follow one particular communication paradigm. It has been partially successful, but it does have its limitations. Bluetooth communication currently does not provide support for routing. It should be noted that some research efforts are under way to accommodate this in the Bluetooth specification. Once the routing provision is given, inter-piconet communication could be enhanced. The issues of handoffs also have not yet been dealt with till now. Although master–slave architecture has aided low cost, the master becomes the bottleneck for the whole piconet in terms of performance, fault tolerance, and bandwidth utilization. Most importantly, Bluetooth communication takes place in the same frequency band as that of WLAN and hence robust coexistence solutions need to be developed to avoid interference. The technology is still under development. Currently, there are nearly 1,800 adopter companies which are contributing toward the development of the technology.

1.6 HOMERF

Wireless home networking represents the use of the radio frequency (RF) spectrum to transmit voice and data in confined areas such as homes and small offices. One of the visionary concepts that home networking intends to achieve is the establishment of communication between home appliances such as computers, TVs, telephones, refrigerators, and air conditioners. Wireless home networks have an edge over their wired counterparts because features such as flexibility (enabling of file and drive sharing) and interoperability that exist in the wired networks are coupled with those in the wireless domain, namely, simplicity of installation and mobility.

The HIPERLAN/2, as mentioned earlier, has provisions for direct communication between the mobile terminals (the home environment). The home environment enables election of a central controller (CC) which coordinates the communication process. This environment is helpful in setting up home networks. Apart from this, an industry consortium known as the Home RF Working Group has developed a technology that is termed Home RF. This technology intends to integrate devices used in homes into a single network and utilize RF links for communication. Home RF is a strong competitor to Bluetooth as it operates in the ISM band.

Technical Features

The Home RF provides data rates of 1.6 Mbps, a little higher than the Bluetooth rate, supporting both infrastructure-based and ad hoc communications. It provides a guaranteed QoS delivery to voice-only devices and best-effort delivery for data-only devices. The devices need to be plug-and-play enabled; this needs automatic device discovery and identification in the network. A typical Home RF network consists of resource providers (through which communication to various resources such as the cable modem and phone lines is effected), and the devices connected to them (such as the cordless phone, printers, and file servers). The Home RF technology follows a protocol called the shared wireless access protocol (SWAP). The protocol is used to set up a network that provides access to a public network telephone, the Internet (data), entertainment networks (cable televis ion, digital audio, and video), transfer and sharing of data resources (such as disks and printers), and home control and automation. The SWAP has been derived from the IEEE 802.11 and the European digitally enhanced cordless telephony (DECT) standards. It employs a hybrid TDMA/CSMA scheme for

channel access. While TDMA handles isochronous transmission (similar to synchronous transmission, isochronous transmission is also used for multimedia communication where both the schemes have stringent timing constraints, but isochronous transmission is not as rigid as synchronous transmission in which data streams are delivered only at specific intervals), CSMA supports asynchronous transmission (in a manner similar to that of the IEEE 802.11 standard), thereby making the actual framing structure more complex. The SWAP, however, differs from the IEEE 802.11 specification by not having the RTS-CTS handshake since it is more economical to do away with the expensive handshake; moreover, the hidden terminal problem does not pose a serious threat in the case of small-scale networks such as the home networks. The SWAP can support up to 127 devices, each identified uniquely by a 48-bit network identifier. The supported devices can fall into one (or more) of the following four basic types:

- Connection point that provides a gateway to the public switched telephone network (PSTN), hence supporting voice and data services.
- Asynchronous data node that uses the CSMA/CA mechanism to communicate with other nodes.
- Voice node that uses TDMA for communication.
- Voice and data node that can use both CSMA/CA and TDMA for channel access.

Home networking also needs strong security measures to safeguard against potential eavesdroppers. That is the reason why SWAP uses strong algorithms such as Blowfish encryption. Home RF also includes support for optional packet compression which provides a trade-off between bandwidth and power consumption.

Because of its complex (hybrid) MAC and higher capability physical layer, the cost of Home RF devices is higher than that of Bluetooth devices. Home RF Version 2.0, released recently, offers higher data rates (up to 10 Mbps by using wider channels in the ISM band through FHSS).

1.2 WIRELESS INTERNET

1.2.1 INTRODUCTION

The Internet has affected the traditional way of information exchange and now almost every city, every town, and every street has access to the Internet. Some basic concepts about the Internet and some fundamental issues that are encountered when a transition is made from the wired domain to the wireless domain and the Mobile-IP framework are discussed in first Topic. Some of the problems faced during a transition from the wired domain to the wireless domain arise due to the fact that the protocols that work very well in the former may perform poorly in the latter. The key issues involved in TCP for wireless networks and an analysis of the current set of proposals to enhance the performance of TCP in the wireless domain are presented in this Chapter.

The classical wired networks have given rise to a number of application protocols such as TELNET, FTP, and SMTP. The wireless application protocol (WAP) architecture aims at bridging the gap at the application level, between the wireless users and the services offered to them.

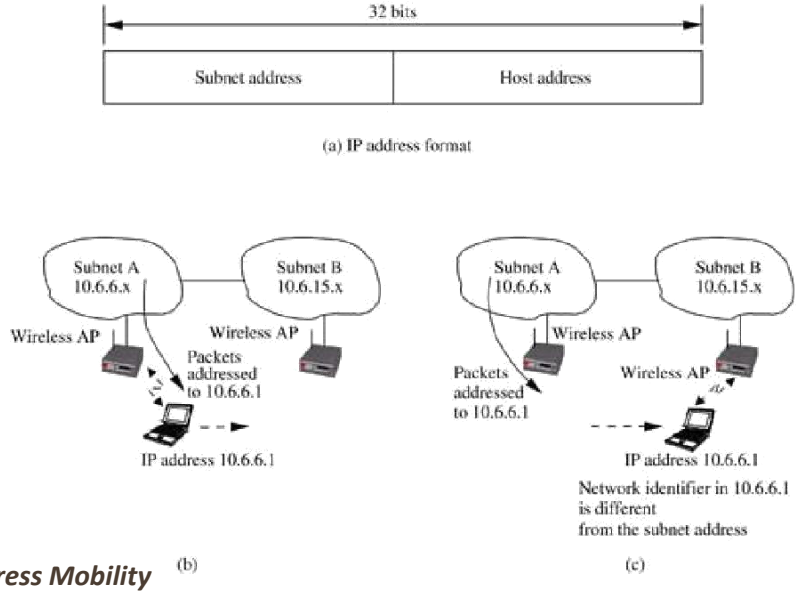
1.2.2 WHAT IS WIRELESS INTERNET?

Wireless Internet refers to the extension of the services offered by the Internet to mobile users, enabling them to access information and data irrespective of their location. The inherent problems

associated with wireless domain, mobility of nodes, and the design of existing protocols used in the Internet; require several solutions for making the wireless Internet a reality. An illustration of wireless Internet with its layered protocol stack at wired and wireless parts is shown in Fig 1.The major issues that are to be considered for wireless Internet are as follows.

a. Address mobility
b. Inefficiency of transport layer protocols
c. Inefficiency of application layer protocols

Fig1 An illustration of wireless Internet.



1.2.2.1 Address Mobility

The network layer protocol used in the Internet is Internet protocol (IP) which was designed for wired networks with fixed nodes. IP employs a hierarchical addressing with a globally unique 32-bit address[1] which has two parts, network identifier and host identifier, as shown in Fig 2 (a). The network identifier refers to the subnet address to which the host is connected. This addressing scheme was used to reduce the routing table size in the core routers of the Internet, which uses only the network part of the IP address for making routing decisions. This addressing scheme may not work directly in the wireless extension of the Internet, as the mobile hosts may move from one subnet to another, but the packets addressed to the mobile host may be delivered to the old subnet to which the node was originally attached, as illustrated in Fig 2 (b) and 2 (c). Hence the traditional IP addressing is not supportive of address mobility which is essential in wireless Internet. Fig 2 shows the mobility of a node (with IP address 10.6.6.1) attached to subnet A (subnet address 10.6.6.x) moving over to another subnet B with address 10.6.15.x. In this case, the packets addressed to the node will be routed to the subnet A instead of the subnet B, as the network part in the mobile node's address is 10.6.6.x. MobileIP[2] is a solution that uses an address redirection mechanism for this address mobility issue in wireless Internet

1 The recently introduced IP Version 6 has a 128-bit address.
Throughout this chapter, Mobile IP refers to the mobility aspect of IP address
2 and MobileIP
refers to one particular solution for Mobile IP.

Fig 2. The address mobility problem.

1 .2.2 Inefficiency of Transport Layer Protocols

The transport layer is very important in the Internet as it ensures setting up and maintaining end-to-end connections, reliable end-to-end delivery of data packets, flow control, and congestion control.TCP is the predominant transport layer protocol for wired networks, even though UDP, a connectionless unreliable transport layer protocol, is used by certain applications. Wireless Internet requires efficient operation of the transport layer protocols as the wireless medium is inherently unreliable due to its time-varying and environment-dependent characteristics. Traditional TCP invokes a congestion control algorithm in order to handle congestion in the networks. If a data packet or an ACK packet is lost, then TCP assumes that the loss is due to congestion and reduces the size of the congestion window by half. With every successive packet loss the congestion window is reduced, and hence TCP provides a
degraded performance in wireless links. Even in situations where the packet loss
is caused by link error or collision, the TCP invokes the congestion control
algorithm leading to very low throughput. The identification of the real cause that led to the packet loss is important in improving the performance of
the TCP over wireless links. Some of the solutions for the transport layer issues include indirect-TCP (ITCP), snoop TCP, and mobile TCP.

1.2.2.3 Inefficiency of Application Layer Protocols

Traditional application layer protocols used in the Internet such
as HTTP,[3] TELNET, simple mail transfer protocol (SMTP), and several markup languages such as HTML were designed and optimized for wired networks. Many of these protocols are not very efficient when used with wireless links. The major issues that prevent HTTP from being used in wireless Internet are its stateless operation, high overhead due to character encoding, redundant information carried in the HTTP requests, and opening of a new TCP Connection with every transaction. Wireless bandwidth is limited and much more expensive compared to wired networks. Also, the capabilities of the handheld devices are limited, making it difficult to handle computationally and bandwidth-wise expensive application protocols. Wireless application protocol (WAP) and optimizations over traditional HTTP are some of the solutions for the application layer issues.

1.2.3 MOBILE IP

Each computer connected to the Internet has a unique IP address, which helps not only in identifying the computer on the network but also routing the data to the computer. The problem of locating a mobile host in a mobile domain is now imminent as the IP address assigned can no longer be restricted to a region.

The first conceivable solution to the above problem would be to change the IP address when the host moves from one subnet to another. In this way, its address is consistent with the subnet it is currently in. The problems with changing the IP address as the host moves is that TCP identifies its connection with another terminal based on the IP address. Therefore, if the IP address itself changes, the TCP connection must be reestablished. Another method would be to continue to use the same IP address and add special routing entries for tracking the current location of the user. This solution is practical if the number of mobile users is small. The quick-fix solutions are inadequate, but they give valuable insight into the nature of the mobility problem and offer certain guidelines for the actual solution. Before providing the solution to the problem, some issues of utmost importance need to be enumerated. These are as follows:

- **Compatibility:** The existing wired Internet infrastructure is well-established today and it is economically impractical to try to alter the way it is working.
- **Scalability:** Wireless communication is the technology for the future, so the solution should be scalable to support a large number of users.

• **Transparency:** The mobility provided should be transparent in the sense that the user should not feel a difference when working in a wireless domain or in a wired one. In Fig 3, mobile node (MN) is a mobile terminal system (end user) or a mobile router. It is the host for which the mobility support is to be provided. At the other end of the network is the system with which MN communicates. This is referred to as the correspondent node (CN), which may be a fixed or a mobile node. In this section, CN is considered to be a fixed, wired node. The node or router to which the MN is connected, which currently enjoys all the network facilities, is known as the foreign agent (FA). The subnet to which the MN's IP address belongs is the home network, and the router or node under whose domain this IP address lies is the home agent (HA).

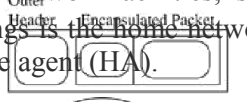
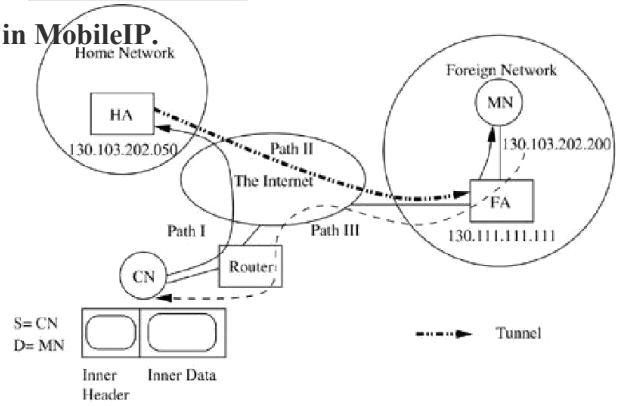


Figure 2.3. Routing in MobileIP.



Suppose MN is currently in the subnet 130.111.*, hence as shown in the figure, 130.111.111.111 becomes the FA for MN. If CN sends a packet to MN, it reaches the HA of MN (130.103.202.050) along Path I.HA cannot find MN in the home network, but if it knows the location of MN, it can send the packet along Path II by creating a *tunnel,* as explained later.

1.2.3. 1 MobileIP

The essence of the MobileIP scheme is the use of the old IP address but with a few additional mechanisms, to provide mobility support. MN is assigned another address, the care of address (COA). The COA can be one of the following types:

1. **Foreign agent-based COA:** The address of the FA to which the MN is connected can be used to locate the MN. The COA of the MN in this case is the address of its current FA.
2. **Colocated COA:** In this case MN acquires a topologically correct IP address. In effect, each MNnow has two IP addresses assigned to it. In this case the CN sends data to the old IP address. TheHA receives this packet and *tunnels* it to the MN using the new IP address.

In the case of FA-based COA, the FA encapsulates the packet and forwards it to MN, while in the case of collocated COA, it is encapsulated at MN. The HA encapsulates the data packet inside

another packet addressed to the COA of MN. This is known as *encapsulation* and the mechanism is known as *tunneling.* Path II in Fig .3 shows the tunnel using the FA-based COA. Though the problem is solved, it has been done with a high degree of inefficiency.

Registration with the HA: This section discusses how the COA of an MN is communicated to its HA. This is done through the process of *registration.* When an MN moves to a new location, it tries to find the FA. This is done using the agent advertisement packet or agent solicitation packet. Registration involves authentication and authorization of the MN by the HA. In case of the collocated COA, there is no intermediate FA. MN simply sends the registration request to its HA, which authenticates it and sends back a registration reply.

Reverse Tunneling

It appears that there should not be any problem for the MN in sending a packet to the CN following path III. However, there are other practical constraints that play an important role here.

1. **Ingress filtering:** There are some routers which filter the packets going out of the network if the source IP address associated with them is not the
   subnet's IP address. This is known as ingress filtering where the MN's packet may get filtered in the foreign network if it uses its
   home IP address directly.
2. **Firewalls:** As a security measure, most firewalls will filter and drop
   packets that originate from outside the local network, but appear to have a source address of a node that belongs to the local network. Hence
   if MN uses its home IP address and if these packets are sent to the home network, then they will be filtered.
3. **Time to live (TTL):** The MN should be able to communicate
   transparently with all the CNs that it can communicate with while at home. Hence, in case of *triangular routing*, the TTL for the packets must be reduced only by one, up to the point where the packet is tunneled home.
   Firewalls and ingress filtering have made a simple solution complicated. Therefore, to avoid these problems the idea of *reverse tunneling* is used, that is, MN encapsulates its packets using the source address of the encapsulated packet as its COA and destination as HA. The routing of packets from MNto CN takes place via the non-shortest path (as shown in Figure 1.3), that is, MN to HA to CN or vice versa is called *triangular routing*. This method, though not efficient, does work in practice.

1.3.3.2 Simultaneous Bindings

Simultaneous bindings is a feature of MobileIP that allows an MN to register more than one COA at the same time, that is, the HA allows MN to register more than one COA. MN can also deregister a specificCOA. In such a situation, the HA must send multiple duplicated encapsulated data packets, one to eachCOA. The idea behind the use of simultaneous binding is to improve the reliability of data transmission.

1.3.3.3 Route Optimization

The packets sent to and from the HA are routed on non-optimal paths, hence the need for optimizations [1]. The CN is assumed to be mobility-aware, that is, it has the capability to deencapsulate the packets from the MN and send packets to the MN, bypassing the HA. The following are some of the concepts related to optimization strategies.

• **Binding cache:** The CN can keep the mapping of MN's IP address

and COA in a cache. Such a cache is called a binding cache. Binding cache is used by the CN to find the COA of the MN in order to optimize the path length. Like any other cache, this may follow the update policies such as least recently used and first-in-first-out.

• **Binding request and binding update:** The CN can find the binding using a binding request message, to which the HA responds with a binding update message.
• **Binding warning:** In some cases, a handoff may occur, but CN may continue to use the old mapping. In such situations, the old FA sends a binding warning message to HA, which in turn informs the CN about the change, using a binding update message.

1.3.4 MobileIP Variations – The 4 × 4 Approach

As discussed in Section 1.2.3.1, MobileIP is a general-purpose solution to the mobility problem over IPv4. It uses encapsulation as a primary technique and thus introduces a huge overhead (approximately 20 bytes per packet). In the MobileIP scheme, the MN is dependent on the FA to provide a COA. The presence of the FA in all transactions prevents the MN from being able to perform any kind of optimization, and it is unable to forgo the MobileIP support even when it is not required. The key factors that affect any optimization scheme are the permissiveness of the network and the capabilities of the communicating nodes. In the following strategy presented, it is presumed that theMN does not depend on the FA for any support and it is able to acquire a COA from the subnet that it is present in.

Goals of Optimizations

Any optimization scheme should try to ensure guaranteed delivery, low latency, and low overhead. Deliverability is to be understood in terms of the traditional datagram network that provides only a best-effort service. The latency issue mainly deals with the route that is being followed by the packet from the source to the destination, either in terms of the hop count or the delay. The overhead in the MobileIP scheme is essentially the packet encapsulation overhead.

The 4 × 4 Approach

The strategy presented here provides four options for packets directed from the MN to the CN (OUT approaches) and four more options for packets directed from the CN to the MN (IN approaches). The set of options can be provided as a 4×4 [2] matrix to the hosts, which can decide on the appropriate combination depending on the situation. The IN and OUT strategies are summarized in Tables 1.1 and 1.2, respectively. s and d represent the outer source and destination in the encapsulated packet while S and D represent the inner source and destination of the packet (refer to Figure 1.3). Indirect transmission refers to the routing of packets between the CN and MN involving the HA, whereas direct transmission bypasses the HA. In Table 1.1 the four IN strategies are listed along with the respective source and destination fields, and the assumptions made and restrictions on usage of the strategies. For example, IN-IE uses the traditional Mobile IP mechanism and works in all network environments irrespective of security considerations, while IN-DT is applicable for short-term communication wherein the mobility support is compromised.

In Table 1.2, the four OUT strategies are listed. For example, OUT-IE uses the traditional MobileIP reverse tunneling mechanism and works in all network scenarios, while OUT-DH avoids encapsulation overhead but can be used only when the MN and CNare in the same IP subnet.

Table 1.1. The IN strategies in 4 × 4 approach

IN Strategy	s	d	S	D	Notes	Acceptable Combinations
Incoming Indirect Encapsulated (IN-IE)	IP address of HA	COA of the MN	IP address of the CN	Home IP address of the MN	1. Highest overhead 2. Guaranteed delivery 3. Uses tunneling 4. CN need not be mobility aware	OUT-IE OUT-DE OUT-DH
Incoming Direct Encapsulated (IN-DE)	IP address of CN	COA of the MN	IP address of the CN	Home IP address of the MN	1. CN is mobility aware 2. No tunneling	OUT-DE OUT-DH
Incoming Uses Home Address	Not applicable	Not applicable	IP address of the CN	Home IP address of the MN	1. No encapsulation 2. Usable when there	OUT-DH only

OUT Strategy	s	d	S	D	Notes	Acceptable Combinations
Outgoing Indirect Encapsulated (OUT-IE)	COA of the MN	IP address of HA	Home IP address of the MN	IP address of the CN	1.Highest overhead 2. Guaranteed delivery 3. No ingress filtering 4. CN need not be mobility aware	IN-IE only
Outgoing Direct Encapsulated (OUT-DE)	COA of the MN	IP address of CN	Home IP address of the MN	IP address of the CN	1. CN is mobility aware 2. No tunneling	IN-IE IN-DE
Outgoing Direct Home Address (OUT-DH)	Not applicable	Not applicable	Home IP address of the MN	IP address of the CN	1. No encapsulation 2. Usable when there are no security constraints at the intervening routers 3. MN and CN on the same subnet	IN-IE IN-DE IN-DH
Outgoing Direct uses Temporary Address	Not applicable	Not applicable	COA of MN	IP address of the CN	1. MN cannot receive packets addressed to its original IP address	IN-DT only

Comparison and Evaluation of the Strategies

Having seen the four approaches for each of the two directions of packet transfer, different combinations of these strategies can be considered. Though there seem to be 16 combinations, some of them are inapplicable and some are redundant. There are also restrictions on when the approaches are valid; the characteristics of the situation will determine which approach to choose. The choice of a particular strategy can be made on a per session basis or on a packet-to-packet basis, as desired by the entities involved in the conversation. Tables 1.1 and 1.2 also show the acceptable combinations of the strategies.

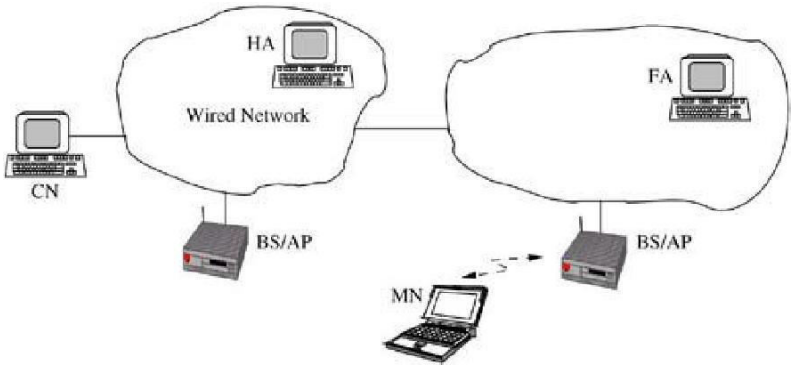
1.3.5 Handoffs

A handoff is required when the MN is moving away from the FA it is connected to, and as a result the signals transmitted to and from the current FA become weak. If the MN can receive clearer signals from another FA, it breaks its connection with the current FA and establishes a connection with the new one. The typical phases involved in handoff are measuring the signal strength,
ecisions regarding where and when to hand off, and the establishment of a new connection breaking the old one.

Classification of Handoffs

The issues in handoffs are on the same lines as those in cellular networks. Handoffs can be classified in three ways [3] based on functionalities of the entities involved, signaling procedure, and number of active connections. Function-based classification is based on the roles of the MN and FA during the handoff. Figure 4.4 shows the MN, BS, FA, and CN in the handoff scenario.

Figure 1.4. Entities in wireless Internet handoff scenario.



Here, handoffs can be classified into four categories as follows:

1. **Mobile initiated handoff:** In this case, the handoff is managed by the MN. The MN measures the signal strength, decides the target base station (BS), and triggers the handoff.
2. **Mobile evaluated handoff:** This is similar to the previous case except that the decision on the handoff lies within the network, perhaps with the BS.
3. **Network initiated handoff:** In this case, the network (BS) decides where the MN should be handed over. Also, only the network measures the

signal strength of the uplink and the MN has very little role to play.

4. **Mobile assisted handoff:** The MN assists the network in the network initiated scenario by measuring the downlink signal strength. This is typically to avoid a *black hole* scenario. A black hole scenario occurs when the channel properties tend to be asymmetric. (Usually wireless
channels are assumed to have the same properties in both uplink and downlink, but in certain circumstances the throughput on one of the directions may be significantly less than the other. This scenario is referred to as a black hole.)

The second kind of classification is based on the number of active connections, where the handoffs are classified into two types: the hard handoff (only one active connection to the new or the old FA) and the soft handoff (has two active connections during the handoff). Signaling procedure-based handoffs are classified into two types depending on which FA (old FA or new FA) triggers the handoff along with MN.

• **Forward handoff:** In this case, MN decides the target BS and then requests the target BS to contact the current BS to initiate the handoff procedure.

• **Backward handoff:** In this case, MN decides the target BS and then requests the current BS to contact the new one.

Fast Handoffs

A typical handoff takes a few seconds to break the old connection and establish the new one. This delay may be split into three components [4]: delay in detection of a need for a handoff, layer2 handoff (a data link connection that needs to be established between the new FA and MN), and layer3 handoff orregistration with HA. The first two components cannot be avoided; however, the delay due to the third can be reduced. Also, if the above operations are parallelized, the total delay will be reduced. Two techniques called pre- and post-registration handoffs are employed to perform the above operations. The difference lies in the order in which the operations are performed. In the case of the pre-registration handoff, the registration with the HA takes place before the handoff while the MN is still attached to the old FA, while in the case of the post-registration handoff, registration takes place after the MN is connected to the new FA. In this case, the MN continues to use the old FA, tunneling data via the new FA until the process of registration is completed.

1.3.6 IPv6 Advancements

The various optimizations provided over IPv4 (IP version 4) in order to avoid the inefficiencies in routing MN's data were discussed in Section 1.3.4. IPv6 (IP version 6) has a built-in support for mobility to a great extent. The features [5] are listed below:

• Route optimization is a built-in feature of IPv6.

• IPv6 has fields for specifying both new (COA) and home (IP) address. So problems that lead to reverse tunneling can be avoided.

• The problem of *ingress filtering* is also solved due to the above.

• Control packets such as those used in route optimization can be piggy-backed onto the data packets.

• *Detection of black holes:* Sometimes it might happen that the signals of one of the links (uplink or downlink) become weak while the other link has a good signal strength. Such a phenomenon is known as a *black hole* because data can go in one direction but cannot come out in the other. In such cases, a handoff may be required. IPv6 allows both MN and BS to detect the need for a handoff due to creation of black holes.

• IPv6 avoids overheads due to encapsulation because both the COA and the original IP address are included in the same packet in two different fields.

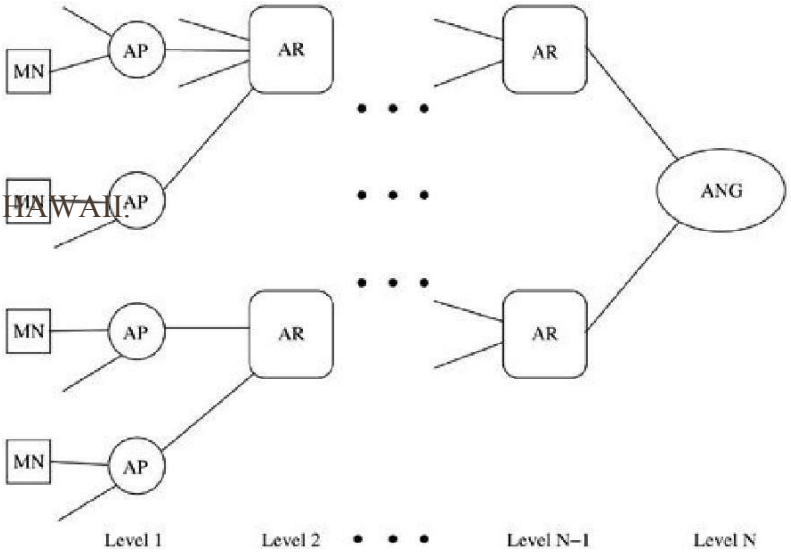
Apart from these, IPv6 allows 2^{128} addresses, thereby solving the IP address shortage problem, and includes advanced QoS features. It also supports encryption and decryption options to provide

authentication and integrity.

1.3.7 IP for Wireless Domains

MobileIP is only a solution to the mobility of IP address problem, it is not a specific solution for wireless, especially cellular domains. The following
discussion addresses certain protocols that are IP-based and suited for the wireless domain as well. In particular, we consider an approach which is terminal independent, that is, an approach aimed at giving a uniform service to both hosts that have the MobileIP capability as well as legacy hosts. The terminal independent mobility for IP (TIMIP) [6] strategy is based on two main protocols for the wireless networks, namely, HAWAII [7] and CellularIP. Figure 1.5 gives the hierarchy of routers in the HAWAII, CellularIP, and TIMIParchitectures. The access point (AP) is a router that is at the first level of the hierarchy and this is in direct communication with the MN over the wireless interface. Access routers (AR) are interior routers in the tree. The access network gateway (ANG) is the router at the root of the tree that acts as the interface between the wireless (TIMIP) domain and the core wired IP network.

Figure 1.5. Hierarchical routers.



HAWAII stands for handoff aware wireless access Internet infrastructure. The infrastructure identifies two categories of mobility to be handled, micromobility (intra-domain) and macromobility (inter-domain), where domain refers to a part of the network under the control of a single authority, such asAR and ANG. The objective of the infrastructure is to solve the QoS and efficiency issues that are not addressed by MobileIP.

CellularIP

CellularIP offers an alternative to the handoff detection problem by using the MAC layer information based on the received signal strengths to detect handoffs, instead of using the network layer information. The routing nodes maintain both a paging cache and a routing cache; the routing cache is a mapping between an MN's IP address and its current location in the CellularIP domain. The paging cache is preferred for nodes that receive or send packets relatively infrequently, and it is maintained by paging update packets sent by the MN whenever it crosses between two APs. The routing cache will be updated whenever the MN has a packet to send. The MN will send the packet to the closest AP and this will update all routing caches all the way up to the ANG. It is to be noted that during a handoff or just after the handoff, packets meant for the MN will be routed to both the old as well as the current AP in charge of the MN for a time interval equal to the routing cache timeout.

TIMIP

In the terminal independent mobility for IP (TIMIP) approach, the emphasis is on providing a uniform service to both MobileIP-capable MNs as well as the legacy terminals. The MobileIP capability of the legacy terminals will be provided by the ANG. The ANG will keep track of the information regarding each of the MNs in its domain such as the MN's MAC and IP addresses, the MobileIP capabilities, and the authentication parameters.

Whenever an MN arrives in the TIMIP domain, a routing path has to be created in the domain so that all packets intended for this host can be efficiently routed. This will cause a trigger of updates to ensure route reconfiguration in the entire hierarchy. The ARs not involved in the route will be unaware of the new path to the MN. As a result, the default strategy for any packet in the TIMIP domain, that is, for any IP address that is unknown at a particular AP or AR, will be to route it to the ANG.

• **Micromobility:** Whenever the MN moves within the same TIMIP domain, it is referred to as micromobility. The route updates and the corresponding acknowledgments will propagate up the hierarchy until the crossover AR is reached. The old path needs to be deleted in all the routing tables of the nodes. Now the crossover AR will send a route update packet addressed to the MN, and this packet will propagate down the tree until the old AP in charge of the MN is reached.
• **Macromobility:** Similar to CellularIP and HAWAII, TIMIP relies purely on MobileIP to support macromobility. The ANG acts as the MobileIP proxy on behalf of the MN that does not have MobileIP capability, and does all the MobileIP signaling that the MN would have normally done. For the normal MobileIP capable MNs, however, the ANG performs the role of a FA. The TIMIP approach also provides for seamless mobility through the context transfer framework. The context transfer essentially ensures that the data loss during handoff is minimized and this is transparent to the MN and the CN.

1.3.8 Security in MobileIP

The wireless domain is inherently insecure. Any data that needs to be transmitted has to be broadcast and anyone who can hear this can read it irrespective of the destination address.

Security Problems

The common security problems that may arise in wireless networks are as follows:

• **Registration request by a malicious node:** This is a problem because a malicious node can pose as a legitimate MN and use the MN's IP address for registration, thereby enjoying all the facilities meant for the MN.
• **Replay attacks:** Many times the above problem is solved by making the registration process encrypted. Though this appears to avoid the first problem, the malicious node may copy the MN's registration packet, which is encrypted when the MN tries to register with the FA. Though this packet cannot be decoded by this malicious node, it can certainly use this packet for registering itself as theMN at a later point of time, and hence enjoy all the facilities at the cost of the MN.
• **Tunnel hijacking:** In this case, the malicious node uses the tunnel built by the MN to break through the firewalls.
• **FA can itself be a malicious node.**

The MN and HA share the same security association and use the *message digest 5* (MD5) with 128-bit encryption. To circumvent the problem of replay attacks the MN and HA use a shared random number[4] (called Nonce) and this random number is sent along with the encrypted registration request. On registration,
the HA verifies the random number and issues a new random number to be used for the next registration. Hence, even if the packet is copied by the malicious node, it becomes useless for a replay attack, as at the time of the next registration the random number would have changed anyway.

1.3.9 MRSVP - Resource Reservation

The following section describes a reservation protocol used to provide real-time services to mobile users. A major problem is that mobility affects the QoS adversely. Hence there is a need for advance reservations to be made on behalf of a mobile host at future locations that it is likely to visit. We notice that the current RSVP[5] structure is far from adequate and examine the proposed scheme. Resource reservation protocol (RSVP) is a resource reservation setup protocol designed for multicast, multimedia data streams or flows (RFC 2205). A flow is specified by attributes such as source-destination pair, average data rate, latency, and QoS (RFC 1363).

Overview

The usual QoS parameters are delay, loss, throughput, and delay jitter. Whenever an MN moves across from one agent to another, there is obviously a change in the data flow path due to the handoff. The delay is likely to change due to the change in the data flow path and also due to the fact that the new location may vary from the old location with respect to congestion characteristics. Again, if the new location is highly congested, the available bandwidth is less, hence the throughput guarantees that were provided earlier may be violated. In addition, under extreme cases there may be temporary disconnections immediately following a handoff, which causes significant data loss during the transit.

Requirements of a Mobility-Aware RSVP

A fundamental requirement is that an MN must be able to make advance reservations along data flow paths to and from locations that it is likely to visit in the lifetime of a particular connection or session. Such a protocol has to have information that we refer to as the MSPEC, which is the set of locations from which the MN requires reservations. The definition of the MSPEC may be either statically done or there may be additional options to update it dynamically while the flow is active. A hypothetical MRSVP [9] has two types of reservations: ACTIVE and PASSIVE. An ACTIVE reservation is a normalRSVP-like reservation that is on the data flow path from the current location of the MN. A PASSIVE reservation is made along all paths to and from other locations in the MSPEC of the MN. The path along which the reservation will be made is the path specified by the MobileIP protocol. Passive reservations become active reservations whenever there is a active sender or receiver involved in that data flow path. The paths along which passive reservations have been made can be used by other flows with weaker QoS guarantees, but appropriate action needs to be taken when the passive flow turns into an active one.

MRSVP - Implementation

In this section, we describe a basic implementation of the MRSVP framework. We have to identify proxy agents (PAs) that will make reservations on behalf of mobile senders and receivers. There are two types of PAs: remote and local. A local proxy agent (LPA) is that to which the MN is currently attached. Every other agent in the MSPEC will be a remote proxy agent (RPA).

The sender periodically generates ACTIVE PATH messages, and for a mobile sender the PAs will send PASSIVE PATH messages along the flow path to the destination. Similarly, the PAs for a mobile receiver send the PASSIVE RESV messages while the receiver itself sends the ACTIVE RESV message.

The framework also defines additional messages such as JoinGroup, RecvSpec, SenderSpec, and SenderMSpec [9]. The key issues in the implementation are as follows:

- The identification of proxy agents (local and remote) that will perform the reservations on behalf of an MN.
- The identification of flow anchors (proxy agents), a Sender Anchor when the MN is a sender and a ReceiverAnchor when the MN is a receiver, that will act as fixed points in the flow path.
- The establishment of both active and passive reservations (by the remote proxy agents) for the

MNaccording to the MSPEC.

• The actual message sequences that lead to the reservation depends on the type of the flow and the strategy adopted. A detailed discussion of the protocol can be found in [9].

The MRSVP scheme is an initial approach to providing QoS guarantees within the MobileIP framework. The scheme considers both unicast as well as multicast traffic for all types of senders and receivers. The significant contribution of the approach is the notion of PASSIVE reservations that exist virtually on future routers that the MN's data flow is likely to use, but will turn into real flows when theMN moves into the new domain.

1.4 TCP IN WIRELESS DOMAIN

The topics discussed so far addressed the network layer modifications that are necessary to make an efficient transition from the wired to the wireless domain. The wireless domain is not only plagued by the mobility problem, but also by high error rates and low bandwidth. Obviously there needs to be a higher layer abstraction that would perform the error recovery and flow control. The traditional TCP, which guarantees in-order and reliable delivery, is the classical wired networks transmission protocol. Since the transition to the wireless domain should be compatible with the existing infrastructure, there is need for modifications of the existing protocols. This is the correct approach rather than resorting to a completely new set of protocols.

1.4.1 Traditional TCP

TCP provides a connection-oriented, reliable, and byte stream service. The term connection-oriented means the two applications using TCP must establish a TCP connection with each other before they can exchange data. It is a full duplex protocol, meaning that each TCP connection supports a pair of byte streams, one flowing in each direction. TCP includes a flow-control mechanism for each of these byte streams that allows the receiver to limit how much data the sender can transmit. TCP also implements a congestion-control mechanism.

TCP divides the data stream to be sent into smaller segments and assigns sequence numbers to them. The sequence number helps the receiver to provide the higher layers with in-order packet delivery, and also detect losses.

The sliding window mechanism employed by TCP guarantees the reliable delivery of data, ensures that the data is delivered in order, and enforces flow control between the sender and the receiver. In the sliding-window process, the sender sends several packets before awaiting acknowledgment of any of them, and the receiver acknowledges several packets at a time by sending to the transmitter the relative byte position of the last byte of the message that it has received successfully. The number of packets to be sent before the wait for acknowledgment (window size) is set dynamically, that is, it can change from time to time depending on network conditions.

Because the major cause of packet loss in the wired domain is congestion, TCP assumes that any loss is due to congestion.

The TCP congestion control mechanism works as below. Initially, the TCP sender sets the congestion window to the size of one maximum TCP segment [also known as maximum segment size (MSS)]. The congestion window gets doubled for each successful transmission of the current window. This process continues until the size of the congestion window exceeds the size of the receiver window or the TCP sender notices a timeout for any TCP segment. The TCP sender interprets the timeout event as network congestion, initializes a parameter called *slow start threshold* to half the current congestion window size, and resets the congestion window size to one MSS. It then continues to double the congestion window on every

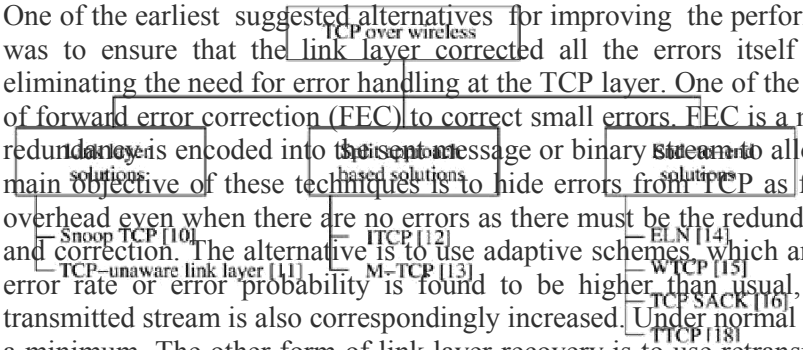
successful transmission and repeats the process until the congestion window size reaches the slow start window threshold. Once the threshold is reached, theTCP sender increases the congestion window size by one MSS for each successful transmission of the window. This mechanism whereby the congestion window size is brought down to one MSS each time network congestion is detected and then is incremented as described above is referred to as *slow start.*

Another important characteristic of TCP is fast retransmit and recovery. If the receiver receives packets out of order, it continues to send the acknowledgment for the last packet received in sequence. This indicates to the sender that some intermediate packet was lost and the sender need not invoke the congestion control mechanism. The sender then reduces the window size by half and retransmits the missing packet. This avoids the slow start phase.

1.4.2 TCP Over Wireless

The adaptation of TCP to congestion causes a lot of problems in the wireless domain. The wireless domain has high packet loss and variable latency, which may cause TCP to respond with slow start. Bandwidth utilization is further reduced due to retransmission of lost packets.

One of the earliest suggested alternatives for improving the performance of TCP over wireless networks was to ensure that the link layer corrected all the errors itself over the wireless interface, thereby eliminating the need for error handling at the TCP layer. One of the suggestions in this category is the use of forward error correction (FEC) to correct small errors. FEC is a means of error control coding wherein redundancy is encoded into the sent message or binary stream to allow self-correction at the receiver. The main objective of these techniques is to hide errors from TCP as far as possible. However, FEC incurs overhead even when there are no errors as there must be the redundant parity bits to allow error detection and correction. The alternative is to use adaptive schemes, which are dynamic in the sense that when the error rate or error probability is found to be higher than usual, the redundancy introduced into the transmitted stream is also correspondingly increased. Under normal circumstances, the overhead is kept to a minimum. The other form of link layer recovery is to use retransmissions at the link layer. This incurs the overhead only on error. However, the link level recovery mechanism may cause head-of-the-line blocking, wherein the recovery mechanisms employed for one data stream consume the network resources and prevent others from being able to transmit packets. Some researchers have advocated the use of the retransmit when FEC capability is exceeded.



The most accepted role of the link layer strategy would be one in which the link layer helps TCP error recovery by providing "almost in order delivery" of packets. Not all connections can benefit from link level retransmission as it is dependent on the nature of the applications. Several alternatives have been proposed to alter the existing TCP protocol to suit the wireless domain. The simplest idea would be to design a new TCP protocol for the wireless domain, but this will be incompatible with the wired domain. The following sections discuss various approaches to Improve TCP performance in the wireless domain. Figure 1.6 provides a classification of the existing approaches.

Figure 1.6. Classification of approaches for TCP over wireless.

1.4.3 Snoop TCP

The central idea used in snoop TCP [10] is to buffer the data as close to MN as possible in order to minimize the time for retransmission. The BS just snoops the packets being transmitted in both directions and recognizes the acknowledgments. The BS buffers the packets transmitted but does not acknowledge on behalf of MN. It simply removes the packet from the buffer when it sees an acknowledgment. If BS gets a duplicate acknowledgment (DUPACK) or no acknowledgment for quite some time, then it retransmits from the buffer after discarding the duplicate acknowledgment. This is to avoid unnecessary retransmissions from CN. The BS does not send acknowledgments to the CN on behalf of the MN, in order to retain the end-to-end semantics that traditional TCP provides. When the data transmission is from MN to CN, if
the BS detects a gap in the sequence numbers acknowledged by the CN, it sends a NACK or negative acknowledgment to the MN to indicate loss over the wireless link.

1.4.4 TCP-Unaware Link Layer

This strategy particularly aims at simulating the behavior of the snoop- TCP protocol without requiring the link layer at the BS to be TCP-aware (hence the name TCP-unaware link layer even though TCP requires some information from the link layer). The usage of delayed DUPACKs [11] imitates snoop-TCP without requiring the link layer at BS to be TCP-aware. At the BS, as in snoop-TCP, link layer retransmission is used to perform local error recovery. But unlike snoop-TCP, where retransmissions are triggered by TCP DUPACKs, here retransmissions are triggered by link level ACKs. The MN reduces the interaction between the link layer and TCP using delayed DUPACKs. The advantages of this scheme are that the link layer need not be TCP-aware, it can be used even if headers are encrypted, which is not possible in snoop-TCP, which needs to look into the headers to see the sequence numbers, and it works well for small round trip times (RTTs) over the wireless link. The most significant disadvantage of this mechanism is that the optimum value
of DUPACK delay is dependent on the wireless link, and this value is crucial in determining the performance.

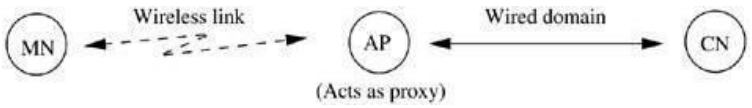
1.4.5 Indirect TCP:

This approach involves splitting of the TCP connection into two distinct connections, one TCP connection between the MN and BS and another TCP connection between the BS and the CN.

In this context of TCP over wireless, the terms BS and AP are used interchangeably. Such a division splits the TCP connection based on the domain, the wireless domain, and the wired domain. The traditional TCP can be used in the wired part of the connection and some optimized version of TCP can be used in the wireless counterpart. In this case, the intermediate agent commonly known as the access point (AP) acts as a proxy for MN.

The indirect TCP (ITCP) mechanism is shown in Figure 1.7. Loss of packets in the wireless domain, which would otherwise cause a retransmission in the wired domain, is now avoided by using a customized transport protocol between the AP and MN which accounts for the vagaries of the wireless medium. The AP acknowledges CN for the data sent to MN and buffers this data until it is successfully transmitted to MN. MN acknowledges the AP alone for the data received. Handoff may take a longer time as all the data acknowledged by AP and not transmitted to MN must be buffered at the
new AP.

Figure 1.7. Indirect TCP.



1.4.6 Mobile TCP

The most common problem associated with the wireless domain is that quite often the connection between MN and BS is lost for small intervals of time. This typically happens when MN moves behind a huge building or MN enters offices where the signals are filtered. In such cases, the sender will keep transmitting and times out eventually. In case of ITCP, the data buffered at AP may grow too large in size. It may also lead to slow start.

In such situations the sender needs to be informed. This situation is handled in mobile TCP (M-TCP) [13] by the supervisory host (the node in the wired network that controls a number of APs) which advertises the window size to be one, thus choking the sender and hence avoiding slow start. Connection may be resumed when MN can be contacted again. When the supervisory host receives aTCP packet, it forwards it to the M-TCP client. Upon reception of an ACK from M-TCP client, the supervisory host forwards the ACK to the TCP sender. Hence M-TCP maintains the end-to-end TCPsemantics even though the TCP connection is split at the supervisory host. When the M-TCP client undergoes a temporary link break, the supervisory host avoids forwarding the ACK of the last byte to the sender and hence the sender TCP goes to the persist state by setting the window size to zero. This avoids retransmission, closing of the congestion window, and slow start at the sender. For more details on mobile TCP, the reader can refer to [13].

1.4.7 Explicit Loss Notification

Typically, the problem with TCP lies in the fact that it does not know the exact cause for packet loss, and hence has to invariably assume congestion loss. An ideal TCP simply retransmits the lost packets without any congestion control mechanism. The MAC layer, however, can identify the reason for the packet loss. Once the MAC layer detects that either a handoff is about to occur or realizes that the actual cause of the packet loss is not congestion, then it immediately informs the TCP layer of the possibility of a non-congestion loss. The crux of the strategy is to detect loss at MN and send an explicit loss notification (ELN) to the sender. The sender does not reduce window size on receiving the ELN as this message implies that there was an error and not congestion. This technique avoids slow start and can handle encrypted data. However, the protocol layer software at the MAC layer of MN needs to be changed. Further, the information conveyed by the MAC layer may not always be reliable. For more details on ELN, the reader can refer to [14].

1.4.8 WTCP

WTCP [15] aims at revamping the transport protocol for the wireless domain using (a) rate-based transmission at the source, (b) inter-packet separation at the receiver as the congestion metric, (c) mechanisms for detecting the reason for packet loss, and (d) bandwidth estimation, as some of the underlying principles. A unique characteristic of WTCP is the attempt to separate the congestion control and reliability mechanisms. WTCP uses separate sequence numbers for congestion control and reliability mechanisms in order to distinguish the two. The reliability mechanism involves a combination of selective and cumulative acknowledgments, and takes into account the reverse-path characteristics for determining the ACK frequency.

1.4.9 TCP SACK

The selective retransmission strategy [14] is more complex and requires more buffer space at the end-points. Hence TCP traditionally uses cumulative acknowledgments and the go-back-N strategy. Using selective retransmit reduces the overhead of retransmission on errors and therefore cannot be ruled out for use in wireless domains. The TCP with selective ACK scheme
(TCP SACK) [16], [17] improvesTCP performance by allowing the TCP sender to retransmit packets based on the selective ACKs provided by the receiver.

1.4.10 Transaction-Oriented TCP

The TCP connection setup and connection tear-down phases involve a huge overhead in terms of time and also in terms of the number of packets sent. This overhead is very costly, especially if the size of the data is small. An alternative for such transactions is transaction-oriented TCP (TTCP) [18]. The motivation behind this approach is to integrate the call setup, the call tear-down, and the actual data transfer into a single transaction, thereby avoiding separate packets for connecting and disconnecting. However, the flip side to the strategy is that changes must be made to TCP, which goes against some of the fundamental objectives that the changes to TCP must be transparent and must not affect the existing framework. Table 1.3 shows a summary of the various approaches discussed so far. The next section briefly describes the impact of mobility on the performance of TCP.

Table 1.3. Summary of proposed protocols to improve the performance of TCP over wireless

Feature	Snoop TCP	TCP-Unaware Link Layer	Mobile TCP	ITCP	ELN	WTCP	TCP SACK	TTCP
Changes in:								
AP	Yes	Yes	Yes	Yes	No	No	No	No
CN	No	No	No	No	Yes	Yes	Yes	Yes
MN	Yes	No	Yes	Yes	Yes	Yes	No	No
Retransmitting Node	AP	AP	NA*	AP	NA	NA	NA	NA
Single Point Failure	No	No	No	Yes (AP)	No	No	No	No
Handoff Latency	Low	Low	Low	Low	High	High	High	High
Security	Breach at AP	No breach	NA	Breach at AP	No breach	No breach	No breach	No breach
End-to-End Semantics	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Retransmissions by Intermediate Nodes	Yes	Yes	No	Yes	No	No	No	No
Slow Start	Yes	Yes	No	NA	No	No	No	Yes
Buffer at AP	Yes	Yes	No	Yes	No	No	No	No

*Not Applicable

1.4.11 Impact of Mobility

Handoffs occur in wireless domains when an MN moves into a new BS's domain (a cell in the cellular context). If the link layer ensures reliable delivery and guarantees zero loss during a handoff, then TCPwill be totally unaware of the handoff and no measures need to be taken at the transport layer to support handoff. The only exception to this is when the handoff latency is too large and exceeds theTCP timeout; then the transparency of handoffs to TCP is lost.

Fast Retransmit/Recovery

The usual problem associated with handoffs is that the handoff may lead to packet loss during transit, either as a result of the intermediary routers' failure to allocate adequate buffers or their inability to forward the packets meant for
the MN to the new BS. The result of the packet loss during handoff is slow start. The solution involves artificially forcing the sender to go into fast retransmission mode immediately, by sending duplicate acknowledgments after the handoff, instead of going into slow start. The advantage of the strategy is its simplicity and the fact that it requires minimal changes to the
existing TCP structure. However, the scheme does not consider the fact that there may be losses over the wireless links.

Using Multicast

Multicast has been suggested to improve the performance of TCP in the presence of handoffs [10]. The idea is similar to the one used in MRSVP [9], where the MN is required to define a group of BSs that it is likely to visit in the near future. These include the current cell (or the current BS) the MN is attached to and also the cells (BSs) likely to be visited by it. These BSs are then directed to join the multicast group, the address being the unique multicast address assigned to the MN. Packets destined for MNwill have to be subsequently readdressed to the multicast group. In the implementation, only one BS is actually in contact with the MN and is responsible for transmitting the packets to it. If the rest of the BSs in the multicast group are able to buffer the packets addressed to the multicast address, then the loss of packets during the handoff can be significantly minimized. There is a trade-off between buffer allocation at the BSs and the loss during handoff. In practical situations, the number of buffers allocated can be minimized by buffering only when a handoff is likely to occur.

1.5 WAP

WAP stands for wireless application protocol. This name is a misnomer, because WAP represents a suite of protocols rather than a single protocol. WAP has today become the *de facto* standard for providing data and voice services to wireless handheld devices. WAP aims at integrating a simple lightweight browser also known as a micro-browser into handheld devices, thus requiring minimal amounts of resources such as memory and CPU at these devices. WAP tries to compensate for the shortfalls of the wireless handheld devices and the wireless link (low bandwidth, low processing capabilities, high bit-error rate, and low storage availability) by incorporating more intelligence into the network nodes such as the routers, Web servers, and BSs. The primary objectives of the WAP protocol suite are independence from the wireless network standards, interoperability among service providers, overcoming the shortfalls of the wireless medium (such as low bandwidth, high latency, low connection stability, and high transmission cost per bit), overcoming the drawbacks of handheld devices (small display, low memory, limited battery power, and limited CPU power), increasing efficiency and reliability, and providing security, scalability, and extensibility.

1.5.1 The WAP Model

WAP adopts a client-server approach. It specifies a proxy server that acts as an interface between the wireless domain and core wired network. This proxy server, also known as a WAP gateway, is responsible for a wide variety of functions such as protocol translation and optimizing data transfer over the wireless medium. Figure 4.8 illustrates the client-server model that WAP employs. The WAP-enabled handset communicates with a Web content server or an origin server [that may provide hypertext markup language (HTML)/common gateway interface (CGI) content] via a WAP gateway. It is at the WAP gateway that the convergence of the wireless and wired domains actually occurs. The gateway receives WAP requests from the handset, and these have to be converted into suitable HTTP requests to be sent to the origin server. If the origin server cannot provide the required information in wireless markup language (WML) form, then there must be an additional filter between the server and the gateway to convert the HTML content into WAP-compatible WML content. The gateway may additionally perform functions such as caching and user agent profiling as part of some optimization measures. This is also known as *capability and preference information.* By means of user agent profiling, the MN specifies its characteristics such as hardware characteristics, software capabilities, and user preferences, to the server so that the content can be formatted appropriately to be displayed correctly.

1.5.2 The WAP Protocol Stack

The WAP protocol stack is designed in a layered fashion that allows the architecture to provide an environment that is both extensible and scalable for application development. The WAP architecture

allows other services to access the WAP stack at well-defined interfaces. Figure 4.9 gives an overview of the different layers in the WAP protocol suite and also their basic functionalities. This section provides a brief description of some of the important layers in the protocol stack.

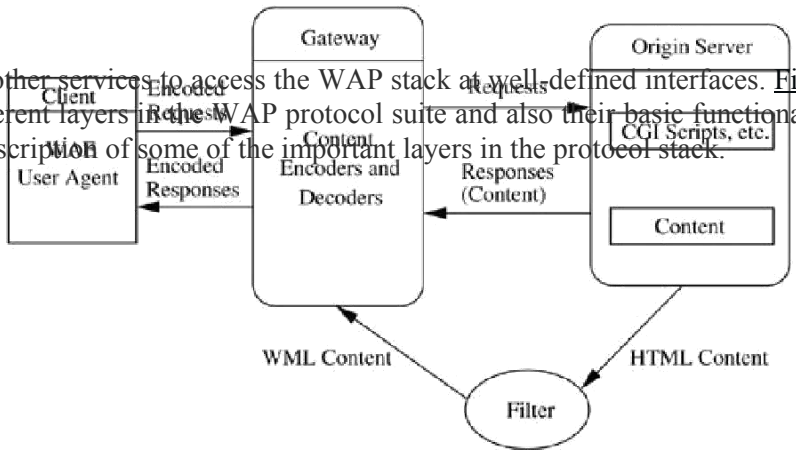
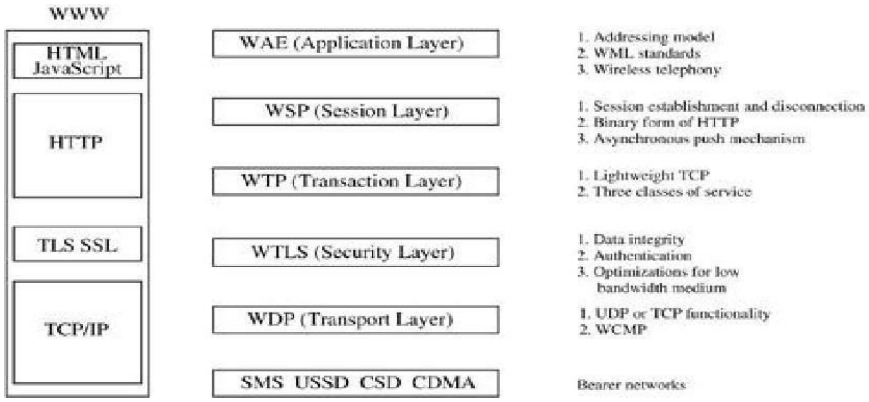


Figure 1.9. The WAP protocol stack.



The Wireless Application Environment

The wireless application environment (WAE) has a number of components that address specific issues in the application environment. The WAE provides for an addressing model for accessing both the WWW URLs and other resources specific to the wireless domain using uniform resource identifiers (URIs). The WAE uses WML as the standard markup language, which can be construed as an efficient binary encoded form of the traditional HTML. The WAE also provides a compact scripting language analogous to JavaScript. The WAE also provides for a set of telephony applications through the wireless telephony application interface (WTAI).

Wireless Session Protocol

The wireless session protocol (WSP) establishes a reliable session between the client and the server and also ensures that the session is released in an orderly manner. The push mechanism is a fundamental component of the WAP programming model aimed at reducing the number of requests made by the client to the server. A data server will asynchronously push the information to the registered client(s) efficiently using this mechanism. This is especially useful in multicast and broadcast

applications. The WSP provides the equivalent of HTTP in the WWW domain. The core of the WSP design is a binary form of HTTP. A session may be suspended to save power at the clients, but the session reestablishment follows only a small procedure that avoids the overhead of starting a full-fledged session afresh.

Wireless Transaction Protocol

The wireless transaction protocol (WTP) can for all practical purposes be viewed as a lightweight version of TCP. A transaction is defined as a request/response cycle. The WTP has no explicit setup and tear-down phases like TCP, as this would cause a tremendous overhead. There are no security options at the transaction layer in the WAP stack. WTP defines three categories or classes of service:

1. Class 0: Unreliable send (push model) with no ACK. There is no retransmission in case the message is lost. This is essentially a connection-less service.
2. Class 1: Reliable push service, where a request is sent and the responder sends the data as an implicit acknowledgment to the request. The responder maintains this state for some time to handle possible retransmissions.
3. Class 2: This is the classical request-data-ACK cycle providing a two-way reliable service.

Wireless Transport Layer Security

The objective of the wireless transport layer security (WTLS) is to provide transport layer security between the WAP client and a WAP server. WTLS is based on the industry standard transport layer security (TLS) protocol with certain features such as datagram support, optimized handshake, and dynamic key refreshing. The primary objectives of WTLS are data integrity, privacy, authentication, and denial of service (DoS) protection. WTLS has capabilities to detect and reject data that is not successfully verified; this protects servers from DoS attacks.

Wireless Datagram Protocol

The wireless datagram protocol (WDP) defines the WAP's transport layer in the protocol suite. TheWDP has an adaptation layer that is bearer-specific that helps optimize the data transfer specific to a particular bearer service (such as SMS, USSD, CSD, and CDMA). If the underlying bearer service uses the IP standard user datagram protocol (UDP), then there is no necessity for a separate functionality at the WDP layer as UDP itself is used. The wireless control message protocol (WCMP) is responsible for providing the error-handling mechanisms analogous to Internet control message protocol (ICMP).

1.5.3 WAP 2.0 and i-mode

The i-mode (information-mode) system, developed in Japan and a major competitor to WAP, has three main components: a transmission system, a handset, and a language for designing Web pages. The transmission system consists of the existing mobile phone network (which is circuit-switched) and a new packet-switched network. Voice transmission uses the existing mobile phone network while data transmission uses the packet-switched network and is billed based on the number of packets transmitted as opposed to connection time. i-mode uses a subset of HTML called as cHTML (compactHTML). In contrast, WAP 2.0 was developed by the WAP Forum and is likely to use packet-switched network. WAP 2.0 has new features such as multimedia messaging, pull (request for data, then receive the data) as well as push model (asynchronous data transfer, without requiring explicit requestmessages, such as stock prices), integrated telephony, interoperability with WAP 1.0, and support for plug-ins in the browser. Unlike i-mode, WAP 2.0 charges the users based on connection time.

1.6 OPTIMIZING WEB OVER WIRELESS

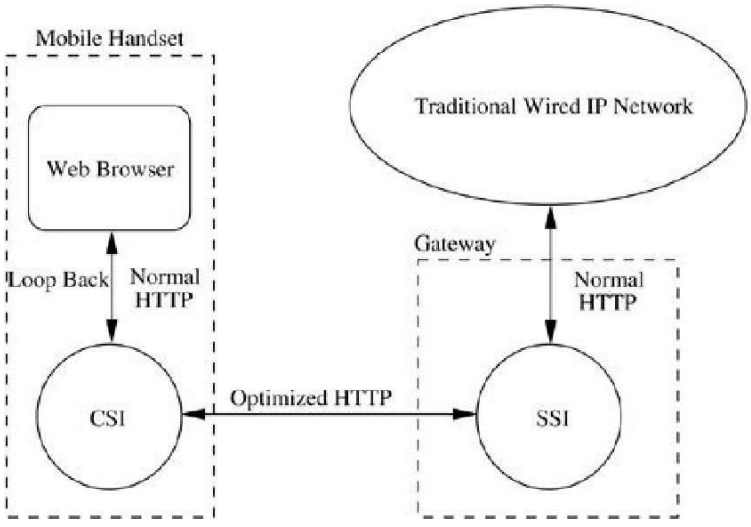
The limitations of wireless networks that provide the motivation for such optimizations are low

bandwidth, low reliability, high latency, and high cost per byte transferred. Integrating Web access over wireless devices would have to take into account the drawbacks of the wireless medium and the capabilities of the devices. Systems such as WebExpress [19] are aimed at optimizing routine repetitive browsing; many of the mechanisms suggested may not be suitable for random browsing (*i. e*., there are no perceivable trends in the Web accesses). Web browsers must offer a good interface for the wireless devices, keeping in mind the network, memory, processing power, and power consumption constraints.

1.6.1 HTTP Drawbacks

The main protocol on which the Web operates today is the hypertext transfer protocol (HTTP), which is optimized mainly for the wired world. It has a lot of overhead, but it is acceptable when the network bandwidth is an inexpensive resource as in typical wired networks compared to wireless networks.HTTP has drawbacks such as high connection overhead (a new TCP socket is opened for every newHTML object), redundant capabilities transmission (information regarding the browser capabilities is included in every HTTP request), and verbosity (HTTP is ASCII-encoded and hence inherently verbose). The WebExpress system suggests that an Intercept model be applied for Web access over wireless interfaces. This allows the number of requests sent over the wireless channel to be optimized, and also avoids the connection setup overhead over the wireless interface. There are two main entities that are introduced into the system: the client side interface (CSI) and the server side interface (SSI). The CSI appears as a local Web proxy co-resident with the Web browser on the wireless rendering device, say, a mobile phone or a PDA. The communication between the CSI and the Web browser takes place through the loopback feature of the TCP/IP suite (wherein the host sends a packet to itself using an IP address like 127.0.0.1). The communication between the CSI and the SSI is the only interaction over the wireless network and this uses a reduced HTTP, as discussed later. The SSI communicates with the Web server over the wired network. The SSI could typically be resident at the network gateway or the FA in MobileIP. The intercept model (Figure 4.10) is transparent to browsers and servers, and is also insensitive to changes in HTTP/HTML technology.

Figure 1.10. The intercept model.



1.6.2 Optimizations

Four main categories of optimizations that can improve the performance of Web access systems over wireless channels can be identified. These are:

• **Caching:** Current caching technologies are suited for wired applications. Cache objects are either purged at the end of the session or may persist across sessions. But it is advantageous to have cached

data persist across browser sessions, as this increases cache hit ratios. Appropriate cache coherency methods are added to detect and change old information.

• **Differencing:** For transaction processing (involving forms) caching techniques do not help as different replies to the same application server are often different. Still, the fact that these replies tend to be similar can be exploited to reduce the network traffic over the wireless interface. A base object carries fundamental features that do not change across transactions and is
created and maintained by both the client and server interfaces. Whenever a new transaction takes place, the server computes the difference stream and only the difference stream is transmitted.

• **Protocol reduction:** This approach aims at reducing the overhead of repeated setup and tear-down of TCP/IP connections for each Web-object to be transmitted. This can be eliminated by establishing a single TCP/IP connection between the CSI and the SSI that will persist for the entire session. The connection setup/tear-down overhead is on the local and wired connections only.

• **Header reduction:** HTTP requests are prefixed with headers that indicate to the origin server the rendering capabilities of the browser and also the various content formats handled by it. The alternative to this is that the CSI sends this information in the first request and SSI records this information.

For every subsequent request sent by the CSI, the SSI automatically inserts this capability list into each packet meant for the origin server.

1.7 SUMMARY

This chapter focused on the issues in wireless networks that are pertinent to the higher layers in the protocol stack, the network layer, the transport layer, and the application layer. The various aspects of the wireless Internet, that is, extension of the services offered by the Internet to the wireless domain, were discussed. Mobile IP aims at providing network connectivity to mobile hosts, and it is in a larger sense not restricted to wireless networks. The inefficiencies of MobileIP routing can be tackled in both generic techniques such as the optimizations incorporated in IPv6 and specific techniques as suggested in the 4×4 approach. The network layer also has to address the issues of security, accounting, and handoffs, as these are of great significance in wireless networks; some of the relevant issues were also discussed.

This chapter also discussed the issues in adaptation of TCP to the wireless domains, as it has been shown that the existing transport framework would perform miserably when used in its current form (optimized to work with high bandwidth, low-error wired networks). Most of the solutions involved some capability at the BS to buffer packets and also act on behalf of the MNs to send ACKs/NACKs. The strategies discussed in this chapter were broadly classified into various categories based on the role of the BS. The WAP architecture specified provides for an efficient, interoperable, and scalable framework for developing and using applications in the wireless domain. WAP 2.0 added more features to the previously existing WAP 1.0 protocol.