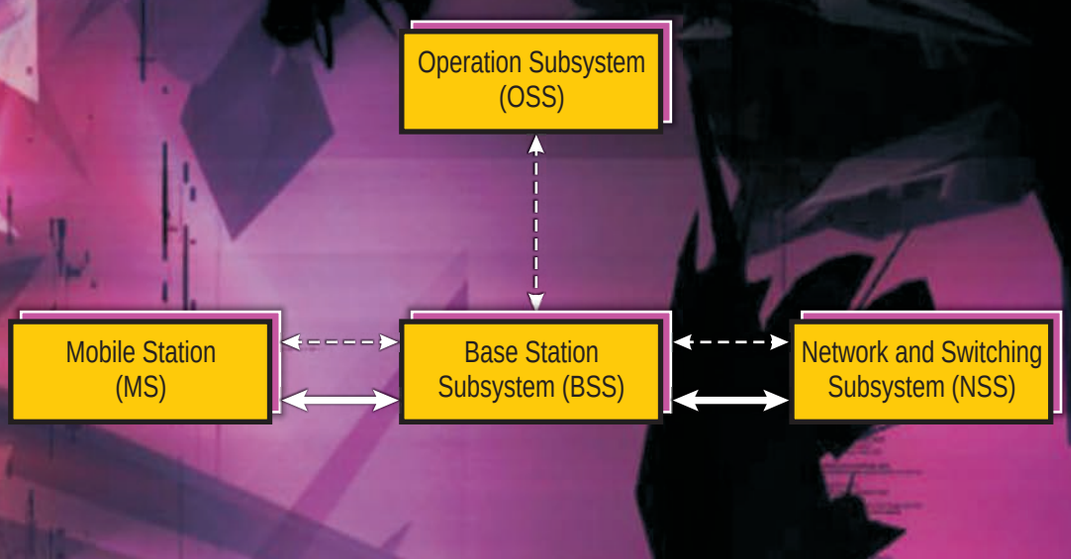
Eastern Economy Edition

Fundamentals of Mobile Computing



Prasant Kumar Pattnaik
Rajib Mall

PHI

FUNDAMENTALS OF
MOBILE COMPUTING

PRASANT KUMAR PATTNAIK
Associate Professor
School of Computer Engineering
KIIT University
Bhubaneswar

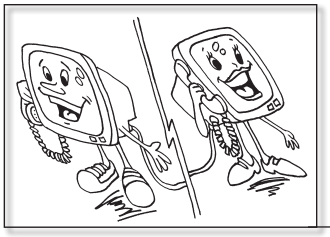
RAJIB MALL
Professor
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

PHI Learning Private Limited
New Delhi-110001
2012

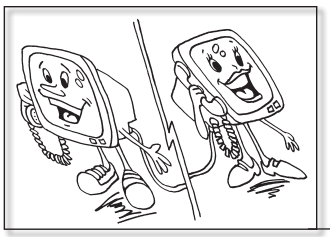
FUNDAMENTALS OF MOBILE COMPUTING
Prasant Kumar Pattnaik and Rajib Mall

ISBN-978-81-203-4632-1

Published by Asoke K. Ghosh, PHI Learning Private Limited, M-97, Connaught Circus, New Delhi-110001 and Printed by Raj Press, New Delhi-110012.



Contents



Preface

This introductory text in the area of mobile computing is primarily based on the classroom teachings of the first author. Mobile computing is progressing at a rapid pace on account of the continual development of new technologies and systems. Mobile computing is also included in the curricula of several universities. The authors, therefore, decided to work on an introductory textbook on this subject for the benefit of the students and the teachers. The authors fervently hope that the readers will find this book interesting as a whole and benefit from it by learning about the fundamental range of topics covered. The authors in no way claim that the material in the book is comprehensive.

The text is organized into eleven chapters and a glossary. Chapter 1 provides an overview of the basics of wireless technologies and computer communications; it explains how a cellular communication system functions—the technology that has revolutionized society.

Chapter 2 elucidates what mobile computing is and how it differs from wireless networking. Several generations of cellular communication technologies are discussed.

Chapter 3 describes how the medium access control (MAC) protocol works in regulating the use of a shared physical channel among a set of nodes. The chapter brings out, in essence, that protocol design for mobile ad hoc networks is much more difficult and complex than that for wireless LANs.

Chapter 4 provides an overview of the working of mobile internet protocol (Mobile IP). Chapter 5 reviews some basic aspects of the TCP/IP protocol suite. It brings out problems that might arise when TCP is used, as it is, in mobile wireless networks. It then discusses the adaptations that have been extended to make TCP work satisfactorily in the mobile environment.

Chapter 6 on mobile databases explains how a mobile database is an adaptation of a traditional database to accommodate many of the issues

that a mobile environment imposes, such as disconnections, low battery power, etc. A transaction model for a mobile computing application is also discussed.

Chapter 7 is devoted to mobile ad hoc networks (MANETs). It discusses routing protocols that have been proposed for use in MANETs, their security vulnerabilities and the security measures incorporated at different protocol layers. The routing protocols used in wireless sensor networks, the challenges faced in the design and operation of wireless networks and the various areas of their applications are all discussed in Chapter 8.

The features required of a mobile operating system, including those available in popular commercial OSs, are discussed in Chapter 9. Chapter 10 discusses J2ME and Android SDK that facilitate application development for mobile phones.

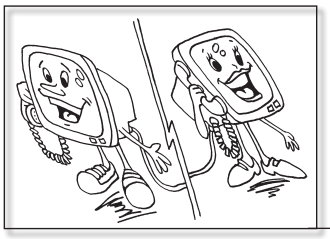
Finally, in Chapter 11 an example of mobile computing infrastructure, that is, mobile commerce (M-commerce) has been explained in simple words.

The authors warmly welcome all readers to send their suggestions for improvements to the book. All suggestions from readers towards enhancements of the contents and the presentation style will be duly acknowledged in the next edition of the book.

The first author is thankful to his wife Bismita and daughter Prasannakshi, and the second author is thankful to his wife Prabina and daughter Mithi for patiently and selflessly permitting them to devote time towards the development of the book. Besides, both the authors profusely thank their colleagues and friends for their cooperation and help during the preparation of the manuscript.

Prasant Kumar Pattnaik
patnaikprasant@gmail.com

Rajib Mall
rajib@cse.iitkgp.ernet.in



1

Basics of Communication Technologies

In this chapter, we review a few basic concepts related to telecommunications and computer communications that form the essential infrastructure required for building knowledge in the area of mobile computing. In particular, we focus here on the basics of wireless technologies and computer communications. These concepts form the prerequisites that we implicitly assume in the treatment of this text.

Mobile computing refers to the computational tasks performed by mobile users using their handsets. Since the handsets have very limited processing power and memory, these devices by themselves do not have the capability to carry out any significant and meaningful computations and can only serve as the front-end for invoking remote applications. Mobile computation, therefore, inevitably involves the invocation of applications running on remote servers. In other words, mobile computation is usually achieved by the interaction of a front-end application running on the mobile handset with a server, seamlessly, through the medium of wireless communication. Therefore, a study of mobile computation involves the study of the invocation mechanisms at the client end, the underlying wireless communication, and the corresponding server-side technologies.

1.1 Mobile Handsets, Wireless Communications, and Server Applications

Currently the technologies forming the client-side (hand-helds) and the server-side computations are in a state of flux and are rapidly evolving. However, the wireless networking technologies have undergone significant advancements over the last few decades and are nearly maturing. Wireless networking has its roots in radio communications. In wireless

communication, individual users with handsets may directly communicate with each other over a radio link or via several radio and wired links formed through some intermediaries such as base stations and fixed line networks. When all the intermediaries are located on the ground, then the communication system is referred to as a *terrestrial radio system*. If at least one of the intermediaries is satellite borne, then it is referred to as a *satellite radio system*.

Radio transmission was first discovered by Marconi in the year 1895, that is, more than a century back. It would not be an overstatement to say that since then the radio communication has revolutionized the entire human civilization. Starting with the elementary, ground-breaking point-to-point wireless communication achieved by Marconi, there has been continually the induction of new and increasingly sophisticated technologies. Sophisticated communication techniques such as conventional radio, television, digital packet radio, cellular phone, wireless LAN, and wireless-based Internet access have rapidly been developed over a relatively short span of time. These have impacted our daily life, culture, and society in no mean ways. More recently, a technology that has revolutionized the human society and impacted almost everybody, is the *cellular communication technology*.

1.2 Cell Phone System

Within a short span of less than two decades, more than half the world's population today owns cell phones. The cell phones have achieved remarkable acceptance all over the world due to their major advantages over the traditional phones. Besides providing the flexibility of communicating while on the move, the cell phone system also provides data services such as Short Message Service (SMS), Multimedia Messaging Service (MMS), and even email and web browsing while on the move. Further, unlike a traditional network, in a mobile telecommunication network, even when the user moves, the current location of the mobile phone is maintained by the network so that any voice call or SMS can be easily sent to the handset wherever it may be. This has made it possible for people to communicate and carry out important work anytime and from anywhere. For this reason, mobile computing is often referred to as *ubiquitous computing* as well. Today, we can say that cellular communication is influencing virtually everyone's life in some way or the other. Mobile computing is poised to achieve higher levels of success in the current decade.

Before we proceed further, a brief narrative of how a cellular communication system functions is presented in Box 1.1 to enable readers to familiarize themselves with the architecture of the system.

BOX 1.1 How a cellular communication system functions

Cellular telephony is based on radio frequency communications. A cellular communication system divides the service areas into different geographic zones called *cells* (see Fig. 1.4). Each cell can support a certain maximum number of simultaneous user connections. Consequently, more cells are required per unit area in cities, simply because of more people dwelling in cities and hence more number of simultaneous calls occurring per unit area. Each user has a mobile handset that has its own radio transmitter and receiver antenna to establish connection to the local base station. The base station in turn connects to a switching centre. The switching centre act as a middleman between two mobile handsets and establishes connections between them when requested. As a user moves from one place to another, his call is handled by the switching centre of the base station located in another cell site, thus, providing a consistent, high quality signal. When a user travels outside his home network, he can still make calls. This is possible by a facility called *roaming* and it works because the base station in the local cell is connected by wires with all other networks. This makes it possible for a user to get his call almost anywhere. If someone dials a user's number, the network finds him and gives him the ring tone. We will elaborate this narrative and give a more detailed overview of the cellular communication architecture in Section 1.8.

1.3 Types of Telecommunication Networks

A popular way of classifying telecommunication networks is into *voice networks* and *data networks*. The voice networks were the first ones to be developed and formed the predominant telecommunication networks in the nineteenth and twentieth centuries. In a voice network, analog traffic is usually modulated on a carrier signal for spectral and transmission efficiency. The traditional telephone networks that carry voice traffic in analog form are an important example of a voice network. *Data networks* are of much more recent origin. The data networks carry data in digital form. The data signals are also modulated on an analog carrier signal for spectral and transmission efficiency. The term *data* here refers to any information such as text, documents, picture, movie, sound, etc., which needs to be first coded into a bit stream. Of course, analog signals such as voice, video, etc., have to be first quantized into digital data for coding into bit streams.

It is worth noting that the traditional voice networks such as PSTN (public switched telephone networks) relied on circuit switching, whereas data networks are more recent and are based on the store-and-forward packet switching mechanism. A big advantage of the packet-switched networks is that they make efficient use of the transmission medium and therefore are much more cost effective compared to circuit-switched networks. On the other hand, an inherent shortcoming of using a store-and-forward network for digitized voice signal transmission is that in a store-and-forward network, every switch inspects the packet to determine

its destination and based on this information the network sends the packet on its way forward. Further, depending on the traffic conditions, packets can get stored at a switch for a significant amount of time causing queues to get built up at various switches. The cumulative delays that a packet undergoes at various switches can be hundreds of times longer than the propagation delay. So, the different packets of a voice call can undergo different amounts of delay causing degraded sound quality. In a voice call, even a moderate delay at a switch can cause crackling or popping sounds at the receiver. Unless the switches are built using very high-speed hardware, the receiver of a call would hear such distracting noises. This was possibly a reason why packet switching was not used for voice calls until recently. The modern day networks deploy high-speed, powerful switches and routers. This has made it possible for voice to be transmitted over packet-switched networks and hence VoIP (Voice over Internet Protocol) has become a very popular and cheap medium of communication. Computer networks are predominantly based on the packet-switching technology.

1.4 Computer Networks

Several types of computer networks are in use today. Of these, we discuss here Controller Area Networks (CANs), Local Area Networks (LANs), and internetworks.

1.4.1 Controller Area Networks

A Controller Area Network (CAN) is essentially a very small network that is typically used to connect the different components of an embedded controller. The end-to-end length of a CAN is usually less than 50 metres. Since the propagation time of a CAN is very small, it behaves more like a local bus in a computer.

To understand the genesis of CAN and its operation, consider the present-day automobiles. The area of automotive electronics has becomes fairly sophisticated for providing support to several activities such as engine management, fuel injection, active suspension, braking, lighting, air conditioning, security and central locking. A considerable amount of information exchange among the various automotive components takes place when the engine is operational. The conventional method of networking the components in older models of cars was the point-to-point wiring of the different electrical components such as motor generator, lamps, battery, ignition system, etc. As cars became more sophisticated, the use of such a naive scheme would have required several kilometres of wiring, adding not only to the cost of manufacturing but also contributing to severely reduced reliability.

The limitations of fixed point-to-point wiring techniques in handling the demands of modern automated cars and other embedded applications, gave rise to the development of CAN. A special requirement placed on CAN is that it should be able to effectively handle noise. Automotive components such as electric motors, ignition systems, as well as RF transmissions, are heavy producers of noise. Another requirement imposed on CAN is the use by it of the 12-volt power supply that was mandated by the conventional 12-volt automotive power supply. CAN specifies only the physical and data link layers of the ISO/OSI model while the higher layers are left open for specific implementations.

Because of its robustness, the use of CAN has extended beyond its automotive origins and can now be found in diverse application areas such as industrial automation systems, trains, ships, agricultural machinery, household appliances, office automation systems and elevators. Now CAN is an international standard under ISO 11898 and ISO 11519–2.

1.4.2 Local Area Networks

A Local Area Network (LAN) is typically deployed in a building or a campus and is usually privately owned. For example, a LAN can be used to connect a number of computers within an organization to share data and other resources such as files, printers, FAX services, etc. LANs typically operate at data rates exceeding 10 Mbps and many present-day LANs (gigabit Ethernets) operate at 1 Gbps.

1.4.3 Internetworks

Several LANs can be interconnected using switches to realize internetworks or internet in short. Figure 1.1 shows four separate LANs interconnected using switches to form an internet.

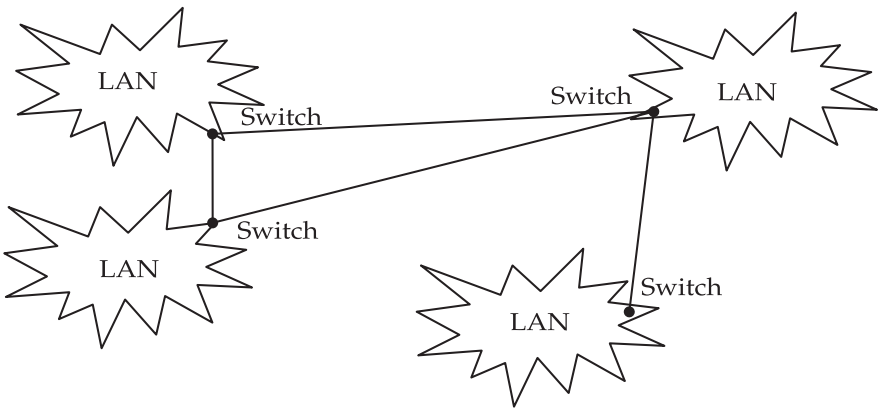


Figure 1.1 *Structure of an internet.*

In an internet, a node in a LAN communicates with a node in another LAN using packet switching. Packet switching implies that messages are divided into relatively small units of data called packets at the sender's end before they are sent over the network. Each packet is transmitted individually and traverses the network using the store-and-forward principle. A packet can even follow a route different from that taken by other packets of the same message to reach its destination. The individual packets are routed through the network based on the destination address contained within each packet. Once all the packets forming a message arrive at the destination, they are ordered into the original message. The store-and-forward principle followed in packet switching allows the same data path to be shared among message transmissions from different sources. The networks of this type are known as connectionless, in sharp contrast to the dedicated connections realized in connection-oriented networks. The two major packet-switching modes are, therefore, known as *connectionless* (also known as datagram switching) and *connection-oriented* (also known as virtual circuit switching). We do not discuss these here and any reader unfamiliar with these can refer to a first-level computer networking book. Wide area networks (WANs) are important examples of packet-switched computer networks. The most prominent WAN is the Internet.

The Internet, sometimes referred to as "the net," is a worldwide system of computer networks—a network of networks in which users at any one computer can, if they have permission, get information from any other computer in the net using protocols such as ftp, http, etc. A user can even talk directly to other users at other computers in the Internet using VoIP. We must, however, be aware of the following two distinctions: the term internet is used to denote a single logical network realized from a collection of LANs by using an internetworking protocol, while the Internet refers to the worldwide collection of interconnected networks which evolved from the ARPANET project. It uses the Internet Protocol (IP) to link the various physical networks spread all over the world into a single logical network. The Internet began in 1962 as an attempt towards realizing a resilient computer network for the US military and since then has grown into a global network of more than a billion computers connected to several of thousands of smaller computer networks, based on a common addressing scheme called the IP address.

1.5 Traditional LAN

A traditional LAN architecture is schematically shown in Fig. 1.2. It is the same as the bus architecture discussed in Section 1.6.1. There is a single shared channel (bus) and only one node can transmit at any time. The exact time when a node can transmit on the channel is determined by the *access arbitration policy* of the network. The *transmission control policy*

determines how long a node can transmit. These two policies together are called the *access control policy* and form the Media Access Control (MAC) layer protocol. In other words, the MAC protocol in a LAN consists of two parts: an access arbitration part that determines when a node can use the channel and the transmission control policy decides how long it can use it, once it starts using it. Let us first review some basic aspects of LANs that are crucial to understanding our subsequent discussions.

1.6 LAN Architectures

Two major LAN architectures are being used: the bus architecture and the ring architecture. These two architectures use different access control techniques. We first briefly discuss these two LAN architectures, and then discuss the associated access control techniques.

1.6.1 Bus Architecture

The schematic representation of a bus-based network architecture is shown in Fig. 1.2. In a bus-based architecture, nodes are connected to the network cable using T-shaped network interface connectors. The terminating points are placed at each end of the network cable. There is a single, shared channel (bus) for which the transmitting nodes contend to gain access to it. In a bus architecture, nodes communicate using broadcasting. The most commonly used protocol for access control in traditional bus networks is the Carrier Sense Multiple Access with Collision Detection (CSMA/CD). These networks are also called multiple access networks. In these networks, when two or more nodes transmit packets simultaneously, the transmissions overlap in time and the resulting signal gets garbled. Such an event is called a collision. A collision entails retransmission of the corrupted frame.

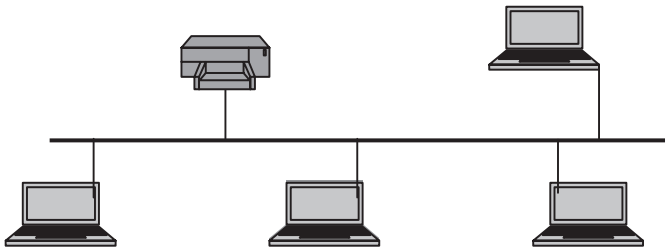


Figure 1.2 *A traditional bus-based LAN.*

In CSMA/CD networks, nodes continuously sense the channel to determine whether the channel is idle. A node transmits a packet only if it senses the channel to be idle. But, this does not guarantee that there

will be no collisions. Several nodes might sense the channel to be idle at the same time instant and start transmitting simultaneously, resulting in a collision. Also, even after a node starts transmitting, another node may not sense it till the signal has propagated to it. It should, therefore, be clear that the larger the propagation delay of a network, the higher is the probability of collisions of packets. Thus, a long LAN would suffer from too many collisions and not be usable. This sets a limit for the maximum LAN size.

Ethernet is a LAN standard based on the CSMA/CD media access control protocol. CSMA/CD protocol does not define any specific collision resolution mechanism. On the other hand, Ethernet uses the Binary Exponential Back-off (BEB) algorithm for collision resolution. Due to its ubiquity, high speed, simplicity and low cost, Ethernet has over the years emerged as one of the most preferred LAN protocols. It is, therefore, not surprising that many attempts have been made in the past to develop protocols based on Ethernet to support real-time communication. As far as the real-time communications are concerned, the logical ring architecture possesses significant advantages over Ethernet due to its inherent deterministic access arbitration mechanism in contrast to the collision-based mechanism of Ethernet. In a collision-based network, under high load situations the number of collisions per unit time would increase very rapidly with load, leading to increased retransmissions, rapid drop in throughput and rise in delay. As a result, in Ethernet the delay in message transmission increases rapidly as the traffic increases.

1.6.2 Ring Architecture

In a ring network shown in Fig. 1.3, nodes are placed along the ring. The nodes transmit in turn. Each node usually transmits for a certain predetermined period of time. Therefore, packet transmission delays become predictable and can be made sufficiently small as per requirement. As a result, ring-based architectures are often preferred in real-time applications.

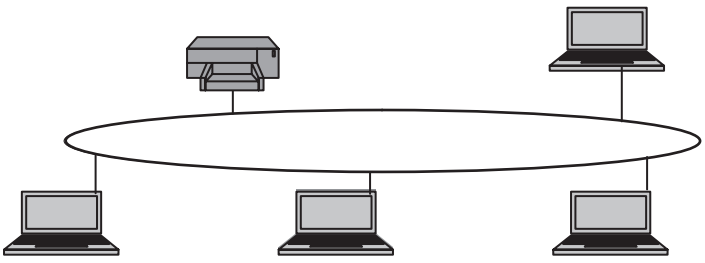


Figure 1.3 *A ring network.*

The ring architecture, however, suffers from a few important problems. First, any break in the ring can bring the whole network down. This makes

reliability of ring networks a major concern. Further, ring is a poor fit to the linear topology normally found in many situations. This made researchers to look for alternative technologies which can have the advantages of both the bus and ring architectures. This led to the development of the token bus architecture. A token bus is a bus-based architecture where the stations on the bus are logically arranged in a ring with each station knowing the address of the station to its "left" and "right".

When the logical ring is initialized, the highest numbered station gets a chance to start its transmission. After transmitting for a predetermined duration, this station passes the transmission permission to its immediate neighbour (left or right as per the convention adopted) by sending a special control frame called a *token*. The token propagates around the logical ring. At any time, only the token holder is permitted to transmit packets. Since only one station at a time holds the token in a ring network, collisions cannot occur. An important point that must be kept in mind is that the physical order in which the stations are connected to the cable need not be the same as the order of the stations in the logical ring. This is because a cable is inherently a broadcast medium. Each station would receive every frame transmitted, and discard those that are not addressed to it. After a node exhausts it assigned time slot for transmission, it hands over the token to its logical neighbour. For this, it transmits a special token, specifically addressed to its logical neighbour in the ring, irrespective of whether that station is physically adjacent to the concerned node or not. The MAC protocol in a token ring network should support adding stations to and removing stations from the logical ring. The protocols based on the underlying ring architecture hold the advantage of being able to guarantee the delivery time of a packet.

1.7 Components of a Wireless Communication System

A wireless communication system is built from various types of basic components. The following are some of these basic types of components.

Transmitter: The input to a wireless transmitter may be voice, video, data or other types of signal meant to be transmitted to one or more distant receivers. This signal is called the baseband signal. The basic function of the transmitter is to modulate, or encode several baseband signals onto a high frequency carrier signal. A modulated high frequency signal can be radiated and propagated more effectively and helps make more efficient use of the radio frequency (RF) spectrum than what the direct transmission of the individual baseband signals can achieve.

Receiver: The receiver receives the modulated signals and reverses the functions of the transmitter component and thereby recovers the transmitted

baseband signal. The antenna of the receiver is usually capable of receiving the electromagnetic waves radiated from many sources over a relatively broad frequency range.

Antenna: The function of an antenna is to convert the electric signal from a transmitter to a propagating electromagnetic RF wave; or conversely, to convert a propagating RF wave to an electric signal in a receiver. In a *transceiver*, a transmitter and a receiver are co-located for supporting full-duplex communications. In this case, the same antenna is usually shared by both the transmitter and the receiver. There are mainly two types of antennas that are used on wireless networks: omni-directional and directional. Omni-directional antennas can receive and transmit over 360 degrees. This can be compared to a light bulb that emits light all over. In contrast, a directional antenna is similar to a flashlight, it focuses light in some direction.

Filters: Filters are a key component present in all wireless transmitters and receivers. They are used to reject interfering signals lying outside the operating band of receivers and transmitters. They also reject unwanted noise signals generated by the amplifier circuitry.

Amplifiers: An amplifier amplifies the strength (usually voltage) of a signal. The important specifications of an amplifier include power gain and the noise figure. The noise figure of an amplifier is a measure of how much noise is added to the amplified signal by the amplifier circuitry. This is most critical in the front-end of the receiver where the input signal is very weak and it is desirable to minimize the noise added by the receiver circuitry. Therefore, it is necessary that the first amplifier in the receiver circuit has as low a noise figure as possible.

Mixers: A mixer is typically used to achieve frequency conversion at the transmitters and receivers. Frequency conversion is required because it is advantageous to transmit signals at a higher frequency. This is achieved by modulating a carrier waveform using the original baseband frequency. When a baseband signal is mixed appropriately with a high frequency on a carrier, it can be easily and efficiently radiated and becomes less susceptible to noise and attenuation. Therefore, the transmission range increases and the received signal quality improves. Further, multiple baseband signals can be mixed with a carrier appropriately in order to efficiently utilize the spectral bandwidth. This forms the essence of any signal modulation technique. When multiple baseband signals modulate multiple carrier frequencies and the different baseband signals are made to occupy non-overlapping bandwidths over a frequency spectrum, a *broadband* signal is obtained.

1.8 Architecture of a Mobile Telecommunication System

A simplified architecture of a mobile telecommunication system has been shown in Fig. 1.4. It has three main components: the core network, the radio access network, and the mobile phones.

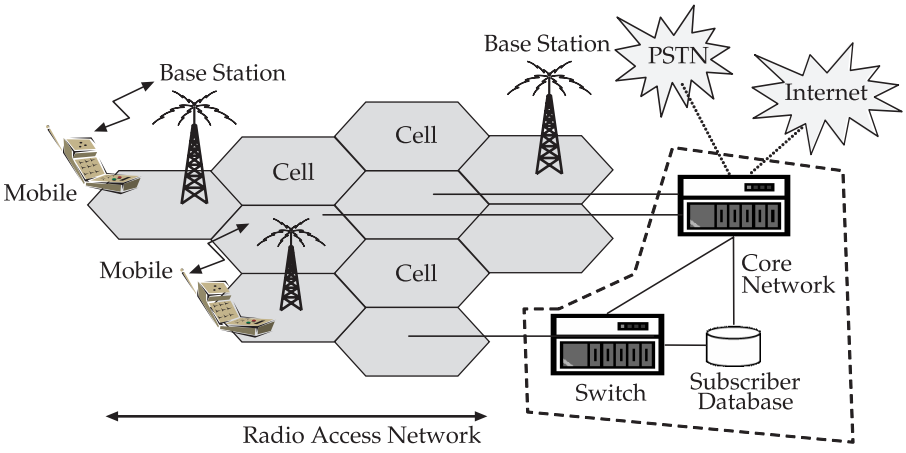


Figure 1.4 *Architecture of a mobile telecommunication system.*

Mobile handsets communicate over the radio access network. The radio access network is primarily composed of the base stations which communicate with the mobile phones using radio frequency electromagnetic waves. As shown in Fig. 1.4, the coverage area is decomposed into hexagonal cells. In each hexagonal cell, one base station is located. Two types of radio channels are usually involved in the communication between a base station and the cell phones: control channels and voice channels. Control channels typically use frequency shift keying (FSK) and are used for transferring control messages (data) between the mobile phone and the base station. Voice channels typically use frequency modulation (FM). A base station typically has two antennas of different characteristics. One antenna is used for receiving and the other for transmitting. The use of the two different types of antennas at the base station increases the ability of the base station to receive the radio signal from mobiles that use very low transmitter power levels. On the other hand, mobile handsets typically use the same antenna for both receiving and transmitting.

The core network interconnects the base stations, switches the mobile switching centre (MSC), and also provides an interface to other networks such as the traditional telephone network (PSTN) and the Internet. The interconnect used in the core network is required to provide high-speed connectivity. Therefore, usually fibre optic cables are used as the interconnect in the core network. But based on the terrain conditions, microwave communication is also sometimes used. This interconnection

in the core network must allow both voice and control information to be exchanged between the switching system and the base station. The MSC is connected to the landline telephone network to allow mobile telephones to be connected to standard landline telephones. The core network is responsible for transmitting voice calls, SMS (Short Message Service), etc. from one phone to another through switches. The core network also maintains a database that contains information about the subscribers and the information about billing.

1.9 Wireless Networking Standards

Standardization is very important to the computer networking domain since many protocols and devices need to interoperate in any practical networking solution. Further, there can be various vendors manufacturing the networking equipment. In the absence of appropriate standards, it would become difficult to interoperate the products manufactured by different vendors. Mainly, three international standardization bodies are responsible for formulating the networking standards: ITU, IEEE and ISO.

The IEEE (Institute of Electrical and Electronics Engineers) is a non-profit, technical professional association of members from over 150 countries. It acts as a standards body. Standards are very important in networking since multiple devices that are often heterogoneous and manufactured by different vendors need to communicate. The IEEE proposes standards for new technologies and maintains the old standards. The IEEE created the 802 group to help standardize the LAN technology. The 802.3 standard from this group defines the requirements that a product must meet for it to be considered "Ethernet". Wireless Ethernet is defined by 802.11. The 802.11 standard is further broken down into more specific certifications, such as 802.11a, 802.11b, and 802.11g. Each of these defines a different method for providing wireless Ethernet. Each protocol specifies various aspects of data transfer that distinguish them from the other protocols.

The 802.11 standards define rules for communication on wireless local area networks (WLANs). The popular 802.11 standards include 802.11a, 802.11b and 802.11g. The 802.11 was the original standard in this family, ratified in 1997. It defined WLANs that operated at 1–2 Mbps. This standard is obsolete today, but its extensions are being used extensively.

Each extension to the original 802.11 appends a unique letter to its name. For example, the standards 802.11a, 802.11b and 802.11g define different types of signal modulation and frequencies of operation as shown in Table 1.1.

The following IEEE 802.11 standards are being used for wireless local area networking:

- 802.11a: 54 Mbps standard, 5 GHz signalling (ratified 1999)
- 802.11b: 11 Mbps standard, 2.4 GHz signalling (1999)

TABLE 1.1 Wireless Networking Standards

Standard	*Data rate*	*Information*
IEEE 802.11	Up to 2 Mbps in the 2.4 GHz band	This specification has been extended into 802.11b.
IEEE 802.11a (Wi-Fi)	Up to 54 Mbps in the 5 GHz band	Products that adhere to this standard are considered "Wi-Fi Certified." Eight available channels. Less potential for RF interference than 802.11b and 802.11g. Better than 802.11b at supporting multimedia voice, video and large-image applications in densely populated user environments. Relatively shorter range than 802.11b. Not interoperable with 802.11b.
IEEE 802.11b (Wi-Fi)	Up to 11 Mbps in the 2.4 GHz band	Products that adhere to this standard are considered "Wi-Fi Certified." Not interoperable with 802.11a. Requires fewer access points than 802.11a for coverage of large areas. Offers high-speed access to data at up to 300 feet from base station. 14 channels available in the 2.4 GHz band (only 11 of which can be used in the U.S. due to FCC regulations) with only three non-overlapping channels.
IEEE 802.11g (Wi-Fi)	Up to 54 Mbps in the 2.4 GHz band	Products that adhere to this standard are considered "Wi-Fi Certified." May replace 802.11b. Improved security enhancements over 802.11. Compatible with 802.11b. 14 channels available in the 2.4 GHz band (only 11 of which can be used in the U.S. due to FCC regulations) with only three non-overlapping channels.
IEEE 802.16 (WiMAX)	Specifies WiMAX in the 10 to 66 GHz range	Commonly referred to as WiMAX or less commonly as WirelessMAN or the Air Interface Standard, IEEE 802.16 is a specification for fixed broadband wireless metropolitan access networks (MANs)
IEEE 802.16a (WiMAX)	Added support for the 2 to 11 GHz range.	Commonly referred to as WiMAX or less commonly as WirelessMAN or the Air Interface Standard, IEEE 802.16 is a specification for fixed broadband wireless metropolitan access networks (MANs)
Bluetooth	Up to 2 Mbps in the 2.45 GHz band	No native support for IP, so it does not support TCP/IP and wireless LAN applications well. Not originally created to support wireless LANs. Best suited for connecting PDAs, cell phones and PCs in short intervals.

(*Contd.*)

TABLE 1.1 Wireless Networking Standards (Contd.)

Standard	*Data rate*	*Information*
HiperLAN/1 (Europe)	Up to 20 Mbps in the 5 GHz band	Only in Europe. HiperLAN is totally ad-hoc, requiring no configuration and no central controller. Does not provide real isochronous services. Relatively expensive to operate and maintain. No guarantee of bandwidth.
HiperLAN/2 (Europe)	Up to 54 Mbps in the 5 GHz band	Only in Europe. Designed to carry ATM cells, IP packets, Firewire packets (IEEE 1394) and digital voice (from cellular phones). Better quality of service than HiperLAN/1 and guarantees bandwidth.
Open Air	Pre-802.11 protocol, using frequency hopping and 0.8 and 1.6 Mbps bit rate	OpenAir is the proprietary protocol from Proxim. All OpenAir products are based on Proxim's module.

- 802.11c: Operation of bridge connections (moved to 802.1D)
- 802.11d: Worldwide compliance with regulations for use of wireless signal spectrum (2001)
- 802.11e: Quality of Service (QoS) support (not yet ratified)
- 802.11F: Inter-Access Point Protocol recommendation for communication between access points to support roaming clients (2003)
- 802.11g: 54 Mbps standard, 2.4 GHz signalling (2003)
- 802.11h: Enhanced version of 802.11a to support European regulatory requirements (2003)
- 802.11i: Security improvements for the 802.11 family (2004)
- 802.11j: Enhancements to 5 GHz signalling to support the Japan regulatory requirements (2004)
- 802.11k: WLAN system management (in progress)
- 802.11l: Skipped to avoid confusion with 802.11i
- 802.11m: Maintenance of 802.11 family documentation
- 802.11n: 100+ Mbps standard improvements over 802.11g (in progress)
- 802.11o: Skipped
- 802.11p: Wireless Access for the Vehicular Environment
- 802.11q: Skipped
- 802.11r: Fast roaming support via Basic Service Set transitions
- 802.11s: ESS mesh networking for access points
- 802.11T: Wireless Performance Prediction—recommendation for testing standards and metrics
- 802.11u: Internetworking with 3G/cellular and other forms of external networks

- 802.11v: Wireless network management/device configuration
- 802.11w: Protected Management Frames security enhancement
- 802.11x: Skipped (generic name for the 802.11 family)
- 802.11y: Contention Based Protocol for interference avoidance.

1.10 Wireless Local Area Networks (WLANs)

A brief history of evolution of wireless networks is given in Box 1.2. Today, Wireless Local Area Networks (WLANs) provide connectivity between computers over short distances using the wireless medium. Typical indoor applications of WLANs may be in educational institutes, office buildings and factories where the required coverage distances are usually restricted to less than a few hundred feet. In the absence of obstructions and with the use of suitable antennas, ranges of up to a few kilometres can be obtained. Wireless networks are especially useful when it is impossible or prohibitively expensive to carry out wiring within or across buildings, or when only temporary access is needed between computers. WLANs are useful to provide connectivity among portable computers. As an example, consider an educational institute where the students may carry their own laptop and use those in classrooms, library or lounge as and when they require.

BOX 1.2 History of wireless networks

The concept of communication without wires is not new. Smoke signals and tribal drums were used to communicate over short distances without cords or wires. Eventually, communications over long distances became possible through wires. Claude Chappe invented the telegraph in 1792 and Alexander Graham Bell first sent voice transmissions over wire in 1876. However in 1894, near Bologna, Italy, wireless communication was born. Gulielmo Marconi tapped out a message, causing a bell to ring on the other end of the room without using any wire. Scientists began searching for ways to broadcast speech using Marconi's wireless. In 1906, Reginald Fessenden did it by using amplitude modulation. In 1935, the American engineer Edwin Armstrong introduced FM (frequency modulation) radio waves, which used less power and achieved reception of higher quality signals. On October 13, 1983, the first call on a commercial cellular system was made in Chicago.

Most commercial WLAN products use a number of technologies in accordance with the international standards specified by the Institute of Electrical and Electronic Engineers (IEEE). These WLANs operate with maximum data rates of up to 54 Mbps, which are much lower than the data rates that can be achieved with the wired Ethernet connections. WLANs almost universally use Internet Protocols (such as TCP/IP) for communication between computers. These protocols are basically a set of standard rules to facilitate communication between computers. The

HIPERLAN standard provides for WLAN operation with bit rate up to 20 Mbps. This can be observed from Table 1.1. WLANs operate in specific frequency ranges depending on the availability of spectrum in specific countries.

A typical WLAN is comprised of a few important components. These components are shown in Fig. 1.5 and are briefly discussed below.

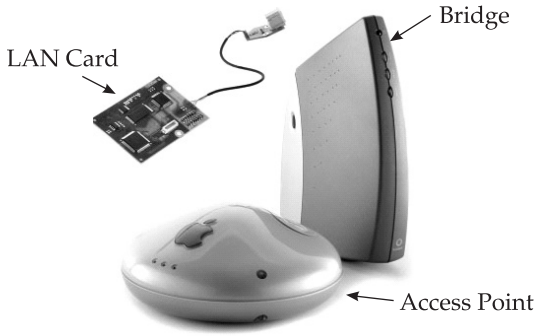


Figure 1.5 *Wireless LAN card, access point and bridge.*

Access point: It is a radio receiver/transmitter (also called transceiver) that connects to the wired network. These are typically mounted on the roofs at different locations of a building. You can spot them if you carefully observe the roof of a building having wireless LAN. The transceiver exchanges signals with the wireless LAN card in desktop or notebook PCs. A single access point can support a small group of users. It is connected to a wired network through cables and provides the connectivity between wireless devices and the wired network.

Wireless LAN cards: End-users access the WLAN through WLAN adapters (wireless network interface cards) in their hand-helds. The LAN card used to be mounted on the motherboard of a computer. Now, it is inbuilt into the motherboards.

Bridge: It is used for connecting two LANs that may be in two different buildings or on two separate floors within the same building.

1.10.1 Wireless LAN Architecture

One access point can provide support to a small group of users and can perform within a range up to a few hundreds of feet. The access point (or the antenna attached to the access point) is usually mounted on roof tops but may be mounted elsewhere, if it is practical to do so, as long as the desired radio coverage is obtained. End users access the wireless LAN through wireless-LAN adapters, which are integrated within the hand-held computers. Wireless LAN adapters provide an interface between the client's network operating system and the air waves via an antenna.

One of the main concerns of users of wireless LANs is the apparent reduction in privacy and security. To address this issue, WLANs uses multiple levels of security to prevent unauthorized access to network resources. Figure 1.6 shows the architecture of an infrastructure-based IEEE 802.11 network.

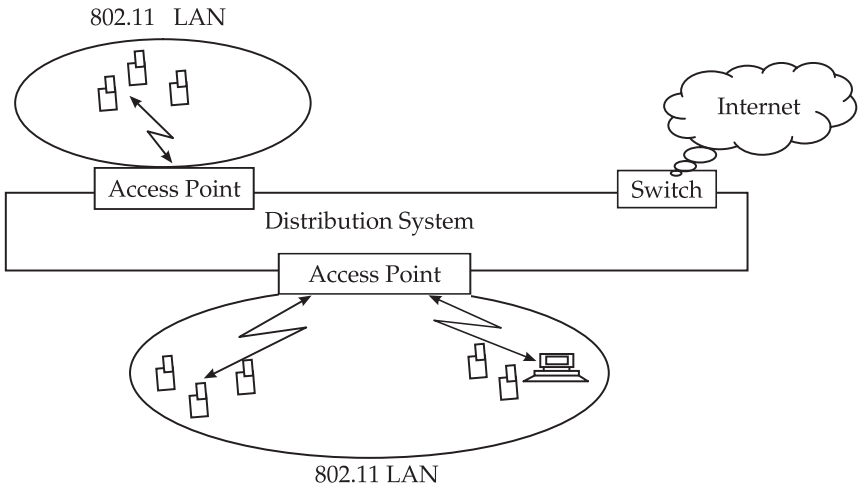


Figure 1.6 *Architecture of an infrastructure-based IEEE 802.11 network.*

In Fig. 1.6, observe that the access points are connected through a distribution system which usually uses a fibre optic cable. As shown, mobile nodes are connected to access points. The 802.11 standard stipulates that the distribution system may be of any technology, such as Ethernet, token ring, or any other network type. The majority of actual installations, however, utilize Ethernet (802.3). An 802.3-based distribution system (also referred to as the "wired backbone") consists of switches or hubs that tie together users (PCs and access points) equipped with 802.3 network interface cards (NICs). The switch or hub is somewhat analogous to an 802.11 access point. The main difference is that a hub or switch provides connection over a physical medium and an access point provides wireless connectivity. In the example of Fig. 1.6, two 802.11 LANs are connected via a distribution system. A distribution system can also be used to increase network coverage through roaming between cells. The distribution system also provides connectivity to the Internet through a switch.

1.10.2 Applications of Wireless LANs

Applications of the wireless LAN are numerous. In the following, we discuss only the four important application scenarios of WLANs.

1. *Campus wireless LANs:* Several organizations, hotels, retail outlets, warehouses, factories, research centres, and educational institutions are among many others who are recognizing the value of flexibility and connectivity provided by wireless LANs. Users with laptops, PDAs, palmtops can have access to the organization's intranet and to the Internet from any location in the campus. An effective way to make the network-based services available to them is by creating a wireless network.
2. *Streamlining inventory management:* Stock control and inventory management are the important activities in almost every manufacturing organization and retail outlet. Traditionally, stockkeeping was done manually, with the stockkeeper going around the godown and noting down the stock levels, counting various item and tallying them manually for inventory control. Human errors can lead to stockkeeping mistakes resulting in various types of financial losses. With wireless LAN, devices such as hand-held scanners, keypads and bar code readers can be linked to database applications and printers. Instead of being a laborious time consuming task, inventory management becomes automatic, making stock and inventory control more manageable.
3. *Providing LAN facilities in difficult-to-wire areas:* Wireless LANs can coexist with wired LANs and help in extending LAN connectivity to locations where providing wired connections is difficult.
4. *WLAN connectivity to geographically dispersed computers:* Networks of wireless LANs are being used for providing high bandwidth links to connect different branches of an organization spread over a city. These links between WLANs are being set up as Metropolitan Area Networks (MANs) for e-governance or for linking various government organizations, information kiosks, public utility places, and various branches of a private organization, etc.

1.10.3 Advantages of Wireless LANs over Wired LANs

With wireless LANs, users can access shared information without looking for a place to plug in their network cable, and network managers can set up networks. Wireless LANs offer the following advantages over the traditional wired networks:

1. *Mobility:* Wireless LAN systems can help users get information at any place in their organization. Such types of support and service are not possible with the wired networks.
2. *Simplicity and speedy deployment:* The installation procedure of a wireless LAN system is simple because it eliminates the need to run wires through walls. A wireless LAN facility can be set up in an area in a matter of few hours.

3. *Flexibility:* Wireless technology allows the network to be accessible where wiring is difficult to lay. Consider an airport lounge, the passengers can connect their computers to the network just sitting at their seats.
4. *Cost effectiveness:* While the initial investment required for wireless LAN hardware can be higher than the cost of wired LAN hardware, but in the long run the cost benefits of WLAN are reasonable because of the environment which often requires frequent movements and dynamic changes.

1.11 Bluetooth Technology

The Bluetooth technology has become a popular means to implement a Personal Area Network (PAN). A PAN helps to interconnect a set of computerized devices that an individual person might require. For example, a PAN makes it possible to network various appliances used in daily life such as a microwave oven, fridge, air conditioner, mobile phones, etc. This connectivity makes some meaningful applications to become possible. For example, the fridge might automatically send the list of items that are out of stock to the mobile, which can remind a housewife when shopping in the mall. In the present standard, using wireless PAN, a dynamic group of less than 255 devices can be made to communicate within an area of less than 10 metres in diameter.

BOX 1.3 How was bluetooth technology named!

The king of Denmark, Herald, was nicknamed Bluetooth since he was fond of blueberries and as per legend, he ate so many of them that his teeth got stained and turned blue. He was credited to have united Denmark and Norway regions. The Bluetooth technology was named after the king, since it unifies (that is, provides seamless connectivity) multiple devices.

A prominent advantage of Bluetooth is that it can be used to get rid of the mesh of wires that are required for interconnecting various devices that are positioned near each other. For example, the mouse, the camera, the printer, etc. can all communicate with the computer using Bluetooth connectivity, making it possible to get rid of the mesh of wires. Such a Bluetooth connectivity among a set of devices is called a *piconet*. A piconet is essentially a very small (pico means very small) network as illustrated in Fig. 1.7.

A Bluetooth piconet is based on a master-slave communication architecture. One of the computer-enabled devices is designated as the master and the other devices become the slaves. In a piconet, one master device can interconnect with up to seven active slave devices using

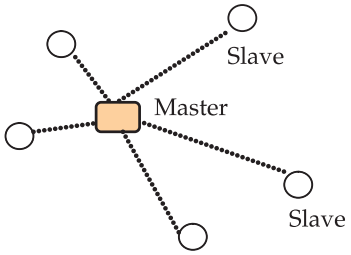


Figure 1.7 *A Bluetooth piconet.*

Bluetooth to form an ad hoc network. In Bluetooth communication, the slave units only respond to commands from the master. This allows the Bluetooth MAC to be simple, efficient, and non-contention based. In a piconet, data transfer rates of up to 2.1 Mbits are possible. A *scatternet* is formed when an ad hoc network of more than one piconets is formed.

1.11.1 Protocol Stack of Bluetooth

Bluetooth protocol stack makes possible the communication of both data and control among many devices in a PAN. The protocol stack is schematically presented in Fig. 1.8. We first give a brief overview of the protocol stack, and then give a more detailed description of a few specific protocol layers.

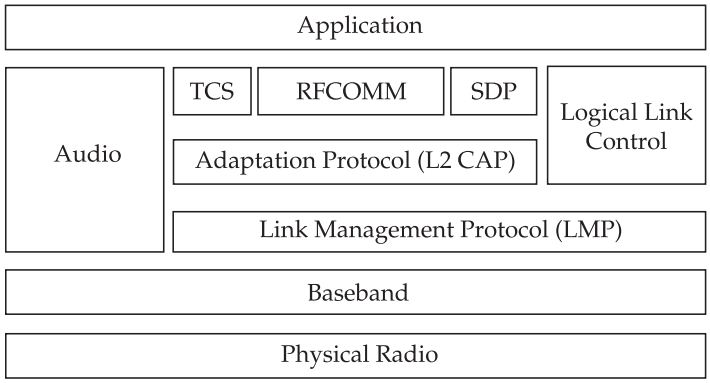


Figure 1.8 *Bluetooth protocol stack*.

The lowermost layer is called the Radio Frequency (RF) layer. Radio modem is specified in this layer. This link controls the packet but the bit levels are specified by the baseband layer. The Link Management Protocol (LMP) is capable of configuring links with other devices. It provides power modes, traffic scheduling, packet format, authentication and encryption. Logical Link Control and Adaptation Protocol (L2CAP) provide connection-oriented and connectionless data services to the upper layers of the protocol

stack. Control information may also be exchanged through it. L2CAP may include services such as segmentation and re-assembly of data packets. The Service Discovery Protocol (SDP) enables two or more Bluetooth devices to support a particular service. In the following, we briefly describe the functionalities of the different layers of the Bluetooth protocol stack.

Radio layer: It defines the requirements on the Bluetooth transceiver device that provides the required wireless support for master-slave communications. It specifies the transceiver characteristics such as the specific modulation technique used (GFSK, i.e. Gaussian Frequency Shift Keying), interference performance, radio frequency tolerance, etc. The Bluetooth radio is a low-power transceiver with a range of up to 10 metres and uses spread spectrum transmission by frequency hopping.

Baseband: The Baseband is the physical layer of the Bluetooth. As can be seen in Fig. 1.8, this layer lies on top of the Bluetooth radio layer in the Bluetooth stack. The functionalities of this layer include link establishment and power control. It also defines the packet format and timing issues. The baseband protocol implementation is called a Link Controller (LC).

Link Management Protocol (LMP): LMP is a data link layer protocol and is responsible for link setup between the Bluetooth devices. The specific responsibilities that it carries out include link setup, authentication, and link configuration. Two types of links are possible: Asynchronous Connectionless Links (ACL) and Synchronous Connection-Oriented (SCO). The ACL link is used for transmitting the packet-switched data received from L2CAP. The SCO link is used for real-time data involved in voice telephony applications.

Logical Link Control and Adaptation Protocol (L2CAP): This layer provides connection-oriented and connectionless data services to the upper layers of the protocol stock. Using the services of this layer, the higher level protocols can transmit and receive L2CAP messages that can be up to 64 kilobytes in length. L2CAP forms the Bluetooth data packets through message segmentation while transmitting and also forms messages from the received Bluetooth data packets.

Services Discovery Protocol (SDP): Each Bluetooth device maintains an SDP server application that keeps track of the services available on that device. By using this protocol, a mobile application can discover which services are available on which device and can also determine the characteristics of those available services. Thus, it becomes possible to establish connection between two or more Bluetooth devices for specific services. By using SDP, mobile applications can get information such as (i) device information, (ii) services, and (iii) service characteristics.

Radio Frequency Communication (RFCOMM): RFCOMM is a simple set of transport protocols that are built on top of L2CAP, and helps realize

TCP/IP connectivity. It also provides an emulated RS-232 serial port with other Bluetooth devices, thus simplifying connectivity with many simple and small devices without using wires.

Telephony Control Protocol Specification (TCS): TCS defines the call control signals for establishing a voice connection between Bluetooth devices. This protocol also considers the mobility related issues of hand-held devices. In order to avoid confusion with the popular transport layer protocol, i.e., the Transmission Control Protocol (TCP), this telephone control protocol specification is usually denoted by the abbreviation TCS.

SUMMARY

We first reviewed some of the basic concepts in computer networking to provide the background necessary to understand the rest of this text. Using traditional networking, people were unable to access information on the networks or run remote application while they were moving. We pointed out that wireless networking provides the flexibility of network access to the mobile user, and over the last two decades or so it has proved to be a powerful and indispensable technique for a general user. However, several shortcomings such as the overall security of information, disconnections, and degradation of signals as a user moves, need to be adequately addressed. We briefly reviewed the important features of WLANs, a popular means adopted today to provide wireless connectivity to users.

While the advantages of wireless networking are numerous, possibly the most severe handicap of wireless networking is the security of information. The security issues arise from the basic broadcast nature of wireless communication. Several techniques have been implemented to address the security issues, which we will discuss in the subsequent chapters.

FURTHER READINGS

Cisco Systems, Inc., "Wireless Technologies." *Internetworking Technologies Handbook* (http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/wireless.pdf)

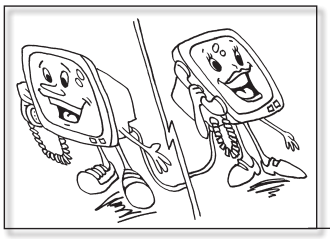
Joel Conover, "First Things First—Top 10 Things to Know About Wireless," *Networking Computing*, July 2000.

Rajib Mall, *Real Time Systems: Theory and Practice*, Pearson Education, New Delhi, 2008.

Rhoton, John "The Wireless Internet Explained", *Digital Press*, 2001.

EXERCISES

1. What is the difference between a data network and a voice network? What are their relative advantages?
2. What types of communication services are usually offered on data and voice networks?
3. What are the main difficulties that would be experienced if digitized voice signals are transmitted over a data network? How can these difficulties be overcome?
4. What is Controller Area Network (CAN)? How is it different from a LAN? What are the important applications of a CAN?
5. Explain why there is a bound on the length and the number of nodes that can be connected to a LAN.
6. Determine the minimum packet size for a CSMA/CD network operating at 1 Gbps and having a length of 100 metres. Clearly show your assumptions and calculations.
7. What is a transceiver? What is the role of a transceiver in a wireless communication network?
8. Explain the different hardware components that are used in a wireless network.
9. What is wireless LAN (WLAN)? Explain the basic wireless LAN architecture using a suitable schematic diagram.
10. What are the important advantages of a wireless LAN (WLAN) over a traditional wired LAN?
11. Discuss the architecture of a mobile telecommunication network using a suitable schematic diagram.
12. Briefly explain how a call can be set up between two mobile phones.
13. Why are standards necessary in networking? Briefly explain any LAN standard.
14. What do you understand by signal modulation? Briefly explain how it is achieved. Why is it necessary to modulate a baseband signal on a carrier signal, before transmitting it?
15. What is the difference between a baseband signal and a broadband signal? Using a suitable schematic diagram, explain how a broadband signal is obtained.
16. What is Bluetooth? How is Bluetooth useful in mobile computing? Describe the protocol stack of Bluetooth.
17. What is a personal area network? Explain the role of Bluetooth in personal area networking using a suitable example.
18. What are piconets and scatternets? Mention some important applications of these networks.
19. Discuss the architecture of a mobile telecommunication system using a suitable schematic diagram.



2

Introduction to Mobile Computing and Wireless Networking

A few important breakthroughs achieved over the last few decades have revolutionized the way people use computers. First, about two decades ago the advancements made in the field of miniaturization of electronic circuits made it possible to pack powerful processing units and significant memory into portable laptops. More recently, the computing elements have become more powerful and at the same time shrunk to fit into palmtops that people can carry effortlessly. In another development, high speed data communication facility has been made available to portable computing platforms, largely through advancements in the areas of computer communication and wireless networking technologies. These two developments have contributed to the formation of the discipline of mobile computing. This has made it possible for the users to take advantage of a host of innovative services available in these platforms and use them ubiquitously. As a result, people from all walks of life have started to carry mobile handsets with them wherever they go and are able to perform meaningful computations with them.

2.1 What Is Mobile Computing?

Mobile computing (sometimes called ubiquitous computing and also at times called nomadic computing) is widely described as the ability to compute remotely while on the move. This is a new and fast emerging discipline that has made it possible for people to access information from anywhere and at anytime. We can also view *mobile computing* as encompassing two separate and distinct concepts: mobility and computing. Computing denotes the capability to automatically carry out certain processing related to service invocations on a remote computer. Mobility, on the other hand, provides the capability to change location while communicating to invoke

computing services at some remote computers. The main advantage of this type of mobile computing is the tremendous flexibility it provides to the users. The user need not be tethered to the chair in front of his desktop, but can move locally or even to far away places and at the same time achieve what used to be performed while sitting in front of a desktop.

2.2 Mobile Computing vs. Wireless Networking

We must distinguish between mobile computing and wireless networking. These two terms are not synonymous. While mobile computing essentially denotes accessing information and remote computational services while on the move, wireless networking provides the basic communication infrastructure necessary to make this possible. Thus, we can say that mobile computing is based on wireless networking and helps one to invoke computing services on remote servers while on the move: be it be office, home, conference, hotel, and so on.

It should be clear that wireless networking is an important ingredient of mobile computing, but forms only one of the necessary ingredients of mobile computing. Mobile computing also requires the applications themselves—their design and development, and the hardware at the client and server sides. In fact, we can say that mobile computing subsumes the area of wireless networking. Consequently, to be able to understand the subtle issues associated with mobile computing, in addition to studying the different aspects of mobile computing applications, their design and development, we need to have a good knowledge of the basics of wireless communications technologies. Wireless networking is increasingly replacing traditional networks because of the low setup time and low initial investment required to set up the wireless network. As we discuss later in this chapter, wireless networks appear in various forms such as WLANs (Wireless LANs), mobile cellular networks, personal area networks (PANs), and ad hoc networks, etc.

Wireless networks can be classified into two basic types. One is an extension of wired networks. It uses fixed infrastructures such as base stations to provide essentially single hop wireless communication with a wired network as illustrated in Fig. 2.1 or a two-hop wireless cellular communication with another mobile as explained earlier in Fig. 1.4.

The other type of wireless network is an ad hoc network. An ad hoc network does not use any fixed infrastructure and is based on multi-hop wireless communication as shown in Fig. 2.2.

One popular example of a fixed infrastructure wireless network is a Wireless LAN (WLAN) that implements the IEEE 802.11 protocol. Observe from Fig. 2.1 that only the last hop is through the wireless medium. An access point (AP) provides the last hop connectivity of the mobile nodes to a wired network. All communication goes through APs which perform

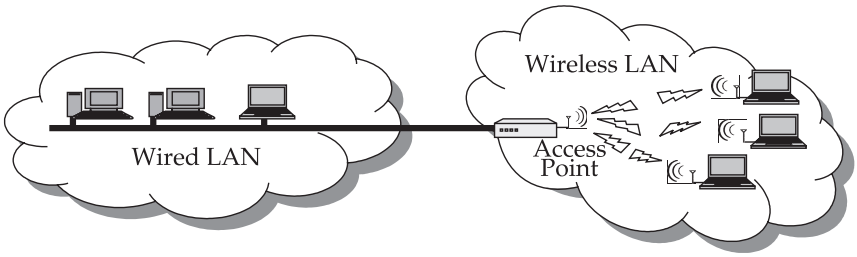


Figure 2.1 *Wireless network based on fixed infrastructures.*

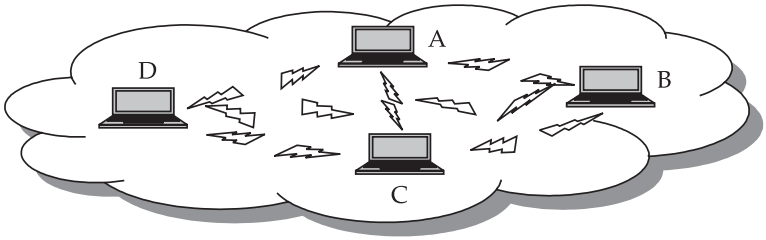


Figure 2.2 *Wireless network having no fixed infrastructures.*

bridging† between the wireless and the wired mediums. A station must be recognized by an AP to be able to connect to the network. The AP may require authentication and this in turn is used as the basic means to keep out the unauthorized users. In an infrastructureless network, the communication between hosts occurs directly or via a few intermediate nodes that form the hops. For example, station A in Fig. 2.2 can communicate with station C using either the hops A–B, B–C or A–D, D–C.

A recent development in this context, is wireless networking of various types of devices using the Bluetooth technology. As already discussed in Chapter 1, the Bluetooth technology can also be used to establish direct wireless connection of cell phones with devices such as printers, cameras, scanners, laptop and desk computers. Bluetooth is gradually replacing cables and infrared as the dominant way of exchanging information between devices. One of the objectives of the Bluetooth technology is to enable users to easily connect to a wide range of personal computing and telecommunication devices, without the need to buy, carry, or lay out cables. In fact, the Bluetooth technology enables setting up of personal area networks (PANs) known as piconets and ad hoc networks known as scatternets. It provides opportunities for rapid deployment of ad hoc connections, and the possibility of automatic, transparent connections between devices. It promises to eliminate the need to purchase additional or proprietary cabling and configuration exercises needed to connect the individual devices.

† A network bridge connects multiple network segments at the data link layer (Layer 2) of the OSI reference model.

An ad hoc network is also known as a Mobile Ad hoc Network(MANET). It is a collection of mobile nodes that form a network on the fly without requiring the support of any fixed infrastructure. Wireless sensor networks are a special type of wireless ad hoc networks. Mobile ad hoc networks and sensor networks are discussed in more detail in Chapters 7 and 8 respectively.

2.3 Mobile Computing Applications

Mobile computing technology makes it possible for people to send or extract information while on the move. For example, a stock broker travelling in a car may wish to issue stock transaction orders from a mobile phone or to receive share price quotations. As can be guessed, ease of deployment and scalability are two important positive points in favour of data transmissions over the wireless medium. But, it is not without some shortcomings. When data is being transmitted on air, all the wireless devices present in the transmission range can receive the data. This, therefore, opens up very difficult security issues that must be overcome to ensure privacy of data.

2.4 Characteristics of Mobile Computing

A computing environment is said to be "mobile", when either the sender or the receiver of information can be on the move while transmitting or receiving information. The following are some of the important characteristics of a mobile computing environment.

Ubiquity: The dictionary meaning of ubiquity is *present everywhere*. In the context of mobile computing, ubiquity means the ability of a user to perform computations from anywhere and at anytime. For example, a business executive can receive business notifications and issue business transactions as long he is in the wireless coverage area.

Location awareness: A hand-held device equipped with global positioning system (GPS) can transparently provide information about the current location of a user to a tracking station. Many applications, ranging from strategic to personalized services, require or get value additions by location-based services For example, a person travelling by road in a car, may need to find out a car maintenance service that may be available nearby. He can easily locate such a service through mobile computing where an application may show the nearby maintenance shop. A few other example applications include traffic control, fleet management and emergency services. In a traffic control application, the density of traffic along various roads can be dynamically monitored, and traffic can be directed appropriately to

reduce congestions. In a fleet management application, the manager of a transport company can have up-to-date information regarding the position of its fleet of vehicles, thus enabling him to plan accurately and provide accurate information to customers regarding the state of their consignments. Location awareness can also make emergency services more effective by automatically directing the emergency service vehicles to the site of the call.

Adaptation: Adaptation in the context of mobile computing implies the ability of a system to adjust to bandwidth fluctuation without inconveniencing the user. In a mobile computing environment, adaptation is crucial because of intermittent disconnections and bandwidth fluctuations that can arise due to a number of factors such as handoff, obstacles, environmental noise, etc.

Broadcast: Due to the broadcast nature of the underlying communication network of a mobile computing environment, efficient delivery of data can be made simultaneously to hundreds of mobile users. For example, all users at a specific location, such as those near a railway station, may be sent advertising information by a taxi service operator.

Personalization: Services in a mobile environment can be easily personalized according to a user's profile. This is required to let the users easily avail information with their hand-held devices. For example, a mobile user may need only a certain type of information from specific sources. This can be easily done through personalization.

2.5 Structure of Mobile Computing Application

A mobile computing application is usually structured in terms of the functionalities implemented. The simple three-tier structure of a mobile computing application is depicted in Fig. 2.3. Figure 2.4 shows a specific scenario of the types of functionalities provided by each tier. As shown in these figures, the three tiers are named presentation tier, application tier and data tier.

We now briefly explain the roles of the three tiers of a mobile computing application.

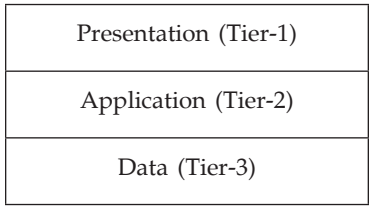


Figure 2.3 *The three tier structure of a mobile computing application.*

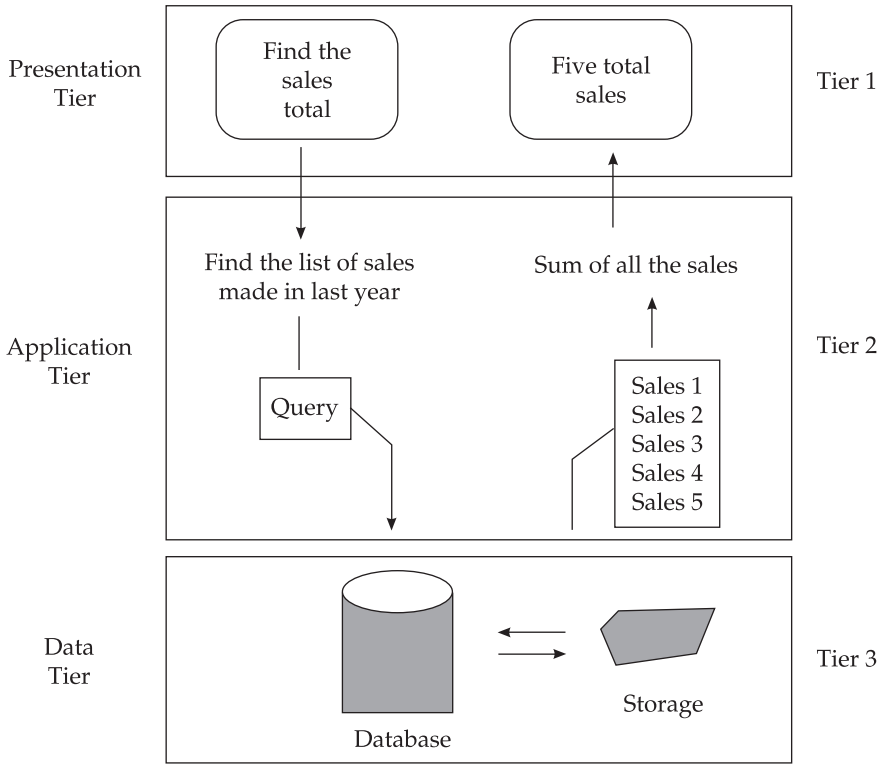


Figure 2.4 *Functionalities provided by each tier structure of a mobile computing application.*

Presentation tier

The topmost level of a mobile computing application concerns the user interface. A good user interface facilitates the users to issue requests and to present the results to the them meaningfully. Obviously, the programs at this layer run on the client's computer. This layer usually includes web browsers and customized client programs for dissemination of information and for collection of data from the user.

Application tier

This layer has the vital responsibility of making logical decisions and performing calculations. It also moves and processes data between the presentation and data layers. We can consider the middle tier to be like an "engine" of an automobile. It performs the processing of user input, obtaining information and then making decisions. This layer is implemented using technology like Java, .NET services, cold fusion, etc. The implementation of this layer and the functionality provided by this layer should be database independent. This layer of functionalities is usually implemented on a fixed server.

Data tier

The data tier is responsible for providing the basic facilities of data storage, access, and manipulation. Often this layer contains a database. The information is stored and retrieved from this database. But, when only small amounts of data need to be stored, a file system can be used. This layer is also implemented on a fixed server.

2.6 Cellular Mobile Communication

In a cellular mobile system, the area of coverage is split into cells as shown in Fig. 2.5. Even though the cells have been shown to be non-overlapping hexagons for simplicity, but in reality cell shapes are irregular and do overlap to some extent. A base station (BS) is located at the centre of each cell. The BS in a cell receives communications from all mobile handsets in the cell and forwards the data to the appropriate handset. Thus, a base station keeps track of the calls of all handsets in its cell. When a mobile handset while still continuing a call, moves to another cell, the BS "hands-off" the call to the BS in the new cell. When a cell covers a crowded area having too many users, then the users can experience frequent call drops. To overcome this problem, such cells are split into smaller cells.

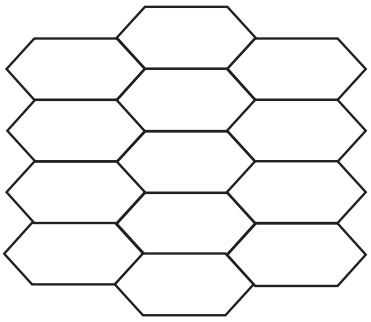


Figure 2.5 *Cell structure in a cellular mobile communication system.*

Initially the focus of cellular mobile communication was voice communication. But today cellular phones provide many services based on data communication too. These include electronic mail, Internet access and running a variety of mobile applications. The term mobile communication has a much wider connotation than that of cellular communication, and includes wireless LANs and ad hoc networks. However, due to the overwhelming popularity of mobile phones, cellular communication and mobile communication are at times used interchangeably.

2.6.1 Generations of Cellular Communication Technologies

Mobile communication technology has advanced at a very rapid pace over the last five decades. The gradual technology improvements over the last four decades can be roughly demarcated into four generations. Each generation essentially provides higher data rate and additional capabilities, as shown schematically in Fig. 2.6. This figure does not show the data rates of technologies before GSM, since these were analog techniques that did not support the data communications facility. The fourth generation (4G) of technology provides a substantial order of magnitude improvements in data speeds, but is not yet widely implemented. The important characteristics of the various generations of cellular mobile systems have been summarized in Table 2.1. As can be seen from the table, each passing generation of mobile cellular system brought about significant advancements to the technology, causing the quality of the services to improve and the number of service offerings to increase, and at the same time the cost to the customer to drop drastically. We briefly discuss these different generations of mobile cellular communication systems in the following.

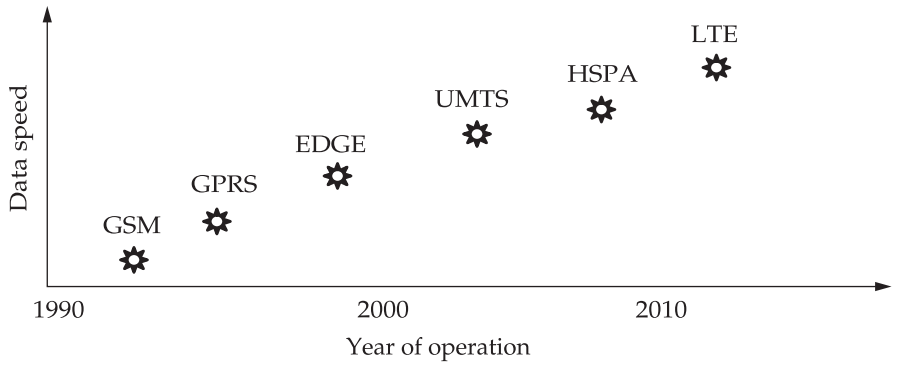


Figure 2.6 *Summary of mobile technology advancements.*

First generation

The first generation (1G) cellular system was designed in the late 1960s, but was commercially deployed in the early 1980s. The first commercial 1G system in the United States was known as Advanced Mobile Phone System (AMPS). It became operational in 1982 and supported only voice calls. This was a completely analog system. In an analog system, analog signals are transmitted by modulating them on a higher frequency carrier signal, without first converting the signal into digital form through quantization. In a completely analog system, it is difficult to support SMS and other data services. Also, the signals from different users cannot be intermixed on the same channel, and have to be transmitted in clearly separated channels.

TABLE 2.1 A Summary of the Important Characteristics of the Different Generations of Cellular Wireless Communication Systems

Mobile gen.	*Features*	*Standards*	*Data speed*
1G	Analog transmissions, primarily supported voice communications.	NMT, AMPS, TACS	600–1200 bps
2G	Digital transmissions, improved performance by letting multiple users share a single channel.	GSM	9.6 kbps
2.5G	Enhanced multimedia and streaming video, web browsing.	GPRS	28 kbps or higher
3G	Enhanced multimedia and streaming video capabilities.	UMTS, HSPDA, EDGE, W-CDMA	384 kbps or higher

In the 1G system, the available frequency spectrum was split into a number of sub-bands (or channels), each of which was used by a different caller. These systems typically allocated one 25 MHz frequency band for the signals to be sent from the base station to the handset (incoming signal), and a second different 25 MHz band for the signal transmitted from the handset to the base station (outgoing signal). Figure 2.7 shows a frequency band split into five sub-bands (channels). Though for simplicity, we have shown the different channels to be adjacent to each other, each channel was separated from the adjacent channels by a spacing of about 30 kHz. This was called a guard band. The use of guard bands was one of the causes of inefficient spectrum usage and resulted in the reduced number of simultaneous calls that could be supported. This problem was overcome in the subsequent generations of technologies.

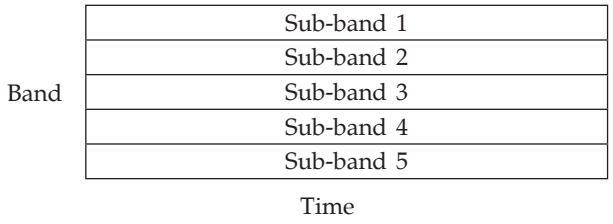


Figure 2.7 1G *frequency band split into five sub-bands.*

Different 1G standards were used in different countries:

- AMPS (Advanced Mobile Phone System) in the USA
- NMT 450 (Nordic Mobile Telephone) in various European countries
- TACS (Total Access Communications System) in the UK

The 1G systems were of multiple access type, since once a caller hanged up, another caller could use the same frequency. For this reason, the 1G technology was also called Frequency Division Multiple Access (FDMA). It was possible to reuse the same frequencies in the non-adjacent cells, because the transmitter power output was restricted. For example, the cells shown shaded in Fig. 2.8 could use the same set of frequencies.

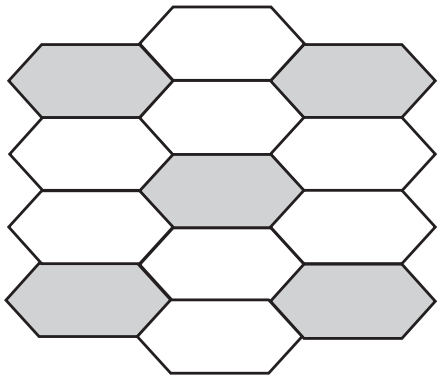


Figure 2.8 *Cell structure and frequency allocation of 1G systems.*

When a caller crossed a cell boundary, the channel being used might not be made available in the new cell as it might already be in use in some other cell. During handoff, a different channel was possible to be allocated in this case, otherwise the call got dropped if none were available. Beside the number of callers that could be accommodated being low, the voice quality was poor due to analog transmission. Also, it provided no security at all, since any one could hear a call by tuning into a channel.

Second generation

As already pointed out, the 1G technology had many disadvantages. The major drawback was the small number of simultaneous calls that could be made and the high risk of call drops during handoffs. Calls in 1G were expensive because of the inherent inefficient usage of the bandwidth spectrum and hence very few could afford to use a cell phone. Further, the 1G networks were not capable of providing several useful services such as caller identity and SMS. The disadvantages of 1G systems were overcome by the second generation (2G) cellular systems.

The 2G systems encoded voice and other information digitally before transmitting them. Digital transmission has many advantages over analog transmissions. These include noise immunity and better bandwidth utilizations. The 2G system offered significant advancements in the evolution of the mobile cellular technologies. Hence the 2G technology rapidly replaced the 1G technology because of the drastic reductions in the cost of phone calls and availability of a wider range of services coupled with substantial improvements in the quality of services. Also, SMSs became possible.

However, we must remember that the 2G technology is in many respects an extension of the 1G system and many of the principles involved in a 1G system also apply to 2G. For example, they both use the same cell structure. However, there are many differences. For example, they use different signal modulation techniques. 2G uses CDMA (Code Division Multiple Access) and TDMA (Time Division Multiple Access) as channel access technology, while 1G used FDMA.

The 2G mobile system deployment started in the 1990s, and two competing standards existed. In North America, the IS-95 standard was adopted which used Code Division Multiple Access (CDMA) and could multiplex up to 64 calls per channel in the 800 MHz band. In Europe and elsewhere, operators adopted the Global System for Mobile Communication (GSM) standard, which used Time Division Multiple Access (TDMA) to multiplex up to 8 calls per channel in the 900 and 1800 MHz bands.

The first commercial deployment of Global System for Mobile Communication (GSM) was done in 1992. It supported higher voice quality and provided data services such as SMS and e-mail. We will discuss the GSM system in more detail, later in this chapter. In 1993, another 2G system, known as CDMAone, was standardized and commercially deployed in South Korea and Hong Kong in 1995, followed by the United States of America in 1996.

2.5 Generation

General Packet Radio Service (GPRS) is an extension of GSM and is considered to be the 2.5 generation technology. As indicated by the name, it is based on packet switching compared to circuit switching used in 2G. This was a significant improvement over 2G and helped to reduce call costs dramatically. Another important advantage of GPRS is that it allows users to remain connected to the Internet without incurring additional charge and supports multimedia capabilities including graphics and video communications. GPRS deployments began in 2000, followed by EDGE in 2003. EDGE enhances the capabilities of GPRS, allows faster Internet browsing, and makes it possible to use streaming applications. Though this technology provided faster data rates over 2G systems, it is called 2.5G because it did not offer the multi-megabit data rates which are the characteristics of the 3G system.

Third generation

The 3G systems are often referred to as IMT-2000 (International Mobile Telecommunications-2000) systems since this was made a global standard by ITU. The 3G systems support much higher data transmission rates and offer increased bandwidth, which makes them suitable for high-speed data applications as well as for high quality traditional voice calls. The

3G systems can be considered to be purely data networks, since voice signals are converted to digital data, this results in speech being dealt with in much the same way as any other form of data. The 3G systems use packet-switching technology, and provide cheaper calls while giving better average call quality than that of the 2G systems, but they do require a somewhat different infrastructure compared to the 2G systems. The 3G networks made it possible for service providers to offer many innovative applications and services such as email, instant messaging and video telephony, multimedia gaming, live-video buffering, and location-based services among others. The first 3G network was deployed in Japan in 2001 by DoCoMo.

UMTS (Universal Mobile Telephone System) is one of the 3G mobile systems that was developed within the ITU's IMT-2000 framework. UMTS was developed mainly for the GSM networks, so that these could be easily upgraded to UMTS networks. In UMTS, coverage is provided by a combination of a variety of cell sizes ranging from "in building" *pico cells* to *global cells* provided by satellites.

Even though it was expected that the UMTS specification would become a single global standard for 3G systems, it did not turn out that way. Now many different versions of 3G systems have come into existence and each one evolved from some existing 2G system. The main 3G technologies that are prevalent include UMTS and CDMA2000. European countries have adopted UMTS, while the USA uses CDMA2000.

Fourth generation

A 4G system provides a faster data rate than that of 3G (at least 10 times faster) and makes mobile broadband Internet access possible. The 4G system has made possible high speed Internet access from smartphones and laptops with USB wireless modems. A few applications that could not be supported in earlier generations of the cellular phone systems, have now become possible in 4G including IP telephony, gaming services, high-definition mobile TV, video conferencing and 3D television. The 4G technology is expected to help solve the last mile problem that prevents the mobile users from running applications that are available on wired networks. There are at present two competing 4G standards: Mobile WiMAX standard and the Long Term Evolution (LTE) standard.

In the following section, we provide a brief overview of the working of a few cellular wireless technologies that are being popularly used at present.

2.7 Global System for Mobile Communications (GSM)

GSM (Global System for Mobile Communications) is at present being used in India. It is possibly the most successful digital mobile system to have ever been used till now. An important characteristic of the GSM system

is that it provides data services in addition to voice services, and yet is compatible to 1G systems.

GSM networks operate in four different radio frequencies. Most GSM networks operate either in the 900 MHz or in the 1800 MHz frequency bands. Some countries in the American continent (especially the USA and Canada) use the 850 MHz and 1900 MHz bands because the 900 MHz and 1800 MHz frequency bands are already allocated for other purposes. The relatively rarely used 400 MHz and 450 MHz frequency bands are assigned in some countries, notably Scandinavia where these frequencies were previously used for the first generation systems. In the 900 MHz band, the uplink frequency band is 890–915 MHz, and the downlink frequency band is 935–960 MHz.

2.7.1 GSM Services

GSM provides three main categories of services. These are:

(i) Bearer services
(ii) Teleservices
(iii) Supplementary services

In the following, we elaborate these different categories of services.

Bearer services

Bearer services give the subscribers the capability to send and receive data to/from remote computers or mobile phones. For this reason, bearer services are also known as *data services* (see Box 2.1). These services also enable the transparent transmission of data between GSM and other networks like PSTN, ISDN, etc. at rates from 300 bps to 9600 bps. These services are implemented on the lower-three layers of the OSI reference model. Besides supporting SMS, e-mail, voice mailbox, and Internet access, this service provides the users with the capability to execute remote applications. GSM supports data transfer rates of up to 9.6 kbps.

BOX 2.1 GSM bearer services

The GSM data services are named *bearer* services. Consider the following example: Suppose a customer requires to send a data file such as a picture to a computer at the office that is connected to a public telephone network. In this example, the bearer service provides 9.6 kbps circuit-switched data transfer. The handset dials the office computer telephone number and establishes a connection with it via the modem. When the office computer modem accepts the call, the customer's handset begins to send data directly on the telephone line channel at 9.6 kbps.

Bearer services permit either transparent or non-transparent, and either synchronous or asynchronous modes of data transmission. We elaborate these in the following.

- The transparent bearer services use the functions of the physical layer of transmission of data leading to constant delay and throughput if no transmission errors occur. There is a mechanism called FEC (Forward Error Correction) to increase the quality of data transmission.
- The non-transparent bearer services use protocols of the second and third layers to implement error correction and flow control. They use transparent bearer services in addition to a Radio Link Protocol (RLP). This protocol comprises mechanisms of high level data link control.

Teleservices

GSM provides both the voice-oriented teleservices and the non-voice teleservices, as discussed below.

Telephony: The main goal of GSM was to provide high quality digital voice transmission, offering the bandwidth of 3.1 kHz of analog phone systems. Special codecs are used for voice transmission, while other codecs are used for the transmission of analog data for communication with traditional computer modems used in fax machines.

Emergency number: The same number is used throughout an area. This service is free of cost and mandatorily provided by all service providers. This connection will automatically be set up with the closest emergency centre.

Short message services: This service offers transmission of text messages of sizes up to 160 characters. SMS services use the signalling channels, making possible the duplex system of the sending and receiving the SMSs messages.

Fax: In this service, using modems fax data is transmitted as digital data over the analog telephone network according to the ITU-T Standards T.4 and T.30.

Supplementary services

GSM provides certain supplementary services such as user identification, call redirection, and forwarding of ongoing calls. In addition, standard ISDN features such as 'close user groups' and 'multiparty' communication are available.

2.7.2 System Architecture of GSM

A GSM system consists of three main subsystems:

(i) Radio Subsystem (RSS)

(ii) Networking and Switching Subsystem (NSS)
(iii) Operation Subsystem (OSS)

A schematic of the functional architecture of a GSM system is shown in Fig. 2.9. The different components of this architecture are briefly explained in the following.

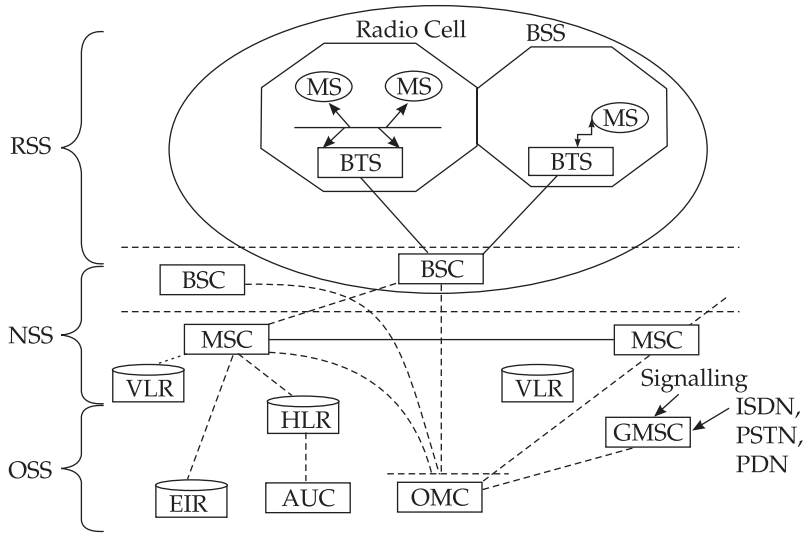


Figure 2.9 *Functional architecture of a GSM system.*

Radio subsystem (RSS)

This subsystem comprises all the radio specific entities. That is, the mobile stations, the base station subsystems, the base transceiver station and the base station controller. We briefly explain the important components of the radio subsystem in the following:

Mobile Station (MS): A mobile station (MS) or cell phone contains two major components: the subscriber identity module (SIM) and the mobile device. The SIM is a removable smart card. Each mobile device has a unique identifier that is known as its IMEI (International Mobile Equipment Identity). Apart from the telephone interface, an MS also offers other types of interfaces to the users such as USB, Bluetooth, etc. Despite its small size, a SIM card is a very important component of a GSM network. It contains all the subscription information of a subscriber and holds the key information that activates the phone after it is powered on. It contains a microcontroller to primarily store and retrieve data from the flash storage on the SIM. Identification information is stored in the SIM card's protected memory (ROM) that is not accessible or modifiable by the customer. Additional flash memory is included in the mobile device to allow storage of other information such as addresses, pictures, audio and video clips, and short

messages. The SIM card contains many other identifiers and tables such as card type, serial number, a list of subscribed services, and a Personal Identity Number (PIN).

Base Station Subsystem (BSS): A GSM network comprises many BSSs. Each BSS consists of a Base Station Controller (BSC) and several Base Transceiver Stations (BTSs). We will explain these components subsequently. A BSS performs all functions necessary to maintain radio connections to an MS, as well as does coding/decoding of voice.

Base Transceiver Station (BTS): A BTS comprises all radio equipment such as antenna, signal processors and amplifiers that are necessary for radio transmission. It encodes the received signal, modulates it on a carrier wave, and feeds the RF signals to the antenna. It communicates with both the mobile station and the BSC.

Base Station Controller (BSC): A BSC manages the radio resource of the BTSs in the sense that it assigns frequency and time slots for all MSs in the area. It also manages the handoff from one BTS to another within the BSS. The BSC also multiplexes the radio channels onto the fixed network connection to the Mobile Switching Centre (MSC).

Network and switching subsystem (NSS)

This subsystem forms the heart of the GSM system. It connects the wireless networks to the standard public networks and carries out usage-based charging, accounting, and also handles roaming. NSS consists of a switching centre and several databases as described below.

Mobile Switching Center (MSC): An MSC can be considered to form the heart of a GSM network. An MSC sets up connections to other MSCs and to other networks such as Public Data Network (PDN). An MSC is responsible for the connection setup, connection release, and call handoff to other MSCs. A Gateway MSC (GMSC) is responsible for gateway functions, while a customer roams to other networks. It also performs certain other supplementary services such as call forwarding, multiparty calls, etc.

Home Location Registers (HLRs): A HLR stores in a database important information that is specific to each subscriber. The information contains subscriber's IMSI, pre/post paid, user's current location, etc.

Visitor Location Register (VLR): It is essentially a temporary database that is updated whenever a new MS enters its area by roaming. The information is obtained from the corresponding HLR database. The function of the VLR is to reduce the number of queries to the HLR and make the user feel as if he were in his home network.

Operation subsystem (OSS)

The operation subsystem contains all the functions necessary for network operation and maintenance. It consists of the following:

Operation and Maintenance Centre (OMC): It supervises all other network entities. Its functions are traffic monitoring, subscribers, security management and accounting billing.

Authentication Centre (AuC): It protects against intruders targeting the air interface. The AuC stores information concerned with security features such as user authentication and encryption. The AuC is related to the HLR.

Equipment Identity Register (EIR): It is essentially a database that is used to track handsets using the IMEI. It helps to block calls from stolen, unauthorized, or defective mobiles.

2.7.3 GSM Security

Security in GSM is broadly supported at three levels: operator's level, customer's level and system level. These three levels help oversee aspects such as correct billing to the customer, avoiding fraud, protecting services, and ensuring anonymity. The following are a few important features associated with providing security in GSM networks.

Authentication

The purpose of authentication is to protect the network against unauthorized use. In the GSM context, it helps protect the GSM subscribers by denying the possibility for intruders to impersonate authorized users. A GSM network operator can verify the identity of the subscriber, making it highly improbable to clone someone else's mobile phone identity.

Authentication can be achieved in a simple way by using a password such as Personal Identification Number (PIN). This method is not very secure in GSM networks as an attacker can "listen" the PIN and easily break the code.

Confidentiality

A GSM network protects voice, data and sensitive signalling information (e.g. dialed digits) against eavesdropping on the radio path. Confidentiality of subscriber-dialled information in the GSM network is achieved by using encryption techniques prescribed by the GSM designers. Data on the radio path is encrypted between the Mobile Equipment (ME) and the BTS which protects user traffic and sensitive signalling data against eavesdropping.

Anonymity

A GSM network protects against someone tracking the location of a user or identifying calls made to (or from) the user by eavesdropping on the radio path. The anonymity of the subscriber on the radio access link in the GSM network is achieved by allocating Temporary Mobile Subscriber Identity (TMSIs) instead of permanent identities. This helps to protect against tracking a user's location and obtaining information about a user's calling pattern.

2.8 General Packet Radio Service (GPRS)

GPRS when integrated with GSM, significantly improves and simplifies Internet access. It transfers data packets from GSM mobile stations to external packet data networks (PDNs). Packets can be directly routed from the GPRS mobile stations to packet switched networks making it easy to connect to the Internet.

GSM uses a billing system based on the time (duration) of connection, whereas GPRS uses a billing system based on the amount of transmitted data rather than the duration of the connection. So, users can remain continuously connected to the system, and yet get charged only for the amount of transmitted data.

2.8.1 GPRS Services

GPRS offers end-to-end packet-switched data transfer services which can be categorized into the following two types:

(i) Point-to-Point (PTP) service
(ii) Point-to-Multipoint (PTM) service.

The PTP service is between two users and can either be connectionless or connection-oriented. The PTM is a data transfer service from one user to multiple users. Again, there are two types of PTM services. One is multicast PTM where the data packets are broadcast in a certain area and the other is group call PTM where the data packets are addressed to a group of users.

2.8.2 GPRS Architecture

GPRS architecture introduces two new network elements, called GPRS Support Node (GSN) and the Gateway GPRS Support Node (GGSN). A GSN is essentially a router. All GSNs are integrated into a standard GSM architecture. The GGSN is the interworking unit between the GPRS network and the external packet data network (PDN). The GGSN contains

routing information for GPRS users, performs address connection and tunnells data to a user through encapsulation. In Fig. 2.10, the GGSN is connected to an external network and it transfers packets to the SGSN through an IP-based GPRS backbone network.

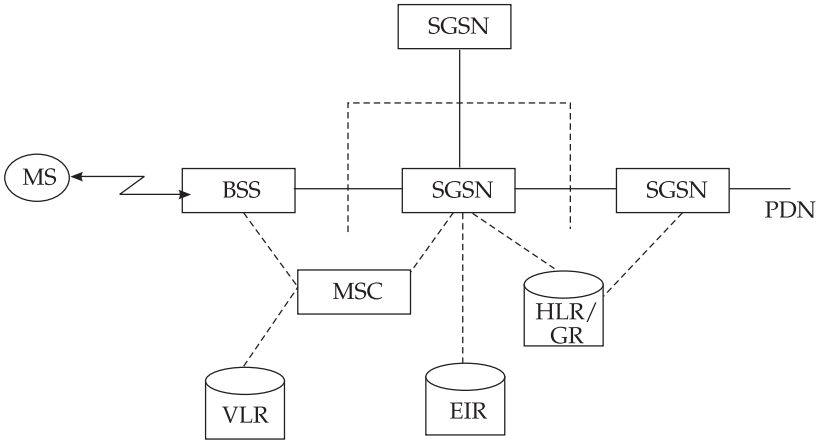


Figure 2.10 *GPRS architecture reference model.*

As shown in Fig. 2.10, SGSN (Serving GPRS Support Node) helps support MS. The SGSN is connected to BSC through frame relay and it is at the same hierarchy level as the MSC. The GPRS Register (GR) is a part of HLR which stores all the relevant GPRS data. In a part of HLR which stores all the relevant data of GPRS in a mobile IP network, GGSN and SGSNs can be compared with home agent and foreign agent respectively. The data packets are transmitted to the BSS and finally to the MS through the GGSN and SGSN. The MSC as we have already discussed is responsible for data transport in the traditional circuit-switched GSM.

2.9 Universal Mobile Telecommunications System (UMTS)

CDMA2000 and UMTS were developed separately and are two separate ITU approved 3G standards. In these networks, coverage is provided by a combination of various cell sizes, ranging from "in building" pico cells to global cells provided by satellites, giving service to the remote regions of the world.

The UMTS was developed mainly for countries with GSM networks, and it is expected that all GSM networks will be upgraded to UMTS networks. Because it is a new technology, a whole new radio access network has to be built. An important advantage of UMTS is that it gives significantly enhanced capacities to operators.

The UMTS specification has been designed so that the UMTS systems are compatible with GSM networks. Therefore, the UMTS networks can easily work with any existing GSM/GPRS network. The UMTS systems use different frequency bands, so the BTSs do not interfere with each other.

Let us now discuss the dissimilarities between these networks. The UMTS networks are different from the 2G networks in the following respects:

Higher speech quality: In addition to speech traffic, the UMTS supports the advanced data and information services and can be called a true multimedia network.

Higher data rate: The UMTS supports 2 Mbps data rate, which is much higher than that supported by the 2G mobile systems.

Virtual home environment (VHE): A user roaming from his network to other UMTS networks will not feel any discontinuity or service difference, thus giving a "feeling" of being in the home network. In contrast, in a 2G network, a user is registered to a visitor location and is also charged a roaming overhead.

2.9.1 UMTS Network Architecture

The UMTS network architecture can be divided into three main elements:

User Equipment (UE): The User Equipment (UE) is the name by which a cell phone is referred to. The new name was chosen because of the considerably greater functionality that the UE incorporates compared to a cell phone. It can be thought of as both a mobile phone used for talking and a data terminal attached to a computer with no voice capability.

Radio Network Subsystem (RNS): The RNS is the equivalent of the Base Station Subsystem (BSS) in GSM. It provides and manages the wireless interface for the overall network.

Core Network: The core network is the equivalent of the GSM Network Switching Subsystem (NSS).

2.10 Mobile Phone and Human Body

Extensive research has been conducted to study the possible health effects of continuous exposure to low intensity RF fields produced by mobile phones. The overall evidence suggests that mobile phone usage of less than 10 years does not pose any increased risk of brain tumour. The effect of still longer use is unclear due to non-availability of data. Any conclusion

therefore is uncertain and tentative. From the available data, however, it does appear that there is no increased risk for brain tumours in long-term users as well, with the exception of acoustic neuroma (see Box 2.2) for which there is limited evidence of a weak association. Available studies suggest that self-reported symptoms are not correlated to an acute exposure to RF fields, but considering that the studies are carried out over a rather limited duration, it is very difficult to draw any firm conclusions. Currently available studies on neurological effects and reproductive effects have not indicated any health risks at exposure levels below guidelines.

BOX 2.2 Acoustic neuroma

Acoustic neuroma is a non-cancerous tumour that develops on the nerve that connects the ear to the brain. The tumour usually grows slowly. As it grows, it presses against the hearing and balance nerves. At first, there may be no symptoms or mild symptoms. They can include the following:

- Loss of hearing on one side
- Ringing in ears
- Dizziness and balance problems

SUMMARY

Mobile computing is an umbrella term used to describe technologies that enable people to access network services—anyplace, anytime, and anywhere. The term ubiquitous computing and nomadic computing are often used synonymously with mobile computing. Mobile computing essentially involves accessing and manipulating information over a wireless network using a hand-held device. The challenges in designing effective mobile computing solutions include providing acceptable service in the presence of low available bandwidth, intermittent disconnections, poor security, and difficulty in assigning fixed addresses to hosts. Unlike their wired counterparts, the design of software for mobile devices must consider resource limitation, battery power and display size. Consequently, new hardware and software techniques must be developed. For example, applications need to be highly optimized for space, in order to fit in the limited memory on the mobile devices. For Internet enabled devices, the good old TCP/IP stack is not very suitable, it is computationally expensive for hand-held devices, does not make efficient use of the available bandwidth and is not optimized for low power consumption.

Digital cellular standards meet the current requirements of voice communications and are being upgraded to meet the future demands in mobile multimedia applications. The 3G mobile networks represent an evolution in terms of capacity, data speeds and new service capabilities from second generation mobile networks to provide an integrated solution for mobile voice and data with wide area coverage.

FURTHER READINGS

Abdelsalam Helal, Bert Haskell, J. Carter, R. Brice, D. Woelk and M. Rusinkiewicz, *Anywhere Computing Concepts and Technology*, Kluwer Academic Publishers, September 1999.

Adelstein, Gupta, Rechard III, and Schwiebert, *Fundamentals of Mobile and Pervasive Computing*, TMH.

Akyildiz, I.F. and J.S.M. Ho, " Dynamic Mobile User Location Update for Wireless PCS Networks", *ACM/Baltzer Wireless Networks Journal*, 1995.

Arkady Zaslavsky and Zahir Tari, "Mobile Computing Overview and Current Status", *Australian Computer Society*, 1998.

Christian Bettstetter, Hans-Jorg Vogel, and Jorg Eberspacher, "GSM Phase 2 + GPRS Architecture, Protocol and Air Interface", *IEEE Communication Survey*, 1999, Vol. 2.

Jan A. Audestad, "Network Aspect of GSM System", *EURCCON*, June 1988.

Max Stepanov, "GSM Features and Security", July 2004.

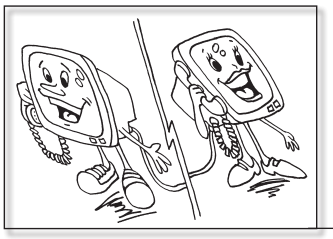
Rappaport, T., *Wireless Communication and Principles*, Pearson Education.

Tuan Huynh and Hoang Nguyen, "Overview of GSM and GSM Security", 2003.

EXERCISES

1. State **true** or **false** against each of the following statements. Give appropriate reasons for your answer.
   (i) Mobile computing and wireless networking are synonymous terms.
   (ii) The Bluetooth technology is an example of an infrastructure-less network.
   (iii) WLAN is an example of an infrastructure-less network.
   (iv) A sensor network is essentially based on a single hop wireless communication.
   (v) 2G cellular phones support electronic mail services.
   (vi) In the Global Packet Radio Service (GPRS), the customer is charged for Internet access based on connection time rather than on total information download.
   (vii) GSM stands for Group of Special Mobiles.
   (viii) UMTS networks can easily work with the existing GSM/GPRS networks.
2. What is mobile computing? Mention at least three applications of mobile computing.

3. Distinguish between mobile computing and wireless networking.
4. Give an overview of the working of current mobile cellular phones. Briefly explain the distinguishing features of various generations of wireless cellular networks.
5. Briefly explain how the mobile cellular communication has evolved over different generations of technology.
6. What do you understand by 2.5G? Mention a few characteristic features of this technology. How is it different from 2G and 3G technologies?
7. Distinguish between infrastructure-based networks and infrastructure-less networks with the help of suitable schematic diagrams.
8. Briefly write about the Bluetooth technology. Give at least two examples of its use.
9. Explain the architecture of a mobile computing environment. Define the functions of the presentation tier, application tier and data tier of mobile computing environment.
10. Briefly discuss the important functional differences between 1G, 2G, and 3G cellular networks.
11. Compare 1G and 2G cellular wireless communication technologies.
12. Is 3G cellular wireless technology superior to 2G technology? Justify your answer.
13. Explain how a GSM network provides security to the customers.
14. What are the advantages of GPRS over GSM?
15. What is UMTS? Describe the functions of HLR and VLR in call routing and roaming?
16. Identify at least three similarities and three dissimilarities between a GMS network and a UMTS network.
17. What do you mean by Virtual Home Environment (VHE)? Explain how VHE is realized in 3G networks.
18. Do mobile phones affect the human body negatively? Explain your answer.
19. Discuss the advantages and disadvantages of supporting TCP/IP in a mobile computing network.
20. Identify the main reasons as to why a mobile handset is compact and lightweight and yet provides a large number of features such as roaming, camera, audio and video play and record, Internet browsing, etc., while the traditional landline phone handsets are bulky and provide only limited features.
21. What is the difference between analog and digital transmissions in the context of mobile communications? What are their relative advantages?



3

MAC Protocols

In a wireless network, multiple nodes may contend to transmit on the same shared channel at the same time. In this situation, the transmitted data would get garbled unless a suitable medium access arbitration scheme is deployed. Usually, it is the responsibility of the medium access control (MAC) protocol to perform this task. The MAC protocol is a sublayer of the data link layer protocol and it directly invokes the physical layer protocol.

The primary responsibility of a MAC protocol is to enforce discipline in the access of a shared channel when multiple nodes contend to access that channel. At the same time, two other objectives of any MAC protocol are maximization of the utilization of the channel and minimization of average latency of transmission. However, a MAC protocol must be fair and ensure that no node has to wait for an unduly long time, before it is allowed to transmit. In the following, we identify the various characteristics desirable of any MAC protocol.

3.1 Properties Required of MAC Protocols

The design of a MAC protocol depends upon the specific environment in which it would be used and the specific design requirements to be met. In spite of the wide variations in the characteristics of different protocols, however, in a general sense a good MAC protocol needs to possess the following features:

- It should implement some rules that help to enforce discipline when multiple nodes contend for a shared channel.
- It should help maximize the utilization of the channel.
- Channel allocation needs to be fair. No node should be discriminated against at any time and made to wait for an unduly long time for transmission.

- It should be capable of supporting several types of traffic having different maximum and average bit rates.
- It should be robust in the face of equipment failures and changing network conditions.

Many MAC layer protocols for wireless networks have already been proposed, standardized, and are in use. Also, many other protocols that work with improved efficiency or overcome some problems in specific wireless environments, are being proposed by researchers and practitioners. At present, IEEE 802.11 has emerged as a popular and standard MAC protocol for wireless networks. IEEE 802.11-based network cards and routers are available in the market that can be used to inexpensively and easily set up wireless LANs (commonly referred to as Wi-fi hotspots). As we discussed in Chapter 1, wireless networks can be divided mainly into two categories: (a) infrastructure-based wireless networks that include the WLANs, and (b) infrastructure-less wireless networks that include the mobile ad hoc networks (MANETs). Though the MAC protocols for these two environments have many things in common, MAC protocols for Infrastructure-less networks are surprisingly much more complex as they have to address certain additional problems that arise in the infrastructure-less environments. We discuss MAC protocols for both these environments in this chapter.

In Section 3.2, we first discuss a few basic issues concerning the MAC protocols in a wireless network. In Section 3.3, we discuss a taxonomy of MAC protocols for wireless networks. In the three subsequent sections, we discuss the three important categories of MAC protocols: the fixed assignment, the random access, and the reservation-based protocols. Finally, we discuss the MAC protocols that have been specifically designed for ad hoc networks.

3.2 Wireless MAC Protocols: Some Issues

A MAC protocol in a wireless medium is much more complex than its wired counterpart. First, a collision detection scheme is difficult to implement in a wireless environment, since collisions are hard to be detected by the transmitting nodes. Also, in infrastructure-less networks, the issue of hidden and exposed terminals make a MAC protocol extremely inefficient unless special care is taken to overcome these problems. We elaborate the hidden and exposed terminal problems in the following:

3.2.1 The Hidden and Exposed Terminal Problems in an Infrastructure-less Network

The *hidden terminal* problem arises when at least three nodes (A, B, and C), as shown in Fig. 3.1, communicate among each other. As shown in this figure,

B is in the radio range of A, and B is also within the radio range of C. However, the nodes A and C are not in the radio range of each other. Note that if both A and C start to transmit to B at the same time, the data received at node B would get garbled. Such a situation can arise because A and C are "hidden" from each other, because they are outside each other's transmission range. In this situation, when one node starts to sense the medium before transmission, it cannot sense that the other node is also transmitting. This creates a very difficult and important arbitration problem that a MAC protocol needs to resolve.

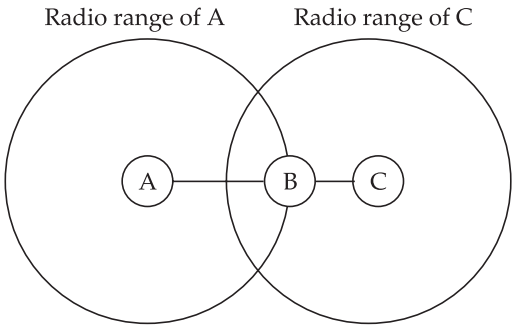


Figure 3.1 *Hidden terminal problem.*

A related problem called *exposed terminal* could arise in a scenario such as that depicted in Fig. 3.2. MAC protocols usually inhibit transmission when transmission from another terminal is detected. As a result, node A will not be able to transmit to any node when B is transmitting to C. On the other hand, had A transmitted to D, it would have been received correctly by D and B's transmission would have also been correctly received at C. The problem arose only because A and B are within each other's transmission range, though the destination nodes are in the transmission range of only one of the nodes. In other words, the problem occurs because A is exposed to B's transmission. The overall effect of this problem is that it leads to inefficient spectrum usage as well as unnecessary transmission delays unless these are carefully addressed by a wireless MAC protocol.

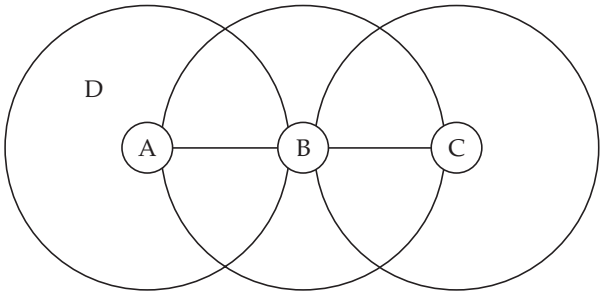


Figure 3.2 *Exposed terminal problem.*

3.3 A Taxonomy of MAC Protocols

A large number of MAC protocols have been proposed. These MAC protocols can be broadly divided into the following three categories:

(i) Fixed assignment schemes
(ii) Random assignment schemes
(iii) Reservation-based schemes

The fixed assignment schemes are usually called circuit-switched schemes. In the fixed assignment schemes, the resources required for a call are assigned for the entire duration of the call. On the other hand, the random assignment schemes and the reservation schemes are called packet-switched schemes. The random assignment schemes are comparable to the connection-less packet-switching schemes. In this, no resource reservations are made, the nodes simply start to transmit as soon as they have a packet to send. In the reservation schemes, a node makes explicit reservation of the channel for an entire call before transmitting. This is analogous to a connection-based packet-switching scheme. The reservation-based MAC schemes are suitable to handle calls with widely varying traffic characteristics. In the following sections, we discuss these three categories of MAC protocols in some more detail.

3.4 Fixed Assignment Schemes

A few important categories of fixed assignment MAC protocols are the following:

- Frequency Division Multiple Access (FDMA)
- Time Division Multiple Access (TDMA)
- Code Division Multiple Access (CDMA)

We briefly discuss these techniques in the following subsections.

BOX 3.1 An analogy to the fixed assignment solution to the multiple access issues of a shared medium

An analogy may be drawn to the fixed assignment solution to the multiple access issues of a shared medium in the following way: Consider a students' common room (channel) in which many students want to communicate with each other. If the students want to avoid cross-talk in the ongoing process, then either the students could take turns in speaking (i.e. time division), or they could speak at different pitches (i.e. frequency division), or they could speak in different languages (i.e. code division). The last analogy captures the essence of CDMA, when the students who are speaking the same language understand each other, but the rest of the students cannot. In CDMA, each communicating pair shares a decryption code using which lets them understand only the communication between them. In this case many codes occupy the same channel, but only the users who share a specific code will be able to understand each other.

3.4.1 Frequency Division Multiple Access (FDMA)

In FDMA, the available bandwidth (frequency range) is divided into many narrower frequency bands called channels. Figure 3.3 shows a division of the existing bandwidth into many channels (showns as Ch 1, Ch 2, etc.). For full duplex communication to take place, each user is allocated a forward link (channel) for communicating from it (mobile handset) to the base station (BS), and a reverse channel for communicating from the BS to it. Thus, each user making a call is allocated two unique frequency bands (channels), one for transmitting and the other for receiving signals during the call. Obviously, when a call is underway, no other user would be allocated the same frequency band to make a call. Unused transmission time in a frequency band that occurs when the allocated caller pauses between transmissions, or when no user is allocated a band, goes idle and is wasted. FDMA, therefore, does not achieve a high channel utilization.

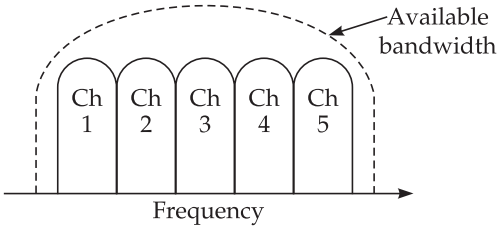


Figure 3.3 *Channels in Frequency Division Multiple Access (FDMA) scheme.*

3.4.2 Time Division Multiple Access (TDMA)

TDMA is an access method in which multiple nodes are allotted different time slots to access the same physical channel. That is, the timeline is divided into fixed-sized time slots and these are divided among multiple nodes who can transmit. Note that in this case, all sources use the same channel, but take turns in transmitting. Figure 3.4 shows the situation where time slots are allocated to users in a round robin manner, with each user being assigned one time slot per frame. See Box 3.2. Obviously, unused time slots go idle, leading to low channel utilization.

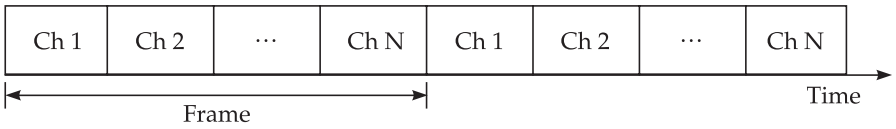


Figure 3.4 *Channels in Time Division Multiple Access (TDMA) scheme.*

BOX 3.2 TDMA scheme

In TDMA, each user of the channel owns the channel for exclusive use for one time slot at a time in a round robin fashion.

3.4.3 Code Division Multiple Access (CDMA)

In CDMA, multiple users are allotted different codes that consist of sequences of 0 and 1 to access the same channel. As shown in Fig. 3.5, a special coding scheme is used that allows signals from multiple users to be multiplexed over the same physical channel. As shown in the figure, three different users who have been assigned separate codes are multiplexed on the same physical channel.

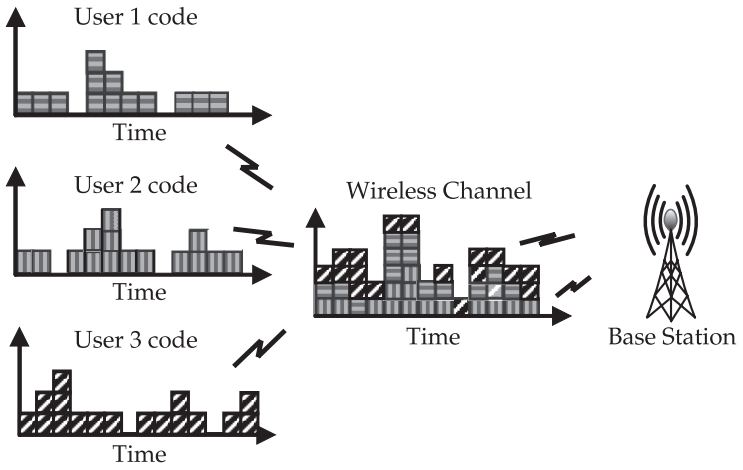


Figure 3.5 *Schematic of operation of Code Division Multiple Access (CDMA).*

In the following, we elaborate the CDMA technology. In CDMA, multiple users use the same frequency at the same time and no time scheduling is applied. All the senders send signals simultaneously through a common medium. The bandwidth of this medium is much larger than the space that would be allocated to each packet transmission during FDMA and the signals can be distinguished from each other by means of a special coding scheme that is used. This is done with the help of a frequency spreading code known as the *m*-bit pseudo-noise (PN) code sequence. Using *m* bits, $2^m - 1$ different codes can be obtained. From these codes, each user will use only one code.

BOX 3.3 How to distinguish transmission from different nodes

Two vectors are said to be orthogonal if their inner product = 0. Let **p** and **q** be two vectors and suppose **p** = (2, 5, 0) and **q** = (0, 0, 17), then the inner product of **p**∗**q** = (2∗0+5∗0+0∗17) = 0.

For good autocorrelation, binary 0 is represented as –1 and binary 1 is represented as +1. Let the binary sequence be 1001, then the representation sequence is +1–1–1+1.

It is possible to distinguish transmissions from different nodes by ensuring some properties on the codes. A code for a user should be orthogonal (that is, non-interfering) to the codes assigned to other nodes. The term "orthogonal" means that the vector inner product is zero, and good autocorrelation uses the bipolar notation where a code sequence of binary 0 is represented as –1 and binary 1 is represented as +1. See Box 3.3. On the receiving end, only the same PN sequence is able to demodulate the signal to successfully convert the input data.

BOX 3.4 Pseudorandom sequence generator

To generate a series of pseudorandom numbers, a seed (or starting point) is required. Based on the selected seed, the next number can be generated using a deterministic mathematical transformation or can be generated probabilistically. In CDMA, a code actually denotes a starting point (seed) for a pseudorandom sequence generator (PRSG). PRSG generates a series of bits at a frequency which is much higher than the actual user data (such as digitized voice). These bits are XORd with the user data and subsequently the results are transmitted. This occurs in the case of multiple transmitters.

If someone listens to this signal with the help of a suitable wideband receiver, the person will hear something similar to what is produced by random noise. All the other users who are on the same frequency will send a similar signal, but with a different PRSG seed. So these apparent random noises will all coexist in the same band of frequencies, but would not interfere with each other. This is due to the reason that the exact frequency of any transmitter at any instant (which is in effect determined by the seed) is almost always unique. Error correction takes care of occasional bit errors.

The receiver is aware of the PRSG starting point for each transmitter. It hears just one of the transmitters by correlating the noise it receives, against its own PRSG, which is also running with the same seed. It is slightly similar to FDMA in this sense, but the difference is that the transmitters do not stay on one frequency. They hop around many times per bit of user data. The pseudorandom sequence determines this hopping, rather than a fixed assignment to each transmitter.

For simplicity, we assume that all nodes transmit on the same frequency at the same time using the entire bandwidth of the transmission channel. Each sender has a unique random number key, and the sender XORs the signal with this random number key. The receiver can "tune" into this signal if it knows the pseudorandom number. Consider an example, where X, Y are the transmitters and Z is a receiver. Sender X_data = 1 and X_Key = (010011). Its autocorrelation representation is (–1, +1, –1, –1, +1, +1). The signal to be calculated at sender X is Xs = X_data ∗ X_key = +1∗X_key = (–1, +1, –1, –1, +1, +1). Similarly, sender Y_data = 0 and Y_key = (110101) and the signal to be sent at Y is Ys = –1∗Y_key = –1∗(+1, +1, –1, +1, –1, +1) = (–1, –1, +1, –1, +1, –1). The signal received by receiver Z is Xs + Ys = (–1, +1, –1, –1, +1, +1) + (–1, –1, +1, –1, +1, –1) = (–2, 0, 0, –2, +2, 0). At the receiver, in order to receive the data sent by sender X, the signal

Z is dispread. So now if Z wants to get information of sender X data, then Z∗X_key = (–2, 0, 0, –2, +2, 0)∗(–1, +1, –1, –1, +1, +1) = 2 + 0 + 0 + 2 + 2 + 0 = 6 > 0 (positive), that is the original bit was a 1. Similarly, the information of sender Y data may be obtained as Z∗Y_key = (–2, 0, 0, –2, +2, 0)∗(+ 1, +1, –1, +1, –1, +1) = –2 + 0 + 0 – 2 – 2 + 0 = –6 < 0 (negative). So the Y data original bit was a 0.

3.5 Random Assignment Schemes

There are a number of random assignment schemes that are used in MAC protocols. A few important ones are the following:

- ALOHA
- Slotted ALOHA
- CSMA
- CSMA/CD
- CSMA/CA

3.5.1 ALOHA Scheme

It is a simple communication scheme that was developed at the university of Hawaii. The basic (also called pure) ALOHA scheme, is a simple protocol. If a node has data to send, it begins to transmit. Note that the first step implies that Pure ALOHA does not check whether the channel is busy before transmitting. If the frame successfully reaches the destination (receiver), the next frame is sent. If the frame fails to be received at the destination, it is sent again. The simple ALOHA scheme works acceptably, when the chances of contention are small (i.e., when a small number of senders send data infrequently). However, the collisions can become unacceptably high if the number of contenders for transmission is high. An improvement over the pure ALOHA scheme is the slotted ALOHA. In the slotted ALOHA scheme, the chances of collisions are attempted to be reduced by enforcing the following restrictions. The time is divided into equal-sized slots in which a packet can be sent. Thus, the size of the packet is restricted. A node wanting to send a packet, can start to do so only at the beginning of a slot. The slotted ALOHA system employs beacon signals that are sent at precise intervals that mark the beginning of a slot, at which point the nodes having data to send can start to transmit. Again, this protocol does not work very well if the number of stations contending to send data is high. In such cases, the CSMA scheme (described next) works better.

3.5.2 The CSMA Scheme

A popular MAC arbitration technique is the Carrier Sense Multiple Access (CSMA). In this technique, a node senses the medium before starting to transmit. If it senses that some transmission is already underway, it defers its transmission. Two popular extensions of the basic CSMA technique are the collision detection (CSMA/CD) and the collision avoidance (CSMA/CA) techniques.

Unlike that in a wired network, in a wireless network the CSMA/CD technique does not work very well. In the CSMA/CD technique, the sender starts to transmit if it senses the channel to be free. But, even if it senses the channel to be free, there can be a collision (why?) during transmission. In a wired network, the implementation of a collision detection scheme is simple. However, in a wireless network it is very difficult for a transmitting node to detect a collision, since any received signal from other nodes would be too feeble compared to its own signal and can easily be masked by noise. As a result, a transmitting node would continue to transmit the frame, and only the destination node would notice the corrupted frame after it computes the checksum. This leads to retransmissions and severe wastage of channel utilization. In contrast, in a wired network when a node detects a collision, it immediately stops transmitting, thereby minimizing channel wastage.

In a wireless network, a collision avoidance scheme works much better compared to a collision detection-based scheme. A collision avoidance scheme is based on the idea that it is necessary to prevent collisions at the moment they are most likely to occur, that is, when the bus is released after a packet transmission. We explain the reason for this in the following. During the time a node is transmitting on the channel, several nodes might be wanting to transmit. These nodes would be monitoring the channel and waiting for it to become free. The moment the transmitting node completes its transmission, these waiting nodes would sense the channel to be free, and would all start transmitting at the same time. To overcome such collisions, in the collision avoidance scheme, all nodes are forced to wait for a random time and then sense the medium again, before starting their transmission. If the medium is sensed to be busy, a node waiting to transmit waits for a further random amount of time and so on. Thus, the chance of two nodes starting to transmit at the same time would be greatly reduced.

3.6 Reservation-based Schemes

A basic form of the reservation scheme is the RTS/CTS scheme. In an RTS/CTS scheme, a sender transmits an RTS (Ready to Send) packet to the receiver before the actual data transmission. On receiving this, the

receiver sends a CTS (Clear to Send) packet, and the actual data transfer commences only after that. When the other nodes sharing the medium sense the CTS packet, they refrain from transmitting until the transmission from the sending node is complete.

In a contention-based MAC protocol, a node wanting to send a message first reserves the medium by using an appropriate control message. For example, reservation of the medium can be achieved by transmitting a "Ready To Send" (RTS) message and the corresponding destination node accepting this request answers with a "Clear To Send" (CTS) message. Every node that hears the RTS and CTS messages defers its transmission during the specified time period in order to avoid a collision. A few examples of RTS-CTS based MAC protocols are MACA, MACAW, MACA-BI, PAMAS, DBTMA, MARCH, S-MAC protocols which have specifically been designed for sensor networks. In the following, we discuss MACA as a representative protocol belonging to this category of protocols.

3.6.1 MACA

MACA stands for Multiple Access Collision Avoidance. MACA solves the hidden/exposed terminal problems by regulating the transmitter power. A node running MACA requests to use the medium by sending an RTS to the receiver. Since radio signals propagate omni-directionally, every terminal within the sender's radio range will hear this and then refrain from transmitting. As soon as the receiver is ready to receive data, it responds with a CTS.

Figure 3.6 schematically shows how MACA avoids the hidden terminal problem. Before the start of its transmission, it sends a Request To Send (RTS). B receives the RTS that contains the sender's name and the receiver's name, as well as the length of the future transmission**.** In response to the RTS, an acknowledgment from B is triggered indicating Clear To Send (CTS). The CTS contains the names of the sender and receiver, and the length of the planned transmission. This CTS is heard by C and the medium is reserved for use by A for the duration of the transmission.

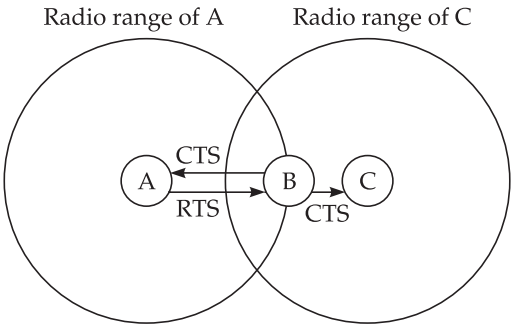


Figure 3.6 *Hidden terminal solution in MACA.*

On receipt of a CTS from B, C refrains from transmitting anything for the time indicated in the CTS. Thus a collision cannot occur at B during data transmission, and the hidden terminal problem is solved.

Though this is a collision avoidance protocol, a collision can occur during the sending of an RTS. Both A and C could send an RTS at same time. But an RTS occurs over a very small duration compared to the duration of data transmission. Thus the probability of collision remains much less. B resolves this contention problem by acknowledging only one station in the CTS. No transmission occurs without an appropriate CTS.

Figure 3.7 schematically shows how the exposed terminal problem is solved in MACA. Assume that B needs to transmit to A. B has to transmit an RTS first as shown in Fig. 3.7. The RTS would contain the names of the receiver (A) and the sender (B). C does not act in response to this message as it is not the receiver, but A responds with a CTS. C does not receive this CTS and concludes that A is outside the detection range. Thus C can start its transmission assuming that no collision would occur at A.

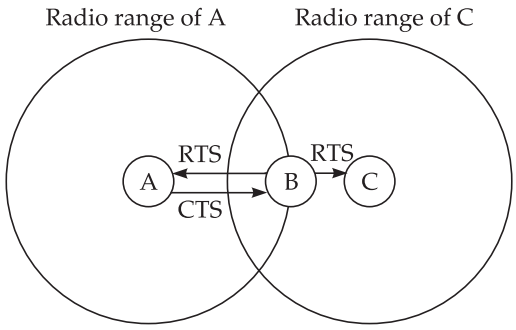


Figure 3.7 *Exposed terminal solution in MACA.*

3.7 The 802.11 MAC Standard

The IEEE 802.11 is the most widely used standard for WLANs today. Low cost WLAN cards can be easily purchased from the market to set up a WLAN. The IEEE 802.11 standard defines the functional aspects of the medium access control (MAC) sublayer. The IEEE 802.11 defines separate standards for infrastructure-based and ad hoc networks. The network interface cards can be set to work in either of these two modes.

3.7.1 Infrastructure-based Mode

This mode of usage is widely used and helps provide Wi-Fi hot spots in a campus to access the Internet. It is based on the CSMA/CA protocol due to the obvious advantages of this protocol over the CSMA/CD schemes

in a wireless environment, which we discussed in Section 3.5. In CSMA, a station that wants to transmit will first have to listen to the channel for a pre-determined period of time for checking if there is any transmission activity occurring on the channel. The station would transmit if it senses the channel to be idle. The transmission by the channel is deferred, if it senses the channel to be busy. CSMA can be considered a greedy protocol since a node trying to transmit, puts the packet out on the channel the moment it finds that the channel is idle. CSMA uses collision avoidance to improve the system's performance by trying to be less greedy. In case, the channel is sensed to be busy when it tries to transmit, then the transmission will be deferred for a "random" interval, thereby reducing the probability of further collisions on the channel. CSMA/CA is used in the 802.11-based wireless LANs.

The RTS/CTS (Request to Send/Clear to Send) is an optional mechanism used by the 802.11 to reduce frame collisions caused by the hidden node problem. As we discussed, a Request to Send (RTS) packet is sent by a sender and a Clear to Send (CTS) packet is sent by the intended receiver. The CTS alerts all the nodes within the range of both the sender or the receiver to keep quiet for the duration of the transmission.

The IEEE 802.11 has been extensively used because of its installation and operational simplicity and low cost. However, it should be borne in mind that it can be used only for a single-hop communication. That is, all nodes can communicate as long as they are in the transmission range of a single access point. This limitation can be overcome by multi-hop ad hoc networking. The 802.11a operates at 54 Mbps. Note that in 802.11, since the transmitter power is constant and limited, the data rate decreases when the transmission range increases.

3.8 MAC Protocols for Ad Hoc Networks

Ad hoc networks are infrastructure-less, self-organizing networks of mobile computers typically carried around by people. These computers are equipped with radio-enabled network interface cards for communication purposes. Every node can communicate directly with the other neighbouring nodes who can "hear" its transmissions. The computers detect other computers in their neighbourhood, dynamically forming links that set up the network. A link is also called one "hop". In addition to communicating with other computers in the neighbourhood, a computer can communicate with remote computers not in the vicinity by letting intermediate computers relay information, thereby forming a multi-hop path to the required destination computer. In forming a multi-hop path to a destination computer, a complexity that arises is on account of the mobility of the computers. Thus, a multi-hop path can become stale (that is, invalid) after some time, or might even break while a communication is still going on. Thus, no fixed infrastructure is required for setting up an ad hoc network.

Wireless multi-hop ad hoc networks are comprised of lightweight mobile radio frequency bands which are a scarce resource and therefore these are shared among all hosts. Medium Access Control (MAC) protocols for ad hoc networks can be either centralized or distributed. For instance, Bluetooth uses a centralized scheme where the master of a network assigns turns for transmission to slaves. The IEEE 802.11 is a MAC protocol that can utilize both schemes, but in ad hoc networks the distributed scheme is used. In this scheme, nodes contend for transmission turns, but if more than one node wants to transmit at the same time, a random node is chosen for transmission. The important categories of MAC protocols for MANET are discussed below.

There are essentially two broad categories of MAC protocols for ad hoc networks. The first category controls shared medium access by letting terminals compete asynchronously. In this protocol, every node is free to transmit any time. But upon detection of a collision, the access arbitration policy is invoked to avoid the collision. The second type of MAC protocol divides the medium into channels so that each competing node uses a different channel, thereby avoiding collisions. One way to achieve this is by dividing the transmission time (slots), and inserting a frequency band between terminals and requiring them to synchronize in order to ensure that they never use the same slot, frequency or code. In ad hoc networks, only the terminals that are present in the neighbourhood have to be synchronized. This is because the non-neighbour terminals cannot interfere with each other. Garcia et. al have named the first type "contention-based" protocols and the second type "contention-free" protocols. The RTS-CTS based schemes are contention-based protocols and the multiple access protocols (FDMA/TDMA/CDMA) are contention-free protocols. In Fig. 3.8, we have shown this classification of the protocols.

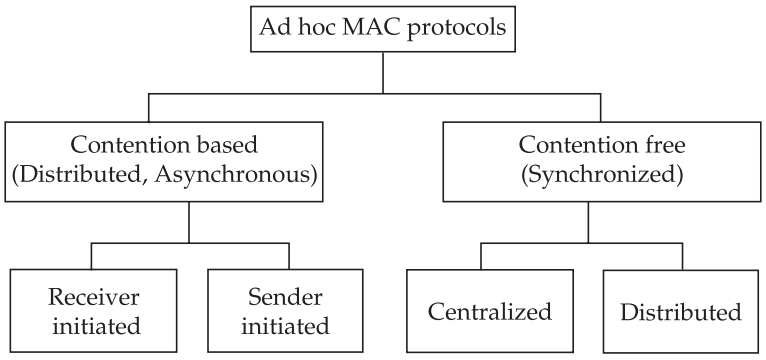


Figure 3.8 *A classification of MAC protocols for ad hoc networks.*

In another category of protocols called fixed-assignment schemes, no contention is involved. These protocols use techniques such as FDMA,TDMA and CDMA. Protocols using this kind of medium access methods are

contention-free because every node can transmit without worrying about collisions.

As discussed, the RTS-CTS based schemes are contention-based. Contention-based protocols incur additional overhead due to the reservation messages. In a contention-based protocol, the delay that a message undergoes is not deterministic, since it would depend on whether the other nodes are trying to send messages at that time or not. Further, in a contention-based protocol, as the network traffic increases, the number of contentions increase. This in turn causes the throughput to degrade dramatically. A positive aspect of the contention-based protocols is their simplicity and relative easy of implementation.

Contention-free protocols overcome many of the shortcomings of the contention-based protocols. Packets in a contention-free protocol are not delayed on account of collisions. However, contention-free protocols are difficult to implement in a multi-hop ad hoc network. We therefore do not discuss contention-free ad hoc network protocols. In general, contention-free protocols for wireless networks can be both centralized and distributed. For instance, Bluetooth piconet uses a contention-free centralized MAC protocol in which time slots are divided among terminals by the master. Also, cell-based, single-hop, mobile phone networks use centralized contention-free protocols. Interestingly, ad hoc networks that have controlling masters use centralized contention-free protocols. Therefore, Bluetooth scatternet would also qualify as a contention-free centralized MAC protocol.

Multi-access protocols do not need a centralized master but the neighbouring nodes must agree on the specific resource (time, frequency, or code) to use, so that there is no conflict with the other nodes. The problem of assigning resource (time, frequency, code) to different pairs of nodes in a multi-access protocol, is equivalent to the well-studied problem in graph theory called the *edge colouring problem* in which no two edges (links) can be assigned the same colour (out of a finite set of colours) if they share a common vertex. We will not discuss the details of any MAC protocol and restrict our discussions to the conceptual level only.

Much of the research done for MANET may also apply to sensor networks since both operate as multi-hop wireless networks with power constraints. MANET, however, focuses on device mobility, while sensor networks normally have limited or no mobility. Long studied problems in wireless networks, such as the hidden terminal and exposed node problem, also exist in sensor networks, so protocol designers must handle these issues in addition to the characteristics unique to sensor networks. Researchers now have the challenge to solve existing problems from traditional wireless networks under the constraints introduced by the limited resources available in WSN. In this, nodes define common active/sleep periods. The active periods are used for communication and the sleep ones for saving energy. This approach requires that nodes maintain a certain level of synchronization to keep the active/sleep periods common to all nodes. During the active periods, nodes contend for the medium using

CSMA, IEEE 802.11 DCF, etc. However, the use of common active/sleep periods may not be suitable for applications with irregular traffic, because nodes use contention inside active periods, which would be prohibitive when nodes wake up without communicating, and may cause collisions when there is high traffic that cannot be absorbed by the initially envisaged size of the active periods.

SUMMARY

A MAC protocol regulates the use of a shared physical channel among a set of nodes. It arbitrates among the contending nodes. We discussed the three important categories of solutions: fixed-assignment, random access, and reservation-based. Protocol design for mobile ad hoc networks is much more complex than that for wireless LANs, since nodes in such networks are constrained by factors such as low power, limited bandwidth, intermittent link errors, and hidden and exposed terminal problems. The IEEE 802.11 protocol has been implemented in low cost network cards and access points, and has become widely accepted for the design of Wireless LANs to provide Wi-Fi hot spots.

FURTHER READINGS

Deng, J. and Z. Haas, "Dual Busy Tone Multiple Access (DBTMA): A New Medium Access Control for Packet Radio Networks", Florence, Italy, 1998.

Dam, T.V. and K. Langendoen, "An Adaptive Energy-efficient MAC Protocol for Wireless Sensor Networks", in *SenSys '*03, New York, USA: ACM Press, pp. 171–180, 2003.

Heidemann, J., W. Ye, and D. Estrin, "Medium Access Control with Co-ordinated Adaptive Sleeping for Wireless Sensor Networks", *IEEE/ACM Transactions on Networking*, 12, No. 3, pp. 493–506, 2004.

Karn, P., "A New Channel Access Method for Packet Radio", *ARRL/CRRL Amateur Radio 9th Computer Networking Conference*, Sept. 22, 1990.

Lichun Bao, J.J. Garcia-Luna-Aceves, "Hybrid Channel Access Scheduling in Ad hoc Networks", *Proceedings of the 10th IEEE International Conference on Network Protocols*, 2002.

Rhee, I., A. Warrier, M. Aia, and J. Min, "ZMAC: a Hybrid MAC for Wireless Sensor Networks", New York, USA: ACM Press, 2005, pp. 90–101.

Rom, R. and M. Sidi, "Multiple Access Protocols—Performance and Analysis", Springer-Verlag, New York, 1990.

Singh, S. and C.S. Raghavendra, "PAMAS: Power-Aware Multi-Access Protocol with Signalling for Ad hoc Networks", *ACM Computer Communication Review*, July 1998.

Sunil Kumar, Vineet S. Raghavan, and Jing Deng, "Medium Access Control Protocols for Ad hoc Wireless Networks: A Survey", *Ad Hoc Networks, Elsevier Science*, Vol. 4, 2006, pp. 326–358.

Talucci, F., M. Gerla, and L. Fratta, "A Receiver Oriented Access Protocol for Wireless Multihop Networks", *Proceedings of IEEE PIMRC*, 1997.

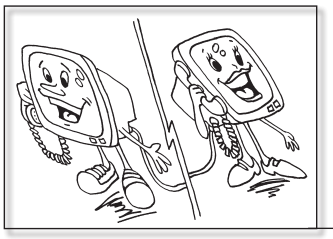
Toh, C.K., *Ad Hoc Wireless Mobile Wireless Networks*, Prentice Hall PTR, 2002.

Vaduvur Bharghavan, Alan Demers, Scott Shenker, and Lixia Zhang, "MACAW: A Media Access Protocol for Wireless LANs", *ACM SIGCOMM Computer Communication, Review*, Vol. 24, Issue 4, October 1994.

EXERCISES

1. What is the role of a MAC protocol? At which ISO/OSI layer does it operate?
2. What is a hidden terminal? What problem does it create during wireless communications? Explain your answer using a suitable schematic diagram.
3. When does the exposed terminal problem arise? Explain your answer using a suitable example.
4. What are the principal responsibilities of the MAC protocols? How do MAC protocols for wireless networks differ from those in wired networks?
5. What are the broad categories of MAC protocols? Name one popular protocol from each of these categories.
6. Explain the working of a contention-based MAC protocol. Give two examples of contention-based MAC protocols.
7. What are the different categories of MAC protocols. Identify the situations under which protocols from one category would be preferable over the other categories. Explain the working of a reservation-based MAC protocol.
8. Why are collision-detection based protocols not suitable for wireless networks?
9. Name one MAC protocol that is used in mobile ad hoc networks. Briefly explain its working.

10. Name one MAC protocol that is used in sensor networks. Briefly explain its working.
11. What is MACA protocol? In which environment is it suitable? Briefly explain its working. How does MACA protocol solve the hidden/exposed terminal problems?
12. What do you mean by a schedule-based MAC protocol? Name a schedule-based MAC protocol. Briefly explain its working.
13. What is FDMA? Briefly explain its working and at least one of its important applications.
14. What is TDMA? Briefly explain its working and at least one of its important applications.
15. Explain the basic scheme of the CDMA protocol. What is the role of a pseudorandom sequence generator in the working of the CDMA protocol?
16. Briefly explain the IEEE 802.11 standard and discuss its application.
17. Do you agree with the following statement: "In CSMA/CD protocol, when two nodes transmit on a shared medium, a collision can occur only when two nodes start transmitting exactly at the same time instant." Explain your answer.
18. Identify the specific reasons as to why the MAC protocols designed for infrastructure-based wireless networks may not work satisfactorily in infrastructure-less environments.



4

Mobile Internet Protocol

The Internet is built on top of a collection of protocols, called the TCP/IP protocol suite. Transmission Control Protocol (TCP) and Internet Protocol (IP) are the core protocols in this suite. IP is responsible for routing a packet to any host, connected to the Internet, uniquely identified by an assigned IP address. This raises one of the most vexing issues caused by host mobility. In the traditional IP addressing scheme, each LAN is assigned an address. The nodes in the LAN are assigned an address based on the LAN address. In the traditional IP addressing scheme, when a host moves to a different location, it may move to another network. As a result, it needs to change its IP address. This is an unworkable proposition for routing messages to a host, as it would keep changing its address as it moves from one network to another.

In this context, a Mobile Internet Protocol (Mobile IP) was proposed by the Internet Engineering Task Force (IETF). The mobile IP allows mobile computers to stay connected to the Internet regardless of their location and without changing their IP address. In other words, Mobile IP is a standard protocol that extends the Internet Protocol by making mobility transparent to applications and to higher level protocols like TCP.

Every mobile user likes to have a continuous network connectivity irrespective of its physical location. The traditional IP does not support user mobility. Mobile IP was created by extending IP to enable users to keep the same IP address while travelling to a different network. Box 4.1 explains how the necessity of mobile IP arose with the introduction of mobile computers to the Internet.

4.1 Mobile IP

The packet delivery to and from a mobile node has been schematically shown in Fig. 4.1. In this figure, a correspondent node (CN) is connected

BOX 4.1 Evolution of mobile IP

The IP defined in RFC 791 is the widely-used version of the Internet Protocol. Interestingly, however, the present form of IP is not version 1 of IP but version 4. The TCP/IP suite evolved through three earlier versions, and was split into TCP and IP layers in version 4. So, when you use IP today, you are referring to IP version 4 (IPv4). It is therefore usual that "IP" means "IP version 4. Despite the careful design of IPv4, it was later recognized that IP suffers from several shortcomings. For example, it would not be able to support the enormous number of users that are expected to use Internet in a couple of years. Also IP does not distinguish among the different applications, and treats all applications equally. However, the quality of service (QoS) requirement of different applications may be different. For example, a streaming video requires that video frames be transmitted without delay jitters, whereas applications such as e-mail can tolerate considerable delay. This needed the development of a new version of IP. In fact, a new version of IP has already been developed. It is formally called Internet Protocol version 6 (IPv6) and also sometimes referred to as IP Next Generation or IPng.

The Internet was started at a time when mobile computers did not exist. As a result, the present Internet lacks support for mobile users travelling across the world. To address this issues, mobile IP was proposed. Mobile IP (MIP) supports mobility at Internet Layer (Network Layer).

via a router to the Internet, and the home network and the foreign network are also connected via a router, i.e. the home agent (HA) and foreign agent (FA), respectively, to the Internet. Therefore, home agent (HA) is implemented on the router connecting the home network with the Internet, a foreign agent (FA) is also implemented on the router connecting the foreign network with the Internet. The tunnel for the packets towards the mobile node starts at the home agent and ends at the foreign agent, again here the foreign agent has the care-of-address (COA).

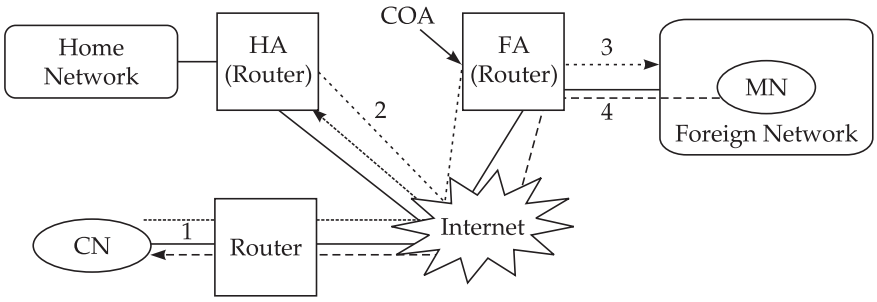


Figure 4.1 *Packet delivery to and from a mobile node.*

It can be observed in Fig. 4.1 that several terminologies are associated with mobile IP. We explain these terminologies in the following subsection.

4.1.2 Terminologies—Mobile IP

Various terminologies associated with mobile IP and used in Fig. 4.1 are explained below:

Mobile Node (MN): A mobile node is a hand-held equipment with roaming capabilities. It can be a cell phone, personal digital assistant, laptop, etc.

Home Network: The home network of a mobile device is the network within which the device receives its identifying IP address (home address). In other words, a home network is a subnet to which a mobile node belongs to as per its assigned IP address. Within the home network, there is no need of mobile IP.

Home Address (HA): The home address of a mobile device is the IP address assigned to the device within its home network.The IP address on the current network is known as home address.

Foreign Agent (FA): The foreign agent is a router in a foreign network that functions as the point of attachment for a mobile node when it roams to the foreign network. The packets from the home agent are sent to the foreign node which delivers it to the mobile node.

Foreign Network: The foreign network is the current subnet to which the mobile node is visiting. It is different from home network. In other words, a foreign network is the network in which a mobile node is operating when away from its home network.

Correspondent Node (CN): The home agent is a router on the home network serving as the anchor point for communication with the mobile node. It tunnells packets from a device on the Internet, called a correspondent node (CN), to the roaming mobile node.

BOX 4.2 Tunnelling process

The packet is forwarded by the home agent to the foreign agent. When the packet comes to the foreign agent (care-of-address), it delivers the packet to the mobile node. This process is called *tunnelling*. Tunnelling has two primary functions: encapsulation of the data packet to reach the tunnel endpoint, and decapsulation when the packet is delivered at that endpoint.

Care-of-Address (COA): It is the address that is used to identify the present location of a foreign agent. The packets sent to the MN are delivered to COA.

The COA can be any of the following two types:

(a) *Foreign agent COA:* The COA is an IP address of foreign agent (FA).
(b) *Co-located COA:* When the mobile node (MN) acquires a temporary IP address, that address acts as the COA.

BOX 4.3 Care-of-Address (COA)

In real life, if a person is not in his own house and living in a temporary location, that location is called his care-of-address (C/O) but, here, the care- of-address defines the current location of the mobile node from an IP point of view. All IP packets sent to mobile nodes are delivered to the care-of-address (COA), i.e. not directly to the IP address of the mobile node.

Note: The co-located address (temporary IP address) can be acquired using services like dynamic host configuration protocol (DHCP).

Home Agent (HA): It is located in home network and it provides several services for the MN. HA maintains a location registry. The location registry keeps track of the node locations using the current care-of-address of the MN.

Agent Discovery: During call establishment it is necessary for a mobile node to determine its foreign agent. This task is referred to as *agent discovery*. The following two discovery methods are popularly used: (1) Agent advertisement, and (2) Agent solicitation.

In the following, we briefly explain these two agent discovery methods.

1. *Agent advertisement:* Generally the foreign and the home agents advertise their presence through periodic agent advertisement messages. An agent advertisement message, lists one or more care-of-addresses and a flag indicating whether it is a home agent or a foreign agent. Agent advertisement is a popularly used method in agent discovery.

2. *Agent solicitation:* In case a mobile node (MN) does not receive any COA, then the MN should send an agent solicitation message. But it is important to monitor that these agent solicitation messages do not flood the network. A mobile node can usually send up to three solicitation messages (one per second) as soon as it enters a new network. The basic purpose of the solicitation messages sent by a mobile node (MN) is to search for a foreign agent (FA). For a highly dynamic wireless network in which MNs move at great speed, even a time interval of the order of a second between these messages is too long. If an MN does not receive any address in response to its solicitation messages, then to avoid network flooding, the MN should exponentially reduce the rate of sending the solicitation messages.

Tunnelling and encapsulation

Tunnelling establishes a virtual pipe for the packets available between a tunnel entry and an endpoint. Tunnelling is the process of sending a packet via a tunnel and it is achieved by a mechanism called encapsulation.

Encapsulation refers to arranging a packet header and data in the data part of the new packet. On the other hand, disassembling the data part of an encapsulated packet is called decapsulation. Whenever a packet is sent from a higher protocol layer to a lower protocol layer, the operations of encapsulation and decapsulation usually take place.

4.2 Packet Delivery

Let us consider the situation, where the corresponding node (CN) wants to send an IP packet to a mobile node. CN sends the packet to the IP address of the mobile node as shown in step 1 of Fig. 4.1. The IP address of the MN is the destination address, whereas the address of CN is the source address.

The packet is passed to the Internet that does not have any information about the MN's current location. So the Internet routes the packet to the router of the MN's home network. The home agent examines the packet to determine whether the MN is present in its current home network or not. In case that MN is not present, then the packet is encapsulated by a new header that is placed in front of the existing IP header. The encapsulated packet is tunnelled to the COA, which act as the new destination address and the HA acts as the source address of the packet as shown in step 2 of Fig. 4.1.

The encapsulated packet is routed to the foreign agent which performs decapsulation to remove the additional header and forwards the decapsulated packet to the MN, which is the actual destination, as specified by the source node (CN), shown in step 3 of Fig. 4.1.

The MN after receiving the packet from CN, forwards a reply packet to the CN by specifying its own IP address along with the address of the CN as shown in step 4 of Fig. 4.1. The MN's IP address acts as the source address and the CN's IP address acts as the destination address. The packet is routed to the FA. After receiving the packet, FA forwards the packet to CN.

4.3 Overview of Mobile IP

The goal of mobile IP is to enable packet transmission efficiently without any packet loss and disruptions in the presence of host and/or destination mobility.

Consider a scenario adapted from that discussed by Kozierok[†] to explain the mobile IP. Suppose a person working as a business development executive for a company needs to take care of many regional offices in India

†Kozierok, M. Charles, *TCP/IP Guide: A Comprenhensive, Illustrated Internet Protocol Reference*, Starch Press, 2005.

and abroad. His home office is in Delhi where he spends about 40% of his time. The rest of the time he spends between the other offices, say, Kolkata, Mumbai, Chennai, Kathamandu and Singapore.

A problem that arises in this context is: how does he make arrangements so that he would continue to receive postal mails regardless of his location? If we can answer this, we can easily understand how IP works in the context of a mobile device.

There are two broad categories of solutions to this problem being faced by the business executive: (i) address changing, (ii) decoupling mail routing from his address. It would be difficult for the business development executive to inform about his changed address to all those who are likely to write letters to him each time he moves. Also, by the time, he would have informed everyone about his new address, it would have become time for the address to change again. And he certainly cannot decouple the routing of mail from his address, unless he has set up his own personal postal system.

A practical solution to this problem is mail forwarding. Let us say that he leaves Delhi for Singapore for a couple of months. He will inform the Delhi post office that he will be in Singapore. The Delhi post office would intercept his mails headed for his normal Delhi address, relabel them, and forward them to Singapore. Depending on where he is staying, this mail might be redirected either straight to a new address in Singapore, or to a Singapore post office where he can pick it up. If he leaves Singapore to go to another city, say, Kathamandu, he would just call the Delhi post office and tell them about his new location. When he gets back to home office, he will cancel the forwarding arrangement and get his mail as usual. The advantages of this system are many. It is a relatively simple mechanism to understand and implement. A positive aspect of this scheme is that it is transparent to everyone sending mails; they still send mail to him at his Delhi address and the mail reaches wherever he goes. The handling of the forwarding process is taken care by the Delhi post office as well as the post office where he presently gets located; the rest of the postal system does not even know that anything out of the ordinary is going on.

There are some disadvantages of this approach too. The Delhi post office may allow occasional forwarding for free, but would probably charge a fee if the business development executive needs this service on a regular basis. He may also need a special arrangement in the city where he travels to. He has to keep communicating with his home post office each time he moves. And perhaps most importantly, every piece of mail has to be sent through the system twice—first to Delhi and then to wherever he moves, which is inefficient and introduces additional delay in delivering and also loads the postal system.

Mobile IP works in a manner very similar to the postal mail forwarding system. Each network can be considered like a different "city", and the internetwork of routers is like the postal system. The router that connects any network to the Internet is like that network's "post office", from an IP perspective.

The mobile node is normally resident on its home network, which is the one indicated by the network ID in its IP address. Devices on the internetwork always route using this address, so the pieces of "mail" (datagrams) always arrive at a router at the device's "home". When the device "travels" to another network, the home router ("post office") intercepts these datagrams and forwards them to the device's current address. It may send the datagrams straight to the device using a new, temporary address, or it may send them to a router on the device's current network (the "other post office", Singapore) for final delivery. The mobile node's home router serves as the home agent and the router in Singapore as the foreign agent. The mobile has been assigned a temporary "care-of address" to use in Singapore (which in this case is a co-located care-of-address). As per mobile IP terminology, the home agent tunnells the packet to the COA.

BOX 4.4 An analogy with Mobile IP

Consider a second analogy with Mobile IP. Suppose you are moving from one residential flat to another. What do you do?

- Leave a forwarding address with your old post office.
- The old post office forwards mail to your new post office, which then delivers it to you.

The steps used in the operation of mobile IP are the following:

Step 1: The remote client sends a datagram to the MN using its home address. It reaches the home agent (say Delhi) as usual.

Step 2: The home agent encapsulates that datagram in a new packet and sends it to the foreign agent (say Singapore).

4.4 Desirable Features of Mobile IP

Some of the features required of mobile IP are the followings.

Transparency: A mobile end-system should continue to keep its IP address and there should not be any disruption of communication after any movement. In other words, the IP address is to be managed transparently and there should not be any effect of mobility on any ongoing communication.

Compatibility: Mobile IP should be compatible with the existing Internet protocols.

Security: Mobile IP should, as far as possible, provide users with secure communications over the Internet.

Efficiency and Scalability: In the event of worldwide support, there can be a large number of mobile systems in the whole Internet. This should neither result in large number of messages nor should it incur too much computational overhead. It should also be scalable to support billions of moving hosts worldwide.

4.5 Key Mechanism in Mobile IP

Mobile IP is associated with the following three basic mechanisms:

- Discovering the care-of-address
- Registering the care-of-address
- Tunnelling to the care-of-address

A schematic diagram of Mobile IP is shown Fig. 4.2. The specific protocols used by the basic mechanisms have also been shown. Observe that the registration process works over UDP and the discovery protocol over ICMP.

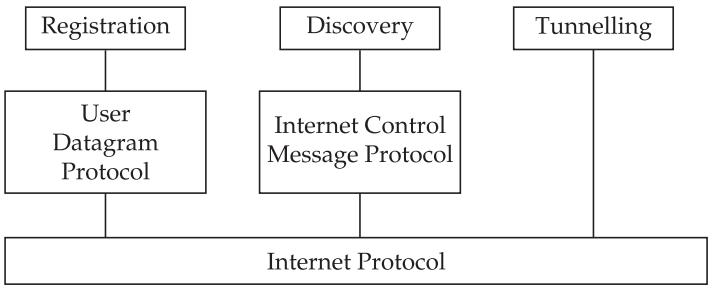


Figure 4.2 *A schematic model of Mobile IP.*

Discovering the care-of-address

Each mobile node uses a discovery protocol to identify the respective home and foreign agents. The discovery of the care-of-address consists of four important steps.

1. Mobile agents advertise their presence by periodically broadcasting the *agent advertisement* messages.
2. The mobile node receiving the *agent advertisement* message observes whether the message is from its own home agent and determines whether it is on the home network or on a foreign network.
3. If a mobile node does not wish to wait for the periodic advertisement, it can send out *agent solicitation* messages that will be responded to by a mobility agent.

The process of agent advertisements, involves the following activities:

- Foreign agents send messages to advertise the available care-of-addresses.
- Home agents send advertisements to make themselves known.
- Mobile hosts can issue agent solicitations to actively seek information.
- If a mobile host has not heard from the foreign agent to which its current care-of-address belongs, it takes up another care-of-address.

Registering the care-of-address

If a mobile node discovers that it is on the home network, it operates without requiring any mobility services.

If a mobile node obtains a care-of-address from a foreign agent, then this address should be registered with the home agent. The mobile node sends a request for registration to its home agent along with the care-of-address information whenever the home agent receives the registration request information. The routing table is updated and it sends back the registration reply to the mobile node. The mobile node makes use of the registration procedure to intimate the care-of-address to a home agent. These steps are schematically shown in Fig. 4.3. We now elaborate these different steps of the registration process.

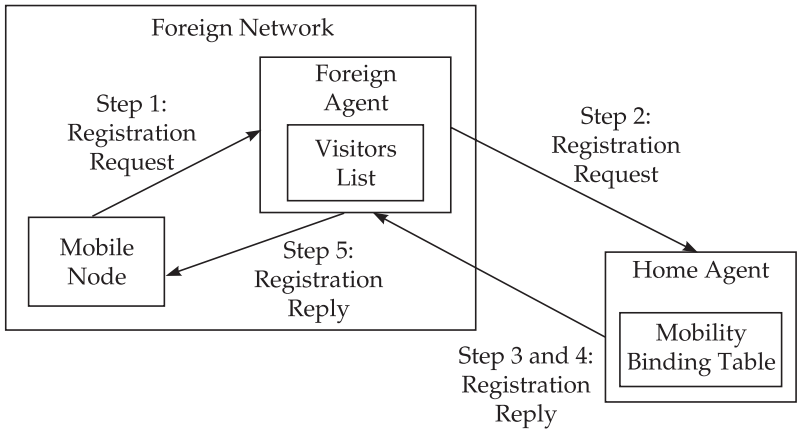


Figure 4.3 *Registration process in Mobile IP.*

The registration process shown in Fig. 4.3 consists of the following steps:

1. If the mobile node is on a new network, it registers with the foreign agent by sending a *registration request* message which includes the permanent IP address of the mobile host and the IP address of its home agent.
2. The foreign agent in turn performs the registration process on behalf of the mobile host by sending a Registration Request containing the

permanent IP address of the mobile node and the IP address of the foreign agent to the home agent.
3. When the home agent receives the Registration Request, it updates the mobility binding by associating the care-of-address of the mobile node with its home address.
4. The home agent then sends an acknowledgement to the foreign agent.
5. The foreign agent in turn updates its visitors list by inserting the entry for the mobile node and relays the reply to the mobile node.

Box 4.5 Security in Mobile IP

Security is very important in Mobile IP as mobile nodes are often connected to the Internet via wireless links which are very vulnerable to security attacks. For example, during the registration procedure the home agent should be convinced that it is getting the authentic registration request from a mobile node. Mobile IP solves this problem by specifying a security association between the home agent and the mobile node.

Tunnelling to the care-of-address

Tunnelling takes place to forward an IP datagram from the home agent to a care-of-address. This involves carrying out the following steps:

- When a home agent receives a packet addressed to a mobile host, it forwards the packet to the care-of-address using IP-within-IP (encapsulation).
- Using IP-within-IP, the home agent inserts a new IP header in front of the IP header of any datagram.
- Destination address is set to the care-of-address.
- Source address is set to the home agent's address.
- After stripping out the first header, IP processes the packet again.

The tunnelling operation in mobile IP and IP-within-IP encapsulation (embedding) are shown in Fig. 4.4.

4.6 Route Optimization

In the mobile IP protocol, all the data packets to the mobile node go through the home agent. Because of this there will be heavy traffic between HA and CN in the network, causing latency to increase. Therefore, the following route optimization needs to be carried out to overcome this problem.

- Enable direct notification of the corresponding host
- Direct tunnelling from the corresponding host to the mobile host
- Binding cache maintained at the corresponding host

The mobile IP scheme needs to support the four messages shown in Table 4.1.

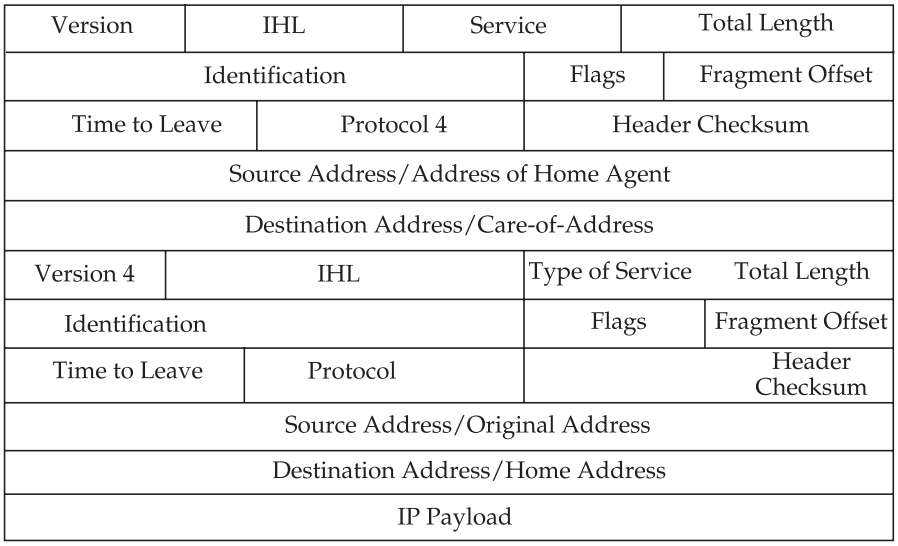


Figure 4.4 *IP encapsulation in mobile IP.*

The association of the home address with a care-of-address is called binding.

TABLE 4.1 Messages Transmitted in Optimized Mobile IP

Message type	*Description*
1. Binding request	If a node wants to know the current location of a mobile node (MN), it sends a request to home agent (HA).
2. Binding acknowledgement	On request, the node will return an acknowledgement message after getting the binding update message.
3. Binding update	This is a message sent by HA to CN mentioning the correct location of MN. The message contains the fixed IP address of the mobile node and the care-of-address. The binding update can request for an acknowledgement.
4. Binding warning	If a node decapsulates a packet for a mobile node (MN), but it is not the current foreign agent (FA), then this node sends a binding warning to the home agent (HA) of the mobile node (MN).

4.7 Dynamic Host Configuration Protocol (DHCP)

DHCP was developed based on bootstrap protocol (BOOTP). DHCP provides several types of information to a user including its IP address. To manage dynamic configuration information and dynamic IP addresses, IETF

standardized an extension to BOOTP known as dynamic host configuration protocol (DHCP). The DHCP client and server work together to handle the roaming status and to assign IP address on a new network efficiently. The DHCP server allocates an IP address from a pool of IP addresses to a client.

BOX 4.6 The BOOTP protocol

The BOOTP protocol is used for booting (starting) computers from the network. These are popularly used in case of diskless computers. Whenever a client requests an IP address from the server machine, BOOTP searches a table which matches to its physical address. The BOOTP protocol does not handle the mobility-related issues since it cannot address the following problems:

- When a host moves from one network to another network.
- When a host seeks a temporary Internet Protocol (IP) address.

4.7.1 Significance of Dynamic Host Configuration Protocol

DHCP is an extension to the BOOTP and compatible with it. For example, if a host is running BOOTP, it can also request configuration (example: static configuration) from a DHCP server node. The importance of DHCP in a mobile computing environment is that it provides temporary IP addresses whenever a host moves from one network to another network.

DHCP supports the following three important mechanisms for IP address allocation:

Automatic allocation: In automatic allocation, DHCP assigns a permanent IP address to a particular client.

Dynamic allocation: In dynamic allocation, DHCP assigns IP address to a client for a specific period of time.

Manual allocation: In manual allocation, a client's IP address is assigned by the network administrator, where the DHCP is used to inform the address assigned to clients.

SUMMARY

It is necessary to maintain the same IP addresses of a mobile host even when it moves. The traditional IP design does not support mobility. So, whenever a computer changes its location, it would need a new IP address to be set on it according to the network address of the new network to which it has moved in. This is inconvenient, but once given a new IP, there would be no way of receiving the messages sent to its old IP address.

Mobile Internet Protocol (Mobile IP) has become a standard protocol to allow users to maintain connectivity with their home IP addresses regardless of their physical movement. Without getting into too much details, this chapter gives an overview of the working of mobile IP.

FURTHER READINGS

Jiannong Cao, Liang Zhang, S.K. Das, and Henry Chan, "Design and Performance Evaluation of an Improved Mobile IP Protocol," *INFOCOM* 2004.

Johnson, D., C. Perkins, and J. Arkko, "Mobility Support in IPv6," Request for Comments (Proposed Standard) 3775, Internet Engineering Task Force, June 2004. Available online at http://www.rfc-editor.org/rfc/rfc3775.txt.

Liu Yu, Ye Min-hua, and Zhang Hui-min, "The handoff schemes in mobile IP," Vehicular Technology Conference, *57th IEEE Semiannual Volume 1*, 22–25 April 2003, pp. 485–489.

Perkins, E. Charles, "IP Mobility Support for IPv4," Request for Comments (Proposed Standard) 3344, Internet Engineering Task Force, August 2002. Available online at http://www.ietf.org/rfc/rfc3344.txt.

Perkins, E. Charles, "Mobile IP," *Communications Magazine*, IEEE Vol. 40, Issue 5, May 2002, pp. 66–82.

Rajeev Koodli, "Fast Handovers for Mobile IPv6," Internet Draft, Internet Engineering Task Force, draft-ietf-mipshop-fast-mipv6-03.txt, October 2005. Available online at http://www.ietf.org/internet-drafts/draft-ietf-mipshop-fast-mipv6-03.txt.

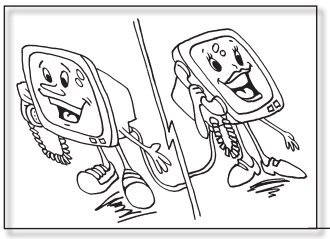
Rajeev Koodli and Charles E. Perkins, "Mobile IPv4 Fast Handovers," Internet Draft, Internet Engineering Task Force, draft-ietf-mip4-fmipv4-00.txt, February 2006. Available online at http://www.ietf.org/internet-drafts/draft-ietf-mip4-fmipv4-00.txt.

Stefan Raab and Madhavi W. Chandra, *Mobile IP Technology and Applications*, Cisco Press, 2005.

EXERCISES

1. Explain why the traditional IP cannot be used in a mobile network. What are the main differences between the traditional IP and the mobile IP? How does mobile IP support mobile hubs?
2. Explain the limitations of IPv4 and how are they overcome by IPv6.

3. Explain the following terms associated with mobile IP:
   (a) Home Address
   (b) Mobile Node
   (c) Foreign Agent
   (d) Foreign Network
   (e) Home Network
4. Write short notes on the following:
   (a) Correspondent Node
   (b) Care-of-Address
   (c) Agent Discovery
   (d) Tunnelling and Encapsulation
5. Explain the operation of mobile IP with the help of a suitable schematic diagram and by using suitable examples.
6. What are the disadvantages of mobile IP?
7. Explain the discovery of care-of-address in the context of movement of a mobile to a foreign network.
8. Explain the agent advertisement procedure of mobile IP.
9. What do you mean by agent solicitation? Why are agent advertisement messages needed?
10. What do you mean by encapsulation and decapsulation in the context of mobile IP? Explain why are these needed.
11. Give a brief account of route optimization in mobile IP.
12. What do you mean by the term *binding* of a mobile node?
13. What are the main functions of DHCP? Why is DHCP needed? Can it be used when nodes are mobile? Explain your answer.
14. Explain how mobile IP is different from DHCP.
15. State some applications of DHCP.
16. Explain the significance of Dynamic Host Configuration Protocol. Give examples of situations where it is useful.



5

Mobile Transport Layer

In mobile computing applications, Transmission Control Protocol (TCP) is possibly the most popular transport layer protocol. In fact, TCP is the *de facto* standard transport layer protocol for applications that require guaranteed message delivery. TCP is a connection-oriented protocol. UDP (User Datagram Protocol), on the other hand, is a connectionless protocol in the TCP/IP protocol suite and does not guarantee reliable data delivery. However, when the traditional TCP is used in mobile computing networks, it operates in a highly inefficient and unsatisfactory manner. TCP, therefore, needs several special adaptations to make it suitable for use in wireless networks. This issue forms the focus of the discussions in this chapter.

BOX 5.1 A rough analogy of connectionless transmission

A rough analogy of connectionless transmission is like posting a long letter by splitting it into several small parts, writing these parts using a number of postcards, assigning sequence numbers to postcards, and posting these in a post box. At the receiver-end, some postcards might be received out of sequence, a few may arrive at the destination after undue delay, and some postcards might even get misplaced before they arrive at the destination. On the other hand, in a connection-oriented protocol, explicit connection is set up between a sender and a receiver before a message is split into parts and sent across. This connection establishment enables the receiver to know beforehand how many postcards are expected to arrive, and based on this information, request for any missing postcards can be issued. Thus, it can be guaranteed that the message will be fully reconstructed at the receiver end.

In this chapter, we first provide a brief overview of the TCP/IP protocol suite and identify the factors that render TCP operation in mobile wireless networks unsatisfactory. Subsequently, we outline how TCP has been extended to make it suitable to work in mobile wireless networks.

5.1 Overview of TCP/IP

The TCP/IP protocol suite was developed by DARPA in 1969 to provide seamless communication services across an internetwork consisting of a large number of different networks. The TCP/IP protocol suite is a collection of a large number of protocols. The protocol suite has been named after the two most important protocols of the protocol suite: Transmission Control Protocol (TCP) and Internet Protocol (IP). A few important protocols in the TCP/IP suite and the specific protocol layer at which they operate have been depicted in Fig. 5.1. The diagram also depicts the specific lower layer protocols that are invoked by a protocol.

As shown in Fig. 5.1, the TCP/IP protocol stack consists of four layers of protocols. The four layers of the protocol are: Application layer, Transport layer, Internet layer, and Network interface layer. Of the four layers, TCP/IP does not define any specific protocol for the network interface layer, but allows any of the standard protocols to be used at this layer. For this reason, in Fig. 5.1 we do not show any specific protocol for the network interface layer.

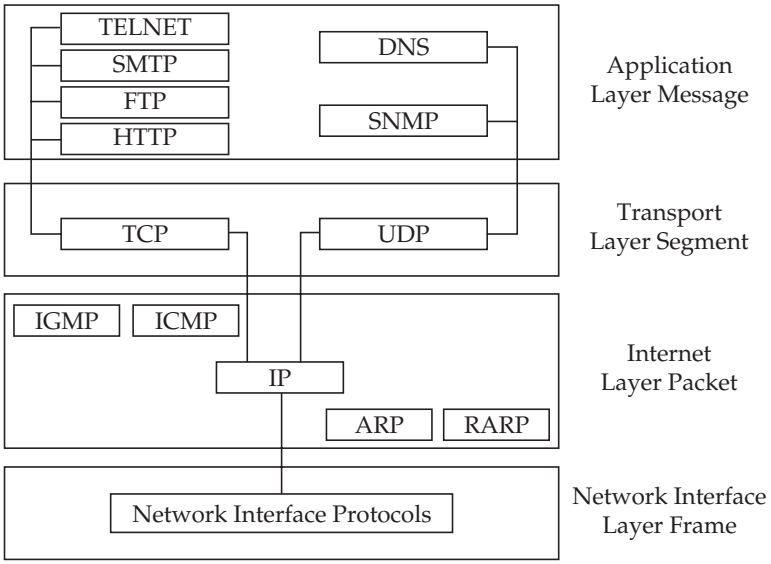


Figure 5.1 *TCP/IP protocol stack.*

The application programmers and end-users are mainly concerned with the application layer protocols. The application layer protocols, in turn, make use of the services provided by the lower layer protocols. An application layer protocol requiring to send a message to another application (that may possibly be running on a different host either in the same local network or in some remote network) makes use of a transport

layer protocol and passes it with the message to be transmitted. The specific transport layer protocol converts the message into small parts and attaches certain information to it. The transport layer protocol first converts a message into segments and passes these segments to the Internet layer protocol (IP). The IP layer protocol attaches certain information to the segments such as the destination host address to form *packets*. We can say that a TCP segment is carried in one or more IP packets. The IP passes on the packets to the network interface layer protocol, which in turn converts them to frames by adding certain additional information to the packets such as checksum and then transmits them on the network.

The reverse operation takes place when a frame arrives at a host. The network interface layer protocol removes the information added by the corresponding network interface layer protocol at the sender-end and passes on the packet to the IP layer. The IP layer protocol at the destination removes the information added by the IP layer at the sender's end and gets back the segments and passes these to the transport layer protocol. The transport layer protocol at the receiver strips the information added by the transport layer protocol at the sender, reconstructs the message and sends it to the application layer.

Note that the application layer deals with messages; the transport layer deals with segments; the internet layer deals with packets; and the data link layer deals with frames. We will discuss the details of the operations of a few important protocols of the TCP/IP suite in the subsequent chapters.

Over the last two decades, the Internet has seen almost exponential growth and now the Internet applications have become ubiquitous. The Internet-based applications are developed predominantly by using the client-server paradigm. In a typical Internet-based application deployment scenario, the server is an application providing certain services and a client that typically runs on a web browser, is primarily the requester of services. TCP has now become the *de facto* transport layer protocol for client-server communications.

5.2 Terminologies of TCP/IP

In the following, we briefly discuss a few of the protocols and terminologies associated with the TCP/IP protocol suite.

***TCP** (Transmission Control Protocol):* On the sending side, TCP is responsible for breaking a message into small parts, adding sequence numbers and certain other information and after this, making them known as segments. TCP passes the segments to the lower layer protocol for transmission over the network. While at the receiver's end, TCP assembles the segments when they arrive and reconstructs the message. TCP is a reliable protocol. Whenever a packet is lost or corrupted during transmission, TCP detects it

and requests the sender for retransmission. Thus, retransmission is used as the primary mechanism by TCP for reliable data delivery to the destination.

***IP** (Internet Protocol):* At the host machine of an application sending a message, IP is responsible for constructing packets (also called datagrams) from the segments it receives from the transport layer protocol by adding the destination host address and then passes these on to the lower layer protocol for transmitting. On the receiver's side, it deconstructs the segments and then passes these to the transport layer protocol.

***HTTP** (Hyper Text Transfer Protocol):* The HTTP protocol is used for communications between a web server and the client-side application running on a web browser.

***SMTP** (Simple Mail Transfer Protocol):* The SMTP protocol is used for sending and receiving e-mails by a mail client.

***MIME** (Multipurpose Internet Mail Extensions):* The MIME protocol lets the SMTP encode multimedia files such as voice, picture, and binary data in e-mails and transmit them across TCP/IP networks. SMTP has been designed to handle only the text contents in e-mails. MIME helps e-mails to include non-text contents such as picture, voice, and binary data files by encoding the binary data in the ASCII text format.

***FTP** (File Transfer Protocol):* The FTP protocol is used to transfer files between the computers.

***SNMP** (Simple Network Management Protocol):* The SNMP protocol is used for administration and management of computer networks. The network manager uses tools based on this protocol to monitor network performance.

***ICMP** (Internet Control Message Protocol):* The ICMP protocol runs on all hosts and routers and is mainly used for reporting errors such as a non-reachable host.

***ARP** (Address Resolution Protocol):* The ARP protocol is used by IP to find the hardware address (also called the physical address) of a computer based on its IP address. The hardware (physical) address is stored in the ROM (Read Only Memory) of the computer's network interface card. It is also known as MAC (Media Access Control) address and also as an Ethernet hardware address (EHA).

***RARP** (Reverse Address Resolution Protocol):* The RARP protocol is used by IP to find the IP address based on the physical (MAC address) address of a computer.

***BOOTP** (Boot Protocol):* The BOOTP protocol is used for booting (starting) a diskless computer over a network. Since a diskless computer does not store the operating system program in its permanent memory, the BOOTP

protocol helps to download and boot over a network, using the operating system files stored on a server located in the network (discussed in Section 4.7).

Routers: A router is responsible for *routing* the packets that it receives to their destinations based on their IP addresses, possibly via other routers.

DNS: It stands for **D**omain **N**ame **S**ystem (or **S**ervice or **S**erver). It is a software service available on the Internet that is responsible for translating domain names into IP addresses. We use domain names while accessing any website since these are alphabetic character strings that are much easier to remember compared to the conventional IP address specification using dot-separated numerical values. Of course, when we specify a website (URL) using its domain name, a DNS service hosted on the Internet translates the domain name into the corresponding IP address, since, after all, the Internet works using IP addresses. For example, the domain name *www.iitkgp.ernet.in* might get translated by the DNS to *144.16.192.245.*

IP Addresses: Each computer must have an IP address before it can be meaningfully connected to the Internet. A packet gets routed to its destination based on its IP address.

***IGMP** (Internet Group Management Protocol):* The IGMP protocol is used by hosts to exchange information with their local routers to set up multicast groups. A setup of multicast groups allows efficient communication, especially for video streams and certain gaming applications. The routers also use the IGMP to check whether the members of a known group are active or not.

5.3 Architecture of TCP/IP

As already stated in Section 5.1, the TCP/IP protocol consists of four layers as shown in Fig. 5.2. These layers are: Application layer, Transport layer, Internet layer, and Network access layer. The functionalities of each of these layers are discussed below:

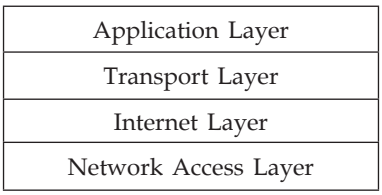


Figure 5.2 *TCP/IP protocol layers.*

Application layer: The protocols at this layer are used by applications to establish communication with other applications which may possibly be running on separate hosts. Examples of application layer protocols are http, ftp, and telnet.

Transport layer: It provides reliable end-to-end data transfer services. The term end-to-end means that the end points of a communication link are applications or processes. Therefore, sometimes protocols at this layer are also referred to as host-to-host protocols. Remember that there can be several applications or processes running on a host. Thus, to identify the end point, it is not only the computer that needs to be identified, but also the exact process or application that would receive the message needs to be identified. This is efficiently accomplished by using the concept of a port number. An application or a process specifies a port number on which it would receive a message. Once a message reaches a host, it is demultiplexed using the port number at the transport layer for delivery to the appropriate application. The transport layer provides its services by making use of the services of its lower layer protocols. This layer includes both connection-oriented (TCP) and connectionless (UDP) protocols.

Internet layer: The Internet layer packs data into data packets that are technically known as IP datagrams. Each IP datagram contains source and destination address (also called IP address) information that is used to forward the datagrams between hosts and across networks. The Internet layer is also responsible for routing of IP datagrams. In a nutshell, this layer manages addressing of packets and delivery of packets between networks using the IP address. The main protocols included at the Internet layer are IP (Internet Protocol), ICMP (Internet Control Message Protocol), ARP (Address Resolution Protocol), RARP (Reverse Address Resolution Protocol) and IGMP (Internet Group Management Protocol).

Network access layer: The functions of this protocol layer include encoding data and transmitting at the signalling determined by the physical layer. It also provides error detection and packet framing functionalities. As we will discuss later in Section 5.6, the functionalities of this layer actually consist of the functionalities of the two lowermost layers of the ISO/OSI protocol suite, namely data link and physical layers. The data link layer protocols help deliver data packets by making use of physical layer protocols. A few popular data link layer protocols are Ethernet, Token Ring, FDDI, and X.25. Ethernet is possibly the most common data link layer protocol. The physical layer defines how data is physically sent through the network, including how bits are electrically or optically signalled by hardware devices that interface with a network medium, such as coaxial cable, optical fibre, or twisted pair of copper wires.

5.4 An Overview of the Operation of TCP

When a client-server application runs on hosts that are wide apart, data transmission between the client and the server may span multiple networks. These networks are called sub-networks. For data routing, the Internet Protocol (IP) requires that each host in the network should have a unique address. Identification of hosts is not enough for data delivery, the packets must be forwarded to the exact application (or to a process in an application) requiring the packet. Within each host, every process is identified by a port number based on which the TCP can deliver data/ information to each relevant process.

BOX 5.2 Multiplexing and demultiplexing

A host can run many client and server applications. Therefore, these different applications can send/receive data concurrently and independently. As a result, data sent by different applications need to be *multiplexed* together, before these are sent on the network. Similarly, the TCP receives segments that may correspond to different applications running on a host. Therefore, on receiving a segment from its lower layer, TCP has to decide as to which application is the recipient. This is called *demultiplexing*. TCP performs multiplexing and demultiplexing by using *port* numbers.

Usually a message in the form of a block of data is passed to TCP by the sending application. The TCP breaks it into many small parts and attaches certain control information (called TCP header) to each small part. Each small part of the data along with the TCP header is called a segment (Fig. 5.3).

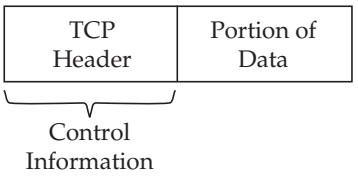


Figure 5.3 *The structure of a TCP segment.*

The TCP header includes several items of information including the following:

(i) Destination Port
(ii) Checksum
(iii) Sequence number

For a more detailed treatment of TCP operation, the reader is referred to Behrouz A. Forouzan[1].

1. Behrouz A. Forouzan, *Data Communication and Networking*, 4th ed., TMH, 2007.

IP datagram

An IP packet is also called a datagram. A datagram is of variable length which can be up to 65,536 bytes. It has two fields, namely header and data. The structure of an IP datagram is schematically shown in Fig. 5.4. In the following, we discuss some of the important fields of an IP datagram.

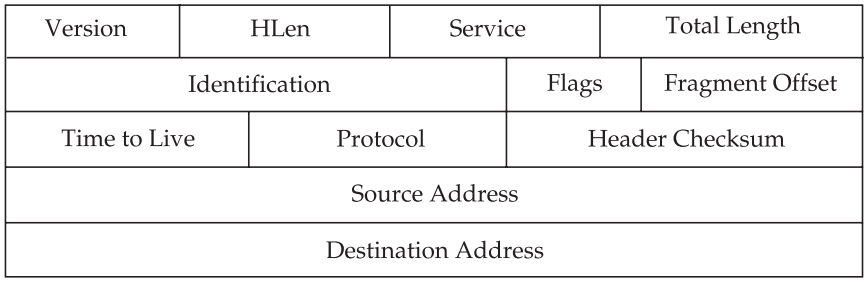


Figure 5.4 *IP datagram structure.*

Version (Ver): The IP version number is defined in this field, e.g. IPV4 or IPV6.

Header length (Hlen): It defines the header length as multiples of four bytes.

Service type: It has bits that define the priority of the datagram itself.

Total length: This field is allotted 16 bits to define the length of IP datagram.

Identification: It is mainly used to identify fragmentation that belongs to different networks. 16 bits are allotted for this job.

Flags: It deals with fragmentation of the data.

Fragmentation offset: It is a pointer to the offset of the data in the original datagram.

Time to live: This field is used to define the total number of hops that a datagram has to travel before discarding the operation.

Protocol: This field has 16-bits. It defines which upper layer protocol data is encapsulated at that time, say, for example TCP or UDP or ICMP, etc.

Header checksum: It has a 16-bit field to check the integrity of the packets.

Source address: It is a four byte ($4 * 8 = 32$) internet address to define the original source.

Destination address: It is a four byte ($4 * 8 = 32$) internet address to identify the destination of datagram.

The IP datagrams are sent to the link layer and become ready for transmission to the first sub-network in the path to its destination. The IP datagrams are also called packets.

Port address

In a client-server application, often the client and server programs are located on different host machines. The client program usually uses a temporary port number and the server program uses a well-known (or permanent) port number. These port numbers are used for identification of the application. A few well-known ports used by some of the popular TCP/IP protocols and by user applications are given in Table 5.1.

TABLE 5.1 A Few Commonly Used Well-known Port Numbers

Protocol	*Port*
TELNET	23
SMTP	25
RPC	111
DNS	53

Data encapsulation

When the TCP segments are handed over to Internet Protocol (IP layer), this layer appends an IP header containing the relevant control information. A segment after this additional control data is added, is called an IP datagram (see Fig. 5.4). The network access layer also appends its own header and now the segment becomes known as a frame/packet. The packet header includes important information such as the following:

(i) Facilities requests (such as priority)
(ii) Destination sub-network address

5.5 Application Layer Protocols of TCP

The application layer protocols of the TCP/IP protocol suite are shown pictorially in Fig. 5.5. The following are the three important application layer protocols.

Simple Mail Transfer Protocol (SMTP): It provides an 'electronic mail' function, that is used for transferring messages between different hosts. Originally, SMTP could handle text messages only. MIME helps transmit multimedia data within an e-mail by encoding the binary multimedia data in the ASCII format.

File Transfer Protocol (FTP): FTP is mainly used for transferring files from one host to another based on a user command. FTP allows both binary and text file transfers. Each FTP connection opens two TCP connections, one for data transfer and the other for transfer of control commands such as put, get, etc.

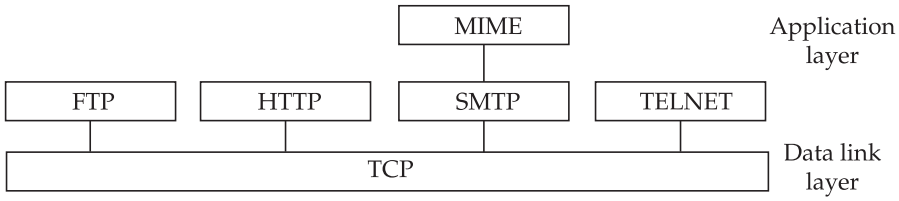


Figure 5.5 *Application layer protocols in TCP/IP protocol suite.*

TELNET: This application layer protocol lets users use a remote log-on facility, using which a user can log-on to a remote system. Both FTP and TELNET make use of the TCP layer. TCP forwards these data over the network by invoking the IP layer and the IP layer in turn invokes the link layer protocol. A problem with this type of transmission is that it becomes easy to sniff (secretly hear) the data by using the publicly available TCP sniffer programs such as "TCPdump". Due to this, at present, most users and applications use sftp (secure ftp) and ssh (secure shell) protocols. These protocols essentially serve to encrypt before passing the data on to the TCP layer. These protocols also perform decryption after receiving the data.

5.6 TCP/IP versus ISO/OSI Model

It is important to realize the difference between the TCP/IP protocol suite and the ISO/OSI model. As already mentioned, TCP/IP was developed by DARPA (Defence Advance Research Program of the US government) in 1969. The aim was to integrate all the computers efficiently. Thus TCP/IP networking predates ISO/OSI which was formulated in 1976. Naturally, TCP/IP makes no reference to ISO/OSI model and splits the network functionalities into layers quite differently.

A rough correspondence between TCP/IP protocol layers and ISO/OSI layers has been shown in Fig. 5.6. Observe that the Internet layer roughly corresponds to the network layer of the ISO/OSO model. The network access layer encompasses the data link and physical layers. The TCP/IP protocol suite does not define specific data link layer protocols to be used and can work on any data link protocol such as token ring and Ethernet.

5.7 Adaptation of TCP Window

The TCP primarily deploys a flow control technique to control congestion in a network. Traffic congestion occurs when the rate at which data is injected by a host into the network exceeds the rate at which data can be delivered to the network. A flow control technique helps adapt the rate

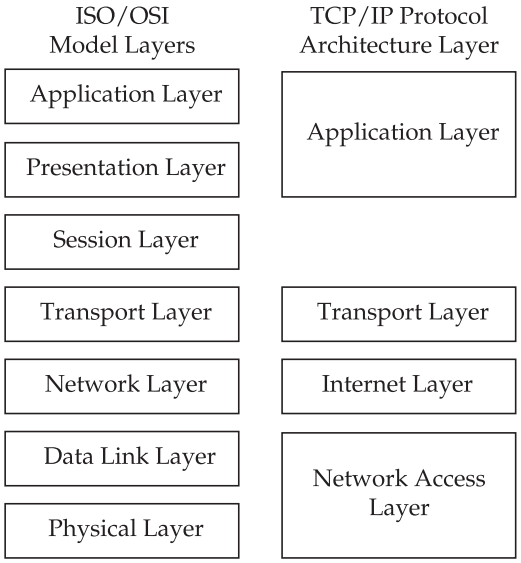


Figure 5.6 *A comparison of TCP/IP and ISO/OSI models.*

of data transmission by the TCP at the sending host end. The flow control technique helps to prevent the build-up of congestion in the network and at the same time helps to prevent buffer overrun at the slow receivers.

If data transmissions occur at a much faster rate than what the network infrastructure can comfortably support, then data packets get built up at the routers. When the buffers at routers start to overflow, the packets start getting lost. Additionally, if data transmissions by a sender take place at a much faster rate than what a slower receiver can handle, then the receiver's buffer starts to get flooded and hence the packets get lost. TCP handles both these causes of packet loss by reducing the rate at which data is transmitted at the sender's end. Thus, a receiver uses the flow control mechanism to restrict how fast a sender can transmit. However, to provide an acceptably fast data transmission service, once congestion disappears the transmission rate at the sender's end needs to be increased to a suitable value. Thus we can say that a flow control technique helps TCP dynamically adjust the transmission rate at the sender's end, reducing the transmission rate as congestion starts to develop and increasing it as congestion starts to disappear. The flow control mechanism deployed by TCP (called the sliding window protocol) is primarily based on the concepts of congestion window and advertised window. We now briefly describe these techniques.

When a sender starts to send data packets, the receiver indicates an advertised window (or receiver window) to the sender while sending acknowledgements. The advertised window is usually set equal to the size of the receive buffer at the receiver. The sender uses the advertised window size obtained from the receiver to determine the maximum amount of data

that it can transmit to the receiver without causing buffer overflow at the receiver. In other words, to prevent buffer overflow at the receiver, the data packets transmitted by a sender without having received acknowledgments for them should not exceed the size of the buffer available at the receiving end.

For each segment sent, a sending host expects to receive an acknowledgment. A congestion window indicates the maximum number of segments that can be outstanding without the receipt of the corresponding acknowledgement before the TCP at the sender's end pauses transmitting and waits for an acknowledgement to arrive. The TCP at a sender's end pauses if the number of segments for which the acknowledgement is outstanding becomes equal to the congestion window. A sender sets the congestion window size to 1 and keeps on increasing it until duplicate acknowledgements are received or until the number of outstanding packets becomes equal to the size of the advertised window.

Upon receipt of an acknowledgement, TCP detects packet loss using Retransmission timeout (RTO) and duplicate acknowledgements. After transmitting a segment, a TCP sender sets the retransmission timer for only one packet. If an acknowledgement for the packet is not received before the timer goes off, the packet is assumed to be lost. RTO is dynamically calculated. Timeouts can take too long. Again the TCP sender assumes that a packet loss has occurred if it receives three duplicate acknowledgements consecutively. In TCP, when a receiver does not get a packet that it expects in a sequence but gets an out of order packet, it considers that the expected packet might have got lost and it indicates this to the sender by transmitting an acknowledgment for the last packet that was received in order. Thus, three duplicate acknowledgement are also generated if a packet is delivered at least three places beyond its in-sequence location.

In wired networks, packet losses primarily occur on account of congestions encountered in the transmission path. However, in a wireless environment packet losses can also occur due to mobility and channel errors. In wired networks, bit errors are rare. On the other hand, wireless networks are vulnerable to noise. Noise can cause intermittent bit errors. Further, there can be intermittent disconnections due to fading and also due to obstructions that may be encountered by a mobile host. Further, packets may get lost during handoff. An intermittent disconnection may cause the TCP at the sender's end to time out for an acknowledgement and cause it to retransmit. This would cause unnecessary retransmissions to occur, even though the packet may be buffered at a router. Also, various additional causes of packet loss can result in a high rate of packet loss in a mobile wireless network. These losses would be interpreted by TCP as symptoms of congestion and force it to operate in slow start. This would cause the network to operate inefficiently and result in unacceptable slow data transmission.

BOX 5.3 Adaptive transmission control mechanism

When many packets are transmitted to a single receiver and the rate with which these packets are transmitted is higher than the processing rate of the destination host or an intermediate router, the buffers of the router (or the destination as the case may be) get filled quickly. This results in dropping packets at the affected router or the destination. When a sender realizes that some packets might have been dropped based on acknowledgements not arriving before timeout or based on the receipt of three duplicate acknowledgements, possibly due to congestion, it tries to retransmit the missed packets. Though this mechanism may overcome packet losses, it does not resolve the congestion problem. The congestion problem is resolved through an adaptive transmission control mechanism called flow control.

5.8 Improvement in TCP Performance

TCP was designed for traditional wired networks. If used as it is, a few shortcomings become noticeable. We first review a few relevant aspects of traditional TCP. Then we discuss how TCP has been extended to work efficiently in a mobile environment.

5.8.1 Traditional Networks

In the wired networks, packet losses are primarily attributable to congestions that get built-up in the network. To reduce congestion, TCP invokes the congestion control mechanisms. Congestion control is primarily achieved by reducing the transmission window, which in turn results in slower data transfer. The important mechanisms used by TCP for improving (tcp-reno model) performance are given below.

Slow start

The slow-start mechanism is used when a TCP session is started. Instead of starting transmission at a fixed transmission window size, the transmission is started at the lowest window size and then doubled after each successful transmission. The rate of doubling is tied to the rate at which acknowledgements come back. Thus, the doubling of window size occurs at every round trip time (RTT). RTT is the time that elapses between a segment is transmitted by a sender and the corresponding acknowledgement is received. If congestion is detected (indicated by duplicate acknowledgements), the transmission window size is reduced to half of its current size and the congestion avoidance process starts. We can say that this mechanism of rate doubling and reduction to half the previous value is nothing but a binary search technique deployed to determine the 'right' transmission window size.

The slow-start process begins by the sender setting the transmission window size to one, i.e., transmiting one segment to the receiver. The sender does not transmit the next segment until it receives an acknowledgement for the previous segment. Once the acknowledgement is received, the sender becomes sure that the congestion window (network capacity) is at least one segment. To determine the exact congestion window size, the sender doubles the transmission window size. It transmits two segments and after arrival of the two corresponding acknowledgements, it again increases the transmission window size by two and sets it equal to four, and so on. Increments that occur to the size of the congestion window are, thus, exponential. A congestion window is doubled every time the acknowledgements arrive smoothly. This exponential growth of congestion window stops at the congestion threshold.

Congestion avoidance

The congestion avoidance algorithm starts where the slow start stops. Once the congestion window reaches the congestion threshold level, then after that if an acknowledgement is received the window size is increased linearly, i.e., the window size doubling is avoided. From this point, the TCP increases its transmission rate linearly by adding one additional packet to its window at each transmission time. If congestion is detected at any point, the TCP reduces its transmission rate to half the previous value. Thus the TCP seesaws its way to the right transmission rate. Clearly, this scheme is less aggressive than the slow-start phase (note the linear increase against the exponential growth in slow start).

Fast retransmit/fast recovery

Usually, a sender initiates a timer after transmitting a packet and sets the timeout value (RTO). RTO is calculated based on RTT. The sender waits for an acknowledgement of a transmitted packet from the receiver until the timer expires. When the timer expires, it retransmits the packet. However, there exists another situation under which a sender can retransmit a packet. This mechanism is called fast retransmission. In this, the retransmission is not triggered by a timer, it is rather triggered by the receipt of three duplicate copies of an acknowledgement for a packet received from the sender. Since duplicate acknowledgements also arise when a segment is received out of order, the sender waits for three copies of acknowledgements for the same packet. This is taken by the sender as the confirmed indication of a missed packet for starting to retransmit the particular packet. When retransmission occurs, the congestion window size is reduced by half. For example, if the current congestion window size is four segments, then it is set to two segments. Once the lost segment has been retransmitted, TCP tries to maintain the current data transmission rate by not going back to slow start. This is called fast recovery. In fast recovery, the congestion window size is incremented by three since the retransmission occurred after the third

duplicate acknowledgement. This is construed to be the indication that three packets would have been successfully buffered at the receiver end. Thus, in fast recovery, compensation for the segments that have already been received by the receiver is carried out. If the acknowledgements are received smoothly, it is considered to be the indication that there is no congestion.

5.8.2 TCP in Mobile Networks

In Internet, TCP is the de facto standard transport protocol. It has been remarkably successful in supporting the diverse applications which drive the Internet's popularity. A few of such applications are access to information hosted across various websites, file transfer and email. The performance of TCP is considered satisfactory for wired networks, but it suffers from serious performance degradation in wireless networks. Wired networks are significantly different from wireless mobile networks. The main differences are much lower bandwidth, bandwidth fluctuations with time and also as a mobile host moves, higher delay, intermittent disconnections, high bit error rate, and poor link reliability. An implication of these differences is that packet losses are no longer only due to network congestion, they may well be also due to intermittent link failures, bit errors due to noise, or handoffs between two cells. Therefore, the traditional TCP assumptions are not valid in mobile (wireless) environments. This leads to poor performance of TCP in mixed wired-wireless environments. Several modifications at the transport layer have been proposed and studied in recent years to deal with the problems. To understand these works, we need to first consider single-hop wireless networks (such as wireless LANs). Based on this, we shall discuss multi-hop wireless networks (such as mobile ad hoc networks). A few important mechanisms used to improve TCP performance over mobile wireless networks are discussed below.

5.8.3 TCP in Single-hop Wireless Networks

We first discuss the modifications proposed to TCP to make it effective in single-hop wireless networks.

Indirect TCP (I-TCP)

This protocol was proposed by Bakre and Badrinath[2]. It segments the connection between the fixed host and the mobile host into two different connections: the wired part and the wireless part (Fig. 5.7). The wired connection exists between the fixed host (FH) and the base station (BS) and the wireless part connection exists between the BS and the mobile

2. Bakre, A. and B.R. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts," *Proceedings 15th International Conference on Distributed Computing Systems* (ICDCS), May 1995.

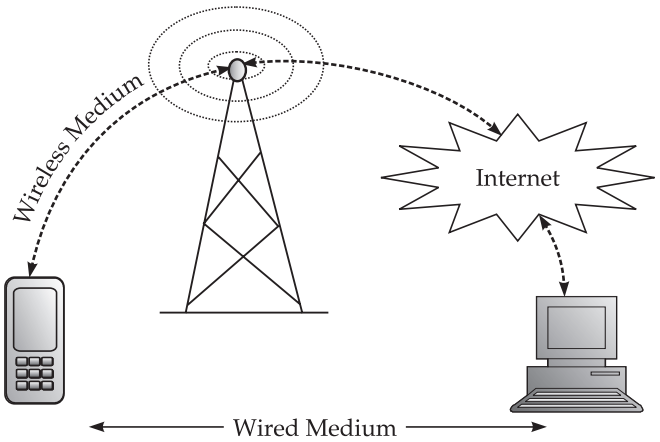


Figure 5.7 *A schematic of working of indirect-TCP.*

host (MH). Thus, the base station maintains two separate TCP connections: one over the fixed network and the other over the wireless link. The wireless link usually has poor quality communication, but it gets hidden from the fixed network at the BS. When a packet is sent by FH to MH, the packet is received by BS first and then BS transmits it to the MH over the wireless link. If the mobile host moves out of the current BS region, the whole connection information and responsibilities that are with the current BS are transferred to the new BS.

The advantage of the split connection approach of I-TCP is that it does not need any changes to be made to the standard TCP protocol. By partitioning a TCP connection into two connections in I-TCP, the transmission errors in the wireless part would not propagate into the fixed network, thereby effectively achieving an increase in bandwidth over the fixed network. An important disadvantage of this scheme is that I-TCP does not maintain the semantics of TCP as the FH gets the acknowledgement before the packet is delivered at MH. I-TCP does not maintain the end-to-end semantics of TCP and assumes that the application layer would ensure reliability.

Fast retransmission

This approach was suggested by Caceres et al.[3] to overcome the delay in transmissions caused due to intermittent disconnections such as those that occurs when a mobile host (MH) moves to a foreign agent (FA) during a TCP communication. As already discussed, TCP transmission behaviour after a disruption depends on its duration. The extremely short disruptions (lasting for a time much less than RTO) would appears as short bursts of

3. Caceres, R. and L. Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments," *IEEE JSAC*, Vol. 19, No. 7, July 2001.

packet losses. In this case, the TCP retransmits those packets for which the timeout occurs and recovers them without slow-start. However, for long disruptions (lasting for a time much greater than RTO), TCP resorts to slow-start. This results in inefficiency. As soon as a mobile host registers at a foreign agent, it starts sending duplicate acknowledgements. As is standard with TCP, three consecutive acknowledgements for the same TCP segment are inferred as a packet loss by the host-end TCP, and it is also inferred that the connection is live, thereby causing the fast retransmit behaviour of TCP.

The advantage of this scheme is that it reduces the time for the MH to get reconnected, otherwise FH would wait for RTO unnecessarily. The disadvantage of this approach is that it does not propose a general approach for TCP communication in mobile wireless networks. For example, it does not address the specific error characteristics of the wireless medium.

Snooping TCP (S-TCP)

Balkrishnan et al.[4] proposed a protocol that improves TCP performance by modifying the software at the base station while preserving the end-to-end TCP semantic. The modified software at the base station is known as *snoop*. It monitors every packet that passes through the TCP connection in both directions, that is from MH to FH and vice versa. It buffers the TCP segments close to the MH. When congestion is detected during sending of packets from the FH to MH in the form of a duplicate acknowledgement or the timeout, it locally retransmits the packets to MH if it has buffered the packet and hides the duplicate acknowledgement.

An advantage of snooping TCP is that it maintains the TCP semantics by hiding the duplicate acknowledgements for the lost TCP segment and re-sends the packets locally. However, it also suffers from higher overheads incurred when MH moves from its current BS to a new BS, the packet buffered at the current BS need not be transferred to the new BS.

Mobile TCP (M-TCP)

This protocol for mobile cellular networks was proposed by Kevin Brown et al.[5] In mobile wireless networks, users would badly suffer from unacceptable delays in TCP communications and frequent disconnections caused by events such as signal fades, lack of bandwidth, handoff, unless these are explicitly handled by the protocol. The M-TCP protocol tries to avoid the sender window from shrinking or reverting to slow-start when bit errors cause a packet loss, as is attempted in I-TCP and snooping TCP.

---

4. Balakrishnan, H., V. Padmanabhan, S. Seshan, and R. Katz, "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links, "*Proceedings of ACM SIGCOMM'96*, August 1996.
5. Brown, K. and S. Singh, "M-TCP: TCP for Mobile Cellular Networks," *ACM Computer Communication Review*, Vol. 27, No. 5, October 1997.

In this protocol, as in I-TCP, the TCP connection between the fixed host and the mobile host is segmented into wired and wireless parts—the wired part connection between the fixed host (FH) and the supervisory host (SH) and the wireless part connection between the SH and the mobile host (MH). Many MHs are connected to SH through several base stations as shown in Fig. 5.8. The SH supervises all the packets transmitted to MH and the acknowledgements sent by MH. It is also used as an interface between FH and MH and vice versa.

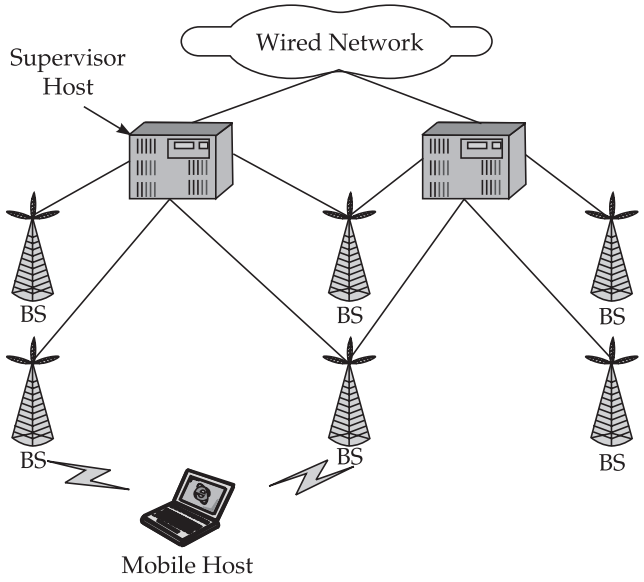


Figure 5.8 *A schematic of operation of the M-TCP protocol.*

When a packet is sent to FH by MH using SH, the wired part uses the normal unmodified TCP and the wireless part uses the modified version of TCP known as M-TCP to deliver data to MH. This packet is acknowledged only when the MH receives the packet. Thus, it maintains the TCP semantics, unlike the I-TCP. In case the acknowledgement is not received by FH, SH decides that MH is disconnected and sets the sender FH window size to zero. This prevents re-transmission. When SH notices that the MH is connected, it sets the full windows size of the sender FH. When MH moves from its current SH region to a new SH region, a state transfer take places, so that the new SH can maintain TCP connection between FH and MH.

Freeze-TCP

The basic idea in this scheme is to "freeze" the TCP senders' streams, little before a disconnection is to occur. This is done by artificially sending a "Zero Windows Advertisement" informing the sender that the receiver cannot

receive data at the moment. When the sender resumes its connectivity, the receiver can unfreeze the sender by sending the value of its actual receive window. The most interesting advantage of this proposal is the avoidance of the slow-start period upon re-establishment of connectivity. The freeze-TCP method has the advantage that it does not require the involvement of the intermediate nodes and hence it can be used if the IP payload is encrypted. This method offers a promising approach for use in Virtual Private Networks (VPNs).

5.8.4 TCP in Multi-hop Wireless Networks

The TCP-F (TCP feedback) protocol has been proposed for extending TCP to multiple-hop networks. In a mobile ad hoc network, a sender MH sends a packet to destination MH through the intermediate MH, since all the nodes of networks are MH. Again the MHs are free to move arbitrarily which changes the network topology unpredictably. If the normal TCP runs over in this network, there may be significant performance degradation at the transport layer. This happens because the normal TCP is unable to distinguish between packet loss resulting from link failure and packet loss due to congestion on the network. As a result, the normal TCP invokes the congestion control mechanism, even if a packet loss occurs due to link failure. When this happens, TCP waits for a longer time to retransmit the lost packet and this slows down the rate of transmission. The TCP-F proposed by Chandran et al.[6] addresses the problem caused by link failure due to mobility by performing a freezing action and limiting the re-transmissions. For simplicity, consider that a source MH is sending packets to a destination MH. As soon as an intermediate MH detects the disruption of route due to mobility of MH along that route, it sends a route failure notification (RFN) packet to the source MH and records that event. Each intermediate MH that receives the RFN packet invalidates the particular route and prevents the incoming packets intended for the destination MH passing through that route. If the intermediate node MH knows of an alternate route to destination MH, this alternative route can now be used to support further communication and the RFN is discarded. Otherwise, the intermediate MH propagates the RFN towards the source MH. On receiving the RFN, the source MH completely stops sending further packets (new or retransmission), then it marks all its existing timers as invalid and freezes the send windows. The source MH remains in this state until it is notified of the restoration of the route through route re-establishment notification (RRN) packets.

A summary of the proposed enhancements to the TCP to support host mobility is presented in Table 5.2.

---

6. Chandran, K., S. Raghunathan, S. Venkatesan, and R. Prakash, "A Feedback-based Scheme for Improving TCP Performance in Ad Hoc Wireless Networks," *IEEE Personal Communications*, 8(1): 34–39, February 2001.

TABLE 5.2 A Comparative Study of a few Important Protocols for Mobile Applications

TCP approach	*Mechanism used*	*Merits*	*Demerits*
Indirect TCP (I-TCP)	Segments the TCP connection into two	• Simple • Isolation of wire and wireless links is possible	• Loss of the TCP semantics • Security problem
Snooping TCP (S-TCP)	Snooping of data and acknowledgements	• Transparency • MCA interaction	• Inadequate isolation of the wireless links • Security problem
Mobile TCP (M-TCP)	The segmented TCP connection can choke the sender through window sizes	• End-to-end segment is maintained • Handles frequent disconnections	• Poor isolation wireless links • Security Problem
Fast retransmission Fast recovery	It avoids slow-start after any roaming	• Simple • More efficient	• Not transparent • Mixed layers
Freeze-TCP	It freezes the TCP, later it resumes the TCP after reconnection	• Works even when there are long interruptions	• Changes in TCP • MAC dependent

SUMMARY

In this chapter, we first reviewed some basic aspects of the TCP/IP protocol suite as the de facto protocol for the Internet and for a host of other applications. We then identified the important problems that might arise when TCP is used as it is in mobile wireless networks. We also discussed the adaptations that have been extended to make TCP work satisfactorily in the mobile environment.

FURTHER READINGS

Ahuja, A., S. Agarwal, J.P. Singh, and R. Shorey, "Performance of TCP over Different Routing Protocols in Mobile Ad Hoc Networks," *IEEE Vehicular Technology Conference,* Vol. 3, pp. 2315–2319, Tokyo, 2000.

Anantharaman, V., S.J. Park, K. Sundaresan, and R. Sivakumar, "TCP Performance over Mobile Ad Hoc Networks: A Quantitative Study,"

'Was to appear in *Wireless Communications and Mobile Computing Journal (WCMC), Special Issue on Performance Evaluation of Wireless Networks*, 2003.

Bhagwat, P., P. Bhattacharya, A. Krishna, and S.K. Tripathi, "Enhancing Throughput over Wireless LANs Using Channel State Dependent Packet Scheduling," *IEEE INFOCOM'96*, San Francisco, March 1996.

Dyer, T.D., and R.V. Boppana, "A Comparison of TCP Performance over Three Routing Protocols for Mobile Ad Hoc Networks," *ACM Mobihoc*, October 2001.

Fu, Z., P. Zerfos, H. Luo, S. Lu, L. Zhang, and M. Gerla, "The Impact of Multihop Wireless Channel on TCP Throughput and Loss", *IEEE INFOCOM'03*, San Francisco , March 2003.

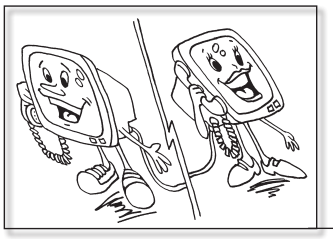
Holland, G. and N.H. Vaidya, "Analysis of TCP Performance over Mobile Ad Hoc Networks," *MOBICOM'99*, Seattle, August 1999.

Xiang Chen, Hongqiang Zhai, Jianfeng Wang, and Yuguang Fang, "A Survey on Improving TCP Performance over Wireless Networks," Dept. of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, http://winet.ece.ufl.edu/~zhq/book05chen1.pdf.

EXERCISES

1. State **true** or **false** against each of the following statements. Briefly explain your answer in each case.
   (i) TCP is a peer-to-peer, connection-oriented protocol.
   (ii) The link layer is responsible to deliver data packets by making use of lower layer protocols and also provides error detection and packet framing functionalities.
   (iii) TCP guarantees that data will be delivered without loss, duplication or transmission errors.
   (iv) TELNET encodes multimedia data using the MIME protocol.
2. Explain the following terms associated with TCP/IP stack:
   (a) IP, (b) HTTP, (c) SMTP, (d) MIME, (e) FTP, (f) SNMP, (g) ICMP, (h) ARP, (i) RARP, (j) DNS, (k) IP Addresses, (l) IGMP
3. Write short notes on:
   (a) TELNET, (b) FTP, (c) SMTP, (d) TCP/IP vs. ISO/OSI protocol model.
4. Explain the layered architecture of the TCP/IP protocol suite and compare it with the ISO/OSI architecture.

5. Answer the following with respect to missing and duplicate segments in TCP operation.
   (a) What can cause segments to be missed at the receiver-end and also cause duplicate segments to arise? Explain your answer using a suitable scenario of operation.
   (b) How exactly is a missing segment detected in TCP? Explain the specific actions that take place when a missing segment is detected.
6. What is slow-start in TCP operation? Explain its working. How does slow-start help improve the performance of TCP?
7. What problems would occur if the traditional TCP is used in mobile wireless environments? Discuss how TCP can be adapted to work efficiently in a mobile network environment.
8. Explain Indirect-TCP (I-TCP) with the help of a suitable schematic diagram.
9. Why are the I-TCP acknowledgements and semantics not end-to-end? What are the implications of this?
10. What is the snooping TCP approach in mobile wireless networks? Discuss its advantages.
11. How are handoffs handled in snooping TCP?
12. Briefly discuss the M-TCP approach of extending TCP to work efficiently in mobile wireless networks. How does M-TCP maintain end-to-end semantics?
13. Explain the working of Freeze-TCP. What are the shortcomings of this protocol?
14. Why do congestions occur in a network? Explain how does TCP detect and handle congestions.



6

Mobile Databases

Recent developments in technologies have made it possible for a mobile user to have uninterrupted information and service access at his disposal. For example, a user can even browse Internet or carry out online banking transactions while travelling in a vehicle. This has been made possible, among other things, by the mobile database technology that allows the applications for hand-held devices to access data stores even while on the move. The mobile database technology makes data from database applications available ubiquitously to a mobile worker. An employee using a hand-held device can link to his corporate network, download data, work offline, and then connect to the application seamlessly. Consider another example. Using the mobile database technology, a courier delivery worker can collect signatures after each delivery and send the delivery information to a corporate database as soon as the delivery is made.

Traditional database systems are difficult to use in a mobile computing environment. Traditional databases make implicit assumptions that do not hold in a mobile computing environment. To realize effective mobile databases, the special characteristics of mobile computing should be considered in the database design. For example, mobile transactions have to cope with unexpected disconnections caused due to mobility. The transaction processing system also needs to cope with the severely restricted resources of the mobile hosts, including low battery life, slow processor speed, and limited memory. Mobile applications are required to react to frequent changes in the environment such as changing locations, high variability of network bandwidth, and dynamically changing data. Taking all these limitations and constraints into consideration, the different issues in database design such as data dissemination, data replication, transaction models, query processing and concurrency need to be investigated to determine the specific changes that are required. Before discussing these issues, we need to keep the basic structure of a mobile application in mind. Figure 6.1 presents a schematic representation of invocation of the services

of an application in a simple mobile computing system. As can be seen in this figure, a mobile host (MH) in a wireless cell is connected to a Mobile Support Station (MSS) that has a wireless interface. The hosts on the fixed network do not have a wireless interface. The MSS is connected to the fixed network though a wired fixed network. A mobile client can invoke a service of an application server on the fixed network. The application server, in turn, accesses the database servers using a connectivity protocol.

A few specific requirements identified for a mobile database system are enumerated below[1].

- A mobile database should have a small footprint to fit into a resource-constrained mobile device.
- It should be able to run without the services of a database administrator.
- It should easily interoperate with the large enterprise databases.
- A mobile database needs to support rather simple insert, delete, update, and query functionalities.

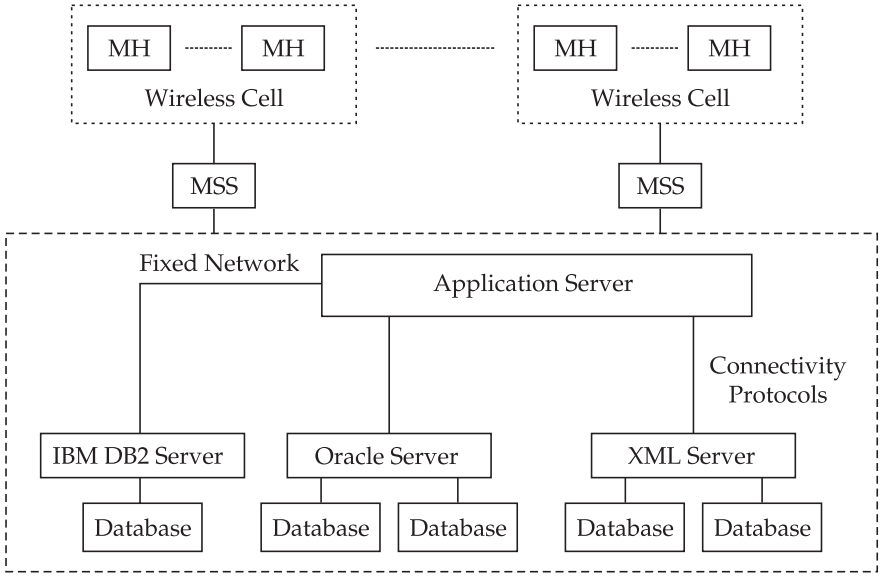


Figure 6.1 *A schematic of service invocation in a mobile computing environment.*

The commercially available mobile database systems allow operations across a wide variety of platforms and data sources. They also allow users with a hand-held device to synchronize with the Open Database Connectivity (ODBC) database content, and personal information management data and e-mail from Lotus Development's Notes or Microsoft's Exchange. These database technologies usually support query-by-example (QBE) or SQL

1. Lu, E. and Y. Cheng, "Design and Implementation of a Mobile Database for Java Phones," *Computer Standards and Interfaces*, Elsevier, Vol. 26, pp. 401–410, 2004.

instructions. Some of the commercially available Mobile relational database systems includes IBM's DB2 Everywhere 1.0, Oracle Lite, and Sybase's SQL Anywhere.

A few important factors that need to be considered for service invocation in a mobile computing environment are as follows:

- User time is highly valuable.
- Connection time requirement should be minimum.
- The number of bytes, or packets, transferred in the unit of charges is computed in digital cellular systems and therefore needs to be minimized.
- Time-of-day based charges may vary based on whether communication occurs during peak or off-peak periods.
- Energy of the mobile handsets is limited. Battery power is a scarce resource and should be optimized.

6.1 Issues in Transaction Processing

Database applications are normally structured into transactions. A *transaction* is a unit of operation that changes a database from one consistent state into another consistent state. Normally, different transactions on a database operate in an interleaved manner. Interleaving the access of different transactions to the database can remarkably improve the throughput and resource utilization of the database. Therefore, one important aim in the design of database systems is to maximize the number of transactions that can be active at a time.

Let us now examine how the **ACID** properties (**A**tomicity, **C**onsistency, **I**solation, **D**urability) are ensured in a database in the presence of interleaved execution of transactions. While each transaction preserves the *consistency* of the database at its boundaries, a transaction that fails to complete might cause the integrity of the database to be violated. Rollback protocols are used to ensure *atomicity* and *durability* properties when a transaction fails to complete. Isolation can be ensured through the use of locking and rollback protocols.

A particular sequencing of actions of different transactions is called a *database schedule*. However, concurrent execution of transactions can lead to some database schedules that take the database to an inconsistent state. A database is in a *consistent state* if it satisfies all the defined semantic integrity constraints. A *transaction model* defines the framework for the definition and execution of transactions. The integrity constraints on a database are given by the well-known ACID properties, meaning that a transaction should be an atomic, consistent and recoverable unit and should not interfere with other transactions that are executed concurrently. In some application scenarios such as those in a mobile environment, the

BOX 6.1 Issues in transaction processing

A transaction consists of many read and write operations on the database. In a concurrent transaction processing environment, arbitrary transaction execution may lead to violation of database integrity. To exclude the violation of integrity constraints, the concept of transaction serializability has been introduced. The execution of a set of concurrent transactions is said to be serializable, when the concurrent execution of the transactions is equivalent to some serial execution of these transactions. To successfully complete a transaction, all operations must be performed on the database in some serializable order and a recovery protocol must be executed. But, how does a DBMS ensure serializability? In order to ensure serializability, a DBMS needs to enforce the following constraints.

Atomicity: Either all operations of a transaction are reflected on the database or none at all. That is, all the operations of a transaction are together treated as a single indivisible unit.

Consistency: A transaction should transform the database from one consistent state to another consistent state.

Isolation: The effect of a transaction should be such that it appears to be executed in isolation. That is, its intermediate results must not be seen by other transactions that commit before it.

Durability: Changes made to the database by a committed transaction must persist even after any database failure, rollback, etc. that may take place.

ACID properties turn out to be too stringent for realizing any meaningful database-based application in practice and therefore these properties require to be relaxed.

6.2 Transaction Processing Environment

In this section, we discuss the different transaction processing environments.

6.2.1 Centralized Environment

This is the oldest data-processing system in which a large program manages both the tasks of initiating a transaction by interacting with users and processing it. In this environment, the overheads associated with the processing of a transaction are rather low since only a single user executes the transactions. Recovery is also a trivial issue here, since rollback recoveries do not arise.

6.2.2 Client-server Environment

In this environment, the execution of a transaction and transaction initiations are done by a server and a client respectively. In this environment, the server is typically concurrent. That is, many clients can submit transactions simultaneously.

6.2.3 Distributed Environment

In this environment, data is distributed over a network. That is, either a transaction can get fully executed at any arbitrary node, or different parts of transaction can be executed at different nodes. Atomicity and durability of a transaction become nontrivial issues in this environment.

6.2.4 Mobile Environment

This is a special type of distributed environment and can accommodate user movements while issuing the transactions, and the system reports the results. Even though it as an extension of the distributed model, it becomes important to satisfactorily address many nontrivial issues that are not present in a typical distributed computing scenario. These issues are discussed in the following:

Movement of users

The movement of users introduces several complications as far as the database operations are concerned. The complications arising due to user mobility must be satisfactorily addressed. Users tend to move from one cell to another while maintaining the network connection. To support this, issues such as tracking users, and maintaining the network connection while crossing the cell boundaries must be addressed. Recovery of transactions is complicated as it becomes a tricky issue to determine the location at which the logs are to be maintained. The options available are:

Maintaining logs at the mobile host: This may be difficult due to lack of sufficient resources such as memory and processing power. As a result, storing the logs at the mobile hosts and initiating the recovery in case of failure would be an unacceptable overhead. Also, as the failure rate of a mobile host is high, this approach gets further complicated and hence not considered satisfactory.

Maintaining logs at mobile host's base MSS: This approach incurs commu-nication overhead because each transaction log must be updated, introducing one additional round trip delay, which is undesirable.

Maintaining logs at the current MSS: In this approach, logs are maintained at the current MSS for all transactions submitted from its cell. When an MH moves out of the cell, all logs are transferred to the new MSS. This approach generates a very high network traffic due to the transfer of logs. Also, it introduces additional complications such as how much resources should be allotted to a mobile host, particularly the foreign ones, and what to do with the logs if an MH fails.

Disconnections

A fairly complicated issue in the operation of a mobile database arises due to the temporary disconnections of mobile users arising out of attempts to save power. In a typical distributed environment, the non-availability of a host is considered to be a failure. But here one must distinguish between a failure and a planned disconnection. In a planned disconnection, depending on the application, the mobile user can perform some operations by downloading data beforehand in anticipation. This is popularly called *data hoarding*. Another way to deal with the disconnections is by migrating transaction processing to a non-mobile computer.

One more approach to handle the non-availability of an MH for a temporary period is by maintaining proxy agents at the MSS. A proxy agent represents the MH during its absence and participates in communication and finally hands over control to MH on its reappearance. In this approach, it is necessary to satisfactorily take care of creation of proxy agents, participate in communication, hand over control and ensure self-destruction.

But when an MH disappears for a long duration, the proxy agent does not know what to do. Maintaining it in the wait state may lead to unnecessary wastage of resources of the MSS. Collecting messages on behalf of an MH in its absence or use of mailboxes may be an appropriate solution. When a mobile host is not available in the network, all messages destined for it can be routed to its mailbox, which will be retrieved by the corresponding MH after reconnection. For implementation of this concept, a base MSS must be present for each MH, which again is a very difficult proposition.

Poor communication media

The bandwidth allocated to mobile users could be very low. Even when sufficient bandwidth is available, many kinds of noise such as interference from other traffic, atmospheric conditions, etc. may corrupt the data, thus causing multiple retransmissions. Transmission over a wireless medium consumes more power than that over a wired link and power is a scarce resource for mobile hosts. For this reason, mobile hosts tend to disconnect from the network whenever there is no data to send or receive in the near future. This phenomenon is known as *dozing*.

BOX 6.2 Energy consumption in wireless communication

A wire serves as a guided medium, whereas in wireless transmission power is radiated in all directions. This causes the radiated energy density to fall off exponentially with distance. Also, the impedance of air is much higher than that of a wire. These are the two important reasons for higher energy consumption in wireless communication compared to that in wired communication.

Processing power

Mobile hosts are generally equipped with less powerful CPUs in order to conserve power. Due to this reason, compute intensive applications such as a database server are difficult to use. This constraint also forces one to rethink the way computations are done. For example, it may be required to find ways to reduce the client-side computations.

Memory

Mobile hosts usually have a very limited main memory as well as secondary storage. Nonvolatile memory devices such as hard disks consume significant power for running the motor that drives the read/write heads. Also, hard disks are usually bulky and heavy. Therefore, all the important applications and system software requiring storage on permanent memory may be put into a flash memory, which consumes much less power.

Battery power

Battery power is possibly the most precious and scarce resource of a mobile host. Not only batteries store very limited amounts of energy, often recharging them becomes difficult. Most of the battery energy is spent on non-computational tasks such as transmitting and receiving data over the wireless medium and some energy is also consumed by the secondary storage and the input and output devices.

User interfaces

The I/O devices consume significant power and often are bulky. One needs to find new ways to optimize the power they consume and also the space they occupy. As a substitute for keyboard, one can use pen-based input devices, which occupy less space but consume more computing power for lengthy inputs, particularly for SQL type statements. Thus, a graphical/symbolic interface with pointing capabilities may suit this environment.

Security

Since mobile handsets are carried across several networks, the prevention of impersonation of one machine by another is problematic. For example,

when a mobile computer enters a new cell, the specific data it sends and receives are subject to theft and unauthorized copying for replay at a later date.

A mobile computer visiting a new environment consumes some resources of that environment. These resources includes network bandwidth, CPU time, and disk storage of the base station. An accounting of this must be made for both billing purposes and for limiting the impact of visiting mobiles on the performance of a foreign environment.

Another security concern is how to prevent the visiting mobile computer from abusing the foreign environment, in term of the bandwidth and the disk storage it uses.

Asymmetry in communication

In a mobile computing environment, there is a difference in uplink and downlink bandwidths due to the fact that mobile clients are not able to transmit at very high speed, due to power restrictions and rather small data transmission requirements of the mobile handsets. The bandwidth in the downstream direction (server to mobile handsets) is far greater than that in the upstream direction (handset to server).

6.3 Data Dissemination

Data dissemination deals with the delivery of data from a data repository (server) to a large number of clients. The data dissemination model in a mobile environment is usually different than its wired counterpart due to the asymmetry in communication.

The model of sending data to clients in a mobile computing environment is usually push-based. In a push-based data dissemination, data is sent to the clients by the server without waiting for specific requests from the client. In a push-based transmission, not only does the server save itself from handling large numbers of similar queries from a large number of clients, but also can propagate information whose existence may be unknown to the clients.

On the negative side, push-based data dissemination unless judiciously used can overwhelm the clients with unwanted information. Therefore, the usefulness of a push-based data dissemination technique to a large extent depends on the accuracy of predicting the requirements of various clients. A simple solution usually used is to allow the clients to subscribe to specific services or provide their profile, based on which the information can be disseminated to them.

6.4 Transaction Processing in Mobile Environment

In general, mobile transaction processing should provide seamless transaction processing during disconnections, and preserve the system correctness by allowing limited inconsistencies among the data copies. It is important to recognize that clients can read or write data from any one of the servers and move over to a new location and carry out further operations from there. For example, a client can write at one server and later read the item at a different server. The client can obtain inconsistent views unless the different servers have been synchronized. The following sections describe how each of the ACID properties has been redefined. The relaxation of one property may impact the other ACID properties.

6.4.1 Atomicity Relaxation

In this modified scheme, an MH is allowed to submit 'pieces' of transactions from different cells according to their movements. This approach weakens the atomicity property and requires the ability to break a transaction into sub-transactions that can be concurrently executed and interleaved with the sub-transactions of other transactions while guaranteeing the ACID properties.

Several techniques are available for decomposing a transaction into sub-transactions according to different and at different levels of granularity required and depending on the type of the transaction. That is, the read only transactions may be submitted as a whole at a unique MSS or may be split during processing as per the movements of the MH that submitted the transaction. Transactions that update data items can be split into mutually independent sub-transactions. This decomposition ensures that the split sub-transactions can be concurrently executed at different MSSs and their execution order does not impact the successful commitment of the original transaction. The splitting in this case is done according to Bernstein's conditions. Each transaction, Si, has a write set Wi and a read set Ri. Every two sub-transactions, Si and Sj, of a transaction T must satisfy the following sets of conditions to guarantee their independence.

$$Wi \cap Wj = \Phi$$
$$Ri \cap Wj = \Phi$$
$$Wi \cap Rj = \Phi$$

A runtime support must be designed to compute the decomposition of transactions.

6.4.2 Consistency Relaxation

Under this approach, the database is logically partitioned into "clusters", based on either some semantic properties or location proximity. Data in

the same cluster must be strictly consistent, whereas a "bounded" degree of inconsistency is tolerated among the clusters, according to a relaxed definition of consistency. In this approach, a mobile host can download one cluster before disconnection and execute transactions on this data by preserving ACID properties and after reconnection it can be copied to the original database.

6.4.3 Isolation Relaxation

Some transaction models have been devised for mobile environments in which the isolation property is not guaranteed. That is, the intermediate results of a transaction can be observed by other concurrent transactions. This is usually a side effect of the relaxation of other ACID properties.

6.4.4 Durability Relaxation

Durability of committed transactions is mainly affected by the possibility of MHs autonomously operating on data. A disconnected MH can only commit a transaction locally if this transaction does not conflict with other transactions executed on the same host while the host is disconnected. On reconnection, the transactions are globally executed on different MHs.

6.5 Data Replication

Data replication is the process of maintaining a defined set of data in more than one location. In a mobile computing environment, data replication is essential since this would allow the user to continue operating during temporary disconnections and would also improve the transaction success rate and transaction throughput. In order to achieve this, data must be replicated at multiple places including partial replication of the relevant data on the mobile database of the MH. An MH needs to have its own complete or partial copy of the relevant data from the main database so that it can process data locally. In general, data replication provides the following two advantages.

Data availability: Data is made available locally to the remote databases. This avoids the need to connect to a single central database through slow and less reliable connections.

Faster response times: Replication improves the response times for data requests since requests are processed on a local server. Further, local processing decreases the workload at the central database server.

The number of replications of a database in a mobile computing environment can be large. There is not much issue in replicating the read-only (or cached) copies, however, the key issue is in managing the replication of the updateable (or core) copies. Of course, the number of core copies need to be kept small to contain the overhead of synchronization.

6.6 Mobile Transaction Models

Mobile transactions access remote data through low bandwidth wireless connections, and access local data in the disconnected mode. Reliability and consistency management are the main goals of mobile transaction models. Some existing models of mobile transaction management include Reporting and Co-Transactions model (Chrysanthis, 1993), Kangaroo Transaction model (Margaret Dunham et al., 1997), Clustering model (Pitoura and Bhargava, 1994), Isolation-Only Transactions (Satyanarayan et al., 1990) and Two-tier Transaction model (Gary Shih, 2002).

6.7 Rollback Process

Consider a transaction which has started operation, and the modification of a database is in progress. Now if an error occurs due to mobility, then all the transactions may fail. In this case, rather than leaving the database with partial results, the transaction will have to undo the updates and leave the database with older values that were available before the transaction started. Such a reversal to the initial state is known as *rollback* process.

6.8 Two-phase Commit Protocol

In a mobile environment, during transaction processing, databases, and computer networking, an appropriate commit protocol needs to be used at the transaction commit stage. The two-phase commit protocol (2PC) is a type of an atomic commitment protocol. It is a distributed algorithm that coordinates all of a distributed atomic transaction to decide whether to commit or abort the transaction. The protocol achieves its goal even in cases of temporary system failure (involving either process, network node, or communication failures), and is widely used (Bernstein et al., 1987). The two phases of the algorithm are the *commit–request phase*, in which a *coordinator* process takes the necessary steps for either committing or aborting the transaction. In the *commit phase*, the coordinator decides whether to abort the transaction.

Note: The two-phase commit (2PC) protocol should not be confused with the two-phase locking (2PL) protocol, which is a concurrency control protocol.

6.9 Query Processing

An efficient database system management should minimize the number of uplink requests (queries). For this, the client needs to cache some commonly used data in its local memory. Also in a service invocation by a mobile handset (client), the location-specific results become important.

6.9.1 Location-dependent Querying

Several attempts have been made to integrate Global Positioning System (GPS) information into queries to create location-dependent services. Examples of such services include local news, train or flight information, etc. In web browsing, the client browser may insert the location information in the query to seamlessly obtain location-specific services.

6.9.2 Query Optimization

During a transaction with a database, queries typically require gathering and updating information in a database. Efficient query processing is achieved by using query decomposition and query optimization. A query can be decomposed into a set of algebraic expressions by using the following techniques:

- Appropriate analysis
- Normalization (conjunctives as well as disjunctives)
- Semantic analysis

During query optimization, a query is parsed and normalized. The parser ensures that the language, say SQL, syntax is correct. Normalization ensures that all the objects referenced in the query exist. Permissions are checked to ensure that the user has the permission to access all tables and columns in the query.

A query parser parses and translates a given query into an immediate form such as a relational algebraic expression. The parser achieves this based on both the syntax of the query and the semantics of the query. A parse-tree of the query is constructed and then translated into a relational algebra expression.

BOX 6.3 Query decomposition operations

We explain the basic query decomposition operations using a few simple examples:

Normalization: Transform the query to a normalized form.

Example: Consider the following query: Find the names of employees who have been working on project P1 for 12 or 24 months?

- The query in SQL:
  SELECT ENAME
  FROM EMP, ASG
  WHERE EMP.ENO = ASG.ENO AND
  ASG.PNO = "P1" AND
  DUR = 12 OR DUR = 24
- The qualification in conjunctive normal form:
  EMP.ENO = ASG.ENO ^ ASG.PNO = "P1" ^ (DUR = 12 v DUR = 24)
- The qualification in disjunctive normal form:
  (EMP.ENO = ASG.ENO ^ ASG.PNO = "P1" ^ DUR = 12) v
  (EMP.ENO = ASG.ENO ^ ASG.PNO = "P1" ^ DUR = 24)

Analysis: Detect and reject "incorrect" queries; possible only for a subset of relational calculus. Example of the type of incorrect query: Checks whether the attributes and the relation names of a query are defined in the global Schema or Checks whether the operations on attributes do not conflict with the types of the attributes, e.g., a comparison > operation with an attribute of type string.

Elimination of redundancy: Simplify the query by eliminating redundancies, e.g., redundant predicates. Redundancies are often due to semantic integrity constraints expressed in the query language.

Semantical analysis: Identify the incorrect queries. Example of semantically incorrect query: Checks whether the components contribute in any way to the generation of the result.

The purpose of the data analysis is to improve the speed of answering queries on that data. Analysis produces summary information that can either answer a query without consulting the data itself, or else modify the query to a form that the data server will be able to process more quickly. This query-modification operation using the knowledge of the data is known as Semantic Query Optimization.

A mobile computing environment poses several challenges in query processing. For example, as per the mobile-computing model, the route between a pair of hosts may change over time. This may change the location-specific results. So the mobile-computing model may directly affect the database query processing. In the case of distributed query processing, the communication costs play an important role in the query optimization process. The issue of host mobility, therefore, complicates the optimization process.

6.10 Recovery

A database recovery step gets the database back to the last stable state that was reached before a failure. A transaction processing system (TPS) may fail for many reasons. These reasons could be varied and may include, among other things, battery wear out, system failure, human errors, hardware failure, incorrect or invalid data, computer viruses, software application errors, natural or man-made disasters. As it is not possible to prevent all TPS failures, a TPS must be able to detect failures immediately after they occur. For this, the TPS must be able to detect and correct errors when they occur and restore the database to a stable state. A TPS recovery takes place by a recovery manager, based on the backup, journal, and checkpoint information. We explain these concepts in the following:

Journal: A journal maintains an audit trail of transactions and database changes. The transaction logs and the database change logs are maintained. A transaction log records all the essential data for each transaction, including data values, the time of transaction and the terminal number. A database change log contains the before and after copies of records that have been modified by transactions.

Checkpoint: A checkpoint is a record of the state of the database system at a specific time. A checkpoint record contains the necessary information to restart the system. In case of a failure, it is possible to resume processing from the most-recent checkpoint. Of course, the processing performed after the last checkpoint would be lost. Therefore, checkpoints should be taken frequently.

Recovery Manager: A recovery manager is a program which restores the database to a correct state from which the transaction processing activities can be resumed, in case there is a failure.

SUMMARY

A mobile database is expected to handle the special issues that arise in case of transactions invoked by mobile clients. The client communicates with the server using a wireless connection. A mobile database is an adaptation of a traditional database to accommodate many of the issues that a mobile environment imposes, such as disconnections, low battery power, etc. A cache is maintained to hold the frequently used data and transaction results so that they are not lost due to connection failures. We discussed a transaction model suitable for a mobile computing application.

FURTHER READINGS

Barbara, D., "Mobile Computing and Databases—A Survey," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, No. 1, pp. 108–117, January 1999.

Bernstein A., Philip, Vassos Hadzilacos, Nathan Goodman, *Concurrency Control and Recovery in Database Systems*, Chapter 7, Addison Wesley Publishing Company, ISBN 0-201-10715-5, 1987.

Chrysanthis, P.K., "Transaction Processing in Mobile Computing Environments," *Proceedings of IEEE Workshop on Advances in Parallel and Distributed Systems*, pp. 77–82, 1993.

Dirckze, R.A., L. Gruenwald, "A Toggle Transaction Management Technique for Mobile Multidatabases", *Proceedings of the CIKM 98*, Bethesda, MD, USA, pp. 371–377, 1998.

Gary Shih and Simon S.Y. Shhim, *A Service Management Framework for M-Commerce Applications*, Kluwer Academic Publishers, the Netherlands, 2002.

Lu, E. and Y. Cheng, "Design and Implementation of a Mobile Database for Java Phones," *Computer Standards and Interfaces*, Elsevier, Vol. 26, pp. 401–410, 2004.

Margaret H. Dunham, Abdelsalam Helal, and Santosh Balakrishnan, *A Mobile Transaction Model that Captures Both the Data and Movement Behavior*, Baltzer Science Publishers, 1997.

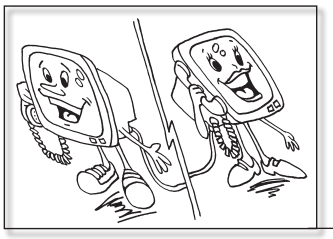
Pitoura, E. and B. Bhargava, "Building Information Systems for Mobile Environments". In Third International Conformation and Knowledge Management, CIKM' 94, Gatheburg, MD, USA, ACM, pp. 371–378, November 1994.

Satyanarayan, M., J.J. Kistler, P. Kumar, M.E. Okasaki, E.H. Siegel, and D.C. Steere, "Coda: A Highly Available File System for a Distributed Workstation Environment", *IEEE Transaction on Computers*, Vol. 39, No. 4, pp. 447–459, April 1990.

EXERCISES

1. Explain the different issues that arise in a transaction processing system having mobile clients for satisfactory operation of an application. How are these issues resolved?
2. What do you mean by serializability of transaction executions? How does a DBMS ensure that the transactions are serializable?

3. What do you understand by concurrent transactions? Using one example, explain why arbitrary concurrency in transaction processing should not be allowed.
4. Briefly explain why wireless communication requires higher energy compared to that in wired networks.
5. What is data hoarding in a mobile database environment? Explain why is it required?
6. What is dozing in a mobile computing environment? Why does dozing occur? What are its implications in mobile transaction processing?
7. What do you mean by atomicity relaxation?
8. What is a mobile database? How is it different from a traditional database?
9. What do you understand by data dissemination in a mobile computing environment? Discuss how data dissemination can be achieved in the mobile computing environment.
10. Explain the rollback process during the execution of a mobile transaction. How is it different from a traditional database rollback?
11. Explain the two-phase commit process in the context of mobile transactions.
12. How does a database handle failure that occurs during a mobile transaction processing?
13. Explain why wireless transmission of data consumes more power compared to transmission of data over a wired medium.
14. What problems might occur if a traditional database system is used in a transaction processing system having mobile clients? Explain your answer using suitable examples.
15. Why is data replication important in a mobile computing application? Discuss the issues of data replication in mobile applications.



7

Mobile Ad Hoc Networks

In recent years, the mobile ad hoc networks are becoming very popular. We discussed in Chapter 1 that a wireless LAN can be set up with the help of certain fixed infrastructures such as access points and routers. However, in many situations such as that in a remote village or on a railway station no networking infrastructures may exist, yet the user of a mobile device might still like to communicate with another user who might be in the same area or elsewhere. In this situation, if the mobile devices present in the area can somehow discover each other and collaborate to set up a wireless network among themselves, then the communication can become possible. However, it is necessary that such a network should be self-configuring (set up and configured automatically without requiring any user intervention), since the network configuration keeps changing due to the movements of mobile devices. This is important since mobile ad hoc networks are realized through short-lived links among the mobile devices. Such temporary or ad hoc networks that are established and maintained on the fly and work without the support of any form of fixed infrastructure such as a base station, are known as Mobile Ad hoc NETworks (MANETs).

MANETs are traditionally defined as self-configuring networks set up among the hand-held devices of mobile users. However, of late several specialized MANETs such as Wireless Sensor Networks (WSNs) and Vehicular Ad hoc Networks (VANETs) have emerged. Each of these specialized ad hoc networks is suitable for a specific kind of application. Being ad hoc networks after all, all these networks share some basic characteristics. However, there exist significant differences among them with respect to their operation, design, and applications. The focus of this chapter is to discuss some essential aspects of MANET applications and operations. Considering their widespread acceptance and use, we discuss WSNs in the next chapter as a special type of MANET. The term mobile ad hoc network (MANET) was proposed by IETF in the year 2002. In this text,

we shall use the terms mobile ad hoc networks (MANETs) and wireless ad hoc networks interchangeably.

This chapter is organized as follows: In Section 7.1, we discuss the basics of mobile ad hoc networks. In Sections 7.2 and 7.3, we present the characteristics and applications of MANET. In Sections 7.4 to 7.7, we discuss some basic issues pertaining to ad hoc networks. In Section 7.8, we discuss various routing issues in MANETs. In Sections 7.9 and 7.10, we discuss VANETs (Vehicular Ad hoc NETworks) as a special type of MANET. In Sections 7.11 and 7.12, respectively, we discuss the various security issues and attacks in MANETs.

7.1 A Few Basics Concepts

In the preceding chapters, we discussed various aspects of wireless networks and the vital role played by networking infrastructures such as hubs, routers, and base stations in their operation. However, such networking infrastructures may not be available in many situations such as a disaster-hit locality or a remote location as we have already pointed out.

7.1.1 How Is an Ad Hoc Network Set Up without the Infrastructure Support?

Let us try to imagine how a set of mobile devices can communicate with each other in the absence of any form of fixed networking infrastructures such as hubs, routers, base stations, etc. It is not hard to visualize that in this situation, a network can be established through cooperation among the devices themselves. In a simplistic realization of this concept, a mobile device wanting to communicate can forward its packets to its neighbours, and the neighbour nodes in turn can forward those to their neighbours, and so on until the destination is reached. This essentially forms a simple ad hoc network. Based on this idea, Fig. 7.1 illustrates a schematic model of a simple ad hoc network comprising three mobile devices, named S, R, and D. In this figure, suppose the mobile device S wants to communicate with the device D. Assume that S and D are not within the transmission range of each other and cannot directly communicate with each other. However, they can take the help of node R to relay packets from each other. Observe that Fig. 7.1 depicts a protocol stack to highlight that even though R is primarily an independent device and not a networking infrastructure, yet R is acting as some sort of a router operating at the network (or Internet) layer to facilitate communication.

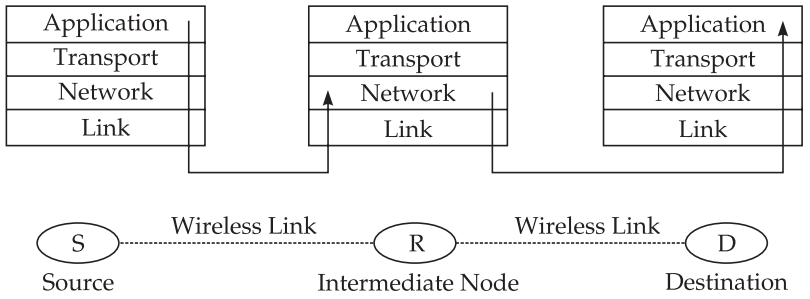


Figure 7.1 *A schematic model of a mobile ad hoc network.*

7.1.2 Why Is Routing in a MANET a Complex Task?

In a wired network, a router determines the path that needs to be followed by a packet based on the information contained within the IP address of the destination, and uses this information to forward a packet towards its destination. In an ad hoc network, such a simple and efficient routing protocol is difficult to deploy. First, it is very difficult to have a global identifier assigned to every node that would also indicate the route to the node. This is because the nodes keep on moving and the identity would also have to change, which would incur an inordinately large overhead. Further, even if a route between a pair of nodes is somehow determined, routes become quickly obsolete since they dynamically get built and also get dissolved. Suppose a route has already been determined between a source and a destination and packets are getting transmitted on this route. Some nodes forming this route may move away even as the packets are getting transmitted, thus disrupting the communication.

In a nutshell, we can say that in a MANET the topology of the network and consequently the routes between different devices change dynamically as nodes move away or fail. This is in contrast to any wired network. As a result, packet routing is a critical and complex issue in MANETs and a central topic in any study of MANETs. Consequently, a satisfactory routing protocol for ad hoc networks should be able to cope with factors such as link breakages, and ensure that messages get transmitted efficiently and seamlessly even when nodes move or shut down due to hardware malfunctioning, depletion of battery energy, etc.

7.2 Characteristics of Mobile Ad Hoc Networks (MANETs)

There are several characteristics that distinguish a MANET from an infrastructure-based network. Naturally, these characteristics must be

carefully considered while trying to improvise new network protocols or extend the traditional network protocols for use in a MANET. A few of these important characteristics of MANETs are described below.

Lack of fixed infrastructure: Lack of any specific networking infrastructure is possibly the most distinguishing characteristic of a MANET. In the absence of any fixed networking infrastructure, a pair of nodes can either communicate directly when they are in the transmission range of each other, or they can communicate using a multi-hop communication that gets set up through several devices located between them. Based on this characteristic alone, cellular networks and wireless LANs cannot be considered to be MANETs.

Dynamic topologies: Since the devices in a MANET are allowed to move arbitrarily, the network topology can change unpredictably. The rate of topology change depends on the speed of movement of the mobile devices. The speed of movement of a mobile device can vary greatly with the time of the day and the specific MANET application being considered.

Bandwidth constrained, variable capacity links: Wireless links have significantly lower capacity than their wired counterparts. Further, factors such as fading, noise, and interference can change the available bandwidth of a wireless link arbitrarily with time. Consequently, the bandwidth of a link can change arbitrarily with time.

Energy constrained operation: The nodes in a MANET rely on battery power. These batteries are small and can store very limited amounts of energy. On the other hand, transmissions and processing required during routing involve expenditure of substantial amount of energy causing the batteries to get rapidly drained out, unless the routing protocol is carefully designed. Therefore, energy conservation is usually considered to be an important objective of MANET routing protocols.

Increased vulnerability: MANETs are prone to many new types of security threats that do not exist in the case of their wired counterparts. Many of these threats arise due to the underlying wireless transmissions and the deployment of collaborative routing techniques. Further, there are increased possibilities of eavesdropping, spoofing, denial-of-service attacks in these networks. Further, it is very difficult to identify the attacker since the devices keep moving and do not have a global identifier. Besides, nodes are vulnerable to capture and compromise. We discuss these security threats in greater detail in Section 7.10.

Other characteristics: Other distinguishing characteristics of a MANET include a distributed peer-to-peer mode of operation, multi-hop routing, and relatively frequent changes to the concentration of nodes over any specific area.

On account of the above mentioned characteristics of MANETS, a suitable protocol for MANETs needs to operate under several constraints. We point out these constraints in the following Subsection 7.2.1.

7.2.1 MANET Operational Constraints

The nodes in a MANET have low processing capabilities and these are connected by low bandwidth wireless links. To cope with low processing capabilities of the nodes and the limited bandwidth availability, a suitable routing protocol needs to be adopted to make efficient utilization of the available bandwidth. However, an appropriate routing protocol for a MANET should keep the computational and communicational overheads low, since the nodes in a MANET have low computational capability, storage capacity and battery power. Out of these, battery power is possibly the most scarce resource. To conserve communication resources, a suitable routing protocol for a MANET should not make use of frequent flooding or even make use of periodic updated messages. Since nodes in a MANET are likely to be mobile, in case a route in use dissolves due to the rapidly changing topology, a routing protocol for an ad hoc network should be able to find an alternate route very quickly.

7.3 Applications of MANETs

There is at present a widespread interest in MANETs among practitioners and researchers due to the significant advantages they offer for certain types of applications. A MANET can be set up quickly since no fixed infrastructures need to be deployed. Thus, in any situation where fixed infrastructure becomes difficult to be set up because of security, cost, inaccessibility of the terrain, or safety-related reasons, ad hoc networks become the preferred choice. Of the large number of applications that are possible with MANETs, a few example applications are defence-related operations and disaster management applications. Some example applications for which ad hoc networks are being used are described below.

Communication among portable computers

Miniaturization has allowed the development of many types of portables and computerized equipment, which have become very popular. Many of these portables work meaningfully when connected to some network, possibly a LAN or the Internet. For this, the portables are typically required to be within the range of some wireless hub. Satisfaction of this requirement would, however, drastically reduce the flexibility and the mobility of the devices. As an example, consider a lecture room where no networking infrastructures exist. In this case, using MANET the audience can exchange

notes, and also can surf the Web if at least one of the hand-held devices has access to Internet, for example, through a data card. If the mobile devices are present in sufficient density, network connections among them can be established seamlessly to form a MANET over which the nodes can communicate and carry out the network operations.

Environmental monitoring

A popular category of applications of MANETs is the collection of the various types of data about the environment in which they are deployed. Continuous data collection from remote locations is considered important for several applications such as environmental management, security monitoring, road traffic monitoring and management, etc. Miniaturized sensors have proved to be an effective means of gathering environmental information such as rainfall, humidity, presence of certain animals, etc. As planting sensors in the environment is becoming common, ad hoc networking is gaining more ground and becoming more prevalent. In this environmental monitoring application, a large number of sensors nodes are deployed in the environment. Such ad hoc sensor networks can be deployed to collect data from remote locations and the sensor nodes can even respond to some commands issued by the data collection centre. MANETs efficiently handle the introduction of new sensors into an already operational sensor network as well as can handle dynamic disconnections of nodes. Since each sensor acts as a hub, the range over which the sensors can be spread is tremendously increased, because they do not have to be deployed around some central station. For this application, use of energy efficient protocols becomes imperative since it can help increase the life span of the network.

Military

The present-day military equipment have become quite sophisticated, have many automated parts and contain one or more computers. This opens up the scope of setting up an ad hoc network consisting of various military equipment deployed in a frontline battle field. Ad hoc networking of these equipment can allow a military setup to take advantage of an information network among the soldiers, vehicles, and military information headquarters. For example, an ad hoc network can be automatically set up at a battlefront among the equipment, and the hand-held devices can collect information from and disseminate command to the frontline personnel.

Emergency applications

Ad hoc networks do not require any pre-existing infrastructure. These networks, therefore, can be deployed easily and rapidly in emergency situations such as a search and rescue operation after a natural disaster, and for applications such as policing and fire fighting. In these situations,

ad hoc networks can be set up on the fly. As an example, consider the situation where a severe earthquake has hit a locality. All the communication infrastructure would be destroyed and it would not be possible to immediately re-establish communication using the traditional infrastructure-based networks such as telephones, mobile phones, Internet, etc. In such a scenario, an ad hoc network can be quickly set up to provide network connectivity to rescue personnel in order to facilitate the rescue operation.

7.4 MANET Design Issues

Before we discuss some important MANET protocols, we point out below a few important issues that are relevant to the design of suitable MANET protocols.

Network size and node density

Network size and node density are the two important parameters of a MANET that need to be considered while designing an appropriate routing protocol for a network. Network size refers to the geographical coverage area of the network and network density refers to the number of nodes present per unit geographical area. For larger networks, clustering is essential to keep the communication overheads low. The cluster size as well as a specific clustering solution for a network would, to a large extent, depend on node density.

Connectivity

The term connectivity of a node usually refers to the number of neighbours it has. Here a neighbour of a node is one that is in its transmission range. The term connectivity is also sometimes used to refer to a link between the two nodes. The term *link capacity* denotes the bandwidth of the link. In a MANET, both the number of neighbouring nodes and the capacities of the links to different neighbours may vary significantly.

Network topology

The topology of a network denotes the connectivity among the various nodes of the network. Mobility of the nodes affects the network topology. Due to node mobility, new links can form and some links may get dissolved. Other than mobility, nodes can become inoperative due to discharged batteries or hardware failures, and thereby cause changes to the topology. The rate at which the topology changes needs to be appropriately considered in the design of an effective network.

User traffic

The design of a MANET is carried out primarily based on the anticipated node density, average rate of node movements, and the expected traffic. The traffic in a network can be of various types. A network protocol should leverage the characteristics of specific traffic types that are expected to improve its performance. The common traffic types are the following:

- Bursty traffic
- Large packets sent periodically
- Combination of the above two types of traffic

Operational environment

The operational environment of a mobile network is usually either urban, rural and maritime. These operational environments support the Line of Sight (LOS) communication. But, there can be a significant difference in the node density and mobility values in different operational environments, requiring different designs of mobile networks to suit an operational environment.

Energy constraint

As already mentioned, no fixed infrastructure exists in a MANET; the mobile nodes themselves store and forward packets. This additional role of mobile nodes as routers leads to nodes incurring perennial routing-related workload and this consequently results in continual battery drainage. Though this overhead is indispensable if the network is to be kept operational, the energy spent can be substantially reduced by allowing the nodes to go into a sleep mode whenever possible.

7.5 Routing

As we have already discussed, packet routing is usually a much more complex task in an ad hoc network compared to that of an infrastructure-based network. The main complications arise on account of continual topology changes and limited battery power of the nodes. Recall that we discussed these issues in Section 7.4 and a few other issues that are inherent to MANETs. When the destination node is not in the transmission range of the source node, the route has to be formed with the help of the intervening nodes in the network.

As we know, the purpose of routing is to find the best path between the source and the destination for forwarding packets in any store-and-forward network. In a traditional network, routing is a relatively easy task because the routes to nodes can be uniquely and efficiently identified based

on the subnet structure encoded in IP. In a MANET, the nodes making up a route may themselves move or shut down due to low battery energy, in the process making the knowledge about routes at various nodes to quickly become obsolete. It is therefore necessary to find a new route each time a node needs to transmit a message, making routing an expensive and difficult task.

Based on the above discussions, it may be seen that:

- Traditional routing protocols would not be suitable in an ad hoc network.
- Each node in an ad hoc network needs to have routing capability and also needs to participate in routing to keep the network operational.

We can now state that whenever there is an incoming packet in a MANET:

(a) Forward the packet to the next node (hop).
(b) While forwarding the packet, the sender needs to ensure that:
  (i) The packet moves towards its destination.
  (ii) The number of hops/path length is minimized.
  (iii) Delay is minimized.
  (iv) The packet loss is minimized.
  (v) The packet does not move around the network endlessly.

Several types of routing protocols have been proposed for MANETs. Different routing protocols essentially implement the above steps (a) and (b) while meeting the constraints inherent to the network, such as low energy consumption, through the deployment of various techniques. We will now review the essential concepts of a traditional routing technique. Later, we will build upon these concepts to introduce the routing protocols for ad hoc networks. No simple IP-address based routing is possible in a MANET due to the continual topology changes on account of node movements.

7.6 Essentials of Traditional Routing Protocols

Before understanding the MANET routing protocols, it is necessary to have a clear understanding of the routing mechanisms deployed in a traditional network. It will help us appreciate the specific changes made to traditional routing protocols to support the specific requirements of an ad hoc network. Two important classes of routing protocols for traditional networks are the *link state* and the *distance vector*. These two protocols are extremely popular in packet-switched networks. Both these protocols require a node to determine the next hop along the "shortest path" towards a given destination. The shortest path is computed according to some specific cost metric such as the number of hops in the route.

7.6.1 Link State Protocols (LSP)

The term *link state* denotes the state of a connection of one router with one of its neighbours. A neighbour of a router is one with which it can directly communicate without taking any help from the intervening routers. Each router determines its local connectivity information, and floods the network with this information with a *link state advertisement*. As a router in the network receives this link state advertisement, it stores this packet in a link state packet database (LSPDB). This storage of link state advertisements in an LSPDB is in addition to the routing table that each router maintains. It is easy to see that all routers in the network will have identical LSPBDs. Based on the bits and pieces of information stored in its LSPDB, each router constructs the connectivity information for the entire network as a graph using the Dijkstra's shortest path algorithm. Once a router constructs this graph, it computes the routing table from this and uses it in all its routing decisions. Thus, we can say that a router in the link state protocol bases its routing decisions on messages (link state advertisements) received from other routers in the network regarding their link states or the state of its connectivity with other routers.

A basic characteristic of a link state routing protocol is that every router constructs a graph representing the connectivity between the various nodes in the network based on the information received from other routers. We can think of this construction of the map of the entire network from bits and pieces of information received from other routers as similar to the solution of a zig-saw puzzle by putting together the different pieces of the puzzle. This graph representing the network is usually constructed in the form of a tree with the local router forming the root of the tree. The graph captures the shortest path route from the root to any other router. Once a node constructs this tree, it computes the best path from itself to every other node in the network and stores this information in the form of a routing table. This contrasts with the distance-vector routing protocols, in which each router shares its routing table with its neighbours. But in a link state protocol, only the connectivity related information is exchanged between routers, and no complete routes are exchanged.

In a link state protocol, each router periodically determines the state of its links to its neighbours by exchanging *hello packets* with them across all its network interfaces. Figure 7.2 shows how a router is connected to other routers through links established by its network interfaces. Based on the reply received from its neighbours, the router determines the state of the link in terms of the delay and other characteristics. Subsequently, the router forms a short message called the *link state advertisement* and sends it to its neighbours. A link state advertisement is also sent whenever a router experiences a connectivity change. A link state advertisement message contains:

- The identity of the router originating the message.
- The identities of all its neighbours.

- The delays along various links to its neighbours.
- A unique *sequence number*, which is formed by increasing the count every time the router forms a new link state advertisement.

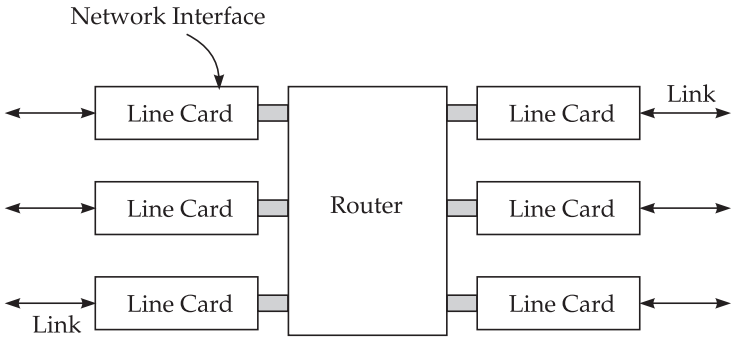


Figure 7.2 *Schematic diagram of a router.*

This link state advertisement is then flooded throughout the network as follows: A router sends a copy of a link state advertisement to all of its neighbours. A router receiving this message examines the sequence number of the last link state advertisement from the originating router by consulting its LSPDB. If this received link state advertisement is more recent, it replaces the last message with the currently received message in its LSPDB, and also forwards a copy of this link state advertisement to each of its neighbours.

Using the latest link state advertisements stored in its LSPDB, a router can easily reconstruct the network topology in the form of a tree by using the Dijkstra's iterative shortest path algorithm that we discuss below. This algorithm constructs the shortest path tree edge by edge, at each step adding one new edge, corresponding to the construction of the shortest path to a router. During the construction of the link state tree, if there are any inconsistencies among the reported link state advertisements from different routers, then the same have to be resolved. If one router reports that it is connected to another, but the other node does not report that it is connected to the first, then that link is not included on the tree.

Let us now discuss the construction of the link state tree for a router from the link state advertisements stored in its LSPDB. A router maintains two data structures: a tree containing nodes which are *done*, and a list of *candidates*. This tree is essentially a shortest path first (SPF) tree. The algorithm starts with both data structures empty. It first adds itself to the tree and thus is at the root. Then a greedy iterative algorithm based on the Dijkstra's algorithm is deployed to repetitively carry out the following by examining the link state advertisements that a router received in the past.

All routers which are connected to the router just added to the tree, excepting any routers which are either already in the tree or in the candidate list, are added to the candidate list.

The delays from each router in the candidate list to every other router in the tree are compared. The candidate router having the shortest delay is moved into the tree and attached to the appropriate neighbour router. Whenever a router is moved from the candidate list into the tree, it is removed from the candidate list.

The above two steps are repeated till there are no more routers left in the candidate list. If different routers somehow have maps that are inconsistent, then *routing loops* can form. In the simplest case of formation of a routing loop, two neighbouring routers determine that the other is the best path to a given destination, and therefore packets traverse endlessly between them. Routing loops involving more than two nodes are also possible.

The reason for the formation of routing loops is fairly simple. Each node computes its shortest-path tree and its routing table without interacting in any way with any other node. Therefore, if two nodes start with different maps, it is easy to have scenarios in which routing loops involving multiple nodes are created. Examples of these situations can be easily constructed and hence are not discussed here.

Once the network topology has been determined in the form of a shortest path tree, a router forms its routing table and uses it to find the best route to any destination. Based on this, it can determine the optimal next hop for each destination in the network. Two widely used link state protocols in traditional networking are OSPF (Open Shortest Path First) and IS-IS (Intermediate System To Intermediate System).

7.6.2 Distance Vector (DV) Protocols

The distance vector protocols get their name from the fact that they base their routing decisions on the distance to the destination in terms of the number of hops that a packet will have to traverse to reach its destination. You might wonder what exactly the term *vector* denotes. The term vector here means that routes are advertised as a vector (distance, direction), where distance is the number of hops between the two nodes and direction is defined in terms of the next hop router to which the packets need to be forwarded. The distance vector protocols are based on the well-known Bellman-Ford algorithm.

The distance vector routing protocols share everything they know about the various routes in the network with their neighbours by broadcasting their entire route table. Each node advertises its entire routing table to its immediate neighbours only. A router transmits its routing table that has been formed from its own perspective. That is, it represents the routes to various routers from itself. For example, "router A is at a distance of five hops away, in the direction of the neighbour router X." As this statement implies, each router learns routes from its neighbouring routers' perspectives, and based on this forms its own perspective of the routes and

then advertises the routes from its own perspective. As a router receives the routing information of its neighbouring nodes, it updates its own routing table by examining the received information and in turn informs its own neighbours of the changes. This is also referred to as "routing by rumour" because routers are relying on the information they receive from other routers and have no way to determine if the information is actually valid. A number of techniques have been suggested to cope with instability and inaccurate routing information.

As we have seen, the routers using the distance vector protocol do not have knowledge of the entire path that a packet would take to reach its destination. Instead, they just know the following vector:

1. Direction in (or the specific network interface over) which a packet should be forwarded.
2. Its own distance from the destination.

The DV protocol is, therefore, based on calculating the distance and the direction to any router in a network. The cost of reaching a destination is calculated using the various route metrics. The two popular distance vector routing protocols are RIP (Routing Information Protocol) and IGRP (Interior Gateway Routing Protocol). RIP uses the hop count of the destination whereas IGRP takes into account the other information such as node delay and available bandwidth. RIP supports the cross-platform distance vector routing, whereas IGRP is a Cisco Systems proprietary distance vector routing protocol.

As we have already mentioned, the distance vector routing protocol requires each router to periodically send its entire route tables to all its neighbours. A notable exception to this convention is the Cisco's Enhanced IGRP (EIGRP). Though EIGRP is a distance vector protocol, it does not require transmitting updates periodically. Further, the updates are not broadcast and do not contain the full route table.

7.7 Routing in MANETs: A Few Basic Concepts

In this section, we present a few basic concepts and definitions, based on which we will discuss the MANET routing protocols. We first highlight a few important differences of MANET routing from the traditional routing protocols that are deployed in wired networks. This will help us to better appreciate the routing protocols in MANETs.

7.7.1 Routing in MANETs vs. Routing in Traditional Networks

The following are the three important ways in which a MANET routing protocol differs from routing of packets in a traditional network.

- In a MANET, each node acts as a router, whereas ordinary nodes in a traditional wired network do not participate in routing the packets.
- In a MANET, the topology is dynamic because of the mobility of the nodes, but it is static in the case of traditional networks. Thus, the routing tables in a MANET quickly become obsolete, making the routing process complicated.
- In the simple IP-based addressing scheme deployed in wired networks, the IP address encapsulated in the subnet structure does not work because of node mobility.

In order to cope with the above three important differences, each MANET node needs to carry out the following important routing tasks: route discovery and route maintenance. We explain these tasks in the subsequent sections.

Types of communications

In a network, a node can initiate the following types of communications:

Unicast: In this, a message is sent to a single destination node.

Multicast: In this type of transmission, a message is sent to a selected subset of the network nodes.

Broadcast: In this type of transmission, a message is sent to all the nodes in the network. Since unrestrained broadcast communications can choke a MANET, applications usually do not use broadcast communication.

As we shall see, different protocols support different types of communications with different levels of proficiency. The exact routing protocol to be deployed in a network is determined based on the results of a traffic analysis characterizing the transmissions in the network into unicast, multicast, and broadcast. Since the broadcast transmission is normally not deployed in a MANET, the routing protocols in a MANET can broadly be classified into unicast and multicast types.

7.7.2 A Classification of Unicast MANET Routing Protocols

Unicast routing protocols in MANETs are classified into proactive (table-driven), reactive (on-demand) and hybrid protocols. This classification is based on how a protocol manages to determine the route correctly in the presence of topology changes.

Proactive (table-driven) protocols

A proactive routing protocol is also known as a *table-driven* routing protocol. In this protocol, each node in a routing table maintains information about routes to every other node in the network. These tables are periodically updated in the face of random network topology changes. An example of

a proactive (table-driven) protocol is the Destination Sequenced Distance Vector (DSDV) protocol. Since each node knows the complete topology, a node can by itself determine the best route to a destination based on its local information. Since the topology can change due to node movements and nodes shutting down, the routing table needs to be updated periodically. Therefore, a proactive protocol usually generates a large number of control messages to keep the routing tables up-to-date at different nodes. This routing overhead may take up a large part of the available bandwidth. Especially for networks with a large number of nodes and very high node mobility, the control messages may consume almost the entire bandwidth. Obviously, this protocol is not suitable for large networks as the size of the routing table may become excessively large due to a node maintaining routes to each and every other node in the network.

Reactive (on-demand) protocols

A reactive routing protocol is also known as an on-demand routing protocol, since in this protocol nodes do not maintain up-to-date routes to different destinations, and new routes are discovered only when required. When a node does not have knowledge about any route to a specific destination, it uses a flooding technique to determine the route. These protocols were designed to reduce the large overheads incurred by the proactive protocols. This efficiency is achieved by maintaining information for active routes only and doing away with a large number of route update messages. Two examples of on-demand routing protocols are:

(i) Dynamic source routing (DSR)
(ii) Ad hoc on-demand distance vector routing (AODV)

Hybrid routing protocols

Hybrid routing protocols have the characteristics of both proactive and reactive protocols. These protocols combine the good features of both the protocols. The hybrid routing protocols are designed to achieve increased scalability by allowing nodes with close proximity to work together to form some sort of a backbone to reduce the route discovery overheads. This is mostly achieved by proactively maintaining routes to nearby nodes and determining routes to far away nodes only when required using a route discovery strategy. Most hybrid protocols proposed to date are zone-based, which means that the network is partitioned or seen as a number of routing zones by each node. An example of a hybrid routing protocol is the Zone Routing Protocol (ZRP).

7.8 Popular MANET Routing Protocols

Based on the classification of the routing protocols discussed in Section 7.7.2, we now discuss a few popular MANET routing protocols.

7.8.1 Destination-Sequenced Distance-Vector Routing Protocol

Destination-Sequenced Distance-Vector Routing (DSDV) is an important MANET routing protocol. It is based on the table-driven (proactive) approach to packet routing (Barbora, 1999). It extends the distance vector protocol of wired networks just as the traditional algorithm makes uses of the classical Bellman–Ford routing algorithm. An improvement made here is the avoidance of routing loops through the use of a number sequencing scheme. In DSDV, each node in a MANET maintains a routing table in which all of the possible destinations and the number of hops to each destination are recorded. Hence, routing information is always readily available, regardless of whether the source node requires a specific route or not.

Each node maintains information regarding routes to all the known destinations. The routing information is updated periodically. This can be considered a shortcoming of the protocol since it deprives a node from going into sleeping mode. Also, there is a traffic overhead even if there is no change in network topology. Further, nodes maintain routes which they may never use.

A sequenced numbering system is used to allow mobile nodes to distinguish stale routes from new ones. Updated routing tables are exchanged periodically among the nodes of the network to maintain table consistency. A naïve table exchange approach would generate a lot of control traffic in the network, leading to inefficient utilization of the network resources. To alleviate this problem, DSDV uses two types of route update packets. The first is known as *full dump*. This type of packet carries all the available routing information and can require multiple network protocol data units (NPDUs) to be transmitted. During periods of occasional movement, these packets are transmitted infrequently. Smaller incremental packets are used to disseminate only the information that has changed since the last full dump. These incremental broadcasts usually fit into a standard NPDU, thereby decreasing the amount of traffic generated. The mobile nodes maintain an additional table where they store the data received through the incremental routing information packets from various nodes.

New route broadcasts contain the address of the destination, the number of hops to reach the destination, as well as a unique sequence number. The route labelled with the most recent sequence number is always used. Up on a change, a node might receive several messages from different sources. The weighted average time that routes to a destination will fluctuate before the route with the best metric received. Mobiles also keep track of this settling time of routes. By delaying the broadcast of a routing update by the length of the settling time, mobiles can reduce network traffic and optimize routes by eliminating the suboptimal ones.

Important steps in the operation of DSDV

The important steps in the operation of DSDV are summarized below:

1. Each router (node) in the network collects route information from all its neighbours.
2. After gathering information, the node determines the shortest path to the destination based on the gathered information.
3. Based on the gathered information, a new routing table is generated.
4. The router broadcasts this table to its neighbours. On receipt by neighbours, the neighbour nodes recompute their respective routing tables.
5. This process continues till the routing information becomes stable.

DSDV incorporates a sequenced numbering scheme. Each routing advertisement comes with a sequence number. Within an ad hoc network, advertisements may propagate along many paths. Sequence numbers help a node to consider the advertisements in the correct order. This avoids the loops that may form while using the unchanged distance vector algorithm.

Figure 7.3 shows an example of a MANET. Table 7.1 is the routing table of the node N_4 at the moment before the movement of nodes . The metric field in the routing table helps to determine the number of hops required for a packet to traverse to its destination. The install time indicates when the entry was made. It is used to delete stale entries from the table.

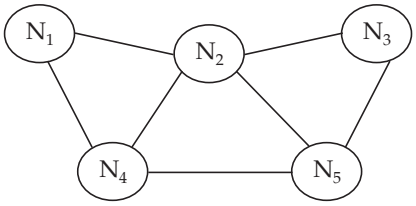


Figure 7.3 *An example of a MANET topology at a given instant of time.*

TABLE 7.1 DSDV Routing Table for the MANET of Figure 7.3 at Node N_4

Destination	*Next hop*	*Metric*	*Sequence no.*	*Install time*
N_1	N_1	1	321	001
N_2	N_2	1	218	001
N_3	N_2	2	043	002
N_5	N_5	1	163	002

7.8.2 Dynamic Source Routing (DSR) Protocol

Dynamic Source Routing (DSR) (Johnson and Maltz 1996, 2001) protocol was developed to be suitable for use in a MANET having a reasonably small diameter of about 5 to 10 hops and when the nodes do not move very

fast. DSR is a source initiated on-demand (or reactive) routing protocol for ad hoc networks. It uses source routing, a technique in which the sender of a packet determines the complete sequence of nodes through which a packet has to travel. The sender of the packet then explicitly records this list of all nodes in the packet's header. This makes it easy for each node in the path to identify the next node to which it should transmit the packet for routing the packet to its destination.

In this protocol, the nodes do not need to exchange the routing table information periodically, which helps to reduce the bandwidth overhead associated with the protocol. Each mobile node participating in the protocol maintains a *routing cache* which contains the list of all routes that the node has *learnt*. Whenever a node finds a new route, it adds the new route to its *routing cache*. Each mobile node also maintains a sequence counter called *request id* to uniquely identify the last request it had generated. The pair < source address, request id > uniquely identifies any request in the ad hoc network. DSR works in two phases: (i) Route discovery and (ii) Route maintenance. We discuss these two phases in the following.

Route discovery

Route discovery allows any host to dynamically discover the route to any destination in the ad hoc network. When a node has a data packet to send, it first checks its own routing cache. If it finds a valid route in its own routing cache, it sends out the packet using this route. Otherwise, it initiates a route discovery process by broadcasting a route request packet to all its neighbours. The route request packet contains the source address, the request id and a route record in which the sequence of hops traversed by the request packet, before reaching the destination is recorded.

A node upon getting a route request packet does the following. If a packet does not have the required route in its routing cache, it forwards the packet to all its neighbours. A node forwards a *route request* message only if it has not yet seen it earlier, and if it is not the destination. The *route request* packet initiates a *route reply* upon reception either by the destination node or by an intermediate node that knows a route to the destination. Upon arrival of the *route request* message at the destination, this information is piggybacked on to the *route reply* message that contains the path information and is sent to the source node.

The route discovery process is schematically shown in Figure 7.4. As shown in the figure, suppose a node N_1 wishes to send a message to the destination node N_8. The intermediate nodes are N_2, N_3, N_4, N_5, N_6, N_7. The node N_1 initiates the route discovery process by broadcasting a *route request* packet to its neighbours N_2 and N_3. Note that each node can have multiple copies of the route request packet arriving at it. The propagation of route reply is shown in Figure 7.5, and the acknowledgement messages from destination to source are indicated by thick arrows.

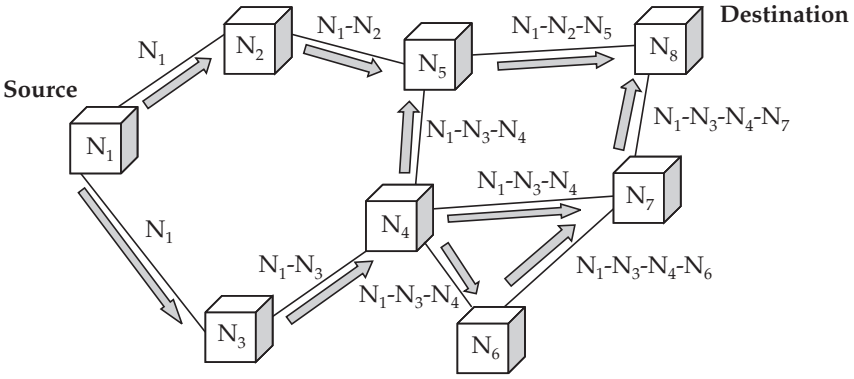


Figure 7.4 *An example of the route discovery process in DSR.*

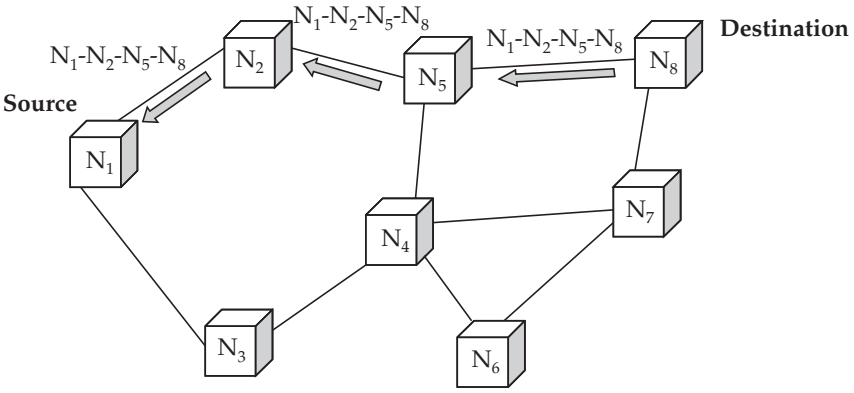


Figure 7.5 *An example of the propagation of route reply in DSR.*

Route maintenance

A known route can get broken either due to the movement of some nodes making up the route or the battery of a node forming part of the route getting exhausted. Route maintenance is the process of monitoring the correct operation of a route in use and taking any corrective action when needed. When a host (source) while using a route, finds that it is inoperative, it carries out route maintenance. Whenever a node wanting to send a message finds that the route is broken, it would help if it already knows of some alternative routes. Since the nodes do not exchange any routing information in this protocol, whenever a node detects that one of its next hop neighbour node is not responding, it sends back a route error packet containing its own address and the address of the hop that is not working. As soon as the source node receives the RouteError message, it deletes the broken-link-route from its cache. If it has another route to the destination, it starts to retransmit the packet using the alternative route. Otherwise, it initiates the route discovery process again.

7.8.3 Ad Hoc On-demand Distance Vector (AODV)

The route discovery and route maintenance activities in AODV (Perkins et al., 2001) are very similar to those for the DSR protocol. AODV does make use of hop-by-hop routing, sequence numbers and beacons. The node that needs a route to a specific destination generates a *route request*. The *route request* is forwarded by intermediate nodes which also learn a reverse route from the source to themselves. When the request reaches a node with route to destination, it generates a *route reply* containing the number of hops required to reach the destination. All nodes that participate in forwarding this reply to the source node create a forward route to destination. This route created from each node from source to destination is a hop-by-hop route. Recollect that DSR includes the complete route in packet headers. The large headers can substantially degrade the performance, especially when the data content of packets is small. AODV attempts to improve upon DSR by maintaining routing tables at the nodes, so that the data packets do not have to contain the routes. AODV retains a positive feature of DSR, in that the routes are maintained only between those nodes that need to communicate. If a link break occurs while a route is being used to transmit a message, a route error message is sent to the source node by the node that observes that the next link in the route has failed.

7.8.4 Zone Routing Protocol

The Zone Routing Protocol (ZRP) (Haas, 1997) is a hybrid protocol. It incorporates the merits of both on-demand and proactive routing protocols. A routing zone is similar to a cluster. However, unlike clusters, zones can overlap. A routing zone comprises a few MANET nodes within a few hops from the central zone. Within a zone, a table-driven routing protocol is used. This implies that regular route updates take place only within a zone. Each node, therefore, has a route to all other nodes within the zone. If a destination node happens to be outside the source's zone, ZRP employs an on-demand route discovery procedure which works as follows. The source node sends a route request to the border nodes of its zone, containing its own address, the destination address and a unique sequence number. Border nodes are those nodes which are some predefined number of hops away from the source. Each border node checks its local zone for the destination. If the destination is not a member of its local zone, then the border node adds its own address to the route request packet and forwards the packet to its own border nodes. When the destination node is reached in this process, a route reply is sent on the reverse path, back to the source. The source node uses the path saved in the route reply packet to send data packets to the destination.

7.8.5 Multicast Routing Protocols for MANET

As we have already discussed, multicast is the delivery of a message to a group of destination nodes in a single transmission as shown in Figure 7.6. For efficient operation of a multicast routing protocol, it is necessary to minimize the unnecessary packet transmissions as well as minimize the energy consumption. In order to achieve this, a multicast transmission should not be approximated by multiple unicast transmissions.

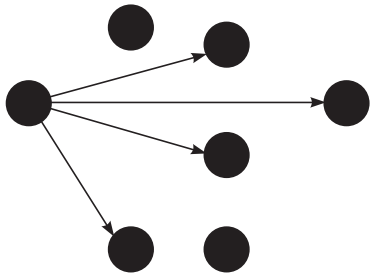


Figure 7.6 *Multicast transmission.*

Efficient multicast routing is much more difficult to achieve in a MANET compared to any other network. This is due to host mobility, broadcast nature of wireless environment, and interferences from various noise sources. The proposed MANET multicast routing protocols either modify the conventional tree structure, or deploy a different topology between group members. The popular MANET multicasting protocols are either tree-based or mesh-based:

Tree-based protocol

Tree-based schemes establish a single path between any two nodes in the multicast group. These schemes require minimum number of copies per packet to be sent along the branches of the tree. Hence, they are bandwidth efficient. However, as mobility increases, link failures trigger the reconfiguration of the entire tree. Another complicacy is that when there are many sources, a node either has to maintain a shared tree, losing path optimality, or maintain multiple trees resulting in storage and control overhead. An examples of this category of protocol is the Multicast Ad Hoc On-demand Distance Vector (MAODV) (Royer and Perkins, 1999) Routing Protocol.

Mesh-based protocol

Mesh-based schemes establish a mesh of paths that connect the sources and destinations. They are more resilient to link failures as well as to mobility. The major disadvantage of this scheme is that multiple copies of the same packet are disseminated through the mesh, resulting in reduced packet

delivery and increased control overhead under highly mobile conditions. An examples of this category of protocol is the On-demand Multicast Routing Protocol (ODMRP) (Lee et al., 2002).

7.9 Vehicular Ad Hoc Networks (VANETs)

A Vehicular Ad Hoc Network (VANET) is a special type of MANET in which moving automobiles form the nodes of the network. VANETs were initially introduced for vehicles of police, fire brigades, and ambulances for safe travelling on road. In this network, a vehicle communicates with other vehicles that are within a range of about 100 to 300 metres. Multi-hop communication often results in rather large networks. In a city or a busy highway, the diameter of the network can be several tens of kilometres. Any vehicle that goes out of the signal range of all other vehicles in the network is excluded from the network. A vehicle that was outside the communication range of all other vehicles of a VANET can come in the range of a vehicle that is already in the network and as a result can join the network.

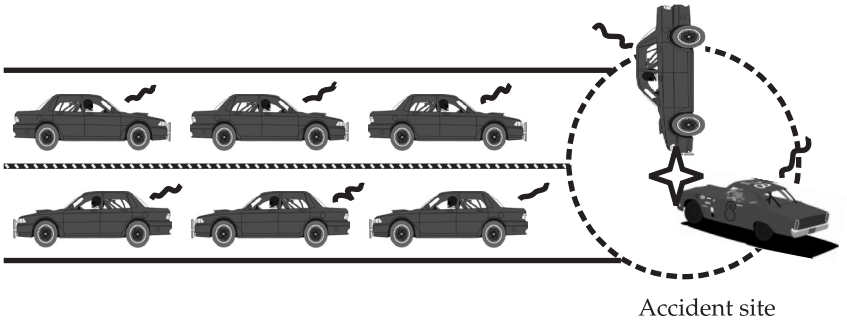


Figure 7.7 *A VANET use scenario.*

A VANET can offer a significant utility value to a motorist. A few important uses of a VANET are described below:

- A VANET can help drivers to get advance information and warnings from a nearby environment via message exchanges. For example, in the scenario shown in Figure 7.6, two vehicles are involved in a collision. The trailing vehicles get advance notification of the collision ahead on the road. The driver can also get advance information on the road condition ahead, or a warning about the application of emergency electronic brake by a vehicle ahead in the lane.
- A VANET can help disseminate geographical information to the driver as he continues to drive. For example, the driver would be notified of the nearby food malls or petrol refilling stations, map display, etc.

- Drivers may have the opportunity to engage in other leisurely tasks, such as VoIP with family, watch news highlights, listen to series of media files known as podcasts, or even carry out some business activities such as participate in an office video conference session.

7.10 MANET vs VANET

A MANET, as we have already defined, is a collection of mobile nodes that communicate with each other over bandwidth constrained wireless links without any infrastructure support. In this sense, we can consider a VANET to be a special category of MANET. The nodes are mobile in VANETs as well as in MANETs. However, the VANET nodes (vehicles) can communicate with certain roadside infrastructures or base stations. Further, the node mobility in a VANET is constrained to the road topologies, whereas the movement of nodes in a MANET is more random in nature. Considering that vehicles move over large distances at relatively high speeds, a VANET undergoes fast topological changes. Another important difference is that in a MANET, power is a major constraint but in VANET the battery power available in a vehicle is quite adequate. The issues such as the relatively larger size of VANETs compared to MANETs and the relatively high speed with which vehicles move, need to be appropriately considered for the design of an effective VANET.

7.11 Security Issues in a MANET

As discussed earlier, MANETS are fundamentally different from both wired networks and infrastructure-based wireless networks. The nature of MANETs not only introduces new security concerns but also exacerbates the problem of detecting and preventing anomalous behaviour. In a wired network or in an infrastructure-based wireless network, an intruder is usually a host that is outside the network and therefore could be controlled through a firewall and subjected to access control and authentication. In a MANET, on the other hand, an intruder is part of the network, and therefore much more difficult to detect and isolate.

Dynamic topological changes and the inherent wireless communications in a MANET, make it vulnerable to different types of attacks. At the physical layer, an intruder can easily cause jamming or overload the available network resources beyond their capacities, thereby effectively paralysing it. Wireless links can get jammed and the batteries at the nodes can get depleted by such overloading, causing breakdowns of the network. Attackers can also disturb the normal operation of routing protocols by modifying the headers of packets. The intruder may insert spurious

information while routing packets, causing erroneous routing table updates and thereby leading to frequent misroutings.

A few important characteristics of ad hoc networks that can be exploited to cause security vulnerabilities are the following:

Lack of physical boundary: Each mobile node functions as a router and forwards packets from other nodes. As a result, network boundaries become blurred. The distinction between nodes that are internal or external to a network becomes meaningless, making it difficult to deploy firewalls or monitor the incoming traffic.

Low power RF transmissions: It is possible for a malicious node to continuously transmit and monopolise the medium and cause its neighbouring nodes to wait endlessly for transmitting their messages. Also, signal jamming can lead to a denial-of-service (DoS) attack.

BOX 7.1 Security issues in MANETs

Simple attacks on a MANET may be in the form of broadcasting false information or stealing information by eavesdropping. The injection of false or malicious messages into a network can have serious repercussions. More destructive attacks can be orchestrated by trojan horses and viruses bringing down almost all the nodes in the network. There can be physical attacks as well. These include wiretapping, theft of equipment, and storage media.

In contrast to wired networks where the network devices are usually kept behind locked doors, ad hoc network equipment is usually carried around as small battery-powered devices. This makes them more vulnerable for attacks, since they are easier to capture and carry away. Again it can be quite hard to intercept wired media without getting noticed because the media itself might be hard to get to and to intercept signals in the cables without drawing attention is difficult. In the wireless medium, tapping is as easy as just putting up an antenna, usually small enough to be noticed.

Limited computational capabilities: Nodes in an ad hoc network usually have limited computational capabilities. It therefore becomes difficult to deploy compute-intensive security solutions such as setting up a public-key cryptosystem. Inability to encrypt messages invites a host of security attacks such as spoofing as well as several forms of routing attacks.

Limited power supply: Since nodes normally rely on battery power, an attacker might attempt to exhaust batteries by causing unnecessary transmissions to take place or might cause excessive computations to be carried out by the nodes.

Characteristics of secure ad hoc networks

We discuss below a few important characteristics of a secure ad hoc network. Different types of attacks on the network attempt to breach one

or more of these security features. A security solution should ensure that these characteristics are not compromised. A secure ad hoc network should have the following characteristics:

- *Availability:* It should be able to survive denial-of-service (DoS) attacks.
- *Confidentiality:* It should protect confidentiality of information by preventing its access by unauthorized users.
- *Integrity:* It should guarantee that no transferred message has been tampered with.
- *Authentication:* It should help a node to obtain guarantee about the true identity of a peer node.
- *Non-repudiation:* It should ensure that a node having sent a message, cannot deny it.

7.12 Attacks on Ad Hoc Networks

Out of the large varieties of attacks that are possible in a MANET, many are essentially routing attacks. Several other attacks target the wireless transmission in a MANET, especially the network connectivity. In this section, we first discuss a classification of various types of MANET security attacks and then discuss a few types of attacks that are rather common. Various attacks on a MANET can be classified into passive and active attacks, based on the means of attack. Passive attacks target to monitor and steal the data exchanged in the network, without disrupting the network operations. It becomes very difficult to identify these attacks since these do not have any perceivable symptoms. These attacks can be reduced by using suitable encryption techniques. An active attack, on the other hand, is destructive and disturbs the normal functionality of the network. Table 7.2 shows a classification of various types of security attacks against MANETs into active and passive types. We discuss these different types of attacks in this section. Attacks can also be classified according to the specific layer of the protocol stack that they target as shown in Table 7.3. Each attack can be considered to be exploiting the vulnerabilities at one or more layers of the MANET protocol stack while most attacks target certain security vulnerabilities at specific protocol layers. The multilayer attacks are those that exploit the vulnerabilities existing at more than one protocol layer. We discuss a few of the important types of security attacks in the following.

TABLE 7.2 Passive and Active Attacks in a MANET

Passive attacks	*Active attacks*
Snooping, eavesdropping, traffic analysis, monitoring	Wormhole, black hole, grey hole, resource consumption, routing attacks

TABLE 7.3 Attacks at Different Layers of a MANET Protocol Stack

Layer	*Attacks*
Application layer	Malicious code, repudiation, data corruption
Transport layer	Session hijacking, SYN flooding
Network layer	Wormhole, black hole, fabrication attack
Data link layer	Resource consumption
Physical layer	Traffic analysis, monitoring, disruption, jamming, interceptions, eavesdropping
Multilayer	Denial-of-Service (DoS), impersonation, replay

Routing loop

By sending tampered routing packets, an attacker can create a routing loop. This will result in data packets being sent around endlessly, consuming bandwidth and causing dissipation of power for a number of nodes. In this attack, the packets are prevented from reaching their intended recipients and thus it can be considered to be a type of denial-of-service (DoS) attack.

Malicious code attacks

A malicious code can be a virus, worm, spyware, or a Trojan. In a MANET, an attacker can propagate malicious code and can slow down the nodes, overload the network, or even crash the nodes.

Repudiation attack

Repudiation attack refers to the denial of participation in a communication. In this attack, a malicious user can deny a credit card or bank transaction.

SYN flooding attack

In this attack, an attacker creates a large number of half-opened TCT connections with the victim nodes by sending a large number of SYN packets to them. This causes the TCP connection tables of the victim nodes to overflow.

Session hijacking

In a typical session, all the communications are authenticated only at the beginning of the session. The attacker can spoof the IP address of a node that has just started a session and hijack the session from the victim and perform a DoS attack.

Fabrication attack

In AODV routing, when a node detects a broken link while forwarding a packet (possibly because the next hop node has either moved or has shut

down), it sends a route error message towards the packet sender. In the fabrication attack, a malicious node sends a false route error message to the packet sender, even when the next hop link is not broken.

Black hole

In this type of attack, a node can set up a route to some destination via itself, and when the actual data packets are received from other nodes, these are simply dropped. This node forms a black hole, to which data packets enter but never leave.

Grey hole

A special case of the black hole attack is the grey hole attack. In this attack, the attacker selectively drops some kinds of packets that pass through it but not the others. For example, the attacker might forward routing packets but not the data packets. This type of attack is more difficult to detect compared to a black hole attack.

Partitioning

In this kind of attack, the attacker partitions a network by causing some nodes to split up from the other nodes. That is, one set of nodes is not able to communicate with the other set of nodes. By analysing the network topology the attacker can choose to make the partitioning between the set of nodes that causes the most harm to the system. This attack can be accomplished in many ways, such as by tampering routing packets as in the previous attacks. It can also be launched through some physical attack such as radio jamming.

Blacklist

This attack tries to exploit a loophole in security mechanisms. Some ad hoc routing protocols try to tackle this security problem by keeping a list of perceived malicious nodes. Each node has a blacklist of, what it thinks, bad nodes and thereby avoids using them when setting up routing paths. An attacker might try to get a good node blacklisted, causing the other good nodes to add this node to their respective blacklists and so avoid it.

Wormhole

In a wormhole attack, a direct link (tunnel) between the two nodes is established. This is referred to as *wormhole link*. The direct link can be established by making use of a wired line, a long-range wireless transmission, or an optical link. Through the wormhole link, one node eavesdrops messages at one end, and tunnells them through the wormhole link to the other node which then replays them. The tunnel essentially emulates a shorter route through the network and so naive nodes prefer

to use it rather than the alternative longer routes. Once a wormhole is established, a malicious node can use it for traffic analysis or make a denial-of-service attack by dropping certain data or control packets. When this attack targets specifically the routing control packets, the nodes that are close to the attackers are shielded from any alternative routes with more than one or two hops to the remote location.

Dropping routing traffic

It is essential that in an ad hoc network, all nodes participate in the routing process. However, it is possible that a node may act selfishly and process only the routing information that is related to itself either maliciously or to conserve energy. This behaviour/attack can create network instability or can even segment the network.

7.13 Security Attack Countermeasures

Countermeasures against the various types of attacks we discussed, have been proposed. We summarize the different security measures incorporated in the different layers of the network protocol in Table 7.4. Though cryptographic techniques are powerful techniques for ensuring confidentiality, authentication, integrity, and non-repudiation, these are ineffective against jamming. Spread spectrum technology such as frequency hopping is a promising countermeasure against the signal jamming type of attacks. In this technique, the transmission frequency changes randomly. Directional antennae can be deployed as a countermeasure against signal jamming.

TABLE 7.4 Security Measures at Different Protocol Layers

Layer in protocol stack	*Security measures incorporated*
Data link layer	Use of spread spectrum transmission and directional antennae
Network layer	Use of authentication measures and keeping track of the trusted nodes
Transport layer	Securing and authenticating end-to-end communications through data encryption techniques
Application layer	Detection and prevention of virus, worms, malicious code through code analysis.

Maintenance of a trust rating of various nodes is a promising technique to overcome many of the routing attacks. In this technique, every node maintains a trust rating of various nodes and packet transmission is carried out using a selective flooding technique.

SUMMARY

Mobile ad hoc networks are becoming very popular. Their use in every day applications is rapidly increasing. In this chapter, we briefly discussed the important characteristics of ad hoc networks and compared them with infrastructure-based wireless networks. We then discussed their various applications, the challenges in designing an effective MANET, and the importance of designing effective routing protocols in proper functioning of an ad hoc network. We identified why the routing task in a MANET is much more complex than that in a traditional network. We then discussed a few important routing protocols that have been proposed for use in MANETs. We discussed VANETs, as a special type of MANET, which are rapidly emerging as popular networks having many applications. Though MANETs are handy and appear to be a promising networking technology, they suffer from many more security vulnerabilities than the infrastructure-based networks do. To be practically useful, the security concerns for these networks need to be satisfactorily understood and addressed. We finally reviewed the security issues in MANETs and discussed the overall approaches adopted by a few security solutions.

FURTHER READINGS

Abolhasan, M., T. Wysocki, and E. Dutkiewicz, "A review of routing protocols for mobile ad hoc networks", *Ad Hoc Networks*, Vol. 2, pp. 1–22, 2004.

Aggelou, G. and R. Tafazolli, "RDMAR: A Bandwidth-efficient Routing Protocol for Mobile Ad Hoc Networks". In: *ACM International Workshop on Wireless Mobile Multimedia* (WoWMoM), pp. 26–33, 1999.

Belding-Royer, E.M. and C. Perkins, "Evolution and future directions of the ad hoc on-demand distance-vector routing protocol", *Ad Hoc Networks*, Vol. 1, pp. 125–150, 2003.

Haas, J., "A new routing protocol for the reconfigurable wireless networks", *Proc. of IEEE Int. Conf. on Universal Personal Communications*, pp. 562–566, 1997.

Jacquet, P., P. Muhlethaler, T. Clausen, A. Laouiti, A. Qayyum and L. Viennot, "Optimized link state routing protocol for ad hoc networks", *IEEE INMIC*, Pakistan, 2001.

Jiang, M., J. Ji, Y.C. Tay, "Cluster based routing protocol", *Internet Draft*, draft-ietf-manet-cbrp-spec-01.txt , 1999.

Johnson, B. David and David A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks". In: *Mobile Computing*, T. Imielinski and H. Korth, Eds., Kluwer Academics Publishers, pp. 153–181, 1996.

Johnson, D.B. and D.A. Maltz, "DSR: The Dynamic Source Routing Protocol for Multihop Wireless Ad Hoc Networks". In: *Ad Hoc Networking*, C.E. Perkins, Ed., Addison Wesley, Boston, 2001.

Lee, S., W. Su, and M. Gerla. "On-demand Multicast Routing Protocol in Multihop Wireless Mobile Networks", *Mobile Networks and Applications*, Vol. 7, Issue 6, December 2002.

Marina, M.K. and S.R. Das, "Routing Performance in the Presence of Unidirectional Links in Multihop Wireless Networks". In: *Proc. of the 3rd Symposium of Mobile Ad Hoc Networking and Computing* (MobiHoc), Lausanne, Switzerland, June 2002.

McQuillan, John M., Isaac Richer, and Eric C. Rosen, "The New Routing Algorithm for the ARPANet," *IEEE Trans. on Comm.*, 28(5), pp. 711–719, 1980.

Mohan, P. Madhan, Johnson J. James, Murugan K., and V. Ramachandran, "A Comparative and Performance Study of On-demand Multicast Routing Protocols for Ad Hoc Networks", College of Engineering, Guindy (CEG), Anna University.

Perkins, C.E. and T.J. Watson, "Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers". In: *ACM SIGCOMM'94 Conference on Communications Architectures*, London, UK, 1994.

Perkins, E. Belding-Royer and S. Das, "Ad hoc On-demand Distance Vector (AODV) Routing", *IETF Draft*, November 2001.

Royer, E.M. and C.E. Perkins, "Multicast Operation of the Ad Hoc On-demand Distance Vector Routing Protocol", *Proceeding of the fifth annual ACM/IEEE international conference on mobile computing and networking*, pp. 207–218, 1999.

Siva Ram Murthy, C. and B.S. Manoj, *Ad Hoc Wireless Networks: Architectures and Protocols*, Prentice Hall PTR, New Jersey, USA, May 2004.

Thomas Kunz and Ed Cheng, "On-demand Multicasting in Ad Hoc Networks: Comparing AODV and ODMRP", Carleton University.

Toh, C., "A novel distributed routing protocol to support ad hoc mobile computing". In: *IEEE 15th Annual International Phoenix Conf.*, pp. 480–486, 1996.

Zhang, Yongguang, Wenke Lee, and Yi-an Huang, "Intrusion detection for wireless ad hoc networks", *Wireless Networks* 9, pp. 545–556, 2003.

Zhou, L. and Z. J. Haas, "Securing Ad Hoc Networks", *IEEE Network* 13(6), pp. 24–30, 1999.

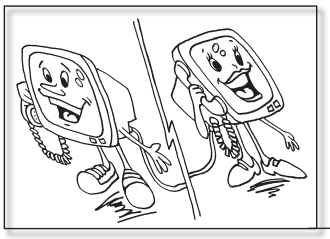
EXERCISES

1. What is an ad hoc network? Why the traditional routing strategies cannot be deployed in a MANET straightaway? Compare the MANET routing strategies with the routing strategies of traditional networks.
2. Define the characteristics of MANETs that are important determinants in the design of an effective ad hoc network.
3. What do you mean by dynamic topology of a MANET? What problems does it cause in the design of a routing protocol? How are these problems addressed in a popular MANET routing protocol?
4. Briefly explain why the traditional packet routing protocols for wired networks cannot be used straightaway in a MANET. Discuss how the routing protocols for traditional wired networks have been extended to work in a MANET.
5. Describe at least three applications of mobile ad hoc networks.
6. What are the important design constraints on a MANET? Explain the implications of these constraints.
7. Can cellular networks and wireless LANs be considered as ad hoc networks? Explain your answer.
8. Discuss why routing is a more challenging problem in MANETs when compared to wired networks. Explain the different types of routing protocols in MANETs.
9. What is "counting-to-infinity" problem? How is this problem addressed in a MANET?
10. What are the competing issues that are addressed by a routing protocol in a MANET? How are these achieved?
11. Briefly explain the important classes of MANET routing protocols and compare their relative advantages. Compare them with respect to network overhead, routing quality and routing time.
12. What do you mean by unicast, multicast and broadcast communications in a MANET? Give examples of typical applications of each of these types of communication.
13. What is the difference between reactive and proactive routing in MANETs? Give examples of each category of protocols.
14. What is a hybrid routing protocol? What is its advantage over the other classes of routing protocols?
15. Explain the working of the Destination-Sequenced Distance-Vector (DSDV) routing protocol using suitable examples.
16. Explain how the security threats in MANETs are different from a wired network.

17. Explain Dynamic Source Routing (DSR). Clearly highlight the route discovery and route maintenance operations.
18. What is a wireless sensor network? How is it different from a MANET? Explain an application of a wireless sensor network.
19. Describe the security measures that can be incorporated in each layer of protocol stack of MANETs.
20. Write short notes on:
    (a) Characteristics of a secure ad hoc network
    (b) Security attacks in ad hoc networks.
21. Discuss the important security threats in a MANET. What are the factors responsible for limited security in MANETs?
22. Explain each of following types of attacks associated with ad hoc networks:

    (a) Routing loop (b) Black hole (c) Grey hole (d) Partitioning (e) Blacklist (f) Wormhole (g) Dropping routing traffic
23. Give an example of a core-based multicast routing protocol for MANETs. Explain its working.
24. Mention the important differences between a mobile ad hoc network and a cell phone network.
25. Explain any one routing technique that can be used in a mobile ad hoc network.
26. What are the different categories of routing protocols for mobile ad hoc networks?
27. Explain the factors that make mobile ad hoc networks more vulnerable to security attacks compared to the traditional networks. Also, explain the major types of security attacks that are possible in a mobile ad hoc network. How can each of these types of attacks be overcome?
28. What is the difference between active and passive security attacks in a MANET? Give an example of each of these types of attacks.
29. What do you understand by active and passive attacks on a MANET? Name at least five different types of security attacks in a MANET. Classify these into active and passive attacks. Discuss the security measures against these attacks.
30. For every layer of MANET protocol stack, discuss at least one type of security attack that exploits a vulnerability at that layer.
31. A major task of the designer of a wireless sensor network is prolonging the life of the network. Explain how this is achieved while designing a MANET.
32. Explain why traditional IP-based routing protocols may not perform satisfactorily in mobile ad hoc networks.

33. What do you mean by size and node density of a MANET? Explain these two terms and discuss how these two parameters impact the design of a MANET.
34. What is a VANET? Explain how does it differ from a traditional MANET? Explain any one application of a VANET.
35. Explain some routing attacks that are possible at the application layer of a MANET.



8

Wireless Sensor Networks

A wireless sensor network (WSN) is a self-organizing network of tiny sensor nodes. Each sensor node (also called mote) usually senses certain physical characteristics of its environment such as temperature, sound, vibration, precipitation, etc. and then transmits the sensed information to the user of the network. A WSN is called self-organizing since it gets setup and configured automatically without requiring any form of manual intervention.

A schematic diagram depicting the working of a WSN is shown in Fig. 8.1. As can be observed from the figure, each node of a WSN senses some specific data pertaining to the environment according to the user's specific requirements and the aggregate data is transmitted to the user of the network. For transmission of the required data, the nodes set up an ad hoc network among themselves.

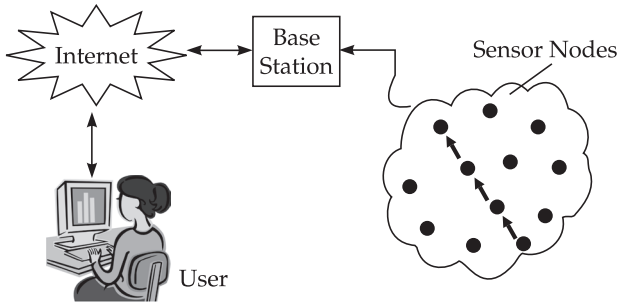


Figure 8.1 *A schematic model of a wireless sensor network*.

If one analyses the emergence of WSNs, it can easily be observed that the two most important contributory factors have been the recent progress in wireless communication technology and digital electronics. While advancements in wireless communications have made it possible for tiny

sensor nodes to transmit data at low-power levels, advancements in digital electronics have led to the development of sensor nodes that are small, powerful and yet inexpensive. For example, a reasonably sophisticated sensor node might cost less than rupees one hundred. The low cost of the sensor nodes makes it possible to deploy them in large numbers in a WSN.

BOX 8.1 The mote

The name mote was given by a few scientists from the university of California at Berkeley. These scientists were working on a project, called *smart dust*, which was funded by the Defense Advanced Research Projects Agency's (DARPA). The aim of the Smart dust project was to shrink the size of the sensor nodes (motes) to the size of dust. At this size of the motes, they can have very special uses. For example, they can be ingested to help diagnose diseases. Since they can float in air and liquids due to their small size, they can be used to monitor inaccessible places such as estimating the size of an underground oil well or the pollution level over a city.

WSNs are multi-hop wireless networks, similar to ad hoc networks. However, these networks have sensor devices as nodes in contrast to hand-held computers in ad hoc networks. Also, several thousands of nodes are deployed in a WSN of size of a few hundred square metres, compared to a few tens of nodes in the case of a comparable ad hoc network.

In the following section, we discuss the main differences between WSNs and mobile ad hoc networks (MANETs).

8.1 WSN vs. MANET

Though there are significant differences between the two, WSNs can be considered to be a special type of MANET. This is because WSNs are ad hoc networks after all. Also, the network topology of both MANET and WSNs changes dynamically. MANET topology changes due to user movements, whereas the topology of a WSN changes continually as nodes keep failing and node batteries keep getting depleted. In spite of these superficial similarities, there are significant differences between these two types of networks as enumered below.

1. *Size and sophistication of the sensor nodes:* As already discussed in Chapter 7, a MANET is usually made up of hand-held devices such as laptops, palmtops, personal digital assistants (PDA) and other such devices equipped with wireless transceivers. Compared to a hand-held device, a sensor node is much more limited in power, computational capacity, and memory. Therefore, sensor nodes cannot perform complex computations such as encryption and run

complex routing algorithms, which a hand-held device can. The sensor nodes usually restrict themselves to sensing, aggregating, and transmitting, and carrying out only very simple network routing operations.

2. *Node density:* The number of sensor nodes in a WSN is usually several orders of magnitude higher than the nodes in an ad hoc network of comparable size. The number of nodes in a WSN is typically several thousands over an area of a few hundred square metres compared to several tens of nodes in a MANET over a comparable area. Thus, sensor nodes are much more densely deployed in a WSN.

3. *Node reliability:* Once the sensor nodes get deployed, it is hard to access them for repair or replacement. Therefore, as the battery power of a node gets depleted, it quickly becomes incapacitated. The issue gets more complicated since these nodes are deployed in the open and often under harsh weather conditions, thus increasing their failure rates. Hence, sensor nodes are more prone to failures than the hand-held devices. On the other hand, nodes in a MANET do not fail as frequently since they have larger batteries and also batteries are easily recharged. Of course, at present some types of sensor nodes are able to recharge themselves from solar energy, from vibrations, or from any existing temperature differentials.

4. *Broadcast versus point-to-point communication:* Sensor nodes mainly use broadcast communication, whereas most ad hoc networks are based on some form of point-to-point communication.

5. *Node identity:* Sensor nodes do not usually have global identification (ID) because of the large number of sensors that are deployed. Assigning and maintaining the global identifies in a dynamic network would consume too much of computation and communicational resources. Thus, it is very difficult for a user of WSN to address or command sensor nodes either individually or in small groups. MANET nodes, in contrast, do have IDs by which they can be addressed and communicated to.

6. *Node movement:* Nodes in a WSN are usually static, whereas nodes in a MANET move as desired by the user.

7. *Transmission range:* Sensor nodes have tiny batteries and therefore can only afford to transmit at very small power levels. Also, they are equipped with much simpler wireless transceiver equipment compared to MANET nodes. As a result, the distance over which a sensor node can directly communicate is many times smaller than the distance over which a MANET node can communicate.

8. *Bandwidth:* The bandwidth available to a node in a WSN is typically of the order of 10 kbps, whereas the nodes in a MANET typically

transmit at much higher rates. This is because the sensor nodes do not have a DSP processor and as a result cannot perform digital signal modulation.

9. *Communication protocol:* The communication protocols used in MANETs incur too much computing and transmission overheads to be meaningfully deployed in a WSN. WSNs deploy much simpler protocols.
10. *Information flow:* The information flow in a WSN is typically well-structured and the flow is in one direction only. This is so because the nodes essentially collect information from the environment and transmit it to a server. This characteristic has been exploited to develop simpler and efficient communication protocols for WSNs. On the other hand, in a mobile ad hoc network, the flow of information is usually chaotic.

8.2 Applications

WSNs were conceived about two decades ago essentially for military applications such as battlefield surveillance and enemy tracking. But, nowadays WSNs have become popular in several civilian areas such as habitat monitoring (Cerpa et al., 2001 and Mainwaring et al., 2002), environment observation and forecasting (Biagioni and Bridges, 2002), health applications (Akyildiz et al., 2002), and home and office applications (Srivastava et al., 2001). The following are a few important applications of WSNs:

Military applications

WSNs have been used to provide the following services in battlefields:

Monitoring enemy troop formations: Wireless sensor networks can be used by a military commander to gather data about enemy formations. The gathered data could include troop strength, location of troops, amount and type of war equipment, etc. Such collected information can be used by the commander to decide about counter deployments.

Battlefield surveillance: To monitor enemy intrusions or to detect the presence of enemy troops, sensors can be randomly deployed in inaccessible and critical regions. This way, intrusions and army advancements can be easily determined. Based on the surveillance data, safe routes or paths for the offensive attacks can be worked out.

Great Duck Island (GDI) system

This is an example of WSN deployment in a habitat monitoring application. The researchers from Intel Research laboratory deployed a mote-based

sensor network in Great Duck Island, Maine, USA, to monitor the behaviour of storm petrel (small sea birds). The important features of their designed network were as follows:

Internet connectivity: The sensor network at Great Duck Island (GDI) was designed to be accessed via the Internet. It was also possible to monitor and manage the sensor networks deployed in the remote sites over the Internet.

Tree topology: The nodes were arranged in clusters and the clusters organized in a tree topology. To start with, three to four clusters of 100 fixed sensor nodes were deployed.

Life of the sensor network: The sensor network was designed to function for at least a couple of months without the need for recharging. However, GDI had sufficient solar power to recharge the sensor nodes that were fitted with small solar panels.

PODS—A remote ecological sensor network

PODS was a research project in the University of Hawaii that built a wireless sensor network to investigate as to why endangered species of plants grow in one area but not in neighbouring areas. They deployed secret sensor nodes, called Pods, in Hawaii Volcanoes National Park. Each pod consisted of a computer, radio transceiver and environmental sensors including a high resolution digital camera. These pods relayed sensor data to the user via a wireless link to the Internet.

Automated Local Evaluation in Real Time (ALERT)

ALERT is an automated flood alarm system, probably the first well-known deployment of a real wireless sensor network. ALERT provides important real-time rainfall and water level information to evaluate the possibility of potential flooding. ALERT sensor networks are usually equipped with meteorological sensors, such as water level sensors, temperature sensors, and wind sensors. Data is transmitted from the sensor site to the base station. A flood forecast model is deployed to process the received data and issue automatic warnings. Currently, ALERT is deployed across most of the states of western United States such as California and Arizona.

Smart Sensors and Integrated Microsystems (SSIM)

SSIM is a biomedical project about the artificial retina. In this, a retina prosthesis chip (prosthesis is an artificial replacement of a missing body part) consisting of 100 micro sensors is built and implanted within the human eye. This allows a patient with no vision or limited vision to see at an acceptable level. Feedback control, image identification and validation are achieved using wireless communication with a processor.

8.3 Architecture of the Sensor Node

A typical architecture of a sensor node is shown in Fig. 8.2. As can be seen, the sensor nodes pass information to the controller through an ADC (analog-to-digital signal converter). The controller processes this information and often filters it before transmitting it with the help of the transmitter.

In the following, we briefly explain the different components of the sensor node architecture.

Controller

The controller is possibly the most important component of the sensor node. Besides responding to certain user commands, it processes the sensor information, transmits it on the network, and also performs several network-related operations. A typical controller consists of an 8-bit 4-MHz microcontroller with slow 10-Kbps communication capability and 8-Kbyte read-only memory and a 512-byte RAM. While a controller is commonly designed using a microcontroller, SoC (System on Chip) and ASIC (Application Specific Integrated Circuit) are the other alternatives.

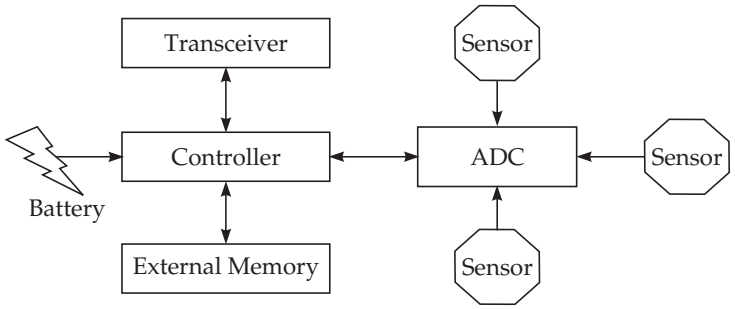


Figure 8.2 *Architecture of a sensor node.*

Transceiver

In a transceiver, the functionality of both the transmitter and receiver are combined into a single device. Most WSNs are based on radio frequency communications among the sensor nodes. WSNs typically use the licence-free communication frequencies such as the 915-MHz band or the 2.4-GHz band. Even when idle, most sensors consume power at a rate that is almost equal to the power consumed while receiving signals. This is because citcuits such as the amplifier and filter continue to operate as usual. Consequently, when a sensor node is not transmitting or receiving, it is better to completely shut down the transceiver rather than leave it in the idle mode.

External memory

From an energy perspective, the most appropriate memory is the on-chip memory of a microcontroller or SoC. For storage of large amounts of data, flash memories are used due to their low cost and high storage capacity. Magnetic memories are rarely used due to their high energy consumption, volume, and weight.

Power source

A sensor node consumes power for sensing, communicating and data processing. Typically, more energy is required for data communication than for any other process. Current sensors are able to renew their energy from solar radiation or from vibrations. Two power saving policies are being used. These are the Dynamic Power Management (DPM) policy and the Dynamic Voltage Scaling (DVS) policy. DPM helps conserve power by shutting down those parts of the sensor node that are idle. The DVS scheme is based on varying the power levels based on the work load.

Sensors

Sensors are designed to exploit certain physical principles to produce a measurable response to a change in some physical parameter such as temperature or pressure. A sensor should be small in size and consume negligible energy. Sensors fall into different categories based on whether they are passive or active or omni-directional or narrow-beam sensors. Passive sensors sense data without disturbing the environment in any way and consequently consume almost no energy. Active sensors probe by perturbing the environment in some way. Examples of active sensors are sonar, ultraviolet devices, and radar. Naturally, these sensors need battery power for their operation. Narrow-beam sensors sense in a well-defined direction such as a camera. Omni-directional sensors cannot be ascribed any direction to their measurements.

8.4 Challenges in the Design of an Effective WSN

Effective design of WSNs must take into account various constraints and limitations that these networks have to work with. Besides the limited processing capability and available memory, these networks work with limited bandwidth and limited network coverage. Further, new application scenarios lead to new types of challenges. The following subsections identify areas that constitute some of the important challenges in the design of an effective WSN.

Energy: Sensor nodes are powered by batteries and it is usually difficult to replace them when depleted since the sensor nodes are usually deployed in

remote and hostile environments. So the hardware, the network protocols and the applications must be designed to minimize energy consumption so as to increase network life.

Location discovery: In many applications, routing protocols need the geographic location of each sensor node to forward data among the networks. Location discovery protocols must be designed in such a way that location discovery can be done accurately and with minimum overhead. Use of WSNs becomes too expensive for applications.

Cost: Cost is another important factor that needs to be optimized during the design of WSNs. If the initial cost is high, popularization of WSNs would be difficult.

Security: An important application of sensor networks is the deployment in mission-critical tasks such as military surveillance, security, etc. However, the secure operation of sensor networks is even more difficult compared to MANETs due to the severe resource limitations of sensor nodes. Security solutions are constrained when applying them to sensor networks. For example, the encryption of transmitted data incurs significant processing overhead and power consumption. Further, sensor nodes are more susceptible to tampering and capture since they function totally unattended. Some of the important issues that need to be addressed in a security context are: secure routing, authentication, key establishment and trust setup, and prevention of physical attacks against sensor nodes.

8.5 Characteristics of Sensor Networks

The following subsections identify some important characteristics of sensor networks.

Multi-hop wireless communication: Due to low energy and simple antennae, the wireless coverage of sensor nodes is extremely small. Therefore, multi-hop communication is the norm in these networks.

Energy efficient operation: A key objective of WSNs is to maximize the life of the network.

Self-organization: This is an important characteristic of WSNs and helps it to cope with node failures. Failures may occur due to exhaustion of battery power and several other reasons such as hardware or software failure. The appearance of obstacles in communication paths and the addition of nodes to the network would also require self configuration.

Collaboration and in-network processing: Depending on the application, it is often required that a group of sensor nodes interact among themselves to detect an event by meaningfully processing their collective information.

Data redundancy: In a sensor network, it is not possible to assign a global identifier to each node due to the sheer number of nodes deployed. Lack of such global identifiers along with random deployment of sensor nodes, makes it hard to select a specific set of sensor nodes to be queried. Therefore, it is very difficult to give a command for obtaining data from the sensor nodes that are spread over a specific area or are part of the network. For this reason, data is usually transmitted from every sensor node within the deployment region with significant redundancy.

8.6 WSN Routing Protocols

In a wireless sensor network, nodes do not have any global identification associated with them. Further, nodes fail frequently due to a variety of reasons leading to dynamic changes to the network topology. As nodes fail, some network links get destroyed while some new links may get established when a new sensor node is introduced. It is easy to see that such an environment cannot be served efficiently by routing protocols developed for wired networks. It also needs to be realized that the network protocols for MANET are not suitable due to their limited computing power, lack of a DSP processor, limited battery energy, and the simplicity of antenna of the sensor nodes. In the following subsection, we classify the large number of WSN routing protocols that have been suggested based on the protocol used and the network structure.

8.6.1 Classification Based on Protocol Operation

In WSNs, there are essentially four main types of protocols that have been proposed. These protocols are briefly discussed below.

Multipath-based routing: In a network such as WSN, where the topology changes dynamically, established paths dissolve sometimes even before the full message transmission is complete. Since a path setup does take time and is expensive, it is necessary to set up redundant paths and maintain them as backups. Multipath-based routing protocols have at least one alternative path (from source to destination). However, this approach increases energy consumption and traffic. The redundant paths are kept alive by sending periodic messages. An example of such a protocol is Directed Diffusion (DD) paradigm (Intanagonwiwat et al., 2000 and Estrin et al., 1999).

Query-based routing: In these protocols, the destination nodes query data through the network. The node(s) containing the required data transmit it back to the node that has initiated the query along that path. An example of a protocol deploying this technique is Rumor routing (Braginsky and Estrin, 2002).

Negotiation-based routing: This class of routers use negotiation as the basis for cooperation between the competing entities. Inter domain routing is often driven by self-interest and is based on a limited view of the internet, which may be detrimental to the stability and efficiency of routing. An example of this class of routers is Sensor Protocols for Information via Negotiation (SPIN) (Heinzelman et al., 1999).

QoS-based routing: In some applications, data is required to be delivered within a certain period of time from the moment it is sensed, otherwise it would become stale and be of little use. Therefore, bounded latency for data delivery is a condition for time-constrained applications. In WSNs, however, the conservation of energy is often considered relatively more important than the quality of data sent. As the energy gets depleted, the network may be required to reduce the quality of the results in order to reduce the energy dissipation in the nodes and hence lengthen the total network lifetime. An example of a QoS-based routing protocol is Sequential Assignment Routing (SAR) (Akyildiz et al., 2002 and Sohrabi et al., 2000).

8.6.2 Classification Based on Network Structure

In general, routing techniques deployed in a WSN can be classified, based on the network structure, into either flat routing, hierarchical routing, or location-based routing.

Flat routing: In this type of routing, each node performs routing. An example of flat routing is the gradient-based routing technique (Schurgers and Srivastava, 2001).

Hierarchical routing: These protocols are also known as cluster-based routing. In these protocols, different nodes can play different roles in the network based on their remaining battery powers. Nodes having higher battery energy are used to process and send the information, while low-energy nodes only focus on sensing activity. Examples of such protocols include Low Energy Adaptive Clustering Hierarchy (LEACH) (Heinzelman et al., 2000), Power-Efficient Gathering in Sensor Information Systems (PEGASIS) (Lindsey and Raghavendra, 2002 and Lindsey et al., 2001).

Location-based routing: In these protocols, the nodes are addressed by their location. Distances to neighbouring nodes can be estimated by the received signal strengths (RSS) or by GPS receivers. In this type of protocol, the positions of sensor nodes are exploited to route data to the network. The distance between the neighbouring nodes can be estimated on the basis of the incoming signal strength. The relative coordinates of neighbouring nodes can be obtained by exchanging location information between neighbours. If there is no activity, nodes go to sleep to save energy. To maximize energy saving, it is necessary to have as many sleeping nodes

in the network as possible. Examples of such protocols include Geographic Adaptive Fidelity (GAF) (Xu et al., 2001) and Geographic and Energy Aware Routing (GEAR) (Yu et al., 2001).

In the following subsections, we briefly review a few important protocols belonging to the above three categories.

8.6.3 Directed Diffusion (DD)

In this protocol (Intanagonwiwat et al., 2000 and Estrin et al., 1999), each sensor node tags its data with one or more attributes. A node interested in some data floods the network with an interest packet. Interests are flooded over the network. When a node receives an interest packet from a neighbour, it responds by sending data to the neighbour to retrace the path of the interest packet. Each node only knows the neighbour from whom it got the interest. It is possible that each node would receive the same interest from more than one neighbour. In this case, multiple paths can be set up from the source node to the destination node. Among these paths, one or a few high rate paths are defined and the other paths remain low rate. Depending on the number of paths that are reinforced, directed diffusion can be single-path or multi-path routing. If a better path emerges, the protocol updates the paths. Further, directed diffusion allows intermediate nodes to cache and aggregate data to improve data accessibility and energy efficiency.

8.6.4 Rumor Routing

In rumor routing (Braginsky and Estrin, 2002), each node maintains a neighbour list and an event table. When a node witnesses an event, it adds the event to its event table. Nodes that have recently observed an event, generate an agent with a certain probability. An agent is a long-lived packet that roams through the network, propagating information. Each agent carries a list of events it has encountered, along with the number of hops to that event. When it arrives at a node, it synchronizes its list with the node's list. The agent traverses the network for some number of hops, and then dies. Any node can generate a query that is destined to a particular event. If it has a route to the event, then it will transmit the query. If it does not, it will forward the query to a random neighbour. If the query has not reached the destination node when the query Time-to-Live (TTL) expires, then it will retransmit the query, or flood the query. The goal of rumor routing is to avoid expensive flooding operations. Unlike directed diffusion (DD), which tries to find an optimal path by flooding queries for gradient setup, rumor routing sends a query on a random walk until it finds the event path.

8.6.5 Sequential Assignment Routing (SAR)

Sequential Assignment Routing (SAR) (Akyildiz et al., 2002 and Sohrabi et al., 2000) uses the QoS information in its routing decisions. This protocol creates trees rooted at one-hop neighbours of the sink by taking the QoS metric, the available energy on each path and the priority level of each packet into consideration. By using created trees, multiple paths from sink to sensors are formed. One of these paths is selected according to the available energy resources and QoS on the path.

8.6.6 Low Energy Adaptive Clustering Hierarchy (LEACH)

Low Energy Adaptive Clustering Hierarchy (LEACH) (Heinzelman et al., 2000) was proposed by Heinzelman. It is a very popular routing protocol for WSNs that is based on setting up a hierarchical structure. This protocol divides the network into clusters (sections). In each cluster, an elected sensor node acts as the cluster head. The task of a cluster head is to manage communication among other cluster heads. A cluster head can process the collected data before relaying to the other cluster heads. The cluster heads collaborate to send data to the base station.

8.6.7 Power-Efficient Gathering in Sensor Information Systems (PEGASIS)

In PEGASIS (Lindsey and Raghavendra, 2002 and Lindsey et al., 2001), each node communicates only with the immediate neighbour. When a node on the chain receives data from a neighbour, it aggregates the data with its own data and sends the data to the next neighbour on the chain. Rather than multiple cluster-heads sending data to the base station as in LEACH, only one node on the chain is selected to transmit the data to the base station.

8.6.8 Geographic and Energy Aware Routing (GEAR)

This protocol is based on the knowledge of a node's own as well as its neighbours' locations and energy information (Estrin and Govindan, 2001). It conserves energy by minimizing the number of interests in directed diffusion (DD) by only considering certain regions rather than sending the interests to the whole network.

8.6.9 Geographic Adaptive Fidelity (GAF)

The Geographic Adaptive Fidelity (GAF) (Xu et al., 2001) is an energy-aware location-based routing algorithm. Though this class of protocols is

designed primarily for mobile ad hoc networks, it has been used for sensor networks as well. The network area is first divided into fixed zones forming a virtual grid. Inside each zone, nodes collaborate with each other to play different roles. For example, nodes elect one sensor node to stay awake for a certain period of time while the others go to sleep mode.

8.7 Target Coverage

The coverage problem is one of the active issues of the WSNs that determines how efficiently the sensor network is being covered by a set of sensor nodes. This problem deals with the QoS of the network ensuring that the particular sensor network is monitored or observed by at least one sensor node. It may be broadly classified into the following categories:

- Area coverage where the sensor nodes are deployed.
- Target coverage where the sensor nodes are deployed to cover a specific set of targets or points.

In the target coverage problem, a fixed number of targets (t_1, t_2, …, t_m) need to be continuously monitored by a number of sensor nodes (s_1, s_2, …, s_n), while the lifetime of the network needs to be maximized. There are a specific number of targets which are required to be covered by a set of sensor nodes. Each target needs to be monitored by at least one sensor node. Since sensor nodes are provided with only limited resources and cannot withstand extreme environmental conditions, they are deployed in large numbers for exceeding the actual requirements.

Some of the possible strategies to cover a specific set of targets are described below.

8.7.1 Some Strategies for Target Coverage

Activating all the sensors at a time

In order to cover all the specific targets, the simplest method is to activate all the sensor nodes deployed for coverage at the same time. However, this simultaneous activation of sensor nodes may make them exhaust their energy together. So, the lifetime of the network is the least in this scheme.

Formation of disjoint set covers of sensor nodes

An energy efficient method to cover targets is to make the sensor nodes alternate between active and sleep modes (Cardei and Du, 2005). Of the different sensors covering a target, some go to sleep whereas the remaining stay active. This way there is no disruption of the monitoring process, whereas different nodes get sleep at different times. During the active

mode, a sensor node is capable of keeping track of a target and collect important information pertaining to it. Disjoint set covers of active sensor nodes are made in such a manner that each set cover covers all the targets. These set covers are activated one after another till the sensor nodes are out of their energy. So, the main idea behind this protocol is that rather than making all the sensor nodes active at a time, these can be partitioned into a number of disjoint set covers with each set cover capable of monitoring the targets.

Formation of non-disjoint set covers of sensor nodes

Another energy efficient method for target coverage (Cardei et al., 2005) is to make the sensor nodes part of more than one set cover, with the sensor nodes alternating between the active and sleep modes. Several non-disjoint set covers of active sensor nodes are made to activate successively where each set cover is capable of keeping track of all the specific targets until the time of energy exhaustion of sensor nodes. It results in much more energy efficient monitoring compared to the disjoint set cover method of target coverage, thus maximizing the network lifetime to a greater extent.

Partial target coverage

Till now, we have mentioned full coverage of targets which ensures that all the required targets are to be covered. Information collected by a subset of targets may also be beneficial (Zorbas et al., 2009). A number of set covers of sensor nodes may be activated for partial coverage of the targets, not necessarily covering all the targets. The gathered information about a subset of targets by the sensor nodes may prove to be very important where the main objective is to maximize the lifetime of the network.

K-coverage

In the k-coverage ($k \geq 1$) approach (Zorbas et al., 2009), a minimal subset of the sensor nodes is determined that can keep track of the required targets, with each target being covered by at least k sensor nodes. The activated nodes collectively gather information about the specific targets and send it to the base station.

SUMMARY

In this chapter, we discussed sensor networks and the challenges in their design and operation. We also reviewed the routing protocols used in WSNs. A wireless sensor network consists of a large number of tiny autonomous, lightweight and low cost devices called sensor nodes. A sensor node has characteristics such as limited sensing and limited computations

and hence can communicate only within short distances. The sensor nodes are constrained in terms of energy, memory, computational capability and communication bandwidth.

These sensors monitor different conditions at different locations, such as temperature, humidity, lighting condition, pressure, soil makeup, noise levels and help to find some current characteristics of certain objects such as the presence or absence, mechanical stress, speed, direction, size, etc.

WSNs are practically becoming important. Various areas of applications of sensor networks include critical missions, civil, healthcare, and environmental disciplines, etc. The low cost of these nodes helps in their deployment ranging from thousands to million nodes. These nodes can be deployed either in a random fashion or in a pre-designed fashion.

FURTHER READINGS

Akyildiz, I.F., et al., "Wireless sensor networks: a survey", *Computer Networks*, Vol. 38, pp. 393–422, March 2002.

Akyildiz, I.F., W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks", *IEEE Communications Magazine*, 40(8), pp. 102–114, August 2002.

Biagioni, Edoardo and Kent Bridges, "The application of remote sensor technology to assist the recovery of rare and endangered species". In: *Special issue on Distributed Sensor Networks for the International Journal of High Performance Computing Applications*, Vol. 16, No. 3, August 2002.

Braginsky, D. and D. Estrin, "Rumor Routing Algorithm for Sensor Networks". In: *Proceedings of the First Workshop on Sensor Networks and Applications (WSNA),* Atlanta, GA, October 2002.

Cardei, M. and D.Z. Du, "Improving Wireless Sensor Network Lifetime through Power Aware Organization", *ACM Wireless Networks*, Vol. 11, No. 3, pp. 333–340, May 2005.

Cardei, M., M.T. Thai, Y. Li, and W. Wu, "Energy-efficient target coverage in wireless sensor networks," *Proceedings of IEEE INFOCOM'05*, Miami, FL, pp. 1976–1983, 2005.

Cerpa, A., J. Elson, D. Estrin, L. Girod, M. Hamilton, and J. Zhao, "Habitat monitoring: Application driver for wireless communications technology". In: *Proceedings of the 2001ACM SIGCOMM Workshop on Data Communications* in Latin America and the Caribbean, April 2001, 2001.

Estrin, D., et al., "Next Century Challenges: Scalable Coordination in Sensor Networks". In: *Proceedings of the 5th annual ACM/IEEE international conference on Mobile Computing and Networking (MobiCom'99)*, Seattle, WA, August 1999.

Heinzelman, W., A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless sensor networks". In: *Proceedings of the Hawaii International Conference System Sciences,* Hawaii, January 2000.

Heinzelman, W., J. Kulik, and H. Balakrishnan, "Adaptive protocols for information dissemination in wireless sensor networks". In: *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99),* Seattle, WA, August 1999.

Intanagonwiwat, C., R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks". In: *Proceedings of the 6th Annual ACM/IEEE. International Conference on Mobile Computing and Networking (MobiCom'00),* Boston, MA, August 2000.

Lindsey, S. and C.S. Raghavendra, "PEGASIS: Power Efficient Gathering in Sensor Information Systems". In: *Proceedings of the IEEE Aerospace Conference,* Big Sky, Montana, March 2002.

Lindsey, S., C.S. Raghavendra and K. Sivalingam, "Data Gathering in Sensor Networks using the Energy*Delay Metric". In: *Proceedings of the IPDPS Workshop on Issues in Wireless Networks and Mobile Computing,* San Francisco, CA, April 2001.

Mainwaring, Alan, Joseph Polastre, Robert Szewczyk, David Culler, and John Anderson, "Wireless sensor networks for habitat monitoring". In: *ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02),* Atlanta, GA, September 2002.

Schurgers, C. and M.B. Srivastava, "Energy efficient routing in wireless sensor networks". In: *MILCOM Proceedings on Communications for Network-Centric Operations: Creating the Information Force,* McLean, VA, 2001.

Schwiebert, Loren, K. Sandeep, S. Gupta, and Jennifer Weinmann, "Research challenges in wireless networks of biomedical sensors". In: *Mobile Computing and Networking*, pp. 151–165, 2001.

Sohrabi, K., et al., "Protocols for self-organization of a wireless sensor network," *IEEE Personal Communications*, Vol. 7, No. 5, pp. 16–27, October 2000.

Srivastava B., Mani, Richard R. Muntz, and Miodrag Potkonjak, "Smart kindergarten: sensor-based wireless networks for smart developmental problem-solving environments". In: *Mobile Computing and Networking*, pp. 132–138, 2001.

Xu, Y., J. Heidemann, and D. Estrin, "Geography-informed energy conservation for ad hoc routing". In: *Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'01)*, Rome, Italy, July 2001.

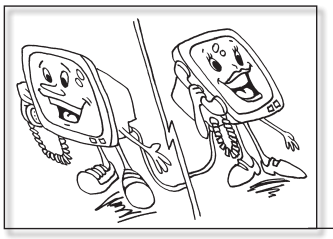
Yu, Y., D. Estrin, and R. Govindan, "Geographical and Energy-Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks," *UCLA Computer Science Department Technical Report, UCLA-CSD TR-01-0023*, May 2001.

Zhou, Zongheng, Samir R. Das, and Himanshu Gupta, "Connected k-coverage problem in sensor networks", *Proceedings of the International Conference On Computer Communications and Networks (ICCCN 2004)*, October 11–13, 2004, Chicago, IL, USA 2004.

Zorbas, D., D. Glynos, and C. Douligeris, "Connected partial target coverage and network lifetime in wireless sensor networks", *Wireless Days (WD)*, 2009 2nd IFIP, pp. 1–5.

EXERCISES

1. Compare WSNs with MANETs with respect to their characteristics, applications and challenges to their commercialization?
2. Discuss a classification of the routing protocols for sensor networks.
3. Write short notes on each of the following:
   (a) Multipath-based Routing (b) QoS-based Routing
   (c) Hierarchical Routing (d) Location-based Routing
4. What is flat routing? How is it different from hierarchical routing?
5. Explain hierarchical routing and query-based routing in a wireless sensor network.
6. What is a wireless sensor network (WSN)? Discuss some important applications of WSNs.
7. Discuss the architecture of a sensor node in a wireless sensor network.
8. Explain the similarities and differences between a wireless sensor network and mobile ad hoc network.
9. Explain the constraints and limitations on the operation and design of WSNs.
10. Explain the main differences between the routing protocols for WSNs and MANETs. Compare and contrast them with the routing protocols for traditional wired networks. Discuss the main reasons behind the differences in the routing approaches.
11. Why is design of a WSN considered to be more challenging than that of a MANET? Explain your answer.
12. Explain the alternative WSN designs that can be proposed for monitoring a finite number of discrete targets.
13. Do you agree with the statement: "Wireless sensor networks are a special type of mobile ad hoc networks." Justify your answer.



9

Operating Systems for Mobile Computing

Of late, smartphones are being used not only to make phone calls, but also to carry out a host of other useful operations such as to make video conference calls, send multimedia messages, take pictures, play media files, browse World Wide Web (WWW), run remote applications, etc. This sophistication of the present-day smartphones enables multiple tasks to be run on the device. Consequently, a powerful operating system has become an essential part of every smartphone. However, we need to keep in mind that the capability of the operating system that is used in a mobile handset varies a great deal with the degree of sophistication of the device. In a mobile handset, the operating system (OS) performs two main responsibilities.

9.1 Operating System Responsibilites in Mobile Devices

9.1.1 Managing Resources

An important responsibility of the operating system of a mobile device is to facilitate efficient utilization of the resources of the device by performing multiple tasks. The resources that are managed by the operating system include processor, memory, files, and various types of attached devices such as camera, speaker, keyboard, and screen. Typically, a mobile device is expected to run multiple applications at the same time and each application may in turn require running multiple tasks. A task can have multiple threads. A few examples of such applications include voice communication, text messaging, e-mail, video play, music play, recording, web browsing, running remote applications, etc. As an example scenario of usage of a smartphone, consider the following: a person might be listening to music, at the same time he might answer an incoming call, and an SMS might

arrive at the same time which he might like to look-up while the call is still on. Such a scenario requires concurrent execution of multiple tasks. When multiple tasks contend to use the same set of resources, the OS acts like a traffic cop—ensuring that different tasks do not interfere with each other.

9.1.2 Providing Different Interfaces

The operating system of a mobile device on the one hand provides a highly interactive interface to the user of the device and on the other interfaces with other devices and networks. An important interface concerns control, data, and voice communications with the base station using different types of protocols. Besides, an OS takes care of recognizing inputs from the keyboard, sending outputs to the display screen, and interfacing with peripheral devices such as other mobile devices, computers, printers, etc.

For the sake of brevity, we shall refer to the operating system used in a mobile hand-held device as a mobile OS. At present, the mobile OS marketplace is dominated by Symbian, Android, Windows mobile, Palm OS, iOS, and Blackberry OS. In this chapter, after discussing a few basic concepts relevant to mobile OS and the special constraints under which a mobile OS is expected to operate, we will survey the evolution and important features provided by a few popular mobile operating systems. Finally, we will provide a brief overview of the operating systems designed for the sensor nodes in a WSN.

9.2 Mobile O/S—A Few Basic Concepts

Traditionally, the operating system is viewed as providing a set of services to the application programs. The operating system is usually structured into a kernel layer and a shell layer. The shell essentially provides facilities for user interaction with the kernel. The kernel executes in the supervisor mode and can run privileged instructions that could not be run in the user mode. During booting, the kernel gets loaded first and continues to remain in the main memory of the device. This implies that in a virtual memory system, paging does not apply to the kernel code and kernel data. For this reason, the kernel is called the *memory resident* part of an operating system. The shell programs are usually not memory resident. The kernel of the operating system is responsible for interrupt servicing and management of processes, memory, and files.

The traditional operating systems such as Unix and Windows are known to have a monolithic kernel design. In a monolithic kernel OS design, the kernel essentially constitutes the entire operating system code, except for the code for the shell. The principal motivation behind this monolithic design was the belief that in the supervisor mode, the operating system

services can run more securely and efficiently. On the other hand, the main problem with the monolithic kernel design is that it makes the kernel massive, non-modular, hard to tailor, maintain, extend, and configure.

Considering the disadvantages of the monolithic kernel design, the microkernel design approach has been proposed. The microkernel design approach tries to minimize the size of the kernel code. Only the basic hardware-dependent functionalities and a few critical functionalities are implemented in the kernel mode and all other functionalities are implemented in the user mode. Most of the operating system services run as user level processes. The main advantage of this approach is that it becomes easier to port, extend, and maintain the operating system code. The kernel code is very difficult to debug compared to application programs.

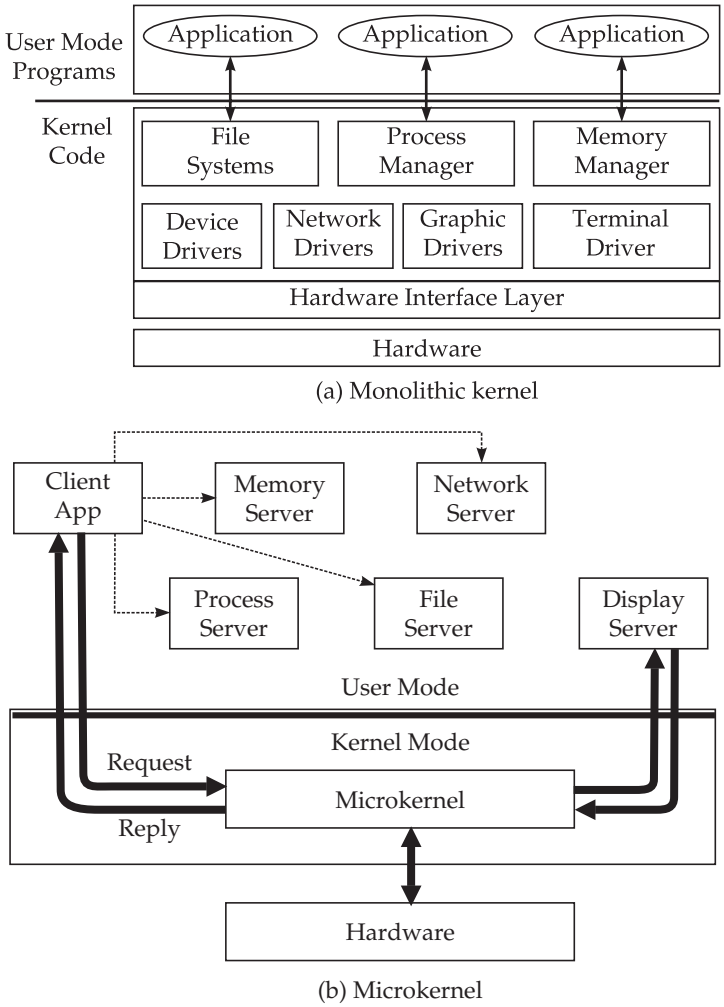


Figure 9.1 *Monolithic design versus microkernel design of an operating system.*

The reason for this is that a bug in a kernel code can crash the system, thus crashing the debugger too. Further, even when some operating system service crashes while being used by a user, it does not bring down the entire system. This is one reason as to why a microkernel operating system could be expected to be more reliable than an equivalent monolithic kernel operating system. The overall architectural difference between a monolithic kernel and a microkernel architecture is schematically shown in Figure 9.1. To restrict the size of the kernel of a mobile OS to the minimum, most mobile OS are, to different extents, based on the microkernel design.

9.3 Special Constraints and Requirements of Mobile O/S

There are a few special constraints under which the operating system of a mobile device needs to operate. There are also a few special features that are required to be supported by a mobile OS, but are not present in traditional operating systems.

9.3.1 Special Constraints

The operating system for a mobile device needs to function in the presence of many types of constraints which are not present in a traditional computer. As an example of such a constraint, consider the fact that a mobile device is powered by severely limited energy stored in a tiny battery. Therefore, an important constraint for a mobile device lies in avoiding complex computations and hence entering into a low power sleep mode as soon as possible. Such a constraint is not usually imposed on a traditional operating system. Further, the number of times that a mobile device is typically turned on per day is significantly higher than that of a desktop or any other computer. Due to this constraint, the mobile OS would need to be loaded (booted) much faster each time after it is switched on, compared to a desktop. Consequently, the kernel of a mobile OS needs to be of a very small size. There are several such constraints which influence the design of a mobile OS. It can be argued that coping with such constraints is an important reason why the operating system of a mobile device needs to differ significantly from a general-purpose operating system. We discuss below some of the important constraints of a mobile OS.

Limited memory

A mobile device usually has much less permanent and volatile storage compared to that of a contemporary desktop or laptop. To cope with the limited memory of a mobile device, the OS must be as small as possible and yet provide a rich set of functionalities to meet user demands. The size of the kernel is, therefore, considered to be a very important figure of merit of a mobile OS.

Limited screen size

The size of a mobile handset needs to be small to make it portable. This limits the size of the display screen. Consequently, new innovative user interfaces need to be supported by the mobile OS to overcome this constraint and minimize user inconveniences. For example, many handsets provide easy configurability of the interface to suit individual preferences, switching between menu and iconic interfaces, etc.

Miniature keyboard

Mobile handsets are either provided with a small keypad or the small-sized display screen is designed to be used as a keyboard in a touchscreen mode using a stylus. In both these arrangements, typing in the documents and entering the string commands is difficult. This mandates the provision of some facility for word completion prompts and availability of capabilities for free form handwriting recognition.

Limited processing power

A vast majority of modern mobile devices incorporate ARM-based processors. These processors are certainly energy efficient, powerful, and cheaper compared to the desktop or laptop processors, yet these are significantly slower. The sizes of the on-chip and off-chip memory are also restricted. The cope with the restricted processing power, storage, and battery power, usually the operating system is made to provide only a limited number of functionalities that are useful in the actual operation of the mobile. Activities such as mobile application development that require use of memory-intensive utility programs, such as editors and compilers, are carried out on a desktop or laptop, and only after the application is completely simulated and tested, it is cross-compiled and downloaded onto the mobile device.

Limited battery power

Mobile devices need to be as lightweight as possible to increase their portability. Due to the severe restrictions that are placed on their size and weight, a mobile device usually has a small battery and often recharging cannot be done as and when required. In spite of the small battery, a mobile phone is expected to support long talk time without the need to recharge frequently. Consequently, the operating system for a mobile device needs to be not only computationally efficient, but also at the same time expected to minimize power consumption. The techniques used by an OS to reduce power consumption include putting the processor and display screen into sleep mode within a few seconds of inactivity, and varying the intensity of transmitted antennae power as per requirement, etc.

Limited and fluctuating bandwidth of the wireless medium

The operating system of a mobile handset needs to run complex protocols due to the inherent problems caused by mobility and the wireless medium. A wireless medium is directly susceptible to atmospheric noise, and thereby causes high bit error rates. Further, the bandwidth of a wireless channel may fluctuate randomly due to atmospheric noise, movement of some objects, or the movements of the mobile handset itself. This can show up as short-term fades. There can be relatively longer-term disconnections due to handoffs. In this context, uninterrupted communication requires a special support for data caching, pre-fetching, and integration.

9.3.2 Special Service Requirements

Several facilities and services that are normally not expected to be supported by a traditional operating system, are mandated to be supported by a mobile OS. We identify a few important ones in the following.

Support for specific communication protocols

Mobile devices are often required to be connected to the base station and various types of peripheral devices, computers and other mobile devices. This requires enhanced communication support. The types of communication protocols used for communication with the base station depend on the generation of the communication technology (1G, 2G, etc.) in which the mobile device is deployed. Considering that a mobile should be usable across the existing technology spectrum, it becomes necessary to simultaneously support two or more generations of technology. For communication with other devices and with computers, TCP/IP and wireless LAN protocols also need to be supported. For web browsing as well as communication with other personal devices such as pen drive and headphones, though mobile devices are equipped with USB and other types of ports, mobility constraints often make infrared or Bluetooth connections preferable. This mandates the operating system to support multiple interfacing protocols and hardware interfaces.

Support for a variety of input mechanisms

A miniature keyboard forms the main user input mechanism for an inexpensive mobile device. Sophisticated mobile devices (smartphones) usually support the QUERTY keyboard. Many recent mobile devices also support touchscreen or even stylus-based input mechanisms along with the handwriting recognition capability. The different input mechanisms to be supported strongly influence the intended primary use of a device as well as the specific customer segment for which it is positioned. These issues dictate the choice and complexity of the user interaction part of the OS to a large extent and to a smaller extent the internal design of the operating

system. A mobile OS needs to support a variety of input mechanisms to make it generic and usable by different manufacturers of mobile devices.

Compliance with open standards

Adhering to an open standard facilitates the development of innovative applications by third-party developers. To facilitate the third party software development as well as to reduce the cost of development and time-to-market by the mobile handset manufacturers, the OS should adhere to open standards. Smartphones come in many different shapes and sizes and have varying screen sizes and user input capabilities. Therefore, the user interface and networking capabilities of a mobile OS need to be designed keeping these diversities in view.

Extensive library support

The cost-effective development of third party applications requires extensive library support by the OS. At the minimum, the expected library support includes the availability of programmer callable primitives for email, SMS, MMS, Bluetooth, multimedia, user interface primitives, and GSM/GPRS functionalities.

9.4 A Survey of Commercial Mobile Operating Systems

It is a challenging task to design a mobile OS with a set of core capabilities that are expected to be supported by mobile devices and with a consistent programming environment across all smartphones that install the OS. The mobile OS has to also facilitate third party development of application software and yet allow manufacturers of different brands of mobile devices to build their choice set of functionalities for the users. A few popular mobile OS are discussed below.

9.4.1 Windows Mobile

Before discussing the important features of the mobile operating systems available from Microsoft, we first give a historical perspective. Considering the proliferation of embedded devices, Microsoft Corporation developed an operating system in the year 1996 targeted specifically at these devices. Since then, this operating system has undergone several enhancements and modifications over successive generations. The main feature of Windows CE operating system, which sets it apart from other traditional operating systems, is the support that it provides for deterministic scheduling of time-constrained tasks. Based on the Microsoft's Windows CE operating system, the company designed the Pocket PC 2000 operating system in the year 2000. It was targeted for PDAs and not mobile phones.

Since the usage of mobile phones was increasing at a brisk rate, Microsoft introduced its *Windows Mobile* operating system in the year 2003. It was developed based on Pocket PC 2000 and was targeted specifically as an operating system for mobile phones which the different cell phone vendors can use in their cell phones. Unlike Apple'e iOS and RIM's Blackberry OS which are essentially confined to use on Apple's iPhone and the Blackberry, Microsoft intended the Windows mobile operating system to be used across a wide cross section of mobile phone manufacturers. Since the *Windows Mobile* operating system was developed to be suitable for use on multiple vendor platforms, Microsoft defined a hardware specification for hand-held computers that can run its *Windows Mobile* operating system in order to simplify the design of the operating system and to reduce the number of versions of the operating system. It was also intended to make the cell phones manufactured by different vendors appear uniform.

Microsoft later renamed its Pocket PC operating system to *Windows Mobile Classic*. Windows mobile classic operating systems support touchscreen-based user interface but do not support any phone capability. Consequently, Windows Mobile is now a family of three operating systems: *Windows Mobile Standard* and *Windows Mobile Professional* are targeted for use in smartphones, and *Windows Mobile Classic* is not targeted for cell phones, but for PDAs. The evolution of Windows Mobile operating system is schematically shown in Figure 9.2.

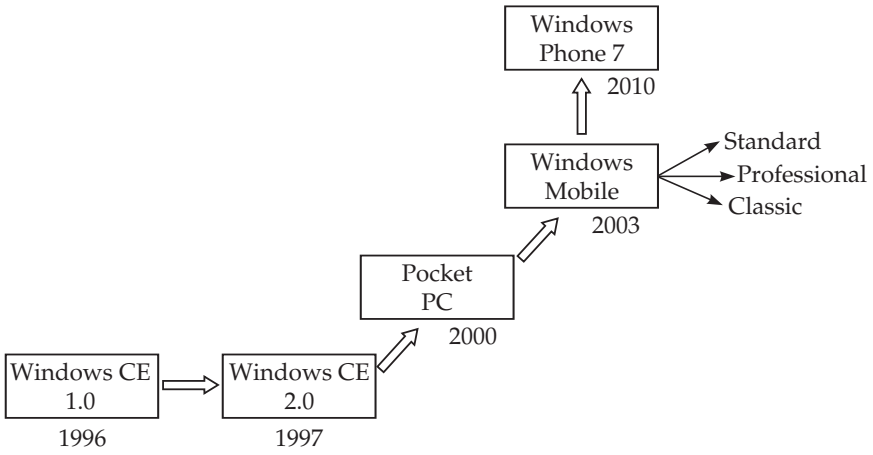


Figure 9.2 *Evolution of Microsoft's mobile OS.*

Windows mobile is designed to have a look and feel that is very similar to the desktop version of Windows. Microsoft obviously planned that since many users are familiar with the desktop version of Windows, they would find it easy to operate Windows mobile. Besides the core capabilities required by a mobile device, many third-party software applications are available for Windows mobile. These software applications can be

purchased via the Windows Marketplace for mobiles. The marketplace is a website maintained by Microsoft, where different application developers can submit their applications for download by the subscribers. Microsoft passes on 70% of the fee received to the developers hosting their applications. Windows mobile has recently been superseded by Windows Phone 7. A significant attempt to increase the market penetration of Windows mobile was the joint announcement in 2011 by Microsoft and Nokia of a partnership between their companies. They announced that Windows Phone 7 operating system would be used as the operating system for Nokia smartphones. In the press conference, they suggested that the mobile OS marketplace would henceforth see a three horse race and named Android and iOS, the main competitors of Windows Phone 7.

Windows Phone 7 is a significant improvement over the Windows mobile operating system. Possibly because of this, Microsoft has not maintained backward compatibility with the Windows mobile operating system, meaning that a mobile application that runs on the Windows mobile may not run on the Windows phone OS. Microsoft has defined the hardware specifications that a Windows Phone 7 device must meet. For example, it should support a screen resolution of 800 × 480 pixels. Windows Phone 7 devices need to have an accelerometer and a compass. Windows phone operating system provides a touchscreen interface with facilities for both command and text input. The operating system detects when a device has been rotated from portrait to landscape orientation.

A few important features of the Windows mobile OS are the following:

- The Graphics/Window/Event manager (GWE) component handles all input and output.
- Provides a virtual memory management.
- Supports security through the provision of a cryptographic library.
- Application development is similar to that in the Win32 environment. This is considered advantageous since many programmers have knowledge of Win 32-based application development.
- At present, it does not provide true multitasking. An application in the background goes into hibernation and gets active only when it comes to foreground. However, it is expected that Microsoft may support true multitasking in the future versions of the Windows Phone operating system.

9.4.2 Palm OS

Palm OS (also known as Garnet OS) is a proprietary operating system that was developed by Palm Computing in 1998 for its highly successful PDA called Palm Pilot. Palm OS was designed for ease of use with the provision of a touchscreen-based graphical user interface. Later, Palm OS was upgraded to facilitate installation in several different mobile devices, such as smartphones of different makes, wrist watches, hand-held gaming

consoles, bar code readers and GPS devices. About a decade ago, Palm OS was a very popular operating system, but has now lost its dominant market position (Lin and Ye, 2009).

The key features of the current Palm OS (named Garnet) are the following:

- It is essentially a simple single-tasking operating system. As a result, only one application can run at a time. The implications of this are many and easily noticeable. For example, if you are on voice communication, you cannot use the calculator, or read an SMS.
- It has an elementary memory management system. To keep the operating system small and fast, Palm OS does not isolate the memory areas of applications from each other. Consequently, any misbehaving application can crash the system.
- Palm supplies Palm emulator, which emulates the Palm hardware on a PC. This allows Palm programs to be developed and debugged on a PC before being run on the Palm hardware.
- It supports a handwriting recognition-based system for user input.
- It supports a facility called HotSync technology for data synchronization with desktop computers.
- It supports sound playback and recording capabilities.
- It incorporates a very simple and rudimentary security model in which a device can be locked by password.
- The different interfaces supported include Serial port/USB, infrared, Bluetooth and Wi-Fi connections.
- It uses a proprietary format to store calendar, address, task and note entries and yet are accessible by third-party applications.

9.4.3 Symbian OS

Symbian operating system was developed through a collaboration among a few prominent mobile device manufacturers including Nokia, Ericsson, Panasonic, and Samsung. Their objective was to develop a single industry standard operating system (Hall and Anderson, 2009). For many years, Symbian was the undisputed leader in the smartphone OS market. It was the operating system used in the handsets manufactured by Nokia, Ericsson, Panasonic, and Samsung. Though Symbian had a vision of a single collaborative effort towards the realization of a versatile mobile OS installable on a wide range of smartphones, in reality, the different participants tried to exert control over the direction of evolution of the operating system and consequently it became huge and unwieldy.

In 2008, Ericsson, Sony, Panasonic, and Samsung pulled out of the collaboration, selling their stake to Nokia. Around the same time, Google announced Android as an open operating system. After this, the market share of Symbian started coming down drastically. To counter this, the Symbian source code was published under Eclipse Public License (EPL)

in February 2010. This event was reported to be the largest codebase transition from proprietary to Open Source in the entire history. Due to the intensifying competition in the mobile handset market, Nokia announced in 2011 that it would move away from Symbian and use Windows Phone 7 OS in its smartphone handsets. This event is likely to cause further drastic erosion to the Symbian installation base.

Symbian OS is a real time, multitasking, pre-emptive, 32-bit operating system that runs on ARM-based processor designs. The inherent design of the Symbian operating system is microkernel-based. The CPU is switched into a low power mode, when the application is not responding to an event. Symbian comes in two major flavours. These flavours have been developed based on demands from various service providers.

(a) **Series 60:** The series-60 platform was until recently the leading smartphone platform in the world. The relatively large sized colour screen, easy-to-use interface and an extensive suite of applications make it well-suited to support advanced features such as rich content downloading and MMS (Multimedia Messaging Service). Series 60 was mainly being used on Nokia's smartphones and Samsung handsets.

(b) **UIQ interface:** UIQ (earlier known as User Interface Quartz) is a software package developed by UIQ Technology for Symbian OS. Essentially, this is a graphical user interface layer that provides capabilities for third-party application developers to develop applications and effortlessly create user interfaces.

A few other important features supported by the Symbian operating system are given below:

- It supports a number of communication and networking protocols including TCP, UDP, PPP, DNS, FTP, WAP, etc. For personal area networking, it supports Bluetooth, InfraRed and USB connectivity.
- It supports pre-emptive multitasking scheduling and memory protection. Symbian is a microkernel-based operating system.
- CPU is switched into a low-power mode when the application is not responding to an event.
- It is optimized for low-power and memory requirements. Applications, and the OS itself, follow an object-oriented design paradigm.
- All Symbian programming is event-based, and the CPU is switched into a low-power mode when the applications are not directly dealing with an event. This is achieved through a programming idiom called active objects.
- Carbide is an Integrated Development Environment (IDE) toolkit that is available for C++ application development on Symbian OS. It essentially works as an Eclipse plug-in and contains editor, compiler, emulator, libraries and header files required for Symbian OS development. Development kits are available at Nokia and the Symbian Foundation websites.

9.4.4 iOS

In January 2007, Apple unveiled its sleek innovative mobile device—the iPhone—causing a storm in the smartphone marketplace. The iPhone was designed to replace Apple's highly successful iPod. Apple had developed iOS as iPhone's operating system and was originally known as iPhone OS, but later renamed iOS. iOS is a derivative of Mac OS. Mac OS was later extended for use in other Apple devices such as the iPod touch, iPad and Apple TV.

iOS is a closed and proprietary operating system fully owned and controlled by Apple and not designed to be used by various mobile phone vendors on their systems. Apple does not license iOS for installation on third-party hardware. However, the overwhelming popularity of iPhone has given iOS a significant market presence. It provided several innovative features that grabbed the market attention. For example, user interactions with OS include gestures such as *swipe*, *tap*, *pinch*, and *reverse pinch*, all of which have specific definitions within the context of the iOS operating system. The other innovative user interactions that are supported by the iOS include internal accelerometers used by some applications for shaking the device as the undo command, rotating the device in three dimensions to switch the display mode from portrait to landscape, etc.

9.4.5 Android

To understand the genesis of Android, we need to dig into a bit of history. The leading search engine company Google's income is dependent on the average number of searches performed per day by people using its search engine. This is because advertisers pay to a web search company based on the average number of user hits that are recorded on its search engine. Prior to Android, Google had no means of directing mobile users to its search site. Even if a mobile owner wanted to search using Google's search engine, the prevalent operating systems of the mobile devices did not facilitate this. For example, Verizon wireless directed customers to its own search engine, as Verizon did not want to miss the extra revenue from this source. Google took a serious view of this situation, since market trend analysts were already predicting that searches from mobile handsets would out pace those from conventional searches. Google was eagerly looking for ways to address this problem. In 2005, Google acquired a small startup company called Android, which was developing an operating system for mobile devices based on Linux.

Google set up the *Open Handset Alliance* in 2007. It is a group of 82 technology and mobile communication companies that are collaborating to develop the Android operating system as an open source software for mobile devices. It facilitated any application developer to write Android applications. Android soon became feature-rich due to its open

source nature. Google could embed its search engine into Android, the way Internet Explorer is embedded into Windows. By giving away the operating system free and focusing on the search advertisement revenue, Google successfully formed a formidable challenge to Windows mobile and Symbian. A Gartner survey indicates that by the third quarter of 2011, Android held 52% share of the mobile OS market compared to that for Symbian's 17% and iOS's 15%.

Starting from 0% market share, in 2008, when the first mobile phone incorporating Android was announced, it has become the dominant player and has shown a remarkable rate of growth in market share and user acceptance. To understand the success of Android, it is important to understand the difficulties that users were experiencing with other operating systems and how Android has helped overcome those.

- Since mobile manufacturers use different user interfaces and interaction styles for their mobile devices depending on the customer segment targeted, Android provided the ability to seamlessly use either a phone-based keyboard or a touchscreen.
- Mobile users expect to browse real web pages, and not the simplified mobile versions of those pages. Many mobile handsets support browsing alternative sites provided by many website operators for mobile handsets with small screens and limited interfacing capabilities. However, Android scores over those operating systems by providing a built-in full web browser capable of rendering full web pages and not just small mobile versions.
- An important handicap of the competing operating systems is the difficulty of development of third-party applications. For example, an application on iOS has to be approved by Apple before it can be offered to the users as outlined in the SDK agreement. Apple does not, in fact, facilitate third party application development and is implicitly promoting a closed proprietary environment, where the internal working of the operating system is not exposed to the developers.
- A prominent advantage that Android holds out is that Android SDK works in Eclipse environment. Since many developers are already exposed to these standard technologies, there is a large pool of developers available for working on projects on the Android platform.
- It provides an RDBMS SQLite for data storage and data sharing across various applications.
- It has several innovative pre-installed applications such as Gmail, Maps, voice search, etc.

Android allows application developers to write code in the Java language. It facilitates the development of applications with the help of a set of core Java libraries developed by Google. Though most applications are developed using Java, applications and libraries can still be written in C and other

languages. The Android code is structured into four different layers as shown in Fig. 9.3. The important features of these layers are discussed below.

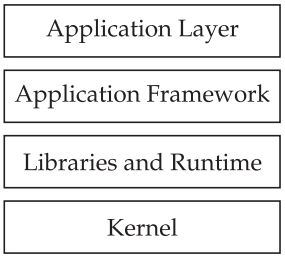


Figure 9.3 *Android software stack.*

Application layer

The Android operating system comes with a set of basic applications such as web browser, email client, SMS program, maps, calendar, and contacts repository management programs. All these applications are written using the Java programming language J2ME. Android applications do not have control over their own priorities. This design is intentional and is intended to help aggressively manage resources to ensure device responsiveness and even kill an application when needed.

Application framework

An application framework is used to implement a standard structure for different applications. The application framework essentially provides a set of services that an application programmer can make use of. The services include managers and content providers. Content providers enable applications to access data from other applications. A notification manager allows an application to display custom alerts on the status bar.

Libraries and runtime

The available libraries are written using multiple languages such as C and C++. These are called through a Java interface. These include a Surface Manager (for compositing windows), 2D and 3D graphics, Media Codecs like MPEG-4 and MP3, an SQL database SQLite and the web browser engine called WebKit.

The Android runtime consists of two components. A set of libraries provides most of the functionalities available in the core libraries of the Java language. The other runtime is the Dalvik virtual machine. Most applications that run on Android are written in Java. Dalvik translates a Java application program into machine code of the mobile device and executes it by invoking the operating system. These can be compiled to ARM native code and installed using the Android native development kit (SDK).

Dalvik VM is not a traditional JVM, but a custom VM designed to run multiple instances efficiently on a single device. Every Android application runs its own process with its own instance of the Dalvik virtual machine.

Kernel

Android kernel has been developed based on a version of Linux kernel. However, it has excluded the native X Window System and does not support the full set of standard GNU libraries. Obviously, this makes it difficult to reuse the existing Linux applications or libraries on Android. Based on the Linux kernel code, Android implements its own device drivers, memory management, process management and networking functionalities. Android is multitasking and allows applications to run concurrently. For example, it is possible to hear music and read or write an email at the same time. This layer is the one that is commonly used by the cell phone users.

Google initially maintained the kernel code they contributed to in the Linux public distribution. Since 2010, Google no longer maintains its Android kernel extensions in the Linux public distribution. Now Google maintains its own code tree. This has marked the branching of Android from Linux code in the public distribution.

9.4.6 Blackberry Operating System

Blackberry operating system is a proprietary operating system designed for BlackBerry smartphones produced by Research In Motion Limited (RIM). Being a proprietary operating system, details of its architecture have not been published. But, at the user level, the very good email system that it deploys is easily noticed. It supports instant mailing while maintaining a high level of security through on-device hardware-based message encryption

9.5 A Comparative Study of Mobile OSs

A comparative summary of the important features of three popular mobile OSs has been presented in Table 9.1 As can be seen, most of the features provided by the three leading mobile OSs—Android, Symbian, and Windows Phone—are more or less comparable. All these operating systems have very small footprint, run on ARM-based processors, and support demand paging. What needs to be borne in mind is that a vast majority of handsets are designed for ARM-based processors.

There are, however, many subtler points on which they differ. For example, both iOS and Windows Phone 7 are proprietary operating systems and not many details about the internal design of these two operating

systems are available. Even iOS does not facilitate the development of third-party applications. However, this has a downside. The design and development of a good operating system by a vendor is not sufficient to warrant its wide customer acceptance. It is advantageous to have a variety of applications available to run on an operating system. As has happened earlier in the computing history, it becomes a losing game for an operating system vendor, if it relies on itself alone to develop a large number of innovative applications on its operating system. In this regard, Android and Symbian being open operating systems, score over the other market players.

TABLE 9.1 A Comparison of the Features of Three Popular Mobile Operating Systems

Feature	*Android*	*Symbian OS*	*Windows Phone 7*
License	Public, Free, and Open Source	Initially was private, later became public.	Proprietary
Footprint	250 KB	200 KB	300 KB
Change of UI	Possible	No	No
Power management	Yes	Yes	Yes
Kernel	Linux with minor changes	Proprietary	Win CE
True multitasking	Yes	Yes	No
Premeptive scheduling	Yes	Yes	Yes
Demand paging	Yes	Yes	Yes
CPU architecture supported	ARM, MIPS, × 86	ARM	ARM

Android provides a flexible UI and is rich in features, being based on the open source Linux. Windows Phone lacks many features provided by Android, and iOS lacks many features even when compared to Windows Phone. For example, it lacks features such as Bluetooth file transfers, file manager, widgets and FM radio, etc.

Besides the above mentioned points, another sore point with the Windows Phone operating system is that it expects the mobile device hardware manufacturers to make their devices compliant to the Windows Phone specification to make them uniform. This is something that the manufacturers of mobile devices do not like as it makes it difficult for them to provide innovative features to attract new customers.

A major reason behind Android's success is that it facilitated competitiveness of hardware makers without good software capabilities.

The royalty-free Linux-based Android has been adopted by all the major Asian handset makers such as Samsung, HTC, LG, etc. as well as Motorola and Sony Ericsson (Farooq and Kunz, 2011).

9.6 Operating Systems for Sensor Networks

A sensor node is tiny and needs to operate in an extremely power-constrained environment. Consequently, it deploys a rudimentary operating system and does not have a kernel mode of operation. It does not support dynamic memory allocation nor does it support virtual memory. It also does not use tasks, signals, and exceptions, but uses functional call in its place. A schematic of the structure of a sensor operating system is shown in Fig. 9.4. Observe that the scheduler invokes the different application components in response to a specific event.

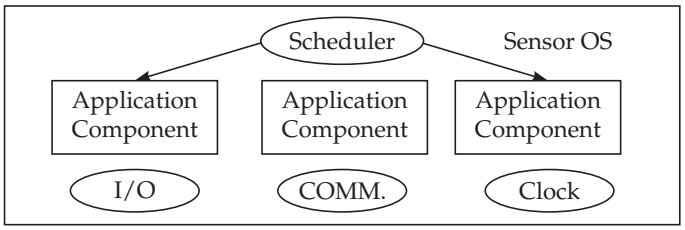


Figure 9.4 *Schematic model of the structure of a sensor operating system.*

The important operating systems that are available for sensor nodes include TinyOS, Contiki OS, Lite oS, and MANTIS (Farooq and Kunz, 2011).

SUMMARY

An operating system is a vital component of any modern mobile handset, especially the high-end ones. The increasing sophistication of mobile OS has made it much easier and less expensive to develop useful mobile applications and integrate them into the phone system. The availability of numerous innovative applications in turn fuels the demand for the smartphones. In this chapter, we first identified the special features required of a mobile OS and the constraints under which these work. We then surveyed and examined the important features of popular commercial mobile OSs.

The marketplace for mobile OSs is at present fragmented and the competition among the various popular OSs is intense. Considering that the features provided by mobile operating systems and the market dynamics are both evolving at a rapid pace, it can be argued that the market shares

of various operating systems is also rapidly fluctuating based on the novel features they are introducing and at the same time a host of several other commercial factors are also coming into play. About a decade ago, Microsoft's Phone operating system was the leader. Later, the leadership position was wrested by Symbian. Now, Android appears to be doing the same. Amid the chaos in the operating system market, even the near future is hard to predict.

At present, each operating system vendor is focusing on offering new and innovative features and no standardization has emerged. This is in contrast to the desktop and server operating systems, where POSIX is an accepted standard. In the absence of standardization, not only the users find it difficult to use phones from different manufacturers, but even the application developers and the smartphone hardware makers face considerable challenges.

FURTHER READINGS

Farooq, Omer Muhammad and Thomas Kunz, "Operating Systems for Wireless Sensor Networks: A Survey," *Sensors*, www.mdpi.com/journal/sensors, 2011.

Hall P., Sharon and Eric Anderson, "Operating Systems for Mobile Computing," *Journal of Computing Sciences*, Vol. 25, Issue 2, pp. 64–71, December 2009.

Kenney, Martin and Bryan Pon, "Structuring the Smartphone Industry: Is the Mobile Internet OS Platform the Key?," *Journal of Industry, Competition, and Trade*, Springer, Vol. 11, pp. 239–261, 2011.

Lin, F. and W. Ye, "Operating Systems Battle in the Ecosystem of Smartphone Industry," *IEEE International Symposium on Information Engineering and Electronic Commerce*, pp. 617–621, 2009.

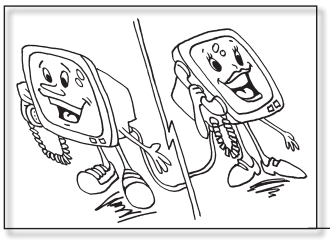
Mall, R., *Real Time Systems: Theory and Practice*, Pearson Education, 2008.

Vaughan-Nichols, S.J., "OSs battle in the smartphone market," *IEEE Computer*, Vol. 36, No. 6, pp. 10–12, June 2003.

EXERCISES

1. Explain the special features that an operating system for a mobile device needs to support compared to the features provided by a traditional operating system.
2. What is a microkernel operating system? Why is microkernel-based design being preferred for developing a mobile OS?

3. How is the operating system for a mobile phone any different from the operating system for a desktop? Name a few important commercial mobile operating systems.
4. Name three commercial operating systems for mobile phones. Outline the important functionalities supported by any one of them which are not supported by a traditional operating system.
5. Explain the principal functions of the operating system of a mobile device. Discuss how an example application can be implemented on a mobile device and the specific operating system services that it makes use of.
6. Compare the features provided by the following mobile operating systems: Android, Symbian, and Windows Phone 7.
7. Discuss the architecture of the Android operating system. Briefly identify the possible reasons as to why it has been able to rapidly improve its market share compared to its peers since its introduction a few years ago.
8. Discuss the important features provided by a modern mobile operating system, the special constraints under which a mobile OS needs to function, and the future trends in the development of mobile operating systems.
9. Using at least one suitable example, explain the flexibilities that a user would be required to sacrifice when a single-tasking operating system is used in the mobile device.
10. What is a microkernel operating system? What are its advantages and disadvantages in the context of the design of a mobile operating system?
11. Briefly explain how an operating system for a sensor network is different from a traditional operating system.



10

Mobile Application Development and Protocols

With the rapid increase in the popularity of mobile computing applications, several tools, techniques, and standards and application layer protocols have become available for developing these applications. As already mentioned, many mobile computing applications are essentially web-based client-server programs. However, mobile application development is considerably different from the traditional client-server program development. The limitations inherent to the mobile wireless networks and mobile handsets pose several challenges to the development of mobile computing applications using the traditional client-server programming techniques. In this chapter, we discuss various protocols and application development frameworks that have been proposed for mobile computing application development.

We first discuss the specific deficiencies of the mobile devices and networks that surface when web-based client-server application developments are attempted on these. Subsequently, we outline the broad approaches that have been proposed to circumvent these deficiencies. Next, we discuss the present technologies of WAP and Bluetooth. Finally, we discuss J2ME and Android-based mobile application development frameworks.

10.1 Mobile Devices as Web Clients

The World Wide Web (WWW) was originally designed to work with traditional computers such as desktops and laptops. The WWW makes implicit assumptions regarding the capabilities of the web clients (desktops and laptops) and the network over which these are run. Considering that the early mobile networks supported only voice traffic and the mobile handsets had rather primitive computing infrastructures and had screens

with only text display capabilities, it was preposterous to even think of web access from mobile handsets. As mobile handsets and mobile networks gradually evolved over time, the mobile networks could support data traffic and the handsets became reasonably powerful. This prompted practitioners and researchers to explore the feasibility of invoking web-based client-server applications from mobile handsets. Even the present-day high-end mobile devices fall far short of the capabilities of the traditional computers on many counts and the latest wireless mobile data networks such as 3G provide data rates that are many times lower than that available to traditional web clients. Another dimension of difficulty arises on account of the frequent short duration signal fading and disconnections that mobile clients (we shall refer to mobile handsets running web-based mobile applications as mobile clients) are susceptible to. It had become clear to the researchers that unless these inherent deficiencies of mobile devices and networks were properly addressed through the development of appropriate protocols and techniques, web access using mobile devices would be frustrating and meaningless.

As we have just pointed out, three main problems need to be addressed before a mobile user can meaningfully access the web and run web applications using a hand-held device. One of the problems pertains to the extremely slow network speeds that mobile devices have to do with—at present network access speeds are restricted to only a few kilobytes per second. In contrast, the typical bandwidth available for web surfing using traditional desktops is a few gigabytes per second or at least several mega bytes per second. This problem of low bandwidth of mobile networks can show up as excruciatingly slow web access to the mobile users. Another difficulty of using mobile devices as web clients, is the small screen that these devices have. A typical screen of a mobile device would not be able to satisfactorily display even a reasonable portion of a normal web page. It is simply not possible to squeeze a desktop-sized screen into the palm-sized screen of a mobile device. Further, the computing capabilities and memory sizes of the hand-held devices are severely limited. If the conventional web pages are to be directly downloaded and displayed, without incorporating any remedial measures to compensate for the miniature screens, the user would find them unusable and get utterly frustrated. The third problem concerns the disconnections and signal fading that occur when a user moves around, rendering any meaningful usage of the Internet futile.

In order to address the above three problems, a large number of approaches have been suggested in recent years. The essential idea behind all the approaches proposed to help provide meaningful web access from mobile hand-held devices is simple. The two problems of low bandwidth and small screen size can be made to solve each other. If a device has a small screen and can only display text and simple graphics, high-speed connectivity is not needed either. With this overall solution scheme in mind, let us now investigate the genesis of the problem. The founders of the WWW (Tim Burners-Lee and other scientists) had envisaged that HTML

pages can be seamlessly displayed across a multitude of devices of greatly varying capabilities. In the pure form of this objective, this can enable people to satisfactorily access the web using their respective mobile devices. However, in the later years, the development of dynamic web content and other advancements made the web move away from this ideal and slowly led to HTML pages incorporating sophisticated graphics, animation, and multimedia contents that implicitly assumed the availability of high bandwidth network access and the presence of significant processing power at the web clients.

After the WWW had an unprecedented success and almost unbelievable user acceptance, the World Wide Web Consortium (W3C) was set up by Tim Berners-Lee (the founder of WWW) at MIT. W3C is a nonprofit group that oversees the formation of various Web standards. Considering the vast popularity of mobile phones and the imminent possibility of web-based information access and application execution from mobile devices, it announced in 1998 the creation of a special version of HTML for mobile devices. This special version of HTML was called compact HTML (C-HTML). In this version of HTML, the advanced features of web such as fonts, frames, tables, graphics, and dynamic content were omitted with the intent of not only saving bandwidth but also freeing the hand-held devices of computational overload. However, care was taken to allow the use of several essential and standard web technologies such as SSL (Secure Socket Layer) for secure communication. Needless to say that this is a rather simple implementation of the overall solution schema that we have pointed out and does not take into account many issues that plague satisfactory web access by mobile clients. C-HTML underwent substantial overhaul and improvements over the next few years and helped define the present web access technology in mobile computing environments. In this chapter, we briefly discuss the progressive evolution of technology that has now made it possible for mobile clients to meaningfully access the web.

10.1.1 HDML (HanDheld Markup Language)

HDML was developed by a company called *Unwired Planet*. Unwired Planet demonstrated that by using a micro-browser developed by it, web pages written in HDML can be meaningfully accessed from a mobile handset. Like C-HTML, HDML required that a web page should avoid the use of complicated features such as tables, frames, dynamic content, etc. Hence it replaced a web page with two text layout metaphors: cards and decks. A card is defined as a *single user interaction*, which in HTML terms would be a web page. However, the analogy of a card with a web page is not exact. For example, a card can contain either one specific information display, a data entry form, or a choice menu. A web page can have many combinations of these basic types of web pages. A single HDML file consists of many cards. So, it is called a deck. An advantage of the card

and deck model in the context of mobile networks is that a click on a web page causes downloading of the entire deck associated with the web page on to the mobile device. This substantially reduces the latency of access. HDML eventually formed the basis of the WAP standard.

For facilitating the development and portability of applications over wireless communication networks, telecommunication companies, Ericsson, Nokia, Motorola, and Unwired Planet founded the WAP forum. The mandate for WAP forum was to develop technologies that would become the de facto standards and would help in the formulation of standards by the appropriate standards bodies. Later the WAP forum became a part of the Open Mobile Alliance (OMA) which has a much wider membership and mandate. OMA is a standards body which develops open standards for the mobile phone industry. In 1998, WML (Wireless Markup Language) was announced. Both HDML and WML share the same basic programming model and functionality. However, the main difference between them is that while WML is XML-based, HDML is not. An important benefit of WML being XML-based is that WML is a widely accepted standard for data interchange. Further, a company can use commercially available XML tools to generate, parse and manipulate WML, and they can also use XSL/XSLT[1] to construct WML decks from XML meta-languages.

10.2 WAP

WAP (Wireless Application Protocol) has now become the de facto standard technology for web access by a mobile client. WAP helps wireless applications to seamlessly access the Internet. WAP differs from C-HTML and HDML in that it is not a markup language—it is a complete stack of protocols. It has been designed not only to overcome specific problems caused by wireless networks during web access (such as high latency and jitter) but also to address those that are caused by the mobile devices such as small screen size and low processing power. This makes WAP much more complicated than HDML or C-HTML, but make it more effective. In fact, the WAP architecture defines an optimized protocol stack for communication over wireless media, a content description language, and a miniature browser. It replaces HTML pages with highly condensed WML pages, called cards. WML has been defined considering the limited display characteristics of the mobile devices.

A traditional web access mechanism has been schematically shown in Figure 10.1. As shown in this figure, the client and the application server communicate using the http *request* and *response* streams. We now compare this mode of web access from a traditional computer with a WAP-based

1. XSL stands for Extensible Stylesheet Language. It is a family of languages that can help to appropriately display XML documents. XSLT stands for XSL Transformations.

web access. The working of a WAP-based web access has been shown in Fig. 10.2. In this, all requests from the mobile device are sent to the WAP gateway. The WAP gateway transforms them into HTTP request-response streams, and sends them to the application/content server. The reverse of this sequence takes place when the application/content server responds to a specific request. It is the responsibility of the WAP gateway to convert a WAP request/response to an HTTP request/response. This conversion is very compute-intensive. To handle this computational workload, the WAP gateway needs to be a powerful computer. It is not difficult to see why WAP is often described as a network-centric protocol—most of the intelligence and computations associated with the WAP protocols are embedded in the network rather than the simple phone. The WAP forum released its first specification of WAP (referred to as WAP 1.0) in the year 1997. The WAP 2.0 was subsequently released. It had backward compatibility with the previous WAP protocol.

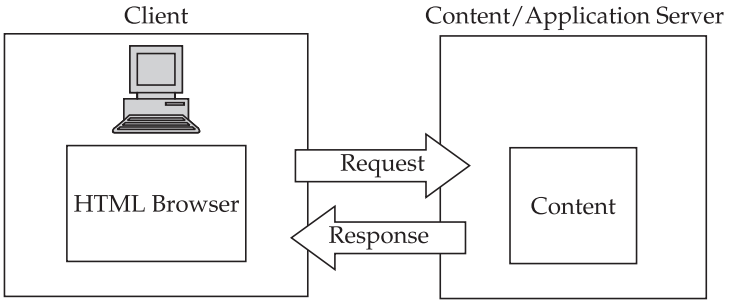


Figure 10.1 *Traditional web access.*

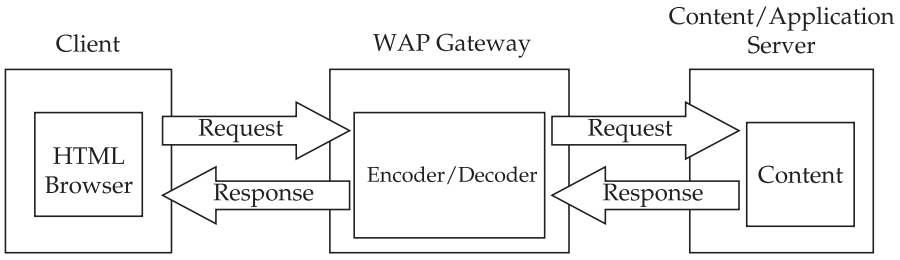


Figure 10.2 *WAP-based web access.*

Although WAP has its own protocol stack, it is designed to be compatible with the Internet. Pages in WAP are converted to the http and TCP protocol at the gateway. WAP 2.0 provides support for the protocols that are counterparts of IP, TCP, and HTTP. It is also flexible and bearer independent—meaning that WAP services can run over any specific wireless data bearer technologies such as SMS, GSM, GPRS, etc. The first web services ran over GSM, requiring the user to dial a specific phone number just as one would call an ISP from a traditional phone.

The architecture of WAP 2.0 protocol stack has been shown in Fig. 10.3. As can be seen from the figure, the WAP protocol operates at several layers. WAP defines its own specific protocols corresponding to the various layers of operation of the Internet. The protocols in the WAP stack and the bearer services in WAP are discussed below.

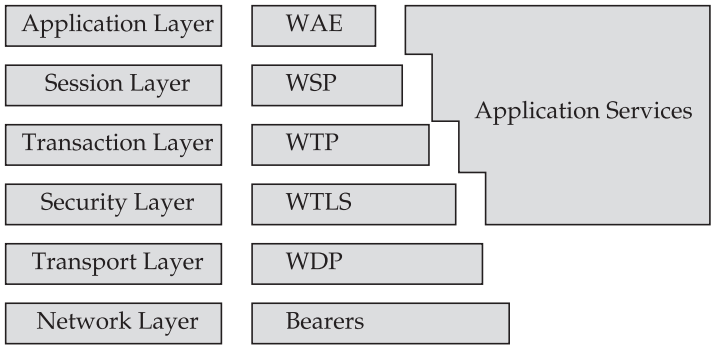


Figure 10.3 *WAP protocol stack.*

Wireless Application Environment (WAE)

WAE includes the micro-browser on the device, WML (the Wireless Markup Language), WMLS (a client-side scripting language), telephony service, and a set of formats for the commonly used data such as images, phone books, and calendars.

Wireless Session Protocol (WSP)

WSP helps establish a web browsing session from a mobile handset. WSP is based on the HTTP protocol and provides the basic session state management, and facilities reliable and unreliable data push (called WAP push).

Wireless Transaction Protocol (WTP)

The WTP layer in the WAP stack can be considered to be the equivalent of the TCP layer of the TCP/IP stack, but it takes into account the availability of low bandwidth by providing different classes of transaction services. WTP transaction services include reliable request and response that have been adapted to the wireless world. WTP handles the problem of packet loss more effectively than TCP. Packet loss is a fairly common phenomenon in wireless technologies due to factors such as atmospheric noise, signal fading, and handoff. The packet losses are often misinterpreted by TCP as network congestion, thereby drastically reducing the network throughput. The problem of packet loss is effectively handled by WTP.

Wireless Transport Layer Security (WTLS)

WTLS is the security layer that is used to transfer data securely between a mobile device and a server. It provides support for data security and privacy, authentication, as well as protection against denial-of-service attacks.

Wireless Datagram Protocol (WDP)

WDP is the bottom-most protocol in the WAP protocol suite. It functions as an adaptation layer in a wireless communication environment that makes every data network look like UDP to the upper layers by providing services for transport of data in the unreliable wireless environment. WDP invokes services of one or more *data bearers* such as SMS, GPRS, CDMA, UMTS, etc.

Bearer Interfaces

A bearer is a low-level transport mechanism for network messages. Considering the diversity of transport technologies, WAP is designed to operate with SMS (Short Message Service) to GPRS (General Packet Radio System), UMTS and IP. GPRS is relatively faster, and provides "always-on" connections for wireless devices. WAP supports circuit-switched bearer services such as dial-up networking using IP and Point-to-Point Protocol (PPP). However, packet-switched bearer services are much better suited than circuit-switched bearer services for mobile devices as they can provide more reliable services in the unreliable wireless connection environment.

10.3 J2ME

Sun Microsystems Ltd. created the Java language in 1995. Even though Internet had already made its appearance by 1995, but none of the available programming languages could effectively work with Internet. Java for the first time made it possible for programs to be run by making it possible for web clients to download and run applets. Besides, it had several other advantages. As we all know, Java was a resounding success and quickly found widespread acceptance. This was followed by Java Enterprise Edition (J2EE) for server side programming. Later Java2 Micro Edition (J2ME) was developed for supporting programming of mobile devices. J2ME has been targeted for use in very small devices (called J2ME devices) such as smartphones, interactive television set-top boxes, pagers, PDAs and other wireless devices. It includes different virtual machines and application programming interfaces (APIs).

One of the identified design objectives for J2ME was to provide portability across several mobile platforms, so that the same application could be run across the mobile devices manufactured by different vendors

and running different operating systems. Mobile hand-held devices have very restricted capabilities compared to the traditional computers. J2ME was designed keeping all the above constraints in mind. Some of the major differences between conventional computers and the J2ME devices are the following:

- Limited processing power
- Limited system memory
- Limited storage capacity
- Small display
- Low battery power
- Limited connectivity to internet

J2ME includes a miniature version of JVM called KVM (K Virtual Machine) which can run small Java programs (called midlets) on the mobile devices. This brought the "write once, run everywhere", principle espoused by Java to the mobile handsets.

The J2ME applications are lightweight and support many useful features. A few important capabilities of J2ME programs are the following:

- Opening UDP connections between two devices.
- Establishing HTTP connections with a server
- Making Socket connections.
- Bluetooth programming .
- Bar code scanning.

Some of the popular applications of J2ME are the following:

- Automotive systems
- Set-top boxes and interactive televisions
- Network-connected consumer devices that use graphic user interface (GUI)

10.3.1 J2ME Configuration

J2ME takes into account the fact that there are a wide variety of mobile devices with varying capabilities and hence the need for supporting a diverse range of features and functions. A J2ME *configuration* targets to make it applicable to devices with a specific range of capabilities. A *profile* on the other hand, selects a configuration and a set of APIs to target a specific domain of applications. The selection of an appropriate configuration and profile are therefore the crucial parts of any application development. Since lightweight appliances do not need the capabilities of the entire Java2 platform, by selecting the appropriate configuration and profile, the resource requirements and the cost of running the applications can be reduced to a large extent. The J2ME configuration parameters include the following:

- Availability of memory space and memory type
- Specification of processor in terms of speed and type
- Network connectivity of the device

J2ME currently defines two configurations:

Connected Limited Device Configuration (CLDC) for handheld devices

CLDC is targeted to the lower-end range of consumer electronic devices that are designed using 16-bit or 32-bit small computing devices with limited memory. These devices usually have between 160 KB and 512 KB of available memory. Usually powered by battery, they have low bandwidth network wireless connections. These devices include pagers, personal digital assistants, cell phones, dedicated terminals, and hand-held devices. These devices use a stripped-down version of the JVM. Java enables the purchase and download of small Java applications (called midlets) to the handsets.

Connected Device Configuration (CDC) for plug-in devices

CDC devices are rather the higher-end devices that use a 32-bit processor, have at least 2 MB of memory available, and implement a complete functional JVM. CDC devices include digital set-top boxes, home appliances, navigation systems, point-of-sale terminals, and smartphones. The programming environment for application software development consists of Java Virtual Machine (JVM) and a core collection of java classes. The J2ME virtual machine could not support all the Java language features or instructions byte codes and the software optimization provided by J2SE (Java 2 Standard Edition) virtual machine due to the limitations in processor and memory of low-end devices. Therefore, J2ME VMs usually incorporate some part of the Java Virtual Machine and a subset of the Java Language syntax.

J2ME is modular and scalable. J2ME consists of four layers of functionalities as shown in Fig. 10.4. We explain the functionality of these layers in the following:

Java Virtual Machine (JVM) layer: Java Virtual Machine (JVM) is implemented in this layer which is customized based on the capability of a particular device and a particular J2ME configuration.

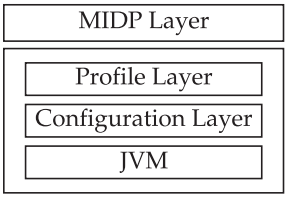


Figure 10.4 *J2ME architecture.*

Configuration layer: The features of Java Virtual Machine (JVM) and the available Java class libraries for a specific category of devices are defined in this layer. For profile implementers, this layer plays a vital role, but is less visible to the users.

Profile layer: The set of application programming interfaces (APIs) available on a particular family of devices is defined in this layer. Applications are written for a particular profile, and applications are portable to any device that supports the specific profile for which it was developed. A device can support multiple profiles. This is the layer that is most visible to users and application providers.

Mobile Information Device Profile (MIDP) layer: The Java APIs which are present in this layer provide reusable functionalities corresponding to user interface, persistence storage, and networking.

10.4 Android Software Development Kit (SDK)

Cell phones are small in size and therefore can easily be carried everywhere like a wallet. Considering their portability and the powerful feature sets that they provide, they have now come to play an important part in today's society. Now with every passing day, mobile phones implement new and innovative functionalities and have come to possess a surprisingly large number of functionalities. For example, many handsets provide facilities for radio and television reception, timer and clock, Internet access, camera, and calculator, just to mention a few. These applications are well known and standard. These are usually developed by the operating system developers themselves and are bundled with the operating system. However, the true power and distinguishing features of a handset come into picture with the third-party developed applications. Therefore, a successful mobile operating system needs to facilitate the development of third-party applications. In this regard, open operating systems stand out in facilitating the development of third-party applications and Android application development tools have now been well received.

10.4.1 Android SDK Environment

The Android SDK (Software Development Kit) is a mobile application development framework using which developers can create applications for the Android platform. The Android SDK provides the tools and libraries necessary to develop applications that can run on Android-based devices. An important advantage of Android SDK is the low processor and RAM requirements. Besides, Android SDK can be installed on almost all common operating systems such as Windows, Mac OS, and Linux. The

SDK comes with an Integrated Development Environment (IDE) and other tools which are required to develop applications. Android SDK converts Java byte code to Android's Dalvik VM bytecode. It is important to realize that while developing android-based applications, the developer codes the applications using Java but the mobile device running Android does not finally run the Java byte code but runs the Dalvik VM byte code. For this reason, applications developed using J2ME cannot be run directly on Android mobile phones although they both use execution environments that are derivations of JVM.

The environment to develop applications for Android consists of the Android SDK, the IDE Eclipse and the Java Development Kit (JDK). After installing the SDK, which is done by simply extracting the downloaded ZIP file in a folder, the path to the SDK has to be set in the path environment variable. Eclipse can be used as the IDE, which also automatically installs the Android SDK as a plug-in. After installing this plug-in, one can start the development of Android applications.

10.4.2 Features of SDK

Using the SDK, one can either run the application on the actual Android device or a software emulator on the host machine. This is achieved by using the Android Debug Bridge (ADB) available with the SDK. ADB is a client-server program and includes three main components:

- A client program which runs on the developer's (called host) machine. One can invoke a client from a shell by issuing an adb command.
- A daemon program which runs as a background process on each emulator or device instance. It is the part that actually manages the communication with the handset or the emulator and helps in executing the application.
- A server program which runs as a background process on the host machine. The server manages communication between the client and an adb daemon that runs on the emulator or the Android handset.

10.4.3 Android Application Components

Application components are the essential building blocks of an Android application. The following are the four components of an Android application.

Activity: Each activity presents a GUI screen of an application. For example, a chat application might have one activity that allows to create a chat, another to view the previous chat sessions, etc. Different activities form a cohesive chat application.

BOX 10.1 Android application development

Android originated from a small software company acquired by Google and is now owned by the OHA (Open Handset Alliance) of which Google is a member. The Open Handset Alliance (OHA) is a consortium of a large number of companies. OHA focuses on the development of open standards for mobile devices. Besides Google, the present members of OHA include HTC, Sony, Dell, Intel, Motorola, Qualcomm, Texas Instruments, Samsung Electronics, LG Electronics, T-Mobile, Nvidia, and Wind River Systems. Android SDK is a free tool in the sense that anyone can use Android SDK to develop applications to run on Android OS. Android makes use of only non-proprietary techniques such as Java programming and extensible mark up languages. This makes it an open tool. Google's Software Development Kit supports a relatively large subset of the Java Standard Edition 5.0 library. This includes the following APIs:

java.io: Contains classes for system input and output through data streams.
java.net: Contains classes for implementing networking applications.
java.security: Contains classes for the security framework.
java.util: Contains several utility classes such as event model, date/time facilities and internationalization.
javax.sound.midi: Contains a collection of classes for input and output, sequencing and Musical Instrument Digital Interface.

A few of the available important Android APIs are the following:

android.graphics: Contains classes for low level graphics.
android.media: Contains classes for media player and recorder functionalities.
android.net: Contains classes for network access that are not included in the normal java.net APIs.
android.opengl: Contains classes for utilities (OpenGL) for 2D/3D graphics.
android.telephony: Contains classes that help receive and monitor phone calls.
android.database.sqlite: Contains classes that help manage a private database.
android.view: Contains classes to handle screen layout.
android.widget: Contains classes for user interface elements to use on the application screen.
android.content: Contains classes for accessing and publishing data on the device.
android.app: Contains classes for high level classes encapsulating the overall Android application model.

Content providers: Content providers are used for reading and writing data that are either private to an application or shared across applications. By using the content provider, an application can query or modify the stored data.

Service: A service denotes a background task and not for interacting through a user interface. For example, a service might play music in the background while the user is interacting with a different application.

Broadcast receivers: The broadcast receiver responds to broadcast announcements by an application. For example, a battery monitoring application might broadcast that the battery is low. Based on this, the music player might reduce the volume or the screen display may be dimmed.

10.4.4 Android Software Stack Structure

To a user of a mobile handset, various functionalities are provided by a cooperative working of a number of application programs and system programs. These collection of programs can be decomposed into a hierarchy of four layers as already discussed in Section 9.4.5 (see Fig. 9.3). The functionalities of these layers are also discussed in that section, which may be referred to for ease of recollection.

10.4.5 Advantages of Android

Application development in the Android platform is becoming popular due to the many advantages that it offers.

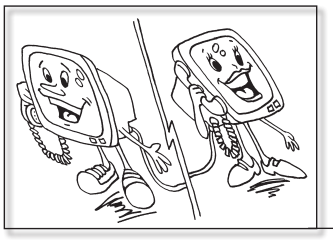
- The mobile platform Android is an open platform and can be ported on almost every type of cell phone.
- The Android SDK to develop applications is possible on every operating system.
- Android requires a low footprint of 250 KB.
- The emulator of the Android platform has a modern design and is easy to use.
- Application installation on the emulator/device is possible via Android Debug Bridge (adb) or via Eclipse (with ADT plug-in)
- Google offers a very good documentation as well as many examples which cover the most basic and important techniques used to get in touch with Android and the application development on it.
- Android supports robust libraries for media access, communication and data transfer.
- Android offers a real database SQLite using which meaningful data manipulation and data sharing across applications is possible.
- Android has an integrated web browser which gives an experience similar to web browsing using a desktop PC.
- Android uses the standardized and open programming language Java.

SUMMARY

In this chapter, we discussed the application layer protocol WAP. We also discussed J2ME and Android SDK that facilitate application development for the mobile phones. These come with a programming environment installable on a Windows or Desktop PC where the application can be developed. After the application is run on a mobile device emulator, it is downloaded on the actual mobile device.

EXERCISES

1. What problems would occur if an attempt is made to enable the mobile user to browse the web by interfacing the mobile networks and the mobile devices directly to the WWW with only trivial extensions to the existing web protocols?
2. Why is HTML-based standard web access not very meaningful in mobile environments? How does WAP address this issue?
3. What is C-HTML? How did C-HTML address problems that arose while accessing traditional web using mobile handsets? What are the shortcomings of C-HTML?
4. What is HDML? What were the problems of traditional web access that HDML tried to solve? How have these problems been solved?
5. Explain WAP 2.0 protocol and its architecture. Briefly explain why WAP is called a network-centric protocol.
6. Distinguish between traditional computing devices and the J2ME devices.
7. What is J2ME? What facilities does J2ME offer to a mobile application developer? Briefly explain the architecture of J2ME.
8. Write a short note on Android software stack.
9. Using a suitable schematic diagram, explain how a traditional web access differs from that done using WAP.
10. Briefly write how an application can be developed using the Android SDK.



11

Mobile Commerce

So far we have discussed the various layers of protocols used in mobile computing. In this chapter, we provide an overview of mobile commerce as an example application of the mobile computing infrastructure. Mobile commerce (M-commerce in short) is an important application of mobile computing. Mobile commerce, in simple words, involves carrying out any activity related to buying and selling of commodities, services, or information using the mobile hand-held devices. M-commerce has over the last decade become extremely popular. The popularity of m-commerce can be traced to the convenience it offers both to the buyers and sellers.

Though buying and selling may appear to be an obvious application of a mobile hand-held device, an important issue in M-commerce is how payments can be made securely and rapidly as soon as a buyer decides to make a purchase. So far, the use of computers and networking in trade related transactions has been limited to automatic teller machines (ATMs), banking networks, debit and credit card systems, electronic money and electronic bill payment systems (E-payment). Each of these modes of transaction has its own advantages and disadvantages and warrants a detailed investigation for use in M-commerce. Mobile payment is a natural evolution of E-payment schemes and has found an important place in M-commerce.

11.1 Applications of M-Commerce

M-commerce applications can be broadly categorized into either B2C or B2B. These two categories of applications are discussed next.

BOX 11.1 Evolution of M-commerce

Money is now an important element of all business and trade. In older times, money did not exist. What existed was a simple "barter system" where things could be exchanged, say, fish for grains, The evolution of currency (money) gave birth to the concept of a "marketplace". In a marketplace, commerce is a function of 4 Ps—Product, Price, Place and Promotions. All these four components play a vital role for a business transaction to take place. Different combinations of 4Ps determine the different forms of commerce. Once the marketplace came into existence, a few pioneers realized that people would be ready to pay extra if products could be delivered at the customer's doorsteps. A small change to two of the Ps, Price and Place, led to the convenience of getting products at customers' homes. This concept delighted the customers and thus, the concept of "Street Vendors" was born. When the postal system came into being, sellers found a new avenue and started using mails to describe their products. It ultimately led to the concept of "Mail Order Cataloguing". A mail order cataloguer buys goods and then sells those goods to the prospective customers. A mail order catalogue is a list of the goods that the cataloguer deals with. From this point, the evolution of the "Teleshopping" networks was inevitable with the development of the Internet. The latest generation of commerce is being done over the Internet. The Internet provides a virtual platform where sellers and buyers can come in contact for sale and purchase of goods and services, even though they may be thousands of miles apart, belong to different countries, and might speak different languages. "E-Commerce" has emerged as a boundary-less trade medium accelerating the pace of globalization.

The Internet has already reached the home of most customers. In this context, the distribution channel has started to assume a new meaning to the e-marketer. With options of paying online through debit and credit cards, on-line transactions have become purely electronic. In this context, the difference between E-commerce and M-commerce is that E-commerce is limited to PC users with an Internet connection, while M-commerce has been adopted by the mobile phone users.

11.1.1 Business-to-Consumer (B2C) Applications

Business-to-consumer (B2C) is a form of commerce in which products or services are sold by a business firm to a consumer. B2C is an important category of mobile commerce applications and is reported to be nearly half of the total M-commerce market (Varshney et al., 2000). A few examples of B2C applications are given below:

Advertising

Using the demographic information collected by the wireless service providers and based on the current location of a user, a good targeted advertising can be done. The wireless service provider may also keep track of the history of the purchases made by customers by directing

advertisements to mobile phones. Customers may also solicit specific advertisements. For example, suppose a consumer in a shop is fascinated by a new electronic product and wishes to buy it but only after getting more details about it. For this purpose, he can view all the relevant advertisements for the product by taking the picture of the bar code using his mobile device.

Comparison shopping

Consumers can use their mobile phones to get a comparative pricing analysis of a product at different stores and also the prices of the related products. For example, suppose consumers visiting a shop can use their mobile phones to access a web-based comparison shopping application. By scanning the bar code on a product, the consumer can see the price of this product at different shops in the adjacent area. After seeing how the product is priced at different shops, the consumer may decide to buy from a shop where it is competitively priced. In a similar manner, consumers can also access product reviews from consumer organizations or customers.

Information about a product

Consumers can access additional information about products through their mobile phones. Assume that a consumer buys some medicine in a pharmacy shop, but cannot read the dosage instructions on the carton given in German and Spanish languages only. The consumer can, however, scan the bar code on the pack using the mobile device to read the dosage instructions in the English language, which he knows.

Mobile ticketing

Mobile phones can be used to purchase movie tickets (called m-tickets) using credit cards. After the payment is received, a unique bar code is sent to the purchaser's mobile phone by an SMS. The purchase can gain entry to the movie hall by showing the bar code downloaded into the mobile device to a bar code reader at the entrance. Consider another example. A customer books a train ticket using the mobile phone. An m-ticket or a text message is sent to the mobile phone. Inside the train, by showing the m-ticket on his mobile phone to the ticket collector, the traveller occupies the seat.

Loyalty and payment services

In this application, mobile phones can replace the physical loyalty cards. Having signed up for a supermarket loyalty scheme, a unique bar code is sent to a consumer's mobile phone. After shopping at the same supermarket, the consumer shows the bar code at the cash counter and accumulates points based on the total amount spent.

Mobile phones can be used to make payments. For example, consumers can buy canned drinks from a vending machine by moving their phones close to an RFID enabled phone reader. Payment is made through the person's mobile phone bill. Consumers pay their bills by simply scanning the bar code on the bill and using their mobile phones to process payment.

BOX 11.2 Radio Frequency Identification

A Radio Frequency Identification (RFID) tag attached to a product, animal, or person for the purpose of identification and tracking, makes use of radio waves. Some tags can be read from several metres away and beyond the line of sight of the reader.

Interactive advertisements

In an interactive advertisement, customers can scan a bar code in an advertisement for a product appearing on a TV screen using their mobile phones. By scanning the bar code, the consumer can order the product by invoking an internet application.

Catalogue shopping

Mobile phones can be used to place orders for products listed in a catalogue. For example, a consumer might receive a catalogue by SMS from a catalogue shopping company. Each product on sale is accompanied by a unique bar code. By scanning the bar codes, the consumer can buy products directly from the catalogue shopping company.

11.2 Business-to-Business (B2B) Applications

Business-to-business (B2B) is a form of commerce in which products or services are sold from a company to its dealers. For example, a company that manufactures TV sets would normally sell it through a dealer network rather than selling the product directly to the consumers. Here, the manufacturer and the dealers are said to be the B2B partners. A few examples of B2B applications of M-commerce are given below.

Ordering and delivery confirmation

In this application, mobile phones can be used by dealers to order products. The orders can be sent to the supplier in a standard format. By scanning the bar code on a product by using the camera of a mobile phone and specifying the quantity required through a simple application, a dealer can automatically re-order goods.

Mobile phones can be used to gather information about the status of consignments during the transport and delivery process. By reading the

bar code on a packet using a mobile device, a truck driver can confirm in real-time that a consignment has been delivered.

Stock tracking and control

Mobile phones can be used to keep track of the stock in a distributed inventory system and send updates to a central database. By using a mobile phone to scan bar codes or RFID tags on products, employees can update the stock in real time. Mobile phones (as opposed to dedicated mobile scanners already used in warehouses) are the particularly attractive tools where the stock is stored in many locations. An example of such an application is the stock control of apparel items warehoused in the various department stores.

Supply Chain Management (SCM)

Information about the supply chain processes can be made available via mobile devices. By scanning an RFID tag using a mobile phone, it is possible for a manager or anyone in the supply chain to check information about a product's state in the supply chain. This kind of accurate information can help manage the business efficiently.

Mobile inventory management

An interesting new B2B application reported in (Varshney et al., 2000), envisages a "rolling inventory" consisting of multiple trucks carrying large amounts of goods. Whenever a store needs certain goods, it locates the nearest truck to take delivery of the required goods. This reduces the amount of inventory and cost for both the producers and the retailers. It also has the potential to drastically reduce the delivery times and help in just-in-time delivery of goods.

11.3 Structure of Mobile Commerce

In mobile commerce, a content provider implements an application by providing two sets of programs: client-side and server-side. The client-side programs run on the microbrowsers installed on the users' mobile devices. The server-side programs, performing database access and computations, reside on the host computer (servers). The architecture of a mobile commerce framework is shown in Figure 11.1. Below, we explain the functionalities of the various layers of the architecture of a mobile commerce framework.

Mobile devices

Hand-held devices essentially present user interfaces to the mobile users. The users specify their requests using the appropriate interface programs,

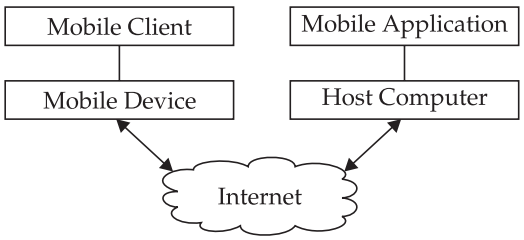


Figure 11.1 *Architecture of a mobile commerce framework.*

which are then transmitted to the mobile commerce application on the Internet. The results obtained from the mobile commerce application are displayed in suitable formats.

BOX 11.3 Features required of a mobile device to enable mobile commerce

To enable mobile commerce to be used widely, a mobile device should provide the following facilities:

- Good Internet connectivity
- Ability to display rich content such as images
- Have a good quality camera with auto focus
- Screen should be able to properly display the bar codes
- Ability to read the RFID tags
- MMS (Multimedia Message Service), SMS (Short Message Service)
- Ability to communicate between the mobile device and the supporting network
- Ability to scan bar codes
- Ability to interact with the Point-of-Sale (PoS) terminals.

Point-of-Sale (PoS) usually means a checkout counter in a shop or supermarket. More specifically, the point-of-sale often refers to the hardware and software used for handling customer purchases at the checkout desks. An example of a PoS terminal is an electronic cash register. Nowadays, the point-of-sale systems are used in almost every supermarket and are used in many retail stores too.

Mobile middleware

The main purpose of mobile middleware is to seamlessly and transparently map the Internet content to mobile phones that may sport a wide variety of operating systems, markup languages, microbrowsers, and protocols. Most mobile middleware also handle encrypting and decrypting communication in order to provide secure transactions.

Network

Mobile commerce has become possible mainly because of the availability of wireless networks. User requests are delivered either to the closest wireless access point (in a wireless local area network environment) or

to a base station (in a cellular network environment). Wired networks are optional for a mobile commerce system. However, host computers (servers) are generally connected to wired networks such as the Internet. So user requests are routed to these servers using transport and/or security mechanisms provided by wired networks.

Host computers

Host computers are essentially servers that process and store all the information needed for mobile commerce applications. Most application programs used in the mobile commerce are hosted on these. These applications usually consist of three major components: web servers, database servers, and application programs and support software. The web servers help interact with the mobile client. The database servers store data. The application program is the middleware that implements the business logic of the mobile commerce application.

11.4 Pros and Cons of M-Commerce

Like any other technology, M-commerce has its own advantages and disadvantages, which are discussed below.

Advantages

The following are the major advantages of M-commerce:

1. For the business organization, the benefits of using M-commerce include customer convenience, cost savings, and new business opportunities.
2. From the customer's perspective, M-commerce provides the flexibility of anytime, anywhere shopping using just a lightweight device. The customer can save substantial time compared to visiting several stores for identifying the right product at the lowest price.
3. Mobile devices can be highly personalized, thereby providing an additional level of convenience to the customers. For example, a repeat order for some items can be placed at the touch of a button.

Disadvantages

The following are the major shortcomings of using M-commerce.

1. Mobile devices do not generally offer graphics or processing power of a PC. The users are therefore constrained to use small screen and keyboard and low resolution pictures and videos.
2. The small screens of mobile devices limit the complexity of applications. For example, the menu choice, and text typing capability are severely constrained

3. The underlying network imposes several types of restrictions. For example, the available bandwidth is severely restricted, and international reach is prohibitively expensive. Therefore, ubiquity of M-commerce is hard to achieve in practice.

11.5 Mobile Payment Systems

Mobile payments are a natural evolution of E-payment schemes. A mobile payment (or M-payment) may be defined as any payment instrument where a mobile device is used to initiate, authorize and confirm an exchange of financial value in return for goods and services. Mobile devices include mobile phones, PDAs, and any other device that connects to a mobile network for making payments. A mobile device can also be used for payment of bills (especially utilities and insurance premiums) with access to account-based payment instruments such as electronic funds transfer, Internet banking payments, direct debit and electronic bill presentment.

An important issue which influences the establishment of the mobile payment procedure is the technical infrastructure needed on the customer side. A sophisticated technology may fail if the customer is not able to handle it with ease. On the other hand, simple procedures based on simple message exchange via short messaging services (SMS) may prove more successful. Thus, at present and possibly in the near future, the important payment solutions will be SMS-based, which can easily be charged to the mobile phone bill of customers. Some other procedures may integrate two or more solutions. It is true that M-payments are still in their infancy. Some important problems dogging the M-payment schemes are those of security, privacy, and guarding against frauds. The challenges for providing secure transactions are many and range from physical theft of a mobile device which can be subsequently used for fraudulent payments. The M-payment solutions are still being developed with standards defined on individual business segments. The other interesting areas related to M-commerce payment not mentioned in this text are the issues of standardization and interoperability. In the following section, we discuss the important types of mobile payment schemes.

11.5.1 Mobile Payment Schemes

Three popular types of M-payment schemes are currently being used:

(a) Bank account based
(b) Credit card based
(c) Micropayment

In each of these approaches, a third party service provider (bank, credit card company, or telecom company) makes a payment on the customer's behalf. An important question that needs to be answered is since the third party incurs an overhead in making the payment, how would it recover the cost. First, the service provider may require pre-payment from users, leading to some financial gain through investment of this fund. A service provider may charge a small amount as service charge, which can decrease with increasing customer base.

Bank account based M-payment

In this scheme, the bank account of the customer is linked to his mobile phone number. When the customer makes an M-payment transaction with a vendor or in a shopping complex, based on a Bluetooth or wireless LAN connectivity with the vendor, the bank account of the customer is debited and the value is credited to the vendor's account.

BOX 11.4 mChek—a new payment scheme

mChek (Dahlberg et al., 2007) is a new payment system that links a debit or credit card, or a bank account, to a mobile phone, allowing one to make payments from the mobile phone. Once registered, the user can pay phone bills, transfer talktime to a friend's account, book tickets for flights, movies, pay water bills, electric bills, etc. from the mobile with the help of simple instructions. This scheme has a tie-up with Airtel which allows Airtel subscribers to download mChek application which provides simple graphic interface to use mChek. There are no charges for downloading as well as using mChek as of now.

Credit card based M-payment

In the credit card based M-payment, the credit card number is linked to the mobile phone number of the customer. When the customer makes an M-payment transaction with a merchant, the credit card is charged and the value is credited to the merchant's account. Credit card based solutions have a limitation, being heavily dependent on the level of penetration of credit cards in a country. Currently, the penetration level of credit cards is rather low but is expected to grow substantially in the coming years.

Micropayment

Micropayment is intended for payment for small purchases such as from vending machines. The mobile device can communicate with the vending machine directly using a Bluetooth or wireless LAN connection to negotiate the payment and then the micropayment is carried out. A customer makes a call to the number of a service provider where the per call charge is

equal to the cost of the vending item. Thus, the micropayment scheme is implemented through the cooperation of the mobile phone operator and a third party service provider. This approach has been used for vending from Coca-Cola machines.

BOX 11.5 Payment settlement solutions

Payment solutions can be categorized on the basis of the payment settlement methods, which are instant-paid, postpaid, prepaid, or a combination of these. In the prepaid solution, customers buy a smart card where the amount equivalent is stored and then they can pay for goods or services desired. Subscription to services can also be considered as the prepaid type of payment. The prepaid type of solution allows privacy to users since at no point of the process is it required to disclose any personal data. The instant-paid solution is that in which payment settlement is done as soon as users confirm the payment as in direct debiting systems. In the postpaid solution, customers pay for goods or services later. Payment by credit card and phone bill is an example of this solution.

11.6 Security Issues

M-commerce is anticipated to introduce new security and privacy risks beyond those currently found in E-commerce systems (Ghosh and Swaminatha, 2001). Users of mobile devices can be difficult to trace because of roaming of the users. Also, the mobile devices go on-line and off-line frequently. Thus, attacks would be very difficult to trace.

Another risk unique to the mobile devices is the risk of loss or theft. A mobile device that is stolen or has fallen into wrong hands can cause frauds that are difficult to track and prevent. A major problem in this regard is the lack of any satisfactory mechanism to authenticate a particular user.

SUMMARY

M-commerce denotes ubiquitously carrying out any activity associated with buying or selling of goods. M-commerce gives added flexibility, cost advantages, and conveniences to both the buyer and the seller of goods. For the sellers, M-commerce provides significant business benefits and market reach. The customers benefit by getting the best deal available. M-commerce even gives the traditional media (newspapers, magazines, TV) the opportunity to use the advertisements appearing in their media in innovative ways such as automated searching of recent advertisements by customers in a shopping complex. The wireless service providers can also see increased traffic from M-commerce. Considering these benefits to all the parties concerned, M-commerce is poised to take off in a big way.

FURTHER READINGS

Dahlberg, T., et al., "Past, present and future of mobile payments research: a literature review," *Electronic Commerce Research and Applications*, doi:10.1016/j.elerap, 2007.02.001.

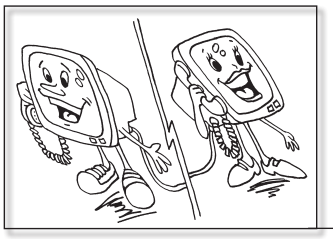
Ghosh K. Anup, and T. Swaminatha, "Software Security and Privacy Risks in Mobile E-Commerce," *Communications of the ACM*, pp. 51–54, February 2001, http://main.mchek.com.

Varshney, U., R.J. Vetter, and R. Kalakota, "Mobile commerce: a new frontier", *IEEE Computer*, Vol. 33, pp. 32–38, October 2000.

Zheng, X. and D. Chen, "Study of mobile payments systems," *IEEE International Conference on E-Commerce*, CEC 2003, pp. 24–27, Digital Object Identifier 10.1109/COEC.2003.1210227, June 2003.

EXERCISES

1. What do you understand by the mobile payment system? Briefly explain an application where mobile payment may be useful. Explain the different payment systems that are available.
2. What do you mean by the 4 Ps of commerce? Explain the different forms of commerce that are obtained by varying the interpretation of the Ps.
3. What do you understand by M-commerce? What are the advantages and disadvantages of M-commerce.
4. Explain any two applications of M-commerce.
5. What do you mean by B2B and B2C commerce? Give examples of M-commerce for these two categories of commerce.
6. What is RFID? Briefly explain the principle of its working. Explain a few applications in which RFID is useful.
7. What is micropayment in M-commerce? How is micropayment achieved?



Glossary

Access Point: It is connected to a wired network through cables and provides the connectivity between the wireless device and the wired network.

Ad Hoc On-Demand Distance Vector (AODV): AODV has the basic route-discovery and route-maintenance like the DSR protocol and uses the hop-by-hop routing, sequence numbers and beacons.

Adaptation: Adaptation implies the ability of a system to adjust to bandwidth fluctuation without inconveniencing the user.

Agent Advertisement: Generally the foreign and the home agents advertise their presence through periodic agent advertisement messages. An agent advertisement message, lists one or more care-of-addresses and a flag indicating whether it is a home agent or a foreign agent. Agent advertisement is a popularly used method in agent discovery.

Agent Discovery: During call establishment it is necessary for a mobile node to determine its foreign agent. This task is referred to an agent discovery. The two discovery methods used are:
1. Agent advertisement and 2. Agent solicitation.

Agent Solicitation: In case a mobile node (MN) does not receive any care-of-address COA, then the MN should send an agent solicitation message.

Amplifier: An amplifier amplifies the strength (usually voltage) of a signal.

AMPS: The first generation (1G) of the cellular system was designed in the late 1960s and was deployed commercially in the early 1980s. The first commercial 1G system in the United States was known as Advanced Mobile Phone System (AMPS). This was a completely analog system.

Antenna: The functions of an antenna are to convert electrical signals from a transmitter to a propagating electromagnetic RF wave, or conversely, to convert a propagating RF wave to electrical signals in a receiver. In a transceiver, a transmitter and a receiver are co-located for full-duplex communications, the same antenna is usually used to both transmit and receive.

Application Tier: Application tier is also known as the middle tier. This layer co-ordinates the application, processes the commands, makes logical decision evaluations and performs the calculations.

ARP (Address Resolution Protocol)**:** The ARP protocol is used by IP to find the hardware (physical) address of a computer network card based on the IP address.

Atomicity: Either all operations of a transaction are reflected on the database or none at all.

Atomicity Relaxation: In this modified scheme, an MH is allowed to submit 'pieces' of transactions from different cells according to their movements. This approach weakens the atomicity property and requires the ability of breaking a transaction so that the sub-transactions can be concurrently executed and interleaved with the sub-transactions of other transactions while guaranteeing other ACID properties.

Automated Local Evaluation in Real Time (ALERT): ALERT is an automated flood alarming system that is probably the first well-known real wireless sensor network deployment. ALERT provides the important real-time rainfall and water level information to evaluate the possibility of potential flooding. ALERT sensor networks are usually equipped with meteorological sensors. Currently, ALERT is deployed across most of the western United States such as California and Arizona.

B2B Ordering and Delivery Confirmation: Mobile phones can be used by dealers to order products. The orders can be sent to the supplier in a standard format. By scanning the bar code on a product with the camera of a mobile phone and using a simple application to state the quantity required, the owner of a small shop can automatically re-order goods.

B2B Supply Chain Management (SCM): Information about the product's state in a supply chain process can be checked via mobile devices in order to manage the business efficiently.

B2B Tracking: Mobile phones can be used to keep track of the stock in a distributed inventory system and send updates to a central database. By using a mobile phone to scan bar codes or RFID tags on products, employees can update the stock in real time.

B2C Advertising: Advertisements are sent directly to mobile phones. For example, suppose a consumer in a shop is fascinated by a new electronic product and wishes to buy it but only after getting more details about

it. For this purpose, he can view all the relevant advertisements for the product by taking the picture of the bar code using his mobile device.

B2C Catalogue Shopping: Mobile phones can be used to place orders for products listed in a catalogue. For example, a consumer receives a catalogue by SMS from a catalogue shopping company. Each product on sale is accompanied by a unique bar code. By scanning the bar codes, the consumer can buy products directly from the catalogue shopping company in a similar way to accessing a web-based catalogue on his mobile phone.

B2C Information About a Product: Consumers can access additional information about products through their mobile phones.

B2C Interactive Advertisement: An interactive advertisement allows customers to scan a bar code appearing on a TV screen using their mobile phone.

B2C Loyalty and Payment Services: Mobile phones can replace the physical loyalty cards. Having signed up for a supermarket loyalty scheme, a unique bar code is sent to a consumer's mobile phone. After shopping at the same supermarket, the consumer shows the bar code at the cash counter and accumulates points based on the total amount spent.

B2C Mobile Ticketing: Mobile phones can be used to purchase and redeem tickets. A traveller can buy movie tickets online.

B2C Shopping: Consumers can use their mobile phones to get a comparative pricing analysis of a product at different stores and also the prices of the related products.

Bandwidth Constrained, Variable Capacity Links: Wireless links have significantly lower capacity than their wired counterparts. Also, due to issues such as multiple access, fading, noise and interference conditions, the bandwidth of the wireless links can change arbitrarily with time.

Black Hole: In this type of attack, a node can set up a route to some destination via itself, and when the actual data packets are received they are simply dropped. This forms a black hole, where data enter but never leave.

Blacklist: Some ad hoc routing protocols try to tackle the security problem by keeping a list of perceived malicious nodes. Each node has a blacklist of, what it thinks, bad nodes and thereby avoids using them when setting up routing paths.

Bluetooth Baseband: It is concerned with connection establishment and other tasks such as packet format, timings, etc.

Bluetooth Link Manager Protocol (LMP): It is responsible for link setup between the Bluetooth devices. It also includes some security issues such as (i) authentication and (ii) negotiation and control of baseband packet sizes.

Bluetooth Logical Link Control and Adaptation Protocol (L2CAP): It enables linking the baseband layer to the upper-layer protocols. This L2CAP can provide (i) connectionless services and (ii) connection-oriented services.

Bluetooth Radio: It specifies the details of wireless support that include (i) frequency (ii) use of frequency hopping (iii) modulation scheme and (iv) transmit power.

Bluetooth Radio Frequency Communication (RFCOMM): It is a cable replacement protocol used in Bluetooth specification. RFCOMM represents a virtual serial port, designed to replace cable technologies. With minimum changes in existing devices, RFCOMM replaces the serial port cables and provides data transport in Bluetooth.

Bluetooth Services Discovery Protocol (SDP): It can supervise and enable connection between two or more Bluetooth devices for a specific service application. SDP includes information such as (i) device information (ii) services and (iii) service characteristics.

Bluetooth Technology: Bluetooth technology enables users to easily connect to a wide range of computing and telecommunications devices, without the need to buy, carry, or connect cables. It provides opportunities for rapid ad hoc connections, and the possibility of automatic, transparent connections between devices. It eliminates the need to purchase additional or proprietary cabling and configure exercises to connect individual devices.

Bluetooth TelCP: Bluetooth also specifies the telephony control protocol. It defines the mobility management functions so that groups of telephony control specification (TCS) devices can be handled. TCS is a bit-oriented protocol that defines call control signalling. It helps to establish speech and data calls between the Bluetooth devices.

BOOTP (Boot Protocol)**:** BOOTP is a network protocol used by a network client to obtain an IP address from a configuration server. BOOTP is usually used by a diskless computer when a computer is starting up. A BOOTP configuration server assigns an IP address to each client from a pool of addresses.

Bridge: It is used for connecting two LANs that may be in two different buildings or on two separate floors within the same building.

Broadcast: In this type of transmission, the message is sent to all the network nodes.

Bus Architectures: In bus-based architectures, nodes are connected to the network cable using T-shaped network interface connectors. Terminating points are placed at each end of the network cable.

Business-to-Business (B2B): Business-to-business (B2B) is a form of commerce in which products or services are sold from a company to dealers.

Business-to-Consumer (B2C): Business-to-Consumer (B2C) is a form of commerce in which products or services are sold from a business firm to a consumer.

Campus Wireless LANs: Several organizations, hotels, retail outlets, warehouses, factories, research centres, and educational institutions, among many others, are recognizing the value of flexibility and connectivity provided by wireless LANs.

Care-of-Address (COA): It is the address that is used to identify the present location of a foreign agent. The packets sent to the MN are delivered to COA.

Centralized Environment: This is the oldest data-processing system in which a large program manages both the tasks of initiating a transaction by interacting with users and processing the transaction. In this environment, the overhead associated with the processing of a transaction is very low as a single user executes the transactions. Recovery is also a trivial issue here, since rollback recoveries do not arise.

Client-Server Environment: In this environment, the execution of a transaction and transaction initiations are done by a server and a client, respectively. The server in this environment is typically concurrent. That is, many clients can be connected to the server at the same time to submit transactions simultaneously.

Code Division Multiple Access (CDMA): CDMA is an access method in which multiple users are allotted different codes of sequences of 0s or 1s to access the same channel. A special coding scheme is used that allows multiple users to be multiplexed over the same physical channel.

Co-located COA: When the mobile node (MN) has acquired a temporary IP address, then that address acts as COA.

Configuration Layer: The configuration layer defines the minimum set of Java Virtual Machine features and Java class libraries available on a particular category of devices. In a way, a configuration defines the commonality of the Java platform features and libraries that developers can assume to be available on all devices belonging to a particular category. This layer is less visible to users, but is very important to profile implementers.

Connected Device Configuration (CDC) for Plug-in Devices: CDC devices use a 32-bit architecture, have at least 2 MB of memory available, and implement a complete functional JVM. CDC devices include digital set-top boxes, home appliances, navigation systems, point-of-sale terminals, and smartphones.

Connected Limited Device Configuration (CLDC) for Hand-held Devices: CLDC is aimed at the low end of the consumer electronics range and designed for 16-bit or 32-bit small computing devices with

limited memory. These devices usually have between 160 KB and 512 KB of available memory.

Consistency Relaxation: Under this approach, the database is logically partitioned into 'clusters', based on some criteria such as certain semantic properties or location proximity. The data in the same cluster must be strictly consistent, whereas a 'bounded' degree of inconsistency is tolerated amongst clusters, according to a relaxed definition of consistency. In this approach, a mobile host can download one cluster before disconnection and execute transactions on this data by preserving the ACID properties and after reconnection it can be copied to the original database if the inconsistencies are below some agreed value.

Consistency: A transaction should transform the database from one consistent state to another consistent state.

Correspondent Node (CN): The home agent is a router on the home network serving as the anchor point for communication with the mobile node. It tunnells packets from a device on the Internet, called a correspondent node, to the roaming mobile node.

Data Replication: Data replication is the process of maintaining a defined set of data in more than one location. It involves copying the designated changes from one location (source or master) to another (target or remote), and synchronizing the data in both the locations.

Data Tier: The data tier consists of database management system and data store.

Data-centric: In many applications of sensor networks, it is not possible to assign a global identifier to each node due to the sheer number of nodes deployed. Lack of such global identifiers along with random deployment of sensor nodes, makes it hard to select a specific set of sensor nodes to be queried. Therefore, data is usually transmitted from every sensor node within the deployment region with significant redundancy.

Denial of Service (DoS): It is possible for a malicious node to continuously transmit and to monopolise the channel use and cause other nodes to wait endlessly.

Destination-Sequenced Distance-Vector Routing (DSDV) Protocol: The Destination-Sequenced Distance-Vector Routing (DSDV) protocol is a table-driven algorithm based on the classical Bellmon–Ford routing algorithm. An improvement made here is the avoidance of routing loops. Each node in an ad hoc network maintains a routing table in which all of the possible destinations within the non-partitioned network and the number of routing hops to each destination are recorded. Hence, routing information is always made readily available, regardless of whether the source node requires a route or not.

DHCP Automatic Allocation: In automatic allocation, DHCP assigns a permanent IP address to a particular client.

DHCP Dynamic Allocation: In dynamic allocation, DHCP assigns an IP address to a client for a specific period of time.

DHCP Manual Allocation: In manual allocation, a client's IP address is assigned by the network administrator, where the DHCP is used to inform the address assigned to clients.

Directed Diffusion (DD): In this protocol, each sensor node names its data with one or more attributes. A destination node sends interests for data, based on these attributes. Interests are flooded over the network. When a node receives an interest from a neighbour, it sends data to the neighbour. Each node only knows the neighbour from whom it got the interest. It is possible that each node would receive the same interest from more than one neighbour. In this way, multiple paths can be set up from the source node to the destination node.

Disconnections: A fairly complicated issue in the operation of a mobile database arises due to the temporary disconnections of mobile users arising out of attempts to save power.

Discovering the Care-of-Address: The discovery of the care-of-address consists of three important steps. (1) Mobile agents advertise their presence by periodically broadcasting through the Agent Advertisement messages. (2) The mobile node receiving the Agent Advertisement message observes whether the message is from its own home agent and determines whether it is on the home network or a foreign network. (3) If a mobile node does not wish to wait for the periodic advertisement, it can send out Agent Solicitation messages that will be responded by a mobility agent.

Distance Vector Protocols: The distance vector algorithm is based on the Bellman–Ford algorithm. Each node advertises its entire routing table to its immediate neighbours only. As a node receives the routing information of its neighbouring nodes, it updates its own routing table by looking at its neighbours' routing table.

Distributed Coordination Function (DCF): It is a distributed function. It can be used with both the infrastructure and the ad hoc (centralized) configurations.

Distributed Environment: In this environment, data is organized in a distributed fashion over a network. That is, either a transaction can get fully executed at any arbitrary node, or some parts of transaction can be executed at a node. Atomicity and durability of a transaction become critical issues in this environment.

DNS (Domain Name Server): Aliases used for IP addresses are called domain names. When we address a website (i.e specify its URL) the domain name is translated to the corresponding IP address by a Domain Name Server.

Dropping Routing Traffic: It is essential in an ad hoc network that all nodes participate in the routing process. However, a node may act selfishly and process only the routing information that are related to itself in order to conserve energy. This behaviour/attack can create network instability or can even segment the network.

Durability: Changes made to the database by a committed transaction must persist even after any failure.

Durability Relaxation: Durability of committed transactions is mainly affected by the possibility of MHs autonomously operating on data. A disconnected MH can only commit a transaction locally if this transaction does not conflict with other transactions executed on the same host while the host is disconnected. On reconnection, the transactions are globally executed on different MHs.

Dynamic Host Configuration Protocol (DHCP): DHCP was developed based on bootstrap protocol (BOOTP). DHCP provides several types of information to the user (client) including its IP address. To manage dynamic configuration information and dynamic IP addresses, IETF standardized an extension to BOOTP known as dynamic host configuration protocol (DHCP).

Dynamic Source Routing (DSR): Dynamic Source Routing (DSR) is a source initiated, on-demand routing protocol for ad hoc networks. It uses source routing, a technique in which the sender of a packet determines the complete sequence of nodes through which a packet has to travel.

Dynamic Topologies: Since nodes are free to move arbitrarily, the network topology may change unpredictably.

Filters: Filters are the key components of all wireless transmitters and receivers. They are used to reject interfering signals lying outside the operating band of receivers and transmitters. They also reject unwanted noise generated by amplifiers.

Foreign Agent COA: The COA is an IP address of a foreign agent (FA).

Foreign Agent: The foreign agent is a router in a foreign network that functions as the point of attachment for a mobile node when it roams to the foreign network. The packets from the home agent are sent to the foreign node which delivers it to the mobile node.

Foreign Network: The foreign network is the current subnet to which the mobile node is visiting. It is different from home network. A foreign network is the network in which a mobile node is operating when away from its home network.

Freeze-TCP: The basic idea in this scheme is to "freeze" the TCP senders' streams a little before a disconnection is to occur.

Frequency Division Multiple Access (FDMA): For systems using Frequency Division Multiple Access (FDMA), the available bandwidth (frequency range) is subdivided into a number of narrower band channels. Each user is allocated a unique frequency band to transmit and receive signals.

FTP (File Transfer Protocol): The FTP protocol is used to transfer files between computers.

General Packet Radio Service (GPRS): GPRS when integrated with GSM, significantly improves and simplifies Internet access. It transfers data packets from GSM mobile stations to external packet data networks (PDNs). Packets can be directly routed from the GPRS mobile stations to packet-switched networks, making it easy to connect to the Internet.

Geographic Adaptive Fidelity (GAF): GAF is an energy-aware location-based routing algorithm, designed primarily for the mobile ad hoc networks but is applicable to sensor networks as well. The network area is first divided into fixed zones forming a virtual grid. Inside each zone, nodes collaborate with each other to play different roles. For example, nodes elect one sensor node to stay awake for a certain period of time while the others go to sleep mode.

Geographic and Energy Aware Routing (GEAR): This protocol is based on the knowledge of node's own as well as its neighbours' locations and energy information. It conserves energy by minimizing the number of interests in DD by only considering a certain region rather than sending the interests to the whole network.

Global System for Mobile Communication (GSM): Possibly, the most successful digital mobile system in current usage is the GSM . The main goal of GSM was to provide a mobile phone system that allows data services and provides voice services compatible to the previous generation systems.

GPRS Architecture: GPRS architecture introduces two new network elements, called the GPRS Support Node (GSN) and the Gateway GPRS Support Node (GGSN).

GPRS Services: GPRS offers end-to-end packet-switched data transfer services which are mainly categorized into two types: (1) Point-To-Point (PTP) service and (2) Point-To-Multipoint (PTM) service.

GPRS: Global Packet Radio System (GPRS) is an extension of GSM and is considered to be 2.5 generation technology. As indicated by the name, it is based on packet switching, compared to circuit switching in the previous generation technologies. An important advantage of GPRS is that it allows users to remain connected to the Internet without incurring additional charge and supports multimedia capabilities including graphics and video viewing.

Great Duck Island (GDI) System: This is an example of Habitat Monitoring Application. The researchers from Intel Research laboratory deployed a mote-based sensor network on Great Duck Island, Maine, USA to monitor the behaviour of storm petrel (small sea birds with tube shaped nostrils).

GSM Anonymity: A GSM network protects against someone tracking the location of the user or identifying calls made to or from the user by eavesdropping on the radio path.

GSM Authentication Centre (AuC): AuC is related to the Home Location Register (HLR). The AuC stores information that is concerned with security features, i.e. user authentication and encryption.

GSM Authentication: The purpose of authentication is to protect the network against any unauthorized use. It also enables the protection of the GSM subscribers by denying the possibility for intruders to impersonate authorized users. The GSM network operator can verify the identity of the subscriber making it infeasible to clone someone else's mobile phone.

GSM Base Station Controller (BSC): The BTSs are managed by the BSC. It reserves radio frequencies, handles the handover from one BTS to another within the BSS. The BSC also multiplexes the radio channels onto the fixed network connections at the interface.

GSM Base Station Subsystem (BSS): A GSM network comprises many BSSs, each BSS contains a BSC and several BTSs. Each BSS is controlled by the Base Station Controller (BSC). The BSS performs all functions necessary to maintain radio connections to an MS, and coding/ decoding of voice.

GSM Base Transceiver Stations (BTS): A BTS comprises all radio equipment such as antenna, signal processing, amplifiers necessary for radio transmission. A BTS can form a radio cell and is connected to an MS via an interface.

GSM Bearer Services: Bearer services can either be connection-oriented or packet-switched. They comprise all services that enable the transparent transmission of data to the network. These services are implemented on the lower-three layers of the OSI reference model.

GSM Confidentiality: A GSM network protects voice, data and sensitive signalling information (e.g. dialed digits) against eavesdropping on the radio path.

GSM Equipment Identity Register (EIR): It contains mobile equipment identity information which helps to block calls from stolen, unauthorized, or defective mobile stations.

GSM Home Location Register (HLR): It stores information that is specific to each subscriber including the user's current location.

GSM Mobile Services Switching Centre: Mobile Services Switching Centres (MSCs) form the fixed backbone network of a GSM system. MSCs set up connections to other MSCs and to the BSCs and also connect to Public Data Network (PDN). MSCs are responsible for connection setup, connection release and handover of connections to other MSCs. They also perform supplementary services such as call forwarding, multiparty calls, etc.

GSM Mobile Station (MS): The mobile stations or cell phones contain two major components: the Subscriber Identity Module (SIM), which is a removable smart card, and the Mobile devices.

GSM Network and Switching Subsystem: This subsystem forms the heart of the GSM system. It connects the wireless networks with the standard public networks and performs and supports usages charging, accounting, and handles roaming of users between different countries. The NSS consists of the switching centre and several types of database.

GSM Operation and Maintenance Centre (OMC): OMC supervises all other network entities. Its functions are traffic monitoring, subscribers management, security management and accounting billing.

GSM Operation Subsystem: The operation subsystem contains all the functions necessary for network operation and maintenance.

GSM Radio Subsystems: This subsystem comprises all the radio specific entities, i.e. the mobile stations, the base station subsystems, the base transceiver station and the base station controller.

GSM Security: Security in GSM is broadly supported at three levels: operator's level, customer's level and system level. These three levels, oversee aspects such as correct billing to the customer, preventing fraud, protecting services, privacy, and anonymity.

GSM System Architecture: A GSM system consist of three main subsystems: Radio Subsystem (RSS), Networking and Switch Subsystem (NSS) and Operation Subsystem (OSS).

GSM Tele Services: GSM provides voice-oriented tele services as well as non-voice tele services. The voice-oriented tele services constitute telephony and the non-voice tele services constitute SMS and FAX.

GSM Visitor Location Register (VLR): It is associated with one or more MSCs and it contains information relating to those subscribers that are currently registered within the MSC area(s) of its associated MSC. VLR is used to store the information of users who are currently in the area of control.

Hierarchical Routing: These protocols are also known as cluster-based routing. In these protocols, different nodes can play different roles in the network based on their remaining battery powers. Higher-energy nodes are used to process and send the information, while the low-energy nodes are used to perform sensing in the proximity of the target.

Home Address: The home address of a mobile device is the IP address assigned to the device within its home network. The IP address on the current network is known as the home address.

Home Agent (HA): It is located in home network and it provides several services for the mobile node (MN). HA maintains a location registry. The location registry keeps track of the node location by the current care-of-address of the MN.

Home Network: The home network of a mobile device is the network within which the device receives its identifying IP address (home address). In other words, a home network is a subnet to which a mobile node belongs to as per its assigned IP address.

HTTP (Hyper Text Transfer Protocol): The HTTP protocol is used for communication between a web server and a web browser.

ICMP (Internet Control Message Protocol): The ICMP protocol is used for error handling in the network.

IEEE MAC Standard: The IEEE 802.11 is the most widely used standard for WLANs today. The IEEE 802.11 standard defines the functional aspects of the medium access control (MAC) sublayer. The nodes in the IEEE 802.11 WLAN can be grouped in one of two configurations: infrastructure (with Access Point) and ad hoc (without Access Point).

IEEE: The IEEE is a non-profit, technical professional association of more than 377,000 individual members in 150 countries. The IEEE acts as a standards body, since in networking, multiple devices that are possibly heterogeneous and manufactured by different vendors need to communicate. The formation of suitable standards is especially important in networking. The IEEE proposes standards for new technologies and maintains the old standards. The IEEE proposed the 802.11 standard for wireless networking. The standard 802.11 is the generic name of a family of standards for wireless networking. The numbering system for 802.11 comes from the IEEE, who uses "802" for many networking standards like Ethernet (802.3).

IGMP (Internet Group Management Protocol): IGMP is a protocol through which hosts exchange information with their local routers. The routers use the IGMP to check whether a known group members are active or not.

Indirect TCP: The Indirect-TCP (I-TCP) protocol suggests that any TCP connection from a mobile host to a machine on the fixed network should be split into two separate connections: one TCP connection between the fixed host and the base station, and the other between the base station and the mobile host.

Integrity of Database: To ensure the integrity of a database, a DBMS should maintain the following constraints:

Intrusion Detection in MANETs: Recognizing and responding to malicious activities, is known as intrusion detection. Intrusion detection is used as a process, which involves technology, people and tools. Intrusion detection is an approach that is complementary with respect to mainstream approaches to security such as access control and cryptography.

IP (Internet Protocol): IP is responsible for delivering data packets across the Internetwork based on their addresses.

IP Address: Each computer must have an IP address before it can meaningfully connect to the Internet. That is, each IP packet must have a destination IP address before it can be sent to another computer.

IP Router: The IP router is responsible for "routing" the packet to its destination, directly or via other routers, based on its IP address.

Isolation Relaxation: Some transaction models have been devised for mobile environments in which the isolation property is not guaranteed. That is, intermediate results of a transaction can be observed by other transactions. This is usually a side effect of the relaxation of other ACID properties.

Isolation: The effect of a transaction should be such that it appears to be executed in isolation, That is, its intermediate results must not be seen by other transactions.

J2ME: The Java2 Micro Edition (J2ME) is meant for tiny devices such as mobile phones, TV set-top boxes, pagers, PDAs, etc.

Java Virtual Machine Layer: This layer is an implementation of a Java Virtual Machine that is customized for a particular device's host operating system and supports a particular J2ME configuration.

Lack of Physical Boundary: Each mobile node functions as a router and forwards packets from other nodes. As a result, network boundaries become blurred. The distinction between nodes, internal and external to a network, becomes meaningless, making it difficult to deploy firewalls, monitor traffic, etc.

LAN Architecture: Two major LAN architectures are being used, the bus architecture and the ring architecture. These two architectures use different access control techniques.

Limited Computational Capabilities: Typically, nodes in ad hoc networks have very limited computational capability. Therefore, handling public-key cryptography during normal operations would be a severe overhead.

Limited Physical Security: Mobile networks are prone to many more types of security threats than those faced by fixed cable networks, mainly due to the wireless transmissions and collaborative routing. There are increased possibilities of eavesdropping, spoofing, denial-

of-service attacks in mobile networks. Also, nodes are vulnerable to capture and compromise.

Limited Power Supply: Since nodes normally rely on battery power, an intruder can exhaust batteries by causing unnecessary transmissions or excessive computations to be carried out by nodes.

Link State Protocols (LSP): In this approach, each node independently calculates the next best logical path from it to every possible destination in the network. This is done periodically or whenever a change is detected in one of the outgoing links.

Location Awareness: A hand-held device equipped with the global positioning system (GPS) can provide information about the current location of a user.

Location-based Routing: In these protocols, the nodes are addressed by their location. Distances to next neighbouring nodes can be estimated by the received signal strengths or by GPS receivers. In this type of protocols, positions of sensor nodes are exploited to route data to the network. The distance between neighbouring nodes can be estimated on the basis of the incoming signal strength. Relative coordinates of neighbouring nodes can be obtained by exchanging such information between neighbours. If there is no activity, nodes go to sleep to save energy. To maximize energy savings, it is necessary to have as many sleeping nodes in the network as possible.

Low Energy Adaptive Clustering Hierarchy (LEACH): Low energy adaptive clustering hierarchy proposed by Heinzelman is a well known routing protocol in WSN based on hierarchical structure. This protocol divides the network into clusters (sections). In each cluster there is an elected sensor node to act as head of the cluster, identified as cluster head. The cluster head's task is to manage communication among member nodes of the cluster, do data processing, and relay the processed sensed data to the base station.

MAC Protocol: When multiple nodes contend to access, i.e. transmit on a shared medium at the same time, the medium access control (MAC) protocol decides which host would be allowed to transmit.

MAC: MAC is a sublayer of link layer and directly invokes the physical layer protocol.

MACA: MACA stands for Multiple Access Collision Avoidance. MACA is designed to solve the hidden/exposed terminal problems mainly by regulating the transmitter power.

Maintaining Logs at the Current MSS: In this approach, logs are maintained at the current MSS for all transactions submitted from its cell. When an MH moves out of the cell, all logs are transferred to the new MSS. This approach generates a high network traffic due to the transfer of logs. Also, it introduces additional complications such as

how much resources should be allotted to a mobile host, particularly to foreign ones, and what to do with the logs if an MH fails.

Maintaining Logs at the Mobile Host: This may be difficult due to lack of sufficient resources such as memory and processing power. As a result, storing the logs at the mobile host and initiating the recovery in case of failure would be an unacceptable overhead. Also, as the failure rate of a mobile host is high, this approach is further complicated and not considered satisfactory.

Maintaining Logs at the Mobile Host's Base MSS: This approach incurs communication overhead because each transaction log must be updated, introducing one round trip delay, which is undesirable.

MANET Hybrid Routing Protocol: Hybrid routing protocols have characteristics of both proactive and reactive protocols. These protocols try to combine the good features of both the protocols.

MANET Proactive Protocols (Table Driven): In a table-driven (proactive) routing protocol, each node maintains routes to every other node. The routing information is usually kept in a number of different tables. These tables are periodically updated if the network topology changes.

MANET Reactive Protocol (On-Demand): Reactive routing is also known as on-demand routing. These routes were designed to reduce the overheads associated with proactive protocols by maintaining information for active routes only. Obviously, they do not have to maintain or constantly update their route tables with the latest route topology.

MANET Unicast Routing Protocols: Unicast routing protocols in MANET are classified into proactive (table-driven), reactive (on-demand) and hybrid types. This classification is based upon how they respond to any change to the network topology.

Mesh-based Protocol: Mesh-based schemes establish a mesh of paths that connect the sources and destinations. They are more resilient to link failures as well as to mobility.

MIME (Multi-purpose Internet Mail Extensions): The MIME protocol lets SMTP encode multimedia files such as voice, picture, and binary data across TCP/IP networks. This helps e-mails to include picture, voice, and binary data files.

Mixers: A mixer is typically used for frequency conversion in transmitters and receivers.

Mobile Commerce: Mobile commerce, essentially involves buying and selling of commodities, services, or information on the Internet using mobile hand-held devices. M-commerce offers consumers significant convenience and flexibility and is playing an increasingly important role in modern times.

Mobile Computing: It is widely described as the ability to compute and communicate while on the move. Mobile computing encompasses two separate and distinct concepts: mobility and computing. Computing denotes the capability to automatically carry out certain processing. Mobility, on the other hand, provides the capability to change location while communicating and computing.

Mobile Equipment (ME): ME is a physical device with unique identifiers called IMEI (International Mobile Identification Number).

Mobile Information Device Profile (MIDP) Layer: This layer is a set of Java APIs that address issues such as user interface, persistence storage, and networking.

Mobile Internet Protocol (Mobile IP): Mobile Internet Protocol (Mobile IP) has been proposed by the Internet Engineering Task Force (IETF). The mobile IP allows mobile computers to stay connected to the Internet regardless of their location and without changing their IP address. In other words, Mobile IP is a standard protocol that builds on the Internet Protocol by making mobility transparent to applications and higher level protocols such as TCP.

Mobile Middleware: The main purpose of mobile middleware is to seamlessly and transparently map the Internet content to mobile phones that may sport a wide variety of operating systems, markup languages, microbrowsers, and protocols. Most mobile middleware also handle encrypting and decrypting communication in order to provide some level of security for transactions.

Mobile Node (MN): A mobile node is a hand-held equipment with roaming capabilities. It can be a cell phone, a personal digital assistant, or a laptop.

Mobile Payment Schemes: Mainly, there are three types of schemes available for M-payment: (a) Bank account based (b) Credit card based and (c) Telecommunication company billing based.

Mobile Payment System: A mobile payment (or M-payment) may be defined as any payment instrument where a mobile device is used to initiate, authorize and confirm an exchange of financial value in return for goods and services.

Mobile TCP: The design of this protocol is based on the observation that, in mobile networks, users will be plagued by frequent disconnection events (caused either by signal fades, lack of bandwidth or by handoff) that must be explicitly handled by the protocol. Unlike I-TCP, M-TCP is another split-connection approach that breaks up a TCP connection between a FH and a MH into two parts: one between the FH and the BS, and the other between the BS and the MH.

Mote: The name mote was given by the scientists from the University of California Berkeley, while working on a project known as Smart

Dust. It was a project funded by the Defense Advanced Research Projects Agency's (DARPA) Network Embedded Software Technology program. The aim of the Smart Dust project was to shrink the size of the sensor nodes (motes) to that of dusts.

Movement of Users: Movement of users introduces several complications as far as database operations are concerned. The complications arising due to user mobility must be satisfactorily addressed.

Multicast: In this type of transmission the message is sent to a selected subset of the network nodes.

Multipath-based Routing: These protocols offer fault tolerance by having at least one alternative path (from source to destination). However this increases the energy consumption and the traffic.

Negotiation-based Routing: It uses negotiation as the basis for cooperation between the competing entities, for some cases of routing between two neighbouring ISPs. Inter-domain routing is often driven by self-interest and based on a limited view of the internet work, which hurts the stability and efficiency of routing.

Network Size and Node Density: Network size refers to the geographical coverage area that could be covered by the network. The number of nodes present per unit geographical area represents network density. Network size and node density are the important parameters that primarily determine the type of routing protocol that would be suitable.

Network Topology: User mobility affects the network topology. Also nodes can become inoperative due to discharged batteries leading to rapid changes to topology.

Operating System (OS): The operating system (OS) is a system software that manages the efficient utilization of the resources of a computer by multiple tasks and provides a user interface communication with base station to the underlying hardware.

Operational Environment: The operational environment of a mobile network refers to a terrain of operations. Common operational environments include urban, rural and maritime. These operational environments support Line of Sight (LOS) communication.

Palm OS: Palm OS is a compact operating system initially developed by the U.S. Robotics' owned Palm Computing, Inc. for personal digital assistants (PDAs) in 1996. Palm OS is designed for ease of use with a touchscreen-based graphical user interface. It is provided with a suite of basic applications for personal information management.

Personalization: Services in mobile environment can be catered according to user's profile. This is required to let the users easily avail information with hand-held devices.

PODS—A Remote Ecological Micro-Sensor Network: PODS was a research project in the University of Hawaii that built a wireless network of environmental sensors to investigate why endangered species of plants grow in one area but not in neighbouring areas. They deployed secret sensor nodes, called Pods, in the Hawaii Volcanos National Park. The Pods consisted of a computer, a radio transceiver and environmental sensors sometimes including a high resolution digital camera, relaying sensor data via the wireless link back to the Internet.

Point Coordination Function (PCF): It is a centralized function used only on infrastructure configurations.

Power-Efficient Gathering in Sensor Information Systems (PEGASIS): In PEGASIS, each node communicates only with a close neighbour. When a node on the chain receives data from a neighbour, it aggregates the data with its own data and sends the data to the next neighbour on the chain. Rather than multiple cluster-heads sending data to the base station as LEACH, only one node on the chain is selected to transmit to the base station.

Presentation Tier: The topmost level of a mobile computing application is the user interface. The main function of the interface is to facilitate the users to issue requests and to present the results to the them meaningfully.

Profile Layer: The profile layer defines the minimum set of application programming interfaces (APIs) available on a particular family of devices. Profiles are implemented upon a particular configuration. Applications are written for a particular profile and are thus portable to any device that supports that profile. A device can support multiple profiles. This is the layer that is most visible to users and application providers.

Query-based Routing: In these protocols, the destination nodes query data through the network. The node(s) containing this required data send it back to the node that has initiated the query along that path.

RARP (Reverse Address Resolution Protocol): The RARP protocol is used by IP to find the IP address based on the physical (MAC address) address of a computer.

Receiver: The receiver receives the modulated signals and reverses the functions of the transmitter component and thus recovers the transmitted base band signal. The antenna of the receiver is usually capable of receiving the electromagnetic waves radiated from many sources over a relatively broad frequency range.

Registering the Care-of-Address: If a mobile node obtains a care-of-address from the distant network (foreign), then it should be registered with the home agent. The mobile node sends a request for registration to its home agent along with the care-of-address information whenever the home agent receives the registration request information.

RFID: Radio Frequency Identification (RFID) tag is attached to a product, an animal, or a person for the purpose of identification and tracking using radio waves. Some tags can be read from several metres away and beyond the line of sight of the reader.

Ring Architecture: A Multi Station Access Unit (MSAU) is a hub or concentrator that connects a group of computers ("nodes" in network terminology) to the ring. The nodes are placed along the ring. The nodes transmit in turn.

Route Optimization: In the mobile IP protocol, all the data packets to the mobile node go through the home agent. Because of this, there will be heavy traffic between the HA and the CN in the network, causing latency to increase. Therefore, route optimization needs to be carried out to overcome this problem.

Routing Loop: By sending the tampered routing packets, an attacker can create a routing loop. This will result in data packets being sent around endlessly, consuming both bandwidth and power for a number of nodes. The packets will not reach their intended recipient and thus can be considered a sort of denial-of-service attack.

Rumor Routing: In rumor routing, each node maintains a neighbour list and an event table. When a node witnesses an event, it adds the event to its event table. Nodes that have recently observed an event generate an agent with a certain probability. An agent is a long-lived packet, roaming through the network and propagating information. Each agent carries a list of events it has encountered, along with the number of hops to that event. When it arrives at a node, it synchronizes its list with the node's list. The agent travels the network for some number of hops, and then dies.

Sequential Assignment Routing (SAR): Sequential Assignment Routing (SAR) is the first protocol for sensor networks that includes the notion of QoS in its routing decisions. This protocol creates trees rooted at one-hop neighbours of the sink by taking QoS metric, energy resource on each path and priority level of each packet into consideration. By using created trees, multiple paths from sink to sensors are formed. One of these paths is selected according to the energy resources and QoS on the path.

Smart Sensors and Integrated Microsystems (SSIM): SIMM is a biomedical project on the artificial retina. In this, a retina prosthesis chip (prosthesis is an artificial replacement of a missing body part) consisting of 100 micro sensors is built and implanted within the human eye. This allows patients with no vision or limited vision to see at an acceptable level. The wireless communication is required to suit the need for feedback control, image identification and validation.

SMTP (Simple Mail Transfer Protocol): The SMTP protocol is used for sending e-mails.

SNMP (Simple Network Management Protocol): The SNMP protocol is used for the administration of computer networks. The network manager uses tools based on this protocol to monitor and assess the network performance.

Snooping TCP: Snooping TCP protocol improves the TCP performance by modifying software at a base station while preserving the end-to-end TCP semantics.

Structure of Mobile Computing Environment: The structure of a mobile computing environment indicates a logical division of the underlying mechanism of the supported functionalities. A simple three-tier structure for mobile computing environment comprises the Presentation tier, the Application tier and the Data tier.

Symbian OS: Symbian OS is a real-time, multitasking, pre-emptive, 32-bit operating system running on ARM processors. Symbian OS from Symbian Ltd. is designed for mobile devices, with associated libraries, user interface frameworks and reference implementations of common tools and runs exclusively on ARM processors.

TCP (Transmission Control Protocol): TCP is a standard transport layer protocol. In mobile computing, TCP is possibly the most popular transport layer protocol. UDP is connectionless and does not guarantee reliable data delivery. But, TCP on the other hand, guarantees reliable data delivery between two applications. TCP needs some special adaptations for use in mobile applications.

TCP (Transmission Control Protocol): TCP is responsible for breaking data down into IP packets before they are sent, and for assembling the packets when they arrive.

TCP Congestion Avoidance: The congestion avoidance algorithm starts where the slow-start stops. From this point, the TCP increases its transmission rate linearly, by adding one additional packet to its window at each transmission time.

TCP Connection-oriented Service: TCP is a connection-oriented protocol. An application requests a "connection" to destination before transferring data. When two processes (A and B) at different sites want to communicate, TCP of A informs the TCP of B and gets approval status from TCP of B. TCP of A and TCP of B exchange data streams in both directions. State information (the state of transmission) is maintained at both ends such as sequence numbers, window size, etc.

TCP Fast-Retransmission: Fast retransmission is based on the observation that TCP encounters unacceptably long pauses in communication during handoffs which cause increased delays and packet losses. This approach addresses the issue of TCP performance when communication resumes after a handoff.

TCP in Mobile Networks: TCP has become the de facto transport protocol in the Internet. It has been remarkably successful in supporting many applications such as web access, file transfer and email.

TCP Port Address: The client program uses a temporary port number and the server program uses a permanent (well-known) port number. These port numbers are used for the identification of the application.

TCP Reliable Service: TCP is a reliable transport protocol. The implication of this is that TCP guarantees that data will be delivered without loss, duplication or transmission errors. For this, TCP uses the following schemes: positive acknowledgement, checksum on both header data, checking sequence numbers to determine the missing segments and detection of duplicate segments.

TCP Slow-Start: The slow-start mechanism is used when a TCP session is started. Instead of starting transmission at a fixed transmission window size, the transmission is started at the lowest window size and then doubled after each successful transmission.

TCP Stream Delivery Service: TCP is a stream-oriented protocol. A stream is essentially any ordered sequence of bytes so that a group of bytes (message) sent in one transmission operation (record) is read exactly as that group at the receiver application.

TCP/IP Application Layer: This layer provides communication between applications running on separate hosts. This layer includes protocols such as http, ftp, telnet, etc.

TCP/IP Data Link Layer: The functions of this layer are related to the logical interfacing between a sub-network and an end system. Its functionalities include encoding schemes and also determining the rate of signalling determined by the physical layer. It also provides error detection and packet framing functionalities. It delivers data packets by making use of lower layer protocols. Ethernet is possibly the most common data link layer protocol.

TCP/IP Internet Layer: It mainly performs data routing and ensures packet delivery at the destination host. It supports communication between two hosts. Remember that there can be several applications or processes running on a host. Once a message reaches a host, it is demultiplexed using the port number at the transport layer for delivery to the appropriate application. In a nutshell, this layer manages addressing of packets and delivery of packets between networks using the IP address.

TCP/IP Transport Layer: It provides reliable end-to-end data transfer services. The endpoints of a communication link are the processes. Thus, to identify the endpoint, not only the computer needs to be identified, but also the port must be identified since the same computer may run several applications. The transport layer provides services

by making use of the services of its lower layer. The details of the underlying networks are hidden from the application layer by the transport layer. Sometimes this layer is also referred to as host-to-host layer. This layer manages the transfer of data by using connection-oriented (TCP) and connectionless (UDP) transport protocols.

TCP/IP: TCP/IP is a large collection of a number of protocols based upon the two important protocols TCP and IP. The Transmission Control Protocol (TCP) and Internet Protocol (IP) are the two important protocols of the TCP/IP protocol suite. The TCP/IP protocol suite was developed by DARPA in 1969.

TCP-F (TCP-Feedback): The TCP-F (TCP-Feedback) protocol proposes that the sender can distinguish between route failure and network congestion. Similar to Freeze-TCP and M-TCP, the sender is forced to stop transmission without reducing the window size upon route failure.

TELNET: It enables the remote log-on facility, by which the user can log-on the systems directly. Both FTP and TELNET pump in data to the TCP layer as it is. TCP forwards the same data on the network through the IP and link layers. As a result, it becomes easy to sniff (secretly hear) the data by using the publicly available TCP sniffer programs such as "TCPdump". Due to this, at present most users use sftp and ssh which essentially serve to encrypt (and decrypt) data before passing on to the TCP layer.

Time Division Multiple Access (TDMA): TDMA is an access method in which multiple users or sources are allotted different time slots to access the same physical channel. The available time slot is divided among multiple sources.

Transaction Model: It defines the framework for the definition and execution of transactions. The integrity constraints on a database are given by well-known ACID properties (Atomicity, Consistency, Isolation, Durability), meaning that a transaction is an atomic, consistent and recoverable unit that does not interfere with other transactions that are executed concurrently. In some application scenarios, such as the mobile environment, ACID properties turn out to be too stringent for realizing any meaningful database in practice and therefore require to be relaxed.

Transaction: It is a unit of operation that changes a database from a consistent state into another consistent state. A database is a consistent state if it satisfies all the defined semantic integrity constraints. A transaction consists of many read and write operations on the database.

Transmitter: The input to a wireless transmitter may be voice, video, data or other types of signals to be transmitted to one or more distant receivers. This signal is called the base band signal. The basic function

of the transmitter is to modulate, or encode, the base band signal onto a high frequency carrier signal.

Tree-based Protocol: Tree-based schemes establish a single path between any two nodes in the multicast group. These schemes require a minimum number of copies per packet to be sent along the branches of the tree.

Tunnelling and Encapsulation: Tunnelling establishes a virtual pipe for the packets available between a tunnel entry and the endpoint. Tunnelling is the process of sending a packet via a tunnel and it is achieved by a mechanism called encapsulation. Encapsulation refers to arranging a packet header and data in the data part of the new packet. On the other hand, disassembling the data part of another packet is called decapsulation. Whenever a packet is sent from a higher protocol layer to a lower protocol layer, the operations of encapsulation and decapsulation take place.

Tunnelling to the Care-of-Address: Tunnelling takes place to forward an IP datagram from the home agent to a care-of-address.

Ubiquity: The dictionary meaning of ubiquity is "present everywhere". Mobile computing allows a user perform computations from anywhere and at anytime.

UMT Core Network: The core network is the equivalent of the GSM Network Switching Subsystem (NSS).

UMT Radio Network Subsystem (RNS): The RNS is the equivalent of the Base Station Subsystem (BSS) in GSM. It provides and manages the wireless interface for the overall network.

UMT User Equipment (UE): The UE is the name by which a cell phone is referred to. The new name was chosen because of the considerably greater functionality that the UE has compared to a cell phone. It can be thought of as anything between a mobile phone used for talking and a data terminal attached to a computer with no voice capability.

Unicast: In this, the message is sent to a single destination node.

Universal Mobile Telephone Standard (UMTS): UMTS is enlisted as the third generation technology (3G) with the voice data and non-voice data (exchanging email, instant messaging and video telephony) capabilities.

Vehicular Ad Hoc Network (VANET): It uses the moving automobiles as the nodes in a ad hoc network.

Virtual Home Environment (VHE): A user roaming from his network to other UMTS operators, will experience a consistent set of services, thus "feeling" to be on his home network, independent of the location. In contrast, in a 2G network, a user is registered to a visitor location and is also charged a roaming overhead.

WAP 2.0: WAP 2.0 protocol focuses on the schemes combining wireless applications with the Internet. The WAP 2.0 provides support for the protocols such as IP, TCP and HTTP. It is also flexible with air interface technologies and their corresponding bearers.

Windows CE: Windows CE is a 32-bit multitasking operating system. Usually, this operating system is delivered on a Read Only Memory.

Wireless Application Environment (WAE): The application layer includes the micro-browser on the device, WML (the Wireless Markup Language), WMLS (a client-side scripting language), telephony service, and a set of formats for the commonly used data (such as images, phone books, and calendars).

Wireless LAN Cards: End-users access the WLAN through WLAN adapters (wireless network interface cards) in their hand-held devices. Nowadays, this card is in-built into the motherboards.

Wireless Local Area Networks (WLANs): WLANs provide connectivity between computers over short distances via a wireless medium. Typical indoor applications of WLANs may be in educational institutes, office buildings and factories where coverage distances are usually less than a few hundred feet.

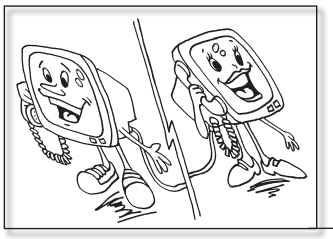
Wireless Sensor Networks: Wireless sensor networks (WSNs) are network systems containing sensor nodes. The sensor nodes can sense certain physical characteristics and can be used to capture environmental information such as temperature, sound, motion, etc.

Wireless Session Protocol (WSP): The session layer provides HTTP functionality, with basic session state management, and a facility for reliable and unreliable data push and pull.

Wireless Transaction Protocol (WTP): The transaction layer provides transport services (one-way and two-way) and the related technologies.

Wireless Transport Layer Security (WTLS): The security layer provides data security and privacy, authentication, as well as protection against denial-of-service attacks.

Zone Routing Protocol: In ZRP, a routing zone comprises a few mobile ad hoc nodes within a few hops from the central node. Within this zone, a table-driven routing protocol is used. This implies that route updates are performed for nodes within the zone. Each node, therefore, has a route to all other nodes within the zone. If the destination node resides outside the source zone, ZRP employs a route discovery procedure.



Index

Fundamentals of Mobile Computing

Prasant Kumar Pattnaik • Rajib Mall

This textbook addresses the main topics associated with mobile computing and wireless networking at a level that enables the students to develop a fundamental understanding of the technical issues involved in this new and fast emerging discipline.

The book first examines the basics of wireless technologies and computer communications that form the essential infrastructure required for building knowledge in the area of mobile computations involving the study of invocation mechanisms at the client end, the underlying wireless communication, and the corresponding server-side technologies.

The book includes coverage of development of mobile cellular systems, protocol design for mobile networks, special issues involved in the mobility management of cellular system users, realization and applications of mobile ad hoc networks (MANETs), design and operation of sensor networks, special constraints and requirements of mobile operating systems, and development of mobile computing applications.

Finally, an example application of the mobile computing infrastructure to M-commerce is described in the concluding chapter of the book.

This book is suitable as an introductory text for a one-semester course in mobile computing for the undergraduate students of Computer Science and Engineering, Information Technology, Electronics and Communication Engineering, Master of Computer Applications (MCA), and the undergraduate and postgraduate science courses in computer science and Information Technology.

KEY FEATURES

- Provides unified coverage of mobile computing and communication aspects
- Discusses the mobile application development, mobile operating systems and mobile databases as part of the material devoted to mobile computing
- Incorporates a survey of mobile operating systems and the latest developments such as the Android operating system

THE AUTHORS

PRASANT KUMAR PATTNAIK, PhD (Computer Science), is Associate Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. He has more than a decade of teaching and research experience. He is a Senior Member of the International Association of Computer Science and Information Technology (IACSIT), Singapore. His areas of interest include mobile computing and cloud computing.

RAJIB MALL, PhD, is Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology Kharagpur. He has vast practical experience in developing industry-oriented software products. Having an academic experience of a decade and half in IIT Kharagpur, Professor Rajib Mall has guided several doctoral dissertations and published over a hundred research articles.

ISBN:978-81-203-4632-1

9 788120 346321

www.phindia.com