1

Introduction

This book focuses on the deployment of multicast in third-generation networks. Multicast is the efficient delivery of data to a group of destinations simultaneously. With multicast, messages are delivered as much as possible only once over each link of the network, creating copies only when the links to the destinations split.

In this chapter, we firstly provide an introduction to cellular mobile communication systems, in particular with respect to the features that distinguish the different generations of mobile communication systems, from analog first-generation to the fourth-generation systems currently in development. Then, we describe several fundamental aspects of data networking that are relevant for multicast. This is followed by an overview of how multicast can be achieved in data networks. We then introduce the basics of Internet Protocol (IP) multicast, the standard for multicast in internetworks. A more detailed description of IP multicast is provided in Chapter 2. Finally, we describe several existing mechanisms for carrying out multicast in third-generation networks. Several of these multicast mechanisms are described in much more detail in later chapters.

COPYRIGHTED MATERIAL

1.1 Cellular Mobile Communication Systems

The mobile communications industry is a relatively young industry. The basic technological concept of the industry lies in using radio waves to transmit data and connect users. Radio is the transmission of signals by modulation of electromagnetic waves with frequencies below those of visible light. Electromagnetic radiation travels by means of oscillating electromagnetic fields that pass through the air and the vacuum of space. Information is carried by systematically changing or modulating some property of the radiated waves, such as amplitude, frequency or phase. When radio waves pass an electrical conductor, the oscillating fields induce an alternating current in the conductor. This can be detected and transformed into sound or other signals that carry information.

The concept of using radio waves for communication dates back to the second half of the nineteenth century, when the German scientist Heinrich Rudolf Hertz demonstrated in 1888 that an electric spark of sufficient intensity at the emitting end could be captured by an appropriately designed receiver and induce action at a distance. This proved for the first time that electromagnetic waves propagate through the air and have the same properties as light. His English forerunner James Clark Maxwell had foreseen this a few

Multicast in Third-Generation Mobile Networks Robert Rümmler, Alexander Gluhak and A. Hamid Aghvami
© 2009 John Wiley & Sons, Ltd

years earlier in 1864. Maxwell's theory of electromagnetic fields claimed the existence of electromagnetic waves and presented four mathematical formulae known today as Maxwell's equations, a set of fundamental equations governing electromagnetism.

Nikola Tesla first demonstrated the feasibility of wireless communications in 1893. He holds the US patent for the invention of the radio, defined as the *wireless transmission of data*. Guglielmo Marconi demonstrated the use of radio for wireless communications by equipping ships with life-saving wireless communications and by establishing the first commercial transatlantic radio service in 1907. Today, the use of radio takes many forms, including wireless and mobile communication of all types, as well as radio broadcasting.

1.1.1 The Cellular Concept

The design objective of early mobile communication systems was to achieve a large coverage area by using a single, high-power transmitter with an antenna mounted on a tall tower, transmitting on a single frequency. While this approach achieved very good coverage, it also meant that it was impossible to reuse the same frequency throughout the system, since any attempts to achieve frequency reuse would result in interference.

The cellular concept was a major breakthrough in solving the problem of spectral congestion and user capacity. It offered very high capacity with limited spectrum without any major technological changes. The cellular concept is a system-level idea that calls for replacing a single, high-power transmitter with many low-power transmitters, each providing coverage to only a small portion of the service area, referred to as a cell. Each base station is allocated a portion of the total number of channels available to the entire system, and nearby base stations are assigned different groups of channels so that all the available channels are assigned to a relatively small number of neighbouring base stations.

The mobile transceivers (also referred to as mobile phones, mobile stations, mobile terminals, handsets or devices) exchange radio signals with any number of base stations. Mobile phones are not attached to a particular base station, but may make use of any one of the base stations provided by the company that operates the corresponding network. The ensemble of base stations covers the landscape in such a way that the user can travel around and carry on a phone call without interruption, possibly making use of more than one base station. The procedure of changing a base station at cell boundaries is called *handover*.

Communication from the Mobile Station (MS) to the Base Station (BS) takes place on the uplink channel or reverse link, and from BS to MS on the downlink channel or forward link. To sustain a bidirectional commmunication between a mobile terminal and a base station, transmission resources must be provided both in the uplink and downlink directions. This can happen either through Frequency-Division Duplex (FDD), whereby uplink and downlink channels are assigned on separate frequencies, or through Time-Division Duplex (TDD), where uplink and downlink transmissions occur on the same frequency, but alternate in time.

FDD is efficient in the case of symmetric traffic. Also, FDD makes radio planning easier and more efficient, since base stations do not interfere with each other as they transmit and receive in different sub-bands. TDD has a strong advantage in the case where the asymmetry of the uplink and downlink data speed is variable. As the amount of uplink data increases, more bandwidth can dynamically be allocated to that, and as it shrinks it

can be taken away. Another advantage is that the uplink and downlink radio paths are likely to be very similar in the case of a slow-moving system.

1.1.2 Propagation Impairments in Cellular Systems

The design of cellular systems is particularly challenging because of the adverse propagation conditions of the radio channel. Three main propagation impairments are usually distinguished. These are pathloss, slow fading or shadowing and fast fading or multipath fading (Brand and Aghvami, 2002).

The pathloss describes the average signal attenuation as a function of the distance between transmitter and receiver, which includes the free-space attenuation as one component, but also other factors come into play in cellular communications, resulting in an environment-dependent pathloss behaviour. Shadowing or slow fading describes slow signal fluctuations, which are typically caused by large structures, such as big buildings, obstructing the propagation paths. Fast or multipath fading is caused by the fact that signals propagate from transmitter to receiver through multiple paths, which can add at the receiver constructively or destructively, depending on the relative signal phases. The received signal is said to be in a deep fade when the paths add destructively such that the received signal level is close to zero. Fades occur roughly once every half-wavelength (Steele and Hanzo, 1999). With wavelengths of 30 cm and less in cellular communication systems, it is clear that multipath fading can result in relatively fast signal fluctuations (Brand and Aghvami, 2002).

1.1.3 Multiple-Access Schemes

Multiple-access schemes allow several devices connected to the same physical medium to transmit over it and to share its capacity. A multiple-access scheme is based on a multiplex method that allows several data streams or signals to share the same communication channel or physical media. Multiplexing is a term used to refer to a process where multiple data streams are combined into one signal over a shared medium. The resources that may be allocated with a multiple-access scheme are frequency bands, time slots, sets of codes or any combination of the three. The basic multiple-access schemes are Frequency-Division Multiple Access (FDMA), Time-Division Multiple Access (TDMA) and Code-Division Multiple Access (CDMA).

In FDMA, each communication is carried over one or two (depending on the duplexing method) narrowband frequency channels. The channel bandwidth and the modulation scheme determine the gross bit rate that can be sustained. With non-ideal filters, guard bands must be introduced between the FDMA channels to avoid so-called adjacent channel interference.

In TDMA, rather than assigning each user a channel with its own frequency, users share a channel of a wider bandwith in the time domain. This is achieved by introducing a framing structure, with each TDMA frame subdivided into a number of slots equal to the number of users that are to be supported. Provided that enough spectrum is available, multiple carriers may be assigned to each cell. Therefore, such TDMA systems typically feature also an FDMA element and are thus in reality hybrid TDMA/FDMA systems (Brand and Aghvami, 2002). TDMA systems must carefully synchronize the transmission

times of all the users to ensure that they are received in the correct slot and do not cause interference.

In CDMA, narrowband signals are transformed through spectrum spreading into signals with a wider bandwidth. As in TDMA, multiple users share the carrier bandwidth, but, as in FDMA, they transmit continuously during the connection. The multiple-access capability derives from the use of different spreading codes for individual users. Because of the spreading of the spectrum, CDMA systems are also referred to as spread-spectrum multiple-access systems. Two basic CDMA techniques suitable for mobile communications are distinguished, namely Frequency-Hopping (FH) and Direct-Sequence (DS) CDMA techniques.

In an FH-CDMA system, a transmitter *hops* between available frequencies according to a specified algorithm, which can be either random or preplanned. The transmitter operates in synchronization with a receiver, which remains tuned to the same centre frequency as the transmitter. A short burst of data is transmitted on a narrowband carrier. Then, the transmitter tunes to another frequency and transmits again. Thus, the receiver is capable of hopping its frequency over a given bandwidth several times a second, transmitting on one frequency for a certain period of time, then hopping to another frequency and transmitting again. Frequency hopping requires a much wider bandwidth than is needed to transmit the same information using only one carrier frequency.

In a DS-CDMA system, a bit stream is multiplied by a direct sequence or spreading code composed of individual chips. They have a much shorter duration than the bits of the user bit stream, and this is why the original signal's spectrum is spread. The bandwidth expansion factor or spreading factor that results from using a transmission bandwidth that is several orders of magnitude greater than the minimum required signal bandwidth is equal to the duration of a bit divided by the duration of a chip.

1.1.4 First- and Second-Generation Systems

Various first-generation cellular mobile communication systems were introduced in the late 1970s and early 1980s. These early systems were characterized by analog (frequency modulation) voice transmission and limited flexibility. The first such system, the Advanced Mobile Phone System (AMPS), was introduced in the US in the late 1970s. Other first-generation systems include the Nordic Mobile Telephone (NMT) and the Total Access Communication System (TACS). The former was introduced in 1981 in Sweden, then soon afterwards in other Scandinavian countries, followed by the Netherlands, Switzerland and a large number of Central and Eastern European countries. The latter was deployed from 1985 in Ireland, Italy, Spain and the UK (Brand and Aghvami, 2002).

While these systems offered reasonably good voice quality, they provided limited spectral efficiency. They also suffered from the fact that network control messages – for handover or power control, for example – are carried over the voice channel in such a way that they interrupt speech transmission and produce audible clicks, which limits the network control capacity (Goodman, 1990).

The breakthrough of mobile telephony into the mass market occurred only in the 1990s with the advent of digital technology and the introduction of second-generation systems. Capacity increase was one of the main motivations for introducing second-generation systems. With digital technology it became possible to increase capacity by relying on low-bit-rate speech codecs and also integrating voice and data. Also, security was

improved, both by means of encryption to provide privacy and authentication to prevent unauthorized access and use of the system. Dedicated channels were used for the exchange of network control information between mobile terminals and the network infrastructure during a call in order to overcome the limitation in network control of first-generation systems.

The Global System for Mobile Communications (GSM) is currently the uncontested standard for second-generation digital cellular communications. GSM, a TDMA-based system with optional slow frequency hopping, has a footprint covering virtually every angle of the world. With a subscriber number close to 500 million and a share of the digital cellular market close to 70 % in early 2001 (Brand and Aghvami, 2002), GSM is truly *the* global system for mobile communications. The General Packet Radio Service (GPRS) is a best-effort packet-switched service, which was designed as an enhancement to existing GSM networks in order to support non-real-time packet data traffic.

In the US, there are essentially two types of second-generation cellular system that are incompatible with each other. The first is a TDMA system called North American Digital Cellular or Digital AMPS (D-AMPS) and referred to as TDMA. The second system, which was launched later, is cdmaOne, the first operational CDMA system (Goodman, 1990). The relevant air interface specifications are the so-called interim standards IS-136 (for D-AMPS) and IS-95 (for cdmaOne).

The first and most popular Japanese second-generation standard is Personal Digital Cellular (PDC). It was later complemented by the Personal Handphone System (PHS), a mixture between mobile and cordless systems, which caters for low mobility, but is popular for certain applications owing to its relatively high data rates of up to 64 kilobits per second (kbps). Both standards are TDMA-based and have not seen wide deployment outside Japan (Brand and Aghvami, 2002).

1.1.5 Third-Generation Systems

Third-generation mobile networks represent the latest phase in the evolution of cellular technology, following from the first-generation analog and second-generation digital systems. Third-generation systems represent a shift from voice-centric services to converged services, including voice, data, video and so forth. In order to allow for advanced services and applications, third-generation networks provide higher capacity and enhanced network functionality.

Already before the launch of second-generation systems, the research community started to think about requirements for a new, third generation of mobile communication systems and about possible technological solutions to meet them. The European Telecommunication Standards Institute (ETSI) was one of the major players regarding the standardization of third-generation systems. It called its third-generation representative Universal Mobile Telecommunications System (UMTS) and established a number of requirements, according to which such a system should be designed.

The International Telecommunications Union (ITU) initially had the intention of controlling the standardization process such that a single system would emerge. With several bodies submitting their proposals for third-generation systems to the ITU in 1998, it soon became clear that the ITU would not be in a position to enforce a unified system. As a result of this, the ITU then advocated the concept of a *family of systems*, defined as a federation of systems referred to as International Mobile Telecommunications

2000 (IMT-2000). Eventually, two main camps formed. The first one is united in the Third-Generation Partnership Project (3GPP), dealing with the standardization of UMTS and the evolution of GSM, and the second one in a similar structure, the Third-Generation Partnership Project 2 (3GPP2), dealing with CDMA2000, an evolution of cdmaOne. The following sections provide a brief overview of third-generation systems.

Third-Generation Operating Modes

The ITU defines a third-generation network as one that delivers, among other capabilities, improved system capacity and spectrum efficiency compared with second-generation systems. The ITU's definition of third-generation technology stipulates, among other things, that these must be capable of supporting data transmission speeds of at least 144 kbps outdoors and 2 Mbps in fixed (indoor) environments (Zanoio and Urvik, 2003).

The IMT-2000 standard of the ITU consists of five operating modes, including three based on CDMA technology. The third-generation modes based on CDMA are most commonly known as Wideband CDMA (WCDMA), CDMA2000 and Time-Division Synchronous CDMA (TD-SCDMA), all of which are single-carrier DS-CDMA air interfaces.

WCDMA is the air interface used in UMTS networks. WCDMA is a wideband DS-CDMA air interface that achieves higher speeds and supports more users compared with the implementation of TDMA used in GSM networks. WCDMA, which operates with a carrier spacing of 5 MHz, can theoretically offer data transmission speeds of up to 2 Mbps.

CDMA2000 is a direct successor to IS-95 or cdmaOne. The CDMA2000 standards are CDMA2000 1xRTT, CDMA2000 Evolution Data Only (EV-DO) and CDMA2000 Evolution Data Voice (EV-DV). The CDMA2000 standard has evolved continually to support new services in a standard 1.25 MHz carrier. CDMA2000 1xRTT, with RTT standing for Radio Transmission Technology, provides data transmission capability with peak data at around 144 kbps. The data rate and cell capacity can be increased using the EV-DO and EV-DV. The data-only system CDMA2000 EV-DO offers theoretical speeds of up to 2.4 Mbps. EV-DV offers both data and voice, but has yet to be deployed anywhere.

TD-SCDMA is the mobile telecommunications standard being pursued in China. TD-SCDMA was incorporated by the 3GPP as part of UMTS Release 4. TD-SCDMA operates with a channel bandwidth of 1.6 MHz and offers theoretical data rates of up to 2 Mbps. TD-SCDMA uses TDD, in contrast to the FDD scheme used by WCDMA. Using the same carrier frequency for uplink and downlink means that the channel condition is the same on both directions, and the base station can deduce the downlink channel information from uplink channel estimates, which is helpful to the application of beamforming techniques. TD-SCDMA is being promoted as China's own third-generation solution, with only a limited number of vendors to offer TD-SCDMA technology. Commercialization of TD-SCDMA lags behind WCDMA by roughly 2 years, with TD-SCDMA networks having yet to be deployed (Tuttlebee and Payne, 2004/2005).

The other two options that are part of ITU's third-generation definition are Enhanced Data Rates for GSM Evolution (EDGE), a digital mobile phone technology that allows increased data transmission rates and improved data transmission reliability compared

with GSM, and Digital Enhanced Cordless Telecommunications (DECT), a standard for digital portable phones, commonly used for domestic or corporate purposes.

CDMA2000 and EDGE are considered as *evolutionary* standards in that they can operate within existing spectrum allocations. WCDMA, TD-SCDMA and DECT, on the other hand, are *revolutionary* standards in that they require new spectrum allocation.

Deployment of Third-Generation Systems

In many European countries, a limited number of licences for radio spectrum required to operate WCDMA was allocated in a sealed bid auction. The license fees for spectrum were very high in some European countries, especially in the UK and Germany, in part due to the initial excitement over the potential of third-generation networks. Roll-out of third-generation networks was delayed in these countries because of this. Additional delays also resulted from the high cost of upgrading equipment for the new systems.

In Europe, the first commercial third-generation services were introduced in the UK and Italy, starting in March 2003. As of June 2007, 200 million third-generation network subscribers had been connected (Wikipedia, 2008a). This accounts for roughly 6.7 % of the 3 billion mobile phone subscriptions worldwide. In the countries where third-generation networks were launched first, namely Japan and South Korea, over half of all subscribers use third-generation networks.

As of December 2007, 190 third-generation networks were operating in 40 countries (The Global Mobile Suppliers Association, 2008). In Europe, the leading country is Italy, with a third of its subscribers migrated to third-generation networks. Other leading countries include the UK, Austria, Australia and Singapore at the 20 % migration level.

1.1.6 Towards Fourth-Generation Systems

It is predicted that, between 2012 and 2015, the current and future evolutions of third-generation networks will not have sufficient capacity to meet traffic demands. Regulatory and standardization bodies are therefore working towards commercial deployment of fourth-generation networks in that timeframe. Fourth-generation systems will be able to provide a comprehensive IP solution based on IPv6, where voice, data and streamed multimedia can be given to users on an *anytime, anywhere* basis, and at higher data rates than previous generations.

It will clearly be difficult to define the dividing line between third-generation and fourth-generation technology. A number of so-called fourth-generation technologies such as Long-Term Evolution (LTE) from 3GPP and Ultra Mobile Broadband (UMB) from 3GPP2 are in fact actually evolutions of third-generation technologies.

3GPP LTE is the name given to a project within the 3GPP to improve the UMTS standard to cope with future technology evolutions. Goals include improving spectral efficiency, lowering costs, improving services, making use of new spectrum and refarmed spectrum opportunities, and better integration with other open standards. The LTE project is not a standard, but it will result in the new evolved Release 8 of the 3GPP specifications, including mostly or wholly extensions and modifications of the UMTS system. UMB is the next-generation 3GPP2 standard air interface. UMB will use Orthogonal Frequency-Division Multiplexing (OFDM) rather than CDMA.

One of the drivers for the popular use of fourth-generation technology has been the aggressive promotion within the industry of the Institute of Electrical and Electronics Engineers (IEEE) 802.16e or WiMax mobile standard. A version of this standard was, however, recently accepted by the ITU as an addition to the IMT-2000 family, and therefore is clearly to be considered together with the other third-generation IMT-2000 technologies.

1.2 Networks and Protocols

A communications network is an interconnected group of computers or devices that have the ability to exchange data. The basic function of any network is to deliver data from a source to one or more receivers. Examples of networks are Local-Area Networks (LANs), which are constrained to small geographic areas, and Wide-Area Networks (WANs), which cover larger geographic areas than LANs. An immense number of technologies have been developed to accomplish this task, including a wide variety of physical media for transmitting data, such as twisted-pair copper wire cable or optical fibre, different ways for organizing network devices, as well as many protocols to govern how everything fits together to form a functioning whole (Kosiur, 1998).

Protocols are the rules that determine how a network operates. Protocols have two important roles. Firstly, they describe the syntax and semantics of messages exchanged within the network, that is, in what format messages are transmitted and what meaning these messages have. Secondly, protocols also define the actions that should be taken upon receipt of a message (Kosiur, 1998).

The Open Systems Interconnection (OSI) reference model is a layered, abstract description for communications and computer network protocol design. It was developed as part of the OSI initiative and is sometimes known as the OSI seven-layer model. As depicted in Figure 1.1, the OSI reference model consists of the application, presentation, session, transport, network, data link and physical layers. A layer is a collection of related functions that provides services to the layer above it and receives service from the layer below it. For example, a layer that provides error-free communications across a network provides the path needed by applications above it, while it calls the next lower layer to send and receive packets.

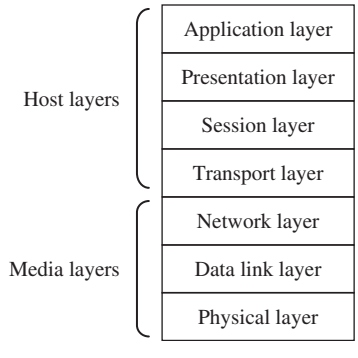


Figure 1.1 Layers of the OSI reference model

The layers in the OSI reference model are briefly described in the following (Zimmermann, 1980):

1. The **application layer** performs application services for the application processes and issues requests to the presentation layer. The application layer provides services to user-defined application processes, and not to the end-user itself.
2. The **presentation layer** establishes a context between application-layer entities, in which the higher-layer entities can use different syntax and semantics, as long as the presentation layer understands both and the mapping between them.
3. The **session layer** controls the dialogues/connections (sessions) between computers. It establishes, manages and terminates the connections between the local and remote application.
4. The **transport layer** provides transparent transfer of data between end-users, providing reliable data transfer services to the upper layers. The transport layer controls the reliability of a given link through flow control, segmentation/desegmentation and error control. Some protocols are state and connection oriented. This means that the transport layer can keep track of the segments and retransmit those that fail.
5. The **network layer** provides the functional and procedural means of transferring variable-length data sequences from a source to a destination via one or more networks. The network layer performs network routing functions, and might also perform fragmentation and reassembly, and report delivery errors.
6. The **data link layer** provides the functional and procedural means to transfer data between network entities and to detect and possibly correct errors that may occur in the physical layer.
7. The **physical layer** defines all the electrical and physical specifications for devices. In particular, it defines the relationship between a device and a physical medium.

1.2.1 Circuit-Switched and Packet-Switched Networks

Networks can be categorized by how data is routed and under what conditions data is accepted by individual network devices. In circuit-switched networks, a physical path is obtained for and dedicated to a single connection between two endpoints in the network for the duration of the connection. In packet-switched networks, small units of data called packets are routed through the network based on the destination address contained within each packet. Circuit-switched networks are often called connection-oriented networks, while packet-switched networks are referred to as connectionless networks.

With circuit switching, the source and destination must establish a connection to exchange data. The advantage of a circuit-switched network is its guaranteed capacity. Once a circuit is established, no other network activity can decrease the circuit's capacity. The main disadvantage of circuit switching is the fixed cost of the circuit, which is independent of the amount of traffic that flows on it. Some examples of circuit-switched networks are the Public Switched Telephone Network (PSTN) and the Integrated Services Digital Network (ISDN).

With packet switching, data between two end-systems is not sent as a continuous stream of bits. Instead, it is divided into small units called packets that are sent one at a time. These packets are multiplexed or allocated different time slots for transmission. More than one source can inject packets over the same wires into packet networks, so that, when

a source is not transmitting, network resources are available for use by other sources. To allow the network to sort out these multiple flows of data, each packet carries an identifier of its destination. Thus, logical paths instead of physical circuits exist between communicating end-systems. The main advantage of packet switching is that multiple communications among end-systems can occur concurrently. The disadvantage is that, as network activity increases, a given pair of communicating end-systems receives less of the network capacity. Some examples of packet-switched networks are Asychronous Transfer Mode (ATM) and X.25 networks.

1.2.2 Internet Protocol Suite

The inability of a single type of network to satisfy all communication requirements necessitates connecting different networks together to create internetworks. The largest such internetwork is the Internet, a worldwide, publicly accessible collection of interconnected computer networks that transmit data by packet switching using standard IP. It is a *network of networks* that consists of millions of smaller domestic, academic, business and government networks. The IP suite is a set of communication protocols that implement the protocol stack on which the Internet and most commercial networks run. The IP protocol suite features various other protocols on top of IP, for example transport protocols such as the Transmission Control Protocol (TCP) and the User Datagram Protocol (UDP). The IP protocol suite is often referred to as TCP/IP, as these are the two most important protocols in the suite. TCP/IP is generally described as having four abstraction layers. The IP suite uses encapsulation to provide abstraction of protocols and services. Generally, a protocol at a higher level uses a protocol at a lower level to help accomplish its aims. The TCP/IP model and related protocols are currently maintained by the Internet Engineering Task Force (IETF).

From lowest to highest, the layers of the TCP/IP protocol suite are the network access layer, the network or internetwork layer, the transport layer and the application layer. The four layers are depicted in Figure 1.2. While the TCP/IP protocols do not fit neatly into all seven layers of the OSI reference model, they provide all the necessary functionality for productive networking.

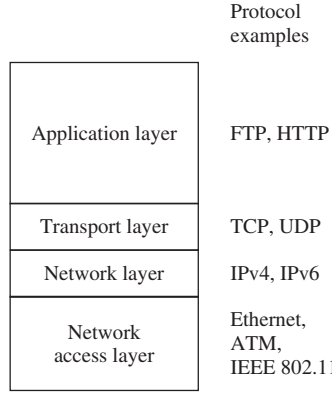


Figure 1.2 Layers of the TCP/IP protocol suite

The layers in the TCP/IP protocol suite perform the following functions (Kosiur, 1998):

1. The **application layer** organizes the data that the network will transfer. Examples of application protocols are the File Transfer Protocol (FTP) used to transfer data from one computer to another through a network and the Hypertext Transfer Protocol (HTTP) used to transfer information on the Intranet and the World Wide Web (WWW).
2. The **transport layer** is responsible for delivering the applications' information to the destination. Examples of transport protocols are TCP and UDP. TCP is a protocol that provides reliable, in-order delivery of a stream of bytes, making it suitable for applications such as file transfer and email. UDP, on the other hand, is a transport protocol that does not guarantee reliability or ordering in the way that TCP does. Packets may arrive out of order, appear duplicated or go missing without notice. Avoiding the overhead of checking whether every packet has actually arrived makes UDP faster and more efficient for applications that do not need guaranteed delivery.
3. The **network layer** or internetwork layer bears the responsibility of understanding the topology of the network and forwarding information through the network to its destination. IP version 4 (IPv4) and IP version 6 (IPv6) are the only standard internetwork-layer protocols used on the Internet. In order to forward data to the appropriate destination, network protocols rely on an addressing scheme for the hosts that are to receive the data. IPv4 is the first version of the protocol to be widely deployed. IPv6 is designated as the successor of IPv4. One of the main changes brought by IPv6 compared with IPv4 is a much larger IP address space that allows for greater flexibility in assigning addresses.
4. The **network access layer** combines the data link and physical layer of the network technology. Examples of network access technologies in widespread use are Ethernet, ATM and IEEE 802.11, the set of standards for Wireless LAN (WLAN) computer communication, developed by the IEEE in the 5 GHz and 2.4 GHz public spectrum bands.

1.2.3 Routing in Internetworks

Routers and switches are decision points in a network – a decision is made as to where in the network packets should be forwarded. The difference between the two types of device is what part of a packet is scanned to make a forwarding decision. In order to make a forwarding decision, switches look at Medium Access Control (MAC) addresses in layer 2 of the OSI reference model protocol stack, whereas routers review the network address found in layer 3 of the protocol stack, the network layer.

A routing protocol helps routers to create a topological map of the network. Routing protocols are part of the internetwork layer of the TCP/IP protocol stack. Routing protocols are different in how they share information and compute routes. With link-state routing protocols, every node constructs a map of the connectivity of the network in the form of a graph showing which nodes are connected to which other nodes. Routers share only the identity of their neighbours, but they flood the entire network with this information. Distance-vector routing protocols periodically share their knowledge of the entire network, but only with their neighbours. As neighbouring routers learn new information, they pass that information on to their neighbours until, slowly but surely, the information makes

its way across the entire network. With both types of routing protocol, routing tables are constructed, which are held in the router's memory. The routing process forwards packets on the basis of routing tables which maintain a record of the routes to various network destinations.

More specifically, link-state routing protocols create the map in three phases: each router meets its neighbours (learn your neighbourhood), routers then share that information with other routers on the network (learn about other neighbourhoods) and, finally, routers combine the information and calculate routes (Kosiur, 1998). Link-state protocols flood the entire network with link-state information, which can lead to unnecessary bandwith usage. This disadvantage is offset by the speed and efficiency of link-state routing. The best known link-state routing protocol is the Open Shortest Path First (OSPF) routing protocol.

1.3 Multipoint Communications

Point-to-multipoint or more simply multipoint communications is a term used in the telecommunications field to refer to communication that is accomplished via a multipoint connection, providing a communication link between a single source and multiple receivers.

Many applications involve one-to-many or many-to-many communications, where one or more sources are sending data to multiple receivers. Such multipoint transmissions can be achieved in three different ways: unicast where a separate copy of the data is delivered to each recipient, broadcast where a data packet is forwarded to all portions of the network and multicast where a single packet is addressed to all intended recipients and the network replicates packets only as needed.

1.3.1 Unicast

In networking, the fundamental method of communication is between two hosts, or unicasting. Such one-to-one sessions offer a great deal of control of the data traffic between the source and receiver, allowing for acknowledgement of receipt, requests for retransmission of data, changes in transmission rate and so on. Figure 1.3 illustrates the point-to-point nature of unicast.

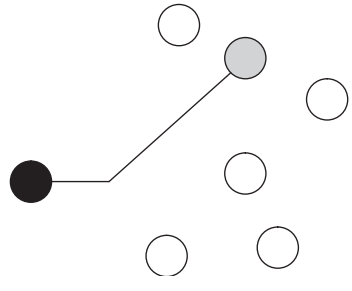


Figure 1.3 Unicast as one-to-one communication

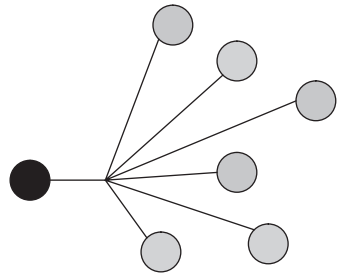


Figure 1.4 Broadcast as one-to-all communication

In multipoint unicasting, a source sends an individual copy of a message to each recipient. If five individuals in a workgroup wish to receive a copy of the same file using FTP, the server would send the file to each of the five recipients separately, using 5 times as much bandwidth as for a single transfer. This is wasteful of network bandwidth.

1.3.2 Broadcast

Under certain circumstances, it can be more efficient to transmit one copy of a message to all network nodes and let the receiving nodes decide if they want the message. Distributing the task of duplicating the packets among the network nodes rather than focusing the task at the sender's host machine is advantageous for the sender. This is referred to as broadcasting. Figure 1.4 illustrates the one-to-all nature of broadcast. There are many network hardware technologies that include mechanisms to broadcast packets to multiple destinations at the same time. With Ethernet, broadcast delivery can be accomplished with a single packet transmission on the wire.

One significant feature of broadcasting is that it relieves the source from the task of duplicating packets that are destined for multiple recipients. The source transmits a single copy of the packet to the appropriate broadcast address and the network devices take over, duplicating the packet as needed to cover the network.

1.3.3 Multicast

Multicast is the delivery of information to a group of destinations simultaneously using the most efficient strategy to deliver the messages over each link of the network only once, creating copies only when the links to the destinations diverge. Multicast falls between unicast and broadcast. Rather than sending data to a single host or to all hosts on the network, multicast aims to deliver data to a select group of hosts. The host group is defined by a specified multicast address. Figure 1.5 highlights the one-to-many nature of multicast.

Once a host group is set up and the sender starts transmitting packets to the host group address, the network infrastructure takes on the responsibility for delivering the necessary data streams to all members of the group. Only one copy of a multicast message passes over any link along the delivery path in the network. Copies of the message are only made when paths diverage at a router, thus helping to conserve bandwidth. In multicast, as in

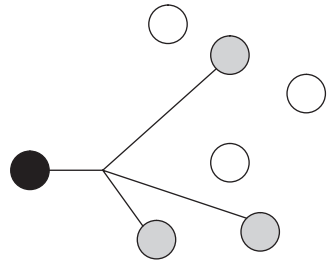


Figure 1.5 Multicast as one-to-many communication

broadcast, the source of a message usually does not know which recipients are within the group or the state of the data delivery.

On an LAN, each host's network interface monitors the LAN and accepts packets addressed to the multicast address that defines the host group to which the packets belong. Unlike broadcasting, multicasting allows each host to choose whether it wants to accept multicast packets. In a WAN, membership information concerning host groups has to be maintained across the entire WAN or internetwork. Procedures for joining a host group and maintaining a host group differ from the LAN case, since routers have to get involved, passing group information among themselves to maintain the required information for multicast within the internetwork.

1.4 IP Multicast

IP multicast is a technique for one-to-many communication over IP infrastructure. In IP multicast, routers create optimal distribution paths in real time for packets sent to a multicast destination address. IP multicast is designed to scale to a large receiver population by not requiring prior knowledge of who or how many receivers there are. Multicast utilizes network infrastructure efficiently by requiring the source to send a packet only once, even if it needs to be delivered to a large number of receivers. The nodes in the network take care of replicating the packet to reach multiple receivers only where necessary. In networking, the term *multicast* is synonomous with IP multicast. Stephen Deering was the first to describe the standard multicast model for IP networks (Deering, 1989, 1991; Deering and Cheriton, 1990). This model describes how end-systems are to send and receive multicast packets. The model includes both an explicit set of requirements and several implicit requirements.

We will briefly describe the key concepts of IP multicast, focusing on the minimum set of functionality required to provide multicast, namely group management and multicast routing. A much more detailed description of multicast group management and multicast routing, as well as other multicast mechanisms such as session announcement and reliable delivery of multicast traffic, is provided in Chapter 2.

1.4.1 Multicast Groups

The model for multicast builds on IP-style semantics as well as open and dynamic groups. A source can send multicast packets at any time, with no need to register, announce or

schedule transmission. IP multicast is based on UDP (not TCP), so packets are delivered using a best-effort policy without reliability or congestion control. Sources only need to know a multicast address. They do not need to know group membership, and they do not need to be a member of the multicast group to which they are sending. A group can have any number of sources. Multicast group members can join or leave a multicast group at will. There is no need to register, synchronize or negotiate with a centralized group management entity.

An IP multicast group address is used by sources and the receivers to send and receive content. Sources use the group address as the IP destination address in their data packets. Receivers use this group address to inform the network that they are interested in receiving packets sent to that group. For example, if some content is associated with a specific multicast group address, the source will send data packets destined for the group to the multicast group address. Multicast receivers will inform the network that they are interested in receiving data packets sent to the multicast group. This is done by having the receiver *join* the multicast group. The protocol used by receivers to join a group is called the Internet Group Management Protocol (IGMP). IGMP is used by IP hosts and adjacent multicast routers to establish multicast group memberships. It is an integral part of the IP multicast specification, operating above the network layer, although it does not actually act as a transport protocol. The IGMP protocol is implemented on the host and router side. The host reports its membership of a group to its local router, with the router listening to reports from hosts and periodically sending out queries to check whether multicast group information has changed.

1.4.2 Multicast Routing

IP multicast does not require a source sending to a given group to know about the receivers of the group. In order to route multicast packets, the network constructs a multicast tree. The multicast tree construction is initiated by network nodes close to the receivers and is thus receiver-driven. This allows it to scale to a large receiver population. The multicast distribution tree is constructed for that group once the receivers join a particular IP multicast group.

There are several different multicast routing protocols, and each one has its own unique technological solution. The Distance-Vector Multicast Routing Protocol (DVMRP) is the earliest protocol for multicast routing. A key concept introduced by DVMRP is the use of separate forwarding trees for each multicast group. This fundamental principle continues to be used in the newer multicast routing protocols. DVMRP is a distance-vector protocol that provides very limited flexibility, functionality and scalability. Nonetheless, the early Internet multicast experiments in the multicast backbone (or MBone for short) were based on DVMRP. DVMRP has largely been superseded by some of the newer protocols.

The next incarnation of multicast routing was an extension of the popular OSPF protocol called MOSPF. OSPF is designed explicitly to be an interior gateway protocol, meaning that it resides within a single autonomous system. Hence, any extensions to OSPF, such as MOSPF, would also reside within the confines of one autonomous system. MOSPF provides an effective means for a single corporation, university or other organization to support multicast routing, but it cannot support wide-scale applications that require the use of the Internet. MOSPF is used sporadically for some specialized applications, but it is not prevalent.

A new breed of multicast routing protocols was developed in the late 1990s. This family of protocols is collectively known as Protocol-Independent Multicast (PIM). The name PIM is derived from the fact that these multicast forwarding protocols are not dependent upon any one specific routing protocol (MOSPF, for example, requires the use of the OSPF unicast routing protocol). Instead, PIM will take advantage of the existing routing tables, regardless of how they were constructed, in order to forward multicast data. There are a few different versions of PIM. One version is called PIM Dense Mode (PIM-DM). As the name implies, this particular rendering of PIM is optimized for densely populated communities of users. The philosophy of PIM-DM is to *push* multicast data towards the users. A router that deploys this protocol simply floods multicast traffic streams to all interfaces (this is similar to broadcast mechanisms). If the downstream routers are not attached to any users that require this particular multicast stream, they will send a *stop* message to the upstream router. This message is called a *prune* message, since the upstream router will then prune its forwarding tree to eliminate that particular branch. PIM-DM routers forward all multicast traffic until a downstream router objects. The most commonly implemented form of PIM is PIM Sparse Mode (PIM-SM). This is a direct contrast to PIM-DM since it invokes a *pull* methodology instead of a *push* technique. This means that PIM-SM routers must specifically request a particular multicast stream before the data is forwarded to them. PIM-SM is well suited for the Internet since it reduces the overhead and bandwidth requirements for multicast data streams.

1.5 Multicast in Cellular Mobile Networks

Only in recent years has resource-efficient multicast for third-generation networks received significant attention in the research community. The goal behind introducing multicast capabilities in third-generation networks is to realize cost efficiencies by transmitting the same data to multiple receivers on shared resources.

In UMTS, multicast can be achieved with the Cell Broadcast Service (CBS), IP multicast and the Multimedia Broadcast/Multicast Service (MBMS). CBS is a standard for GSM and UMTS that allows for simultaneous delivery of messages to multiple users in a specified area. IP multicast is standardized as an optional feature in UMTS networks, but is not efficient in terms of bandwidth consumption. MBMS, on the other hand, standardized by the 3GPP in Release 6, is designed to support efficient broadcast and multicast packet delivery in GPRS and UMTS networks. In CDMA2000 networks, multicast can be achieved by means of IP multicast or the Broadcast/Multicast Service (BCMCS). BCMCS provides multimedia content transmission from a single source to multiple users in CDMA2000 networks. The options for providing multicast in UMTS and CDMA2000 networks are briefly described below.

1.5.1 Cell Broadcast Service

With CBS, fairly small-sized messages are broadcast to all mobile handsets and similar devices within a designated geographical area. The broadcast range can be varied, from a single cell to the entire network. CBS messages originate from a Cell Broadcast Entity (CBE). CBEs are usually connected to a Content Casting Centre (CCC), which is in turn connected to the Cell Broadcast Centre (CBC). CBS messages are then sent from the CBC to the cells, in accordance with the CBS's coverage requirements.

A CBS message page comprises 82 octets, which, using the default character set, equates to 93 characters. Up to 15 of these pages may be concatenated to form a CBS message. Each page of such a message will have the same message identifier (indicating the source of the message) and the same serial number. Using this information, the mobile telephone is able to identify and ignore broadcasts of already received messages. CBS is not widely deployed today. In the USA, for example, most handsets do not have cell broadcast capabilities and the major network operators have not deployed the technology in their networks.

CBS messaging is particularly appropriate for emergency purposes, as it is not as affected by traffic load. It is therefore usable during a disaster when load spikes tend to crash networks. Taking the Tsunami catastrophe in Asia as an example, an operator in Sri Lanka was able to provide ongoing emergency information to its subscribers, to warn of incoming waves, to give news updates, to direct people to supply and distribution centres and even to arrange donation collections by relying on CBS (Wikipedia, 2008b).

1.5.2 IP Multicast

IP multicast is standardized as an optional feature in UMTS networks. With this feature, the IP multicast routing protocol is terminated at the gateway of the UMTS network. As a result, this solution requires that only the UMTS gateway be multicast aware. A similar mechanism applies to CDMA2000 networks. With IP multicast in UMTS, the gateway serves as an IGMP-designated router and performs IGMP signalling on point-to-point packet-data channels. Multicast data is forwarded to the receiver on point-to-point or unicast channels. Only the UMTS terminal and the gateway are multicast aware. This architecture allows the network to treat multicast traffic in the same manner as unicast traffic. With this architecture, however, no bandwidth savings can be achieved in the network. Multicast packets are duplicated at the gateway and transmitted to each multicast group member individually.

This multicast architecture reduces the load on a wireless source that wishes to transmit multicast traffic within the network. The source only needs to send one copy of multicast packet data to the gateway. The gateway is reponsible for forwarding multicast packets on to the multicast distribution tree. The drawback of this design is that the UMTS multicast source does not receive any information from multicast members. When the multicast group does not have any members, the source continues to transmit its multicast data to the gateway. The source is not aware of the empty state of the multicast group. A modified signalling connection between the gateway and the source can avoid this situation.

1.5.3 MBMS for UMTS

MBMS is a standard for multimedia broadcast and multicast content delivery in UMTS networks. MBMS has two operational modes, namely broadcast and multicast. In broadcast mode, transmission takes place regardless of user presence in a defined area, whereas, in multicast mode, solely users that belong to a multicast group are serviced. Consequently, MBMS in multicast mode requires group management, whereas broadcast mode does not.

MBMS classifies different types of user service according to the method of distribution. The three user service types are streaming, file download and carousel services. Streaming services provide a stream of continuous media, for instance audio and video traffic. File

download services are used to deliver binary files that can only be consumed when the file is downloaded in its entirety. Carousel services finally provide content that is retransmitted periodically. Chapter 6 provides a much more detailed description of MBMS.

1.5.4 BCMCS for CDMA2000

BCMCS is a CDMA2000 standard that provides the capability to deliver broadcast and multicast services. As with MBMS, the BCMCS is designed for multimedia content transmission from a single source to multiple users. BCMCS can be used for broadcast services, in which all users within the broadcasting area can receive the information, and multicast services, in which only users that have subscribed to the service can receive the information.

IETF protocols are widely employed in BCMCS. This minimizes the number of new protocols required and maximizes the utilization of well-accepted standards. BCMCS is described in much more detail in Chapter 7.

1.6 Summary

In this chapter, the most important aspects of multicast in third-generation networks have been introduced. Firstly, an introduction to cellular mobile communication systems was given, from first-generation analog systems to third-generation systems and beyond. This was followed by an overview of the most relevant aspects of data networking, covering such aspects as circuit-switched and packet-switched networks, the Internet protocol suite and routing in internetworks. We then described the basics of multipoint communications in data networks. Multipoint communications can be achieved by means of multipoint unicast, broadcast or multicast. With multipoint unicast, packets destined for a group are duplicated and transmitted to group members individually. With broadcast, data is transmitted within a given geographic area to all members of the network. With multicast, data is transmitted to only those recipients that have requested to be members of a given group. We then briefly described the key fundamentals of IP multicast, the technique for one-to-many communication over IP infrastructure. This was then followed by a brief overview of how multicast can be achieved in third-generation networks.

The subsequent chapters will provide much more detail on several key topics. IP multicast is described in Chapter 2, an overview of the third-generation networks UMTS and CDMA2000 is given in Chapter 3, followed by a complete description of the multicast mechanisms for these networks in Chapters 4 to 6. Chapters 8 to 10 then analyse the performance of these multicast mechanisms. Chapter 11 finally looks beyond multicast in third-generation networks and investigates how multicast can be achieved in heterogeneous networks.