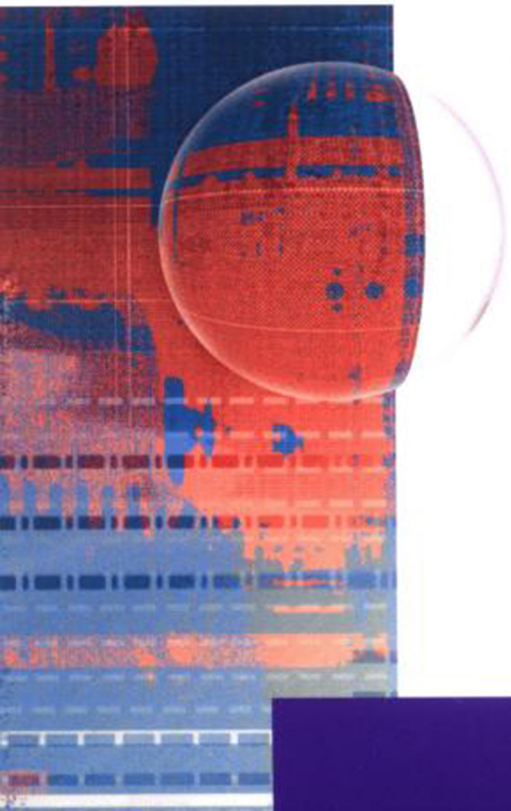
McGraw-Hill NETWORKING
PROFESSIONAL

PACKET BROADBAND NETWORK HANDBOOK



- How to make your network forward-looking and service-ready
- Optical Ethernet and other new trends in transport
- Advanced technologies such as GMPLS, QoS, COPS, and more

HAOJIN WANG

The McGraw-Hill Companies

Library of Congress Cataloging-in-Publication Data

Wang, Haojin.
Packet broadband network handbook / Haojin Wang.
p. cm.
ISBN 0-07-140837-1
1. Broadband communication systems. 2. Packet switching (Data transmission) I. Title.
TK5103.4 .W363 2002
621.382'1-dc21 2002032563

1 2 3 4 5 6 7 8 9 0 DOC/DOC 0 9 8 7 6 5 4 3 2

ISBN 0-07-137006-4

The sponsoring editor for this book was Marjorie Spencer; the editing supervisor was Caroline Levine and the production supervisor was Pamela Pelton. It was set in Vendome ICG by Wayne A. Palmer of McGraw-Hill Professional's composition unit, Hightstown, NJ.

Printed and bound by RR Donnelley.

McGraw-Hill books are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please write to the Director of Special Sales, Professional Publishing, McGraw-Hill, Two Penn Plaza, New York, NY 10121-2298. Or contact your local bookstore.

This book is printed on recycled, acid-free paper containing a minimum of 50% recycled, de-inked fiber.

PART 1

Packet Network Foundations

Part I of this book introduces four widely deployed packet network technologies: X.25, frame relay, asynchronous transfer mode (ATM), and Internet protocol (IP).

Before packet networks, communications technology used circuit-switched telephone networks with dedicated, analog circuits that functioned on a "always on once activated" basis. A dedicated circuit cannot be used for other purposes even if no communications are taking place at the moment. In regard to telephone conversations, it is estimated that on the average a dedicated circuit carried active traffic only 20 to 25 percent of the time and is idle the other 75 to 80 percent. Moreover, other services such as video data streams cannot be efficiently carried on circuit-switched networks.

Packet networks based on packet switching technologies represent a radical departure. The key idea behind packet switching is that a message or a conversation is broken into independent, small pieces of information called *packets* that are either equal or variable in size. These packets are sent individually to a destination and are reassembled there. No physical resource is dedicated to a connection, and connections become virtual, thus allowing many users to share the same physical network resource.

The concept of packet switching is attributed to Paul Baran who first outlined its principles in an essay published in 1964 in the journal *On Distributed Communications*. The term *packet switching* itself was coined by Donald Davies, a physicist at the British National Physical Lab, who came up with the same packet switching idea independently. It is interesting to note that a few decades earlier, a similar discovery in physics by Albert Einstein—that waves of light can be broken into a stream of individual photons—led to the development of quantum mechanics.

Packet networks allow more efficient use of network resources. Each packet occupies a transmission facility only for the duration of the transmission, leaving the facility available for other users when no transmission is taking place.

Packet-switched networks are highly fault-tolerant. From the very start of their development, network survivability was a major design goal. Because packet networks do not rely on dedicated physical connections, packets can be routed via alternative routes in case of an outage in the original communications link.

Packet networks can support bandwidth on-demand and flexible bandwidth allocation. Bandwidth is allocated at the time of communication, and the amount of bandwidth allocated is based on need. In

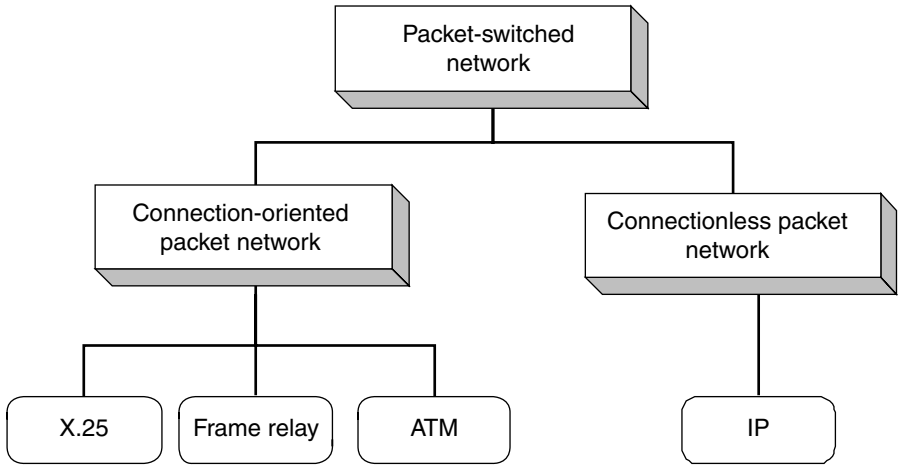


Figure P1-1
Packet network foundations.

contrast, a bandwidth of 64 Kbps is built into the infrastructure of circuit-switched telephone networks.

Since the very first packet-switched network ARPNET was built in 1969, many packet switching technologies have been developed. Among them, four have endured and achieved large-scale deployment: X.25, frame relay, ATM, and IP. The packet network technologies can be generally divided into the two categories shown in Fig. P1-1: connection-oriented and connectionless.

A connection-oriented packet network provides a virtual connection for a communications session between a source and a destination either on a permanent or a temporary basis. Packet networks of this category include X.25, frame relay, and ATM.

Connectionless packet networks are represented by IP. In a classic IP network, packets of the same message may travel different routes and arrive at the destination out of order. The distinction between connection-oriented and connectionless technologies is not absolute: Connection-oriented packet networks such as ATM and X.25 can also provide connectionless service. In addition, the ubiquitous connectionless IP network is moving toward being connection-oriented via new IP network infrastructures such as multiprotocol label switching (MPLS), as will be seen in Part 4 of this book.

CHAPTER 1

X.25 Networks

1.1 Introduction

X.25 is one of the very first standards to elevate packet networking technology to the global level and lay the foundation for later comers like frame relay and ATM. This section, after providing some background information, introduces the X.25 network model and its components.

1.1.1 A Brief History

X.25 is the first generation of public data network standards to serve as a successor to message switching networks and the first public switched data network (PSDN) in parallel to public switched telephone networks (PSTNs). X.25 standards were first defined in 1976 by the CCITT (since renamed the ITU-T). Two major revisions were made subsequently in 1980 and 1984. The *X.25* has become synonymous with a set of standards that together define packet network technology: X.32, X.75, X.3, X.28, and X.29, although the X.25 specification itself merely defines an interface between user applications and an X.25 network edge switch. As used throughout this chapter, the term *X.25* will be used to refer to the overall X.25 network rather than a particular specification, unless explicitly noted otherwise.

X.25 is still one of the most widely used connection-oriented packet networks with guaranteed quality of service (QoS). Its users are mostly business customers with widely dispersed and communications-intensive operations in sectors such as utilities, finance, insurance, retail, and transportation.

An X.25 packet network can be either public or private. Many corporations have determined that it is more economical to establish and use their own telecommunications facilities. In these cases, packet switches are obtained from network equipment providers, and private X.25 networks are set up for the exclusive use of and administrated by specific organizations.

1.1.2 X.25 Network Reference Model

As shown in Fig. 1-1, the X.25 network includes the functions of the bottom three layers of the open systems interconnection (OSI) network reference model (Black 1994): the physical layer, the data link layer, and the

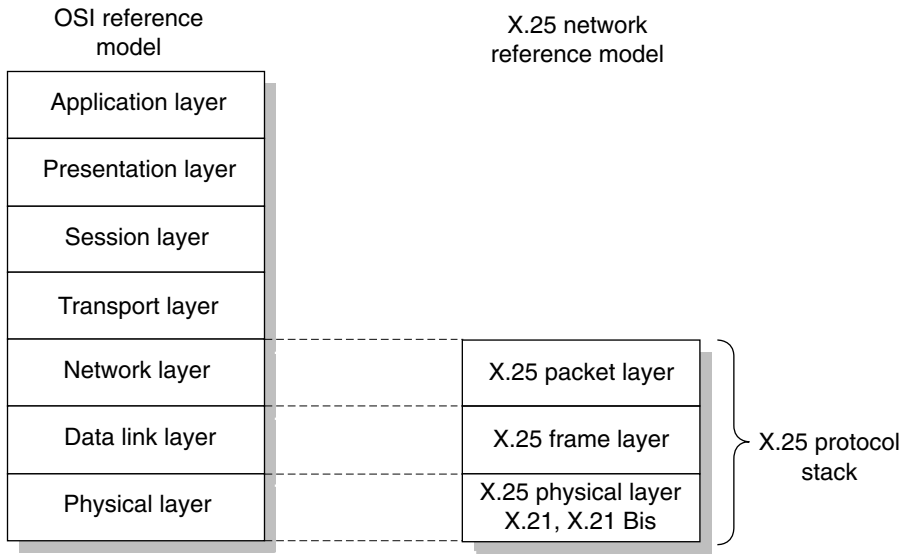


Figure 1-1 X.25 network reference model.

network layer. The X.25 standards focus on the network layer, but offer some specifications for the physical layer and the data link layer as well.

1.1.3 X.25 Network Components

An X.25 network is made up of four types of network elements, with an analogue transmission line connecting them, as shown in Fig. 1-2. In this logical view of an X.25 network, functional components are specified in the X.25 specification. Multiple functional components are often combined into one network device in an actual implementation of the X.25 network. For example, the data-terminal equipment (DTE) and the packet switching equipment (PSE) can be physically combined inside an X.25 switch.

1.1.3.1 PAD The packet assembler/disassembler (PAD) can be viewed as a special network interface provided for character-mode DTEs, such as terminals. When a user sends data to the network, the PAD interface takes a stream of data from a character-mode DTE and assembles it into packets to be sent to the network. At the receiving end, the PAD disassembles packets from the network into streams of data to be sent to a character-mode DTE. The PAD function is often implemented in soft-

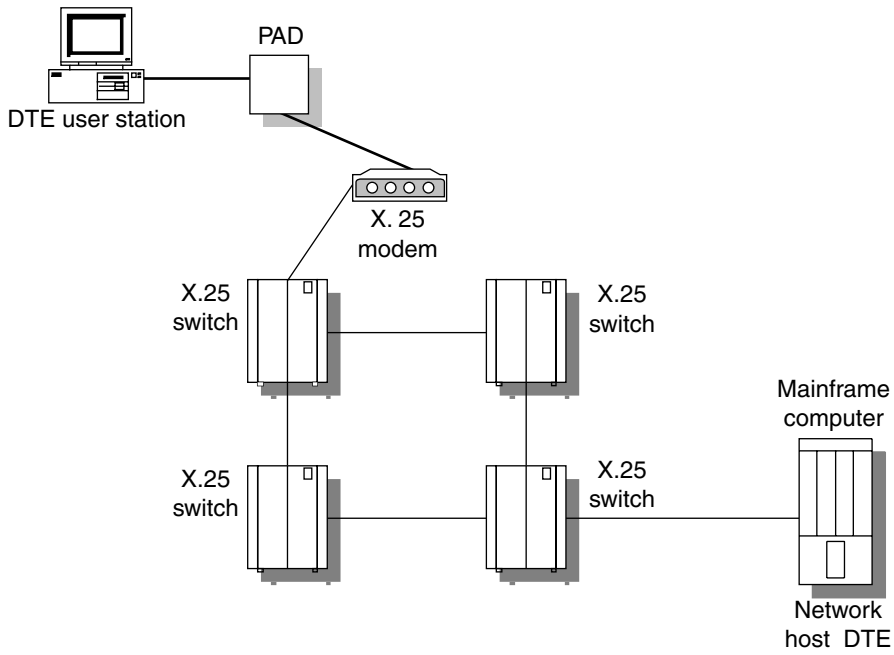


Figure 1-2
An X.25 network overview.

ware that is built into the same device as the DTE (ITU-T 1997a; ITU-T 1997b; ITU-T 2000a; ITU-T 2000b).

1.1.3.2 DTE Data-terminal equipment (DTE) is an interface point between a user equipment and an X.25 network, and it is implemented in a computer or computer-related device (ISO/IEC 1995; ITU-T 2000a). DTE devices such as networked computers are where user applications reside. DTEs are divided into packet-mode DTEs and character-mode DTEs. Packet-mode DTEs are typically computer systems that implement the X.25 protocol in hardware and software and are capable of sending and receiving packets. Character-mode DTEs are asynchronous devices, such as terminals and printers, that send or receive data one character at a time and require a PAD component to interact with other X.25 network components.

1.1.3.3 DCE Data-circuit-terminating equipment (DCE) is a network interface to packet-mode DTEs. The DTE-DCE interface represents the boundary between a user and a network, and a DCE device is often at the edge of a public data network. The DCE function is often built

into a X.25 switch located at the edge of a public X.25 network (ITU-T 1996).

1.1.3.4 PSE Packet-switching elements (PSEs) are packet switches connected over telecommunications facilities (phone lines, for example) in a PSDN. A main function of PSEs is to determine and pass packets to the next switch in a path.

1.2 Physical Layer of X.25 Networks

The physical layer of the X.25 network deals with the transmission medium and provides procedural and functional interfaces between a DTE and a DCE. This layer is specified in the CCITT X.21, X.21-bis, and V.24 recommendations (ITU-T 1998):

- ITU-T Recommendation X.21 specifies the operations of digital circuitry. X.21, initially defined in 1976, specifies the digital signaling interface of how a DTE can set up and clear calls by exchanging signaling messages with DTE (ITU-T 1992). The X.21 interface operates over eight interchange circuits: signal ground, DTE common return, transmit, receive, control, indication, signal element timing, and byte timing. For example, a DTE uses specialized circuits like transmit and control to transmit data and control information. A DCE uses a specialized receiver and indication circuits for data and control information. The functions of the circuits are defined in recommendation X.24, and their electrical characteristics are defined in recommendation X.27.
- ITU-T Recommendation X.21-bis defines an analogue interface to support the access to digital circuit-switched networks using an analogue access line. X.21-bis provides procedures for sending and receiving addressing information to enable a DTE to establish switched circuits with other DTEs that have access to a digital network (ITU-T 1988).
- ITU-T Recommendation V.24 provides procedures to enable a DTE to operate over a leased analogue circuit that connects the DTE to a packet switching node or concentrator.

The physical medium X.25 networks operate on can be either analog or digital transmission lines. One assumption for the X.25 protocols is that transmission facilities like analog lines are inherently unreliable and error-prone. The assumed maximum data rate is up to 64 Kbps.

1.3 Data Link Layer of X.25 Networks

The data link layer of X.25 networks takes a bit stream received from the physical layer and presents to the packet layer a view of an error-free link to transmit packets. X.25 networks adopt the most commonly used high-level data link control (HDLC) protocol for data link layer. This section first provides a brief historical background of data link layer protocols, and then moves on to a detailed description of the HDLC frame format.

1.3.1 Overview of Link Layer Protocols

The responsibilities of the data link layer for X.25 networks (as well as other networks) include the following (ISO/IEC 1997):

- Interfacing the physical layer to receive or send data in a bit stream
- Delineating the received bit stream into link layer frames
- Synchronizing the link to ensure that the receiver is in step with the transmitter
- Detecting transmission errors and recovering from such errors
- Identifying and reporting certain protocol errors to higher layers

Since the early 1970s, the data link layer protocols have repeatedly evolved, and the industry has settled on a few that have achieved wide deployment. In the course of that evolution, data link layers themselves grew from being character-based to being bit-oriented, "character-based" meaning they handled one character at a time with a minimum unit of 8-bit characters. It was IBM in the early 1970s that developed the first bit-oriented data link layer protocol for data communication, called synchronous data link control (SDLC), which allowed the transfer of an arbitrary binary sequence of data without alignment at 8-character

boundaries. SDLC steadily gained broad acceptance, with the ISO and IEC adding enhancements to it. The result was the HDLC protocol known as ISO standard 13239 (ISO/IEC 1997).

A data link layer can be either balanced or unbalanced. In a balanced mode, each station is responsible for both information transmission and error recovery using acknowledgments. In an unbalanced mode, one of the two communicating stations is designated as primary and the other as secondary. The primary station polls the secondary station, which responds with information frames. The primary station then acknowledges receipt of frames from the secondary station.

Several link layer protocols derived from HDLC have also achieved wide deployment:

- *Link access protocol, balanced (LAPB).* LAPB is derived from HDLC and is one of the most commonly used data link protocols. In addition to the other characteristics of HDLC, it provides a mechanism to create a logical link connection for the upper layers. The 1980 revision of the X.25 standards uses LAPB as the link layer protocol.
- *Link access protocol (LAP).* LAP is an earlier version of LAPB and is not very widely used today. The 1976 version of X.25 uses LAP as the data link layer protocol.
- *Link access procedure, D channel (LAPD).* LAPD is derived from LAPB and used for ISDN to transmit data between DTEs through D channels, which are signaling channels as opposed to data channels, and especially between a DTE and an ISDN node.
- *Logical link control (LLC).* LLC is used in Ethernet data link layers and enables X.25 packets to be transmitted through local area network (LAN) channels.

1.3.2 X.25 LAPB

X.25 networks use LAPB as the data link layer protocol, and LAPB is based on HDLC, which is a fundamental component of such packet network technologies as X.25, frame relay, and integrated service digital networks (ISDNs). HDLC, like its predecessor SDLC, is bit-oriented synchronous protocol passing variable-length frames over a point-to-point or multipoint network. HDCL can operate over either dedicated or switched facilities with three possible operating modes: simplex, half-duplex, or full duplex.

Out of many data link layer functions, the X.25 data link layer performs the following two main ones:

- *Data packaging.* It defines a format of data transport unit called *frame* and encapsulates data bits (0s and 1s) into frames, analogous to specifying carts for transporting goods and packaging goods into each cart, to be transported by railroad.
- *A procedure of transporting frames.* It defines a procedure for receiving data and detecting error in received data frames, and for handling any detected error.

The LAPB frame format helps further understand the X.25 data link layer functions. The generic frame format is shown in Fig. 1-3 and has five fields as described below:

- *Flag (8 bits).* A frame always begins and ends with a flag. The flag is an 8-bit sequence (01111110) that delimits a frame. Note that a key function of the data link layer is to delineate a frame by inserting the flag at the beginning and end of a frame. What if the user data contains the same bit pattern as the flag? When the data link layer detects a sequence of five 1s in a row in user data, it inserts a 0 immediately after the fifth 1 in the transmitted bit stream. The data link layer at the receiving end removes inserted 0s by looking for the sequence of five 1s followed by a stuffed 0.
- *Address (8 bits).* The address field indicates the type of frame—a command or a response to a command. It also specifies whether the frame is being sent from a DTE to a DCE or from a DCE to a DTE.
- *Control (8 bits).* The 8-bit field indicates the type of a frame, i.e., a frame that carries user data, signaling data or network maintenance data. LAPB supports an 8-bit control field while the HDLC standard supports optional 16-bit, 32-bit, and 64-bit lengths of the control field. The types of frame are described in more detail later.
- *Information.* This variable-length field contains network layer packets that in turn contain user data. The format of this field depends on the type of frame.
- *Frame check sequence (FCS).* The 16-bit field is set by the frame's transmitter and interpreted by the receiver to detect error in data content.

The X.25 data link layer supports three types of frame: informational, supervisory, and unnumbered. Type is indicated in the control field of the frame.

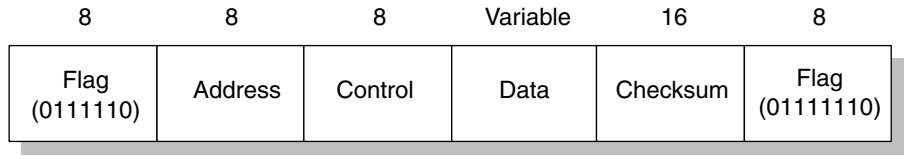


Figure 1-3
HDLC frame format.

1.3.2.1 The Informational Frame This type of frame contains actual user data being transferred. The control field of this kind of frames contains a frame sequence number for the last frame sent and an expected sequence number of the next frame.

1.3.2.2 The Supervisory Frame This type of frame allows a receiver to notify a sender the following status information:

- *Receiver ready acknowledgment.* It is an acknowledgment frame indicating the next frame expected.
- *Reject-negative acknowledgment.* It is used to indicate any transmission error detected and to request retransmission of the frame.
- *Receiver-not-ready.* It notifies a sender to stop sending frames due to a temporary problem at the receiving end.

1.3.2.3 Unnumbered Frame This type of frame provides a means for a DTE and a DCE to set up and acknowledge the HDLC mode and to terminate the data link layer connection. The HDLC standard defines a set of control messages to request and acknowledge the HDLC mode. Three HDLC modes are defined: asynchronous balanced mode (ABM), normal response mode (NRM), and asynchronous response mode (ARM). Note that LAPB uses ABM only. This frame is also used to terminate the data link layer connection.

The link layer operations as defined in LAPB include the following:

- Establishment of a connection between a DTE and DCE, e.g., between a user application and a X.25 switch
- Transfer of data
- Steps for error detection and error recovery
- Teardown of a connection

The data link layer ensures reliable, accurate transfer of data from a sender to a receiver, and only data that is received without error is passed to the packet layer, the layer above.

1.3.3 Data Link Layer Operations

Flow control, error detection, and frame retransmission in case of error are the main operations of the X.25 link layer because the X.25 network is designed to provide error-free frame delivery and flow control at the link layer. This subsection provides a brief overview of link layer operations.

X.25 based on LAPB protocol adopts a frame numbering-based acknowledgment and retransmission scheme to ensure error-free delivery and efficient transmission. In addition, the X.25 switch supports a full duplex link operation that allows two-way simultaneous transmissions. In contrast, one prevalent transmission method is "stop-and-wait," which ensures the receipt of correct packets.

An X.25 switch can continuously send frames up to a limit w without waiting for an acknowledgment. The window size defines the maximum number of frames a sender can send before receiving an acknowledgment. The sender needs to hold all transmitted frames until an acknowledgment frame is received in case of the need for a retransmission. The w is called the *window size* of the link layer control mechanism.

A receiving switch discards a frame if an error in the frame is detected. All the frames after the error frame are discarded regardless of whether each frame is received correctly. The receiver sends a REJ (reject) frame to the sender to indicate the sequence number of the frame with the error. On the transmitting side, the REJ frame indicates the frame sequence number where an error has occurred. The sender retransmits the correct frame and all the frames after it. This approach to retransmission is called *go-back-n*. Note that this error checking and retransmission, if performed, is on a node-by-node basis and could potentially have a significant performance impact if errors occur often.

1.4 X.25 Packet Layer

The data link layer of an X.25 network, after performing the data link layer processing, strips the frame header and passes the data units to the network layer, also known as the packet-to-packet layer of X.25 networks.

This section first introduces a generic packet format and then discusses the important concept of virtual connection. It then uses an operation example to thread all the concepts together to illustrate how an X.25 network works.

The network layer of the X.25 network performs two main functions:

- It takes the data received from the link layer and processes it into units to be presented to the application/user (user data packaging).
- It establishes the procedure for associating two end users/applications.

1.4.1 X.25 Packet Format

X.25 packets are data units seen at the network layer and have the general format shown in Fig. 1-4 (Motorola Codex 1991).

An X.25 packet is carried in the LAPB information field and consists of a header and a user data section. The header section consists of a basic header and an optional extended header. Every packet must contain the basic header. The basic header can be extended for certain packet types, such as the call-setup packets that can specify DTE addresses and additional user facilities. The basic packet header has 3 bytes, divided into three sections of 1 byte each, as shown in Fig. 1-4 and described below.

1.4.1.1 General Format Identifier (4 Bits) The qualifier (Q) bit allows a transport layer protocol to separate control data from user data. It is set by a local DTE to indicate that the data being sent is an X.25 control message. The delivery (D) bit is the delivery confirmation used during the X.25 switched virtual connection setup. The D bit allows a local DTE to request an acknowledgment of data packets from remote DTEs. The default behavior is that a local DCE acknowledges the packets sent by the local DTE. However, when the D bit is set, the acknowledgment must come from a remote DTE. The next two bits specify the

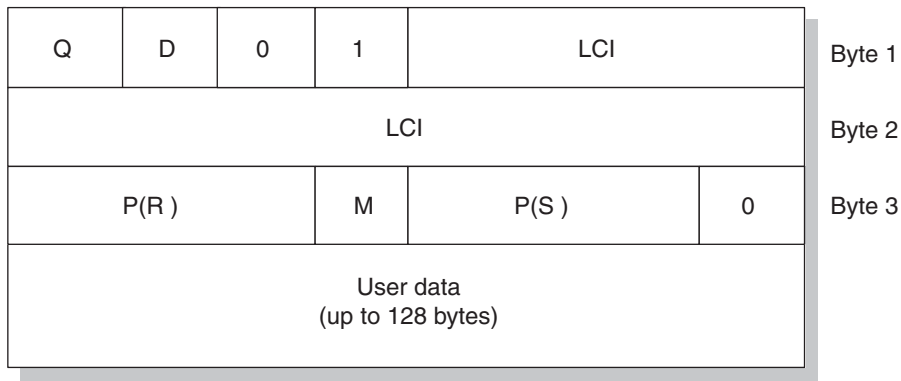


Figure 1-4
Format of X.25 packet.

packet type and packet header length, with 01 indicating a data packet with a 3-octet header.

1.4.1.2 Logical Channel Identifier The 12-bit logical channel identifier (LCI) identifies a virtual circuit (VC) and consists of two parts: a 4-bit logical channel group number and an 8-bit logical channel number. With virtual channel 0 reserved, a DTE-DCE interface can support a maximum of 4095 virtual circuits made up of 16 groups and 256 virtual circuits for each group. There are four types of virtual circuits, each having a block of virtual circuit numbers:

Permanent virtual circuits (PVCs). These are set up permanently and are assigned lowest LCI numbers.

Incoming-only switched virtual circuits (incoming SVCs). These are one-way virtual circuits set up from a DCE to a DTE, not the other way around. The LCI numbers assigned to this type of virtual circuit are higher than PVC LCI numbers but lower than those of the next subset of virtual circuits.

Two-way switched virtual circuits (two-way SVCs). These allow a DTE and DCE to request connections to each other and occupy LCI numbers that are higher than those of incoming SVCs and lower than those of the next set of virtual circuits.

Outgoing-only switched virtual circuits (outgoing SVCs). These allow a local DTE to request a connection to a DCE. The outgoing SVCs have the highest range of LCI numbers.

The third byte of the X.25 packet header has four fields. The packet layers receive and send sequence number fields [P(R)and P(S)] provide a support for a packet traffic and flow control and ensure that packets are received in the proper order. More on this is described later in this section. The More (M) bit, when set to 1, indicates that the information in the user data field is part of contiguous data across several data packets. When set to 0, it indicates that the user data field is the last part of the contiguous data.

The extended packet header has two kinds of fields: the DTE addressing fields and facility fields. The DTE addressing fields contain the address of a called DTE and may optionally contain the calling DTE address as well. The DTE address is defined in the CCITT X.121 recommendation and made up of two parts: the data network identification code (DNIC) and the DTE identifier. The DNIC identifies a public data network and the DTE identifier is a number uniquely identifying a

DTE within a public data network the DTE is connected to. The DTE identifier is normally assigned by a network service provider, as is an IP address or a phone number.

The facility field specifies the X.25 user facilities to be used for a virtual circuit. Two groups of user facilities are specified: essential and additional. Essential facilities are provided by all X.25 public data networks while additional facilities are optional. The type of facility specifies performance parameters such as packet and window size, throughput, etc.

The user data field has the default size of 128 bytes. The 1980 X.25 recommendation allows up to 1024 bytes of user data and the 1984 X.25 recommendation allows up to 4096 bytes of user data. The size of user data field in each packet can be negotiated at call-setup time.

1.4.2 The Concept of Virtual Circuit, PVC, and SVC

Virtual circuit is a key concept of the X.25 network layer and heavily influences the packet networking technologies that came later such as frame relay and ATM. As indicated by the description of the X.25 packet format, a network layer connection is represented and identified by a virtual circuit. A virtual circuit is a logical connection between a source and a destination. A connection is virtual as opposed to dedicated, because the data may pass through different physical links and share the physical facilities with other connections. X.25 network uses an LCI number to identify a segment of a virtual connection between a source and a destination that are identified by the calling and called DTE addresses, respectively.

There are two types of virtual circuits defined by the X.25 standards: permanent virtual circuit (PVC) and switched virtual circuit (SVC). A PVC is a logical association between two TDEs that is permanently held by the network, regardless of whether or not there is data being passed between two DTEs. In contrast, an SVC is a logical connection that is dynamically set up and maintained only for a given time period between two DTEs. SVCs are closed or taken down when a data transfer session is completed and there is no more data to send.

PVCs are normally set up manually, while SVCs are set up using a signaling protocol. The X.25 standard specifies a procedure for SVC setup, as described below.

1.4.3 SVC Operations

There are three phases in an SVC life cycle, as shown in Fig. 1-5. The X.25 standard specifies the detailed steps of operation for each phase. Operational steps of each phase at the network layer are specified for setting up a SVC between a local DTE and a local DCE and between a remote DCE and a remote DTE.

1.4.3.1 Call Establishment The general steps for establishing a call, as shown in Fig. 1-5, are the following:

1. The local DTE generates an X.25 call-request packet and sends it to the local DCE.
2. The request is forwarded through a X.25 public data network and eventually delivered to the remote DTE in the form of an X.25 incoming call packet.
3. After validating the incoming call request and checking its own parameters, the remote DTE generates a call-accept packet to accept the call.
4. The packet is passed through the public X.25 network and arrives at the originating DTE as a call-connected packet. At this point, a virtual circuit has been established and the data transfer phase commences.

1.4.3.2 Data Transfer Data transfer is the process of passing user data packets between two DTEs. It is assumed that at this point a virtual circuit, either permanent or switched, has already been established. The likely general steps are the following:

1. The local DTE creates a data packet and passes it to the local DCE.
2. The local DCE acknowledges receipt of the packet by sending a receive-ready packet.
3. Then the data packet is forwarded through the public X.25 data network and delivered to the remote DCE. The remote DCE passes the data packet to the remote DTE.
4. The remote DTE acknowledges the receipt of the data packet with a receive-ready packet sent to the remote DCE, which in turns forwards the packet back to the local DCE. The local DCE then sends the acknowledgment packet back to the local DTE.

1.4.3.3 Call Disconnect The originating DTE, the terminating DTE, or the X.25 network itself can initiate to close the switched virtual circuit. The likely general steps for disconnecting a virtual circuit, as shown in Fig. 1-5, are the following:

1. The initiating party, assuming it is the originating DTE, generates a clear-request packet and passes the request to the connected DCE.
2. The DCE forwards the packet through the X.25 network to the remote DTE, and the packet arrives at the terminating DTE as a clear-indication packet. The DTE clears the virtual circuit and returns any local resources to the available resource pool. Then the DTE sends back a clear-indication packet.

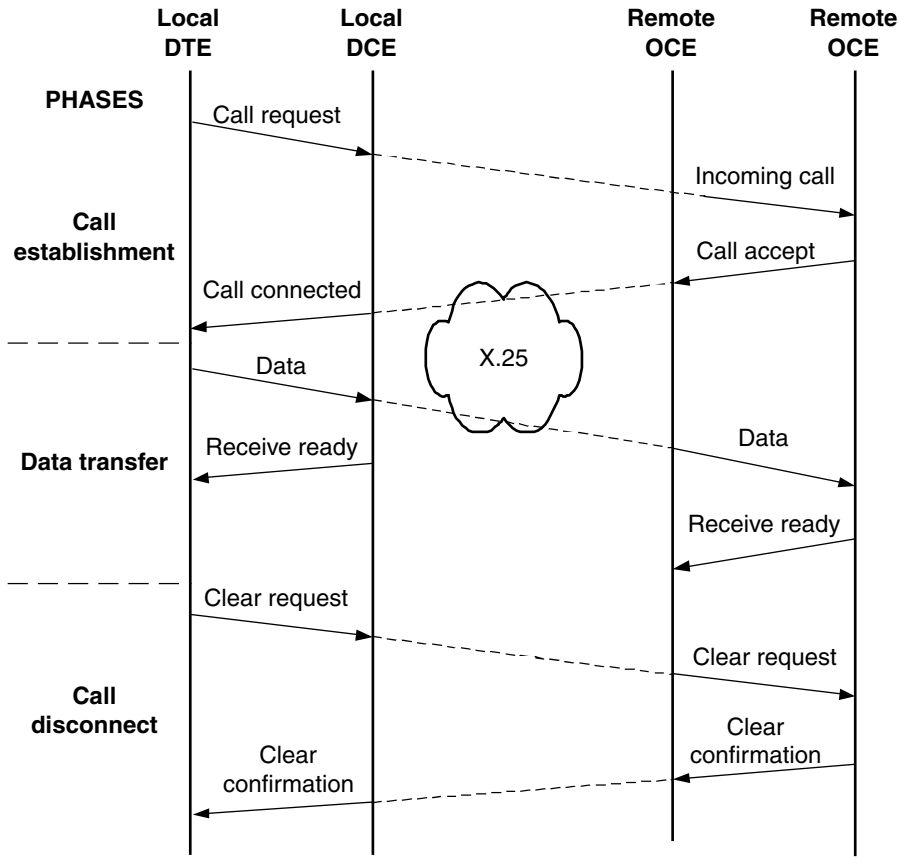


Figure 1-5
Three phases of X.25 SVC life cycle.

3. The clear-indication packet, in a similar fashion, gets forwarded back to the originating DTE, which clears the virtual circuit locally. The SVC has been disconnected.

1.4.4 Traffic and Congestion Control at Packet Layer

In addition to the link layer flow control, the X.25 packet layer provides a packet sequence number-based flow control mechanism between a source and a destination DTE. The packet layer send-sequence number P(S) identifies the current packet with respect to the packet header sequence number modulus. A receiver uses P(R) in the acknowledge packet to indicate the send-sequence number of the next expected packet from the sender. Note that this sequence number-based packet layer flow control is on a per call basis while the LAPB data link layer flow control is on a per link basis.

X.25 also has a window mechanism for flow control. A window at the transmitter defines the maximum number of packets it can send without receiving a packet acknowledgment from the destination. X.25 defines a similar window size as the receiver that specifies how many packets a receiving DTE can accept before issuing an acknowledgment packet.

The reason that both the X.25 link layer and the network layers have a flow control mechanism is that the link layer flow control is concerned with a single link while the packet level is concerned with flow over the whole network.

1.5 X.25 Applications

This section starts with an end-to-end application example that illustrates how an X.25 network works at both the data link and the network layer. It then describes a set of X.25 services before concluding with an overview of the deployments of X.25 technology.

AN END-TO-END X.25 NETWORK OPERATION EXAMPLE

A cash withdrawal transaction at an automatic teller machine connected to a financial database via an X.25 network illustrates how an X.25 net-

work works. Connected to the X.25 network at the other end is a bank computer that stores all customer data and performs customer validations. Assume that a virtual connection is already established using the procedure described in the preceding section. The example is for the purpose of illustration and does not imply any particular implementation.

Step 1. The customer chooses a transaction type and enters an account number and a password. The data is converted into blocks of characters via the PAD that is built inside the computer at the teller machine and then sent to the packet layer. The packet layer creates packets out of blocks of data from the PAD by adding a packet header, which includes fields such as the virtual circuit number on which the packet should be sent, the packet type, and so on. When the packet is complete, it is delivered to the data link layer, which builds a frame from each packet after adding a frame header that includes various kinds of framing information such as frame check sequence (FCS). The frame is then sent to the physical layer that sends the frame, bit by bit, to the local DCE.

Step 2. The local DCE physical layer sends the received bits to the link layer. When the data link layer has collected a recognizable frame, it computes an FCS. It then compares its own FCS with the computed one. If they match, the data link layer removes the framing information and passes the resulting packet to the packet level. If an error is detected in the information field of the frame, however—due to a transmission error, for example—the data link layer sends a retransmission request back to the calling party, which in this case is the computer inside the automatic teller machine. When the data link layer is satisfied with the frame it receives, it strips the frame header fields and passes the rest of the frame to the packet layer.

Step 3. The packet layer looks at the packet header and determines where the packet will be routed next and then sends it back down to the physical layer for routing to the remote DCE. In this fashion of node-by-node handshake, the packet finally reaches the destination DTE, which in this case is a bank mainframe computer that contains all the bank's customer account data.

Step 4. At the destination DTE, the packet is passed from the physical layer to the data link layer and then to the packet layer of the destination DTE. The packet layer in turn passes the data to the application layer and then sends an acknowledgment packet back to the automatic teller machine, the source DTE, to acknowledge the receipt of the ID and password data. The destination DTE may also choose to embed the validation success information in the acknowledgment packet.

Step 5. Once the user transaction is complete, the automatic teller machine initiates the process of clearing the connection and taking down the call, in the steps described in the preceding section.

1.5.1 Additional X.25 Services

On top of the data link and network layer service, the newer version of the X.25 recommendations defines an additional set of services that an X.25 network will support. These services, termed *facilities*, in the X.25 specification, allow greater flexibility for users at X.25 terminals to customize the interface between an X.25 DTE and a public X.25 network (ITU-T 1996). The following is a brief overview of these services:

Flow control negotiation and packet retransmission. This service allows a DTE to change the packet and window sizes for flow control that is used at the interface between a DTE and a local DCE. The service of packet retransmission allows a DTE to initiate a retransmission of an unacknowledged data packet by issuing a DTE reject packet to the network, with a sequence number specified by the reject packet.

Throughput-class negotiation. This service allows a DTE to request a particular throughput class, i.e., the maximum amount of data that can pass through a network in a given time period when the network is saturated. Each throughput class corresponds to a specific amount of data represented in bits per second.

Call barring. This service allows a DTE to reject all incoming calls from the outside and the outgoing calls originating from the higher-layer applications of the node. This service is useful in case of network congestion or for security control.

One-way logical channel. This is a variant of call barring: It allows a DTE to bar incoming and outgoing calls on a specified group of logical channels. The other channels on the DTE are unaffected.

Closed user group. This service allows a set of DTEs to form a group and to exclusively communicate with each other within the group. Any calls from a nonmember are rejected.

Fast select. This service allows a DTE to send up to 128 bytes of user data within a signaling call-setup or call-clear packet. This eliminates the need to send a separate data packet for user data of sizes up to 128 bytes. This service requires that a remote DTE must have subscribed

to the fast-select acceptance service and be willing to accept a calling DTE's fast-select call.

Reverse charging. This service allows a DTE to request that a remote DTE pays for a call. The remote DTE must have subscribed to the reverse-charging acceptance service in order to accept such a reserve charging request.

1.5.2 Deployment of X.25

By some accounts, X.25 remains the most widely deployed packet network technology on a worldwide basis, despite important limitations such as its 64-kb/s speed limit and the cost overheads associated with the extensive error handling incurred by using coaxial and twisted pair copper cable.

Private X.25 networks are typically deployed within large organizations that have widely dispersed and communications-intensive operations in fields such as finance, insurance, transportation, utilities, and retail. X.25 is the network technology of choice, mainly because it offers guaranteed, timely delivery of user data, which is critical to applications like financial transactions.

REVIEW QUESTIONS

1. Discuss the main responsibilities of the data link layer. Describe the two types of data link protocols, i.e., character-based and bit-oriented, and the main differences between them.
2. Describe the relationships between HDLC and LAPB/LAPD.
3. What are the three types of frames defined in the HDLC frame and what purposes do each serve?
4. Discuss the rationale behind the node-by-node acknowledge scheme used at the data link layer of X.25 network.
5. Discuss the X.25 PVC and SVC concepts and the main differences between them.
6. Briefly describe four different types of virtual circuits and how the LCI field of an X.25 packet is related to a virtual circuit type.
7. Describe what the *fast select* service is and under what circumstances the service might be useful.

8. In an intermediate X.25 switch, what is the highest layer of X.25 protocol stack that examines an incoming packet to determine how to switch the packet, the data link layer, or the packet layer?
9. Discuss one of the main reasons for X.25 to remain the network of choice for many large corporations in sectors such as finance, insurance, utility, and retail.

REFERENCES

Black, U. 1994. *X.25 and Related Protocols*. Los Alamitos, CA: IEEE Computer Society Press.

ISO/IEC. 1995. "Information Technology—Telecommunications and Information Exchange Between Systems—High-Level Data Link Control Procedures—Description of the X.25 LAPB-Compatible DTE Data Link Procedures." ISO/IEC 7776. Web site: www.iso.org.

ISO/IEC. 1997. "Information Technology—Telecommunication and Information Exchange Between Systems—High-Level Data Link Control (HDLC) Procedures." ISO/IEC 13239. Web site: www.iso.org.

ITU-T. 1988. "Use on Public Data Networks of Data Terminal Equipment (DTE) Which is Designed for Interfacing to Synchronous V-Series Modems." Recommendation X.21-bis. Web site: www.itu.int/ITU-T/.

ITU-T. 1992. "Interface Between Data Terminal Equipment and Data Circuit-Terminating Equipment for Synchronous Operation on Public Data Networks." Recommendation X.21. Web site: www.itu.int/ITU-T/.

ITU-T. 1996. "Interface between Data Terminal Equipment (DTE) and Data Circuit-terminating Equipment (DCE) for Terminals Operating in the Packet Mode and Connected to Public Data Networks by Dedicated Circuit." Recommendation X.25. Web site: www.itu.int/ITU-T/.

ITU-T. 1997a. "DTE/DCE Interface for a Start-Stop Mode Data Terminal Equipment Accessing the Packet Assembly/Disassembly facility (PAD) in a Public Data Network Situated in the Same Country." Recommendation X.28. Web site: www.itu.int/ITU-T/.

ITU-T. 1997b. "Procedures for the Exchange of Control Information and User Data Between a Packet Assembly/Disassembly (PAD) Facility and a Packet Mode DTE or Another PAD." Recommendation X.29. Web site: www.itu.int/ITU-T/.

ITU-T. 2000a. "List of Definitions for Interchange Circuits Between Data Terminal Equipment (DTE) and Data Circuit-Terminating Equipment (DCE)." Recommendation V.24. Web site: www.itu.int/ITU-T/.

ITU-T. 2000b. "Packet Assembly/Disassembly Facility (PAD) in a Public Data Network." Recommendation X.3. Web site: www.itu.int/ITU-T/.

Motorola Codex. 1991. The *Basics Book X.25 Packet Switching.* Reading, MA: Addison-Wesley.

CHAPTER 2

Frame Relay Networks

2.1 Introduction

2.1.1 A Brief History

Two factors that underlay the fast development and deployment of frame relay technologies in the early 1990s were the fast growing bandwidth requirement and the maturing of transmission technologies such as synchronous optical networks (SONETs) and fiber optical networks. The phenomenal growth of the Internet in the late 1980s and early 1990s by far outgrew the network infrastructure of the time. At that time, X.25 packet-switching networks and proprietary networks built upon private lines carried the load of data traffic. The 64-Kbps bandwidth that X.25 offers paled in face of the fast growing demand. Meantime, the adoption and deployment of optical transmission technologies like high-speed coax cable and SONET provided much more reliable transmission facilities and a higher transmission speed. Frame relay (FR) was designed to take advantage of the newer transmission technologies and provides much higher bandwidth with much more cost-effective transfer of packet data.

Frame relay has its origin in the ISDN standards. In 1988, the ITU-T (then called the CCITT) approved Recommendation I.122 that defines a "framework for additional packet mode bearer services." Recommendation I.122 is a part of the ISDN specifications and actually lays the foundation for frame relay network (ITU-T 1993). It was realized at that time that an ISDN protocol termed *link access protocol—D channel* (LAPD), originally defined for ISDN signaling, could also be used for some other applications. I.122 defines a framework for using LAPD for other applications, and one of these other applications turned out to be frame relay.

Frame relay networks are built on the experience of X.25 and share many similarities with X.25. Thus they can be viewed in many ways as the successor to X.25 data networks. Frame relay looks like a "lean" version of X.25 in many respects: It eliminates the expensive overhead of the X.25 network layer and reduces much of the complexity of the link layer protocol while preserving the basic packet switching concepts of PVC and SVC.

A decade after it was first made available, frame relay remains one of the most widely deployed packet networks so far. This is reflected by the volume of frame relay service revenues, estimated to be close to $13 billion at the end of year 2001.

2.1.2 Frame Relay Network and Protocol Stack

A frame relay network consists of two types of network elements: the frame relay access device (FRAD) and the frame relay switching device. (In X.25 terminology, FRAD is known as data-terminal equipment, or DTE, and frame relay switching devices are known as data-circuit-terminating equipment, or DCE, terms discussed in Chap. 1.) Both elements are illustrated in Fig. 2-3.

FRADs are access points of a frame network and often located at the customer's premises, where frame relay traffic originates or terminates. Examples of FRADs include frame relay access routers, bridges, or workstations that have frame relay interfaces.

Frame relay switching devices do not terminate frame relay traffic but forward frames to the next node along a virtual connection. They are located in a network carrier's backbone, and examples include frame relay switches and routers.

As a data link layer technology, frame relay covers the bottom two layers of the OSI network reference model. The frame relay protocol stacks at the network edge and at intermediate nodes are shown in Fig. 2-1.

At an edge device of a frame relay network, the following three functional modules are defined for the data link layer of the frame relay protocol stack:

- A frame relay service access point (SAP) function that interfaces between the network layer protocol such as Internet protocol/Internetwork Packet Exchange protocol (IP/IPX) and the frame relay data link layer. The SAP converts the network layer data into frame relay frames and vice versa and associates an application with a specific PVC/SVC.
- The frame relay signaling function that maintains the virtual connections between two frame relay systems.
- The frame relay data link function that is responsible for packing the application data into frame relay frames to be transported over physical links.

At an intermediate node of a frame relay network, frames go up no further than the data link layer on the protocol stack before they are forwarded to the next node. Such an intermediate node can be a frame relay router or switch that is connected to at least two other

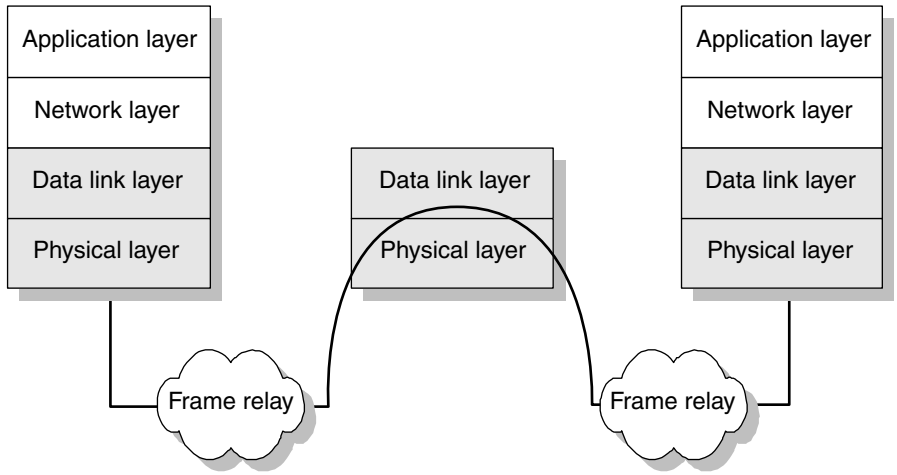


Figure 2-1
Frame relay network reference model.

frame relay devices. The main responsibility of an intermediate node is switching or routing frames to the next frame relay device.

Frame relay can run over physical layer protocols such as DS1, DS3, SONET OC3, OC12, STS1, etc. Frame relay specifies an interface to the physical layer for mapping a frame to each transmission medium.

2.1.3 Frame Relay Standards

There are three major standards organizations mainly responsible for frame relay standards. The first two are ITU-T and the American National Standards Institute (ANSI). ITU-T and ANSI define the frame relay core standards that include the data link layer frame format and signaling standards for PVC and SVC at the network layer, as listed in Table 2-1.

The third frame relay standards organization that plays an important role in the wide adoption and deployment of the frame relay technology is the Frame Relay Forum (FRF), which mainly consists of the frame relay equipment vendors, service providers, and other interested parties in the industry. The main mission of FRF is to promote frame relay technology and interoperability between vendor products by developing and approving implementation agreements (IAs) and conformance tests. IAs have practical importance for frame relay equipment vendors since their products must conform to them if the vendors want those products to be accepted by potential customers. The major FRF IAs are listed in the appendix for this chapter.

TABLE 2-1
Frame Relay Standards Summary

Description	ITU standards
Frame relay core, frame format	Q922 Annex A
Access signaling, PVC signaling, SVC signaling	Q933
Data link service	I.233

2.2 Frame Relay Basics

This section describes key aspects of the data link layer of frame relay. The frame format is introduced first as the context for the frame relay operations to be discussed later.

2.2.1 Frame Structure

The cornerstone of frame relay technology is the frame and its structure. The simplicity of the frame is one of the reasons for the rapid and sustained acceptance and usage of frame relay technology. The frame consists of a header and five other fields (shown in Fig. 2-2) (ITU-T 1992; ITU-T 1993):

Open flag: a 1-byte field that is an HDLC flag to mark the beginning of a frame.

Address: a 2-byte field that is also called the *frame header* and consists of eight different sub-fields as described below.

Frame check sum (FCS): a 2-byte field that allows detection of up to three random bit errors or a burst of sixteen bit errors. This field is directly inherited from the HDLC frame.

Data/information field: the user data field that includes the higher-layer protocol data and end user data, up to the maximum size of 8188 bytes.

Close flag: similar to the open flag field, a 1-byte HDLC flag that marks the end of a frame.

The frame header has total of 16 bits and includes the following frame relay-specific information:

Data link connection identifier (DLCI): a 10-bit field that can identify up to 1024 virtual circuits per interface

Command/response (C/R): a 1-bit field that identifies the type of the frame, either a command/request or a response to a request

Address field extension (EA): a 1-bit field indicating whether an extended address field is present

Forward explicit congestion notification (FECN): a 1-bit flag indicating to the receiver the presence of congestion in the network

Backward explicit congestion notification (BECN): a 1-bit flag indicating to the sender the presence of congestion in the network

Discard eligibility (DE): a 1-bit flag indicating whether the network should discard the frame under congestion conditions

The data link connection identifier (DLCI), a key field in the frame header, identifies logical connections that are multiplexed into a physical circuit. In the basic mode of addressing, the DLCI value is significant to a local frame relay interface only. One implication is that devices at two ends of a connection may use different DLCIs to identify the same virtual connection. A separate DLCI is defined for each frame relay device that is directly connected to the device. A second implication is that the number of nodes in a fully meshed network is limited to no more than 1024. But not all DLCI values are available for user virtual connections and some are reserved:

- DLCI 0 and 1023 are reserved for management.
- DLCI 1 to 15 and 1008 to 1022 are reserved for future use.
- DLCI 992 to 1007 are reserved for layer 2 management of frame relay bearer service.
- DLCI numbers 16 to 991 are available for subscribers for each user frame relay network.

The frame address field is extensible. If the address field extension (EA) bit, which is the first to be transmitted, is set to 0, it indicates that

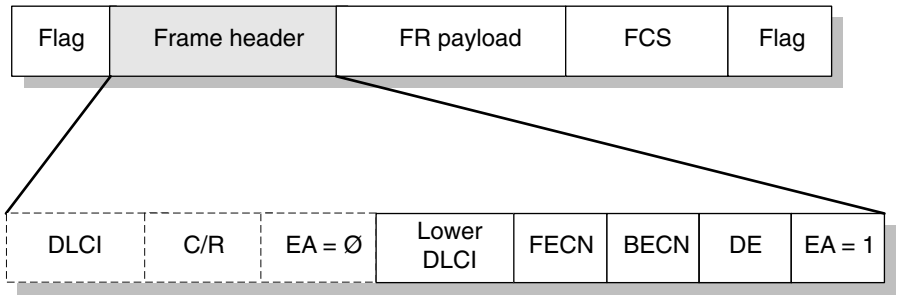


Figure 2-2
Frame relay frame structure.

another byte of address field follows. If set to 1, it indicates the address byte is the last one. Currently all implementations use a 2-byte header, and thus the first EA bit always set to 0 and the second EA to 1, as shown in Fig. 2-2. The EA bit allows the expansion of the address field to accommodate the larger frame relay networks.

2.2.2 FR Virtual Circuits

Frame relay provides a connection-oriented data link layer communication pipe between two DTEs, also known as FRADs. This is a virtual connection, also referred to as a *virtual circuit*, and each VC is uniquely identified by a DLCI. As described earlier, the DLCI at a source and the DLCI at a destination DTE may be two different numbers for the same virtual circuit due to the fact that DLCI is only locally significant.

A frame relay VC provides a bidirectional communication path between a source and a destination DTE. Multiple VCs are multiplexed onto a single physical link at the physical layer to maximize the efficiency of the frame relay network.

There are two types of virtual circuits in frame relay networks: permanent virtual circuits and switched virtual circuits (ITU-T 1995).

2.2.2.1 Frame Relay PVCs Permanent virtual circuits are permanently established connections that are used for frequent and consistent data transfers between DTE devices across a frame relay network. A PVC is normally set up via manual provisioning by operators and stays "up" until it is manually taken down. A PVC can be in one of the two states:

Data transfer. The VC is busy transferring data between DTE devices.

Idle. The virtual connection between a source and a destination DTE device is "up" but not carrying any data at the moment.

PVCs account for the vast majority of the virtual circuits deployed in both public and private data networks up to date. One main advantage of PVCs is their simplicity in deployment and provisioning. PVC is a more suitable choice for applications requiring that a network traffic pattern be relatively stable and predictable, and service offerings—most of which are data services—are relatively simple. Flat billing is preferred by customers.

2.2.2.2 Frame Relay SVCs Switched virtual circuits are temporary connections used for a particular communication session. SVCs are set up and taken down automatically by the network via a simplified version

of ISDN signaling protocol Q931 (ITU-T 1998). In general, a SVC session consists of four operation stages:

Call setup. The virtual circuit is in the process of setup between a source and a destination DTE.

Data transfer. Data transfer over the circuit between DTE devices is in process.

Idle: The connection between DTE devices is "up" but no data is carried over the circuit.

Call termination. The virtual circuit is being terminated.

There were two main motivations for the development of FR SVCs, reducing network operation complexity and offering more flexible, diverse services over frame relay networks. The manual setup and maintenance of PVCs can be labor-intensive and error-prone if the number of virtual circuits is very large. Dynamic setup and tear-down of virtual circuits by the system becomes an appealing proposition. In addition, SVCs are more suitable for complicated service such as voice and fax over frame relay networks that feature more dynamic traffic patterns.

2.2.3 Frame Relay Network Management

Simple network management protocol (SNMP), the standard management protocol defined by the Internet Engineering Task Force (IETF) for the Internet, is the protocol choice for managing frame relay network. IETF RFC 1315 defines a standard managed object model for each frame relay interface at a frame relay DTE (Brown, Baker, and Carvalho 1992).

RFC 1315 models a frame relay DTE as a user-to-network interface (UNI) with many virtual connections to various destinations or neighbors. This view provides a network manager with the ability to group and associate all virtual connections to their corresponding physical connections, and thus allows for simpler diagnostics and troubleshooting.

RFC 1315 defines three groups (also known as tables in SNMP terminology) of managed objects that include the UNI interface [also termed as *data link connection management interface* (DLCMI)], virtual connection, and diagnostic errors. The interface object group models the UNI interface itself with objects such as the interface identifier (ID), address, state, polling interval, and maximum number of supported virtual connections on this interface. The virtual connection object table models virtual connections, one entry in the table per connection, and the managed

objects include data connection link identifier (DCLI) for a connection, frames and octets sent and received on a connection, congestion notifications [forward explicit congestion notifications (FECNs) and backward explicit congestion notifications (BECNs)] received, and performance measurements on the connection. The third table, the error data table, records the error type, the data containing errors, and the times of error occurrence on the interface.

AN END-TO-END APPLICATION EXAMPLE

An application example will help pull together the concepts discussed so far and illustrate how a frame relay network works.

Assume that an organization subscribes to a frame relay PVC service of *N* Kbps to connect two locations over a public frame relay network to allow the employees at the two locations to exchange email and documents over the frame relay network and to support other applications. Two Ethernet LANs are directly connected to the FRADs, which in turn are connected to the public frame relay network, via a backbone frame relay network, as shown in Fig. 2-3.

Also, assume that the enterprise is responsible for provisioning the two local FRADs connected to the public frame relay network. The provisioning tasks include mapping an IP address to a DLCI, and assigning a port for the DLCI. A complete virtual circuit between the frame relay switches at each end is assumed to have been provisioned by the network carrier already.

Assume a large file is sent by one employee at location A to another at location B. The frame relay network operations involve the following steps:

Step 1. Host A sends data in a large file to the frame relay-enabled edge router at the LAN. The IP layer of the frame relay router prepares the data and sends IP packets to the SAP module of the router, which maps the IP address to a DLCI, say, DLCI 101, and passes the information to the frame relay (data link) layer.

Step 2. The source FRAD prepares a frame. The data link layer builds a frame header, filling in a DLCI value according to the mapped DLCI value from the SAP and the local routing table. The frame is then sent out via the port specified at the provisioned routing table.

Step 3. The intermediate frame relay nodes switch the frame. When the frame arrives at the edge frame relay switch that the source FRAD is connected to, the physical layer, after some processing, passes the frame to the frame relay layer.

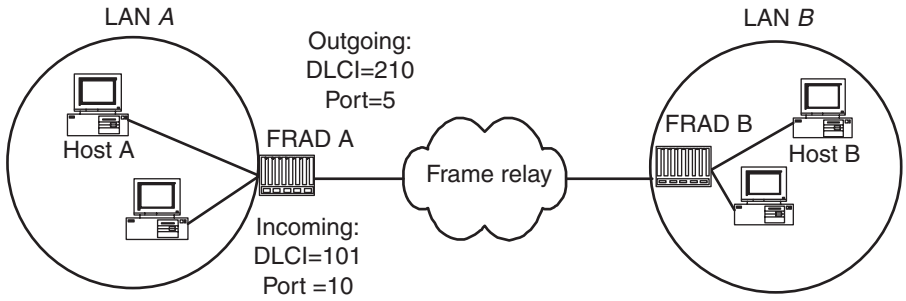


Figure 2-3
A frame relay network example.

The operations at an intermediate frame relay switching device are simple. When a frame comes into a frame relay switch to be sent across the network, the switch does three things:

1. Check the integrity of the frame using frame check sequence. If an error is detected, discard the frame.
2. Look up the routing table for the DLCI in the incoming frame header. If the DLCI is defined, there is a corresponding outgoing DLCI and port specified. In this case, the incoming DLCI 101 is defined and the outgoing DLCI is 210 and outgoing port is 5. Otherwise, if the DLCI is not defined, discard the frame.
3. Relay or forward the frame out via the specified outgoing port to the next connected frame relay switch or the destination FRAD. In a similar fashion, each frame relay switch forwards the frame along the virtual circuit, until it reaches the destination FRAD.

Step 4. The destination FRAD receives the frame. As at the previous nodes, the physical layer performs the layer specific processing first and then passes the frame to the frame relay layer. The header is extracted out for validation and the DLCI is mapped to a particular port and IP address. Then the frame is passed to the IP layer that extracts the IP layer information and finds out the intended destination host. The IP address is then mapped to a medium access control (MAC) address and data is send to the destination host B.

2.2.4 Comparison with X.25 and ATM

Frame relay is a successor technology to X.25 in many ways, and a simple comparison between the two is offered in Table 2-2. Overall, frame relay is characterized by its simplicity and much higher data rate.

First, frame relay is much simpler. It takes advantage of newer and much more reliable transmission links with lower error rates, eliminating many

TABLE 2-2

Comparisons of Frame Relay with X.25 and ATM

	X.25	Frame Relay
OSI layer	Data link and network layer	Data link layer
Transport unit	Packets of variable size, up to 512 bytes	Frames of variable size, up to 4096 bytes
Guarantee of delivery	Yes	No
Physical medium	Copper wire	Copper wire or coax cable
Transmission	Analog	DS1/E1, DS3/E3
Data rate	64 Kbps	64 Kbps to 54 Mbps

of the error-checking services needed by X.25. The elimination of complexity, combined with the presence of digital links, enables frame relay to operate at much higher speeds. In contrast, X.25 was designed to provide error-free delivery using high-error-rate links and is burdened with complicated error-checking mechanisms.

Frame relay is primarily a layer-2 technology with some specifications for the physical layer. This means that frame relay has significantly less processing to do at each node, which improves the throughput by an order of magnitude. In contrast, X.25 is defined for layers 1, 2, and 3 of the OSI network reference model, with focuses on the layer 2 and layer 3 (i.e., on the data link and network layers).

Frame networks use frames with a very simple structure. They contain an expanded address field that enables frame relay nodes to forward frames to their destinations with minimal processing. In contrast, X.25 packets contain several fields used for error handling and flow control, none of which is needed by frame relay.

Frame relay can dynamically allocate bandwidth at both the physical and logical channel levels during a call setup. In contrast, X.25 has only a fixed bandwidth available.

2.3 Frame Relay Network Interfaces and Signaling

An interface in a frame relay network is a point where the information between different types of network devices is exchanged and handshakes take place in order to connect all the pieces together to work as

an interconnected network. The messages plus the exchanges of the messages for this purpose are known as *signaling*.

Overall, frame relay signaling is relatively simple and largely based on the existing signaling standards. This is because frame relay was developed based on a simple rule: Keep the network protocol simple and push other tasks such as retransmission and reliability check to higher layers such as transport control protocol (TCP).

There are two types of interfaces in a frame relay network: user-to-network interface (UNI) and network-to-network interface (NNI), as shown in Fig. 2-4.

2.3.1 FR UNI

The UNI signaling in a frame relay network addresses three main issues: network congestion notification, virtual connection status notification, and SVC connection setup.

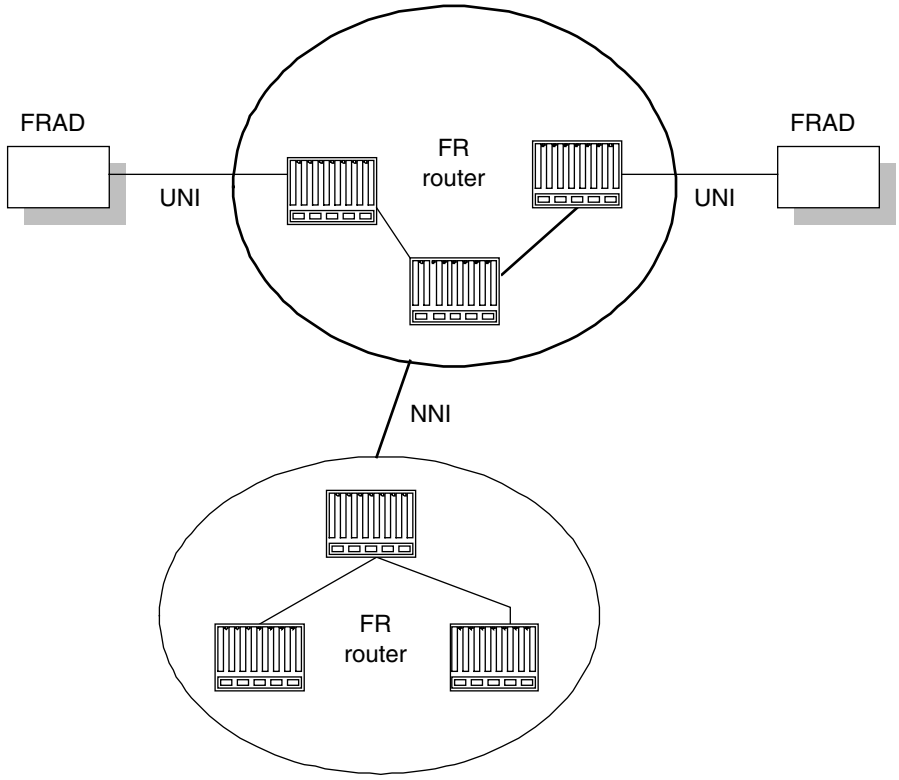


Figure 2-4
Frame relay UNI and NNI interfaces.

2.3.1.1 Congestion Notification There are two types of congestion notification mechanisms: implicit and explicit. The implicit congestion notification uses the congestion control of the higher layer such as TCP layer congestion notification and control. TCP, as will be discussed in Chap. 4, uses a sequence number and an acknowledgment number for a sender and receiver to notify each other of a congestion condition.

Explicit congestion notification uses the mechanism that is built into the frame structure: FECN and BECN fields in each frame that indicate to a sender and a receiver, respectively, the congestion condition.

The FECN and BECN mechanisms together support both source-based and destination-based congestion control protocols. For destination-based congestion control, a sender frame relay node sets the FECN bit in a frame traveling from a sender to a destination when the sender encounters congestion. This enables the destination node to invoke a congestion avoidance procedure such as discarding frames with the discard bit set. For source-based congestion control, a frame relay node sets the BECN bit in frames traveling from a destination to a sender on a bidirectional virtual circuit, so that the source node can adjust the rate at which the frames are being dispatched. The discard eligibility bit, when set to 1, indicates to the network this frame can be discarded in case of congestion. Thus this is, in effect, a simple priority scheme.

2.3.1.2 Virtual Connection Status Notification Two sides of a frame relay UNI need to communicate with each other about the status of the interface and the virtual connections across the interface. This is accomplished through a local management interface (LMI) using designated management frames with a unique DLCI address, as will be explained shortly.

2.3.1.3 SVC Connection Setup SVC signaling provides a mechanism to dynamically set up a virtual connection without the need for an operator's manual intervention. The SVC signaling protocol for a frame relay UNI interface is defined in Frame Relay Forum implementation agreement FRF.4 and its revision FRF.4.1, titled "SVC user-to-network implementation agreement." FRF.4 and FRF.4.1 SVC signaling is based on the existing standard protocols defined by ANSI T1.617 and ITU-T Q.933 (ANSI 1991a; ANSI 1991b; ITU-T 1995).

The SVC signaling mechanism includes the messages for call setup and call disconnect. Call setup includes information about a call, such as source and destination addresses, bandwidth parameters, and call acceptance.

At a high level, the SVC signaling process works as follows. For call setup, the network alerts an intended destination of the incoming call and the destination chooses to either accept or reject the call. If the destination accepts, the frame relay network builds an SVC across the frame relay switches and routers of the network. Once the SVC is established, the source and destination can start the data transfer. When the connection is no longer needed, either destination or source can initiate a procedure to terminate the call and tear down the connection.

2.3.2 Local Management Interface

The frame relay LMI is an important frame relay UNI specification that provides a set of enhancement to the basic frame relay specifications. The initial LMI specification was completed in late 1990s at Frame Relay Forum. There are three versions of LMI—one original specification and two later appendices:

- FRF 1, which has been superseded by FRF1.1. It specifies "keep-alive" messages. LMI "status" messages are one-way: A user device sends query message while the network responds with a status message.
- ANSI T1.617, Annex D. It extends FRF1 to allow two-way status messages for UNI interfaces (ANSI 1991a).
- ITU Q933, Annex A. It specifies the detailed message exchange procedures.

In summary, the two key components of the FR LMI are global addressing and PVC status messaging (ITU-T 1995).

The FR LMI redefines the DLCI field of the original frame structure so that each LMI DLCI value is globally significant rather than being significant only at a local interface, as shown in Fig. 2-5. The LMI frame has a few additional fields that include the following:

LMI DLCI. This is an LMI protocol discriminator containing a value indicating that a frame is an LMI frame.

Message type. It indicates the type of a message: either status query or informational.

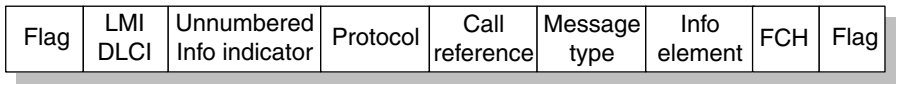


Figure 2-5
LMI frame structure.

Information elements. This field replaces the variable data field of the original frame and contains a variable number of information elements (IEs). Each IE consists of the following three separate subfields:

- *IE identifier:* uniquely identifies an IE
- *IE length:* indicates the length of an IE
- *Data:* variable length of data that may encapsulate the upper layer data

Another key LMI component is a set of PVC connection status messages. The LMI status messages allow a user device like a LAN router to either notify or poll for the connection status. In the "notify" mode, a user device sends a "keep alive" message to notify the network the connection to a network router is still up. In the poll mode, the user device may send a request for a report on the status of all PVCs on a particular port. The network then responds with a "status" message, either in the form of a "keep alive" response or in the form of a full report on the PVSs on the port.

In all, the LMI enables the exchange of the following three types of information:

- A "keep-alive" or heartbeat message in which one party tells another whether it is still up or alive
- The valid DLCIs defined on an interface or a port
- The status of each virtual circuit on a port

The LMI status query messaging mechanism allows for one-way queries and one-way responses. Only a user device can initiate a status query, and the network can only respond with a status message. Note that the one-way communication of the LMI makes the LMI applicable only to UNI, not the network-to-network interface where two-way communication is required.

2.3.3 Frame Relay NNI

The frame relay network-to-network interface allows one frame relay network to communicate with another frame relay network, exchanging information such as status on a connection across an NNI. An NNI can be used between two public frame relay networks or between a public network and a private one.

The frame relay NNI has not received as much attention as the frame relay UNI, because this interface comes into the picture only if one

carrier's network needs to interwork with another carrier's network. ANSI T1.617, Annex D, defines a bidirectional messaging mechanism for two networks to notify each other of PVC status (ANSI 1991a). Note that this bidirectional PVC status signaling works for both NNI and UNI.

2.4 Frame Relay Services

Next we'll present an overview of three frame relay services: data service, interworking with an ATM network, and voice over frame relay (VoFR). Two key factors that have much to do with the success of frame relay networks are the simplicity of the technology and the clearly defined target service. The simplicity of the technology is reflected in the simple protocol stack, easy management, and low network overhead. Frame relay was originally designed and developed for a single service, i.e., the data connectivity to support the fast growth of Internet data traffic. New services were later added to frame relay networks to leverage the large installed base for additional service revenue. Among the new services, interworking with ATM networks and voice over frame relay are two of the most prominent.

2.4.1 Data Service

Frame relay was originally designed and developed primarily for data service, providing LAN-to-LAN connections, and replacing the low-speed, analog-based, leased line networks.

2.4.1.1 Data Service Parameters The frame relay data service can have a range of associated service parameters that further define the service granularity and throughput. The service parameters provide a mechanism for a service provider to offer a service level agreement to its customers in terms of service parameters.

Throughput is one common service parameter that indicates the average number of frame relay information bits transferred per second across a user-network interface in one direction. Frame relay can provide a wide range of throughputs, ranging from 56 Kbps all the way to DS3 rate. Another service parameter is the committed information (or bit) rate (CIR) or committed burst, which indicates the maximum amount of data that a network agrees to transfer under normal conditions during a specified

measurement interval. A third service parameter is circuit excessive burst, which indicates the maximum amount of uncommitted data bits that a network is committed to deliver over the measurement interval.

2.4.1.2 LAN-to-LAN Connection Interconnecting distributed local area networks is one of the prime frame relay applications. Multiple LANs of an international corporation can be connected via a public frame relay network. The LANs of multiple organizations in a customer-suppler relationship can be connected in the same fashion but with access restrictions.

Frame relay provides an economically efficient solution to LAN-to-LAN connection. Before frame relay, connecting distributed enterprise LANs across wide-area networks was an expensive proposition, especially in a mesh configuration. Each additional connection required a physical link between two points across a wide-area network. Frame relay can multiplex multiple virtual connections onto a single physical circuit. If a customer needs an additional connection between two points, it becomes a simple issue of providing an additional virtual circuit.

2.4.1.3 Lease Line Replacement Frame relay can efficiently provide a leased line service. Frame relay, which operates at the data link layer (layer 2), provides a transparent pipe for user data traffic without regard to user data content and the protocols of the higher layers. This enables a large number of customers to migrate legacy traffic such as IBM's System Network Architecture (SNA) from low-speed leased lines onto a high-speed frame relay network. Frame relay simply encapsulates legacy protocols over the frame relay network. This proves to be one of the highly successful applications of frame relay, because legacy protocols such as SNA have a large installed base in both private and public leased line networks.

The encapsulation of legacy protocols inside a frame relay is standardized in IETF RFC 1490 (Bradley et al 1993). Typically, at the network edge, devices like SNA controllers, routers, front-end processors, or FRADs encapsulate SNA protocol data inside the frame relay data field as shown in Fig. 2-6. Frame relay provides a tunnel to carry the data in

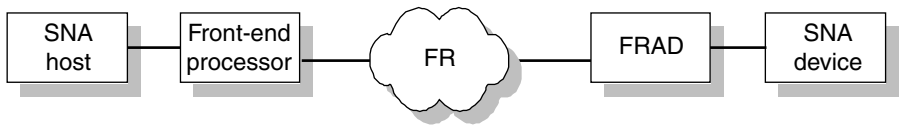


Figure 2-6
The legacy SNA protocol over frame relay. (Frame Relay Forum 1998)

the legacy protocol data unit from a source to a destination without any impact on the legacy data terminals at both ends. In addition to providing a transparent transport tunnel for the protocol data, frame relay nodes can multiplex multiple distinct protocols over a frame relay interface, supporting the interworking of multiple different legacy networks.

2.4.2 Frame Relay and ATM Interworking

Interworking with ATM interworking is another frame relay service that is high on customers' requirement lists. ATM, with its widely deployed base, is well suited for such applications as broadcast video and server farm connections. In addition, ATM can provide much higher speeds for the backbone network. These abilities have prompted both the ATM Forum and the Frame Relay Forum to define standard methods for interworking between a frame relay network and an ATM network.

There are two approaches to interworking between frame relays and ATM networks: tunneling and translation.

The tunneling approach encapsulates frame relay frames inside ATM cells, and end frame relay devices communicate with each other without the need to know an ATM network is in the middle. This is also known as *transport layer interworking*, and an interworking function (IWF) performs the data link layer mapping and encapsulation. Higher-layer protocols at the end-user devices are not affected at all. This approach is used in cases where ATM provides a high-speed pipe and frame relay devices are at both ends.

The second approach, translation, is also known as *service level interworking*. This approach translates signaling and protocols between a frame relay network and an ATM network. It is used in those cases where a frame relay device needs to communicate to an ATM device. This can be thought of as a frame relay network meeting an ATM network half-way, and necessitates all frame relay protocol-specific information being translated into its ATM counterpart at an IWF.

2.4.3 Voice Over Frame Relay

Voice over frame relay is a new breed of frame relay service. Given the large installed base of a frame relay network, network carriers would want to leverage the frame relay infrastructure to carry service other than data to maximize the efficiency of their frame relay network resources.

2.4.3.1 VoFR Components VoFR service involves several key elements: VoFR devices at end-user premises, a multiplexing scheme to allow voice and data service to share the same virtual connection at the same time, differentiated processing of different types of traffic, and a call signaling method.

VoFR-capable devices normally implement voice compression, echo cancellation, jitter processing, frame loss handling, fragmentation, and other processing. The FRF.11 specification, which specifies a VoFR implementation agreement, includes a set of voice compression algorithms that have been widely adopted in the industry (Frame Relay Forum 1997).

Another component of VoFR is a multiplexing scheme. The FRF.11 specification allows multiplexing of up to 255 subchannels onto a single frame relay virtual connection so that single DLCI may carry both voice and data payloads. Furthermore, it provides support for multiple voice payloads on the same or different subchannels within a single frame. This multiplexing scheme provides efficient usage of bandwidth.

To support the real-time-sensitive traffic like voice and fax data, the Frame Relay Forum has developed a frame fragmentation scheme to support different types of delays experienced on the network. Smaller frames on average experience short delays and can be more responsive to real-time traffic. This allows a frame relay device like FRAD or a frame relay switch/router to use a large number of smaller frames to carry voice traffic at access points and then interleave the smaller frames onto large frames on a high-speed link in a backbone network.

2.4.3.2 VoFR Access Model VoFR service supports three access modes, as shown in Fig. 2-7 and specified in FRF.11. The first mode is to have a FRAD directly support an analog phone and a FRAD function as a private branch exchange (PBX) as well as a data hub, as shown in the upper left-hand corner of Fig. 2-7. The second mode is to have a FRAD connected to a transparent switching device like a channel bank, which in turn interfaces the voice access devices like phones or fax machines. In the third mode, a FRAD is connected to a voice-switching device like a PBX.

VoFR access devices also allow for both data and voice traffic on the same virtual connection. As shown in Fig. 2-7, a data device is connected to the same FRAD as the voice devices and they share a virtual connection (DLCI). Also, multiple phones share the same virtual channel connected to the same FRAD. A VoFR-enabled edge router in general also allows a customer to set the traffic priority depending on the

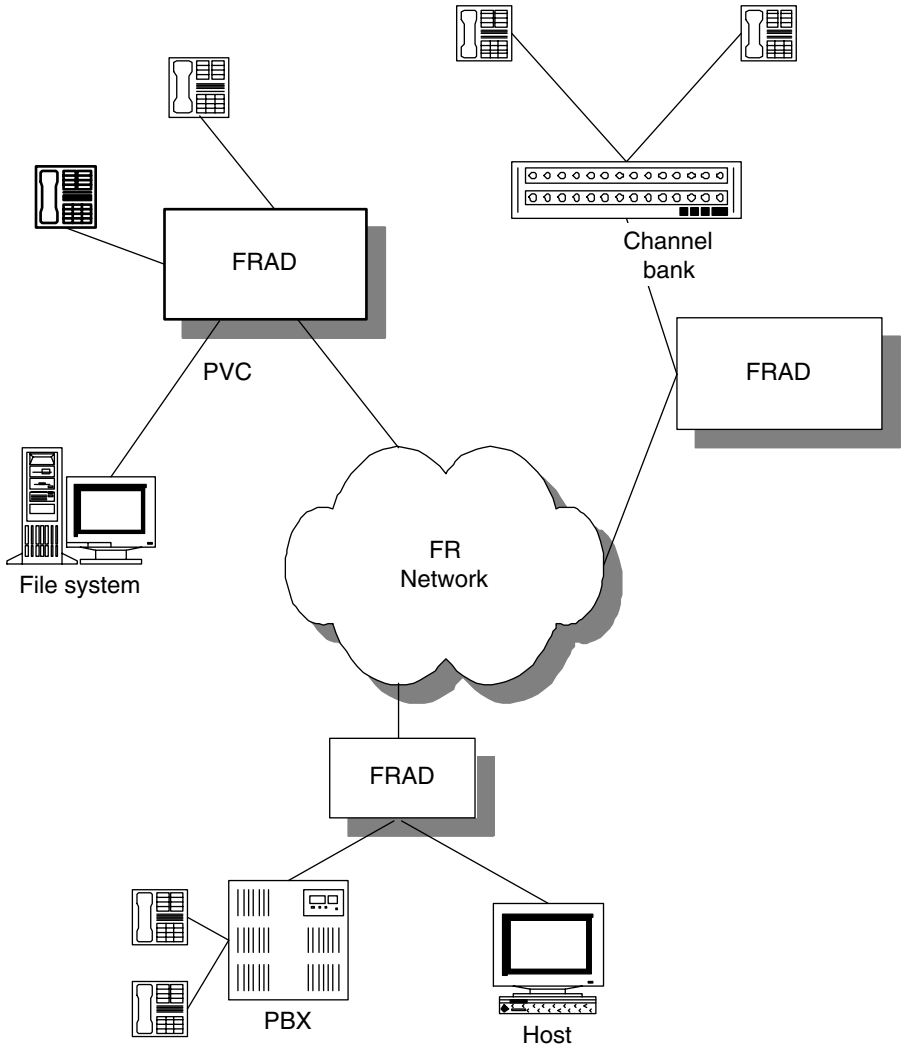


Figure 2-7
Voice over frame relay application scenarios.

customer's needs. For example, a customer can prioritize voice traffic over data traffic or vice versa.

VoFR-capable access devices such as FRADs are capable of converting between analog and digital signals and are also responsible for handling the fragmentation of frames to support voice traffic.

2.4.3.3 Challenges for VoFR VoFR service needs overcome a few key challenges in order to fully realize its potential of carrying voice traffic and combining data and voice onto a single, widely deployed frame relay

network. Frame relay was not originally designed for constant-bit-rate, time-sensitive applications like voice and video and does not have the ability to ensure that frame loss does not exceed a threshold. Nor can it synchronize clocks between sending and receiving devices. Another potential issue is loss of voice quality due to VoFR's use of voice compression. Also, it is not trivial to guarantee the quality or performance of voice traffic, given the current traffic control and prioritization or lack of it on frame relay networks. Finally, standardization of signaling protocols for call setup is another challenge.

Appendix. FRF Specifications for Implementation Agreements

Specification	Title
FRF.1 (1992)	User-to-network implementation agreement
FRF.2 (Aug. 1992)	Frame relay network-to-network interface implementation agreement.
FRF.3 (Dec. 1993)	Multiprotocol encapsulation implementation agreement
FRF.4 (Jan. 1994)	Frame relay user-to-network SVC implementation agreement
FRF.5 (Dec. 1994)	Frame relay/ATM PVC network interface implementation agreement
FRF.6 (Mar. 1994)	Frame relay service customer network management implementation agreement
FRF.7 (Oct. 1994)	Frame relay PVC multicast service and protocol description implementation agreement
FRF.8 (Apr. 1995)	Frame relay/ATM PVC service implementation agreement
FRF.9 (Jan. 1996)	Data compression over frame relay implementation
FRF.10 (Sept. 1996)	Frame relay network-to-network SVC implementation agreement
FRF.11 (May 1997)	Voice over frame relay implementation agreement
FRF.12 (Dec. 1997)	Frame relay fragmentation implementation agreement
FRF.13 (Aug. 1998)	Service level definitions implementation agreement
FRF.14 (Dec. 1998)	Physical layer interface implementation agreement
FRF.15 (Aug. 1999)	End-to-end multilink frame relay implementation agreement
FRF.16 (Aug. 1999)	Multilink frame relay UNI/NNI implementation agreement

Specification	Title
FRF.17 (Jan. 2000)	Frame relay privacy implementation agreement
FRF.18 (Apr. 2000)	Network-to-network FR/ATM SVC service interworking implementation agreement
FRF.19 (Mar. 2001)	Frame relay operations, administration, and maintenance implementation agreement
FRF.20 (June 2001)	Frame relay IP header compression implementation agreement

Note: For more about FRF specifications, go to the Frame Relay Forums Web site at www.frforum.com.

REVIEW QUESTIONS

1. Explain the layer of the OSI network reference model at which a frame relay network operates and the main function of the frame relay SAP.
2. Describe the main differences between the recommendations by ITU/ANSI and those by the Frame Relay Forum.
3. Describe how the field DLCI in the frame header is used and discuss the mechanism in the frame header that allows the extension of the addressing space.
4. Describe the differences between the frame relay permanent virtual connection and switched virtual connection in terms of how each type of connection is set up and taken down.
5. Briefly discuss the signaling protocol used to set up and take down an SVC across a frame relay UNI.
6. Discuss the types of applications that are more suitable for PVCs versus those that are more suitable for SVCs.
7. Explain the differences between the implicit and explicit congestion notification mechanisms of UNI signaling.
8. Discuss the motivations for using voice over frame relay service and the issues facing its use.
9. Describe what frame fragmentation is and what advantages it provides in supporting time-sensitive voice traffic.
10. Describe an application scenario for each of the two different approaches to frame relay-ATM interworking. When should the tunneling approach be used and when should the translation approach be used?

REFERENCES

ANSI. 1991a. "Digital subscriber system No. 1 DSS1 signaling specification for frame relay bearer service." ANSI T1.617. Web site: www.ansi.org.

ANSI. 1991b. "Digital subscriber signaling system No. 1 (DSS1)—Core aspects of frame protocol for use with frame relay bearer service." ANSI T1.618. Web site: www.ansi.org.

Bradley, T., Brown, C., and Malis, A. 1993. "Multiprotocol interconnect over frame relay." IETF RFC 1490. Web site: www.ietf.org.

Brown, C., Baker, F., and Carvalho, C. 1992. "Management information base for DTEs." IETF RFC 1315. Web site: www.ietf.org.

Frame Relay Forum. 1997. "FRF II: Voiceover Frame Relay Implementation Agreement."

Frame Relay Forum.1998. *The Basic Guide to Frame Relay Networking.* Web site: www.frforum.com.

ITU-T. 1992. "ISDN data link layer specification for frame mode bearer services." Geneva, Switzerland: International Communications Union-Telecommunications Standardization Sector. Recommendation Q922. Web site: www.itu.int/ITU-T/.

ITU-T. 1993. "Framework for frame mode bearer services." Recommendation I.122. Geneva, Switzerland: International Communications Union-Telecommunications Standardization Sector. Web site: www.itu.int/ITU-T/.

ITU-T. 1995. "Digital subscriber signaling system No. 1 (DSS 1)—Signaling specifications for frame mode switched and permanent virtual connection control and status monitoring." Recommendation Q933. Geneva, Switzerland: International Communications Union-Telecommunications Standardization Sector. Web site: www.itu.int/ITU-T/.

ITU-T. 1998. "ISDN User-Network Interface Layer 3 Specification for Basic Call Control." Recommendation Q931.

CHAPTER 3

ATM Networks

3.1 Introduction

This section provides background information on ATM. We will cover historical context, ATM standards, and a layered view of the ATM protocol stack.

3.1.1 Motivations for ATM

ATM technology was conceived, developed, and standardized with the grand goal of integrating voice, data, and video networks into a single network. This single network can perform the functions of the existing circuit-switched time division multiplexing (TDM) voice network and IP-based data networks and support new types of services such as video on demand.

ATM is built upon the existing technologies such as X.25 and frame relay. The concept of virtual connection found in X.25 technology is a cornerstone of ATM technology. Switching and multiplexing techniques used in frame relay network also find their way into ATM networks.

ITU-T first defined the concept of broadband ISDN (B-ISDN) in the late 1980s, designating ATM as the basis of B-ISDN in response to anticipated market demands. This at least in part spurred the serious efforts toward ATM standards development that followed. The large-scale deployment of ATM did not materialize until the mid- to late 1990s.

ATM is a technology standardized on a global scale. Two standards bodies, ATM Forum and ITU-T, among many other organizations that have contributed and continue to contribute to the ATM standardization efforts, play a key role in defining the base ATM standards. ITU-T defines the formal base standards and ATM Forum, an industrial forum dedicated to ATM technology, focuses more on the application, interoperability, and implementation level agreements.

3.1.2 Layered View of ATM

The ATM network protocol stack encompasses the bottom two layers of the OSI network reference model, providing the physical layer, data link layer, and part of the network layer functions, as shown in Fig. 3-1.

ATM protocol has a thin physical layer that specifies a mapping of ATM specific structures into the format of each physical layer-specific technology such as DS1, DS3, SONET OC3, etc.

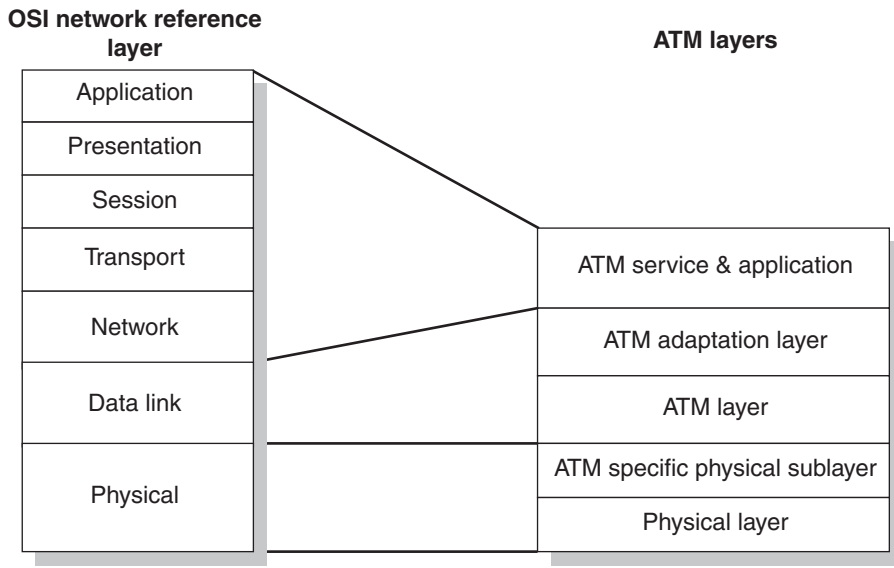


Figure 3-1
ATM protocol layers.

ATM is often referred to as a "layer-2 technology" because the ATM layer itself performs the data link layer functions by packaging user application data into a transportable unit known as a *cell* before transmitting it on physical wire. On top of the ATM layer is the ATM adaptation layer (AAL), which also performs part of the layer-2 functions by preparing ATM cells for the structures suitable for the specific applications and services supported by ATM.

3.2 ATM Basics

This section begins by describing a thin, ATM-specific, physical sublayer that links the ATM layer to the physical layer. Then it focuses on a set of key concepts for the ATM layer, including ATM cells, virtual connection, virtual path, virtual channel, and cell-based switching, before introducing the ATM layer quality of service (QoS) mechanisms.

3.2.1 ATM Physical Layer

The ATM physical layer is responsible for transporting ATM cells over an electrical or optical physical transmission medium connecting ATM

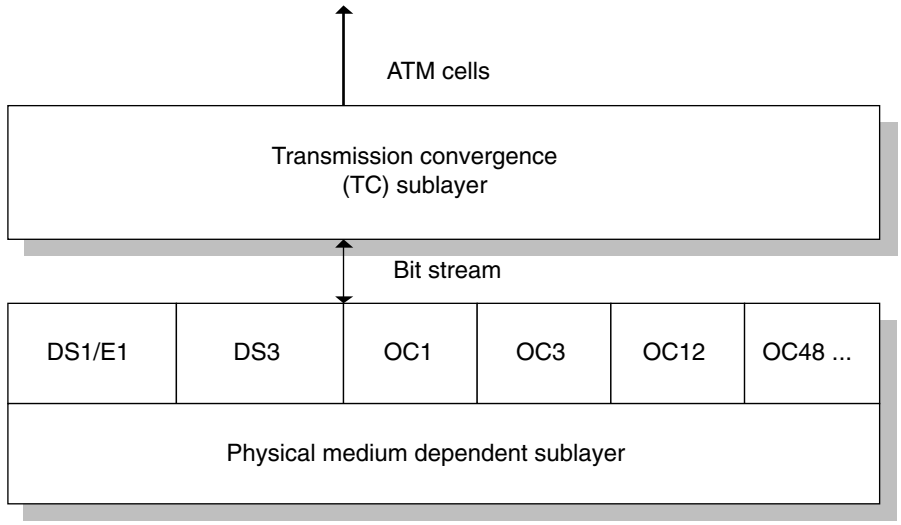


Figure 3-2
ATM physical layer.

devices. The physical layer, as shown in Fig. 3-2, consists of the physical media-dependent (PMD) sublayer and the transmission convergence (TC) sublayer.

The physical media-dependent sublayer represents the physical transmission medium such as DS1/E1, DS3, OC3, OC12, etc., and provides for the actual transmission of bits in ATM networks. This sublayer sees a bit stream and is unaware of anything specific to ATM technology.

The transmission convergence (TC) sublayer is responsible for mapping ATM cells to and from a bit stream provided by the PMD sublayer, and for delivering cells, including the 5-byte cell header, to the ATM layer at a speed specific to each transmission medium.

3.2.2 ATM Layer

The ATM layer is the core of the ATM technology that performs the bulk of the layer-2 functions of the OSI network reference model. The remainder of this section focuses on that layer, which performs the following ATM-specific functions (McDysan and Spohn 1999):

- It constructs cells from higher-layer application data and sends them down to the physical layer for transmission if the traffic originates from a local node.
- It receives cells from the physical layer and processes the cell header, including the cell header error checking.

- It switches and forwards cells using the virtual path identifier (VPI) and virtual channel identifier (VCI) in each cell, updating the cell header with new VPI and VCI values using the configured cross connects.
- It multiplexes and demultiplexes traffic at the cell level.
- It processes the cell headers in terms of cell loss priority, payload type, and the predefined reserved header values. The processing may include taking action on a value in the header such as updating a header value or performing congestion control.

3.2.3 ATM Cells

The concept of the ATM *cell* is at the core of ATM technology. An ATM cell is a data unit for transmission, switching, and multiplexing. In a railroad analogy, an ATM cell is like a cargo car used to transport goods, routed to a different destination, and grouped with other cars destined for different places.

An ATM cell has a fixed length of 53 bytes, a mathematical average of two competing proposals from European and North American standards bodies. Specifically, cell size is determined by the formula (64 + 32)/2 = 48 + 5 = 53 bytes. An ATM cell is made up of two parts, as shown in Fig. 3-3: a cell header of 5 bytes and cell payload of 48 bytes. The 5-byte header is for the ATM layer only. After receiving an ATM cell, the ATM layer examines and strips the 5-byte header and passes the 48-byte payload to the layer above.

The fixed length of ATM cells is motivated by two factors. First, the constant length makes the design of switching fabric hardware simple.

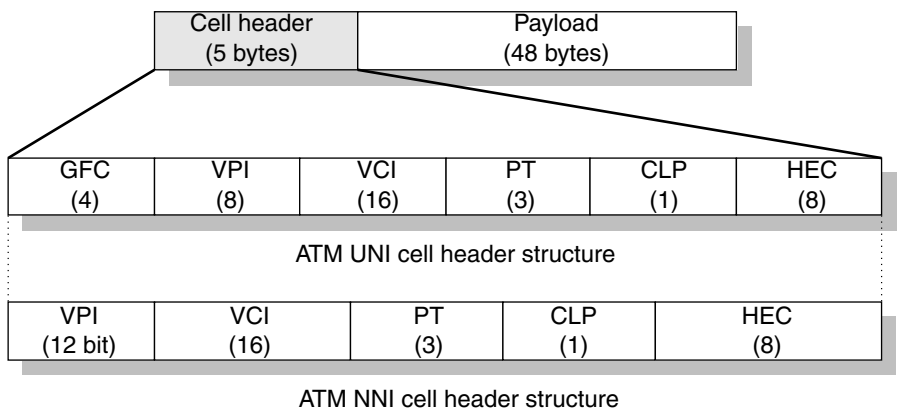


Figure 3-3
The ATM cell and cell header structure.

The variable length of packets such as IP packets requires buffering and more complicated processing than fixed-length cells. Second, it is easier to control the service quality with a fixed-length structure because delay and delay variation, two premier parameters of service quality, are more predictable.

The structure of a cell header is not a fixed one: It depends on where the cell comes from or the types of interfaces. There are two types of interfaces in an ATM network: user-to-network interface (UNI) and network-to-network interface (NNI). An ATM UNI interconnects user side equipment to network side equipment. ATM cells generated at a UNI interface have the cell header structure shown in the top half of Fig. 3-3, while the ATM cells generated at an NNI interface have the cell header structure shown in the bottom half of Fig. 3-3.

The two cell headers are identical, with two exceptions. First, the UNI cell header has a 4-bit generic flow control (GFC) field that is indented for controlling multiple flows from the user side that contend for network resources. Second, the VPI field of a UNI cell header has 8 bits while the same field of the NNI cell header has 12 bits. This reflects the fact that the capacity on the network side is generally much larger than that on the user side.

The generic definitions for the ATM cell header fields are as follows (Perros 2002).

GFC: a 4-bit field only present in the UNI cell header that is intended for controlling traffic flow entering the network from the user side.

VPI: an 8-bit or 12-bit field that uniquely identifies a virtual path across the interface. This field can support up to 256 virtual paths on a UNI (8 bits) and up to 4956 virtual paths (12 bits) on an NNI.

VCI: a 16-bit field that uniquely identifies a virtual channel within a virtual path.

Payload type (PT): a 3-bit field indicating the type of data in the payload field. The 3 bits are used as follows: The first bit indicates the user data or operations, administration, and management (OAM) data; the second bit indicates the upstream congestion condition; the rightmost is an AAL indication bit that is currently used by ATM adaptation layer 5 (AAL5) to indicate the last cell of a packet.

Cell loss priority (CLP): a 1-bit field indicating the loss priority of an individual cell. It can be set either at the user side or the network side; if set to 1, it means the cell has low priority and can be discarded in case of traffic congestion.

Header error checksum (HEC): an 8-bit field to detect any error in the cell header fields that may have occurred in the transmission process.

3.2.4 ATM Virtual Path, Virtual Channel, and Virtual Connection

ATM virtual path (VP), virtual channel, and virtual connection are the key ATM concepts built upon the concept of ATM cell. The basic idea is to flexibly build logical connections on a physical circuit to accommodate the different requirements of different applications. They provide the basic building blocks for ATM networks to build different levels of end-to-end virtual connections.

A physical circuit or line can be configured to have one or more virtual paths, and a virtual path can be configured to have one or more virtual channels. Their relationships are shown Fig. 3-4. A simple analogy is that a physical circuit can be viewed as a big pipe that can contain the next level of pipes (virtual path), which in turn can contain smaller pipes called *virtual channels*.

An end-to-end logical virtual path is termed a *virtual path connection* (VPC), as shown in Fig. 3-4. In the figure, node A is an access ATM node where the ATM traffic originates and node C is an exit node where the ATM traffic terminates. There can be one or more intermediate nodes in between. Thus the interface between node A and node B is a user-network interface, as is the interface between nodes B and C. A virtual path connection consists of one or more virtual path links (VPLs), while each virtual path link connects two nodes. As

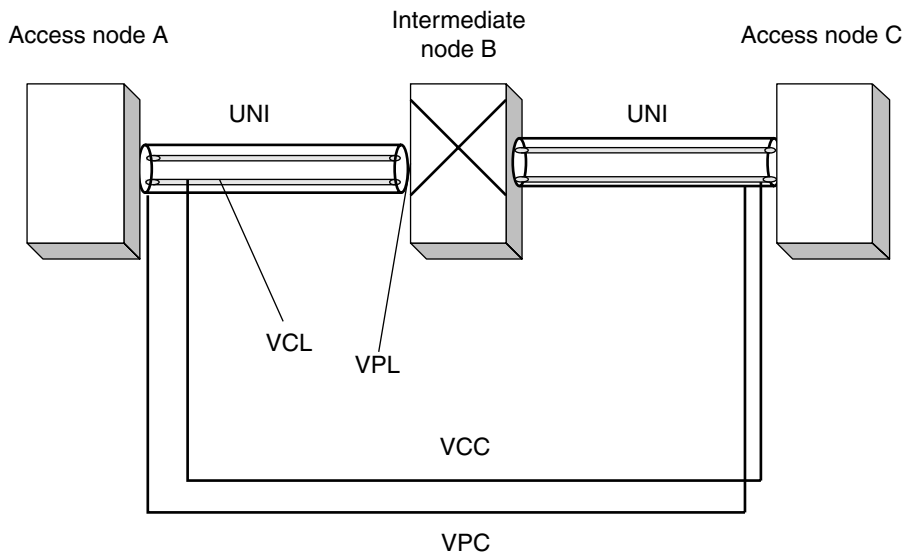


Figure 3-4
Illustration of VPL, VCL, VPC, and VCC.

shown in Fig. 3-4, a physical link connects two physical ports on two adjacent nodes, and a physical link can support multiple virtual path links. A virtual path link is identified by two virtual path link end points on the two connected physical ports.

Similarly, an end-to-end virtual channel is termed a *virtual channel connection* (VCC) and a VCC consists of one or more virtual channel links (VCL). A VCL is identified by two VCL end points identified by two virtual channel identifiers of the respective end points, as shown in Fig. 3-4.

A VPI or VCI that uniquely identifies a VPL or VCL on an interface is only significant locally. An end-to-end virtual path connection or virtual channel connection is not identified by a single identifier but instead by the combination of all component VPIs or VCIs. Since each VPI is unique on a port or an interface and each VCI is unique within a virtual path, the component VPIs or VCIs collectively uniquely identify a virtual path connection or a virtual channel connection across a network. It is up to the administrator of each ATM node to number a virtual path on a port and virtual channels within each virtual path, as long as VPI and VCI are within the defined ranges.

The lengths of VPIs and VCIs in cell headers determine the valid ranges of the VCIs and VPIs. A VPI length of 8 or 12 bits in the cell header determines that the maximum number of virtual paths a UNI physical port can support is no more than 256 and for an NNI port no more than 4984 (2^{12}). A VP can support no more than 2^{16} virtual channels. These are the absolute upper bounds of virtual paths and virtual channels, but in practice many ATM equipment vendors allow far fewer than the maximum. That is because, for a DS1 interface with a line rate of 1.55 Mbps, for example, it is not very practical to have hundreds or even thousands of virtual paths.

Some VPI and VCI values are reserved for special purposes by the ATM standards, and therefore not all VCI and VPI values can be used for user traffic. For example, VCI values 1 through 18 are reserved for signaling or OAM traffic.

A virtual connection can be configured as unidirectional or bidirectional. For a unidirectional virtual connection, traffic can flow only in one designated direction, while for a bidirectional virtual connection traffic flows in both directions.

A virtual connection can be configured as point-to-point or point-to-multipoint, depending on the target application. A point-to-point connection supports applications such as telephony service, tunneling for virtual private networks (VPNs), and dedicated data connection. A point-to-multipoint connection supports broadcast applications.

3.2.5 ATM Cross-Connect and Switching

A cross-connect at an ATM device interconnects two virtual path links or virtual channel links and makes it possible to build an end-to-end virtual connection. ATM cross-connects can take place at one of two levels: virtual path or virtual channel. Cross-connecting two virtual path links basically is a function of mapping one VPI on one port to another VPI on the same or different port. An example may help illustrate the concept of virtual path cross-connect: In Fig. 3-5, which shows an intermediate ATM node, the whole virtual path with VPI = 7 is cross-connected to the virtual path with VPI = 10.

Cross-connecting may take place at a virtual channel level that effectively maps a VCI of one virtual channel to another VCI on the same or different virtual path. As shown in Fig. 3-5, a virtual channel with VPI = 6 and VCI = 25 is cross-connected to a virtual channel identified with VPI = 8 and VCI = 85. All virtual channels within a virtual path follow the same route. Note that the cross-connects are statically configured at the provisioning time.

ATM cell switching is a process of receiving an incoming cell from one port and sending it out through an outgoing port, based on the VPI and VCI values of the cell. This process also includes the operation of replacing the incoming VPI and VCI values with the local values. For example, as video conference traffic arrives at a port with the VPI = 7 at

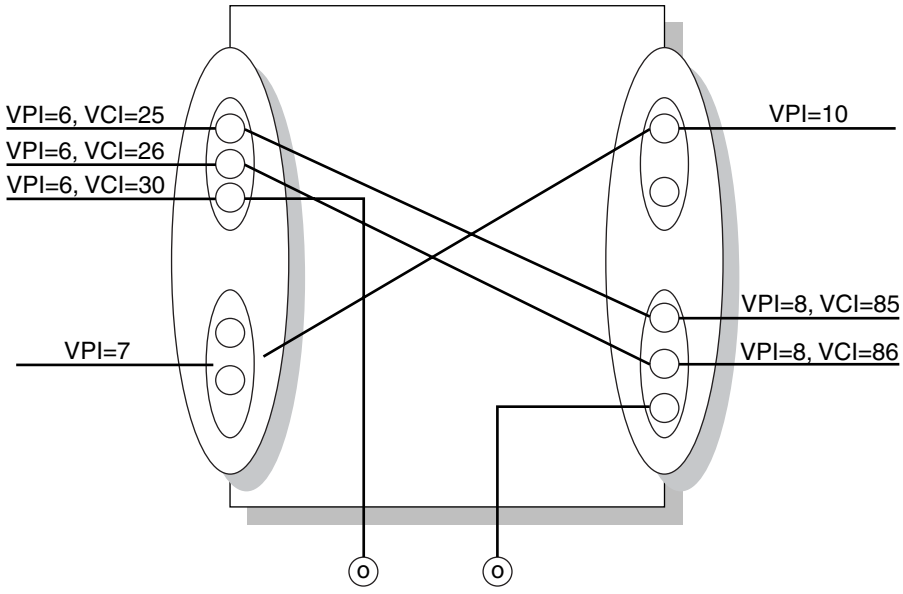


Figure 3-5
Illustration of ATM virtual path and virtual channel switching.

a node, as illustrated in Fig. 3-5, it is automatically switched onto the VP with VPI = 10, and the previous VPI = 7 is replaced with VPI = 10. For the virtual channel level switching, assume again that a phone call traffic arrives at the node with VPI = 6 and VCI = 25. The ATM node switches the data onto the virtual channel with VPI = 8 and VCI = 85 and updates the cell header with the new VPI and VCI values, using the already provisioned VC level cross-connects.

A virtual path or virtual channel can also terminate at a local node if the traffic on a VP or VC is destined for a user on the local node. As shown in Fig. 3-5, the virtual channel with VPI = 6 and VCI = 30 terminates at the local node. A new traffic stream may also originate at the local node.

3.2.6 ATM QoS

ATM defines an elaborate QoS scheme at the data link layer that is difficult to match by any other packet networking technologies. The scheme includes the concept of service category, ATM traffic parameters, and a set of ATM QoS mechanisms (Ibe 1997).

ATM Forum defines five service categories used to describe the types of service supported over ATM networks. The service categories are defined in terms of a set of bit rate parameters. The five service categories are the following:

Constant bit rate (CBR) service. For CBR service, a fixed amount of bandwidth is guaranteed. It supports a tightly constrained cell loss rate and cell transfer delay. This service is used to support real-time-sensitive applications that require a fixed amount of bandwidth such as telephone calls, real-time video, and circuit emulation services.

Real-time variable bit rate (rt-VBR). This service category requires a variable amount of bandwidth, constrained cell transfer delay, and delay variation. It supports real-time-sensitive applications that can tolerate some bit rate variation. Examples of applications include real-time voice and variable-bit-rate video.

Non-real-time variable bit rate (nrt-VBR). Delay variation is not controlled but throughput is guaranteed for this service category. Applications include file transfer and packet data transfer.

Unspecified bit rate (UBR). Also known as *best effort service,* UBR does not require constrained cell transfer delay and delay variation and provides no throughput guarantee.

Available bit rate (ABR). Cell loss can be limited with flow control but delay variation is not guaranteed for this service category.

The ATM Forum also defines a set of traffic parameters to describe ATM QoS from a customer's perspective. This allows a service level agreement to be created between a user and a service provider that specifies a set of QoS parameters the provider guarantees and the traffic characteristics that the user traffic will abide by. The following traffic parameters are defined for a worst-case, bursty traffic scenario:

Peak cell rate (PCR): the maximum number of cells per second an ATM device can transmit

Maximum burst size (MBS): the maximum number of cells an ATM device can transmit at the peak cell rate such that the average of many PCRs over a specified time interval does not exceed the specified sustainable cell rate (SCR)

Sustainable cell rate: the maximum sustainable, average rate at which an ATM device can transmit cells

Cell delay variation tolerance (CDVT): the number of cells for a given unit of time an ATM device can send; it effectively defines the number of consecutive cells that can enter a network for a traffic-policy-conforming connection

There are three major ATM layer mechanisms to achieve QoS. They include the following:

Traffic descriptor. A traffic descriptor consists of a set of traffic contract parameters including peak cell rate, sustained cell rate, maximum burst size, etc. A traffic descriptor can be associated with a virtual path. This allows an ATM node to drop the cells that do not conform to the agreed-upon traffic descriptor.

Cell loss priority. A cell header field that indicates the priority of a cell. Cells with lower priority may be dropped in case of congestion in favor of cells with higher priority.

Connection admission control (CAC). A sophisticated QoS algorithm for deciding whether to accept a connection based on the available resource at a node.

3.3 ATM Adaptation Layer

An ATM adaptation layer (AAL) adapts the raw cells received from the ATM layer into a format suitable for a particular type of service in one direction and maps the service-specific data into raw cells in the other direction. This section describes the five AALs that support different

types of services: ATM adaptation layers 1 (AAL1), 2 (AAL2), 3/4 (AAL3/4), and 5 (AAL5) (ITU-T 1993; ITU-T 1996a).

3.3.1 Overview of AAL

The ITU standards define five classes of service that the ATM adaptation layer supports, named Class A through Class D, as shown in Table 3-1. Note that the class of service concept here is more generic and encompassing than the service category introduced in the previous section, and it may be viewed as a QoS category.

All AAL sub-layers share a common structure that consists of two parts: segmentation and reassembly (SAR) and a convergence sublayer (CS), as shown in Fig. 3-6. Each part has its own protocol data unit (PDU) that consists of a header, a payload, and an optional trailer.

The SAR is mainly responsible for assembling ATM cells received from the ATM layers into a unit suitable for a higher-layer application. In the direction of transmission, it is responsible for segmenting an application-friendly data unit into the format that is suitable for the ATM layer to build into cells to be transmitted by the physical layer.

The convergence sublayer is further divided into two parts: a common part convergence sublayer (CPCS) and a service-specific convergence sublayer (SSCS). The CPCS is mainly concerned with error checking, protocol data alignment, cell delay handling, etc. The SSCS is an optional part and, if present, it handles the data processing specific to each target service such as timing synchronization for TDM circuit emulation service (CES).

3.3.2 AAL1

AAL1, one of the first AALs that ITU defined in the mid-1990s, is intended to support constant bit rate service that has a tightly constrained timing

TABLE 3-1

Class of Services and the Supporting AAL

	Class A	Class B	Class C	Class D
Supporting AAL	AAL1	AAL2	AAL3/4 or AAL5	AAL3/4 or AAL5
Bit rate	Constant	Variable bit rate		
Connection mode	Connection oriented			Connectionless
Applications	Circuit emulation	Packet video, audio	Variable rate data traffic	IP over ATM

Source: McDyson and Spohn (1999).

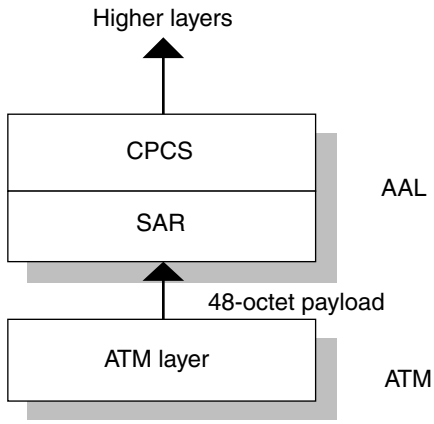


Figure 3-6
A high-level view of ATM adaptation layers.

requirement. One main target application is circuit emulation service, which allows an ATM network to emulate 64-Kbps or higher TDM circuits and carry traditional circuit-switched telephony service with the same high quality.

The AAL1 SAR sublayer has relatively light duty. Its main functions include the following:

- It converts 48-byte SAR PDU to 47-byte CS PDU using a 1-byte SAR header.
- It generates sequence numbering for SAR PDUs at the source and validates the received sequence numbers generated at the destination for bidirectional communications.
- It performs error detection and correction on the sequence numbers.

The convergence sublayer performs many detailed functions to support the transport of TDM circuits, video signals, and high-quality audio signals. They include the following:

- Handling of cell delay variation to deliver AAL1 PDUs at a constant bit rate
- Synchronous recovery of the source TDM clock frequency at the destination end using the synchronous residual time stamp (SRTS) method
- Asynchronous recovery of the TDM clock at the destination using only the received cell stream interarrival times or playback buffer fill
- Transfer of TDM structure information between a source and a destination

3.3.3 AAL2

AAL2, the newest arrival in the ATM AAL family of protocols, is intended to support connection-oriented, real-time-sensitive application such as packetized voice and video. AAL2 as specified in ITU-T Recommendation I.363.2 (ITU-T 2000) was approved in a very short time span in 1997 to meet the market demand. One AAL2 application is the voice over DSL using AAL2 as transport layer for carrying voice between a user device and central office equipment. The key features of AAL2 include the following:

- Support for multiplexing multiple users over a single virtual channel connection
- An 8-bit user channel identifier field, which can support up to 248 simultaneous users on a single virtual channel with some value reserved for maintenance purposes
- Support for variable length of the data field
- Support for variable bit rate of data transfer
- A mechanism for checking and handling user data errors
- Efficient bandwidth usage due to silence detection and suppression for voice traffic as well as idle channel deletion

AAL2 has a common part sublayer (CPS) and a service-specific convergence sublayer, as shown in Fig. 3-7. An operation example will help illustrate how AAL2 works at a conceptual level. Assume that multiple users are making voice calls over one virtual channel at the same time. The voice data arrives at the AAL2 layer from the application layer. Then the following occurs:

1. The SSCS sublayer first converts the voice data into SSCS PDUs, encoding compressed voice information and other application-specific information into the PDUs; then the SSCS PDUs are passed down to the CPS sublayer.
2. The CPS sublayer first builds a CPS packet header and a CPS packet by prefixing a SSCS PDU with a 3-byte header that contains a channel ID identifying an individual user on the VCC.
3. The CPS sublayer collects CPS packets over a specified interval and forms CPS PDUs that each consists of 48 bytes' worth of CPS packets. There is a start field (STF) in the CPS PDU that points to

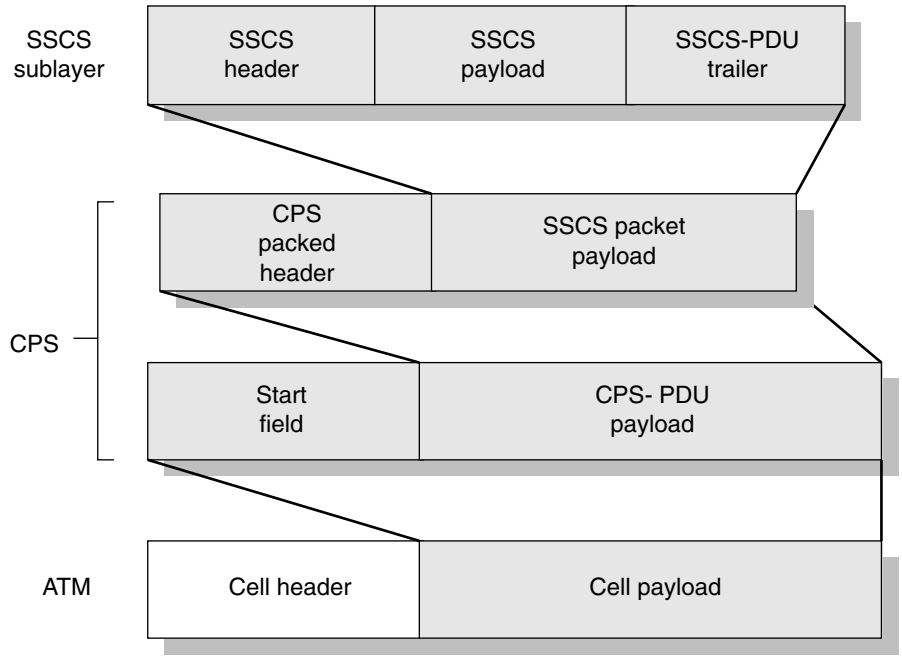


Figure 3-7
AAL2 structure. (McDysan and Spohn 1999)

the next byte that contains the active voice data, skipping the inactive part (e.g., silence in conversation).

4. Finally, the CPS sublayer maps CPS PDUs to the 48-byte ATM cell payload and passes them to the ATM layer, which adds cell headers before sending them to the physical layer for transmission.

3.3.4 AAL3/4

AAL3 and AAL4 originally targeted connection-oriented and connectionless variable bit rate (VBR) services, respectively (ITU-T 1996b). It was realized later that the features they have in common are sufficient to warrant combining the two protocols into one. Up to now, AAL3/4 has not been as widely deployed as other AAL protocols.

AAL3/4 is intended to support packet data service, both connection-oriented and connectionless. AAL3/4 protocol layers conform to the general AAL protocol model as illustrated in Fig. 3-6.

One important function the AAL3/4 SAR sublayer performs is to support multiplexing of up to 1024 logical connections onto a single

ATM virtual channel connection. This allows the multiple users to share the same VCC.

The following scenario illustrates the operations of the AAL3/4 SAR and CPCS sublayers. Assume a sequence of AAL3/4 cells arrives at the ATM layer. Then the following happens:

1. A set of 48-byte cell payloads on the same VCC is passed to the AAL3/4 SAR. In the AAL3/4 SAR header, there is a field indicating the beginning of a message (BOM), the continuation of message (COM), or the end of the message (EOM).
2. When the SAR receives the cell payload that contains the EOM mark, it checks the sequence number, handles errors if any, reassembles the message in packet format, and then passes the reassembled packet to the CPCS sublayer.
3. The CSCP sublayer further packages the data into the interface data unit suitable for the higher layer application and sends them either in a batch or one at a time to the higher layer.

3.3.5 AAL5

AAL1, AAL2, and AAL3/4 are deemed too complicated to support simple data service efficiently. This motivated the development of AAL5, or the Simple Efficient Adaptation Layer (SEAL) as it was originally named. AAL5 is intended to support simple packet data services such as IP packet data that do not have highly constrained delay and delay variation requirements. AAL5 is most suitable for supporting "best-effort" IP data service.

The following scenario illustrates the main operations of the AAL5 SAR and CPCS sublayers. Assume that a user receives an email on an ATM LAN. Then the following occurs:

1. A set of 48-byte cell payloads on a VCC is passed to the AAL5 SAR from the ATM layer. The AAL5 SAR simply checks for the AAL_indicate field in the payload type field of a cell. A nonzero value indicates the cell is the last one of the message that occupies multiple cells and reassembling can begin.
2. The CSCP sublayer builds a CPCS PDU that can have a payload of 1 to 2^{16} bytes. Once the message is re-assembled and padded if necessary, it is passed to the higher-layer application.

3.4 ATM Interfaces and Signaling Protocols

ATM signaling protocols are used mainly to set up virtual connections dynamically across two different types of interfaces. This section introduces the protocols and the two types of interfaces.

There are two types of virtual connections: permanent virtual connection (PVC) and switched virtual connection (SVC). A PVC is statically set up via a manual process. It is permanent because the connections remain in effect until they are manually taken down. The manual setup of a large number of PVCs can be error-prone and tedious and, more importantly, does not scale in large networks.

An SVC is a connection that is dynamically set up and remains in effect for the duration of a call. Once a call is completed, the system automatically tears down the virtual connection. SVCs are more suitable for dynamic applications such as telephony service and video conferencing.

ATM signaling protocols are designed mainly for setting up SVCs. Different network interfaces require different signaling protocols for the SVC setup.

There are two types of interfaces across an ATM network. One is the user-to-network interface (UNI), which is an interface between a user device and a network node. The other is network-to-node or network-to-network interface (NNI), which is an interface between a network node and another node or between one ATM network and another ATM network.

3.4.1 SVC Basics

A few concepts need to be clarified before introducing the SVC signaling protocols and UNI and NNI interfaces.

3.4.1.1 SVC Connections An SVC can be either unidirectional or bidirectional. A unidirectional virtual connection allows traffic to travel only in one direction and is also known as *simplex virtual connection*.

A bidirectional virtual connection allows traffic to go in both directions and is also known as *duplex virtual connection*. A bidirectional SVC actually is a pair of unidirectional SVCs, a forward connection from a calling party to a called party, and a backward connection from the called party to the calling party. The two component connections can

have different traffic parameters with different bandwidths to suit the application needs.

A point-to-point SVC connects one calling party with one called party, and the connection can be either simplex or duplex.

A point-to-multipoint SVC has one root node and one or more leaf nodes. There is one signaling channel between the root node and each leaf node. The user traffic flows only in one direction, from the root to leaf nodes, as specified in UNI 3.1 (ATM Forum 1994), though the backward flow is provided for maintenance traffic. The point-to-multipoint SVCs have been developed to support broadcast type of applications.

Soft permanent virtual channel (SPVC) is a connection between two ATM end hosts that consists of PVCs between an ATM end host and an ATM switch and SVCs between two ATM switches.

3.4.1.2 ATM Address An ATM address uniquely identifies an ATM interface across an entire ATM network. An ATM interface, also known as a *port,* represents a physical entity that must be globally addressable for the signaling purpose such as identifying a calling or called party. One interface must have at least one ATM address but more than one address can be assigned to the same interface.

There are two types of ATM addresses: data network-oriented network service access point (NSAP) address that uses the syntax of ATM end system address (AESA) and the telephony network-oriented ITU E.164 address (ITU-T 1997).

E.164, also commonly known as telephone number, consists of up to 15 digit numbers divided into two parts: country code (CC) and nationally significant number (NSN). The country code contains one to three digits, and the NSN is a domestic phone number such as 972-111-2222, with the first three digits representing an area code, also known as number plan area (NPA) and the second three digits representing a central office code in North America.

The AESA format uses the NSAP address syntax (ISO/IEC 1988), a data network-oriented address defined by the International Organization for Standardization (ISO). The general AESA format has a total of 20 octets divided into two parts, the initial domain part (IDP) and the domain-specific part (DSP), as shown in Fig. 3-8. There are three specific AESA address formats, and the format of each depends on the specific AESA address format. The IDP has two parts: the authority and format identifier (AFI) and initial domain identifier (IDI). The AFI identifies the format of the remainder of the address while the IDI identifies the network address assignment authority. The DSP has a high-order DSP

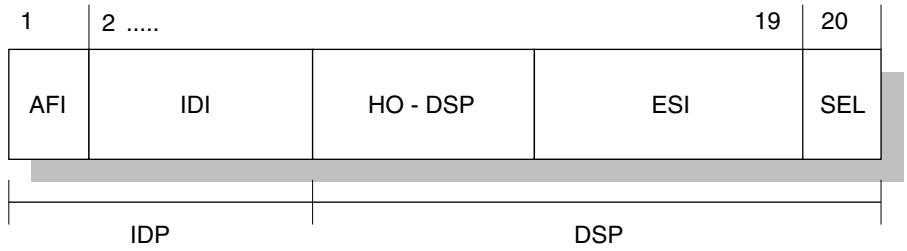


Figure 3-8
The general format of AESA address format (Source: Ref. 10).

(HO-DSP) and low-order DSP part that consists of an end system identifier (ESI) and a selector (SEL) byte. The lengths of both IDP and DSP vary, depending on the type of AESA address.

3.4.2 Introduction to ATM UNI

The ATM UNI specifications define a set of signaling protocol messages and procedures for the message exchanges across ATM UNI interfaces.

3.4.2.1 ILMI Integrated local management interface (ILMI) is a standard interface defined by ATM Forum to allow a user system and a network node to exchange address and management information over a UNI interface (ATM Forum 1996a). It allows a network node to pass to a user system the valid address prefixes for a particular logical ATM UNI, using a management protocol such as simple network management protocol (SNMP). The user system then can build its address by suffixing the address prefix with the ESI and SEL fields and then register its complete address with the network address registration authority. This is a key component for the automatic configuration of an ATM network using ATM SVC.

In addition to the address registration, ILMI also allows a user system to communicate the fault and performance data to the associated ATM network.

3.4.2.2 ATM UNI Overview There are two ATM UNI specifications that are widely in use: ATM UNI 3.1 (ATM Forum 1994) and ATM UNI 4.0 (ATM Forum 1996b). They are more or less based on ITU Q.2931 and related recommendations. Initially two sets of signaling standards, ITU's Q.2931 (ITU-T 1995) and ATM Forum's UNI 3.0, were developed separately without much consideration for interoperability. Then the ATM Forum UNI 3.1 specification became more compatible with the ITU Q.2931 specification. Then, with practical implementation and industrial needs

in mind, ATM Forum based its 1996 UNI 4.0 largely on ITU Q2931 ITU's related Q series recommendations.

The ATM Forum UNI 4.0, which continues to evolve in response to industrial needs, supports the following SVC signaling functions:

- On-demand (switched) connection setup
- Both E.164 and NSAP-based address support
- Switched virtual path service
- Point-to-point connection setup and release
- QoS class request and connection parameter negotiation at setup time
- VPI/VCI selection and assignment
- Support of supplementary service such as call waiting, caller ID presentation, subaddressing, etc.
- Support for narrowband ISDN interworking
- Support multiple signaling channels over one physical UNI
- Available bit rate signaling for point-to-point SVCs

3.4.2.3 Signaling Messages UNI 4.0 defines five sets of signaling protocol messages for the purposes of connection setup, tearing down, and connection maintenance for point-to-point and point-to-multipoint SVCs:

Point-to-point call setup message. These are the messages exchanged between a calling port and a called port to set up a point-to-point SVC. They are based on the ITU-T Q2931 signaling messages to support a two-way hand-shake signaling procedure. The messages include setup, alerting, call proceeding, connect, and connect acknowledgement.

Point-to-point call tearing down messages. These are the messages for taking down a point-to-point SVC. The messages include release and release complete.

Point-to-point status message. These are the messages for querying for the status of a SVC once a point-to-point SVC has been set up. The messages include status inquiry, status, and notify.

Point-to-point signaling link management. These messages support the maintenance of point-to-point SVC. The messages include restart and restart acknowledgement.

Point-to-multipoint connection control. These are the messages used to add or drop a party in a point-to-multipoint SVC. The messages include add party, add party acknowledge, add party reject, party alerting, drop party, drop party acknowledge, leaf setup request, and leaf setup failure.

Each signaling message contains a number of information elements. Some IEs are mandatory while others are optional. There are four generic mandatory IEs for all message types:

- Protocol discriminator
- Call reference
- Message type
- Message length

Other IEs can be optional or mandatory, depending on the type of the message. For example, a SETUP message must have the following mandatory IEs to request a SVC setup:

Broadband bearer capability: an ATM service category such as CBR, rt-VBR, nrt-VBR, UBR, or ABR as described above

Called party number: either an NSAP-based or E.164 ATM address

ATM traffic descriptor: traffic parameters defining the characteristics of traffic on the SVC such as PCR, SCR, and MBS, as described earlier

3.4.2.4 Signaling Procedure The UNI 4.0 signaling protocol defines a set of procedures for exchanging protocol messages between a user device and a network node for SVC call setup, teardown, and status query. The protocol defines the conditions and the sequence of messages the parties involved must follow in sending protocol messages and the procedure for handling error conditions that may occur in the message exchange process.

3.4.3 ATM NNI

The other type of ATM interface, network-to-network or network-to-node interface, is used to set up an SVC between two interior nodes of the same ATM network or between two different ATM networks. In addition to SVC setup, some NNI also perform the function of SVC call routing.

Over the short history of ATM, a number of NNIs have been defined by standards bodies to fulfill different needs. Some of the interfaces are complicated while others are simple. Some enjoy wider deployment than others. The defined NNIs include

Interim Interswitch Signaling Protocol (IISP) by ATM Forum. IISP is a simple protocol, intended to fill the interoperability need of ATM networking before the PNNI standards became mature.

Broadband Intercarrier Interface (B-ICI) by ATM Forum. B-ICI defines the signaling between ATM networks of different carriers, such as the local exchange carrier (LEC) and the interexchange carrier (IXC). B-ICI version 2.0 is defined in parallel to UNI. 3.1 and is not aligned with ATM Forum UNI 4.0.

Broadband ISDN Service User Part (B-ISUP) by ITU. B-ISUP defines a signaling protocol at the network-node interface and matches well with ITU's specifications for the signaling protocol (Q.2931) at the user-network interface in the capabilities it supports.

Private Network-Network Interface (PNNI) by ATM Forum. PNNI aims to support the SVC setup across a multivendor ATM network and automatic network topology discovery for SVC call routing.

The remainder of this section introduces PNNI, one of the most widely deployed NNIs.

3.4.4 PNNI

PNNI was developed to emulate the Ethernet-like "plug and play" user experience of the IP network in the ATM world with automatic topology discovery and wide interoperability. PNNI for all practical purposes can be viewed as consisting of two protocols: a signaling protocol and a routing protocol.

3.4.4.1 PNNI as Signaling Protocol PNNI as signaling protocol uses the ATM Forum UNI 4.0 signaling specification as a basis, augmenting it with additional capabilities of source routing and the ability to revert back to earlier nodes in order to route around an intermediate node that blocks a call request.

The PNNI specification version 1.0 provides the signaling capability for point-to-point SVC setup and release across node-to-node interfaces, with support for both E.164 and NSAP ATM addresses.

3.4.4.2 PNNI as Routing Protocol PNNI defines a logical network hierarchy as the basis for routing. Arguably the PNNI hierarchy is the most complicated part of PNNI, and understanding the motivations behind it helps put it into a proper context.

PNNI routing is a source-based routing as opposed to the hop-by-hop routing seen in the IP network. This choice is based on the experience learned from the IP routing algorithms with global scalability in

mind. The key to source-based routing is having a network-wide topology at each node, and the PNNI logical hierarchy is intended to do just that.

The PNNI hierarchy consists of multiple levels of peer groups. A peer group is a set of PNNI nodes that share the same node state information. A node in a leaf-level peer group represents a physical node or a device. A peer group at the leaf level can have no more than 256 nodes. One way of organizing a peer group is to have an ATM switch and all the ATM devices connected to the switch form one peer group.

Each peer group has a peer group leader (PGL), as shown in nodes A1.3 and A2.1 in Fig. 3-9. Each PGL represents this peer group in the next level up. The peer group leader is selected by the peer nodes within the group based on the configured leadership priority and the node identifier of each node. A set of PGLs forms a parent peer group and, in a recursive manner, a PNNI hierarchy is built. PNNI defines a procedure to automatically build a PNNI hierarchy. Figure 3-9 shows two levels of a PNNI hierarchy. Note that the PNNI nodes in the parent peer groups are logical nodes.

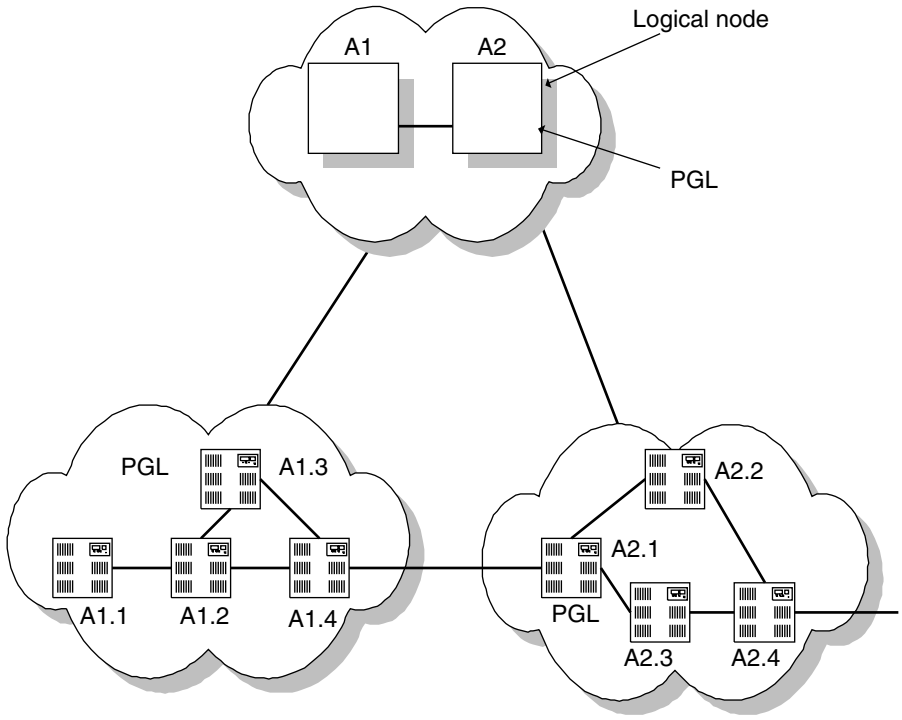


Figure 3-9
An example of PNNI hierarchy with two levels of nodes.

A PGL acts like a go-between between a peer group and its parent group. It has the responsibility of passing on the reachability and status information of the child peer group to the parent peer group nodes. In the same manner, each PGL feeds the information about the nodes of the parent group down to the child peer group nodes. In this manner, each node has the topology of the entire network.

The PNNI source-based routing is straightforward, with each node having network wide topology data: PNNI uses a link-state-based routing algorithm, the same algorithm used in the IP Optimal Shortest Path First (OSPF) (see Appendix A for a detailed description of OSPF). A major difference is that PNNI supports QoS parameter-associated link status for routing and the routing decision is made based on QoS parameters such as available bandwidth and traffic congestion condition on a link.

3.5 ATM Applications

The focus of ATM applications has evolved, along with market conditions and demands, from initial enterprise LAN, to service provider's backbone network, to the public access network used to support broadband access networks like various digital subscriber lines (xDSLs) and wireless networks.

ATM technology is designed to support a variety of applications and services. For the convenience of description, the supported services can be put into three general service categories: voice-based telephony service, data services, and miscellaneous services.

3.5.1 Voice Over ATM

There are four options for carrying voice over ATM networks that have been standardized. Out of the four approaches, as shown in Fig. 3-10, three approaches aim to provide ATM alternatives on the access network side while the ATM trunking option provides an alternative on the network side. Note that on the access network side, none of the solutions directly interface the end user device. The four options include

- Circuit emulation service
- ATM trunking using AAL1 or AAL2

- Loop emulation service (LES) using AAL2
- Voice over IP over AAL5

3.5.1.1 Circuit Emulation Service CES, an early attempt to use ATM networks to provide voice service, uses ATM virtual channel to emulate the TDM 64-Kbps or higher circuits on an ATM network with the same high quality of service.

CES uses AAL1 and the constant bit rate (CBS) service category. That is, a fixed amount of bandwidth is allocated per emulated circuit. CES supports both structured mode and unstructured mode circuit emulation. The former provides up to $N \times$ 64-Kbps circuits where N is 24 for DS1 and 32 for E1. The latter provides a clear channel pipe at the rate of 1.544 Kbps for DS1 or 2.04 Mbps for E1.

The CES interWorking function (IWF) is a key component of CES that enables traditional TDM devices to connect to an ATM switch. It normally resides on customer premise equipment like PBX or T1/E1 multiplexer and allows provisioning of the T1 service over an ATM virtual connection as shown in the upper left corner of Fig. 3-10.

Dynamic bandwidth circuit emulation service (DBCES) is a variant of CES with some distinguishing features such as silence detection and dynamic bandwidth utilization.

One limitation of CES and DBCES is they allow only one circuit per virtual connection. This limits the maximum bandwidth utilization

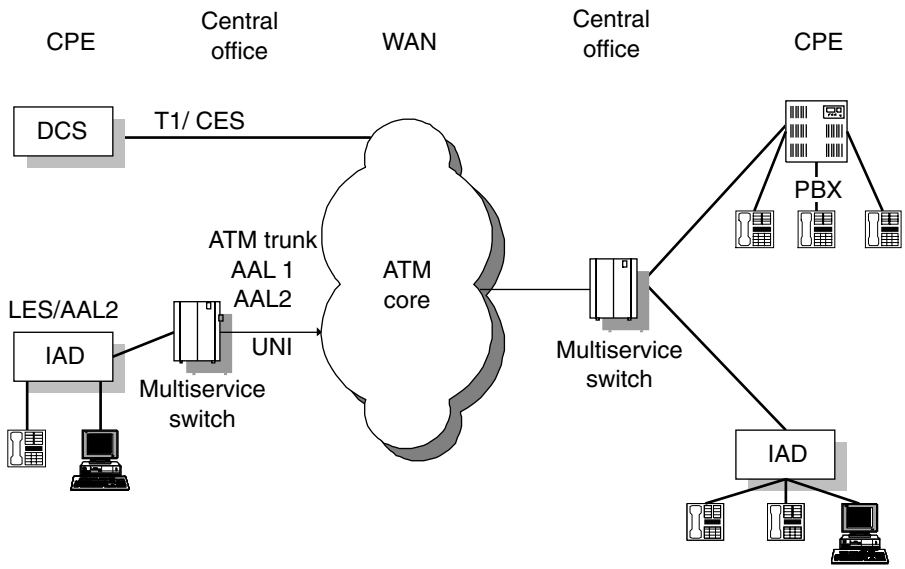


Figure 3-10
Voice over ATM application.

and has motivated the development of other approaches to voice over ATM capable of multiplexing multiple user connections onto one virtual connection.

3.5.1.2 ATM Trunking Service ATM trunking service provides circuits, or "trunks" as they are called in the telecom world, across the ATM backbone that connects two TDM networks. There are AAL1-based ATM trunking and AAL2-based ATM trunking.

AAL1 ATM trunking can be viewed as an extension of the circuit emulation service to a core ATM network that interconnects two narrowband TDM networks. It enables the transport of 64-Kbps voice circuits over an ATM network. A key component of the AAL1 trunking service is a narrowband service IWF that performs mappings between TDM and ATM as shown in Fig. 3-11.

AAL2 trunking service provides support for both switched trunking and nonswitched trunking (ATM Forum 1999). Switched trunking uses SVC and allows the routing of a call to an appropriate AAL2 channel on an on-demand basis. Nonswitched trunking allows only a one-to-one fixed relationship between a narrowband channel and an AAL2 channel. In addition, the AAL2 trunking provides the following capabilities:

- Variable-rate speech encoding
- Idle channel detection and removal
- Silence suppression
- Transport of in-band digital tone multifrequency (DTMF), or dial tone, signals over packets

3.5.1.3 Loop Emulation Service ATM Forum's LES specifications were completed in mid-2000 to meet the needs of broadband access networks such as xDSL and hybrid fiber-coax (HFC) cable (ATM Forum 2000). The LES specifications describe a procedure and the signaling required to support the transport of voice band service across an ATM network using AAL2 to provide a circuit-like voice channel. Since the LES also provides a data channel, it is also interchangeably called broadband LES (BLES) in the xDSL specification of ATM Forum.

Architecture-wise, as shown in Fig. 3-12, a BLES network, derived from the ATM Forum LES reference model, has two main components: a customer premises interworking function (CP-IWF) and a central office interworking function (CO-IWF). These functional modules can be implemented as standalone devices or functional modules inside another device such as an integrated access device (IAD) or a multiservice switch.

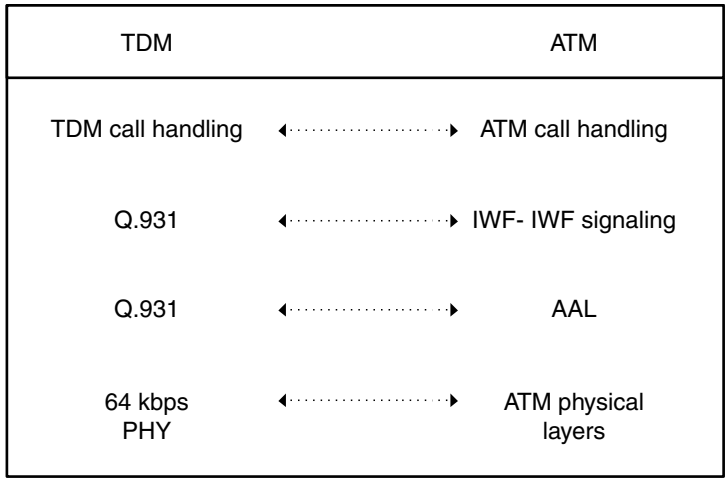


Figure 3-11
The AAL1 trunking IWF functional components.

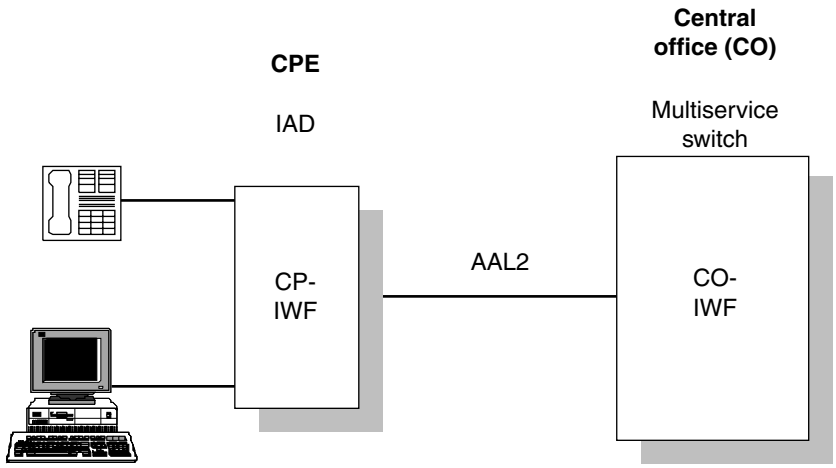


Figure 3-12
An overview of BLES model.

In the specifications, LES supports PVCs, SPVC, or SVCs, but in deployment most implementations so far use only PVCs.

The CP-IWF, which often resides at an IAD, is responsible for converting a TDM channel from the user to a AAL2 virtual channel and for signaling between a customer premise device and a central office switch. For signaling, it currently supports channel associated signaling (CAS) as well as common channel signaling (CCS), though the latter is only at the initial stage of deployment.

The CO-IWF, which often resides at a central office multiservice switch, is responsible for signaling to an ATM network as well as to

customer premise devices. It operates in one of two modes: concentrated and nonconcentrated. In the concentrated mode, the allocation of AAL channels is carried out dynamically and multiple user connections are multiplexed onto a small number of AAL2 channels. In the nonconcentrated mode, each timeslot of the TDM channel is assigned a specific AAL2 channel.

The AAL2 LES supports voice compression, silence suppression, and idle channel cancellation that can further improve the bandwidth efficiency on top of AAL2's efficient bandwidth utilization of variable bit rate service.

3.5.1.4 Voice Over IP over ATM Voice over IP over ATM is another option for carrying voice over an ATM network. Compressed-voice IP packets are carried over AAL5 on an ATM network. Voice signaling for call setup can be achieved with signaling protocols such as Session Initiation Protocol (SIP) or Bearer Independent Call Control (BICC) (Bellcore 2000).

Table 3-2 compares the discussed options for carrying voice over ATM.

3.5.2 IP Data Service over ATM

Part of the original design goal of ATM networks was to provide IP data service. Two widely used approaches to carrying IP over ATM are IP packets over AAL5, also known as *classical IP over ATM,* and next-hop address resolution protocol (McQuerry and McGrew 2001).

3.5.2.1 Classical IP over ATM There are two basic issues involved in carrying IP traffic over ATM: encapsulation of IP packets inside ATM

TABLE 3-2

Comparisons of Options for Voice over ATM

	Voice compression	Silence removal	Idle channel suppression	Support for SVC	AAL used
CES	No	No	No	No	AAL1
DBCES	No	No	Yes	No	AAL1
LES	Yes	Yes	Yes	Yes	AAL2
ATM trunking	No: AAL1 Yes: AAL2	No: AAL1 Yes: AAL2	Yes	Yes	AAL1, AAL2
Voice over IP over ATM	Yes	Yes	No	No	AAL5

ALL5 cells and translation between the IP address and ATM address (Laubach and Halpern 1998).

IETF addresses the encapsulation issue in RFC 1626 by defining a procedure and the maximum transfer unit (MTU) for the mapping between IP packets and AAL5 PDUs (Atkinson 1994). The default MTU size is 9180 bytes, aligning it with the MTU size of switched multimegabit data service (SMDS), another popular scheme for carrying IP packets using HDLC framing over backbone networks.

The second issue of supporting IP data over ATM is resolution of an IP address to a corresponding ATM address. IETF defines an ATM Address Resolution Protocol (ATMARP) in RFC 1577 to map an IP address to an ATM address and to forward IP packets over ATM networks. The basic idea of ATMARP is to define the concept of logical IP subnet (LIS) and use an address server. A group of IP routers and hosts that belong to the same IP subnet sharing the same subnet address and that connect to the same ATM network form a logical IP subnet as the basic unit for address resolution. An address server maps an ATM address to an IP address or vice versa.

The classical IP over ATM does not support "cut-through" routes that bypass intermediate router hops for communications between nodes on the same ATM network but within two different LISs. An additional limitation is that the ATMARP server is a single-point failure as it is currently specified.

3.5.2.2 Next-Hop Resolution Protocol The Next-Hop Resolution Protocol (NHRP) was developed by IETF to overcome the limitation of classic IP over ATM, that is, the latter's lack of cut-through routing between two ATM nodes that are in the same network (Laubach and Halpern 2000).

The basic idea of NHRP is similar to that of classic IP over ATM. In place of LIS, NHRP defines a nonbroadcast multiaccess (NBMA) network to include not only ATM, but also similar nonbroadcast networks such as frame relay and X.25 networks. Each NBMA consists of a set of nodes that are not physically or administratively restricted from communicating directly with each other. Another concept employed in NHRP is the administrative domain. An administrative domain can be a single NBMA, or a large NBMA can be partitioned into multiple administrative domains.

Within a NBMA, there is a set of NHRP servers (NHS), similar to ATMARP servers, and each HNS maintains a "next-hop resolution" mapping table with IP to ATM address mappings of all nodes associated with this NHS. A source node, in need of transmitting data to a destination node within the same network, goes through NHSs in a next-hop fashion

to find an IP to ATM address mapping and then sets up a cut-through SVC to the destination node using the found ATM address.

3.5.3 Other ATM Services

Under the umbrella of other ATM services are included LAN emulation (LANE) service, ATM video on demand (VOD), and ATM interworking with frame relay.

3.5.3.1 LAN Emulation Service LANE is one of the early targeted services of ATM technology. Because of the vast installed base of Ethernet and rapid advances in the Ethernet technology, ATM-based LAN has met only limited success thus far.

ATM LANE emulates a local area network on top of an ATM network. Specifically the LANE protocol defines mechanisms for emulating either an Ethernet or Token Ring LAN. The LANE protocol defines a service interface to the higher layer (i.e., the network layer) and operates in a so-called overlay model where data sent across the base ATM network is encapsulated in an appropriate LAN MAC packet such as an Ethernet packet, as shown in Fig. 3-13.

A simple operation scenario helps illustrate ATM LANE. Assume that an application at a host generates IP packets destined for another application at another host on the same emulated LAN. Then the following happens:

1. The MAC layer of the source host sends the first MAC frame to the ATM layer of the same node that translates the MAC address to an ATM address.
2. The source node, once it has received the destination host ATM address, establishes a data SVC to that host using an established UNI signaling protocol. Then data transfer can proceed from the source to the destination node.

3.5.3.2 ATM Video on Demand (VOD) Service An ATM network is a good option as a transport network for VOD service because of stringent QoS requirements and complicated control functions of VOD service. Both ATM Forum and ITU have developed specifications on VOD network reference models and methods for encoding and carrying video, audio, and data on ATM virtual connections.

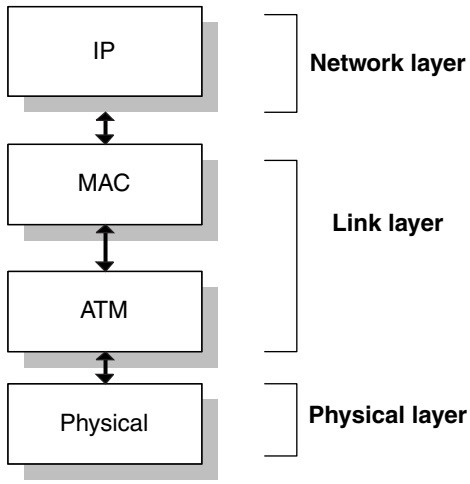


Figure 3-13
ATM LANE overlay model.

In a reference configuration, an ATM network connects a user VOD device (e.g., a set-top box) and a remote content server that provides VOD contents. The user initiates a VOD session that results in SVC being established between the user device and the content server. The user should have VOD session level control functions such as "pause," "fast forward," and "rewind," and should also have interactive experience.

ATM Forum and ITU specifications define how connection control, video, audio, data, and user control streams are encoded and multiplexed over ATM virtual channel connections, and there can be up to three such VCCs using AAL5. Also defined are the encoding standards for bit error rate (BER) delay, and jitter for VOD services.

3.5.3.3 ATM Interworking with Frame Relay ATM interworking with frame relay is an important service because frame relay has a very large installed base around the world. There are two approaches to interworking between an ATM network and a frame relay network: tunneling and translation.

The tunneling approach encapsulates the frame relay frames inside ATM cells, and the end frame relay devices communicate with each other without knowing that an ATM network is in the middle. This is also known as *transport layer interworking*, and in it an interworking function performs the data link layer mapping, signaling mapping, traffic parameter mapping, and the encapsulation. Higher-layer protocols at the end-user devices are not affected at all. This approach is used

in cases where ATM acts as a high-speed pipe and at the two ends of which are the frame relay devices.

The translation approach is also known as *service level interworking*. This approach translates signaling and protocols between a frame relay network and an ATM network. It accommodates situations where an FR device directly communicates with an ATM device, and where an FR network meets an ATM network half-way.

REVIEW QUESTIONS

1. Explain the motivations behind the development of ATM technology.
2. What layers of the OSI network reference model does the ATM protocol stack encompass? Describe the functions performed by the ATM layer in the context of the corresponding OSI layers.
3. Describe the concept of ATM cell and explain why ATM cell size is fixed at 53 bytes, which is not a power of 2 and not even an even number.
4. Describe the concepts of ATM virtual path, virtual path link, virtual channel, virtual channel link and virtual path connection, and virtual channel connections. Discuss the relationships between them.
5. Describe the concepts of ATM cross-connect and cell switching operations.
6. Describe the functions performed by the AAL sublayers in general and the types of services supported by each AAL sublayer.
7. Describe the ATM SVC and its characteristics. What were the motivations behind the development of SVC?
8. Describe the two types of ATM addresses and what they are used for.
9. Describe the two widely deployed ATM UNIs in terms of protocol messages and protocol procedures.
10. Describe the four different ways of carrying voice traffic over an ATM network and discuss which approach is most suitable for ATM access networks and which is most suitable for backbone networks.

REFERENCES

Atkinson, R. 1994. "Default IP MTU for use over ATM AAL5." IETF RFC 1626. Web site: www.ietf.org.

ATM Forum. 1994. "ATM User-Network Interface Specification V3.1" AF-UNI-0010-002. Web site: www.atmforum.com.

ATM Forum. 1996a. "Integrated Local Management Interface (ILMI) Specification." AF-ILMI-0065.000. Web site: www.atmforum.com.

ATM Forum. 1996b. "UNI Signaling 4.0." AF-SIG-0061-000. Web site: www.atmforum.com.

ATM Forum.1999. "ATM Trunking Using AAL2 for Narrowband Services." AF-ATOA 0113. Web site: www.atmforum.com.

ATM Forum. 2000. "Voice and Multimedia over ATM—Loop Emulation Service Using AAL2," version 4. AF-VMOA-0145.000. Web site: www.atmforum.com.

Bellcore 2001. "Bearer Independent Call Control (BICC) over ATM/IP Switching Systems Generic Requirements." GR-3100-CORE. Web site: www.telcordia.com.

Ibe, O. 1997. *Essentials of ATM Network and Services*. Reading, MA: Addison-Wesley.

ISO/IEC. 1988. "Information Processing Systems—Data Communications-Network Service Definition Addendum 2: Network Layer Addressing." International Standard 8348/Addendum 2. Web site: www.iso.org.

ITU-T. 1993. "B-ISDN ATM Adaptation Layer (AAL) Functional Description." Recommendation I.362. Web site: www.itu.int/ITU-T/.

ITU-T. 1995. "Digital Subscriber Signaling System No. 2: User-Network Interface (UNI) Layer 3 Specification for Basic Call/Connection Control." Recommendation Q2931. Web site: www.itu.int/ITU-T/.

ITU-T. 1996a. "B-ISDN ATM Adaptation Layer (AAL) Specification—Types 1 and 2." Recommendation I.363.1. Web site: www.itu.int/ITU-T/.

ITU-T. 1996b. "B-ISDN ATM Adaptation Layer: Type 3/4 AAL." Recommendation I.363.3. Web site: www.itu.int/ITU-T/.

ITU-T. 1997. "The International Public Telecommunication Numbering Plan." Recommendation E.164. Web site: www.itu.int/ITU-T/.

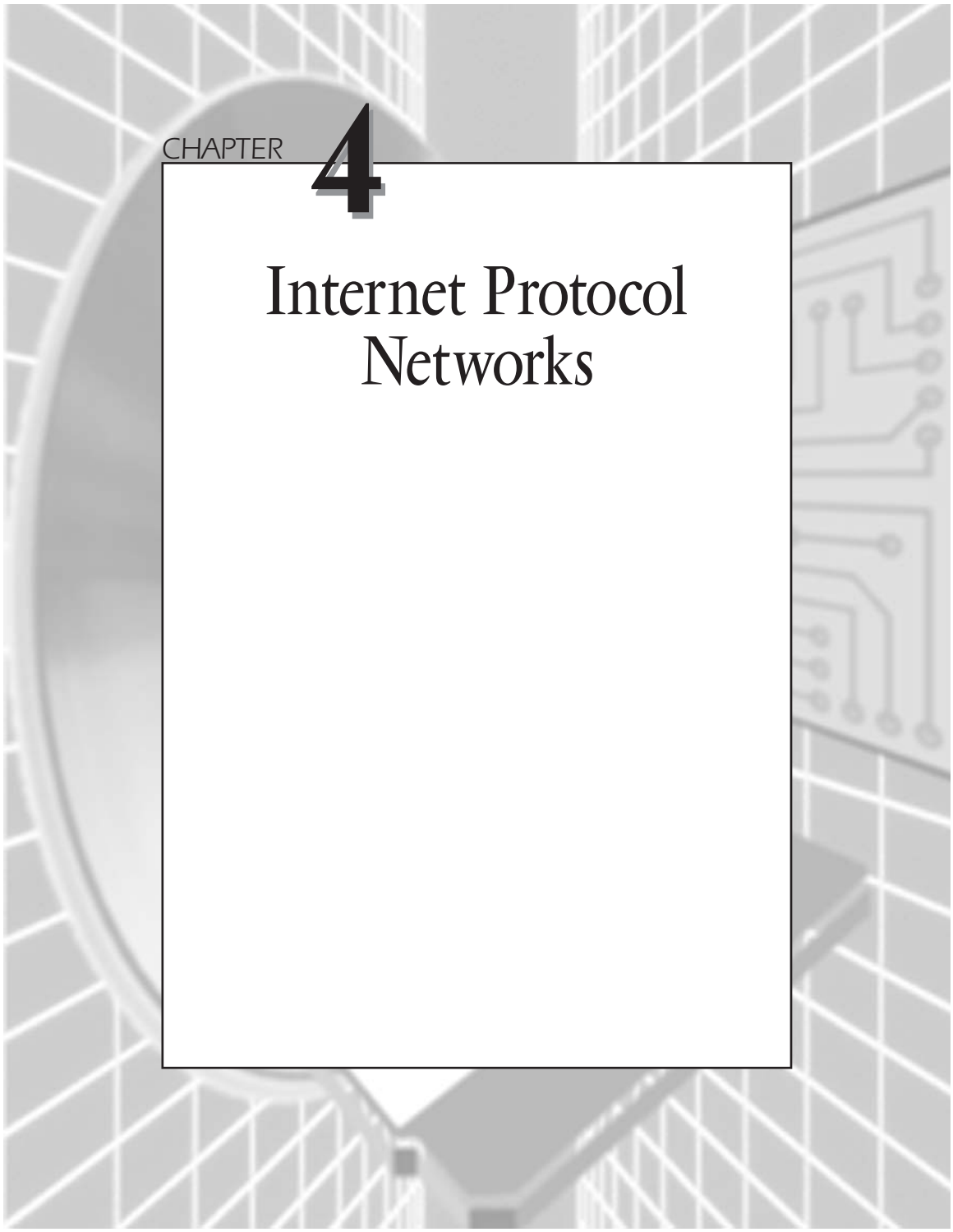
ITU-T. 2000. "B-ISDN ATM Adaptation Layer Specification: Type 2 AAL." Recommendation I.363.2. Web site: www.itu.int/ITU-T/.

Laubach, L., and Halpern, J. 1998. "Classical IP and ARP over ATM." IETF RFC 2225. Web site: www.ietf.org.

McDysan, D., and Spohn, D. 1999. *ATM Theory and Applications*. New York: McGraw-Hill.

McQuerry, S., and McGrew, K. 2001. *Cisco Voice over Frame Relay, ATM and IP*. Indianapolis, IN: Cisco Press.

Perros, H. 2002. *An Introduction to ATM Networks*. New York: John Wiley & Sons.



CHAPTER 4

Internet Protocol Networks

4.1 Introduction

Internet Protocol represents a class of packet network technologies that use connectionless communication protocols. In contrast, X.25, frame relay, and ATM are all based on connection-oriented packet protocols.

4.1.1 A Brief History

The Internet was directly born out of a research program initiated by the U.S. Defense Advanced Research Projects Agency (DARPA) in the early 1970s. The goal of the project was to test computer-based communications for the purpose of joint research and collaboration between geographically dispersed locations, using computers of diverse platforms. The initial network included computers at four university campuses—Stanford Research Institute, the University of Utah, University of California, Santa Barbara, and the University of California, Los Angeles (UCLA). A dominant concern behind the development of the Internet at the time was the need for an alternative communications network in case the nation's telephone network was disrupted in a hostile confrontation with the Soviet Union, which seemed a distinct possibility at the height of the cold war. A fundamental design principle of the IP network was for it to be highly reliable in the face of a network failure. The Internet has gone through many major changes in its short 30 years of history:

- *1962.* Paul Barran at the Rand Corporation developed the idea of package switching, in which communications data is broken into small units called *packets,* each of which is then routed to its destinations individually, where they are all assembled.
- *1969.* ARPANET is formed, linking UCLA, the Stanford Research Institute, the University of California, Santa Barbara, and the University of Utah. Under the sponsorship of DARPA, the network was originally intended to allow scientists to share research related information over a network.
- *1974.* The TCP/IP protocol suite, allowing packet networks to be linked to form a bigger network, was published by Vinton Cerf and Robert Kahn (Cerf and Kahn 1974).
- *1979–1981.* Various research and university networks began to form. The examples include Bitnet (for "Because It's Time") and CSNet.

- *1986.* The National Science Foundation (NSF) launches NSFNet, linking five U.S. supercomputing centers.
- *1987.* The NSF assumes management responsibility for an Internet backbone.
- *1989.* ARPANET was formally shut down in favor of the Internet structure.
- *1991.* The NSF lifts restrictions on commercial use of the Internet.
- *1992.* The World Wide Web is made public by researchers at CERN (a European Laboratory for Particle Physics).
- *1993.* The Mosaic Web browser is first released by the National Center for Supercomputing Applications at the University of Illinois.
- *1995.* The Internet backbone goes commercial.
- *2000 on.* The Internet truly connects the world together and becomes a necessity of daily life. For example, 58 percent of the America population had Internet access in their home in July 2001, according to a survey by Nielsen/NetRating.

4.1.2 Internet Standards

The Internet Activities Board (IAB) was created in 1983 to guide the evolution of the TCP/IP protocol suite and to provide research advice to Internet communities, while the initial research coordination was done under the auspice of a DARPA research program. It now has two primary components: the Internet Research Task Force (IRTF) and the Internet Engineering Task Force (IETF). The former is responsible for organizing and exploring advanced concepts in networking under the guidance of the IAB. The IETF was formed in 1989, and is primarily responsible for the standardization of Internet protocols. It has evolved to become one of the most prominent standards bodies as the Internet takes on an ever more prominent role.

4.1.3 IP Network Layers

This chapter will follow the bottom four layers of the IP network protocol stack, shown in Fig. 4-1, with the corresponding layers of the OSI network model as the reference.

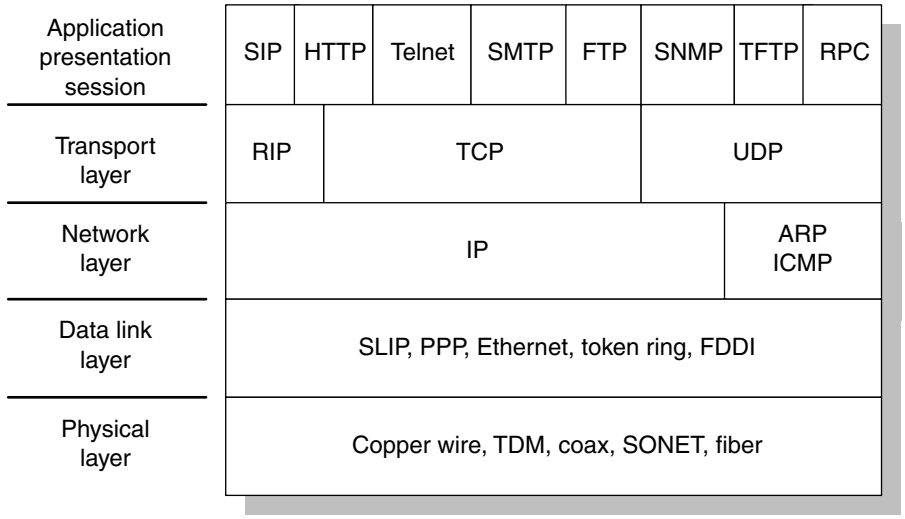


Figure 4-1
IP network reference model.

The physical layer provides the physical connectivity and medium transport. IP traffic can be carried over T1/E1, DS3, SONET, and the fiber optical network. The physical layer is responsible for transferring a stream of bits from point A to point B. This layer encompasses a generic transmission network with no IP-specific features.

The data link layer formats a continuous bit stream from the physical layer into meaningful units called *frames*, and processes the header of a frame to perform the data link layer functions. The functions include error checking and correction, monitoring the physical connectivity, and data compression, among others. Finally, the data link layer prepares the data into packets and sends the packets to the network layer for further processing. As shown in Fig. 4-1, there are two sets of data link layer protocols: one predominantly for wide area networks (WANs) and the other for local area networks (LANs). The link layer protocol, also known as *layer 2 protocols* for WAN, include Serial Line IP (SLIP) and Point-to-Point Protocol (PPP). The layer 2 protocol for LAN is predominantly Ethernet, while fiber-distributed data interface (FDDI) and token rings also had considerable deployment in the early 1990s. IP over other layer 2 technologies such as ATM and frame relay is discussed in Chap. 3 (on ATM) and Chap. 2 (on frame relay), respectively.

The network layer is mainly responsible for routing the received packets to the next stop along the way to the destination. It checks the destination address in the header of each packet against the local routing table and determines the optimal route to the next hop based on criteria such as

link cost or distance. The dominant protocol at this layer is Internet Protocol, for which the IP network based on this protocol is named. IP is a datagram protocol designed to be highly resilient to network failure, but with no guarantee of packet delivery. One other protocol at this layer is Internet Control Message Protocol (ICMP), which allows a router to send control messages to other routers. A familiar utility provided by ICMP is the *ping* function that allows the user to verify the reachability of a remote host. Another protocol on the IP layer is the Address Resolution Protocol (ARP) that directly interfaces the data link layer such as Ethernet and maps a physical address of the data link layer, for example, an Ethernet MAC address to an IP address.

The transport layer protocols enable a system to distinguish one application from another through the associated ports. As shown in Fig. 4-1, the Transmission Control Protocol (TCP) and User Datagram Protocol (UDP), the two most commonly used transport layer protocols on IP networks, are both directly supported by IP. TCP provides reliable delivery of ordered packets to an application. In addition, it also provides packet flow control and prioritized data flow. UDP, on the other hand, only provides unacknowledged packet delivery, but with faster speed. A third protocol at this layer is the Real-Time Protocol (RTP) that is designed to provide real-time delivery of packets to support time-sensitive applications such as multimedia streaming and voice over IP.

On top of the transport layers are applications that encompass the top three layers of the OSI models: the session, presentation, and applications layers. The applications supported by TCP include TELNET, Simple Mail Transport Protocol (SMTP), File Transfer Protocol (FTP), and Hypertext Transfer Protocol (HTTP). Applications supported by UDP include Simple Network Management Protocol (SNMP), and Trivial FTP (TFTP).

This chapter provides a journey along the IP network protocol layers, starting from the bottom layer and moving up to layer 4, the transport layer.

4.2 IP Data Link Layer Technologies

The journey on an IP network starts from the data link layer because that is where the IP-specific protocols and technologies come into the picture. The terms *layer 2* and *data link layer* will be used interchangeably throughout this and many other chapters in the book.

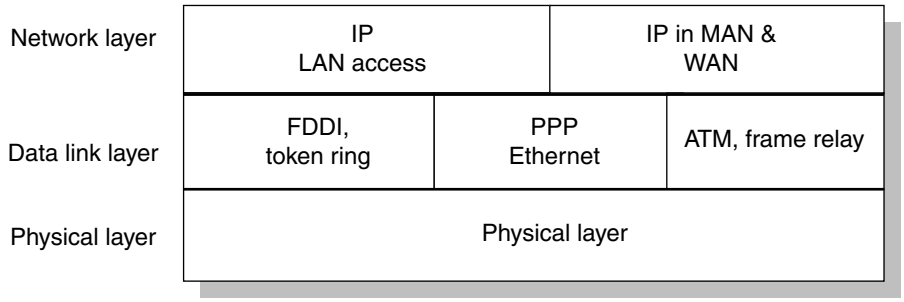


Figure 4-2
Data link layer choices for IP networks.

There are a number of choices for carrying IP traffic at the data link layer, as shown in Fig. 4-2. The data link technologies are divided into two categories based on where each technology is primarily deployed, namely, in a LAN environment or a WAN environment. For a WAN environment, widely used layer 2 technologies for carrying IP traffic include ATM and frame relay. Token ring and FDDI are dedicated to the LAN environment. In the middle, across both LAN and WAN environments, are Ethernet and the Point-to-Point Protocol. Ethernet, initially developed purely for a LAN environment, is being standardized to carry IP traffic in WAN/metropolitan area network (MAN) environments using fiber optical links as well. PPP, originally designed for the dial-up connection between a user and an internet service provider (ISP) network, is also used for carrying IP over SONET networks, or packets over SONET (POS) in the WAN environment.

This section introduces PPP, Ethernet, FDDI, and Token Ring as IP data link layer protocols.

4.2.1 Serial Link IP and Point-to Point Protocol

SLIP was an early version of the data link layer protocol designed at 3Com for users to access the Internet. In 1988, IETF adopted SLIP as a standard link layer access protocol (Romkey 1988). SLIP defines a simple frame that encapsulates the IP datagrams for transmission on the physical layer along with a simple method for delimiting IP packets with a special character. SLIP supports both synchronous and asynchronous transfer of data over dedicated or dial-up lines. It is quite simple, with no mechanisms for error checking or correction, and has no support for protocols other than IP. But soon the phenomenal expansion of the Internet outgrew the SLIP, and PPP is developed to supersede it.

PPP is a data link layer protocol intended to transport multiprotocol datagrams over point-to-point links such as phone lines and optical links. Although it supports datagrams of other protocols, its primary use is for IP datagram transport.

PPP has three components, each providing a different service (Simpson 1994):

- A method for delineating and encapsulating datagrams, based on an established protocol high-level data link control (HDLC).
- A link control protocol for establishing, configuring, and maintaining data link layer connection.
- A set of Network Control Protocols (NCPs) for establishing and configuring different network-layer protocols.

4.2.1.1 Framing and Encapsulation The encapsulation of bit streams from the physical layer is the first responsibility of the data link layer. The data processing operation at the data link layer has two basic steps: framing, and encapsulation. The framing process converts bit streams into meaningful data units called *frames*. The procedure and a frame structure for achieving this are defined in HDLC, a widely used data link layer protocol.

HDLC is an ISO standard that is based on IBM's Synchronous Data Link Control protocol. It defines a frame structure with unique bit patterns to mark the beginning and end of a frame. It has a standard frame structure that can support three types of frames: informational, supervisory, and unnumbered. A more detailed introduction to HDLC is provided in Chap. 1 on X.25 networks.

The encapsulation wraps the data frames obtained from the previous step into datagrams with headers and trailers. The header and trailer of a PPP encapsulation carry control information and padding of the datagram at a fixed size for transmission.

4.2.1.2 Link Control Protocol The Link Control Protocol (LCP) of PPP automatically establishes, configures, and tests a data link connection between two points. To establish a connection between two points, LCP negotiates connection parameters, authenticates the connection, and then negotiates a network layer protocol to use for the connection. An example of a user trying to establish a dial-up connection to an ISP network will help illustrate the LCP operations of connection establishment.

First, the customer personal computer (PC) sends LCP packets to the ISP router to negotiate a set of connection parameters. The parameters

include maximum receive unit (MRU), authentication protocol, quality protocol, magic number, and compression fields. The MRU indicates the maximum size of datagram that the receiving end can handle and the default is set to 1500 octets. The authentication protocols that PPP supports include password authentication and challenge handshake authentication. The magic number is a randomly chosen number to avoid the looped back link. The user can also negotiate the compression of PPP protocol, data link layer control, and address fields.

Once both transmitting and receiving ends have agreed upon the parameters and a connection is established, LCP proceeds with authentication of the user trying to log on to the ISP server. The chosen authentication protocol is used for this purpose.

The next step is to negotiate and configure a network layer protocol—the protocols like IP, IPX, or AppleTalk. After an agreement on a network layer protocol has been reached, the data link layer of both sides sets its NCP accordingly.

4.2.1.3 Network Control Protocol The NCP of PPP is specific to each protocol. It is responsible for supporting a particular network layer protocol. One task of NCP is to allow the assignment and management of IP address. An Internet service provider normally has far fewer IP addresses than the number of customers. It uses NCP to dynamically assign one of the prereserved IP addresses to a dial-up user as the user calls in on a common access number. This allows the ISP to efficiently share a limited number IP addresses among a large number of users.

4.2.2 Ethernet and Other LAN Data Link Layer Protocols

Ethernet was originally defined in a set of IEEE 802.3 standards as a data link layer protocol over copper wire and coax cable transmission media for the LAN environment (IEEE 1996). It has come to be the dominant choice of the LAN protocol for supporting IP traffic. In addition, major extensions to the Ethernet are underway to allow it to go beyond LAN to metro and backbone networks over an optical transmission medium. The extensions are focused on the physical layer, and are discussed in Chap. 7 on the optical Ethernet. Chapter 8 on local area networks provides a general introduction to the Ethernet. This section provides a quick overview of Ethernet as a layer 2 technology, i.e., as the data link layer

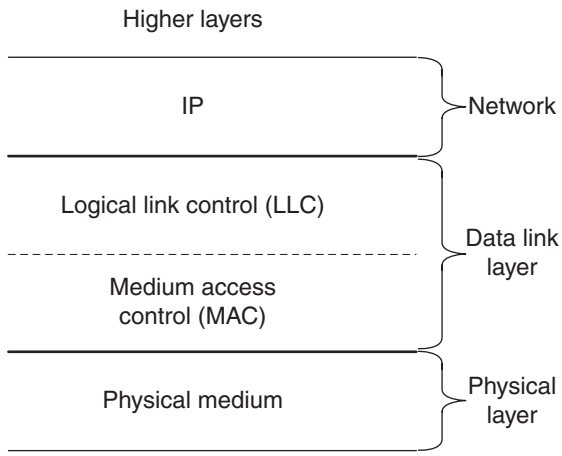


Figure 4-3
Ethernet protocol stack structure.

that consists of medium access control and logical link control sublayers, as shown in Fig. 4-3.

4.2.2.1 Logical Link Control Sublayer The main goal of the IEEE LLC protocol is to shield the differences between different MAC implementations and present a generic link layer service to the network layer. Using the LLC protocol data unit, the LLC provides three types of services to the network layer: unreliable datagrams, acknowledged datagrams, and connection-oriented service.

4.2.2.2 LAN Medium Access Control Sublayer LAN data link protocols differ from each other at the MAC sublayer. That is, the Ethernet MAC sublayer is different from Token Ring's MAC. In general, all MAC sublayers perform two primary functions: data frame assembly and disassembly, and medium access control. Upon receiving an LLC PDU, the MAC sublayer builds a MAC PDU by adding a MAC header and a trailer to each frame. The MAC sublayer also performs error checking and correction.

The Ethernet MAC sublayer uses carrier sense multiple access with collision detection (CSMA/CD) for control of multiple access. Originally Ethernet only supported half-duplex operations, which allows data to travel in only one direction at a time. Later on, the full-duplex mode of operation was added to support simultaneous transmission of data in both directions. In the CSMA/CD, all stations intending to send data have to contend for the right to the transmission medium. If a collision is detected, the sending station tries again after waiting for a random number of milliseconds.

4.3 IP Layer Basics

The journey along the IP network layers comes to the IP layer, also known as the *network layer,* which is one layer above the data link layer. At this layer, the IP header is examined and interpreted and the packet is routed based on the IP address in the packet header. This section introduces the IP packet format and addressing scheme and IP routing basics (Postel 1981).

4.3.1 IPv4 Packet Format

The IP packet consists of two parts: data and packet header. The data part is a variable field holding up to a maximum of 65535 bytes of data. The header consists of the fields shown in Fig. 4-4. When an IP packet is received at a router, the first four fields, the header of the packet carrying the information about the packet, are processed first.

The first four fields of the IP packet header provide general information on the packet:

Version. When a packet is received, an IP node first checks the version field. Currently IP protocol version 4 or IP v4 is almost universally used, but this field will be important when IP v6 starts being deployed.

IP header length (IHL). This field indicates the packet header length in units of 4-byte words. The common header length is 5 bytes, without any option field. This field indicates the number of data words for the optional field if one is present.

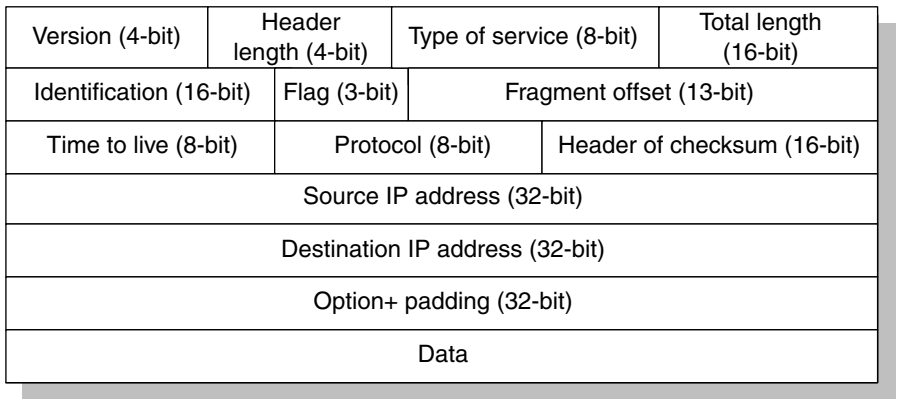


Figure 4-4
IP v4 packet structure.

Type of service (ToS). This field is made up of a 3-bit precedence field and a 3-bit service type; 2 bits remain unused.

Total length. This field indicates the total length of the packet, that is, the header length plus the user data, in units of bytes.

The next three fields—identification, flags, and fragment offset—control the fragmentation and reassembly of IP packets:

Identification. A 16-bit number together with the source address uniquely identifies this packet for the purpose of reassembly of fragmented datagrams.

Flags. A sequence of three flags, with one of the 4 bits unused, is used to control whether routers are allowed to fragment the packet and to indicate the part of a packet to the receiver.

Fragmentation offset. This is a byte count from the start of the received packet, set by any router that performs the fragmentation of the packet.

The remaining fields of the IP packet header are as follows:

Time to live (TTL). The 8-bit time-to-live field specifies the number of hops or links which the packet may be routed over. The TTL is decremented by 1 each time the packet is forwarded from a router. When TTL reaches 0, the packet is declared invalid and thus subject to discarding. Packets of the same application do not necessarily traverse the same route, and potentially can end up in an infinite loop. This field is designed to prevent such situations.

Protocol. The 8-bit protocol field identifies the transport layer protocol such as UDP or TCP and thus enables the protocol stack to know the format of the user data field. The preassigned protocol identifiers include 1 = ICMP, 2 = Internet Group Management Protocol (IGMP), 6 = TCP, and 17 = UDP.

Header checksum. The 16-bit header checksum field ensures the integrity of the packet header fields with a bit-parity check that can be implemented in software or hardware. It is a 2's complement checksum inserted by the sender and updated whenever the packet header is modified by a router. Packets with an invalid checksum are discarded by a router in an IP network.

Source and destination addresses. These two 32-bit address fields identify the addresses of the sender and receiver of the packet, respectively.

Option. The option field can be used to specify items such as security level, source routing, or a request for a route trace. This field is rarely used.

4.3.2 IP v4 Addressing and Subnet Addressing

The IP addressing scheme lies at the heart of IP networking and communications. It is IP addresses such as the source and destination addresses in the IP packet header that globally identify an IP device such as a host, a router, or an interface. Each addressable network element within the Internet must be uniquely identified. Each IP address has 32 bits divided into four groups of digits in the form of "dot decimal": xxx.xxx.xxx.xxx.

The IP address forms a hierarchical structure. The four sets of numbers, each ranging from 0 to 255 ($2^8 - 1$), are classified into four classes of IP address: A, B, C, and D. The address of each class consists of three parts: a prefix that ranges from 1 to 4 bits identifying the address class, a network ID and a host ID, as shown in Figs. 4-5 and Table 4-1. For example, if the first number of your company's IP address is 197, it is a class C IP address with the first 3 bits being binary string 110. The class C IP address can accommodate up to 2^{21} networks with up to 254 hosts within each network.

The four classes (A, B, C, and D) of IP addresses are designed to provide flexibility for various situations. Class A addresses are intended for networks with a very large number of hosts and begin with a bit 0 as class ID. The network identifier part takes up 7 bits and the host address has 24 bits. Obviously class A addresses can accommodate a small number of networks but the largest number of hosts within a network.

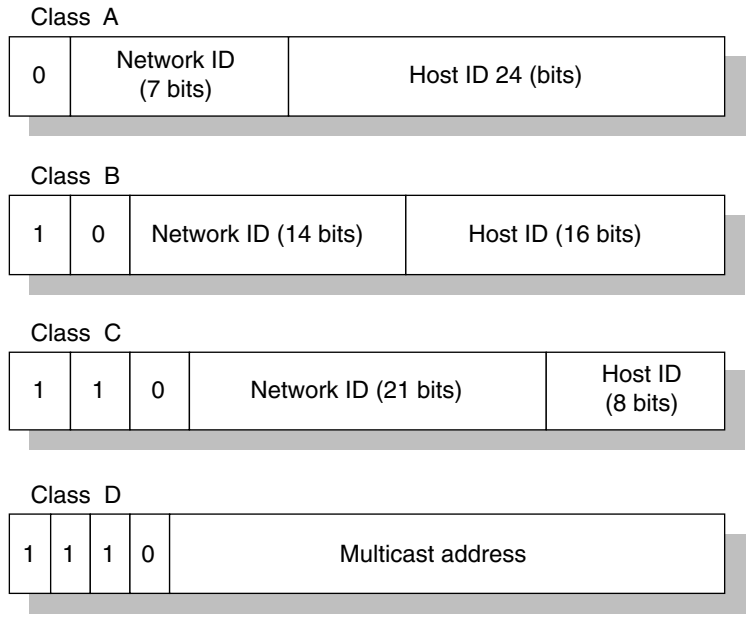


Figure 4-5
Four classes of IP address.

TABLE 4-1
IP Address Structure

Class	Prefix ID bits of the class	Number of network ID bits	Range of first byte network address	Total number of networks	Bits for host address	Number of hosts per network
A	0	7	0–127	2^7	24	2^{24}
B	10	14	128–191	2^{14}	16	2^{16}
C	110	21	192–223	2^{21}	8	2^8
D	1110	NA	224–254	NA	NA	NA

NA = not applicable.

Class B addresses are for medium-size networks such as campus-wide LANs. Their first octet begins with 1 0, and their network addresses range between 128 and 191 (decimal). As shown in Table 2-1, the network identifier part has 14 bits, and the host identifier part has 16 bits. The maximum number of hosts that can connect to a class B network is 2^{16}, and the maximum number of class B networks is 2^{14}.

Class C addresses are for small-size networks. Their first octet begins with a 110, and they have a very large network ID field and a small host field. Up to 2^8 hosts can connect to a class C network, and the network addresses range from 192 to 254.

Class D addresses begin with the octet string 1110, and the remaining 28 bits are known as *multicast addresses*. When IP packets are sent to a group of stations, this is known as *multicasting*. Broadcasting is a variation of multicasting, in which all the stations in a subnet receive broadcast IP packets.

The main advantage of the Internet addressing scheme is its flexibility in arranging networks. For example, if a network has a large number of workstations, class A addresses may be more suitable than other address classes. On the other hand, if there are relatively few workstations in a network, class C addresses may be appropriate.

IP address assignment is a two-step process. First, a central Internet administrative authority assigns the network portion of an IP address, and then a network administrator of the organization is responsible for the assignment of host portion of the address.

The shortage of IPv4 addresses has become an acute issue, amounting to an Internet crisis, even with a huge address space of well over 4 billion Internet addresses (2^{32}). The shortage is partially due to the unplanned usage of the addresses during the earlier days of the Internet when IP addresses were dished out generously to any entities that showed interest.

One short-term solution to the crisis has been to redistribute some of the previously assigned but unused IP addresses. A more significant effort to resolve the IP address shortage problem is to break down the hierarchical structure of the IP address and do away with classes of IP address in IP packet routing. Developed in 1993, this is called *classless interdomain routing* (CIDR), and improves scaling of the routing system as well as provides better utilization of IP address. Routers supporting CIDR treat an IP address as a flat 32-bit number instead of two parts, the network address and the host address.

4.3.3 IPv6 and Its Packet Format and Addressing Scheme

IPv6 is the next generation of the Internet. In a nutshell, it supersedes IPv4 and, in addition to all the services and protocols that IPv4 supports, also provides the following additional features:

Scalability. It provides 128-bit source and destination addresses that accommodate a huge number of IP networks and hosts within each network.

Support for real-time applications. The flow control mechanisms to support real-time applications such as video conferencing and voice applications are built into the IP header structure. The "flow label" enables a router to associate packets with an end-to-end flow and distinguish real-time packets from non-real-time ones.

IPv4 support. Support for all IPv4 routing protocols such as Routing Information Protocol (RIP), Intermediate System-to-Intermediate System (IS-IS) Protocol, OSPF, and Border Gateway Protocol (BGP) with IPv6 extensions.

Security. IPv6 specifications include packet encryption and source authentication.

One main goal of IPv6 is to resolve the IP address shortage issue. Even with the CIDR method to generalize the concept of subnet masking already described, the IP address crisis persists and will only get worse. With the approaching exhaustion of IP addresses, IETF adopted IPv6 (Bradner and Mankin 1995; Hinden 1995) in 1995, in the hope that it will take care of the IP address issue for a long, long time, if not forever.

The IPv6 packet format is simpler than its predecessor, as shown in Fig. 4-6. However, the IPv6 header with 40 bytes is much longer than its predecessor. Its format can be described as follows:

Version. The field is same as in its predecessor, indicating the IP version. This is the very first field examined by a router and has the same number of bits as in IPv4.

Priority. A 4-bit field used together with the flow field to provide flow control and QoS. The 16 values of the 4-bit field are divided into two priority groups: 0 through 7 are for the traffic source that can be flow-controlled in case of congestion, and 8 through 15 are for the traffic sources that can not respond to flow control during congestion. A traffic flow with a higher-priority number is less likely to be discarded than a traffic flow with lower priority in the same priority group in case of a congestion.

Flow label. A 24-bit field intended to support the QoS scheme such as the Resource ReSerVation Protocol (RSVP). RSVP is described in Chap. 18.

Payload length. A 16-bit field that indicates the number of bytes in the user data payload, following the 40-byte header. The 16 bits can indicate a maximum of 56,665 bytes.

Next header. An 8-bit field that indicates the subsequent header extension field. There are six optional header extensions:

- *Hop-by-hop:* a special option that requires processing at every node
- *Source routing:* extended routing like IPv4 source route
- *Fragmentation:* possible packet fragmentation and reassembly
- *Destination support:* optional information to be examined by destination node only
- *Authentication:* integrity and authentication
- *Security encapsulation:* confidentiality

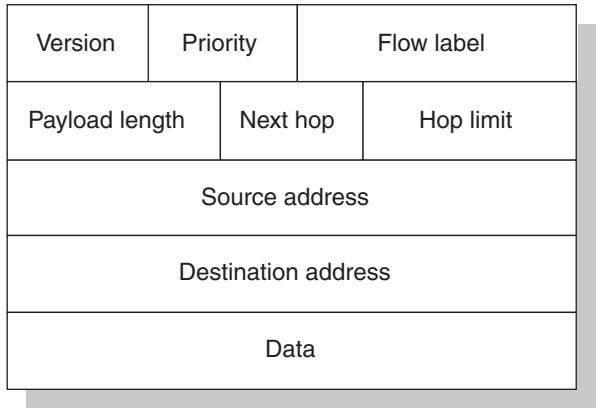


Figure 4-6
IPv6 packet format.

Hop limit. This field specifies the maximum number of hops a packet may travel before reaching the destination. It is equivalent to the TTL field in the IPv4 packet header.

Source and destination address. The source and destination IP address

One major difference between IPv4 and IPv6 is the IP address field with the address length quadrupled for IPv6 packets. The address space of 2^{128} is an astronomically huge number (3.4×10^{27} addresses per person for a world population of 10 billion), with little chance of running out of IP addresses for the foreseeable future. The required IPv6 header fields add up to 40 bytes, but the optional extension fields can make it substantially larger.

4.3.4 IP Layer Operations

The IP layer performs two main operations: IP packet management and packet routing (Comer 2000). The task of packet management involves packet fragmentation and reassembly. Packet routing determines where to send packets and forwards them to the next hop based on the predefined routing criteria.

IP packet fragmentation fits the IP packet into data link layer frames. The data link layer frame size is limited, and the size, which is known as the *maximum transmission unit* (MTU), defines the unit of bits that can be transmitted over a physical network at one time. For example, the Ethernet MTU is 1500 bytes. When the IP layer sends a packet with a size larger than the MTU of the data link layer, the IP layer chops the packet into smaller units that fit into the MTU. This is called *IP fragmentation,* and can take place anywhere in a network: at an end host, a router, or a gateway. The fragmented packet is not reassembled until it reaches the destination node, where the IP layer reassembles the pieces back into an IP packet.

IP routing forwards IP packets from one node to the next, then to the next, all the way to the destination host. An IP packet is processed through all five layers of the IP protocol stack, at both the source and destination hosts, as shown in Fig. 4-7. At each intermediate node, packets are passed only up to the IP layer. Each IP packet header is examined to find out what is the next hop to route the packet to, using the routing table at the local node and the destination address in the IP packet header.

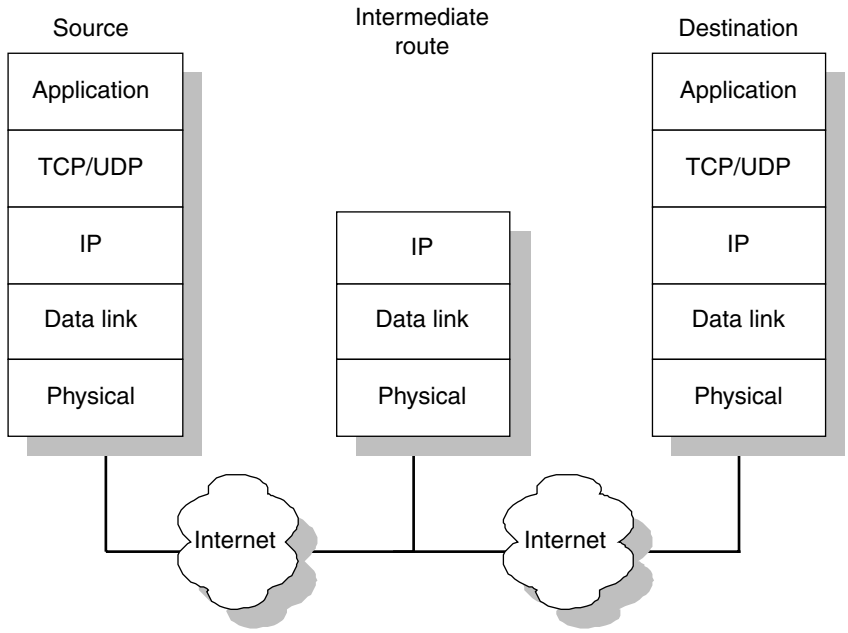


Figure 4-7
IP routing example.

Traditionally, IP routing is done in hop-by-hop fashion. Each IP router has only a local view of all the connected nodes and delivers an IP packet to a connected next hop based on the predefined forwarding criteria. In hop-by-hop fashion, the packet is delivered to the destination node that is connected to the same physical network or within a set of connected networks, known as an autonomous system (AS). If the destination of a packet is *not* on the same physical network as the source, the packet is passed to a gateway on the same network as the source. A gateway is connected to multiple networks. If the destination *is* on a network connected to the gateway, then the packet is passed to the destination network and routed to the destination node in hop-by-hop fashion. Otherwise, the gateway forwards the packet to another gateway. In this fashion, the packet is forwarded until it reaches a gateway that is on the same network as the destination node.

The concept of autonomous system is important for IP routing. IP networks consist of a set of ASs, and each AS is one or more physical networks that are administered as a single unit. The concept of AS is related to the fact that TCP/IP protocols were developed with the ARPANET already in place, and each exiting network is treated as an independent AS. IP routing within an AS and routing between two ASs are handled with different routing protocols.

There are two types of routing protocols used in IP networks: intragateway routing protocol and intergateway routing protocols (Lewis 2000). The intragateway routing protocols, also known as *Interior Gateway Protocols* (IGPs), are used within an autonomous system. The most commonly used IGPs include RIP, IS-IS, and OSPF. The intergateway routing protocols, also known as *exterior routing protocols,* are used between autonomous systems. The widely used exterior routing protocols include Exterior Gateway Protocol (EGP) and Border Gateway Protocol (BGP). Appendix A of this book provides an introduction to the widely used interior and exterior routing protocols.

4.4 IP Transport Layer

Now the journey along the IP protocol stack moves one layer up to the transport layer, which deals with the applications at a source and a destination host. IP packets have been successfully routed to the destination and arrive at the IP layer of a destination host. The packets are passed to the transport layer for further processing after the IP header is stripped from each packet. This section takes a close look at the activities at the transport layer.

The main responsibility of the transport layer is delivery of packets to the application port. There are two basic approaches to accomplishing this, which result in two types of transport layer protocols: connectionless and connect-oriented. The connection-oriented protocol guarantees the in-order delivery of packets. In contrast, the connectionless protocol in general runs faster but without the guarantee of packet delivery. The most widely used transport protocols are the connection-oriented Transmission Control Protocol and the connectionless User Datagram Protocol.

4.4.1 TCP

Transmission Control Protocol is a flexible, connection-oriented protocol. It does not require a particular protocol above it or at the layer below it. It can be used on top of either a connectionless or connection-oriented protocol, though it is in general used on top of a connectionless datagram network such as IP. A quick examination of the TCP packet format will help provide a basis for understanding the TCP operation.

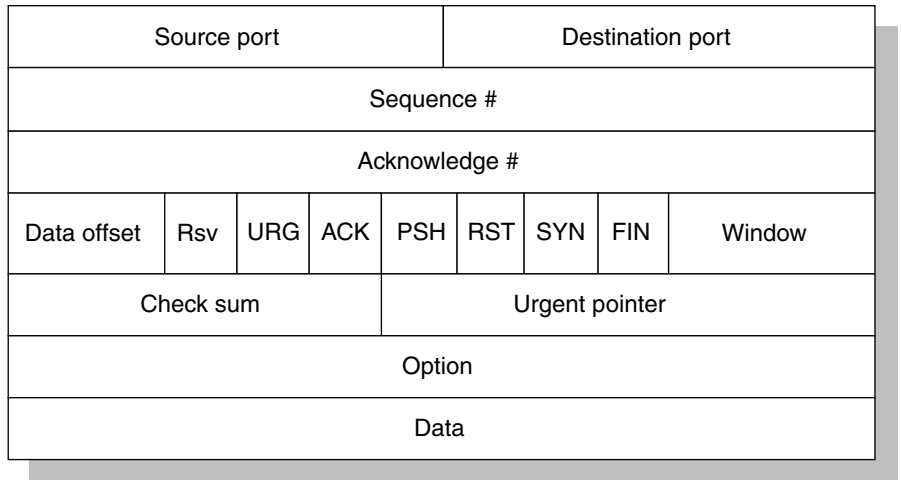


Figure 4-8
The format of TCP packet.

The TCP packet format is shown in Fig. 4-8. It consists of a TCP header part and a data part (DARPA 1981). The fields in the header include the following:

Source and destination port number. A 16-bit field that identifies an application running on the source or destination hosts. Port numbers below 256 are reserved for standard system services as specified in IETF RFC 1700 (Reynolds and Postel 1994). For example, port 1 is for the TCP multiplexer, port 2 is for the management utility, port 13 is for daytime, port 23 is for Telnet, and port 25 is for simple mail transfer (SMT).

Sequence number. A 32-bit number for tracking the data octet of a packet sent by the source. In the case of synchronization or when the SYN bit is set, it is the *initial sequence number.*

Acknowledge number. A 32-bit field sent by the receiver once a connection has been established. This is the sequence number of the data octet, which a receiver expects to receive next. For example, if a receiver has received data up to 1000 octets, the receiver assigns the acknowledged number to 1001 to indicate that the next octet it expects to receive starts at octet 1001.

Data offset. A four-bit field to indicate the length of the TCP header measured in multiples of 32 bits.

Reserved. Reserved for future extension and set to 0.

URG. A 1-bit field for urgent indicator. When this bit is set, the receiver should process this segment immediately, interrupting the current activities.

ACK. A 1-bit acknowledgment filed. This bit allows the sender to explicitly solicit the acknowledgment number from the receiver. If this bit is set, an acknowledge number is sent by the receiver.

PSH. A 1-bit push function field to ask the sender to send whatever data is available without waiting for a full buffer.

SRT. A 1-bit field used for resetting a connection.

SYN. A 1-bit field used for synchronizing sequence numbers. This is used in the beginning to establish a connection by means of a three-way handshake.

FIN. A 1-bit field used for closing a connection.

Window. A 16-bit field that indicates the number of octets a receiver is willing to receive.

Checksum. A 16-bit field to check whether data is corrupted.

Urgent pointer. A 16-bit pointer pointing to data offset following urgent data. This field is significant only if the URG bit is set.

TCP is a connection-oriented protocol that guarantees the end-to-end delivery of packets. The protocol requires that the receiver acknowledge every packet received. The sender sets a timer when a packet is sent out, and resends the packet if it fails to receives an acknowledgment when the timer expires. The acknowledge number in the TCP packet header indicates the sequence number of next packet the receiver expects.

One key function TCP performs is to guarantee the in-order delivery of packets to an application. TCP assumes that the underlying network can be a connectionless datagram network such as IP that may deliver packets out of order, or may deliver duplicate packets. It is the responsibility of the TCP layer to recover and maintain the original order of the packets. TCP achieves this by segmentation and reassembly using the sequence number in the TCP packet header.

The second key function TCP performs is traffic and congestion control. It uses a flow control mechanism for the sending and receiving hosts to negotiate and control the rate at which data is sent and received. TCP uses a sliding window flow control protocol, similar to the one used in X.25 but with a variable window size. The size of the sliding window indicates to the sender how fast it can dispatch the packets to the receiving host in number of bytes per second. The congestion control operates in two phases: the initial phase and steady-state phase. During the initial

phase, the sender starts with a congestion window size equal to that of one TCP segment. Once the sender receives an acknowledgment, the sender doubles the size of the sliding window, in effect doubling the rate of packet dispatch until the rate is too high for the receiver to handle and a packet is lost. This is indicated by the fact that the timer expires before the expected acknowledgment arrives at the sender. Then the sender retransmits the packet and sets the sliding window size back to the size it first started with. From this point on, the protocol enters the steady-state phase. The sender increases the size of the sliding window by 1 as opposed to doubling it, upon the receipt of the acknowledgment of each packet.

4.4.2 UDP

User Datagram Protocol is a connectionless transport layer (i.e., host-to-host) protocol that provides best-effort service for transferring datagrams from one host to another (Postel 1980). It provides an alternative to TCP with the goal of increasing the throughput since there is no network connection to maintain and no overhead for acknowledgment and retransmission.

UDP protocol itself is quite simple. The data payload is enveloped in a UDP header and sent to the IP layer. At the IP layer, an IP header is added. Then at the data link layer, a frame header is appended and the frame is sent across the network. At the receiving end, the frame header is removed first, then the IP layer header, and then the UDP header. The receiver does not send any acknowledge to the sender and sender does not know whether a dispatched packet has reached the destination.

The simplicity of UDP is reflected in the structure of the UDP header, as shown in Fig. 4-9 and described here:

Source and destination port. A 16-bit port address identifying the application on the source and destination hosts.

Length. A 16-bit field indicating the length of this UDP datagram in number of octets, including the UDP header.

Checksum. A 16-bit optional field that has the ones complement of the checksum derived from a pseudoheader. This checksum is for the entire UDP datagram.

Note the absence of the reliability-related fields such as the sequence and acknowledgment numbers in the UPD datagram header. Since the in-order, guaranteed delivery is not a requirement for UDP, the header is shorter and overhead is smaller.

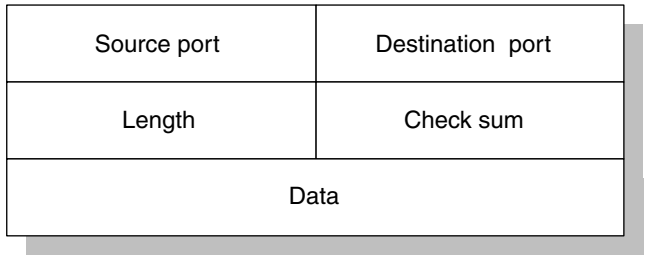


Figure 4-9
UPD packet header structure.

4.5 IP Multicasting

IP multicasting is the transmission of IP packets to a group of end hosts as the destination, as opposed to a single destination host. IP multicasting can support an increasingly wider range of applications such as group email, telecommuting, conferencing, and real-time stream broadcasting. This section describes the basic concepts of IP multicasting, the multicast addressing scheme, multicast routing, and the operations of an end-to-end multicasting applications.

4.5.1 IP Multicasting Concepts

IP multicast supports point-to-multipoint applications such as broadcasting a message to all employees in a corporation, video and audio conferences, mass software upgrades, and multisite remote database replications (Johnson and Johnson 1997). The applications list goes on and on. IP multicasting takes on a more prominent role as Web-based broadcasting applications become prevalent.

IP multicasting is a receiver-based operation: Receivers join a multicast group, and traffic is delivered to all members of that group by the multicasting-capable IP network. One basic requirement of IP multicasting is that only one copy of the message shall pass over any link in a network. Each multicast group is identified by a class D IP address. There is no assumption as to the location of a group member, and the members of the group can be present anywhere on the Internet. A sender need not be a member of the group. The basic idea of multicasting is to have routers listen to all multicast addresses and use multicast routing protocols to manage groups.

There are several extensions to the traditional point-to-point IP network in order to support IP multicasting. The extensions are mainly at the data link layer and IP layer. At the data link layer, the main extension

is to provide a mapping between the multicast address and the physical address (e.g., the MAC address of the Ethernet network) (Deering 1989). This is normally implemented at an interface card of an IP router or host. At the IP layer, a function is required to maintain a list of host group memberships associated with each network interface. The list is updated on a continuing basis as the new applications join or leave the group.

Another IP layer extension is to update the time to live field of the IP packet. IP multicasting uses the TTL field as a scooping parameter to control the number of hops that a multicast packet can travel. As described previously, the TTL field, containing the maximum number of hops allowed for a packet, is decremented by 1 each time the packet is forwarded. In the case of a LAN, TTL is set to 1 to prevent the multicast packets from going beyond the subnet.

Internet Group Management Protocol (IGMP) is a major component of IP multicasting that is designed to allow routers to find out the host group members and join and leave a group. All multicasting-capable hosts and routers must implement IGMP.

4.5.2 Multicasting Addressing

IP multicasting uses a class D address that has 1110 as the IP address prefix, as described in the preceding section. In the "dot decimal" notation, host group addresses range from 224.0.0.0 to 239.255.255.255. There are two types of multicasting addresses defined in IETF RFC 1112: permanent and temporary (Deering 1989). The permanent addresses are administratively assigned or reserved on a permanent basis. For example, the address 224.0.0.1 is for all IP multicast hosts on a LAN, 224.0.0.2 is the multicast address for all routers on a LAN, and 224.0.1.1 is for the Network Time Protocol (NTP). All permanent assigned multicast addresses are listed in IETF RFC 1700 (Reynolds and Postel 1994).

The temporary or transient multicast addresses are assigned and reclaimed dynamically along with the multicast sessions. Transient address spaces range from 224.0.1.0 to 239.255.255.255.

4.5.3 Internet Group Management Protocol

IGMP, a key component of IP multicast, provides a means for hosts to communicate to a router about IP multicasting group information.

IGMP is an IP layer protocol similar to ICMP that is implemented over IP and provides the following two main functions (Fenner 1997):

- It provides the means for a host to tell a connected router about the group memberships on the host.
- It provides the means for a router to query about memberships from a host.

One router on a LAN is designated as the multicast router that performs IGMP protocol functions. To determine if any hosts on the LAN belong to a multicast group, the designated router periodically send a multicast IGMP Query message with IP multicast address 224.0.0.1 and TTL set to 1. In response, each host sends back an IGMP Report message per host group, sent to the group address, i.e., the IP address 224.2.0.1, as shown in Fig. 4-10, so that all group members can read the message.

A host on a LAN can send an unsolicited Report message to join a multicast host group, or leave a group. In Fig. 4-10, host 3 sends an unsolicited Report message to the router to join the host group identified by 224.5.5.5. After the last member leaves the group, the multicast router stops forwarding packets to the multicast address.

4.5.4 Multicasting Routing and Routing Protocols

IP multicasting routing is more complicated than point-to-point routing, because a multicast address identifies a particular transmission session as opposed to a specific physical destination. During the early days of IP multicasting when its application was limited to small networks like a single LAN, a simple approach was to have the source maintain a list of all receivers participating in a session and send a copy of the message to each receiver. This resulted in a very inefficient use of bandwidth because many copies of the same message would traverse the same path through much of the network. In addition, this approach was difficult to scale up.

The new IP multicasting routing is based on the spanning tree concept, which guarantees that only one copy of the message will pass over a link of the network and thus results in much more efficient use of network bandwidth. Since the number of the receivers can be potentially very large and the receivers may be widely scattered on the Internet for a multicast session, the sender will not know about the receivers. In this scheme, one designated router in a network is responsible for routing the multicast traffic. It collects and maintains the information about the

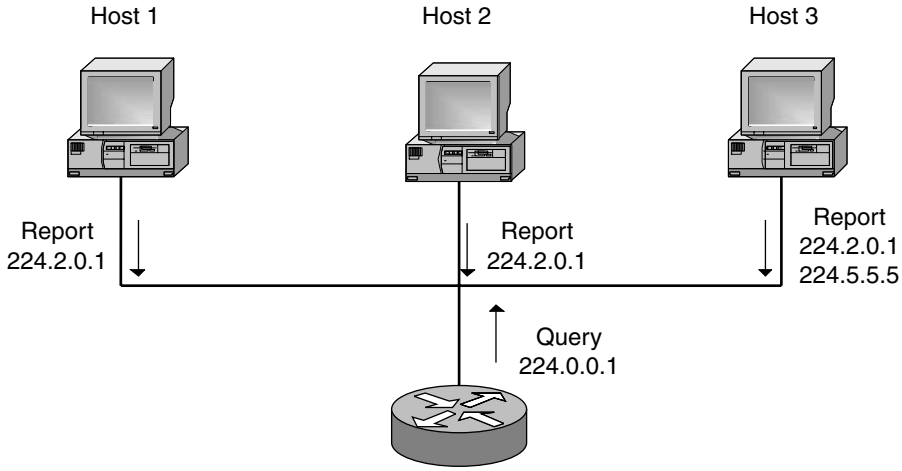


Figure 4-10
Illustration of IGMP protocol operations.

multicast group memberships from other routers and hosts attached to the network. The designated router constructs a spanning tree that connects all members of the IP multicast group. All members of a multicast group are connected to each other in the spanning tree in such a way that there is only one path between every pair of routers and there is no loop in the tree. Messages are replicated only when the tree branches, minimizing the number of copies of the messages that are sent throughout the network. Since the applications can join and leave a multicast group at any time, the spanning tree must be dynamically updated. Branches of the tree on which there is no receiver are discarded or pruned.

Several IP multicasting routing algorithms have been developed to route multicast traffic within a network and between networks. Those routing protocols are based on the existing point-to-point routing protocols. There are two categories of multicast routing protocols: "dense-mode" and "sparse-mode." The dense-mode multicasting routing approach assumes that the multicast group members are densely populated throughout the LAN and that bandwidth is plentiful. The dense-mode routing protocols use periodic flooding of the network with multicast traffic to set up and maintain the spanning tree. Dense-mode multicast routing protocols include Distance Vector Multicast Routing Protocol (DVMRP), Multicast Shortest Path First (MOSPF), and Protocol Independent Multicast-Dense Mode (PIM-DM).

The sparse-mode multicast routing protocol is based on the assumption that the multicast groups are sparsely populated and the group members are widely distributed across the network. Thus the flooding of multicast traffic would not be efficient and the routing protocol of this category uses either source or shared distribution trees. A message to join a multicast

group propagated from the receiver to the source, a rendezvous point or core. The sparse-mode multicast routing protocols include Core Based Tree (CBT) and Protocol Independent Multicast—Sparse Mode (PIM-SM).

END-TO-END MULTICASTING OPERATIONS

We now use an application example to illustrate the basic concepts and operations of IP multicasting described in this section. The application is a Web-based news broadcast to all of those hosts that have subscribed to the news group on the Internet. The source and destinations are far apart, over multiple physical networks, and the members of the group are distributed all over the networks.

1. An application at host A requests to join the news broadcast group. The host sends an IGMP Report message to the designated router to join the host group.
2. The designated router updates its multicast spanning tree based on the new information received from host A.
3. The source sends the news update to the designated router of the network it is attached to. Assume that a spanning tree at this designated router is up to date via the use of one of the multicast routing protocols.
4. The designated router first sends the news update to the next router on the spanning tree, which in turn follows the tree branches. The news update is duplicated only when the tree branches out.
5. The designated router also sends the news updated to the connected border gateway router, which essentially does the same thing: It multicasts the message to the network it is associated with and also sends the news message to the other connected networks. In this fashion, the news update eventually arrives at the designated router host A is connected to. The router then sends a copy of the news update to each member of the host group.

REVIEW QUESTIONS

1. Describe the three components of PPP and the main function of the LCP of PPP. Then discuss which part of PPP is used most when packets are directly carried over SONET or a packet over SONET (POS).

2. Describe how the TTL field of the IPv4 packet header works and why the protocol field is needed.
3. Out of the four classes of IP addresses, which one can accommodate the largest number of hosts within a network and which one can accommodate the largest number of networks?
4. What are the main components in an IP router? Discuss the functions of a router. Does an IP router process a TCP/UDP header?
5. Compare the IPv4 packet structure with that of IPv6. Discuss the differences between the fields other than the IP address.
6. Describe the two types of operations performed at the IP layer and the two types of IP routing protocols.
7. What fields in the TCP header are involved in flow control?
8. Compare the UDP packet header against the TCP packet header, discussing the main differences between them. For an application emphasizing raw throughput rather than the reliability of packet delivery, which protocol should be used?
9. Discuss the main functions of IGMP. An IGMP Report message can be used either in response to an IGMP Query message or when unsolicited. When would an unsolicited IGMP Report message be sent?
10. Describe how the IP multicasting works in a LAN context and a WAN context.
11. Describe the differences between the two categories of multicast routing protocols.

REFERENCES

Bradner, S., and Mankin, A. 1995. "The Recommendation for the IP Next Generation Protocol." IETF RFC 1752. Web site: www.ietf.org.

Cerf, V., and Kahn, R. 1974. "A Protocol for Packet Network Intercommunication," *IEEE Transactions on Communications*, Vol. COM-22, No. 5, pp. 637–648.

Comer, D. 2000. *Internetworking with TCP/IP, Volume 1: Principles, Protocols, and Architectures*, 4th ed. Englewood Cliffs, NJ: Prentice Hall PTR.

DARPA Internet Program. 1981. "Transmission control protocol." IETF RFC 793. Web site: www.ietf.org.

Deering, S. 1989. "Host extension for IP multicasting." IETF RFC 1112. Web site: www.ietf.org.

Fenner, W. 1997. "Internet Group Management Protocol, Version 2." IETF RFC 2236. Web site: www.ietf.org.

Hinden, R. 1995. "IP next generation overview." White paper. Web site: www.sun.com.

IEEE. 1996. "Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Applications." 5th Ed. IEEE 802.3.

Johnson, V., and Johnson, M. 1997. "Introduction to IP multicast routing." White paper. Web site: www.ipmulticast.com.

Lewis, C. 2000. *Cisco TCP/IP Routing Professional Reference*. New York: McGraw-Hill.

Postel, J. 1980. "User Datagram Protocol." IETF RFC 768. Web site: www.ietf.org.

Postel, J. (ed.). 1981. "Internet Protocol—DARPA Internet Program Protocol Specification." IETF RFC 791. Web site: www.ietf.org.

Reynolds, Jr., and J. Postel. 1994. "Assigned Numbers." IETF RFC 1700. Web site: www.ietf.org.

Simpson, W. 1994. "The Point-to-Point Protocol (PPP)." IETF RFC 1661. Web site: www.ietf.org.

PART 2

Broadband Transport Networks

Part II of this book focuses on the digital broadband transport networks that provide physical layer services, namely, transmission of bits and bytes of information, without regard to the upper-layer network protocol. Part II reflects the evolution of the digital transport technologies as shown in Fig. P2-1 and as discussed in the following three chapters: Digital Transmission Systems and SONET (Chap. 5), WDM Networks (Chap. 6), and Optical Ethernet (Chap. 7).

The first-generation digital transmission system, represented by the Digital Signal (DSx) system and T-carrier (e.g., T1, T3, etc.) in North America, used copper wire as the transmission medium and achieved a data rate of over 50 Mbps. The next generation uses optical fiber as the transmission medium, and are represented by the Synchronous Optic Network (SONET) standard in North America and the Synchronous Digital Hierarchy (SDH) standard in Europe; both standards are still widely deployed. The third-generation optical transport network is characterized by signal switching at the optical level and a dense wavelength division multiplexing (DWDM) that combines many wavelengths onto a single optical fiber.

Optical Ethernet combines the ubiquitous Ethernet technology and wave division multiplexing (WDM) network technologies to provide an alternative optical transport network technology that extends the reach of Ethernet into metro and wide-area networks.

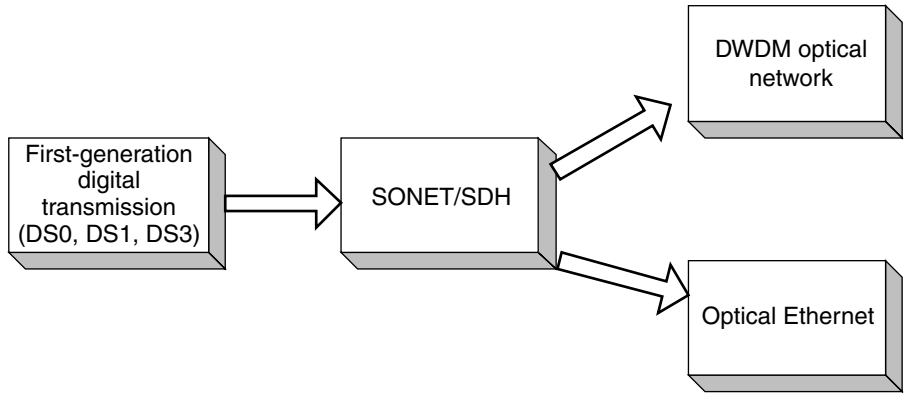


Figure P2-1
Evolution of digital transport network technologies.

CHAPTER 5

Digital Transmission Systems and SONET

5.1 Digital Transmission Systems

Digital transmission systems as described in this section refer to the T-carrier and Digital Signal (DSx) system, which uses copper wire as its transmission medium and electric pulse for digital signals. This section first introduces four key components of a digital transmission system—pulse code modulation (PCM), framing structure, digital multiplexing, and timing synchronization—and then describes the first-generation digital transmission systems: the T-carrier with the corresponding DSx in the United States and its counterpart E-carrier system in Europe.

5.1.1 Digital Transmission Basics

The first-generation digital transmission system was developed and deployed in the early 1960s. It works by first transforming analog signals into digital format (0s and 1s) and then transmitting the digital signals over copper wires. The continuous values of analog signals are represented by discrete 1s and 0s.

There are three key components of a digital transmission system. The first component is a method for converting analog signals into digital signals known as *pulse code modulation*. The second component is a scheme for packaging digital information into recognizable units for transmission and processing. This is known as the *digital framing method*. The third component is a scheme to combine multiple separate streams of digital data on a signal physical wire for transmission and then separate them back into their original streams at the receiving end.

5.1.1.1 Pulse Code Modulation Pulse code modulation converts analog signals of voice speech into digital signals, the first step required for a digital transmission system. There are four steps involved in the PCM process (Gast 2001):

1. *Filtering.* This step filters the frequencies of analog signals that are beyond the human voice range, i.e., between 300 and 3400 Hz, to remove noise.
2. *Sampling.* This step samples analog signals 8000 times per second and outputs pulse amplitude-modulated (PAM) signals, which can be viewed as "normalized" analog signals.

3. *Quantizing.* This step converts the analog to digital signals. Each input PAM signal is quantized into one of 255 amplitudes and then assigned a number between 0 and 127 according a quantizing scale indicating its voice amplitude.
4. *Encoding.* This step encodes the assigned numbers into a digital bit stream (1s and 0s) according to a defined encoding rule.

5.1.1.2 Digital Framing and Encoding Framing, a fundamental step in digital transmission, is a process of packaging a continuous bit stream into recognizable units that can be received and processed appropriately by the receiving end of a transmission system (Black and Waters 2002). The units, known as *frames*, can be viewed as cargo carts, each with a fixed length and width. The data rate of a digital transmission standard determines the frame structure and format.

A DS1 frame is said to contain 24 channels that are arranged into their own individual time slots. Each time slot, or channel, is a sequence of 8 bits, for a total of 192 bits. In order for the receiving equipment to identify the beginning of each frame, a framing bit is added. The framing bit is a signal to the receiving equipment that a new frame is about to be received.

Different framing methods have been developed to group continuous frames with additional information between the frame groups for the purposes of error checking, timing synchronization, and performance, among others. Two well-known framing methods of the DS1 digital transmission standard, for example, include *superframe* or *D4 framing* and *Extended Superframe* (ESF), defined by ANSI.

Line coding with efficient bandwidth utilization, clock recovery, and error checking defines methods for representing digital information with electrical pulse. For example, DS1's Binary 8 Zero Substitution (B8ZS) is an industrial line coding standard that inserts two bipolar violations and two pulses in a specific sequence whenever a bit stream contains a string of eight or more 0s. Each block of eight consecutive 0s is removed and replaced with the B8ZS code. By using this method, the accuracy of user data is continuously validated throughout transmission. Another line coding standard is known as *Alternate Mark Inversion* (AMI): All zero bits are electronically neutral; one bit alternates between positive and negative pulses on the line. Each mark signal is the electrical inverse of the last binary mark. This is sometimes referred to as a *bipolar format* because the mark signal can have either a positive or a negative value.

5.1.1.3 Digital Multiplexing Digital multiplexing, another key component of a digital transmission system, defines a scheme of transmitting more than one digital signal over a physical medium at the same time (Held 1998). Multiplexing can be viewed as interleaving multiple digital streams into a single stream for transmission in order to better utilize the transmission medium. The more digital streams that can be interleaved into one stream, the higher the data rate is. High-speed transmission enables digital multiplexing by combining multiple lower-speed streams into one higher-speed bit stream, analogous to multiple streams of water merging into a big river. Digital multiplexing determines the throughput or data rate of a transmission system.

Digital multiplexing can be performed at either the bit level or the byte level. Bit-level multiplexing interleaves bits from different input streams, 1 bit per stream, while byte-level multiplexing interleaves 1 whole byte from input streams, 1 byte per input stream in turn.

Time division multiplexing (TDM) is the multiplexing scheme used in the T-carrier, E-carrier, and SONET digital transmission systems to be discussed later in this chapter. TDM allocates an equal amount of time, known as a *time slot,* for each component input stream. All the input streams are interleaved on the time scale for transmission. For example, a DS1 frame has 24 DS0s, and 1 bit from each DS0 is sent in turn. Thus DS1 is said to have 24 time slots, one for each DS0. Other multiplexing schemes include frequency division multiplexing (FDM) and code division multiplexing access (CDMA), which are based on frequency and code scale rather than time.

5.1.1.4 Digital Timing Synchronization Timing synchronization is a crucial consideration for digital transmission. Timing (or clocking) is the use of repetitive pulses to keep the bit rate of data constant and to indicate where a 1 ends and where a 0 starts in a digital stream. A transmitter and receiver must keep pulses on the same beat in order for both parties to communicate with each other correctly. The higher the data rate of a digital transmission system, the more accurate the timing synchronization needs to be.

5.1.2 First-Generation Digital Transmission System

The T-carrier system, defined by ANSI as the digital transmission standard for North America and some other parts of the world, consists of T1,

T2, T3, and T4 carrier systems. T1 becomes the building block of higher-rate T-carrier digital transmission systems like T2, T3, and T4.

5.1.2.1 T1-Carrier The T1-carrier is the first-generation electrical digital transmission system, transporting 24 voice conversations over a single twisted-pair wire, with each conversation encoded at 64 Kbps for a total 1.54-Mbps data rate (Black and Waters 2002). The electric pulses are used to represent digital signal 0s or 1s. The number 1 in the term *T1-carrier* refers to the 1 in the data rate of 1.54 Mbps, while the word *carrier* is a telecom term for the transmission medium. The *T* in T1 stands for *terrestrial*, to distinguish it from satellite communications, which were developed at the time.

DS1, closely associated with the T1-carrier, is a digital signal standard that defines a method and rate of signals transported on a T1-carrier. The main difference between T1 and DS1 is that *T-carrier* refers to the physical wire plant (wires, repeater, plug, etc.) that carries the digital signals defined by DS1. The DS1 system defines a digital system rate, a framing scheme, a method of transmission, and a digital coding of information transmitted over the T1-carrier.

T1, since its first deployment more than three decades ago, remains the most widely deployed digital transmission system to date.

5.1.2.2 Digital Hierarchy T1 and DS1 provide a building block to build a higher-rate system whose data rate is a multiple of the DS1 data rate. By using bigger frames that are multiples of the DS1 frame and by multiplexing multiple DS1s, digital transmission systems of higher data rates are built. For example, the T2/DS2 transmission system is built on top of T1/DS1; the T3/DS3 transmission system is built on top of T2/DS2; and so on. This is known as the *North America digital hierarchy*; see Table 5-1 for its technical details.

TABLE 5-1

Digital Hierarchy of DSx and E-Carrier Systems

DSx	DS1 equivalents	Line rate	Corresponding E-carrier	E-1 equivalents	Line rate
DS0		64 Kbps	E0		64 Kbps
DS1/T1	1	1.544 Mbps	E1	1 E1	2.048 Mbps
DS2/T2	4 DS1	6.321 Mbps	E2	4 E1s	8.448 Mbps
DS3/T3	28 DS1s	44.736 Mbps	E3	16 E1s	34.368 Mbps

5.1.2.3 E-Carrier The E-carrier digital transmission system adopted in Europe and some other parts of the world is the counterpart of the T-carrier system adopted in the United States. In a similar fashion to T1, E1 provides the basic building block to build a digital hierarchy: The E2 digital transmission system is built by multiplexing four E1s, the E3 digital transmission system is built by multiplexing four E2s, and so on.

5.1.3 Digital Transmission Network Elements

A digital transport network consists of a digital transmission system and a set of network elements that perform multiplexing, switching, and digital signal boosting. Major digital transport network elements include repeaters, multiplexers, cross-connects, and digital loop carriers (DLC)(ANSI 1995a).

5.1.3.1 Repeater A repeater, also known as a *regenerator,* boosts weakened electric signals so the signals can travel far enough to reach customers' premises. Multiple factors together determine how often a repeater will be used—factors such as type of copper wires, number of wires, and the distance from a central office to a customer's premises. For example, a T1 with four copper wires needs a repeater about 3000 feet out of the central office; then repeaters are spaced out every 5000 to 6000 feet; another repeater is needed about 3000 feet before the wire reaches the customer's premises.

5.1.3.2 Digital Multiplexer and Demultiplexer A digital multiplexer, an important component of a digital transport network, performs the multiplexing function, aggregating multiple lower-speed digital streams into a high-speed stream. A demultiplexer performs the reverse function, separating two or more component streams out at the receiving end. To support two-way communications, places like central offices must have both a multiplexer and a demultiplexer, and the combination of a multiplexer and a demultiplexer is often referred to as a digital *mux.*

5.1.3.3 Digital Cross-Connects One primary function of a digital cross-connect or switch is interconnecting the traffic from different directions and directing each data traffic flow toward the desired destination. It performs functions similar to a railroad intersection where either a whole train or some cars of a train can be switched into a different direction. Cross-connects are located in the interior of digital transport networks and are connected with other transport equipment.

5.1.3.4 Digital Loop Carrier (DLC) A DLC can be viewed as a special-purpose multiplexer/demultiplexer in a transport network: It aggregates 24 channels from individual subscriber premises into a DS1/T1 or vice versa. It can be located at a central office or close to customer premises. A DLC reduces the number of copper wires going into a central office by combining the traffic of 24 channels into one T1.

5.2 SONET Basics

Synchronous optic network (SONET) is a standard digital transmission system that uses light instead of electrical pulses to transmit digital signals over optical fiber. The key components of a SONET transmission system, like that of a digital transmission system, include a SONET frame, a multiplexing scheme, and a method for time synchronization (ANSI 1995b). The European counterpart of SONET is SDH, and is described briefly in this section as well. But we first provide some background information to set a context.

5.2.1 Introduction

Commercial deployment of fiber optics as a communications medium to transport digital signals started in the 1970s. In contrast to electric transmission, which uses electric pulses to indicate digital signals (1s and 0s), optical transmission uses flashes of light to transmit the signals.

SONET is a widely deployed standard for fiber optical transmission networks. The initial optical transmission equipment was all proprietary. In mid-1980s, however, driven by the service providers, the SONET standards were defined by the Exchange Carriers Standards Association (ECSA) on behalf of ANSI, for the telecommunications industries of North America. One goal of the SONET standards is that the service provider be able to mix and match optical transmission equipment from different vendors. A slightly different version, called *Synchronous Digital Hierarchy* was defined by ITU-T for the rest of the world. The SONET standards have been one of the most successful examples of international standardization.

SONET operates at the physical layer of the OSI network reference model, responsible for the physical medium that transports digital data. The SONET physical layer can be viewed as consisting of two sublayers: photonic and framing, as shown in Fig. 5-1.

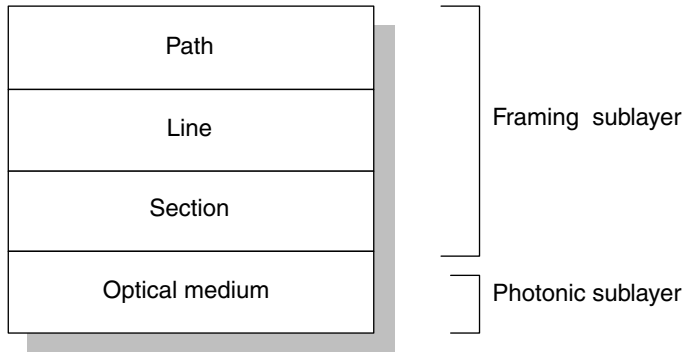


Figure 5-1
Two sublayers of the SONET layer.

The photonic sublayer is concerned with the optical transmission medium and the conversion between optical signals and electric signals. The SONET photonic sublayer operates at 1150- and 1300-μm wavelengths. The framing sublayer is responsible for packaging digital signals into transportable units to be transmitted over the optical medium. This framing sublayer is the focus of this section.

5.2.2 SONET Frame Structure

At the core of SONET are two closely related transmission hierarchies: an optical carrier (OC) hierarchy for optical transmission media and the equivalent synchronous transport signal (STS) hierarchy for electrical transmission. SONET transmission equipment has both electrical and optical components. All digital signals are first in electrical form and then mapped to optical signals. The STS framing structure provides the building block for the optical equivalent and the mapping from STS to the OC level is a function of the photonic sublayer. Thus the STS framing structure, rather than its optical counterpart, is the focus of this section.

The STS frame, just like the frame of the DS1 digital transmission standard, is the unit of digital signal transport and defines the amount of digital information that can be multiplexed onto a physical optical fiber medium. The SONET hierarchy has a base unit, and higher-order SONET frames can be built upon multiples of that base unit (Goralski 2001; Dombrowski and Grise 2000).

5.2.2.1 STS-1 Framing Structure The basic unit or building block of SONET transport is known as synchronous transport signal level 1, STS-1.

The STS structure has two parts: a synchronous payload envelope (SPE) and a transport overhead, as shown in Fig. 5-2. The payload envelope carries data traffic, and the transport overhead section carries the management information. An intuitive way of viewing the STS structure is to look at the whole frame as a cargo cart with a width of 90 columns and height of nine rows. Out of the whole cart, 87 out of 90 columns are the payload envelope while the remaining three columns constitute the transport overhead section.

As with T-carriers, SONET adopts a frame length of 125 μs or a frame rate of 8000 frames per second. Then the STS-1 line rate is derived as follows:

$$90 \text{ columns} \times 9 \text{ rows} \times 8000 \text{ frames/s} \times 8 \text{ bits} = 51.84 \text{ Mbps}$$

The payload envelope provides protocol-independent transport service. Once a payload is built and multiplexed into the SPE, it can be transported and switched through the SONET network without the need to examine or demultiplex the content of the payload envelope at the intermediate nodes. The payload envelope thus can carry any kind of traffic, be it voice, video, or data.

Each STS-1 frame is transmitted row by row and within each row column by column, from the most significant bit to the least significant bit.

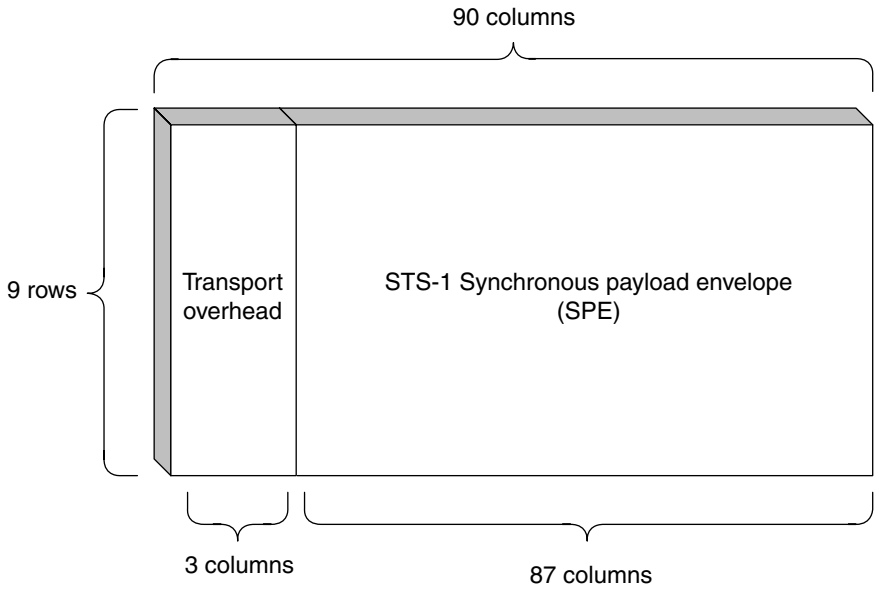


Figure 5-2
STS-1 frame structure. (Source: Ref. 8)

5.2.2.2 STS-1 Overhead Structure The SONET overhead provides a large amount of information for OAM, fault recovery, and connection maintenance. This allows communications between intelligent nodes on the network, enabling the administration, surveillance, provisioning, and control of a network from a central location.

The overhead structure consists of three portions: path overhead, line overhead, and section overhead. The path overhead is a part of the STS-1 payload envelope, while the section overhead and line overhead constitute the transport overhead of an STS-1 frame that occupies the first three columns of a STS-1 frame, as shown in Fig. 5-2. The overhead structure is designed to support the SONET connection hierarchy, which consists of three parts: SONET section, SONET line, and SONET path. The SONET connection hierarchy is introduced in the next section.

The STS-1 path overhead contains the information about the status and performance of a SONET path and has 9 bytes of data that occupy row 1 to row 9 of the first column of the STS SPE. It supports functions such as STS-1 SPE performance monitoring, path status reporting, path trace, and signaling label, while a SONET path is an end-to-end SONET connection. The path overhead data is generated and processed by SONET path terminating equipment like a SONET multiplexer.

The line overhead with 18 bytes of data is generated, processed, and maintained by SONET line terminating equipment such as SONET add/drop multiplexer (ADM) and cross-connects. The information included in the line overhead data includes the line performance monitoring, location of the SPE in the frame, multiplexing or concatenating signals, line maintenance, and automatic protection switch, which will be discussed later.

The STS-1 section overhead with 9 bytes of total capacity that occupies the first row of columns 1 through 9 of the STS-1 frame contains the data that is generated, processed, and maintained by section equipment such as a SONET regenerator. The section overhead data includes frame bytes that mark the beginnings of STS-1 frames, section data communication channel (DCC) bytes that contain the section OAM&P data, and section performance monitoring data.

5.2.2.3 STS-*N* The SONET frame is a flexible structure that allows higher-rate frames to be constructed from STS-1 frames. A higher-order signal can be constructed by interleaving multiple STS streams.

An analogy will help explain the construction of an STS-*N* frame. Imagine that there are *N* STS-1 byte streams merging into one big STS-*N* byte stream. Each STS-1 stream takes its turn to contribute 1 byte at a

time starting from the upper left-most byte of the STS-1 frame, until all the streams give out all their bytes, as shown in Fig. 5-3. Or imagine multiple freight trains coming into a station. A cargo car of each train is taken from each incoming train one at a time to form a new long cargo train.

The STS-*N* frame is structured in a modular way. There is a pointer for each STS-1 frame to point to the beginning of each STS-1 synchronous payload envelope, and therefore the contained STS SPEs do not have to be aligned. Another usage of this structure is that individual STS-1s can be added or dropped along the way, serving customers with smaller bandwidth requirements.

Currently higher signal levels that are commercially available include OC1, OC3, OC12, OC48, and OC192. OC768 is being actively researched, and it should not be long before its commercial deployment if the bandwidth growth in the past 10 years provides any indication. Note that, other than OC3, each higher level of signals is four multiples of the previous lower-level signals. For example, the data rate of OC12 is four times of that of OC3. The line rates of different STS signal levels with the corresponding OC signal levels and the DSx equivalents are listed in Table 5-2.

5.2.2.4 STS-*Nc* Frame STS-*Nc* features a frame structure that is slightly different from that of STS-*N* to achieve a better user data rate. STS-*Nc* SPE consists of $N \times 87$ columns and nine rows of bytes that are switched and transported as a single unit. In contrast, STS-*N* frames are modular

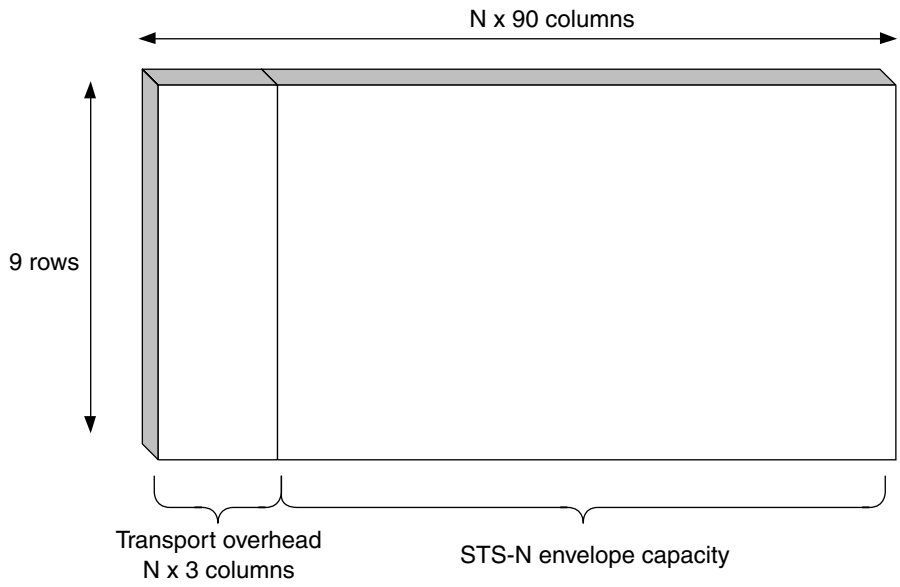


Figure 5-3
STS-*N* frame structure.

TABLE 5-2

SONET Hierarchy and OC-*n* Comparison

STS-*N*	OC-*N*	Line rate	Equivalent DSx
STS-1	OC-1	51.8 Mbps	28 DS1s or 1 DS3s
STS-3	OC-3	155.5 Mbps	84 DS1s or 3 DS3s
STS-12	OC-12	622 Mbps	336 DS1s or 12 DS3s
STS-48	OC-48	2.5 Gbps	1344 DS1s or 48 DS3s
STS-192	OC-192	10 Gbps	5376 DS1s or 192 DS3s
STS-768	OC-768	40 Gbps	21504 DS1s or 768 DS3s

and allow the individual component SPEs to be added or dropped along the way. There is only one set of STS-1 path overheads in the STS-*Nc* SPE. The columns that normally would be used for path overhead can be used for user payload.

In brief, an STS-*Nc* and STS-*N*, say, STS-3 and STS-3c, are similar in that they have the same line rates and same transport overhead. They differ in that STS-3c has a single payload envelope while STS-3/OC-3 has three separable STS-1 payloads, which can be individually added or dropped. Also OC3c has a higher actual capacity of carrying user data than OC3 because the number of path overhead is 1 versus 3 as in the case of OC3.

The STS-*Nc* framing scheme was developed to support high-rate data service between point A and point B where it is not required to add or drop any individual STS-1 SPE along the way.

5.2.3 Virtual Tributary

Virtual tributary (VT) is a variant STS framing structure that supports data rates less than STS-1 (51.85 Mbps). As mentioned above, STS signals are switched and transported at the basic STS-1 rate. To accommodate the traffic of lower rates, the SONET standard also defines the digital signal at a rate below the STS-1 level, known as *virtual tributaries* (which are branches or streams that merge into a bigger stream), using the STS-1 basic framing structure.

5.2.3.1 VT Structure VT is defined to transport traffic below the rate of DS3. Four types of VTs have been defined to accommodate different digital rates. They are VT1.5, VT2.0, VT 3.0, and VT6 and their corresponding DS signals, are listed in Table 5-3.

TABLE 5-3

Virtual Tributary Rates

	Rate	Approximately mapping to	DS signal rate
VT1.5	1.728 Mbps	DS1	1.544 Mbps
VT2	2.304 Mbps	E1	2.048 Mbps
VT3	3.456 Mbps	DS1C	3.125 Mbps
VT6	6.912 Mbps	DS2	6.321 Mbps

Two techniques are key to virtual tributary: VT pointer and bit stuffing. VT payload pointers provide a method to dynamically align a VT SPE within a VT super frame. The bit stuffing technique makes it possible to synchronize various low-speed signals to a common rate before multiplexing.

A VT can be multiplexed into a higher-speed STS-*N* frame but stay visible within a higher-rate frame. An individual VT containing a DS1 can be extracted without demultiplexing the entire STS-1. This provides the flexibility needed to supply services to customers requiring a wide range of data rates.

5.2.3.2 VT Group VT groups have been defined to accommodate mixes of different VT types within an STS-1 SPE. An STS-1 SPE is divided into seven VT groups, with each group allocated 12 SPE nonconsecutive 12 columns by nine rows of bytes in the STS-1 payload envelope columns. Seven VT groups will take up $7 \times 12 = 84$ columns out of the 87 SPE columns. Remember that the first column of an SPE is the path overhead and other two columns are fixed stuff. An STS-1 can have a combination of different groups. But one group can contain only one type of VT, with the following possible combinations:

- Four VT1.5s with three columns per VT1.5
- Three VT2s with four columns per VT2
- Two VT3s with six columns per VT3
- One VT6 with 12 columns per VT6

5.2.3.3 VT Mapping There are two different ways of mapping a user data payload into a VT. Locked-mode VTs bypass the pointers with a fixed byte-oriented mapping of limited flexibility. Floating-mode mappings use the pointers to allow the payload to flow within the VT payload.

There are three different types of floating-mode mapping: asynchronous, bit-synchronous, and byte-synchronous.

5.2.4 SONET Multiplexing

SONET multiplexing, like the multiplexing of digital transmission systems discussed above, is a process of combining multiple lower-rate data streams into one higher-rate data stream to achieve high data throughput and better utilization of the transmission resource. Specifically, the SONET multiplexing process consists of the following general steps (Held 1998):

Mapping. This is an operation of mapping the user data into VT or STS-1 frames and adding stuffing bits (when necessary) and path overhead information.

Aligning. This is a process of locating and then marking the first byte of a virtual tributary with a pointer in the STS path overhead or VT path overhead. This is also a step in preparing the data to be multiplexed.

Interleaving. This is the actual "multiplexing" operation that interleaves multiple lower-order signals into a higher-order signal. There are two types of interleaving: one-stage and two-stage. The two-stage interleaving process accommodates STM-1, the basic European rate of the ITU-T SDH standard, which is equivalent to an STS-3. The STM-1 signals are first byte-interleaved to form STM-3s, and then the STM-3s are byte-interleaved to a higher-order STM-N signal. The one-stage interleaving process directly byte-interleaves N STS-1s to form an STS-N signal without first creating an STS-3 signal.

Concatenation. This step is used for two-stage interleaving. It ensures that combined signals are wholly contained within the boundary of STS-3. If the combined signal data rate is less than or equal to an STS-3c, the combined signal must be wholly contained within an STS-3 frame formed in the two-stage interleaving process. If the rate is greater than an STS-3c, all signals must be wholly contained within blocks that are multiples of STS-3 formed in the first stage of the two-stage interleaving process.

Stuffing. This is a "clean-up" step in the SONET multiplexing process. SONET frames leave rooms to accommodate asynchronous VT signals, and the unused portion of the space must be filled with meaningless "fixed stuff" for a frame before transmission.

Scrambling. The scrambling process applies a formula to the multiplexed digital signal to make the data appear more random. The goal is to guarantee data density and to increase transmission efficiency, because SONET network element (NE) is required to have the capability of deriving clock timing from an incoming optical carrier-*N* (OC-*N*) signal and a signal of constant density leads to clock accuracy. The scrambling step takes place after the multiplexing/interleaving step and before the data transmission.

The other side of the multiplexing process is demultiplexing, the reverse of signal interleaving. Component lower-order signals are extracted from a higher-order signal and distributed to users. The SONET multiplexing process is shown in Fig. 5-4, with the left half of the figure showing the multiplexing and the right half showing the demultiplexing process.

5.2.5 SONET Timing Synchronization

The word *synchronous* in the name SONET is a key concept. Synchronous transmission of digital signals means that all transmissions, even geographically dispersed, happen at the same rate. A simple analogy would be if dancers all over the country danced at exactly the same drum beat so that their moves were synchronized. Note that the word *synchronous* does not mean the signals occur at the same time in terms of a wall clock. Rather it refers to the rate of the pulses in signal transmission.

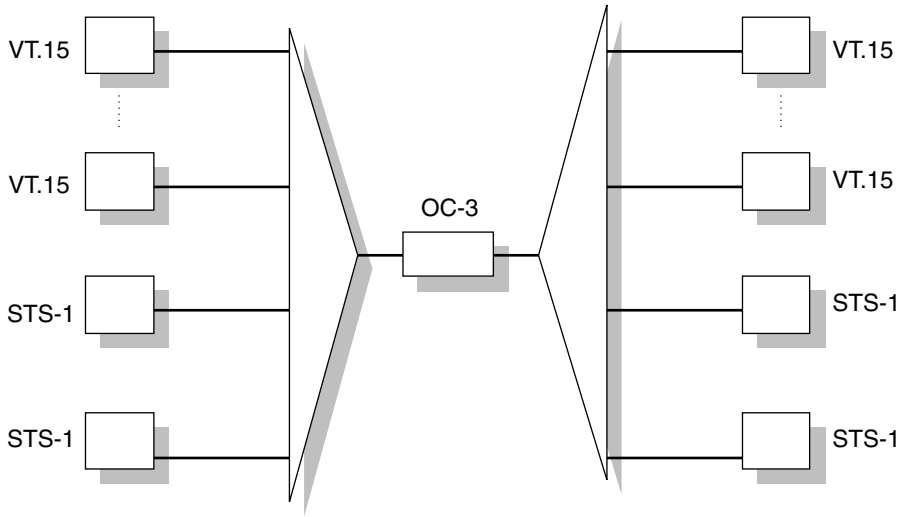


Figure 5-4
The illustration of SONET multiplexing process.

There are three types of communication network based on timing synchronization: synchronous, asynchronous, and plesiochronous. If the transmission of two digital signals is at "almost" the same rate, they are said to be plesiochronous. In asynchronous networks, the transmission and reception of digital signals may not occur at the same pulse rate.

Timing synchronization is crucial for digital communication, and even more so for a high-speed communication network like SONET. Synchronization makes it possible for a transmitter and a receiver to know where a data frame starts and where it ends in order to properly encode and decode the data. Thus synchronous communication minimizes data rate variation by reducing data transmission errors. It also makes multiplexing simpler. Because of its synchronous nature, a digital hierarchy can be built by simply stacking the base signal frames (STS-1) together to form a higher-rate frame.

SONET uses the existing synchronization network, which works as follows: A single clock is designated as the building integrated timing supply (BITS) clock that will have the highest accuracy of any clock in the building, with all other equipment in the building receiving its timing reference from this BITS clock. This BITS clock receives its own timing reference via a DS1 signal from an external source that has higher accuracy. Eventually, all timing references in the building will trace back to a single clock source, known as the primary reference source, also known as the stratum 1 clock, which has the highest accuracy. Different types of equipment have different clock accuracy requirements in terms of stratum level. The clock accuracy requirements for the different stratum levels are listed in Table 5-4. The derived clock slip rate provides another measure of the accuracy of a clock.

A standard SONET NE must support all three major timing modes that define the source of clock timing (ANSI 1999):

External timing. The timing reference is based on an external timing synchronization network that distributes the timing signal based on the DS1 signal as described above.

Free run clock. The timing reference is generated from its own local clock. This mode is recommended only when the external clock source is disrupted since different SONET NEs have different accuracy and stability requirements.

Line time or loop time. The timing reference is derived from an incoming OC-*N* signal.

A SONET network can have clock deviations, known as *phase variations*. There are two types of phase variation: jitter and wander. Clock jitter can

TABLE 5-4

Stratum Clock Accuracy Requirements

Stratum	Minimum accuracy	Slip rate
1	10^{-11}	2.523/year
2	1.6×10^{-8}	11.06/day
3	4.6×10^{-6}	132.48/h
4	3.2×10^{-5}	15.36/min

be caused by multiplexing and regenerator equipment like an ADM or optical repeater. Clock wander may come from temperature differences in different portions of an optical cable, which may cause variations in signal propagation delay. Drift in regenerating laser wavelength and instability in the timing reference can also lead to clock wander.

Jitter is dealt with by means of buffer and filtering and the frame pointer adjustment mechanism of SONET NE. Clock wander is dealt with by tracking the incoming signal and passing the signal on to a phase-locked loop.

5.2.6 SDH

Following the development of the SONET standard by ANSI in North America, ITU-T (then called CCITT) started the efforts of defining an international standard for optical fiber transmission network. Those efforts resulted in the publication of the Synchronous Digital Hierarchy standard in 1989. SDH adapts the SONET framework to the European digital signal rates (ITU-T 2000). This section provides a brief overview of SDH.

The SDH framing structure is very similar to that of SONET, with each frame consisting of two parts, i.e., synchronous payload envelope and path overhead. However, SDH uses a slightly different signal hierarchy. In place of SONET STS signal levels, SDH defines synchronous transfer module (STM) signal levels, with the STM-0 converging at the line rate of 51.8 Mbps, the same as STS-1. However, the base signal level of SDH is STM-1, which is equivalent to SONET's STS-3. A comparison between the signal levels and line rates of SONET and SDH is given in Table 5-5.

SDH defines a set of lower-speed signals called *virtual containers* (VCs), which are directly mapped to SONET VTs, as shown in Table 5-6.

TABLE 5-5

SONET and SDH Digital Hierarchy Comparison

SONET signal	SDH signal level	DSx equivalent	E-carrier equivalent	Line rate
STS-1, OC1	STM-0	28 DS1 or 1 DS3	21 E1s	51.8 Mbps
STS-3, OC3	STM-1	84 DS1 or 3 DS3	63 E1s or 1 E4	155.5 Mbps
STS-12, OC12	STM-4	336 DS1s or 12 DS3	252 E1s or 4 E4s	622 Mbps
STS-48, OC48	STM-16	1344 DS1s or 48 DS3	1088 E1s or 16 E4s	2.5 Gbps
STS-192, OC192	STM-32	5376 DS1s or 192 DS3s	4032 E1 or 64 E4	10 Gbps

TABLE 5-6

SONET VT and SDH VC Mappings

SONET VT	SDH VC	Bit rate
VT1.5	VC11	1.54 Mbps
VT2	VC12	2.05 Mbps
VT3		3.15 Mbps
VT6	VC2	6.31 Mbps

5.3 SONET Network Elements

A SONET transport network consists of optical fiber and three types of transmission equipment. This section first discusses the three types of SONET network equipment in the context of an end-to-end SONET connection and then describes in more detail the SONET add/drop multiplexer and SONET cross connects.

5.3.1 SONET Connections

A SONET transport network consists of three types of equipment: section terminating equipment, line terminating equipment, and path terminating equipment. An end-to-end SONET connection, referring to a connection from one end of a SONET network to the other, consists of sections, lines, and paths, as shown in Fig. 5-5.

A SONET section is a connection between two pieces of SONET section equipment. Examples of SONET section equipment are optical regenerators that strengthen or boost weakened optical signals.

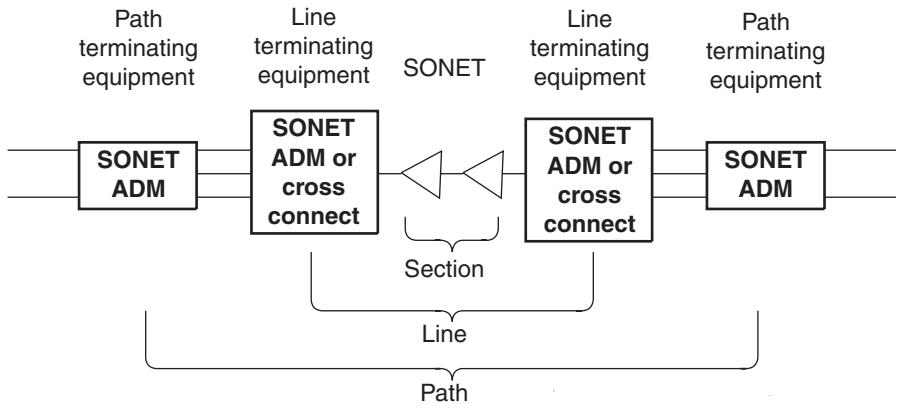


Figure 5-5
An end-to-end SONET connection.

A SONET line is a segment of an end-to-end connection between two pieces of STS line terminating equipment. A line can consist of one or more sections. Line terminating equipment can originate and terminate line signals in the STS-1 payload envelope. It can originate, access, modify, or terminate the line overhead. Examples of STS line terminating equipment include SONET cross-connect equipment.

A SONET path is an end-to-end connection that consists of one or more SONET lines, between two pieces of SONET path terminating equipment. Path terminating equipment can originate, access, modify, or terminate the path overhead. Path terminating equipment is at the edge of a SONET network. For examples, it can assemble 28 DS1 signals, insert path overhead, and form an STS-1 signal. An example of SONET path terminating equipment would be a SONET add/drop multiplexer/demultiplexer.

5.3.2 SONET Add/Drop Multiplexer

SONET ADM is a key piece of SONET transport network equipment that performs two basic functions, as its name suggests (Tektronix 1997):

- Adding, or mutiplexing lower-rate signals (VTs) into a higher-rate signal (OC1, OC3). It can be thought of as merging multiple smaller pipes of byte stream into a bigger pipe.
- Dropping, or demultiplexing a lower-rate stream from a high-rate bit stream. It can be thought of as branching off a local byte stream from a big pipe.

These basic functions of ADM can take on different variations that include

- Drop and continue
- Hairpin
- Multidrop
- Multidrop and continue

A SONET ADM can drop traffic at the rate of a single VT at a local node, without demultiplexing a whole STS-1, because the synchronous framing of SONET make visible the traffic at the DS1 (VT1.5) level.

A SONET ADM can be located at the edge or the center of a SONET network. When used as a terminal ADM at the network edge, it originates or terminates SONET traffic. This type of SONET ADM requires an interface to the non-SONET side of a network to perform the mapping of customer data into STS envelope payloads or vice versa in the other direction. The customer data may take the form of ATM cells, Ethernet frames, or FDDI frames, among others.

An ADM configured as an intermediate node of a SONET network simply performs the variations of add/drop functions as described above.

5.3.3 SONET Cross-Connect and Switch

A primary function of the SONET cross-connect or switch is interconnecting traffic from different directions and directing it all toward the intended destination. Cross-connects or switches are located in the interiors of a SONET networks and are connected with other SONET equipment. Specifically, the functions of a SONET cross-connect include

- Interconnecting a large number of STS-1 signals.
- Traffic grooming, which either aggregates or segregates STS-1s. A SONET cross-connect can combine traffic streams from different locations onto one facility. Conversely it can branch incoming traffic streams into different locations. A SONET cross-connect can be viewed as an intersection train station where traffic coming off one train can be dispersed onto trains leaving for different directions. Or conversely it can be viewed as traffic from multiple different directions converging onto one train.

SONET cross-connects can operate at different granularities and thus can be divided into wideband cross-connects and broadband cross-connects. The interconnection at the finer granularity of the DS1 (VT1.5) level is known as *wideband SONET cross-connects.* A cross-connect operating at the STS-1 level is said to be a *broadband cross-connect.*

5.4 SONET Network Configuration

SONET ring configuration helps SONET achieve a very high degree of reliability and has facilitated the wide deployment of SONET in telecom networks where fault-tolerance is a must (Gorshe 1999). This section discusses various SONET network configurations, with an emphasis on the SONET ring.

5.4.1 Linear Configuration

Linear configuration simply connects a set of SONET ADMs in a chain without connecting the two ends, as shown in Fig. 5-6. Applications of this configuration include connecting a group of central offices or LANs in a metro area. The traffic from one node can be sent to one or more other nodes on the chain via add and drop operations.

The advantage of this configuration is its simplicity and low maintenance. The disadvantages include its lack of resilience in case of a facility fault like a cable cut.

5.4.2 Hub Configuration

The hub configuration features a set of traffic-generating/terminating access nodes that are connected to a common node that can cross-connect and switch traffic between the connected nodes. The applications of this configuration includes a set of LANs that are interconnected via a SONET cross-connect node in a metro area.

The advantages of the SONET hub configuration include the efficient point-to-point connection between nodes and easy expansion of the network. A point-to-point connection between two nodes can be achieved via a central common node that performs cross-connect functions

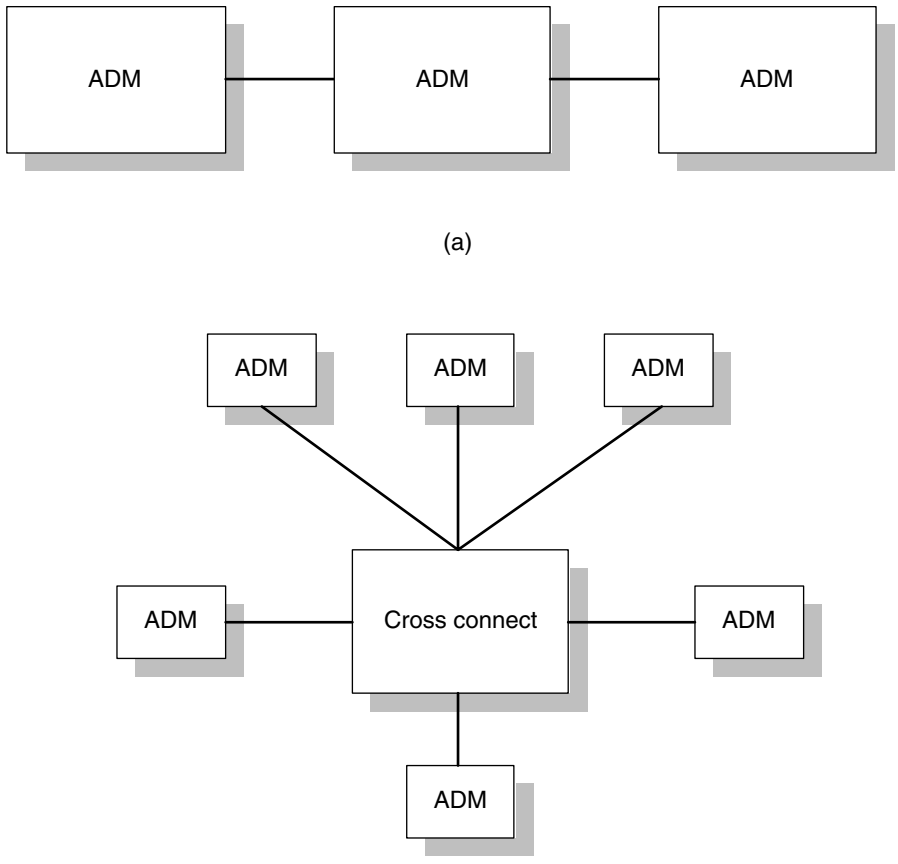


Figure 5-6
SONET topological configurations.

without disturbing any other nodes. In case of fast network growth, additional nodes can be easily added.

The disadvantage of the hub topology is same as that of the linear configuration—its low network survivability in case of a facility fault.

5.4.3 Mesh Configuration

In a mesh configuration, one traffic-generating node is connected to multiple other nodes. This configuration is desirable if there is a heavy point-to-point traffic. One advantage of the mesh topological configuration is the survivability of the network in case of a facility fault at any

given link. Traffic can be routed onto an alternative route to reach the destination node. A related advantage is network-wide traffic balancing: When one link is heavily congested, an alternative link is used. Routing traffic onto a different link requires that each node have the intelligence to switch and cross-connect traffic at the SONET layer.

One application of mesh topology is to interconnect the central offices of one service provider into a high-capacity SONET network. As the optical transport layer becomes smarter, the mesh topology is becoming a real alternative to the popular ring configuration. More on this will be discussed shortly.

5.4.4 Ring Configuration and SONET Reliability

The ring configuration is by far the most popular deployment strategy in metro markets up to now. It is composed of a handful of nodes, two fiber links, and possible interconnection with other rings. One link carries the working traffic while the other lies dormant until needed in case of an outage. In other words, the second ring is used for protection or restoration. The nodes forming a ring typically are SONET ADMs, although SONET cross-connects can also be used. One ring can be interconnected to another, forming a cascade of SONET rings. The most appealing feature of the ring configuration is its ability to recover quickly from a network failure such as a cable cut. Thus, SONET rings are also called *self-healing rings*.

The SONET standard specifies two types of rings to support different protection schemes: unidirectional path-switched rings (UPSR) and bidirectional line-switched rings (BLSR) (ANSI 2000).

On a unidirectional path-switched ring, the newly added traffic or incoming traffic at a node is routed on both fiber rings, which carry the same traffic in the opposite directions. One ring is designated as the *live path* to carry the user traffic to the destination node. The other ring, the protection ring, carries the same copy of data at all times. When any node on the ring detects a network failure, it uses the SONET path layer to signal to other nodes and trigger protection actions.

On a bidirectional line switched ring, user traffic is transmitted and received on the same route in opposite directions on two separate fibers. When a node detects a network failure at a fiber link or another node, it uses line level signaling (in the line overhead) to signal other nodes about the failure and to trigger protection switching. A BLSR can have

two or four fibers. In a two-fiber BLSR, half of the bandwidth is reserved for protection purposes. In a four-fiber BLSR, two are working fibers and the other two provide protection.

Both UPSR and BLSR define a way to use dual rings to provide protection switching. BLSR, which costs more to implement, allows the carrier to use the backup path for customer data until an outage or other problem occurs rather than having it lie idle.

If a connection to the destination is cut, the drop and continue capability of a SONET ADM will attempt alternative routes by dropping the signal to an interconnecting node. If the connection cannot be made through the intermediate node, the signal is repeated and passed along an alternative route to the destination node.

5.5 A New Generation of SONET

The SONET technology is evolving. Facing the new realities and advances in optical technologies such as DWDM, efforts are underway to evolve SONET technology. Most of these efforts have been initiated by a group of startup telecom companies. This section provides a brief overview of the efforts in three areas: mesh SONET topology, intelligent optical switching, and thin SONET.

SONET technology is over 15 years old and has many proven qualities such as high reliability, widely accepted standard status, and interoperability between products of different vendors. SONET deployment is deeply entrenched in telecom networks. However, SONET was designed for traditional TDM-based telecom applications, and when applied in data-centric applications, some of its limitations become apparent.

One limitation is its rigid telecom hierarchy, with a choice of T1 or T3 traffic, but nothing in between. The bandwidth allocation is not very flexible, jumping from 1.5 Mbps to 51 Mbps with few choices in between. In addition, the amount of bandwidth set aside for protection, up to 50 percent of total ring capacity, is a very costly overhead.

SONET's prevalent ring architecture also has its limitations. First, network upgrade is not very efficient in a ring configuration because increasing bandwidth on one link between nodes requires increasing it throughout the ring, even where not needed. It is difficult to add bandwidth selectively. Second, it is not very efficient to scale a ring topology to extend its reach. This is because each ring has a limited

circumference, so scaling up requires many rings to connect over a large area.

5.5.1 SONET Mesh Configuration

A SONET mesh configuration connects every node to every other node in a network, while a SONET ring connects nodes only to a small handful of other nodes on that link. The reemerging interest in SONET mesh configuration was motivated by several factors.

First, the mesh configuration has no distance limitations because traffic flows are switched along typical fiber line length, not rings. This gives an operator the flexibility to deploy increased trunking capacity only on the overloaded routes.

Second, SONET mesh networks can be constructed node by node as customer demand warrants, at different speeds, without the need for SONET equipment such as add/drop muxes at each point. In contrast, SONET rings must operate at the same line rate throughout the network. Increasing SONET capacity requires deploying overlay rings, which leads to "stacked ring" architectures, and requires cross-connects or matched nodes.

Better utilization of bandwidth is a third motivation. As already mentioned, the SONET ring topology keep whole fibers idle for redundancy. A mesh architecture can use intelligent switching technology to create 1:*N* protection. For example, if there are five routes out of a city, using one of the five routes for protection may be sufficient for the performance requirements. Only 20 percent of the total capacity is used for protection, as opposed to a SONET ring, where a 100 percent (1:1 protection) overcapacity is maintained for protection.

5.5.2 Intelligent Optical Switching

The solution to the preceding problem is a wave of optical switches offering strong switch and router intelligence along with a mesh network architecture. The new breed of optical equipment is aimed at making SONET more flexible and giving network operators greater control and granularity over a number of service parameters associated with end-user applications, such as QoS, bandwidth, levels of protection, and so on.

5.5.3 Thin SONET

Also called *light SONET,* thin SONET advocates light SONET layer or partial implementation of SONET layer to relax the rigid telecom rate hierarchy restriction and enhance packet data carrying capacity. One suggestion is to eliminate TDM entirely, encapsulating all traffic (IP, ATM, voice, video, and so on) in frames and running it over DWDM. SONET would be retained for basic physical layer framing and simple functions like failure notification, but QoS would be provided via other technologies.

REVIEW QUESTIONS

1. Describe the four key components of a digital transmission system and the respective functions of each.
2. Describe how a frame structure of a DSx (e.g., DS1, DS2, DS3) is related to the digital multiplexing concept.
3. Describe the basic differences between a T-carrier and a digital signal level, say, T1 and DS1.
4. What is the basic unit transported over a SONET network? Describe the structure of STS-1 and its main components.
5. Describe the main difference between an OC3 and OC3c. For what types of applications is OC3c better suited?
6. The SONET multiplexing process involves some preparation steps and some postprocessing steps. Describe some of these steps.
7. Describe the timing synchronization methods used in SONET and the three clock modes all pieces of SONET equipment are required to have.
8. Describe the basic unit transported on an SDH network and explain the virtual container.
9. Explain the differences between SONET line terminating and path terminating equipment. Provide an example of section terminating equipment, line terminating equipment, and path terminating equipment.
10. Discuss the main advantages and disadvantages of the SONET ring configuration.

REFERENCES

ANSI. 1995a. "Digital Hierarchy—Format Specifications." ASNI T1.107. Web site: www.ansi.org.

ANSI.1995. "Synchronous Optical Network (SONET)—Basic Description Including Multiplex Structure, Rates and Formats." ANSI T1.105. Web site: www.ansi.org.

ANSI. 1999. "Synchronization Interface Standards for Digital Networks." ANSI T1.101. Web site: www.ansi.org.

ANSI. 2000. "SONET—Automatic Protection Switching." ANSI T1.105.01. Web site: www.ansi.org

Black, U., and Waters, S. 2002. *SONET and T1: Architecture for Digital Transport Network,* 2nd ed. Englewood Cliffs, NJ: Prentice Hall PTR.

Dombrowski, G., and Grise, D. 2000. *ATM and SONET Basics.* Fuquay-Varina, NC: APDG Telecom Books.

Gast, M. 2001. *T1: A Survival Guide.* Sebastopol, CA: O'Reily and Associates.

Goralski, W. 2001. *SONET.* New York: McGraw-Hill.

Gorshe, S. 1999. *Handbook of Sonet Technology and Applications.* Boca Raton, FL: CRC Press.

Held, G. 1998. *High Speed Transmission Networking: Covering T/E-Carrier Multiplexing, SONET and SDH.* New York: John Wiley & Sons.

ITU-T. 2000. "Architecture of Transport Networks Based on the Synchronous Digital Hierarchy (SDH)." Recommendation G.803. Web site: www.itu.int/itu-t/.

Tektronix. 1997. "SONET Telecommunication." White paper. Web site: www.tek.com.

CHAPTER 6

WDM Networks

6.1 Introduction

Newer generations of optical networks are characterized by wave division multiplexing and optical signal switching. The term *optical network* as used throughout this chapter refers to optical networks with wave division multiplexing capability.

6.1.1 Historical Background

Fiber optics have been used in communications since the early 1960s. As the technology matures, it has become the choice medium for telecommunications transmission, not only for long-haul networks, but for metro networks, access networks, and even enterprise LANs and new residential lines as well. Optical fiber provides much higher bandwidths than copper cable and is less susceptible to different types of electromagnetic and environmental interference.

For the convenience of description, the optical network evolution can be viewed as consisting of three stages or generations. The first generation is characterized by the use of optical fiber purely as point-to-point transmission medium, an alternative to copper wire to transmit data over long distances and to increase the data rate. An optical fiber carries one optical signal or wavelength. All multiplexing and switching are still done at the electric level. SONET and SDH discussed in Chap. 5 are examples of the first generation of optical networks.

The second generation is characterized by the use of wave division multiplexing, which allows the transmission of multiple wavelengths (also known as *fiber channels*) on a single fiber simultaneously. This virtually transformed optical fiber networks into networks with unlimited transmission capacity. However, the switching of digital signals is still performed at the electric level. The light signals must be converted to electrical signals for signal regeneration and multiplexing and demultiplexing and converted back to optical signals for transmission.

The third generation of optical networks is characterized by optical signal processing. Optical signals are switched, regenerated, and multiplexed at the optical level, and there is no conversion between optical and electrical signals except at the two ends where information is produced and consumed by electronic equipment.

6.1.2 Optical Network Reference Model

The optical layer is at the physical layer of the OSI network reference model. The ITU-T recently defined an optical layer for optical networks, which is at the bottom of the network reference model and provides services to the layer above (ITU-T 2001a).

The optical layer has three sublayers, as shown in Fig. 6-1: the optical amplifier section, the optical multiplex section, and the optical channel. At the top is the optical channel, also called the *light path sublayer,* which is an end-to-end wavelength connection established across an optical network. The optical channel layer is responsible for end-to-end wavelength routing with each optical channel consisting of a number of optical links.

The optical multiplex section sublayer represents a point-to-point link along an optical lightpath. The multiplex link section is responsible for optical link routing and maintenance between two optical multiplexers. The bottom sublayer is the optical amplifier section, which represents a segment between two amplifiers along an optical link. This sublayer is responsible for segment maintenance and operations. The three sublayers of the optical layer are very much analogous to the three sublayers of the SONET layer in concept: the SONET path, line, and section.

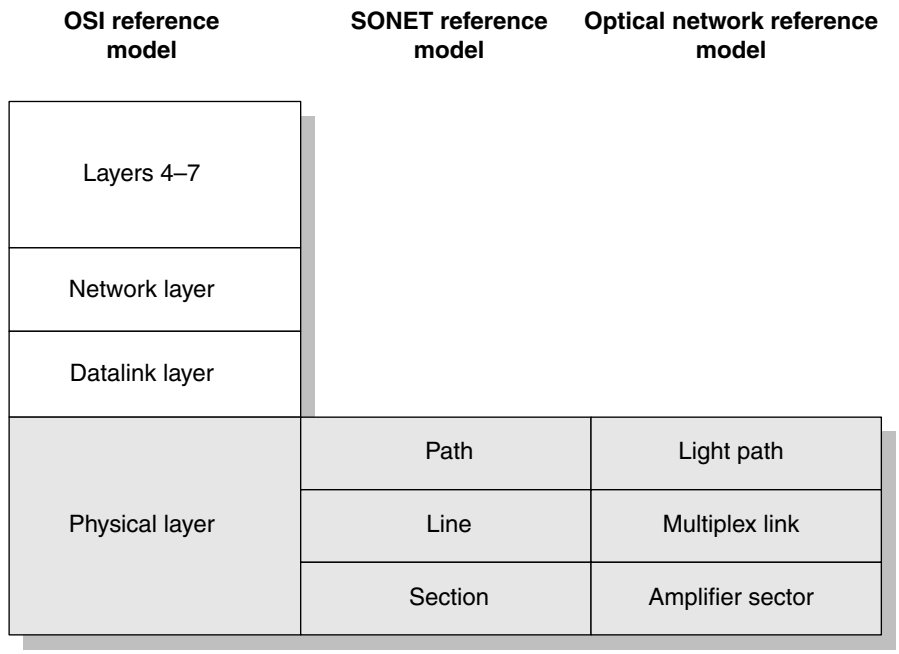


Figure 6-1
Optical layer in reference to OSI and SONET reference models.

The delineation of the sublayers of the optical layer allows the interfaces between the optical layer and the layer above and between the sublayers to be defined and standardized so the implementations of each sublayer can proceed independently of other sublayers.

6.1.3 Basics of Optical Communications

The basic idea of optical communications is to use on/off flashes of light to transmit digital signals 1 and 0 through a silica glass medium, which is also known as *optical fiber*.

A light as we see it actually consists of many different wavelengths, or different "colors." A wavelength is specified by its wave frequency measured in nanometers (nm, 10^{-9} m). The light spectrum visible to humans extends from about 400 nm, which is deep violet, to 750 nm, which is deep red. Thus the spectrum or wavelength range used optical communications, from 1310 to 1550 nm, is invisible to the human eye.

The range of wavelengths useful for telecommunications is not very large, and primarily consists of two regions, one between 1310 and 1380 nm, and the other between 1500 and 1550 nm. The wavelength region between 1550 and 1610 nm is called *long band* or *L-band*. The region between approximately 1530 and 1580 nm is called *conventional band* or *C-band*. These two regions are useful for communications purposes because they are subject to minimal optical signal loss in the transmission process.

6.2 Components of Fiber Optical System

In its simplest form, an optical transmission system consists of four basic elements, as shown in Fig. 6-2 and listed here:

- Optical fiber to carry optical signals
- An optical transmitter to emit light signals
- An optical receiver to receive optical signals
- One or more amplifiers in between to boost weakened optical signals so they can reach the receiver

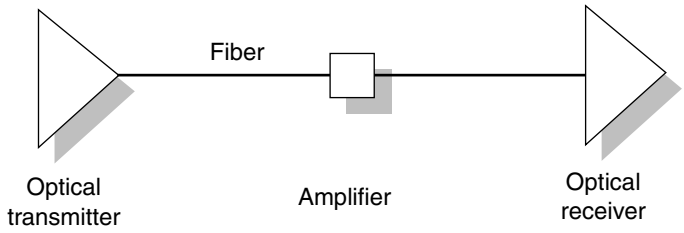


Figure 6-2
A simple fiber optical system.

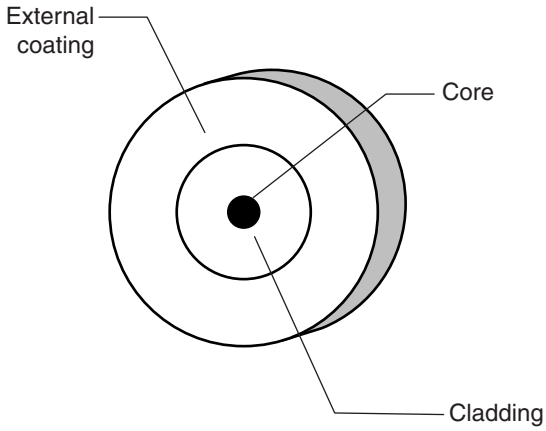


Figure 6-3
Structure of optical fiber.

6.2.1 Optical Fiber

Optical fiber is the transmission medium used to carry optical signals. An optical fiber consists of three layers, as shown in Fig. 6-3: a silica core, a layer of cladding around the core, and a coating that wraps around the cladding. The core, made of pure silica, carries the light signals. The cladding, made of less pure fiber, functions as a shield that forces the light to go straight in one direction. The outer coating, normally made of polymer material, provides protection against the elements and friction and makes the fiber easier to handle.

A fiber line that is currently in use carrying live traffic (light) is said to be "lit." Fiber not in use is said to be "dark." The term *dark fiber* generally refers to unused capacity of fiber that has been installed for long-distance applications that usually reaches up to a few hundred even a thousand kilometers without optical repeaters. The term *dark wavelength* or *dark lambda* refers to unused capacity available on a dense wave division multiplex system.

There are two types of fiber: single-mode and multimode. Each has different physical and transmission characteristics and thus is suitable for different applications.

6.2.1.1 Multimode Fiber Multimode fiber was developed in the early 1970s and is an early generation of optical fiber technology; it has a thick laser core ranging between 50 and 80 μm. Multimode optical fiber is so named because light is reflected along the core at multiple angles or paths or modes.

Multimode fiber can transmit signals only over limited distances. Because the light is propagated along multiple paths, each path has a different length and hence takes a different time to traverse the fiber. These multiple angles or modes cause the signal elements to spread out over time. Consequently, distortions occur that limit the distance over which the integrity of the light signal can be maintained.

Multimode fiber is less expensive and easier to manufacture than other types of optical fiber. Correspondingly, the accompanying connectors and active electronics, namely, the optical transceivers on the switch ports, are less expensive as well, largely because less precision is required in their manufacturing process.

Multimode fiber is best suited for short-distance, in-house wiring. Thus the predominant type of LAN fiber installed within buildings is multimode. The past three decades have resulted in a large installed base of multimode fiber.

Improvements are being made on the continuing basis to increase the performance and bandwidth of multimode fiber. Heading the improvement efforts are ISO and the TIA (Telecommunications Industry Association), which are in the process of specifying new multimode standards with higher bandwidths.

6.2.1.2 Single-Mode Fiber Single-mode fiber is a newer type of fiber developed in the early 1980s that is thinner, purer, and denser than the first multimode fibers. In contrast to multimode fiber, it is able to make use of laser technologies, using light signals that are very dense and pure. The laser core of the single-mode fiber measures between 8 and 10 μm. Light travels in single-mode fiber along a fixed path from one end to the other and can travel much longer distances than multimode fiber without signal regeneration. Single-mode fiber is commonly used between buildings, for metro access, and for long haul applications.

There are three types of single-mode fiber, each designed and engineered at a different wavelength to achieve minimal optical signal loss:

Standard single-mode fiber (SMF). This is the most widely deployed optical fiber in the United States and Europe. It is designed to achieve minimal signal loss at 1310 nm of wavelength.

Dispersion-shifted fiber (DSF). This is designed to achieve minimal signal loss at around 1550 nm. It is in common use in Japan and several other countries.

Nonzero dispersion fiber (NDF). This type of single-mode fiber is also designed around 1550 nm. Instead of achieving zero signal loss at that wavelength, it allows some signal loss to compensate for the penalties caused by the nonlinearity of DSF fiber. This type of fiber is still in an early stage of deployment.

6.2.2 Lasers and Optical Transmitters

The second component of an optical system is the optical transmitter. Laser, which stands for "light amplification by stimulated emission of radiation," is the most commonly used optical transmitter in telecommunications systems that emit light to be carried over optical fiber (Light Reading 2001).

The *laser* is a device made of semiconductor material that is specially designed to emit very precise and intense light of a particular color. A laser emits light when electric current is applied to the semiconductor material packaged inside the laser.

A laser basically consists of two specially designed slabs of semiconductor material on top of each other with another material in between. The material in the middle forms what is known as the "active layer" or "laser cavity," as shown in Fig. 6-4. When electric current flows from the top to the bottom slab, the laser cavity gives out light. Light travels along this layer until it reaches a reflective end, where it bounces back and causes more emission of light. According to the particular design, the light can be reflected a number of times along the active layer before it shoots out. This is a very basic laser design, also known as the *Fabry-Perot laser.*

Three key factors determine the quality of the light generated by a laser:

- *Semiconductor material at the laser cavity layer.* This determines the wavelength of light emitted. Different materials generate different wavelengths.

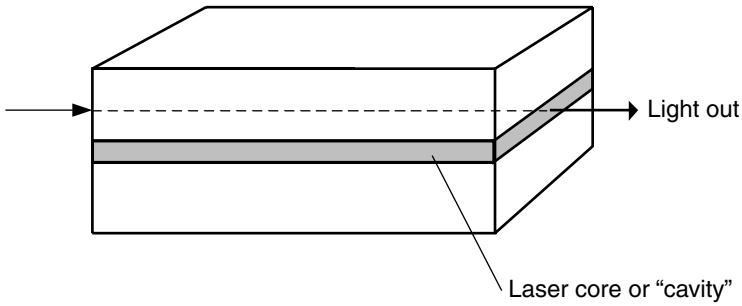


Figure 6-4
Components of a standard laser.

- *Reflective material and methods.* This material reflects light along the active layer, and along with the method used to generate the light determines how strong the light emission is and the distance the light can travel without needing to be regenerated.
- *Amount of electric current applied to the laser.* This determines how stable the lights will be and thus the amount of error there will be in signal transmission.

There are three other types of lasers, in addition to the Fabry-Perot laser, each of which promises to bring unique advantages to optical systems: distributed feedback (DFB) lasers, vertical cavity surface emitting lasers (VCSELs), and tunable lasers.

6.2.2.1 Distributed Feedback Lasers Distributed feedback (DFB) lasers produce a very sharply focused color of light. The main differentiating feature of DFB lasers is that they use a corrugated structure above the active layer of the laser to reflect only a specific wavelength of light. The corrugations serve as a grating, reflecting only a specific desired wavelength back into the middle-layer cavity and emitting a sharply focused light. A DFB layer filters out other wavelengths by allowing them to pass through the corrugated structure. This grating capability is a key function required of an optical add/drop multiplexer.

6.2.2.2 Vertical Cavity Surface Emitting Lasers Vertical cavity surface emitting lasers (VCSELs) are a type of laser that is structured differently from other types of laser. Instead of emitting light from the edge of the middle layer, the laser cavity, VCSEL emits laser light from its surface and has a vertically arranged laser cavity with reflective mirrors arranged vertically within it. One major advantage of this type of laser is that it can be tested while still being manufactured. With other types

of laser, it is not known whether the product is good or not until it is built and ready to use.

6.2.2.3 Tunable Lasers Normally one laser can emit either only one or one in a small range of specific colors of light or wavelengths. With the tunable laser, however, the wavelength the laser emits is tunable within a wide range of wavelengths. The reason for its development was to have one kind of laser that can replace any other laser if it stops working. This can result in substantial cost savings considering the hundreds and even thousands of lasers in use in optical networks.

6.2.2.4 Light Emitting Diodes Besides lasers, there is another type of optical transmitter: light emitting diode (LED). LED is a kind of passive optical device that emits light without requiring the normal maintenance or a power source like a battery. LED provides an inexpensive alternative to laser and is suitable for low data rates and short-distance communications systems.

6.2.3 Optical Receiver

An optical receiver is a device that detects the light flash or optical signal during a bit interval. If it detects a light, it interprets it as a 1; as a 0. Then the receiver converts the optical signal into a usable electrical signal. In addition, a receiver must detect any noise or interference in the signal and filter it out before converting the signal.

An optical receiver has a simple functional structure with an optical amplifier, a photon-detector, and front-end amplifier as shown in Fig. 6-5.

An optical amplifier strengthens the received optical signals before handing them over to the photon-detector for processing. The photon-detector is a key component of the receiver. At a high level, a photon-detector is a tiny device that is made of semiconductor material and can receive

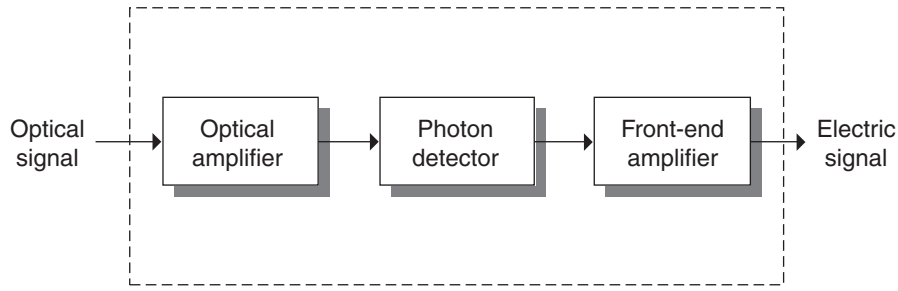


Figure 6-5
Block diagram of an optical receiver.

light. When electrical current is applied to the semiconductor material, the photon-detector turns light into electric current and thus into electrical signals.

The front-end amplifier enhances the converted electrical signals by eliminating as much as possible the noise generated in the optical-to-electrical signal conversion process. The noise can be thermal (heat-related) or some other type. Then the enhanced optical signal is fed into the photon-detector.

6.2.4 Optical Amplifier

Amplification is a key driver technology of optical communications systems. Optical signals (light flashes) diminish over distance, or attenuate. To transmit optical signals over a very long distance, the signals need to be boosted or amplified from point to point. The amount of amplification that takes place is known as "gain" and is usually measured in units of dB (dB means decibel and is calculated as one-tenth of the logarithm of the output power divided by the input power) (Ramaswami and Sivarajan 1998).

Amplification can take place at different places, and different names are given to it depending on where it takes place, as shown in Fig. 6-6. Power amplification takes place right after the optical signals are emitted from the laser. An amplifier in the middle of a link is known as a *line amplifier.* A final boost may be required as an optical signal is about to enter the optical receiver in order to reduce the chance that the signal will be replaced by a neighboring signal. Such amplification at the end of the system is called *preamplification.*

There are different types of amplifiers using different technologies. The following three types are either in wide use in telecommunications systems or are expected to play a major role in the near future.

6.2.4.1 Erbium-Doped Fiber Amplifier An erbium-doped fiber amplifier (EDFA) consists of a few meters of specially processed optical

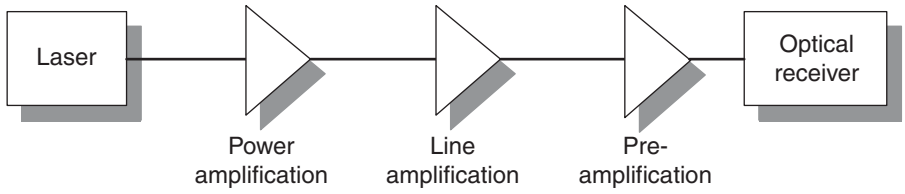


Figure 6-6
Illustration of optical signal amplification.

fiber and a special-purpose laser called a *pump laser*. The piece of fiber is doped with a few parts per million of the rare earth element erbium.

This is how the EDFA amplifier works. A pump laser emits light designed to excite the erbium ions. When a normal optical signal passes through the amplifier, it causes some of those excited ions to fall down to the ground state and give out a photon each. This *stimulated* emission boosts the optical signals because the emitted photons are at the exact same wavelength as the signal and have now become part of the signal. The signal is stronger than before because it now has more photons representing it than before. This process continues down a few meters of the fiber until lots of emitted photons have joined the signal photons and the signal has been amplified.

The EDFA amplifier can work at several wavelengths around 1550 nm. Specially designed EDFA amplifiers can also boost signals between around 1530 and 1580 nm, which as noted previously is known as *C-band* or *conventional-band amplification*. EDFAs can also be designed to give amplification between 1580 and 1610 nm, which is known as *L-band* or *Long-band amplification*. The amount of amplification at different wavelengths can vary, and there is much effort put into EDFA designs to achieve similar levels of amplification at all wavelengths, known as *gain flattening*.

EDFAs are commonly used in submarine optical systems where signals often have to travel thousands of miles underwater. This requires that the amplifiers be made in compact, waterproof packages that can be placed every 50 miles or so along the length of the system. For reasons of reliability, submarine EDFAs tend to be of very simple design. Land-based EDFAs are becoming more common now, as optical networks spread over wider distances on dry land. These systems could be designed more elaborately incorporating gain flattening and other advanced features.

6.2.4.2 Raman Amplifier Raman amplification is a new type of amplifier that uses a specialized pump laser to increase the power of light signals in optical fiber. The amplification works as follows. The pump laser adds a strong wavelength into the normal signal, and the added wavelength amplifies the signal along many kilometers of fiber until the pump signal eventually fades away. It is called *forward pumping* if the pump wavelength is inserted at the beginning of the fiber. But performance is better if the pumping is done from the far end of the fiber, known as *backward pumping* or *counterpumping*. A combination of the two is also used. For each wavelength in a WDM system, a dedicated pump laser and a pump wavelength is required. As with EDFAs, gain flattening is also an issue that requires careful design to achieve. In Raman amplifiers, the

amplification takes place throughout the length of the fiber, rather than all in one place in what is known as a "dedicated box"; therefore, it is also called *distributed amplification.*

Raman amplifiers have several advantages. They are relatively less expensive because the technology does not require special doping of the fiber; they can boost signals for longer distances; and, finally, they can boost multiple fiber channels together. So Raman amplification is emerging as a very promising technology for optical networks, and for long-haul and ultra-long-haul optical networks in particular.

6.2.4.3 Semiconductor Optical Amplifier A semiconductor optical amplifier (SOA) works in a similar way to a basic Fabry-Perot laser. The structure is very similar, with two specially designed slabs of semiconductor material on top and at the bottom and with another material sandwiched in between forming the "active layer." When an electric current is applied to an SOA amplifier, electrons on the active layer are excited, giving out photons ("particles" of light). This amplifies the signals as there are now two photons representing one particular section of a signal where previously there was only one. One major difference from standard lasers is that a laser has very reflective ends to keep light bouncing back and forth within the active layer before emitting the light. With a semiconductor amplifier, the optical signals just pass through the middle core and get amplified. Also in such an amplifier, a light is amplified at as many wavelengths as possible, because an incoming optical signal may have many different wavelengths that all need to be amplified at the same time.

Semiconductor amplifiers do not produce as much amplification as erbium-doped fiber amplifiers in the 1550-nm region of wavelengths. But this type of amplifier can work well around the 1300-nm wavelength region, and thus is very useful for optical networks using fibers of such wavelengths. As the demand for more wavelengths grows, there is a growing demand for amplifiers useful in both the 1300- and 1550-nm regions.

6.3 Wave Division Multiplexing and Optical Switching

At the heart of optical networking technologies are wave division multiplexing and optical switching. The former allows multiple fiber channels

or wavelengths to be transported on the same fiber simultaneously, thus increasing the network capacity to unimaginable limits. The latter allows the traffic on the fiber to be routed from point A to point B.

6.3.1 Wave Division Multiplexing

WDM is a technology that allows a single fiber to carry multiple colors of lights, or multiple wavelengths, at the same time. There are two basic complementary operations involved in multiplexing technology: multiplexing and demultiplexing. At the transmitting end, *n* wavelengths are combined into a single stream of wavelength to be transported over a single fiber. This is called *wavelength multiplexing*. At the receiving end, the combined stream of wavelengths is converted back to the original *n* wavelengths, as shown in Fig. 6-7 (Gandluru 1999). Each wavelength is called a *lambda* or a *fiber channel*. This works very much like time division multiplexing, the frequency division multiplexing, or code division multiplexing.

The key to wave division multiplexing is the ability to combine and separate multiple wavelengths. The optical coupler was first used to combine two light signals into one fiber in 1994. Then optical filters were developed to combine and separate wavelengths. The maximum number of independent wavelengths that can be multiplexed onto a single fiber hinges on how well the wavelength of each source is controlled and on the filters used to combine and separate the light.

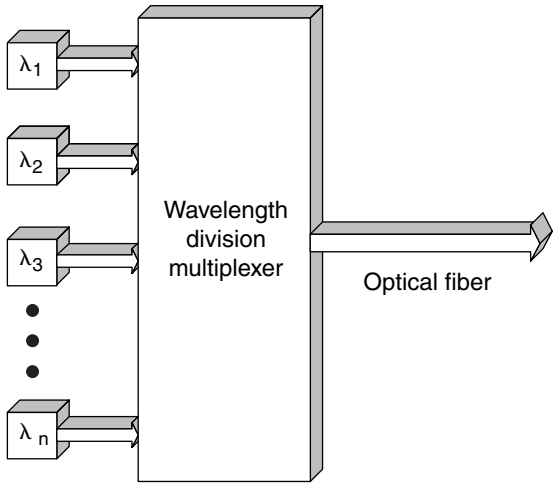


Figure 6-7
Illustration of wavelength division multiplexing.

Wavelengths must be separated from each other so an optical receiver can tell them apart. In general, the wider the spacing between wavelengths, the fewer the wavelengths that are multiplexed onto an optical fiber, and the less the capacity of the system. In contrast, the smaller the spacing between the wavelengths, the higher the capacity of the system and the more difficult and expensive to manufacture the WDM system.

The most commonly seen data rate of wavelength is 10 Gbps (over 10 billion light flashes per second!), though some other rates have also been used and 40 Gbps is being actively worked on. 10 Gbps is becoming the 64kbps equivalent of the fiber world: the optical channel data unit.

The number of wavelengths that are multiplexed onto one fiber is always a power of 2 (2, 4, 8, 16, 32, 64, 128, 256), and that number is fast approaching 256 by the last count. It becomes mind-boggling to just think about a single fiber carrying 2560 Gigabits of data per second!

6.3.2 DWDM and CWDM

Wave division multiplexing technologies can be classified as dense wave division multiplexing (DWDM) and coarse wave division multiplexing (CWDM), based on the number of wavelengths that can be multiplexed onto a single fiber. The boundary between CWDM and DWDM will be an evolving one as the WDM technology evolves. Currently CWDM usually refers to two, four, or eight wavelengths. The first generation of WDM-based networks can only carry a small number of wavelengths or lambdas on a single fiber simultaneously.

WDM systems can also be classified as wide wave division multiplexing (WWDM) and dense division multiplexing based on the spacing between the wavelengths of the WDM system. The wider the space, the less dense the wave division, the easier to manufacture, and the smaller the capacity of the WDM system. The terms *wide wave division multiplexing* and *coarse wave division multiplexing* are often used interchangeably.

Most of today's DWDM systems operate in the wavelength region of 1550 nm, where optical fiber has very low signal attenuation or loss.

A DWDM system requires that the transmission lasers have a very tight wavelength tolerance or small wavelength spacing so that the wavelengths do not interfere with each other. In a DWDM system, the wavelengths are separated by spaces based on a multiple of 0.8 nm, which is also known as *100-Ghz spacing* or the *ITU-Grid* because it is a standard wavelength spacing set by ITU-T (ITU-T 2000b). For example, five wavelengths can be

at 1549.2, 1550, 1550.8, 1551.6, and 1552.4 nm in a DWDM system because they have a 0.8 nm spacing between each of them.

A CWDM system, in contrast, features a much wider spacing between wavelengths, which is normally about 10 nm or more. Thus CWDM systems do not require accurate control of the wavelength of the laser within the transmitter, and the laser transmitters used can be low-cost basic lasers such as a Fabry-Perot or uncooled distributed feedback (DFB) laser.

6.3.3 Optical Switching Technologies

The basic idea of optical switching is very simple: switch a light from an incoming port onto an outgoing port, as shown in Fig. 6-8. The switching capability allows the optical traffic to be routed from an originating node, via many intermediate network nodes, to a final destination node.

Earlier generations of optical switches or routers all perform optical switching in an electrical core: Light pulses are first converted into electrical signals, and the routing information is processed and routing decisions are made by conventional hardware such as application-specific integrated circuits (ASICs). This hybrid optical switch has been necessary because switching at the optical level has not been mature enough during the early stages of optical networks. Also, the electrical cores make managing networks easier, because the standards governing their operation are in place and the required equipment is available.

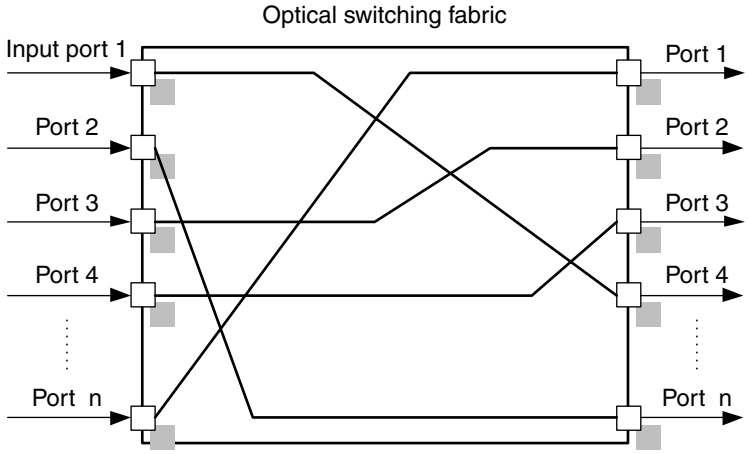


Figure 6-8
Optical switching fabric.

However, there are huge costs associated with optical-electric-optical (OEO) conversion in terms of conversion equipment, performance, and network expandability. The conversion equipment accounts for the lion's share of optical network equipment costs, found in both interfaces and shelf-to-shelf interconnections. The conversion process also introduces considerable delay in signal processing, thus slowing down system performance.

This has led to the development of all-optical switches in recent years that eliminate the need for repeated OEO conversions in a network. In addition, all-optical switches make it easier for the network carriers to upgrade equipment to take advantage of advances in technology. OEO equipment is normally tied to a specific bandwidth and particular multiplexing technology. This means, for example, when the maximum number of multiplexed wavelengths is increased from 32 to 64, all the OEO conversion equipment in a hybrid optical switch must be replaced.

All-optical switching technology is still at an early stage of development. But a wide range of optical switching technologies have been experimented with in recent years, and this section briefly describes seven of them (Light Reading 2001; Krauss 2002).

6.3.3.1 Micro-electro-mechanical System A micro-electro-mechanical system (MEMS) uses tiny mirrors to reflect the light and force it to move in a desired direction; thus it is an outgoing port. These tiny mirrors, no larger in diameter than a human hair, are placed on special pivots so they can be moved in three dimensions. Incoming light is reflected onto the mirror for a different outgoing port, effectively achieving optical switching. Each mirror can flap down to allow the light beams to pass straight over them or flap up to reflect the beam. Hundreds of such mirrors can be placed together in two- or three-dimensional arrays no more than a few centimeters square in size, to form an optical cross-connect.

The advantages of this technology include its widespread use in other industries and longer history than other optical switching technologies. Among the limitations are its low switching speed and its uncertain reliability related to the wear and tear on its moving parts (the reflecting mirrors).

6.3.3.2 Liquid Crystal-Based Optical Switching The basic idea of a liquid crystal-based optical switch is to change the orientation of light beams by applying electric current to a liquid crystal medium and then passing the light through the crystals, thus switching the light to the desired output port. This is the same technology used for laptop computer screens.

Specifically, liquid crystal devices change the polarization properties of light. This takes place in stages. First, an optical filter selects the polarization of the incoming light. Then, the light is fed into the liquid crystal, which alters its polarization. Last, the light hits a passive optical device that steers it in a different direction depending on its polarization. When an electric voltage is applied to the crystal, the molecules are pulled in line with the electric field, and they cannot steer the light any more. This achieves the effect of fixing a beam onto the desired output port.

The advantages of this approach include its relative reliability due to the absence of moving parts, its lower power consumption, and the potential for scaling up to large-size switching.

6.3.3.3 Bubble-Based Optical Switching A bubble-based optical switch uses the same idea as bubble ink jet printers. It uses tiny electrodes in the upper silicon layer of the switching devices to heat up the liquid to form a gas, called *bubbles*, which function as mirrors and bounce light onto alternative paths, achieving the effect of switching the light.

One of the appeals of bubble-based switching is its wide application in the ink jet printer market, which could lead to a low-cost solution of the optical switching problem. But the technology is still in the early developmental stage, and much still needs to be proven.

6.3.3.4 Thermo-Optical Switching The basic idea of the thermo-optical switch is to take advantage of the phase properties of light. That is, the distance unheated light travels is different from that traveled by heated light. By heating a passive splitter, the refractive index can be changed to alter the way in which it divides wavelengths between one output and another. The incoming light is split, sent down two separate waveguides, recombined, and then split one last time. One of the waveguides is heated to change its optical path length. If the two paths are the same length, the light sent down them chooses one exit, while if the lengths are different it chooses the other output port. This effectively achieves optical switching. The heaters must be separated by sufficient distance, at least 100 μm apart, so they will not influence neighboring switches.

Thermo-optical switching is at an early conceptual stage, and still requires much work.

6.6.3.5 Hologram Optical Switching The basic idea behind hologram optical switching is to create a hologram with an electrically energized Bragg grating (a piece of specially processed fiber) inside a crystal. When electrical current is applied, the hologram functions as

a mirror that can be turned on and off inside a crystal to reflect light or allow it through. Light switching is achieved by turning voltage to the hologram on or off. With the voltage on, the Bragg grating deflects the light to an output port. With no voltage, the light passes straight through. For each input fiber, a row of crystals, one for each wavelength on the fiber, is required.

Hologram optical switching, like thermo-optical switching technology, is at an early conceptual stage of development.

6.3.3.6 Electrically Switched Bragg Gratings The basic idea of the electrically switched Bragg grating (ESBG) optical switch is similar to that of hologram-based optical switching: Use Bragg grating (a piece of specially processed fiber) to deflect light onto a desired output port. It uses micro-droplets of liquid crystal suspended in a polymer layer to create Bragg gratings. Then the gratings can be turned on or off by applying electric voltage. With the voltage applied, the Bragg grating functions as a mirror to deflect a specific wavelength of light off the top of the waveguide and onto a desired output port. With no voltage, the Bragg grating disappears and the light passes straight through the waveguide without changing course.

One advantage of this approach is its ability to select a wavelength from a set of wavelengths before switching the light to a different port. Separating and selecting out a wavelength is an important step in optical switching that often is performed by a separate device.

6.3.3.7 Acoustic-Optical Switching The basic idea of acoustic-optical switching is to use sound waves to deflect light. The technology is already in use for applications such as movie screen projectors and lab equipment. It is now being adapted to optical switching. The advantages of this approach include its potential for scaling up to large-size applications, the low attenuation rate of signals it causes, and its fast switching speed.

6.4 Optical Network Configuration and Applications

The deployment of WDM networks is in its early stage, and most of the applications are in the area of long-haul networks. This section first provides an overview of WDM network components and topology, and then describes several application scenarios.

6.4.1 Optical Network Components

The two main components of an optical network are the optical add/drop multiplexer and cross-connects (Henderson 2001).

6.4.1.1 Optical Add/Drop Multiplexer The optical add/drop multiplexer (OADM) is a key component of WDW optical networks that deals with wavelengths as opposed to SONET channels. The primary function of an OADM is to insert wavelengths into and separate them out from an optical channel. An OADM uses techniques such as fiber Bragg gratings to separate wavelengths and a laser to insert them. It operates at the level of the wavelengths of light; the processing of traffic within a particular wavelength is beyond its scope.

An OADM allows the adding or dropping of a subset of a total number of wavelengths. As the number of wavelengths increases, the number of matching components such as Bragg gratings and lasers required can become prohibitively expensive. For example, if an OADM supports 400 wavelengths, 400 sets of components are required to separate and combine all the wavelengths. To reduce the number of wavelengths an OAMD has to deal with, an OAMD may group wavelengths into bands so that only a subset of bands needs to be added or dropped.

6.4.1.2 Optical Cross-Connect The primary functions of an optical cross-connect or switch is interconnecting the wavelengths arriving from different directions and routing them toward the intended destination. The switch is located in the interior of an optical network and is connected with other optical network equipment. In order to route optical signals without converting them to electrical signals, an optical cross-connect must be equipped with the ability to alter the direction of a light path.

An optical cross-connect may switch only a subset of the total number of wavelengths supported at a node. Similar to the scenario for OADM, as the number of wavelengths increases, it becomes prohibitively expensive and complicated to equip one optical cross-connect with a very large number of the components like micro mirrors to route all the wavelengths.

6.4.2 Optical Network Configuration

A WDM optical network can be configured in a ring, mesh, or other topology. The ring is a commonly used topology, as shown in Fig. 6-9.

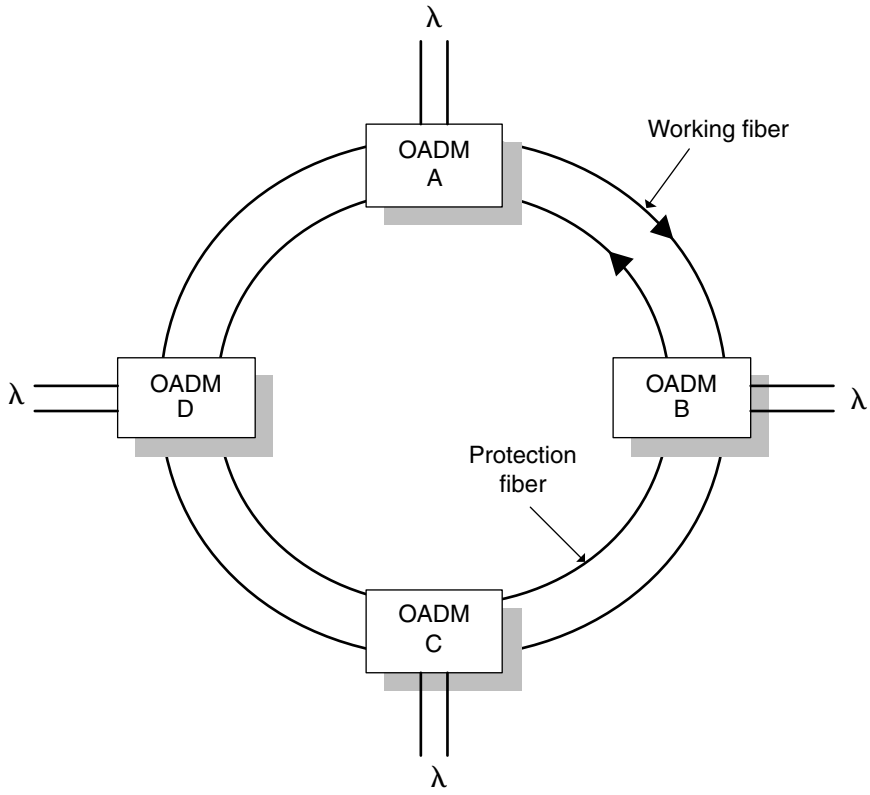


Figure 6-9
A ring topology of an optical network.

This is partly because of the familiarity with SONET ring topology on the service provider's part and partly because of the simplicity and redundancy capability of the ring topology.

6.4.3 Optical Network Applications

A variety of ways to use WDM networks to support a wide range of services, including SONET, ATM, frame relay, and IP, have been proposed and tried. Some applications are more mature than others (IEC 2000).

6.4.3.1 SONET/SDH Network over DWDM The DWDM system is used to support first-generation SONET and SDH optical networks by separating a number of wavelengths out of a fiber to feed the SONET and SDH add/drop multiplexer. A SONET ADM can run over a wavelength or a discrete fiber.

6.4.3.2 ATM over WDM One proposal is to carry ATM traffic directly over WDM systems to achieve both high data rates and QoS. Doing this would avoid the extra layer of protocol conversion needed to multiplex the ATM signals into the SONET data rate first, because the optical layer could carry any type of signals without any additional multiplexing.

6.4.3.3 IP over WDM The proposal to run IP traffic directly over WDM systems attempts to eliminate an additional layer of protocol conversion by bypassing the ATM layer. However, the IP is connectionless without QoS and requires a low-level protocol for framing, like cargo carried in a cargo container.

The primary technique for carrying IP traffic over WDM networks so far has been either using simplified SONET frames or packets over SONET with HDLC frames.

6.4.3.4 Optical Link Monitoring Optical network monitoring is critical to ensuring the correct operations of a network. Currently what can be achieved on an optical link is the monitoring of optical power and perhaps the signal-to-noise ratio. Monitoring for errors at the optical signal level is still not practical. But the established forward error correction methods such as "in-band" and "digital wrapper" can be used to help detect the errors at the optical receiver.

REVIEW QUESTIONS

1. Describe the three sublayers of the OSI optical layer for an optical network and their similarities to the SONET physical layer.
2. Describe the wavelength regions that are suitable for communications systems and explain why these regions are suitable for such applications.
3. Describe the differences between SONET/SDH networks as discussed in Chap. 5 and optical networks discussed in this chapter.
4. Describe the differences between single-mode and multimode fibers and the advantages and the suitable applications of each type of fiber.
5. Describe four different types of lasers and the characteristics of each. Describe the key advantages of tunable lasers and EFB lasers.
6. Compare and describe the differences between the two types of optical transmitter, laser and LED.

7. Describe the basic function of optical amplifiers and where they are used in optical networks.
8. Describe at a high level how each of the three prominent types of optical amplifier—EDFA, Raman, and SOA—works and the advantages of each.
9. Describe the basic operations of wave division multiplexing and the key components of an optical multiplexer.
10. Explain the concept of optical channel spacing, the optical channel data rate, and the relationship between them.
11. Explain DWDM, CWDM, and WWDM, and compare them in terms of wavelength spacing.
12. Describe the basic operations of optical switching. Then compare the two basic approaches to optical switching—hybrid OEO and all-optical switching—in terms of the advantages and disadvantages of each.
13. Describe the basic ideas of MEMS and bubble-based optical switching, and compare the advantages and disadvantages of each.
14. Explain why an OADM and optical switch may only operate on a subset of wavelengths when the total number of wavelengths becomes very large.
15. Describe how a WDM network may support ATM, IP, and SONET services.

REFERENCES

Gandluru, M. 1999. "Optical Networking and Dense Wavelength Division Multiplexing (DWDM)." White paper. Web site: www.cis.ohio-state.edu.

Henderson, P. 2001. "Introduction to Optical Network." White paper. Web site: www.mindspeed.com.

IEC. 2000. "Introduction to Optical Transmission in a Communications Network." Web site: www.iec.org.

ITU-T. 2001a. "Architecture for Optical Transport Networks." Recommendation G.872. Web site: www.itu.int/ITU-T/.

ITU-T. 2001b. "Optical Transport Network Physical Layer Interfaces." Recommendation G.959.1. Web site: www.itu.int/itu-t/.

Krauss, O. 2002. *DWDM and Optical Networks: An Introduction to Terabit Technology.* New York: John Wiley & Sons.

Light Reading. 2001. "Optical Networks-Beginner's Guide." On-line tutorial. Web site: www.lightreading.com.

Ramaswami, R., and Sivarajan, K. 1998. *Optical Networks—A Practical Perspective.* San Francisco: Morgan Kaufmann.

CHAPTER 7

Optical Ethernet

7.1 Introduction

Optical Ethernet, including Gigabit Ethernet and 10 Gigabit Ethernet (10GbE), is intended to extend the reach of Ethernet into backbone networks and metro area networks. Optical Ethernet is identical to LAN-based Ethernet with the exception of the physical layer interface. It is the physical layer interface that is the focus of this chapter. Ethernet as a LAN technology is discussed in detail in Chap. 8 on LANs.

7.1.1 Motivations for Developing Optical Ethernet

Optical Ethernet, 10 Gigabit Ethernet in particular, is a significant development that has the potential to change the networking landscape. Its development for MANs and WANs has been motivated by several factors:

End-to-end network solution. Throughout the relatively short history of packet networks, there has not been an end-to-end network solution based on one single technology. Ethernet potentially can serve as that solution. It is already the dominant networking technology in the LAN environment. Extending it to the WAN and MAN environments seems a natural next step, with optical fiber as the transmission medium. Existing customer networks can be supported in native mode, eliminating format conversions at the network boundaries. This would eliminate the large amount of the network processing that would be needed if different data link protocols were used and, therefore, would reduce the complexity of networking and increase network reliability.

Scalability of Ethernet. Ethernet has scaled from 10 Mbps to 100 Mbps (Fast Ethernet), to 1 Gbps (Gigabit Ethernet), and now to 10 Gbps. Even higher speeds beyond 10 Gbps are on the horizon and are expected to be viable in the future. Thus, network designers can start at much less than 10 GbE and build up as capacity demand expands.

Technological maturity. Ethernet is almost as old as the Internet itself and has a large installed base and a long proven history of deployment and usage. In this sense, optical Ethernet is not a new technology and can avoid the "acceptance test" that any new networking technology normally goes though. Ethernet also provides a means to avoid reengineering an existing network.

Simplicity factor. Ethernet is considered to be a "plug-and-play" solution requiring only a minimum of planning, design, and testing. A source of its simplicity is that Ethernet implementations are standardized, interoperable, and interchangeable. Network management systems can be simplified if the same system can be used at all levels of an end-to-end network. Although most networks have one customer-controlled part and another provider-controlled part, the use of common facilities makes integrated management a practical alternative.

7.1.2 Ethernet Evolution

The Ethernet evolution, based on the transmission medium used and speed achieved can be divided into the following periods:

- 10Base Ethernet, starting from late 1972 to the mid-1990s
- 100Base Ethernet, starting from the mid-1990s to the late-1990s
- 1000Base Ethernet, starting from 1998
- 10Gig Ethernet, starting from 2000

The prefix number in an Ethernet version such as 10 in 10BaseT or 100 in 100BaseT refers to the transmission speed of 10 Mbps and 100 Mbps, for example. The suffix letters such as "T" refer to the medium type. For example, the letter "T" in 10BaseT refers to "twisted pair" copper wire.

With the development of the fast Ethernet standards, the evolution of Ethernet speeded up. Soon after those standards were finalized, work began on 1000Base Ethernet, a ten-fold increase in speed, which was soon followed by the 10-Gigabit Ethernet.

7.1.3 An Overview of Ethernet Architecture

The Ethernet standards cover layer 1 and layer 2 of the OSI network reference model. That is, the physical layer specifies the nuts and bolts of the Ethernet where the cable, connector, and signaling specifications are defined. The layer 2, or data link layer, defines the ways of getting data packets on and off the wire, error detection and correction, and retransmission (Spurgeon 2000).

7.1.3.1 Physical Layer The physical layer of Ethernet consists of three sublayers: the physical medium-dependent (PMD) sublayer, physical medium attachment (PMA) sublayer, and physical coding sublayer (PCS),

as described below. The PMD defines the Ethernet cables, wiring, and other transmission medium characteristics. The PMA defines the type of connectors used to connect an Ethernet device such as an Ethernet Network Interface Card (NIC), hub, or switch to the Ethernet cable. The PCS defines the scheme for encoding and decoding data received from/sent to the PMD sublayer, using a coding scheme appropriate to the medium. Data to be transmitted over an Ethernet cable are encoded first, and then the transmission control bits are added. The decoding reverses the process: The received *n*-bit characters are stripped off the transmission control bits and the *m*-bit data characters are returned to the upper layer.

There are three basic approaches to encoding data onto the physical medium:

- *Synchronous signaling.* In addition to the digital signal, a clock signal is added in parallel so the clock signal will mark the bit boundaries.
- *Asynchronous signaling.* A special bit pattern is added to flag the beginning of a block of bits.
- *Manchester encoding.* Each bit period is divided into two halves. A "1" is defined by a transition from "low" to "high" in the middle of the bit period and a "0" is defined as a transition from "high" to "low" in the middle of the bit period.

The Manchester encoding scheme is used for 10-Mbps Ethernet. Faster Ethernet, Gigabit Ethernet, and 10-Gigabit Ethernet use successively more advanced coding schemes.

7.1.3.2 Media Access Control Sublayer This Ethernet data link layer is generally broken into two sublayers: the logical link control on the upper half and the MAC on the lower half. All IEEE 802 LAN technologies (Ethernet and Token Ring) share the same LLC sublayer, and are discussed in more detail in Chap. 8 on LANs.

The MAC sublayer is responsible for controlling multiple accesses to the shared medium. The Ethernet MAC sublayer uses the carrier sense multiple access with collision detection (CSMA/CD) scheme for medium access control. In CSMA/CD, an Ethernet station sends data only when it finds the shared transmission medium is idle and resends the data if a collision is detected.

Ethernet supports either full-duplex or half-duplex operation. Originally Ethernet only supported half-duplex operations. Half-duplex allows a station to either transmit or receive data but not at the same time while full duplex allows a station to do both at the same time.

7.1.4 Optical Ethernet Standards

The international Institute of Electrical and Electronics Engineers IEEE 802.3z and IEEE 802.3ae committees are mainly responsible for defining Gigabit Ethernet and 10-Gigabit Ethernet standards. In parallel to the formal standards efforts, the two industrial forums, the Gigabit Ethernet Alliance and the 10 Gigabit Ethernet Alliance are the main advocates of the new optical Ethernet technologies. The forums consist of Ethernet equipment vendors, service providers, and other interested parties, and focus on defining the test procedures and processes needed to achieve interoperability between products by different vendors.

7.2 Gigabit Ethernet Introduction

7.2.1 Introduction

The two-year effort of the IEEE 802.3z task force culminated in 1997 Gigabit Ethernet standards. The stated objectives of the Gigabit Ethernet include the following:

- To achieve a speed of 1000 Mbps at the MAC/PLS service interface.
- Preserve the standard 802.3/Ethernet frame format and the minimum and maximum sizes of the frame, so the Gigabit Ethernet is backward-compatible with the 100BaseT and 10BaseT Ethernets.
- To support both full- and half-duplex operation. For half-duplex operation, CSMA/CD is used. For full-duplex operation, standard flow control defined in IEEE 802.3 (IEEE 2001) is used.
- To support both fiber and copper wire as physical media. It was decided to use the recently defined ANSI Fibre Channel standards as the basis for fiber-based media.

In brief, Gigabit Ethernet focuses on the physical layer interfaces and accommodates fiber optical transmission media to take advantage of the newly developed optical technologies and achieve a Gigabit data rate. The data link layer largely remains the same as in the traditional Ethernet.

More details on Ethernet basics can be found in Chap. 8 on LANs.

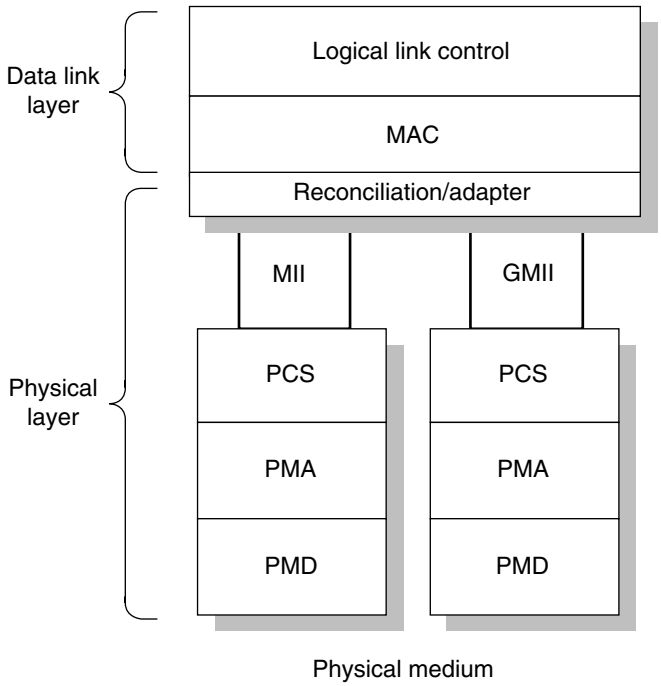


Figure 7-1
IEEE 802.3z Gigabit Ethernet overview. (Gigabit Ethernet Alliance 1996)

7.2.2 Overview of Gigabit Ethernet Architecture

The key components of Gigabit Ethernet, as shown in Fig. 7-1, consist of the three sublayers of the physical layer and an interface to the data link layer (Gigabit Ethernet Alliance 1996; Cisco 2000).

The Gigabit Ethernet can be viewed as combining two technologies: Fast Ethernet and fiber channel. As shown in Fig. 7-2, the physical interface of Gigabit Ethernet largely comes from the ANSI Fiber Channel standards originally defined in 1994 (ANSI 1999) and the data link layer standards in IEEE 802.3, much the same as the physical interface of the Fast Ethernet. Leveraging the existing technologies minimized the technological complexity of Gigabit Ethernet and enabled the Gigabit Ethernet standards to be developed quickly.

7.2.3 Gigabit Ethernet Physical Layer

The physical layer consists of the four sublayers: the PMD, PMA, PCS, and Gigabit media-independent interface (GMII). In Fig. 7-3, the corresponding physical device of each physical sublayer is shown on the right side.

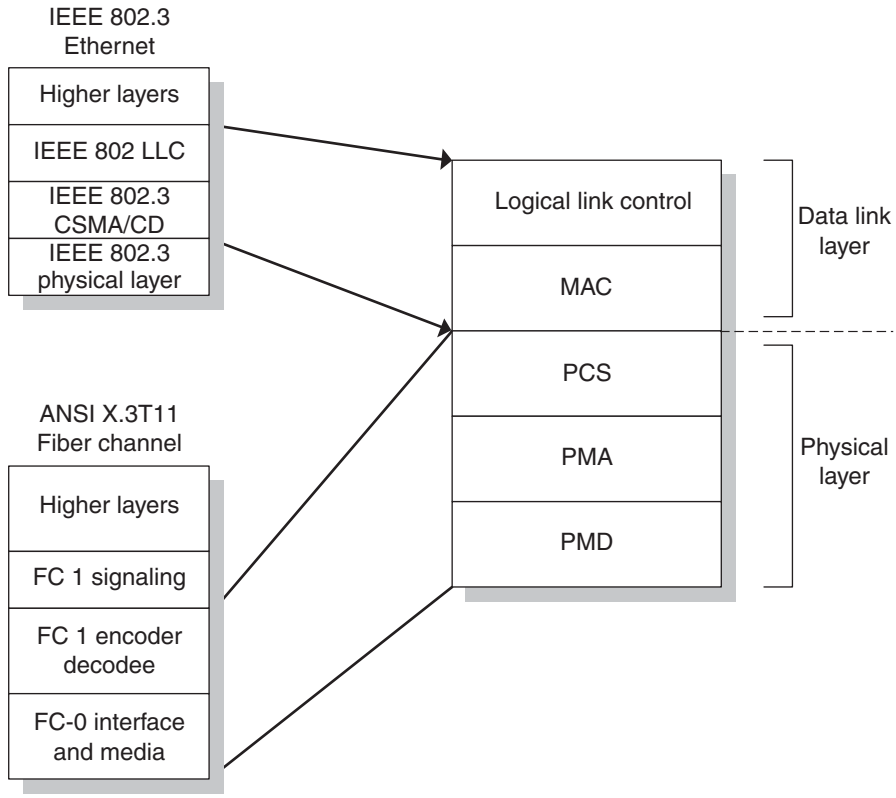


Figure 7-2
Gigabit Ethernet composition.

The first three sublayers from the bottom up are specific to the fiber transmission medium. Anything above the PCS sublayer is medium-independent. This isolates the physical medium to the bottom of the physical layer.

7.2.3.1 Physical Media Dependent The main function of the PMD sublayer is to connect the Gigabit switching equipment to the physical medium or transmission equipment. The supported connector technologies depend on the type of physical medium.

The physical media-dependent sublayer currently allows for 1.062-Gigabaud signaling in full duplex. The Gigabit Ethernet supports three types of physical medium: short-wavelength laser, long-wavelength laser, and short-haul copper. The optical fiber comes in three flavors: multimode fiber (MMF) with a 62.5-μm core, multimode fiber with a 50-μm core, and single-mode fiber.

For both SMF and MMF, the duplex SC connector is supported. For copper cable, the RJ-45 style modular jack is supported. The supported media types and specifications of each are listed in Table 7-1.

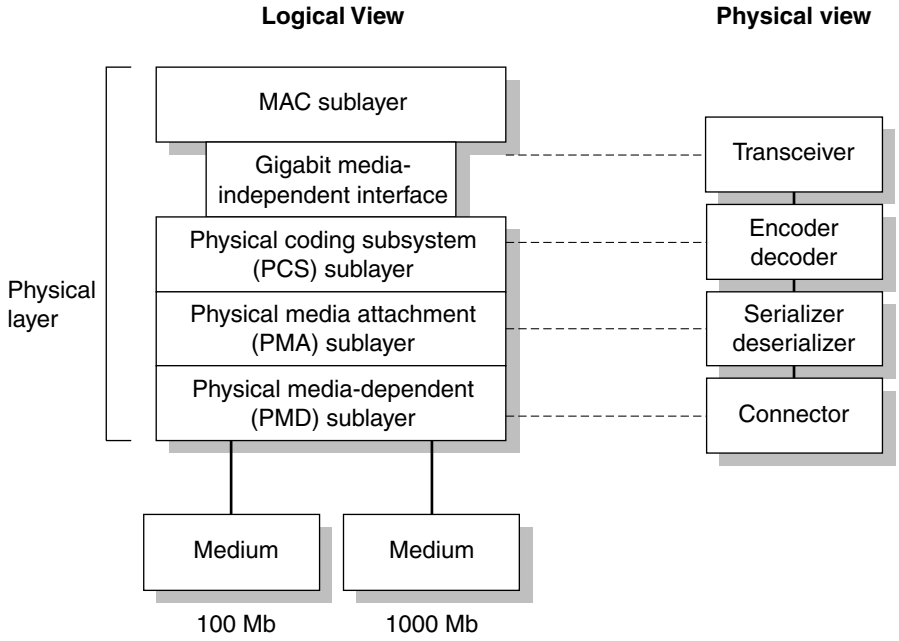


Figure 7-3
Overview of Gigabit Ethernet physical layer.

TABLE 7-1

Supported Media Type for Gigabit Ethernet

	Media type	Media specifications	Minimum distance between two repeaters
1000Base-LX	Multimode fiber (62.5 or 50 μm core)	1250–1270 nm wavelength	316 m
	Multimode fiber	1300 nm	500 m
	Single-mode fiber	1300 nm	3 km
1000Base-SX		770–860 nm wavelength	300 m
1000 Base-CX	Short-length copper	Short-haul copper	25 m
1000 Base-T	Horizontal copper	Four pairs of category 5 or better cabling, 100-ohm impedance rating	100 m

Note: L, S, and C stand for long, short, and copper, respectively. Short and long refer to the long and short wavelengths of optical fiber cable.

7.2.3.2 Physical Media Attachment The main function of PMA sublayer is to serialize and deserialize the digital signals received through connectors at the PMD sublayer into a stream of bits, supporting the encoding scheme appropriate to the medium. Normally the physical media attachment performs 10-bit serialize/de-serialize functions in a

chip. The chip receives 10-bit encode data at 125 MHz from the PCS and then delivers serialized data to the PMD when the data is sent down to the transmission link.

7.2.3.3 Physical Coding Sublayer The main function of the PCS sublayer is to encode and decode data received from or sent to the PMD sublayer, using the coding scheme appropriate to the medium.

This is one of the key components of the Gigabit Ethernet architecture that uses the fiber channel standards for physical layer signaling. Data to be transmitted over fiber is encoded first to add transmission control characters. This is an 8- to 10-bit (8B/10B) mapping process that adds an extra 2 bits per transmission character for control. The decoding reverses the process: The received 10-bit characters are stripped off the transmission control bits and the 8-bit data characters are returned to the upper layer. The control symbols are used to signal conditions such as start of a packet, end of packet, or idle.

With an 8B/10B encoding/decoding scheme, every 8 bits of user data are converted into a 10-bit symbol before transmission over fiber media. The overhead of 2 extra bits per transmission character requires a signal transmission rate of 1.25 Gigabit to transmit 1 Gigabit of user data.

7.2.3.4 Gigabit Media-Independent Interface Unique to the Gigabit Ethernet technology is a GMII that attaches the media access control of the data link layer to the physical layer functions of a Gigabit Ethernet device.

GMII is analogous to the attachment unit interface (AUI) of the 10-Mbps Ethernet and media-independent interface (MII) of the 100-Mbps Ethernet. Unlike AUI and MII, each of which has a connector to allow a transceiver to be attached externally via a cable, there is no connector defined for GMII. The transceiver function is built into most Gigabit Ethernet devices and the GMII becomes an internal component of the device. It is not desirable to expose the GMII as an external interface because of the high frequency of data transfer.

7.2.4 Gigabit Ethernet MAC

The MAC layer defines the frame format and the ways a user accesses the shared transmission media. The Gigabit Ethernet MAC can be viewed as a scaled-up version of the Fast Ethernet MAC with an effective data rate of 1000 Mbps. Gigabit Ethernet supports all standard Ethernet frame formats and is compatible with the installed base of Ethernet and Fast Ethernet products, requiring no frame translation.

7.2.4.1 Gigabit Ethernet MAC Operations The Gigabit Ethernet MAC supports both full-duplex and half-duplex operations. Half-duplex operations support shared connections, using the conventional Ethernet CSMA/CD access and congestion control algorithm.

In full-duplex mode, the Gigabit Ethernet MAC supports point-to-point connection in a switched network (versus a resource sharing network). Gigabit Ethernet MAC uses frame-flow control as defined in IEEE 802.3x for network congestion control (IEEE 2001). The flow control allows a Gigabit Ethernet switch to send a flow control message, a 64-byte packet with unique ID type, to a congested node to stop sending packets for a specified period of time. The flow control packets can operate between two devices only on a point-to-point connection and cannot operate between two devices separated by a switch.

The full-duplex transmission scheme of Gigabit Ethernet in effect raises the bandwidth from 1 Gbps to 2 Gbps on point-to-point links, because a 1-Gigabit data rate is carried each way. Full-duplex operations are ideal for applications on metro networks and for high-speed distributed servers that go beyond the conventional LAN environment.

7.2.4.2 Gigabit Ethernet MAC Extension On top of the existing Ethernet features, Gigabit Ethernet also defines a few new features to address the issues caused by the 10-fold increase in the speed. The new features include carrier extension and frame bursting (IEEE 2001; Seifeit 1999; Saunders 1998).

CARRIER EXTENSION *Carrier extension* is a technique that is used to solve the timing problem associated with CSMA/CD. Ethernet has a minimum frame size of 64 bytes. The reason for a minimum frame size is to prevent a station from completing the transmission of a frame before the first bit has reached the far end of the cable, where it may collide with another frame. Therefore, the minimum time to detect a collision is the time it takes for the signal to propagate from one end of the cable to the other. This minimum time is called the *slot time* or *slot size*.

The issue is that the Gigabit Ethernet transmission speed may be too fast for a station to detect a collision. The original Ethernet restricts the maximum cable length to 2.5 km, with a maximum of four repeaters on a given path. As the bit rate increases, the sender transmits the frame faster. As a result, if the same frame size and cable lengths are maintained, then a station may transmit a frame too fast to detect a collision at the other end of the cable. So, three options are available: either keep the maximum cable length the same and increase the slot time, or keep the slot time the same and decrease the maximum cable length, or both. In Fast

Ethernet, the maximum cable length is reduced to 100 m, leaving the minimum frame size and slot time unchanged.

Gigabit Ethernet uses a bigger slot size of 512 bytes instead of reducing the maximum transmission distance. It maintains the minimum and maximum frame sizes of the original Ethernet frame. To maintain compatibility with the previous Ethernet, the minimum frame size is not increased. If the frame is shorter than 512 bytes, then it is padded with extension symbols. These are special symbols that cannot occur in the payload and are stripped off at the MAC sublayer at the receiving end. This process is known as *carrier extension*.

FRAME BURSTING A problem that comes with carrier extension is the potential inefficient use of bandwidth. For small packets, many padding bytes (i.e., a worst case would be 512 − 64 = 448 bytes) are sent to the destination. For a large number of small packets, the throughput is not that much better than Fast Ethernet.

Frame bursting is a scheme to compensate for the low throughput caused by carrier extension. It works as follows: When a station has a number of packets to send, only the first frame is padded to the slot size of 512 bytes if necessary with carrier extension. Subsequent frames are transmitted back to back, with a predefined minimum interframe gap between frames, until a burst timer expires. Next time the station has data to send, the burst timer is set and the process starts all over again. This can substantially improve throughput in the case of many short frames.

Note that both carrier extension and frame bursting operate only at the MAC sublayer and the logical link control sublayer is not involved.

7.3 10 Gigabit Ethernet

10 Gigabit Ethernet uses optical fiber as the sole transmission medium and targets applications in backbone and metro area networks.

7.3.1 Introduction

The stated goals of the IEEE 802.3ae committee responsible for defining the 10 Gigabit Ethernet standards include the following (10 Gigabit Ethernet Alliance 2001):

- Define two families of physical layer interfaces: one for local area networks operating at a data rate of 10.0000 Gbps, and one for wide

area networks operating at a data rate compatible with the payload rate of OC-192c/ SDH VC-4-64c.

- Define a mechanism to adapt the MAC/PLS data rate to the data rate of the WAN interface.
- Preserve the standard 802.3/Ethernet frame format and the minimum and maximum sizes of the frame, so the 10 Gigabit Ethernet is backward-compatible with Gigabit, 100BaseT, and 10BaseT Ethernet.
- Support full-duplex operation only. The half-duplex for shared connection and CAMA/CD is no longer supported.

In short, 10 Gigabit Ethernet keeps the Ethernet link layer intact and focuses on the physical layer interfaces. Although the development of some details of the interfaces are still in progress, the overall frameworks are sufficiently defined for an overview.

The concept of Ethernet-based communications over long distances is unique to 10GbE and would not be feasible using the original CSMA/CD protocol (i.e., shared media with contention). The restriction to full-duplex operation allows 10 Gigabit Ethernet to operate over long link spans, repeaters, and other transport layers like DWDM or SONET.

7.3.2 Overview of 10 Gigabit Ethernet Architecture

The architecture of the 10 Gigabit Ethernet, as shown in Fig. 7-4, consists of four major components that distinguish it from its predecessors (Bruce 2001; Cisco 2001; IEEE 2002):

- Physical media-dependent sublayer
- A LAN physical layer interface
- A WAN physical layer interface
- An interface to the MAC sublayer

Two key components of the 10 Gigabit Ethernet standards are the two physical layer interfaces to support both LAN and WAN applications. The two families of physical layer interface, each consisting of the PCS, PMA, and PMD sublayers, are a LAN physical interface operating at 10 Gbps and a WAN physical interface operating at a data rate compatible with the payload rate of OC-192c and SDH VC-4-64c. The LAN interface supports 850- and 1310-nm CWDM fiber, earlier types of fiber that are commonly deployed in LAN environments. Table 7-2 lists the distance

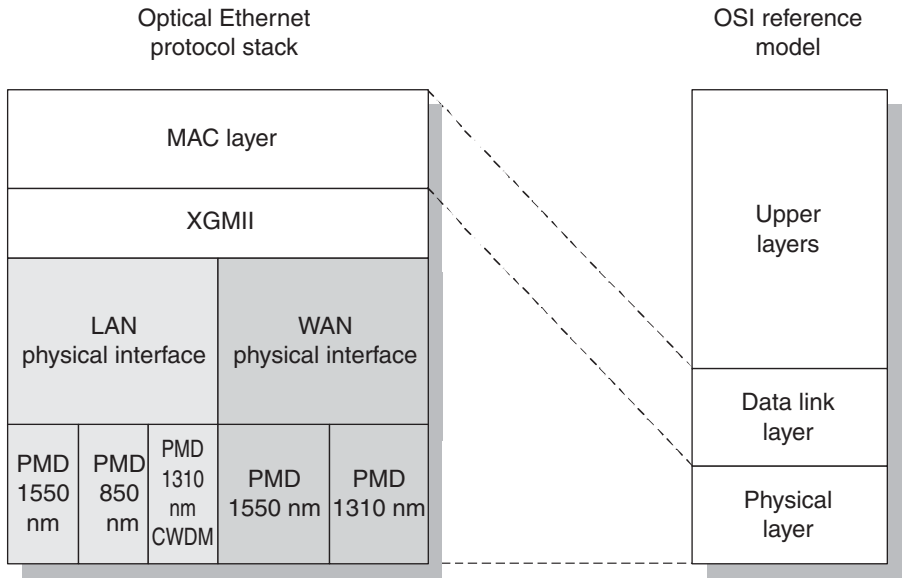


Figure 7-4
10 Gigabit Ethernet architecture overview.

TABLE 7-2

Comparison Between a LAN Physical Layer and a WAN Physical Interface

	10 Gigabit Ethernet LAN physical interface		10 Gigabit Ethernet WAN physical interface
	Serial	CWDM	Serial
MAC rate	10 Gbps	10 Gbps	10 Gbps
PCS	64B/66B	8B/10B	64B/66B SONET framing
PMA	XSBI	XAUI	XSBI
PMD	10Gbase-R 1500-nm DFB 1310-nm FP 850-nm VCSEL	10Gbase-X 1310 CWDM	10Gbase-W 1550-nm DFB 1310-nm FP 850-nm VCSEL
Line rate	10.3 Gbps	10.125 Gbps	9.953 Gbps

FP = Fabry-Pezot; DFB = Distributed feedback; VCSEL = Vertical cavity surface emitting laser.
Source: IEEE (2002).

specifications that are the current targets for 10 Gigabit Ethernet implementations. Note that the symbol PHY is the abbreviation for the physical interface in Ethernet literature.

The concept of two physical interfaces, the LAN physical interface and WAN physical interface, is a true attempt to bridge the two worlds of Ethernet-based IP world and SONET-based TDM world at the physical

layer. The WAN physical interface allows SONET frame and rate compatibility with the widely deployed SONET OC-192 transport infrastructure while the LAN physical interface supports Ethernet-based LAN applications.

7.3.3 Physical Media-Dependent Sublayer

The PMD sublayer defines optical fiber types supported for 10 Gigabit Ethernet, target transmission distances, and required optical multiplexing techniques. The main function of the PMD sublayer is to connect 10 Gigabit switching equipment to the physical medium or transmission equipment, supporting the optical transceiver functions. The PMD sublayer is identical to both the LAN and WAN physical interfaces.

10 Gigabit Ethernet supports at least two types of optical fiber interface: serial and WWDM. A serial optical interface refers to point-to-point fiber link without wave division multiplexing, and there is one light source that transmits the signal over one fiber pair using a low-cost laser. This type of link often uses older short-wavelength, 850-nm, multimode fiber.

The wide wave division multiplexing or coarse wave division multiplexing interface, as described in Chap. 6 on WDM networks, is a type of inexpensive way to multiplex multiple wavelengths onto an optical fiber. Multiple light sources such as the VCSEL array and spacing between wavelengths as large as 20 to 25 nm are used. The transmission fiber can be either single-mode or multimode fiber.

The IEEE 802.3ae Task Force specified three general PMD types to meet the targeted maximum distances. A 1310-nm serial PMD was selected to meet its 2- and 10-km single-mode fiber target. The Task Force also selected a 1550-nm serial PMD to meet its 40-km single-mode fiber objective. In practice, it has already been demonstrated by some vendors that the actual distance can go far beyond the 40-km target.

One contentious issue was whether to support older types of fiber such as multimode fiber and the associated multiplexing technologies. After a long debate, the Task Force decided to adopt multimode fiber at 850-nm wavelength as one of the PMDs. Short-wavelength fibers are inexpensive and widely used to interconnect switching equipment within central offices.

Transmission media other than fiber were not considered for 10 Gigabit Ethernet since 10 Gigabit Ethernet was not expected to connect directly to user end-systems, at least not in the near future, so the standards were initially being restricted to optical fiber. The options being considered

TABLE 7-3

Supported Media Type for 10 Gigabit Ethernet

	Type of fiber supported	Media specifications	Targeted maximum distance (m)
10 GBASE-R	Multimode	850 nm	65
10GBASE-W	Multimode	1310 nm	300
10GBASE-X	Single mode	1310 nm WWDW	10,000
	Single mode	1310 nm serial	10,000
	Single mode	1550 nm serial	40,000

Source: Gigabit Ethernet Alliance (2001).

for the optical layer include multimode and single-mode fiber using serial and parallel links. The supported media types and specifications of each are listed in Table 7-3.

7.3.4 Physical Media Attachment Sublayer

The main function of the PMA sublayer of 10 Gigabit Ethernet, as in the case of Gigabit Ethernet, is to serialize and deserialize the digital signals. The signals are received through connectors at the PMD sublayer and converted into a stream of bits, or 0s and 1s. The PMA is responsible for supporting multiple encoding schemes since each PMD will use an encoding that is suited to the specific media it supports. The PMA sublayer is the same for both LAN and WAN physical interfaces.

7.3.5 10 Gigabit Ethernet LAN Interface

The 10 Gigabit Ethernet LAN physical layer interface mainly consists of the physical link coding sublayer (PCS) and the 10 Gigabit media-independent interface (XGMII) specific to the LAN interface (Fig. 7-5). The sublayers for the LAN and WAN physical interfaces only differ in the PCS sublayer.

The main function of the PCS sublayer is to encode and decode data received from/sent to the PMD sublayer, using the coding scheme appropriate to the medium. The focus of the LAN physical layer interface is on the PCS and the interface to the PMD/PMA sublayer and the interface to the reconciliation/MAC layer.

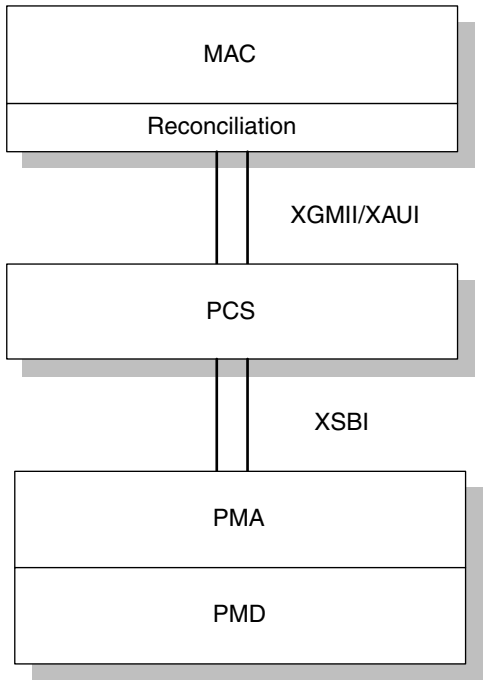


Figure 7-5
10 Gigabit Ethernet LAN physical layer interface overview.

Two types of PCSs have been chosen for the 10 Gigabit LAN physical interface, based on two types of supported PMDs: serial optical link- and CWDM- (also called *WWDM* for wide wave division multiplexing) based encoding schemes. For the serial LAN physical layer interface, a new coding scheme known as *64B/66B* (for 64 bits/66 bits) is used. The 64B/66B coding scheme adds 2 transmission overhead bits to every 64 bits of payload data, and supports high-bandwidth optical transmission. Analysis shows it is at least as robust as the well-tested 8B/10B coding scheme. In addition, 64B/66B has less transmission overhead (2/64 = 3%). This same coding scheme is also used for the 10 Gigabit Ethernet WAN physical layer interface.

Optionally, the 8B/10B coding scheme can also be used to support the appropriate PMD sublayer. The advantages of the 8B/10B-based PCS sublayer is that it is a simple and proven technology used in the widely deployed Gigabit Ethernet switches.

For both the LAN and WAN PCS sublayers, there are two optional interfaces to the MAC layer, XGMII and XAUI. Both interfaces connect an optical PMD module to a MAC module. XAUI is evolved from Gigabit Ethernet's attachment unit interface (GAUI) to the 10 Gigabit rate (the Roman numerical X represents "10 Gigabit"). XAUI is a low-pin-counter,

self-policed serial bus that is designed as an interface extender, extending the PMD interface XGMII up to 20 in and allowing transmission across connectors.

The 10G sixteen-bit interface (XSBI) is an extended version of the OC192 interface developed at the Optical Interworking Forum (OIF). The objective is to leverage an optical interface that is widely deployed in the OC-192 market. This is an interface common to both the LAN and WAN physical interfaces.

7.3.6 10 Gigabit Ethernet WAN Interface

The WAN physical layer consists of the following components: the WAN interface sublayer (WIS), the physical link coding sublayer, and the interfaces to the PMDs and to the MAC layer, as shown in Fig. 7-6.

7.3.6.1 WAN Interface Sublayer The WAN interface differs from the LAN interface mainly by WIS, which is a "thin" version of the OC-192 interface. The 10 Gigabit Ethernet WAN physical layer is defined to be compatible with SONET and yet is not fully compliant to all of the SONET standards. The main objective of WIS is to have a cost-effective link that uses common Ethernet PMDs to provide access to the large installed

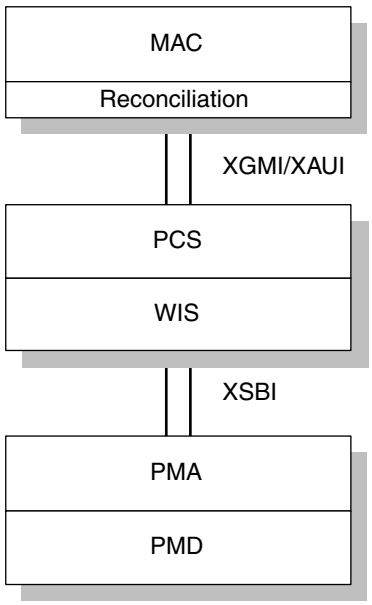


Figure 7-6
10 Gigabit Ethernet WAN physical interface overview.

base of SONET infrastructure and enables IP-based Ethernet switches to be attached to the SONET/SDH and TDM multiplexed networks.

10 Gigabit Ethernet WIS is like SONET in the following aspects:

- It uses SONET infrastructures for layer-1 transport such as SONET ADM, TDM transponders, and optical regenerators.
- It adopts SONET OC-192 line rate of 9.953 Gbps, SONET framing, and some minimal SONET path, section, and line overhead processing.

However, the 10 Gigabit Ethernet WIS is unlike SONET in the following aspects:

- It uses optical transceivers (PMDs) specified by IEE 802.3ae Task Force for 10 Gigabit Ethernet.
- It does not support any SONET DTM hierarchy (i.e., OC-1 at 51.840 Mbps up to OC-192 at 9,953.281 Mbps).
- It avoids SONET grid laser specifications and jitter requirements, mainly to allow for inexpensive implementation.
- It does not implement the SONET stratum clocking scheme. 10 Gigabit Ethernet remains an asynchronous link protocol. Like its predecessors, its timing and synchronization are maintained within each character of transmitted bit stream of data. The receiving hub, switch, and routers can resynchronize the data. In contrast, synchronous SONET/SDH, as described in Chap. 5, requires that all devices share the same clock to avoid transmission errors caused by timing drift between the transmission and reception switches.
- It uses concatenated OC-192c.

7.3.6.2 PCS of 10 Gigabit WAN Interface The main function of the physical link coding sublayer of 10 Gigabit Ethernet is to encode and decode data received from/sent to the PMD sublayer, using the coding scheme appropriate to the medium. The PCS of the WAN interface, like that of the LAN physical interface, accommodates two types of PMDs: serial optical link and WWDM-based encoding schemes.

The WAN physical coding sublayer adopts SONET's 64B/66B framing. Another major difference between the LAN interface and the WAN interface is that the WAN interface does not support an optional 8B/10B coding scheme, because the interface is intended to connect to installed SONET/TDM networks instead of an Ethernet LAN.

7.3.6.3 10 Gigabit Media-Independent Interface XGMII is the physical layer interface to the MAC layer. Like Gigabit Ethernet, the transceiver function is built into all 10 Gigabit Ethernet devices and the XGMII becomes an internal component of the device. It is not desirable to expose the XGMII as an external interface because of the high frequency of data transfer. The XAUI is an optional extension to the XGMII interface to allow the MAC layer to connect directly to the PCSs.

For the WAN PCS sublayer, the interface to PMA is XSBI, which is an extended version of the OC192 interface developed by OIF.

7.3.7 MAC Sublayer

A reconciliation sublayer is attached to the MAC layer, as shown in Fig. 7-6, to adjust the line rate between the LAN and WAN interface line rates. The 10 Gigabit Ethernet MAC layer supports a 10 Gigabit rate while the WAN interface operates at 9.953 Gbps of OC-192. This reconciliation sublayer performs rate matching by inserting extra spaces (i.e., idle characters) between Ethernet frames to bring the effective data rate down to match that of the WAN interface, while still operating at the 10 Gbps clock rate. At the WAN interface, the idle characters are deleted to allow the Ethernet frame stream to be packed into a SONET payload.

The MAC sublayer, the highest layer defined in the 10 Gigabit Ethernet standards, conforms to the existing standards in order to maintain compatibility across all speeds of the previous Ethernet. The scope of the IEEE802.3ae standard at the MAC layer is to define MAC parameters and, if necessary, a minimal augmentation of MAC operation for the full-duplex transfer of LLC and Ethernet frames at 10 Gbps. It remains to be seen whether and how this rate reconciliation function at the MAC sublayer will be standardized.

7.4 Applications of Optical Ethernet

Optical Ethernet is finding its way into LAN, backbone, and metro area networks. This section introduces some of the potential application scenarios. The comparison between Gigabit Ethernet and 10 Gigabit Ethernet shown in Table 7-4 will help summarize the main features of optical Ethernet.

TABLE 7-4

A Summary Comparison Between Gigabit Ethernet and 10 Gigabit Ethernet

	Gigabit Ethernet	10 Gigabit Ethernet
Physical medium	Copper and fiber	Fiber only
PMD	Leverage Fiber Channel PMDs	New optical PMDs
PCS	64B/66B coding; optional 8B/10B encoding system	New 64B/66B coding system
MAC operation mode	Half duplex and full duplex	Full duplex only
MAC extension	Carrier extension and frame bursting	MAC speed throttle
Target max distances	Up to 5 km	Up to 40 km
Target application	LAN and edge of MAN	LAN, MAN, and WAN

7.4.1 Gigabit Ethernet Applications

One significant implications of the Gigabit Ethernet is the deployment of long-distance Gigabit Ethernet using long-wavelength optics on dark fiber that can reach metro distances. For the first time since the inception of Ethernet technology, the Ethernet is capable of going beyond the boundary of local area networks. However, the main thrust of Gigabit Ethernet applications is still in LAN and edge-of-metro networks (Cisco 2000; Chou, Luk, and Ng 2000).

Gigabit Ethernet can be used to form the backbone of a large LAN. It is already widely deployed in the space of local area networks to meet the continuing demands of rapid growth in bandwidth. For example, in a campus network, a Gigabit Ethernet switch can serve as the central switch, interconnecting the installed Fast Ethernet switches and other Gigabit Ethernet switches of scattered departments on the campus, using short-distance multimode fiber.

7.4.2 10 Gigabit Ethernet Applications

10 Gigabit Ethernet can be used to support a number of applications, which may include enterprise LAN interconnections, backend application server connections, inter- and intra-point-of-presence (POP) connections, and fast data transport in metro area networks. To illustrate the LAN and WAN interfaces of 10 Gigabit Ethernet, the LAN- and WAN-oriented application examples are described.

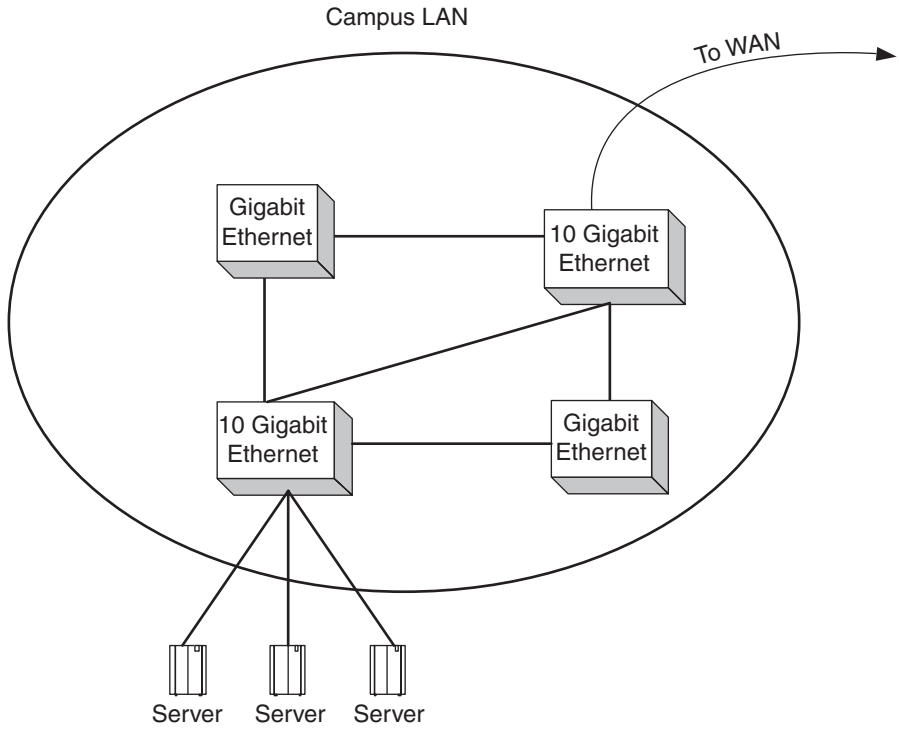


Figure 7-7
10 Gigabit Ethernet LAN application example.

7.4.2.1 10 Gigabit Ethernet in LAN 10 Gigabit Ethernet can be used in LAN backbone to meet the continuing rapid growth in bandwidth. For example, in a campus network, a 10 Gigabit Ethernet switch can serve as the central switch, interconnecting Gigabit Ethernet switches of scattered departments on the campus, using short-distance multimode fiber, as shown in Fig. 7-7.

7.4.2.2 Metro and Wide Area Network Applications 10 Gigabit Ethernet represents a major development for MANs, WANs, and backbone networks. It is not because it is new but because it is old. Ethernet, a technology as old as the Internet itself, for the first time extends beyond LAN boundaries and presents the potential to reach MAN and WAN and unify end-to-end networking under one technology.

10 Gigabit Ethernet can be used to interconnect the core 10 Gigabit Ethernet switches of different LANs. For example, a campus A LAN and campus B LAN can be interconnected via a 10 Gigabit Ethernet backbone, as shown in Fig. 7-8. Between two LANs, in a metro area, enterprise

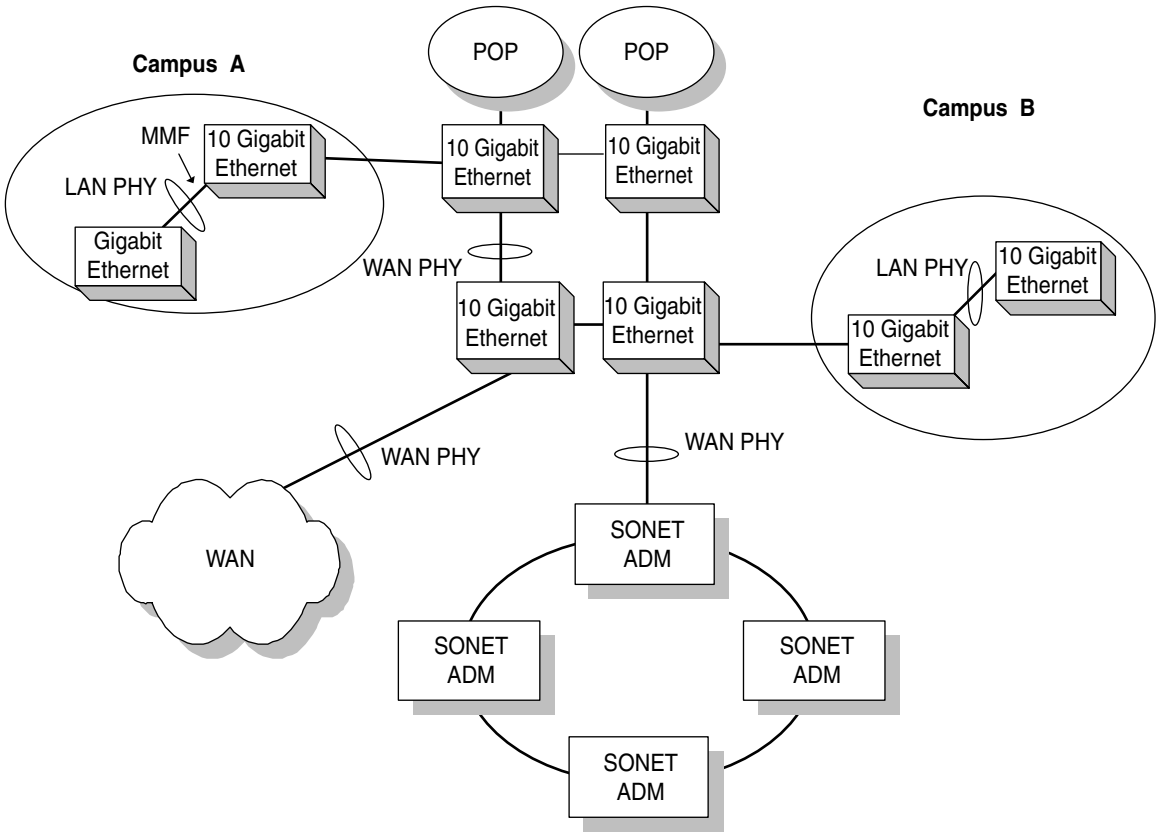


Figure 7-8 The 10 Gigabit Ethernet MAN application examples.

customers may use 10 Gigabit Ethernet over dark fiber to support requirements such as serverless buildings, remote hosting, off-site storage or backup, and disaster recovery.

10 Gigabit Ethernet backbone can also be used as a complementary or cost-effective alternative to the conventional SONET ring, as shown in Fig. 7-8. The 10 Gigabit Ethernet-based backbone can be configured as a mesh or ring. In this case, a 10 Gigabit Ethernet MAN backbone may be used to interconnect a service provider's POP, as shown at the top of Fig. 7-8. A POP is normally a point that collects the traffic from residential or business customers and aggregates them onto a bigger pipe. Gigabit Ethernet switch equipment can be placed at a POP, and then multiple Gigabit Ethernet traffic streams aggregated onto a 10 Gigabit Ethernet pipe.

Applications of 10 Gigabit Ethernet in regard to MANs include network-attached storage (NAS), storage area networks (SANs) and service provider data centers.

REVIEW QUESTIONS

1. Discuss some of the motivations behind the development of optical Ethernet, which is intended for applications in LAN as well as metro area networks and backbone networks.
2. Discuss the reasons optical Ethernet, i.e., Gigabit Ethernet and 10 Gigabit Ethernet, are included in Part II of this book on broadband transport networks rather than some other part.
3. Describe the types of transmission media supported by Gigabit Ethernet.
4. Discuss what issues the Gigabit Ethernet MAC sublayer features such as carrier extension and frame bursting are designed to address.
5. Gigabit Ethernet supports both half-duplex and full-duplex MAC operation modes. Describe the applications for which the half-duplex mode is better suited and the applications for which the full-duplex mode is better suited.
6. Compare the Gigabit Ethernet PMD and 10 Gigabit Ethernet PMD and discuss the differences and similarities between the physical medium supported by each PMD.
7. Discuss why 10 Gigabit Ethernet eliminates the half-duplex operation of the original Ethernet.
8. Discuss the motivations behind the development of the two parallel LAN and WAN physical interfaces defined for 10 Gigabit Ethernet. Compare the WAN and LAN physical interfaces and list the main differences between them.
9. Describe the transmission medium supported by 10 Gigabit Ethernet.
10. Discuss the main function of the reconciliation sublayer between the physical layer interface and the MAC layer in 10 Gigabit WANs.
11. Describe the scope of the 10 Gigabit Ethernet standards as defined by IEEE 802.3ae.

REFERENCES

10 Gigabit Ethernet Alliance. 2001. "10 Gigabit Ethernet Technology Overview," White paper. Web site: www.10gea.org.

ANSI. 1999. "Fibre Channel Physical and Signaling Interface (FC-PH)." ANSI X3.230. Web site: www.ansi.org.

Bruce, T. 2001. "An Introduction to 10 Gigabit Ethernet," White paper. Web site: www.cisco.com.

Chou, T., Luk, H., and Ng, T. 2000. "Gigabit Ethernet." White paper. Web site: www.comm.toronto.edu.

Cisco Systems. 2000. "Introduction to Gigabit Ethernet." White paper. Web site: www.cisco.com.

Cisco Systems. 2001. "10 Gigabit Ethernet Application Overview." White paper. Web site: www.cisco.com.

Gigabit Ethernet Alliance. 1996. "Gigabit Ethernet." White paper. Web site: www.ieee.org/groups/802/3/ae.

IEEE. 2002. "Media Access Control Parameters, Physical Layers, and Management Parameters for 10 Gb/s Operations." IEEE 802.3ae. Web site: www.ieee.org.

IEEE. 2001. "Carrier Sense Multiple Access with Collision Detection Access Method and Physical Layer Specifications." IEEE 802.3. Web site: www.ieee.org.

Robinson, S. 2001. "Manning up for 10 Gigabit Ethernet," *Communication & System Design Magazine,* Vol. 3, No. 4.

Saunders, S. 1998. *Gigabit Ethernet Handbook.* New York: McGraw-Hill.

Seifert, R. 1999. *Gigabit Ethernet: Technology and Applications for High-Speed LANs.* Reading, MA: Addison-Wesley.

Seifert, R. 2000. *The Switch Book.* New York: John Wiley & Sons.

Spurgeon, C. 2000. *Ethernet: The Definitive Guide.* Sebastopol, CA: O'Reilly and Associates.

PART 3

Broadband Access Networks

Part III, the core of this book, is a comprehensive review of current packet broadband access network technologies. A broadband access network provides a broadband connection between a core network and residential or enterprise end users. Traditionally the access network connection has been a 64/56-Kbps telephone line for residential and T1 (1.55 Mbps) for enterprise customers. Access network technologies remained stagnant for decades until the Internet-driven data services created a huge market for packet broadband access networks.

For convenience of description, broadband access network technologies are divided into two categories: "last mile" solutions and "last yard" solutions. "Last mile" solutions provide broadband connections between a backbone network and customers' premises. For example, a fiber network provides broadband access from a central office to the curbs of residential homes or to telecom boxes outside office buildings. "Last yard" solutions start where "last mile" solutions end: they provide broadband connection to user devices. For example, they connect the fiber terminating at a curb or building floor to a PC at a user's desk.

Another dimension of broadband access networks is the distinction between wireline and wireless technologies, as shown in Fig. P3-1.

Some wireline broadband access technologies are transmission-medium-specific while others may operate on more than one transmission medium, as indicated below:

- XDSL: operates on copper wire

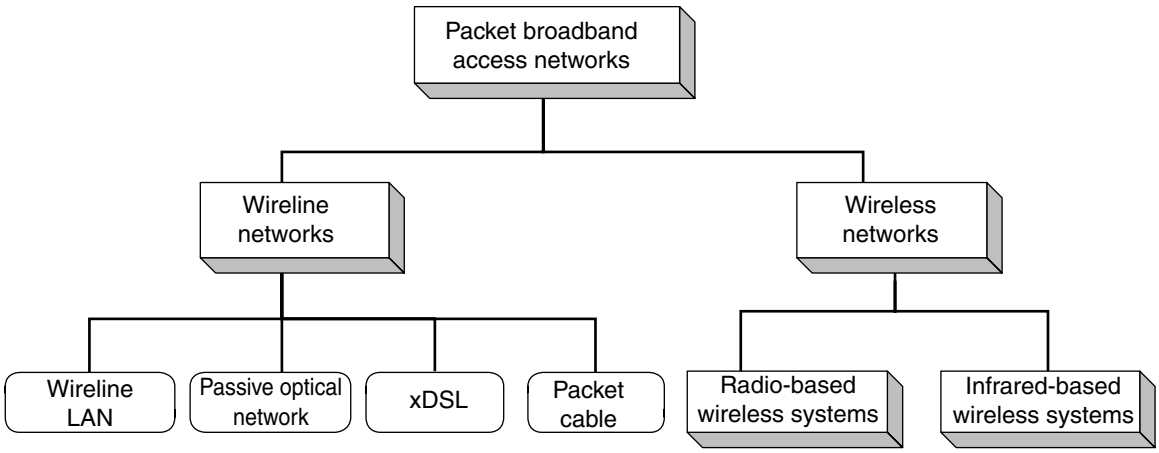


Figure P3-1 Taxonomy of packet broadband access technologies.

- Packet cable: operates on a mix of coax cable and optical fiber
- PON: operates on optical fiber
- LAN: operates on coax cable, copper wire, fiber, or any mix of them

Wireless broadband access networks are further divided into two categories based on the two types of wireless transmission mediums: radio and infrared, as shown in Fig. P3-2. Radio-based wireless technologies that provide "last mile" solutions include

- Wireless LAN as specified in IEEE 802.11
- European wireless LAN standard HiperLAN
- Multipoint, multichannel distribution service (MMDS)
- Local multipoint distribution service (LMDS)
- Broadband wireless access (BWA)
- Broadband satellite network

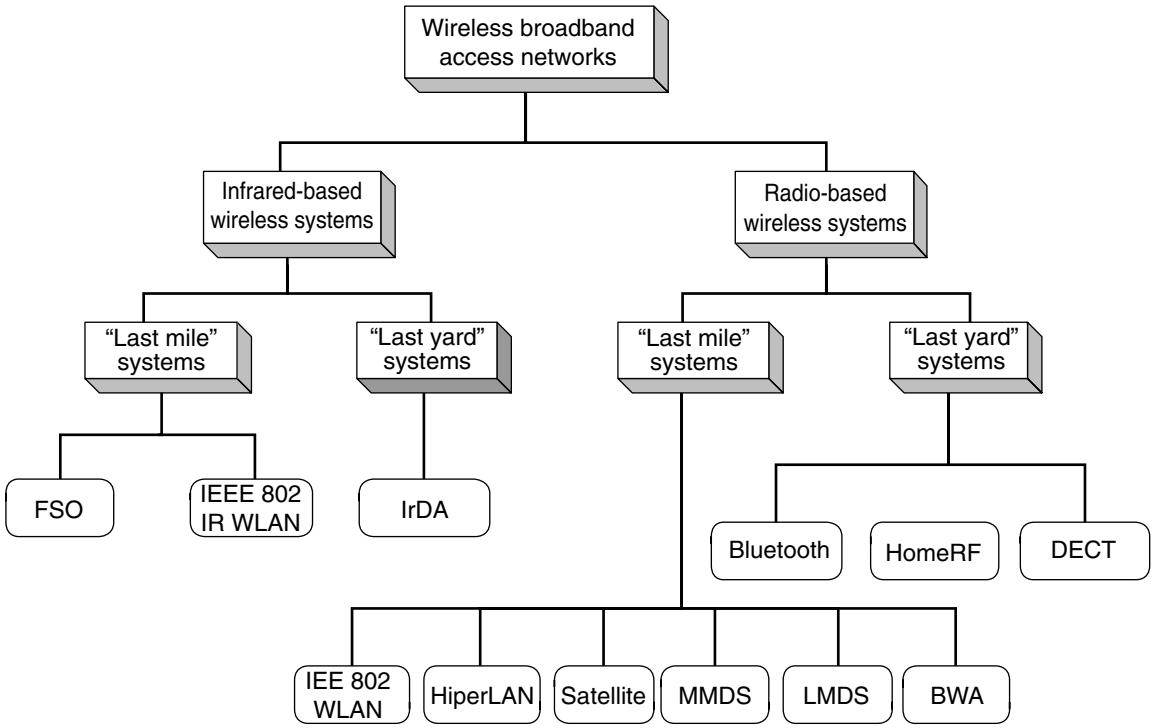


Figure P3-2 Broadband wireless access technologies.

The radio-based wireless access technologies that provide "last yard" solutions include

- *Bluetooth.* Also known as *personal area network* (WPAN), this is a short-range wireless network technology that covers very short distances and connects personal computing devices such as personal digital assistants (PDAs), computers, printers, and others.
- *HomeRF.* HomeRF targets next-generation, innovative applications that take advantage of the increasingly intelligent devices in the home environment.
- *DECT.* Digital Enhanced Cordless Telecommunications (DECT) is a European digital radio access network standard for wireless communications for use in environments like homes or small offices.

Infrared wireless technologies use optical beams, also known as *optical wireless* or *free space optics* (FSO), as the transmission medium. Infrared wireless technologies that provide the "last mile" solutions include

- *IEEE 802.11 infrared-based wireless LAN.* This is the counterpart of radio-based wireless LAN as defined by IEEE 802.11.
- *FSO-based wireless access networks.* These use proprietary technologies to achieve very high bandwidth and provide broadband access to backbone metro area networks for enterprise customers.

The infrared wireless technologies that provide the "last yard" broadband access include IrDA, a standard for connecting appliance devices to computing devices such as PCs within very short distances.

CHAPTER 8

Local Area Networks

8.1 Introduction

A local area network is a high-speed data network that covers a relatively small geographic area. It typically connects workstations, personal computers, printers, servers, and other end-user devices, which are collectively also known as *data terminal equipment*. The common applications of LAN include shared access to devices and applications, file exchange between connected users, and communication between users via electronic mail and others. LANs are also private data networks, because they belong to an organization and are used to carry data traffic as opposed to voice traffic.

This section provides a brief introduction to LAN history, standards, protocol stacks, topologies, and devices.

8.1.1 LAN History and Standards

LAN is a type of broadband packet access network that carries the packet data traffic of an organization. LAN interconnects the end users of an organization to an outside public data network such as the Internet.

The basis of LAN technologies and standards was defined in the late 1970s and early 1980s. LAN technologies really emerged with the Internet itself, and the first widely deployed LAN technology, Ethernet, is almost as old as the Internet itself. The overwhelming majority of the deployed LANs are Ethernet.

IEEE 802, a branch of the International Institute of Electrical and Electronics Engineers (IEEE), is responsible for most of the LAN standards. These standards have also been adopted by other standards organization such as ANSI and ISO. The major LAN standards are listed in Table 8-1.

8.1.2 LAN Protocol Stacks

The LAN protocols operate at the bottom two layers of the OSI network reference model, i.e., at the physical layer and the data link layer, as shown in Fig. 8-1. The physical layer is primarily concerned with the transmission medium and its physical characteristics for digital signal transmission. The data link layer consists of two sublayers, the medium access control (MAC) sublayer and logical link control (LLC) layer. The MAC sublayer is responsible for controlling access to a shared medium by multiple users simultaneously. The LLC sublayer is responsible for

TABLE 8-1
IEEE 802 LAN Standards Summary

IEEE 802 specification	LAN technology	Description
IEEE 802.1 (ISO 15802-2)	General information	Details how the other 802 standards relate to one another and to the ISO OSI reference model.
IEEE 802.2 (ISO 8802.2)	LLC framework	Divides the OSI data link layer into two sublayers and defines the functions of the LLC and MAC sublayers
IEEE 802.3	Ethernet	Defines the CSMA/CD protocol, which is used in Ethernet applications and has become synonymous with Ethernet
IEEE 802.4 (ISO 8802.4)	Token bus	Defines the token-passing bus access method
IEEE 802.5 (ISO 8802.5)	Token ring	Defines the Token Ring access method
IEEE 802.7	Broadband LAN	Recommended practices for broadband LANs
IEEE 802.11	Wireless LAN	Wireless LAN medium access control (MAC) and physical layer specifications
IEEE 802.15	Wireless personal area network (WPAN)	WPAN MAC and physical layer specifications
IEEE 802.16	Broadband fixed wireless metropolitan area networks (MANs)	Air interface specification for fixed broadband wireless access systems
IEEE 802.12	100 VG-AnyLAN	Defines a LAN technology that supports the operations of any existing LAN protocol, including the Ethernet frame format and Token Ring frame format, but not both at the same time

interfacing to the upper layers, such as IP and the Internetwork Packet Exchange protocol (IPX). Any layer above the data link layer is beyond the scope of the LAN protocols.

The IEEE 802 LAN standards are compatible at the upper part of the data link layer, i.e., at the LLC sublayer, but differ from each other at the MAC sublayer and physical layer.

The scope of each LAN protocol may vary. Some cover the entire two bottom layers. For example, Ethernet as defined in IEEE 802.3 (IEEE 2001) and FDDI as defined in IEEE 802.5j (IEEE 1998c) cover the physical layer and both sublayers of the data link layer, as shown on the right-hand side of Fig. 8-1. Other LAN protocols, such as Token Ring and token bus,

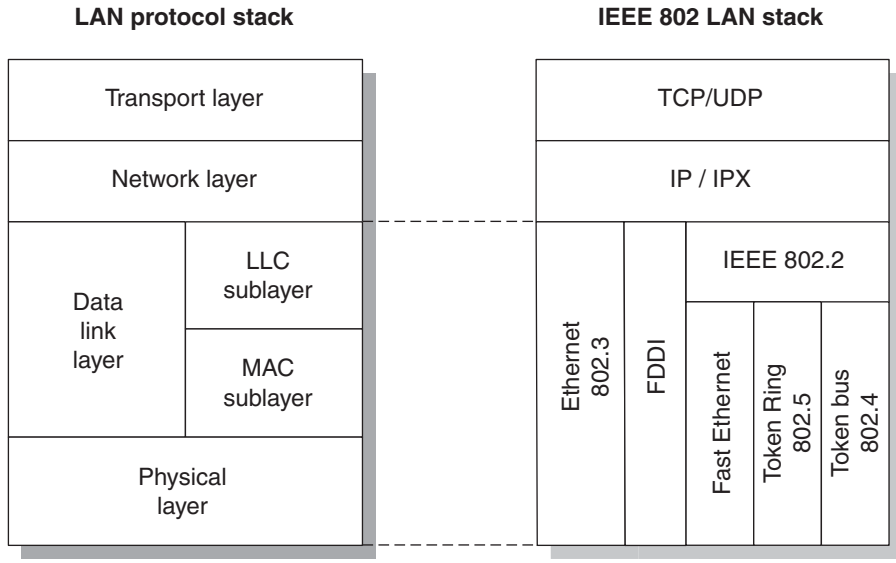


Figure 8-1
The LAN protocol stack.

specify the physical layer and the MAC sublayer while sharing a common LLC specification defined in IEEE 802.2, as shown on the right-hand side of Fig. 8-1.

8.1.2.1 Physical Transmission Medium The LAN transmission medium can be divided into the two general categories of wired and wireless. This chapter focuses only on the wired or wireline LAN technology, while Chap. 9 will describe wireless LAN.

There are basically three types of transmission media used in wireline LAN deployment: copper twisted pair, coaxial cable, and optical fiber. The type of transmission medium determines the data rate and transmission distance.

TWISTED PAIR COPPER WIRE Twisted pair, both shielded and unshielded, is a pair of copper wires that are twisted to increase the transmission distance. It is the least costly of the three wireline LAN media, and one of the most common transmission media currently used in LAN applications. It is primarily used in star and hub LAN configurations in office buildings. The maximum transmission distance of twisted pair cable depends on the target date rate; typically the limit is 100 m without repeater. The data rate of copper twisted pair normally is not as high as that of other transmission media and depends on factors such as transmission distance and the modulation scheme used for transmission. The

longer the transmission distance is, the lower the bit rate is. It is not uncommon to see twisted pair achieve a bit rate of over 1 Mbps for a distance of 100 meters.

COAXIAL CABLE Coaxial cable, whose transmission wire is insulated with dielectric insulating material and braided out conductor, can achieve higher data rates and longer transmission distances. There are two kinds of coaxial cables: thin wire and thick wire, referring to the difference in the cable diameters, thin wire being 0.25 in diameter and thick wire being 0.5 in diameter. Thin-wire coaxial cable reaches shorter distances, typically 200 m with the data rate of over 10 Mbps, while thick-wire cable can reach over 500 m with the same data rate.

OPTICAL FIBER Optical fiber carries data in the form of flashing light beams in a glass fiber, as opposed to electrical signals on a wire. Optical fiber can achieve much higher data rates than coaxial cable or twisted pair over much longer distances. The fiber transmission equipment consists of fiber cable, special electrical-to-optical and optical-to-electrical converters, light emitters such as light-emitting diodes (LEDs) or laser and optical receivers. These transmission components have been much more costly than twisted pair and coaxial cable. However, with the advent of new optical transmission technologies and a massive market for broadband applications, the cost has come down considerably in recent years and the optical fiber is becoming a common choice for LAN deployment.

LANs can use one type of transmission medium or a mix of types. For example, lower-speed twisted pair can be used between a computer and a hub, while coaxial cable can be used between a branch hub and a main hub and high-speed optical fiber cable can be used between a main hub and an outside router.

8.1.2.2 Media Access Control Sublayer A LAN technology must address the issue of resource contention because multiple users share the same transmission medium. A contention occurs when two DTEs transmit data at the same time. There are basically two MAC mechanisms for LAN: carrier-sense multiple access with collision detection and control token.

CSMA/CD The CSMA/CD access control method is used in Ethernet and can be characterized as "listen and send." A network device first listens to the wire when it has data to send, then sends the data when it finds that no other device is sending the data. After it finishes sending the

data, it listens to the wire again to detect if any collision occurs while it transmitted data. A collision occurs when two devices send data simultaneously. If a collision is detected, the device waits for a random amount of time before resending the data. The randomness of the wait period makes the possibility of another collision very small. However, this algorithm is not deterministic, and when the number of users increases to a large enough point, network performance deteriorates drastically owing to the large number of collisions.

The major advantage of CSMA/CD is its simplicity. It is easy to implement and works well in the LAN environment.

CONTROL TOKEN Control token is a special network packet used to control access to a shared transmission medium. A token is passed around a network from device to device. When a device has data to send, it must wait until it has the token, at which time it sends its data. When the transmission is complete, the token is released so that other devices may use the network to transmit their data. A major advantage of token-passing networks is that they are deterministic. In other words, it is easy to calculate the maximum time that will pass before a device has the opportunity to send data. This explains the popularity of token-passing networks in some real-time environments such as factories, where machinery must be capable of communicating at determinable intervals. Token-passing networks include Token Ring and FDDI.

MAC ADDRESS The MAC address is a number that is hard-wired into each LAN card such as the Ethernet Network Interface Card or adapter that uniquely identifies this device on a LAN. The MAC addresses are 6 bytes in length, and are usually written in hexadecimal such as 12:34:56:78:90:AB. The colons in the address may be omitted, but generally make the address more readable. Each manufacturer of LAN devices has a certain range of MAC addresses, just like a range of telephone numbers, that they can use. The first 3 bytes of the address denote the manufacturer.

8.1.2.3 Link Layer Control Sublayer The LLC sublayer, as defined in the IEEE 802.2 standard, mainly hides the differences between various MAC sublayer implementations such as Ethernet, Token Ring, and FDDI and presents a uniform interface to the network layer. This allows different types of LANs to communicate with each other.

The IEEE 802 LLC protocol defines a generic LLC protocol data unit that includes both user data and LLC header. The LLC header contains a control field that in turn contains the fields such as protocol ID and

header type. Also found in the LLC header are source and destination address fields.

8.1.3 Data Transmission Methods

There are three data transmission modes in LAN environments: point-to-point, multicast, and broadcast. In each transmission mode, a single packet is sent to one or more nodes.

In *point-to-point transmission,* which is also known as *unicast,* a single packet is sent from a source to a destination on a LAN. First, the source node addresses the packet by using the address of the destination node. The packet is then sent onto the LAN, and the LAN then passes the packet to its destination.

In *multicast transmission,* a single data packet is copied and sent to a specific subset of nodes on a LAN. First, the source node addresses the packet by using a special type of address, called a *multicast address.* The packet is then sent onto the LAN, which makes copies of the packet and sends a copy to each node that is part of the multicast address.

In *broadcast transmission,* a single piece of data is copied and sent to all the nodes on a LAN. In this type of transmission mode, a source node addresses a packet by using a broadcast address. The packet is then sent onto the LAN, which makes copies of the packet and sends a copy to every node on the LAN.

8.1.4 LAN Topology

A LAN topology defines how the data terminal equipment such as desk top computers, printers, and server computers, and LAN internetworking devices such as switches, routers, and hubs, are interconnected to each other. In general, there are four types of LAN topologies: bus, star, ring, and hub. Each has some advantages, which will now be discussed (Halsall 1996).

8.1.4.1 Bus Topology The bus is one of the most common LAN topologies. A simple bus topology is characterized by a central cable that runs through end-user equipment like computers and servers, as shown in Fig. 8-2(a). A physical connection, also known as a *tap,* is made to the cable for each user terminal or computer to access the network. MAC circuitry and the software implementing the control scheme together allow the connected users to share the common cable and transmission

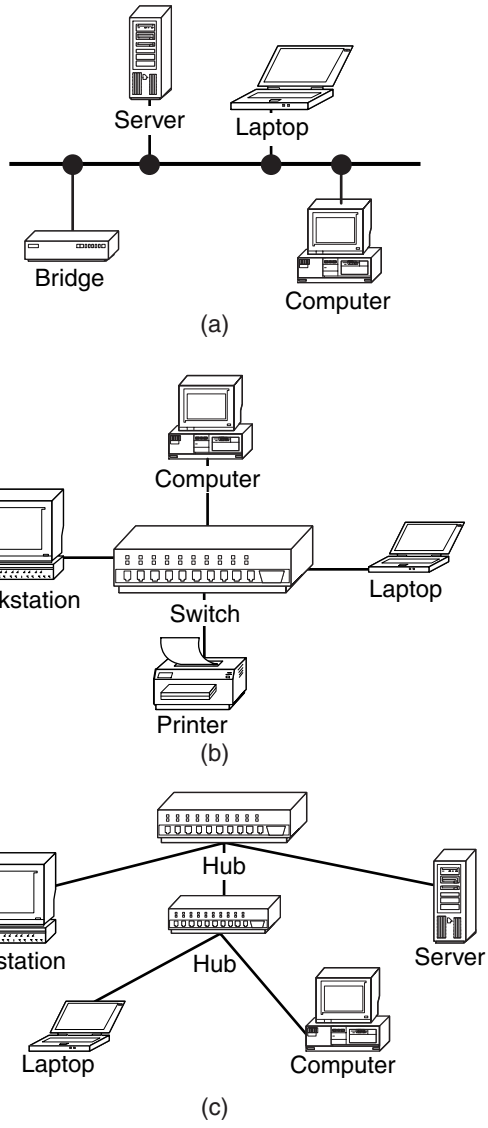


Figure 8-2
Bus, star, and hub LAN topology examples.

bandwidth. A slightly more complicated bus topology may consist of multiple layers of buses. A bus cable can be connected to another bus cable, which in turn may be connected to yet another cable. This forms a topology that looks like an uprooted tree.

8.1.4.2 Ring Topology Ring-based LAN topology is characterized by a cable that passes from one DTE to another until all the DTEs are connected to form a ring or loop. A distinct feature of ring topology is

that traffic travels in one direction only. Between two neighboring user DTEs, it is a direct point-to-point link that carries traffic in one direction only, termed *unidirectional*. Again, medium access circuitry and a control algorithm are built into a DTE and network to allow each DTE a fair chance to access the cable ring.

8.1.4.3 Star Topology In star topology, there is a focal point that is either a switch or a router, and all end-user DTEs are connected to the central point via a point-to-point cable, as shown in Fig. 8-2(b). This is a typical voice-service PBX configuration that is also used to interconnect end-user DTEs, although not as common as other topologies. Compared to the other topologies, star topology has more complicated wiring.

8.1.4.4 Hub Topology A fourth common topology is the hub structure, which is a mix of the ring and bus topologies. A hub topology is simply a bus or ring wiring collapsed into a central unit. A hub does not perform any switching or intelligent processing. All a hub does is simply retransmit all signals received from a DTE to all other DTEs with a set of repeaters inside the hub. As shown in Figure 8-2(c), a hub can be connected to another hub to form a hierarchy of hubs and DTEs that looks like a tree structure.

8.1.5 LAN Internetworking Devices

An internetworking device interconnects two or more other LAN devices. Based on functionality, there are three types of such internetworking devices: repeater, bridge, and router or switch.

8.1.5.1 Repeater A LAN *repeater* is a physical layer device used to connect two LAN cable segments so a LAN will extend further in distance. A repeater essentially boots digital signals to allow a series of cable segments to be treated as a single cable. It receives signals from one network segment and amplifies, retimes, and retransmits those signals to another network segment. A LAN repeater operates at the physical layer without any intelligence to perform any filtering or other traffic processing. In addition, all electrical signals, including electrical noise and errors, are repeated and amplified as well. The total number of repeaters within a LAN is limited due to timing and other issues.

8.1.5.2 LAN Hub A *hub* is a physical layer device that connects multiple user stations, each through a dedicated cable. In some respects, a

hub functions as a multiport repeater. Hubs are used to create a physical star network while maintaining the logical bus or ring configuration of a LAN.

8.1.5.3 LAN Bridge A LAN *bridge* is an internetworking device that interconnects two LAN segments at the data link layer as opposed to the physical layer in the case of a repeater. A bridge must have at least two ports, one receiving incoming frames and one sending outgoing frames. A bridge uses a MAC address to route frames from one segment to another, or even to a different LAN that is the same or different at the physical or MAC layer.

8.1.5.4 LAN Router and Switch A LAN *router* operates at the network layer, interconnecting like and unlike devices attached to one or more LANs. LAN routers normally also support link layer bridging in addition to network layer routing.

A LAN router, as described in Chap. 4 on IP networks and Appendix A, employs routing protocols to dynamically obtain knowledge of destination address prefixes across an entire set of internetworked LANs. A LAN router normally has a packet-forwarding engine that uses a lookup table to identify the physical interface of the next hop toward the destination.

8.2 Ethernet

Ethernet is almost as old as the Internet itself. Since its inception at a Xerox lab in the early 1970s, it has been the dominant protocol for local area networks. By various estimates, Ethernet accounts for somewhere between 80 to 95 percent of worldwide LAN installations.

This section, after first providing some background information, introduces three generations of Ethernet: 10Base Ethernet, Fast Ethernet, and optical Ethernet, with an emphasis on the first two. Gigabit Ethernet and 10 Gigabit Ethernet were described in detail in Chap. 7 in the context of optical transport network, and will be discussed briefly in this chapter in the context of LAN technology.

What is remarkable about Ethernet is its continuity and simplicity. The fundamentals of Ethernet such as Ethernet frame and logical link control, which were already defined for the first generation of Ethernet, have remained largely intact through the rapid technological evolution

of the past two decades. Ethernet is viewed as a kind of plug-play technology because it is relatively simple and can operate with very little manual intervention for configuration and provisioning.

8.2.1 Ethernet Basics

8.2.1.1 A Brief History Ethernet was originally developed by Digital, Intel, and Xerox (DIX) in 1972 and was designed as a "broadcast" system where stations on a network can send messages at will. All the stations may receive the messages, but only one specific station to which the message is directed will respond. Robert Metcalf and David Boggs of Xerox are credited with coming up with first Ethernet design. Ethernet was originally designed to run on any medium (copper wire, fiber, or even radio wave), which is where *Ether* in the term *Ethernet* comes from.

The original version of Ethernet was adopted by IEEE Committee 802.3 (IEEE Project 802 was named after the time Ethernet was set up, in February 1980), and the packet format was standardized, which is known as the IEEE 802.3 Ethernet frame.

The Ethernet evolution, based on the transmission technologies and speed, involved at various times in its development, can be divided into the following periods:

- 10BaseT Ethernet, starting from 1972 to the mid-1990s
- 100Base Ethernet, starting from the mid-1990s
- 1000Base Ethernet, starting from 1998
- 10Gig Ethernet, starting from 2000

An Ethernet version is represented in terms of the transmission speed, the transmission medium, and maybe the transmission distance. The prefix number in an Ethernet version such as 10 in 10Base or 100 in 100Base refers to the transmission speed of 10 Mbps and 100 Mbps. The suffix letter refers to the medium type, while suffix number for earlier versions of Ethernet refers to the maximum transmission distance. For example, the letter T in 10BaseT refers to "twisted pair" copper wire and the number 5 in 10Base5 refers to the transmission distance in hundreds of meters.

8.2.1.2 Ethernet Protocol Stack The Ethernet protocol stack is similar to the general LAN protocol stack as described earlier: It covers the layers 1 and 2 of the OSI network reference model. In addition, Ethernet further

defines three sublayers for the physical (PHY) layer: PMD, PMA, and PCS, which are briefly discussed here (IEEE 2001c).

The physical medium-dependent (PMD) sublayer defines the Ethernet cables, wiring, and other transmission medium-related components. The physical medium attachment (PMA) defines the type of connectors used to connect an Ethernet device such as an Ethernet NIC, hub, or switch to the Ethernet cable. The physical coding sublayer (PCS) defines a scheme appropriate to the medium to encode and decode data received from/sent to the PMD sublayer (Spurgen 2000).

The Ethernet data link layer, like that of other LAN technologies, is broken into two sublayers: the LLC on the upper half and the MAC on the lower half. The MAC deals with getting data on and off the wire and media access control, as shown in Fig. 8-3. The logical link control

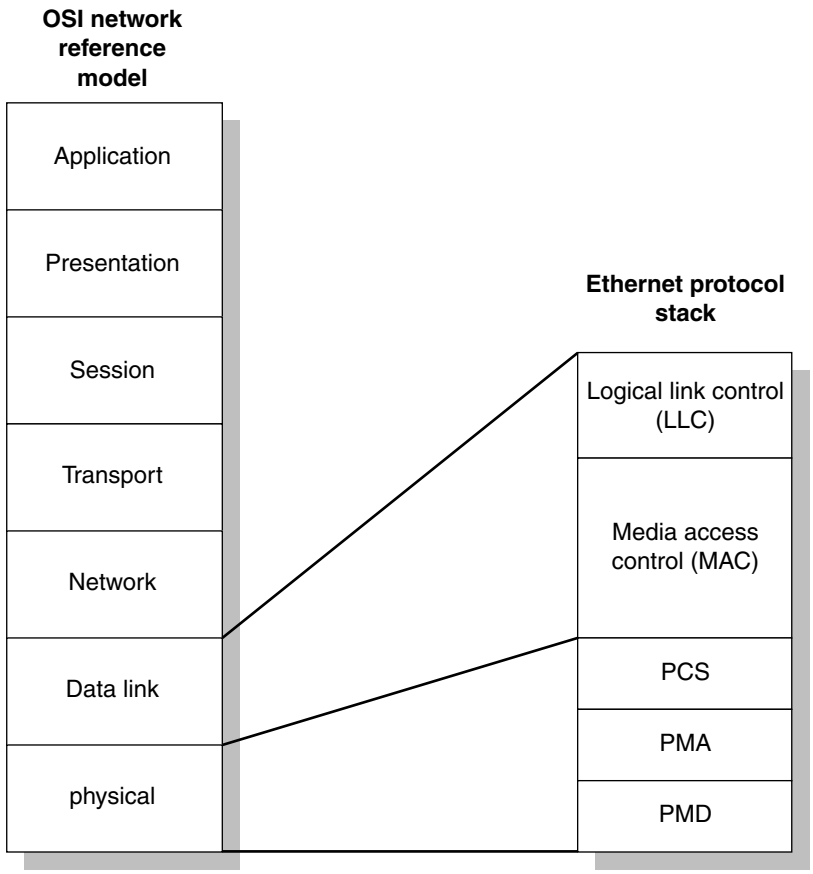


Figure 8-3
The Ethernet protocol stack in reference to the OSI network reference model.

on the upper half of the data link layer deals with error checking and providing a uniform interface to the network layer above.

8.2.1.3 Ethernet Operation Mode Ethernet supports either half-duplex, full-duplex, or both operation modes. Early Ethernet supports only the half-duplex mode of operation, where a station can transmit or receive data but not at the same time. In contrast, a station supporting the full-duplex mode of operation can transmit and receive data simultaneously. It was with the development of Fast Ethernet that Ethernet became able to support both half-duplex and full-duplex modes of operation.

8.2.2 First Generation—10BaseT Ethernet

10BaseT is one of the most popular versions of the first generation of Ethernet, and defines the fundamentals of Ethernet technology upon which later generations of Ethernet have been built. This discussion will cover the area of the physical layer, the media access control sublayer, and the logical link control sublayer.

8.2.2.1 Physical Layer of Ethernet The characteristics of the first generation of Ethernet are summarized in Table 8-2, which includes the transmission medium, transmission distance and data rate, and operation mode.

TABLE 8-2

10Base Ethernet Summary

Standards	IEEE standard—year first released	PMD type	Date rate	Maximal distance in meters	
				Half duplex	Full duplex
10Base5	802.3—1983	Coax cable (thick Ethernet)	10 Mbps	500	Not supported yet
10Base2	802.3—1985	Coaxial cable (thin Ethernet)	10 Mbps	185	Not supported yet
1Base5	802.3—1987	2 pairs of twisted telephone cable	1 Mbps	250	Not supported yet
10Base-T	802.3—1990	2 pairs of category 3 or better UTP cable	10 Mbps	100	100
10Base-FL	802.3—1993	Two optical fibers	10 Mbps	2000	>2000

SOURCE: IEEE 2000.

The transmission medium has evolved from the original thick coax (10base5) to twisted pair copper wire and then to fiber. Twisted pair is the most common choice of cable for the first generation of Ethernet. Unshielded twisted pair (UTP) is one kind of twisted pair that has two copper wires twisted together and is relatively immune to noise.

The physical coding sublayer uses Manchester coding, a common coding scheme at the time of first-generation Ethernet that divides each bit into two halves. A 1 is defined by a transition from "low" to "high" in the middle of the bit period, and a 0 is defined as a transition from "high" to "low" in the middle of the bit period.

Most versions of first-generation Ethernet support only the half-duplex mode of operation and have a transmission distance of around 250 m.

8.2.2.2 Ethernet Frame The Ethernet frame defines a structure to hold user data and to be carried on the physical medium. It consists of two parts: a header and the payload data. Figure 8-4 shows the IEEE 802.3 Ethernet frame format, which includes the following:

- *Preamble.* A 7-byte field containing a series of alternating 1s and 0s used by an Ethernet receiver to acquire bit synchronization and frame timing information. This field is generated by the hardware in an Ethernet device.
- *Destination address.* The MAC address of a receiving Ethernet device.
- *Source address.* The MAC address of a sending device.
- *Type.* A 2-byte field indicating the type of data encapsulated, e.g., IP, ARP, RARP, etc.
- *Payload data.* The data field with length ranging from 46 to a maximum of 1500 bytes.
- *Cyclical redundancy check (CRC).* A 4 -byte field used for error detection.

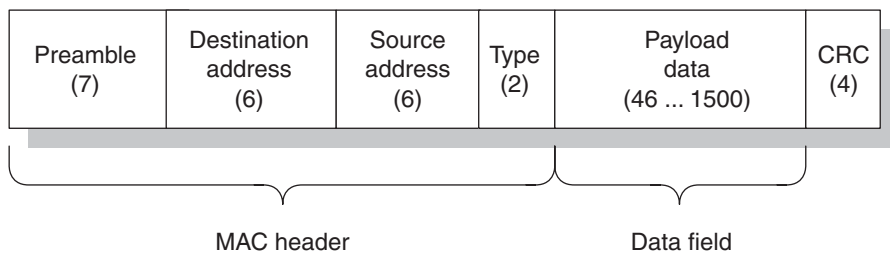


Figure 8-4
The IEEE 802 Ethernet frame structure.

8.2.2.3 Media Access Control Ethernet MAC uses CSMA/CD for access control. By means of carrier sense multiple access, with collision detection, an Ethernet device does the following:

1. Listens to the line before putting a packet "on the wire," and if the line is busy, waits for a predetermined number of seconds before retry
2. When the line becomes idle, transmits while monitoring for collisions
3. If a collision is detected, sends the jam signal and waits for an algorithmically determined number of seconds before resending any packets
4. If the maximum number of transmission attempts is reached, gives up

8.2.3 Second Generation—Fast Ethernet

Fast Ethernet was defined to meet the demands of fast-growing Internet traffic. In the face of fast growth, 10 Base Ethernet became too slow by the early 1990s to meet all the needs of the Internet's data traffic flow. The IEEE reconvened the IEEE 802.3 committee in 1992 to upgrade Ethernet to 100 Mbps. Two competing proposals emerged in the process: one simply aimed at increasing the speed of the existing Ethernet as defined by IEEE 802.3 to 100 Mbps, while the other reworked the old Ethernet with a new architecture. The first proposal resulted in the updated IEEE 802.3 specifications, also known as *Fast Ethernet*, that were approved in 1996. The second resulted in the establishment of the IEEE 802.12 committee and the 802.12 standard specifications in 1995, also known as *100VG-AnyLAN*. This subsection briefly describes Fast Ethernet, while the following subsection discusses 100VG-AnyLAN.

One major change in the Fast Ethernet specifications is that shared medium topologies like the bus topology are eliminated in favor of the star topology in order to decrease transmission collisions and increase network throughput. At the center of the star topology is a switching hub that supports full-duplex operation.

8.2.3.1 Physical Layer The Fast Ethernet specifications define three physical media, or physical medium-dependents: 100Base-T4, 100BaseSE-TX, and 100Base-FX. The 100Base-T4 uses four unshielded

twisted pairs of cable to connect a user station to a hub, a very common situation in office buildings. The 100Base-TX uses two pairs of category 5 unshielded twisted pairs. The 100Base-FX uses a pair of optical fiber cables that are defined by ANSI standards for FDDI. Table 8-3 summarizes the Fast Ethernet physical layer characteristics.

Fast Ethernet adopts a faster coding scheme at the physical signaling sublayer, i.e., the 4-bit/5-bit scheme that uses groups of four data bits as a transmission unit, also called an *encoded symbol,* with the fifth bit as the delimiter.

8.2.3.2 MAC Layer Fast Ethernet retains the original Ethernet MAC layer. All the original frame formats, procedures, and media access control algorithms, i.e., CSMA/CD, remain almost identical. This enables the first-generation of 10-Mbps Ethernet LANs to run over 100 Mbps Fast Ethernet without any changes.

8.2.4 100VG-AnyLAN

The IEEE 802.12 standards, originally approved in 1995, were the result of a competing proposal for upgrading the first generation of Ethernet. The central idea behind 100 VG-AnyLAN is to define a LAN technology that supports the operations of any existing LAN protocol, be it Ethernet frame format and Token Ring frame format, but not both at the same time. The main goals of 100VG-AnyLAN include avoiding the frame collisions of the traditional CSMA/CD access method and providing

TABLE 8-3

100Base Ethernet Summary

Standards	IEEE standard—year first released	PMD type	Maximal distance in meters	
			Half duplex	Full duplex
100Base-TX	802.3—1995	Two pairs of category 5 UTP cable	100	100
100Base-FX	802.3—1995	Two optical fibers	400	2000
100Base-T4	802.3—1995	Four pairs of category 3 or better UTP cable	100	Not supported
100Base-T2	802.3—1997	Two pairs of category 3 or better UTP cable	100	100

SOURCE: IEEE 2000b.

prioritized services on LAN (IEEE 1998a). 100VG-AnyLAN did not achieve wide acceptance in the market, largely due to the overwhelming dominance of Ethernet.

The prioritized service is implemented via a demand priority protocol that utilizes a round robin polling scheme for each station to request a priority for each MAC frame from the repeater. Higher priority is given to delay-sensitive frames, while the best-effort service is given to the rest of the frames.

Collision avoidance is achieved via the exclusive use of a switching hub as opposed to the shared media used by traditional Ethernet. A station can transmit only after it is granted permission to do so by the connected repeater. Thus the access control method is deterministic with no collisions.

8.2.5 Gigabit and 10 Gigabit Ethernet

Gigabit Ethernet and 10 Gigabit Ethernet as transport technologies are introduced in Chap. 7 on optical ethernet, which focuses on the physical layer of both Ethernet technologies. This subsection provides an overview of Gigabit Ethernet and 10 Gigabit Ethernet from the perspective of LAN.

8.2.5.1 Gigabit Ethernet Soon after the Fast Ethernet standards were finalized, the work on 1000Base Ethernet began at the IEEE 802.3z committee. After the specifications were completed, large-scale deployment soon followed.

One primary goal of Gigabit Ethernet, like its predecessor Fast Ethernet, is to alleviate the bandwidth crunch on LANs with 10-fold increase in speed. Gigabit Ethernet also preserves the standard 802.3 Ethernet frame format and the minimum and maximum sizes of the frame, so that it is backward-compatible with 100BaseT and 10BaseT Ethernet.

Gigabit Ethernet supports both full- and half-duplex operations, the same as Fast Ethernet. For half-duplex operations, CSMA/CD is used. For full-duplex operations, the standard flow control defined in IEEE 802.3 is used (IEEE 2001b). At the physical layer, Gigabit Ethernet supports both fiber and copper wire as physical media, although optical fiber is the common choice. It uses the recently defined ANSI Fibre Channel standards as the basis for fiber-based media (ANSI 1998).

Gigabit Ethernet equipment, like Ethernet switch or router equipment, is mainly used for LAN backbone, interconnecting distributed multiple LANs, or connecting a LAN to a backbone IP network.

8.2.5.2 10 Gigabit Ethernet Efforts on the 10 Gigabit Ethernet specifications by the IEEE 802.ae committee were initiated soon after the Gigabit Ethernet specifications were completed. The 10 Gigabit technology clearly targets LAN, the traditional space of Ethernet, and the space beyond LAN such as WAN and MAN. 10 Gigabit Ethernet defines two families of physical layer interfaces: one for local area networks, operating at a data rate of 10 Gbps, and one for wide area networks, operating at a data rate compatible with the payload rate of OC-192c/SDH VC-4-64c. 10 Gigabit Ethernet preserves the standard 802.3 Ethernet frame format and the minimum and maximum sizes of the frame, so that it is backward-compatible with 100BaseT and 10BaseT Ethernet, like Gigabit Ethernet.

One important feature of 10 Gigabit Ethernet is that it supports full-duplex operation only. The traditional Ethernet half-duplex operation for shared connections and CAMA/CD is abandoned.

8.3 Token Ring LAN

Token Ring LAN technology was originally developed by IBM in the 1970s, was originally standardized by the IEEE as the standard IEEE 802.5 in 1985, and then was adopted as ISO 8802.5 (IEEE 1998c). The IEEE 802.5 specification is almost identical to IBM's Token Ring network, with some minor differences. Throughout this chapter, the term *Token Ring* generally is used to refer to both IBM's Token Ring network and IEEE 802.5 network unless noted otherwise.

The Token Ring network is well suited for use in commercial and industrial environments, where predictability of the performance is expected.

8.3.1 Transmission Medium

IBM Token Ring uses twisted pair copper wire as the transmission medium even though IEEE 802.5 does not specify a media type. In more recent deployments, optical fiber cable is also used to extend the size of the ring interconnecting hubs beyond their normal limitations.

With unshielded twisted pair, a very common wiring choice, a Token Ring network can have a maximum of 72 stations or nodes, although in practice the number of nodes is normally smaller. With shielded twisted pair (STP) wiring, the number of attached stations can increase up to 250

in theory. The typical distance of a Token Ring LAN, called an *average ring length* (ARL), is about 100 m, and this distance can be extended 10-fold if optical fiber cable is used between hubs.

The original IEEE 802.5 Token Ring LAN operates at 4 Mbps, but the standard now covers transmission rates up to 16 Mbps.

8.3.2 Token Ring LAN Configuration and Topology

A Token Ring network typically features a ring topology formed from a set of small clusters or stars, as shown in Fig. 8-5. At the center of each star is a multistation access unit (MSAU) with a set of Token Ring stations connected to it. An MSAU is basically a hub device, and each station is connected to it via a twisted pair cable with two wire pairs. One pair receives data and the other is for transmitting data. The MSAUs are connected together with patch cable or optical fiber cable to form a ring.

An MSAU can be passive or active. A passive MSAU merely provides an electrical path for the data to pass through. An active MSAU amplifies the signals passing through it. With active MSAUs, a Token Ring network can extend further in distance.

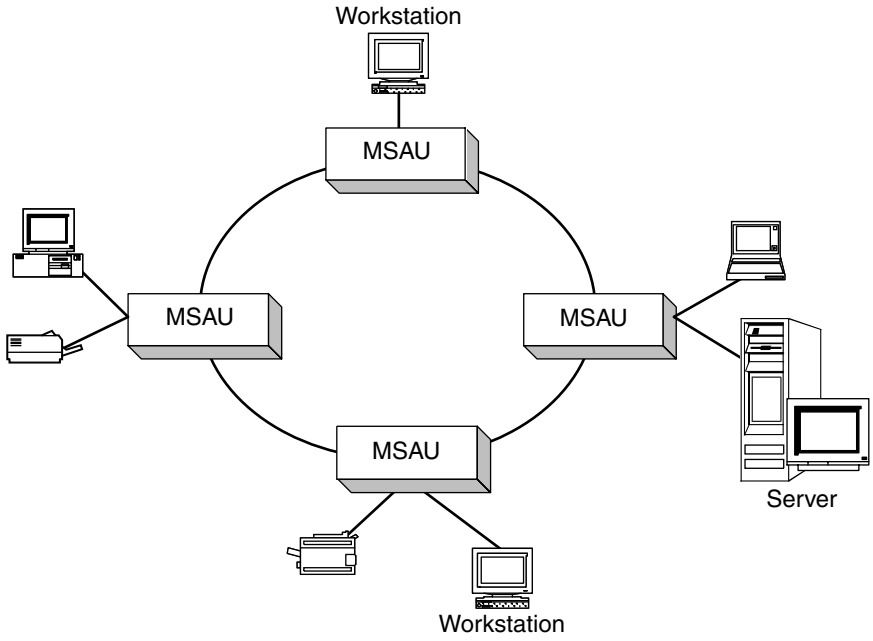


Figure 8-5
A Token Ring network topology.

8.3.3 Media Access Control and Frame Format of Token Ring

As the name of the protocol suggests, the media access method used with Token Ring networks is called *token passing*. This is a deterministic access method that ensures no collisions will occur because only one station can transmit at any given time.

There are two types of frame for Token Ring LAN: token frame and data frame. A token frame is a short frame three octets in length, and can turn into a data frame when the token bit is set to 1, as shown in Fig. 8-6.

The token frame has a start delimiter (SD) and an end delimiter (ED), each with a length of one octet. The access control octet has four fields: priority, token indicator, monitor, and reserved bits. The priority field indicates the frame priority and a station can seize the token only if its own priority is equal to or higher than the token priority. The token indicator bit indicates whether the frame is a token or a data frame. The monitor field prevents any frame from circulating on the ring endlessly. The reserved bits field allows a station with higher priority to reserve the next token to be issued with the indicated priority.

A data frame is a superset of the token frame with additional fields such as destination and source addresses, data, and FCS fields, as shown in Fig. 8-6.

8.3.4 Token Ring LAN Operations

User data travels on the Token Ring network in one direction only, as in other ring topologies. The token frame is passed from node to node. Possession of the token gives a station permission to transmit data.

8.3.4.1 Data Transmission Operations When a station has data to send, it seizes the token first, then changes the token indicator bit to turn the token into the start of a data frame sequence. It then inserts user information and a destination address into the frame and sends it to the next station downstream on the ring.

Each station on the ring examines the data frame and passes it onto the next neighboring station if it is not the intended destination station. The destination station copies the frame for further processing, and sets a bit in the frame to acknowledge receipt of the frame.

The frame continues to circulate the ring until it reaches the sender. The sender removes the frame when it finds that the frame has been "seen" by the destination station. When the sender finishes sending the

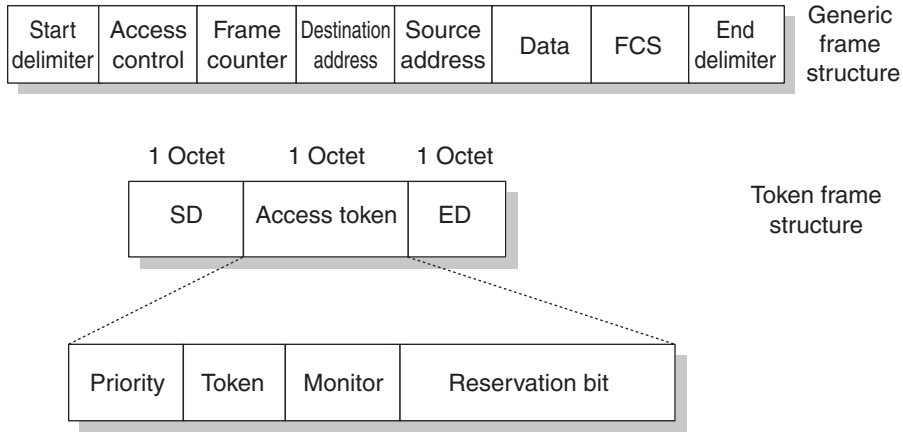


Figure 8-6
Token Ring frame structures.

last frame, it regenerates the token and puts it on the ring to allow other stations on the network to transmit data.

When a data frame is on the ring, no token frame is on the ring at the same time. This prevents two stations from transmitting data simultaneously so that no collisions occur.

8.3.4.2 Priority Token Ring LAN uses a priority system that allows the operator to assign high priority to some stations that can use the network more frequently than others. Briefly, the priority scheme works as follows: The token frame has two fields that control the priority: the priority field and the reservation field. Only those stations with a priority equal to or higher than the priority level contained in the token frame can seize the token. After the token is seized and changed to a data frame, those stations with a priority level higher than that of the transmitting station can reserve the token for the next round of token passing. When the next token frame is generated, it contains the higher-priority level of the reserving station. Once the reserving station finishes sending data, it is responsible for resetting the token frame's priority level to the original level in order to allow other stations a chance to transmit data.

8.3.4.3 Ring Management A station on a Token Ring LAN plays the role of either an active monitor or a standby monitor station. There is only one active monitor on a ring, and it is chosen during a process called the *claim token process*. The active monitor is responsible for maintaining the master clock, issuing a "neighbor notification," which is similar to a keep-alive message, detecting lost tokens and frames and purging the ring to get rid of endlessly circulating frames. Any station on the

ring can be the active monitor station if the current active monitor goes down, via the same claim token process.

8.4 FDDI

Fiber-distributed data interface (FDDI) LAN is another incarnation of Token Ring LAN, defined by ANSI (ANSI 1987, 1988) to fill two needs at the time the protocol was adopted. FDDI is intended to fill the need for a large amount of bandwidth on enterprise LANs and the need for reliable and fault-tolerant networks when enterprises start moving critical applications onto their networks. It was adopted by IEEE as IEEE 802.5 (IEEE 1998c), and by ISO. All the specifications are compatible and completely interoperable.

The FDDI standards define the physical layer and the data link layer of the LAN protocol stack. Specifically, they consist of four separate specifications covering the LAN physical layer protocol, PMD, MAC, and station management.

8.4.1 FDDI Basics

FDDI uses two types of optical fibers as primary transmission media: single-mode fiber, which is more expensive but has higher capacity, and multimode fiber, which is relatively inexpensive but has less capacity. The FDDI specification allows for 2 km between stations using multimode fiber and a longer distance with single-mode fiber, and supports a data rate of 100 Mbps.

The FDDI frame structure is very similar to that of the Token Ring frame structure described earlier, and it can be as large as 4500 octets. Like the Token Ring frame, the FDDI token frame is a subset of a general frame with three fields: a start delimiter, an end delimiter, and a frame control, which have identical fields to the token frame of Token Ring.

8.4.2 FDDI Configuration and Access Control

FDDI uses two counterrotating rings to enhance its fault tolerance capability: a primary ring and a secondary ring. As shown in Fig. 8-7, the secondary ring can be used to provide additional bandwidth or purely as a backup to the primary ring.

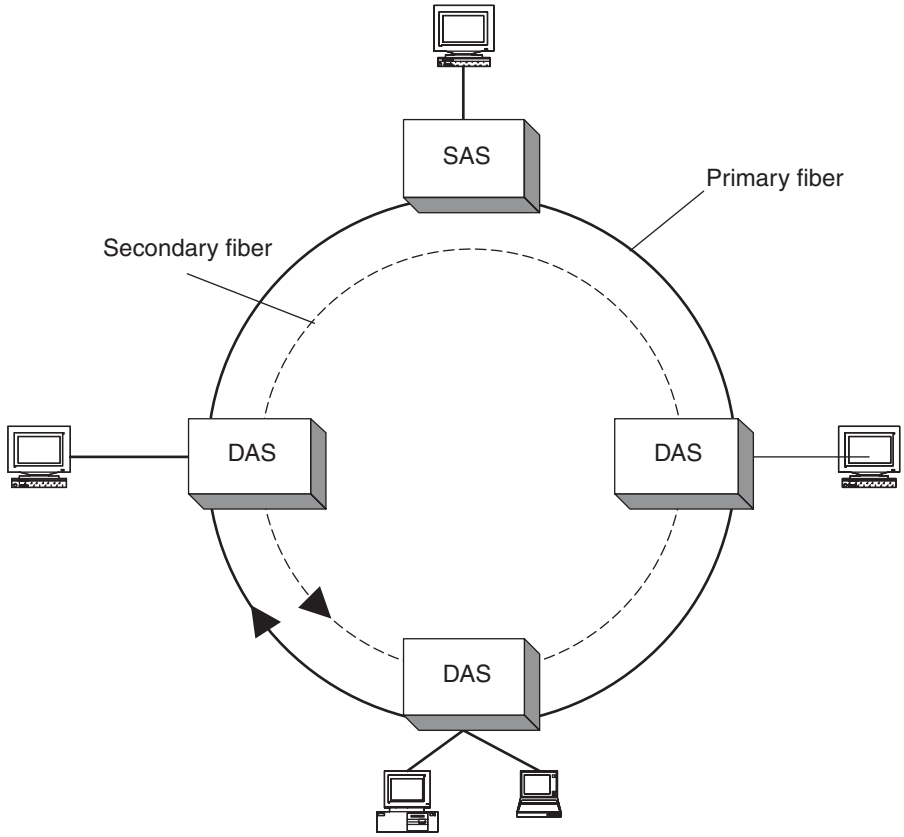


Figure 8-7
Configuration and components of FDDI network.

In an FDDI LAN, there are two kinds of stations: the dual attachment station (DAS), which is connected to both rings, and the single attachment station (SAS), which is attached only to the primary ring. Another FDDI LAN device is the attachment concentrator, which allows multiple DASs or SASs to connect to either ring.

FDDI uses a media access control method that is different from that used by basic Token Ring. As discussed above, Basic Token Ring uses priority and reservation bits in the access control field of the token. In contrast, FDDI uses timed token rotation protocol, which operates as follows: For each rotation of the token, each station computes the time that has expired since it last received the token; this time is called the *token rotation time* (TRT). The TRT includes the time a station needs to transmit any of its waiting frames and the time all other stations in the ring need to transmit any of their waiting frames. TRT will be shorter if the system load is light and longer if the load is heavy. There is a predefined parameter called the *target token rotation time* (TTRT). Upon

receipt of a token, a station computes its TRT and the difference between the TTRT and the just computed TRT. The difference, known as the *token hold time* (THT), decides whether and how long the station can transmit the waiting frames. If the THT is positive, the station can spend up to the amount of time equal to the THT in transmitting data. If the THT is negative, the station cannot transmit any frame for this rotation of the token. This time token rotation protocol prevents a station from holding the token for an excessive amount of time and ensures that all stations have a fair chance to use it. This is the same mechanism the token bus protocol uses.

8.4.3 Station Management

There is one management station that acts as the manager on an FDDI ring, and each station has a station management agent. An agent station communicates with the management station to negotiate TTRT and reports the station status.

8.4.4 CDDI

A standard specification similar to FDDI for copper wire has emerged more recently, called the Copper Distributed Data Interface (CDDI) to be consistent with FDDI naming convention. CDDI is an implementation of the FDDI protocol on the copper medium and supports 100 Mbps over a 100-m distance from a desktop to a concentrator (ANSI 1995).

CDDI was defined by the ANSI X3T9.5 committee. It is officially named the Twisted-Pair Physical Medium-Dependent (TP-PMD) Standard to indicate that the focus of the specification is on the twisted pair physical medium, with rest of the protocol including the MAC algorithm and network configurations identical to that of FDDI.

REVIEW QUESTIONS

1. What are the three media types for LAN? Describe the relationships between transmission distance and data rate.
2. The IEEE 802 LAN standards and protocols cover only the bottom two layers of the network reference model. Describe the responsibilities of each of the two bottom layers in the LAN context.

3. Describe the two media access control methods used for LANs and discuss the characteristics of each.
4. Describe the four LAN topologies and explain which ones are most commonly deployed.
5. Describe the operations of a LAN bridge and the differences between a LAN bridge and a LAN router.
6. Describe the responsibilities of the MAC and LLC sublayers in the LAN protocol stack.
7. Explain why Ethernet is a simple technology in terms of access control methods, network topology, and frame format.
8. Describe the differences between Fast Ethernet and the first generation of Ethernet in terms of transmission media, network topologies, and operation modes.
9. Describe the Token Ring LAN topology and how the token is passed around on a Token Ring LAN.
10. Describe the operations of the CSMA/CD access control method and compare it with the token-passing scheme.
11. Describe how the priority scheme in a Token Ring LAN allows some stations to transmit more data than other stations and how to prevent a frame from circulating the ring indefinitely.
12. Describe how the time token rotation protocol works as used in FDDI and token bus networks. Specifically, discuss how it prevents a station from holding onto the token for an excessive amount of time.
13. Briefly describe CDDI and compare it with FDDI.

REFERENCES

ANSI. 1987. "FDDI—Token Ring Media Access Control (MAC)." ANSI X3.139. Web site: www.anci.org.

ANSI. 1988. "FDDI—Token Ring Physical Layer (PHY) Protocol." ANSI X3.148. Web site: www.anci.org.

ANSI. 1995. "FDDI—Token Ring Twisted Pair Physical Layer Medium Dependent (TP-PMD)." ANSI X3.263. Web site: www.anci.org.

ANSI. 1998. "Fibre Channel: Physical and Signaling Interface—3." ANSI INCITS 303. Web site: www.anci.org.

IEEE. 1990. "Token Passing Bus Access Method and Physical Layer Specifications." IEEE 802.4. Web site: www.ieee.org.

IEEE. 1994. "Distributed Queue Dual Bus (DQDB) Access Method and Physical Layer Specifications." IEEE 802.6. Web site: www.ieee.org.

IEEE. 1997. "IEEE Recommended Practices for Broadband Local Area Networks." IEEE 802.7. Web site: www.ieee.org.

IEEE. 1998a. "Demand-Priority Access Method, Physical Layer and Repeater Specifications." IEEE 802.12. Web site: www.ieee.org.

IEEE. 1998b. "Logical Link Control." IEEE 802.2. Web site: www.ieee.org.

IEEE. 1998c. "Token Ring Access Method and Physical Layer Specification." IEEE 802.5. Web site: www.ieee.org.

IEEE. 1999. "Wireless LAN Medium Access Control (MAC) and Physical Layer Specifications." IEEE 802.11. Web site: www.ieee.org.

IEEE. 2001a. "Air Interface for Fixed Broadband Wireless Access Systems." IEEE 802.16. Web site: www.ieee.org.

IEEE. 2001b. "Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications." 2001 edition. IEEE 802.3. Web site: www.ieee.org.

IEEE. 2001c. "Local and Metropolitan Area Networks: Overview and Architecture." IEEE 802. Web site: www.ieee.org.

IEEE. 2002a. "Media Access Control (MAC) Parameters, Physical Layer and Management Parameters for 10Gb/s Operation." IEEE 802.3ae. Web site: www.ieee.org.

IEEE. 2002b. "Wireless Medium Access Control (MAC) and Physical Layer Specifications for Wireless Personal Area Networks (WPANs™)." IEEE 802.15. Web site: www.ieee.org.

Spurgen, C. 2000. *Ethernet: The Definitive Guide*. O'Reilly and Associations.

CHAPTER 9

Wireless Local Area Networks

9.1 Introduction

Wireless LAN can be used to extend or replace wireline LAN. This section describes the motivations behind the development of wireless LAN, gives a brief history of that development, and describes the scope of wireless LAN.

9.1.1 A Brief History and Wireless LAN Standards

Motivations for wireless LAN include the rapid increase in the number and capabilities of portable computers and hand-held devices with computing capabilities and the increased mobility of users in the LAN environment.

Wireless LAN offers the mobility and increased flexibility needed for group working environments. For example, a group of people with portable computers coming to a meeting forms a wireless LAN on the fly to exchange data and to facilitate communications. Also, in some situations, it is just more economical to use wireless LAN. For example, the cost of installing wireline LANs in places such as old buildings or on factory floors is much higher than installing wireless LANs.

The IEEE 802.11 working group was formed in 1990 to develop a general standard for wireless LAN. Seven years later, the IEEE.802.11 wireless LAN standard was adopted (IEEE 1999a). This base wireless LAN specification defines the physical and media access control layers of LANs with wireless connectivity. Wireless LAN originally called for the use of the 2.4-GHz radio frequency band with a target data rate of 1 to 2 Mbps. But that bandwidth alone was soon found insufficient to meet customer demand. So several extensions to the original IEEE 802.11 standard have either been accepted or are in the process of being developed, listed in Table 9-3. The IEEE 802.11b extension uses a high-frequency band to achieve a data rate up to 11 Mbps. This is also known as *IEEE 802.11 High rate* or *Wi-Fi*. This extended version of wireless LAN was quickly adopted in 1999 and achieved wide acceptance and deployment (IEEE 1999b).

The European counterpart of the IEEE 802.11 wireless LAN specifications is known as *HiperLAN*. Actually there are two specifications, HiperLAN/1 and HiperLAN/2, both of which have been adopted by the European Telecommunications Standards Institute (ETSI). HiperLAN/1

operates in the 5-GHz frequency range, providing a data rate of up to 20 Mbps. HiperLAN/2 operates in the same frequency band, targeting a 54-Mbps data rate.

Globally, wireless LAN operates in one of two frequency ranges that are not regulated and do not require any operating license. One is known as the industrial, scientific, and medical (ISM) band, and is located between 2.40 and 2.49 GHz, with the precise spectrum allocation varying from country to country. The other is located between 5.0 and 5.9 GHz.

9.1.2 WLAN Protocol Stacks

The wireless LAN protocol stack is concerned only with the bottom two layers of the network layer model—the physical and data link layer—the same as its wireline counterpart. The physical layer defines the types of transmission media and transmission and modulation methods. As shown in Fig. 9-1, IEEE 802.11 defines two types of transmission media—radio and infrared. This chapter only describes the radio frequency transmission medium; Chap. 12, on infrared communications and free space optical networks, discusses infrared LAN.

The data link layer consists of two sublayers, the medium access control sublayer and logical link control sublayer. The LLC is the same as its wireline LAN counterpart, the LLC defined in the IEEE 802.2 specifications. The MAC sublayer defines the frame format and methods for access to the medium, which need to take into consideration wireless transmission medium-specific characteristics. This chapter focuses on the physical layer, the MAC sublayer, and the security of wireless LAN.

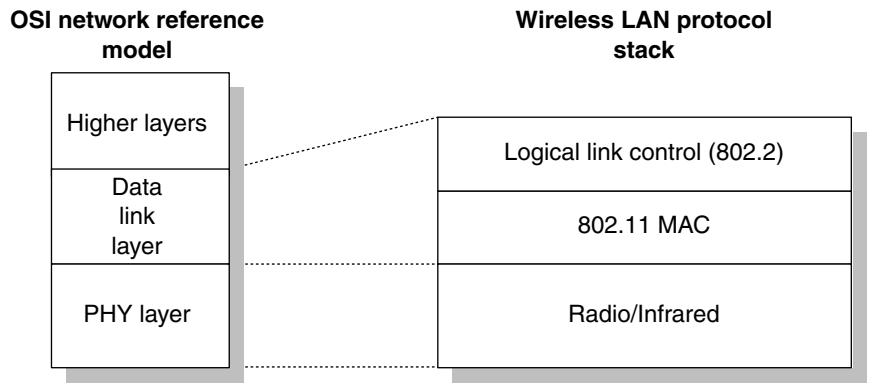


Figure 9-1
Wireless LAN Protocol Stack.

9.2 IEEE 802.11 Wireless LANs

This section describes the physical layer and MAC sublayer of three IEEE 802.11 wireless LAN standards: IEEE 802.11, IEEE 802.11b, and IEEE 802.11a. The term *IEEE 802.11 wireless LAN* is used to refer to both a family of wireless LAN standards developed under the auspices of the IEEE 802.11 committee and to the very first standard in that family (IEEE 802.11). The meaning of the term when used in this chapter will be clear from the context, although in general it will refer to IEEE 802.11.

9.2.1 IEEE 802.11 Wireless LAN

IEEE 802.11 was the first IEEE 802.11 wireless LAN standard, completed in 1997, which provided the foundation in many ways for the wireless LAN standards that followed such as IEEE 802.11b and IEEE 802.11a, although it did not achieved as wide a deployment as IEEE 802.11b.

9.2.1.1 Physical Layer IEEE 802.11 radio-based wireless LANs operate at the 2.4-GHz frequency band, which is one of the unlicensed ISM bands. The specified target data rate is between 1 and 2 Mbps. Different regions around the world allocate this frequency spectrum a little differently. The global spectrum allocation of 2.4 GHz is shown in Table 9-1.

There are two transmission schemes for wireless LAN: direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS). In the United States, the use of spread spectrum technology is mandated in order to limit interference to other applications. The Federal Communications Commission limits transmitter power to 1 W (Stavroulakis 2001).

DIRECT SEQUENCE SPREAD SPECTRUM DSSS is a wireless transmission technique that spreads signals over the whole allocated frequency band

TABLE 9-1

Global Spectrum Allocation of ISM Bands

Country/region	Allocation spectrum, GHz
United States	2.4–2.4835
Europe	2.4–2.4835
Japan	2.471–2.497

Source: Bates (2001).

to minimize the interference from other applications operating in the same frequency bands. A random binary sequence called *spread coding* is combined with the source data by exclusively ORing the data bits with the random binary sequence, then modulated and transmitted. A receiver that knows the spread coding maps back the source data bits. The minimization of cochannel interference is achieved via the randomness of the combined code and the spreading out of the transmission spectrum. The ratio of the number of bits in the binary sequence to the number of data bits is known as the *spreading ratio*. The higher the spreading ratio is, the more resistant the signal is to the interference. The lower the spreading ratio, the more data bandwidth is available to the user. In the United States, the FCC mandates that the spreading ratio be at least 10; the IEEE 802.11 specifications require a spreading ratio of 11 (Burr 2001).

FREQUENCY HOPPING SPREAD SPECTRUM FHSS divides the allocated frequency band into a number of channels. Each channel is of equal bandwidth and is determined by the target data bit rate and the modulation scheme used. A transmitter sends data from each channel for a fixed amount of time, called the *dwell time*. Obviously the transmitter and the receiver must synchronize over the hopping sequence and dwell time or the data will be lost.

There are two kinds of frequency hopping: slow frequency hopping and fast frequency hopping. When the frequency hopping rate is faster than the source data rate, a frequency hopping system is said to be in the *fast frequency hopping mode*, while conversely, if the frequency hopping rate is slower than the data rate, it is in the *slow frequency hopping mode*. In the United States, the FCC requires that the band be split into at least 75 channels and the dwell time be no longer than 400 ms.

One advantage of frequency hopping over DSSS is its flexibility to use an alternative channel within the allocated band. This can be particularly useful with the unlicensed ISM bands because of the possible presence of high-power interference from other applications within the same frequency bands. Another advantage is its resistance to interception and jamming. It is difficult to intercept the signals over a FHSS system because the hopping sequence is random.

9.2.1.2 IEEE 802.11 Medium Access Control Layer The MAC layer ensures that only one transmitter is using the radio frequency at a time. IEEE 801.11 wireless LAN uses a modified version of Ethernet MAC CSMA/CD. Recall that as described in the Ethernet section of Chap. 8,

the "carrier sense" means that a station listens before it transmits any data. If it detects any other station transmitting, it waits and tries again later.

One problem unique to wireless LAN that the MAC layer needs to address is what is known as the *hidden node problem*. As illustrated in Fig. 9-2, station A can reach station B and station B can reach station C, but station A cannot reach station C directly. There is a chance that station C will send data to A while A is in the middle of sending data to C! To avoid this kind of problem, IEEE 802.11 adopts a modified version of CSMA/CD, known as *carrier sense multiple access with collision avoidance* (CSMA/CA).

CSMA/CA uses a four-way hand-shape protocol to ensure that there is only one station transmitting data at a time. Briefly, CSMA/CA works as follows: A station listens to the designated frequency band first. If the frequency channel is busy, it waits a random amount of time before trying again. When the channel becomes available, it sends a *request-to-send* control message that contains the MAC addresses of both the source and the destination stations. Upon receiving the control message, the destination station broadcasts a *clear-to-send* message with the same pair of addresses, but with source and destination reversed. The source station, upon receipt of the reply message, sends a waiting frame. If the frame is positively received, the destination station sends a positive acknowledge message.

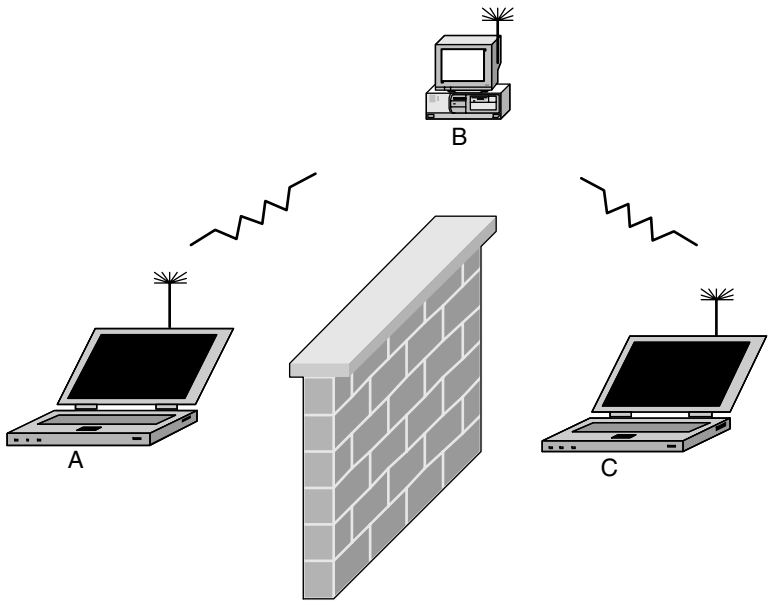


Figure 9-2
Illustration of hidden node issue.

9.2.2 IEEE 802.11b

IEEE 802.11b, an extension to the original IEEE 802.11 wireless LAN standard, was quickly adopted in 1999, not long after the original IEEE 802.11 was published. It was developed mainly to address the issue of the limited bandwidth of the original IEEE 802.11 standard. Also known as *Wi-Fi* (for *wireless fidelity network*), it provides for a raw data throughput of up to 11 Mbps.

IEEE 802.11b deals with the physical layer and the bottom half of the data link layer, i.e., the MAC layer. The upper half of the data link layer, the logical link control sublayer, is covered in the IEEE. 802.2 specification common to both wireless and wireline LANs.

9.2.2.1 Physical Layer IEEE 802.11b operates in the 2.4-GHz frequency band, i.e., the same ISM band with 80 MHz of available spectrum as the original IEEE 802.11 specification.

The IEEE 802.11b physical layer only uses the DSSS radio transmission scheme, while the IEEE 802.11 standard allows a choice of either DSSS or FHSS physical layer transmission, as described above. IEEE 802.11b provides for up to a 11-Mbps raw data rate by using complementary code keying (CCK) with quadrature phase shift keying (QPSK) modulation. In addition, IEEE 802.11b defines a dynamic rate shifting scheme that allows data rates to be automatically adjusted based on noise conditions. In other words, an IEEE 802.11b device can transmit at lower speeds—5.5, 2, or 1 Mbps—when the noise level is high. The device can then automatically "shift" to a high-speed mode when the noise condition improves.

Each 802.11b DSSS channel takes up about 22 MHz of bandwidth with 25 MHz of channel spacing between channels to avoid interference. So the available spectrum can accommodate up to three noninterfering IEEE 802.b channels for a given wireless LAN.

The IEEE 802.11b physical layer is further divided into two sublayers, the Physical Layer Convergence Protocol (PLCP) sublayer and the PMD sublayer. The PMD deals with the wireless transmission medium and the wireless coding. The PLCP presents a common interface for the MAC sublayer and provides the carrier sense and a Clear Channel Assessment (CCA) service to the MAC sublayer. PLCP has two structures: a long and a short preamble to support difference types of service. The long preamble is mandatory for an IEEE 802.11b device to provide data services. The optional short preamble is intended to support real-time sensitive services such as voice, voice over IP (VoIP), and stream video by improving the efficiency of a network's throughput when transmitting this type of data.

9.2.2.2 MAC Sublayer The MAC sublayer provides the interface between the physical layer and the higher layers on a user host and is responsible for the transmission error correction and control of access to the shared radio-link medium.

For error-free transmission, IEEE 802.11b MAC defines two features: cyclic redundancy check and packet fragmentation. CRC ensures that the data is not corrupted in transit. Packet fragmentation allows a large packet to be broken up into small pieces when sent over the air. This reduces the chance for retransmission because the probability of a packet getting corrupted decreases with smaller packet size. This also reduces the retransmission time when data are corrupted because smaller packets can be transmitted more quickly.

The IEEE 802.11b standard specifies one mandatory medium control access method and two optional ones. The mandatory access control method is the same CSMA/CA defined for IEEE 802.11. CSMA/CA is also known as the *distributed coordination function* (DCF) because each station listens for other users independently.

One optional access control method is the point coordination function (PCF), which is used to set up an access point as a point coordinator. With PCF, the point coordinator assigns priority to each client in a given transmission frame and the clients transmit data based on the priority. Another optional access control method is request to send/clear to send (RTS/CTS), a four-way handshake protocol that addresses the hidden node problem described above.

9.2.3 IEEE 802.11a

IEEE 802.11a, another extension to the original IEEE 802.11 wireless LAN standard, is intended to raise the data bandwidth even further than that of IEEE 802.11b. It is different from IEEE 802.11 and IEEE 802b only in the physical layer, and this section highlights the differences.

9.2.3.1 Physical Layer The IEEE 802.11a standard is designed to operate in the 5-GHz frequency range. In the United States, the FCC has allocated 300 MHz of spectrum for unlicensed operation in the 5-GHz block, which is known as the *unlicensed national information infrastructure* (UNII) *band*. Of the allocated 300 MHz, 200 MHz is at 5.15 to 5.35 GHz, with the other 100 MHz at 5.725 to 5.825 GHz. The allocation of the UNII band in Europe is slightly different, as shown in Table 9-2.

The allocated 300-MHz spectrum is divided into three working "regions" with different maximum power requirements and target

applications: The first 100 MHz in the lower-frequency range is restricted to a maximum power output of 50 mW, while the second 100-MHz section has a 250-mW maximum power output. The top 100 MHz is targeted for outdoor applications, with a maximum of 1-W power output. In comparison, the unlicensed ISM band (IEEE 802.11 and IEEE 802.11b) over which wireless LANs operate has only 83 MHz available and thus is relatively more crowded than the 5-GHz unlicensed band.

The target on-air data rate of the IEEE 802.11a wireless LAN is 54 Mbps. However, the on-air data rate can be quite different from the user data throughput, which is the data rate available for end-user applications. Taking into account the overheads of the network protocol, the typical user data throughput is around 30 Mbps. The network protocol overhead for wireless LAN is generally considerably higher than that of its wireline counterpart, due to the more complicated MAC layer message exchanges of wireless LAN described above.

Note that IEEE 802.11b and IEEE 802.11a devices cannot interwork together, for the obvious reason that they use different frequency spectrums.

9.2.4 Comparisons

Table 9-3 compares the three IEEE 802.11 wireless LAN standards discussed in this chapter.

TABLE 9-2

Global 5-GHz Band Allocation

Country/region	Allocation spectrum, GHz
United States	5.15–5.35 5.725–5.825
Europe	5.15–5.35 5.470–5.725
Japan	5.150–5.250

SOURCE: Bates (2001)

TABLE 9-3

Comparison of IEEE 802.11 Wireless LAN Standards

Standard	Modulation	Frequency spectrum, GHz	Raw data rate, Mbps
IEEE 802.11 (1997)	BPSK, QPSK	2.4	1–2
IEEE 802.11b (1999)	CCK	2.4	Up to 11
IEEE 802.11a (2001)	OFDM	5.15–5.25, 5.25–5.35, 5.725–5.825	Up to 54

Note: BPSK: binary phase-shift keying; OFDM: orthogonal frequency division multiplexing.

9.3 IEEE 802.11 Wireless LAN Architecture and Operations

This section describes the wireless LAN architecture and operations generic to all three IEEE 802.11 wireless LAN technologies, i.e., those covered by original IEEE 802.11, IEEE 802.11b, and IEEE 802.11a.

9.3.1 Wireless LAN Components and Configurations

An IEEE 802.11 wireless LAN consists of two types of network devices: client devices and access point. As shown in Fig. 9-3, a client device can be a computer equipped with a wireless network interface card capable of both transmitting and receiving radio signals at a designated frequency band.

IEEE 802.11 wireless LANs have one of two types of configuration: infrastructure and ad hoc. The infrastructure configuration is of the traditional fixed wireless type, where the stations are stationary. In the ad hoc configuration, the stations are more mobile.

9.3.1.1 Infrastructure Configuration In an infrastructure configuration, an access point in the wireless LAN performs the functions of a bridge and a base station. It serves as a bridge between the wireless and a wireline network. For example, an access point with an interface card can be directly connected to a wireline Ethernet LAN. In addition, an access point also serves as base station with both a transmitter and a receiver (or transceiver) that can transmit and receive radio signals to/from the client devices at the predefined radio frequency band. An access point normally has a distance range on the order of 500 ft indoors and 1000 ft outdoors (Webb 1998; Zuo 1999).

An access point with the associated client devices forms a basic service set (BSS), in IEEE 802.11 terminology, which is equivalent to a cell in cellular architecture. A single BSS may constitute a wireless LAN, but often a wireless LAN consists of multiple BSSs, as shown in Fig. 9-3.

An extended service set (ESS), in IEEE 802.11 terminology, is a set of interconnected BSSs, with all their access points connected to a wireline network backbone. A *distribution system* refers to a wireline LAN that connects the wireless LAN to a backbone network or computer server.

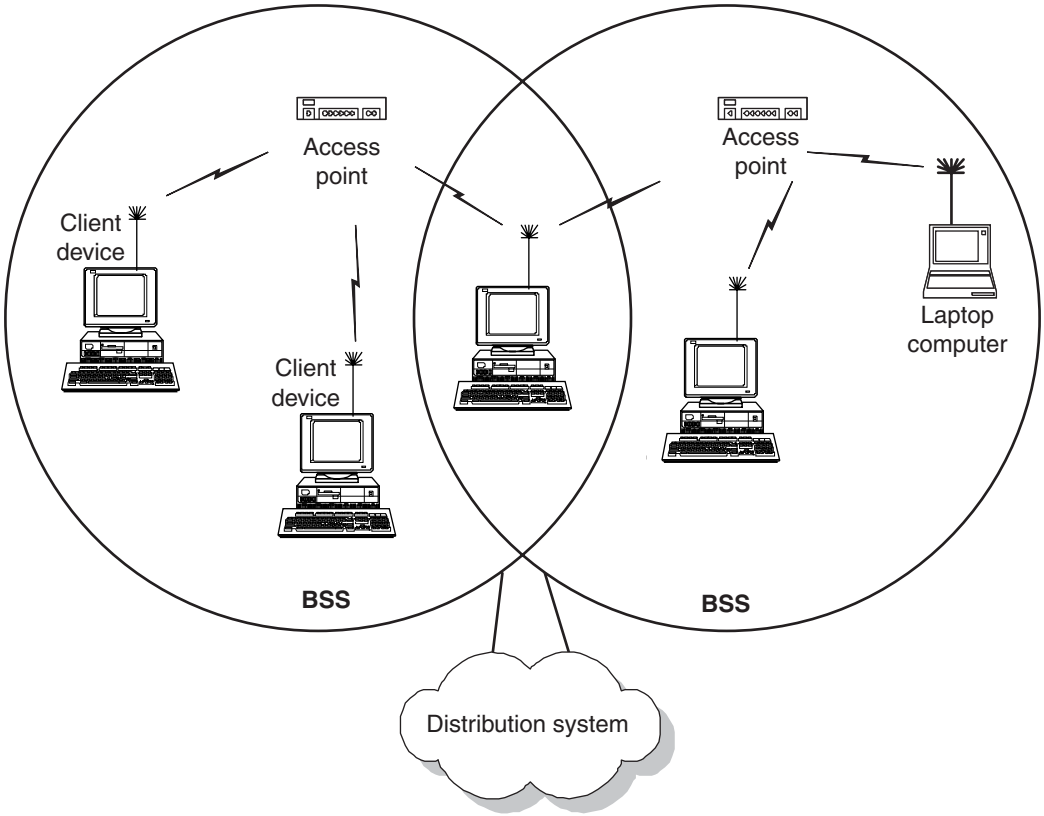


Figure 9-3 *A sample configuration of wireless LAN.*

9.3.1.2 Ad Hoc Configuration An ad hoc configuration is a wireless LAN that is formed on the fly. In certain circumstances, users may desire to build a wireless LAN without any infrastructure like an access point and connected backbone network. For example, say a group of users brought their laptops to a meeting and want to connect their computers at the meeting. To address this need, the IEEE 802.11 standard defines an optional "ad-hoc" configuration or operation mode where a set of stations forms a network "on the fly." An example of an ad hoc configuration is shown in Fig. 9-4 (Perkins 2000; Prem 2000).

In the ad hoc configuration, there are two types of relationships between stations: master-slave and peer-to-peer. In the master-slave relation, one station is elected or selected as the "master" to perform the base station duties. Algorithms such as the "spokesman election algorithm" are defined for this purpose.

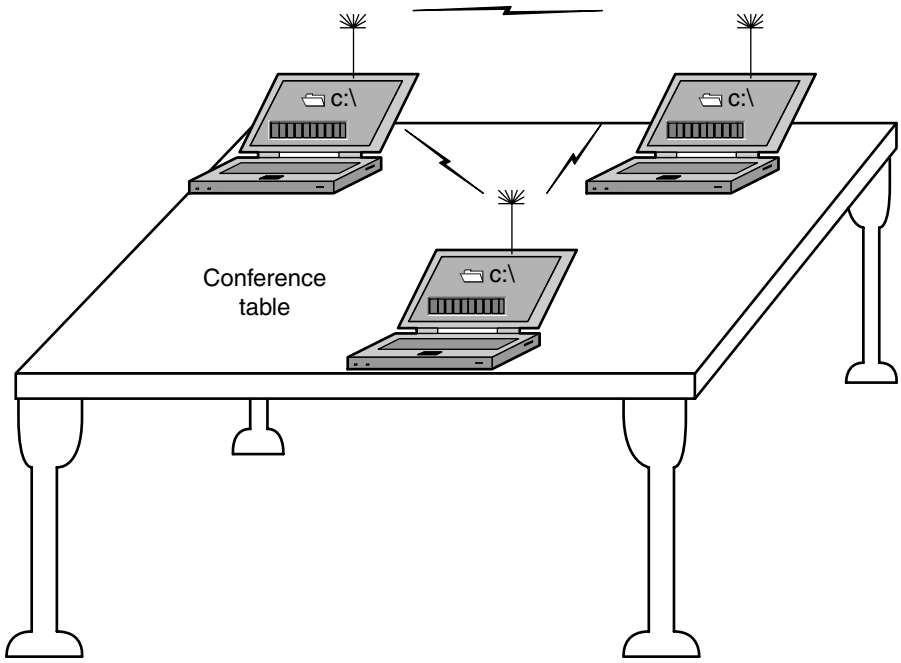


Figure 9-4
A sample ad hoc configuration of IEEE 802.11 wireless LAN.

In a peer-to-peer relation, the "broadcast and flooding method" is used to identify who is who in the network and to exchange the information.

9.3.2 Wireless LAN Operations

This section briefly describes the operations of an IEEE 802.11 wireless LAN, which include the following procedures: synchronizing with the access point as a station first powers up, authenticating the station's identity before it joins the BSS, connecting the station to the access point, transmitting the data, handoff—or roaming—when the station moves from one BSS area to another, and managing power needs.

9.3.2.1 Synchronizing Station and Access Point When a station first powers up or enters a BSS area, it needs to synchronize with the access point—or another station acting as the access point in the case of an ad hoc configuration—to join a BSS. It joins in one of two ways:

1. It passively waits for a beacon frame from the access point. The access point periodically broadcasts beacon frames to all stations. Contained in the beacon frame is the synchronization information.

2. It actively sends a probe request message to find an access point, then waits for a probe response message from an access point.

Either method can be used. The passive approach uses less power but may take longer to synchronize with the access point.

9.3.2.2 Authentication The next step is an authentication process to verify the identity of a station before it joins the BSS. The process includes the log-in and password verification. The null authentication algorithm is the default one while the shared key authentication algorithm is optional. More details on the security framework of the wireless LAN are provided in Sec. 9.4.

9.3.2.3 Association Process This is the last step before the station can start transmitting and receiving data. This is the process where the station and the contacted access point exchange their capabilities, positions, and other information.

9.3.2.4 Data Transmission and Clock Synchronization A station that has gone through synchronization, authentication, and the association process then follows the rules specified in the MAC algorithm to transmit and receive data.

All stations need to keep up synchronization with the access points for purposes such as clock synchronization and power save. The access point periodically transmits a kind of frame known as a *beacon frame* that contains the frame transmission time at the access time. Each station can adjust its clock based on the reception time of the beacon frame.

The beacon frame is also used in power management, as explained in Sec. 9.3.2.6.

9.3.2.5 Handoff The process of a station moving from one BSS area into another without losing the connection and having to reestablish it is known as *roaming*. Roaming is one of the least defined areas in the IEEE 802.11 standards. The standards only specify the message format for handing off from one access point to another and leave the handoff protocol to equipment vendors. A group of vendors including Lucent have jointly defined the Inter-Access Point Protocol (IAPP) to fill the void and to standardize the handoff process between access points of wireless LANs.

9.3.2.6 Power Management All IEEE 802.11 devices have two operating modes: awake and doze. A station in the awake mode is fully powered

and can receive data packets at any time. The doze mode saves power. A station informs the access point of its intention to go into the doze mode before doing so. The transmitter and receiver are powered off for a dozing station. The access point is responsible for buffering data for dozing stations. A dozing station wakes up periodically to check for a beacon frame alerting it to the presence of queued data packets. If there is data waiting, the station sends a poll frame to get the data.

9.4 IEEE 802.11 Wireless LAN Security

As wireless LAN technologies mature, security is becoming one of the key factors that affect a customer's decision of whether to deploy one and where. This section first provides an overview of the security framework of IEEE 802.11 that is generic to all IEEE 802.11 standards, then describes the potential security issues and some security enhancements (IEEE 1999a).

9.4.1 Authentication

Two authentication methods are defined for IEEE 802.11: open system and shared key. In an open system, any station may request authentication. The receiving access point may grant authentication to any requesting station or only to those stations on a user-defined list. An open system uses a default null authentication algorithm that involves a two-step process: identify assertion and request for authentication followed by an authentication result.

In a shared key system, only those stations with the knowledge of a secret key can be authenticated. The IEEE 802.11 standard assumes that the secret key distribution is delivered to all participating stations over a secure channel independent of the IEEE 802.11 wireless communications channel.

9.4.2 Data Encryption

Wired equivalent privacy (WEP), the security framework of IEEE 802.11 wireless LAN, defines the use of 40-bit secret keys for both authentication

and data encryption. A station may maintain two sets of shared keys: a per-station unicast session key and a multicast global key. The majority of current IEEE 802.11 implementations support shared multicast global keys, but per-station unicast session keys are expected to be supported in the near future.

WEP is intended to provide to wireless LAN encryption protection equivalent to that provided to its wireline counterpart. WEP defines a symmetric algorithm in which the same key is used for cipher and decipher.

9.4.3 Potential Security Issues and WEP2

One of the main potential issues is the small WEP key space, which is traditionally 40 bits in length and vulnerable to recovery attempts by intruders. A recovered authentication key can be used for replay attack and authentication forgery.

Dissociation, reassociation, and beacon messages are not authenticated on IEEE 802.11 wireless LAN. A forged dissociation or reassociation request can effectively cause a denial-of-service attack.

Another potential problem is known as *plaintext attack,* which takes advantage of the known plaintext format of many IP messages such as ICMP, ARP, and TCP acknowledgment (ACK) messages. The attacker may recover and alter the packet contents to route the traffic to some place other than the intended destination.

The lack of a centralized authentication, authorization, and accounting support in wireless LANs also makes authorization and authentication less flexible.

In light of the exposed weakness of the original WEP, the IEEE-802.11i Task Group defined a second version of WEP, known as *WEP2,* in the year 2000 that uses a 128-bit key instead of the 40-bit key to increase key space and make key recovery more difficult. Another major change is that the new key can be changed periodically via IEEE 802.1X reauthentication to avoid stale keys (IEEE 2001).

9.4.4 Wireless LAN Security Enhancements

Several potential attacks that were not foreseen when the IEEE 802.11 WEP was initially specified have since become of serious concern. Given the fact that security is a critical factor for customers deciding whether

to install wireless LAN, several enhancements to the existing IEEE 802.11 security mechanisms are under consideration.

9.4.4.1 IEEE 802.1X Security The IEEE 802.1X standard, approved in June 2001, defines a port-based network access control protocol to provide authenticated network access for IEEE LANs including Ethernet, Token Ring, FDDI, and wireless LAN. The basic scope of the standard is to provide a mechanism for authentication and key management. One of its key design goals is to integrate IEEE 802.1 standards with open standards for authentication, authorization, and accounting, including remote-access dial-in user service (RADIUS) and Lightweight Access Directory Protocol (LDAP)(IEEE 2001).

IEEE 802.1X is not a single authentication method, and thus is not an alternative to WEP, 3DES (Triple Data Encryption Standard), or any other cipher. Rather it uses Extensible Authentication Protocol (EAP) as its authentication framework and can support a wide range of authentication methods such as certificate-based authentication, passwords, smartcards, and token cards, among others.

The IEEE 802.1X standard is an access "port" level security mechanism and does not involve per-packet encapsulation. It entails little performance overhead and can be installed on the existing IEEE 802.11 access points and stations via firmware upgrade. IEEE 802.1X entities can be managed via RADIUS and Simple Network Management Protocol. Via RADIUS, IEEE 802.1X provides authentication on a per-user basis. Other per-user services include tunneling, rate limits, and virtual LAN.

9.4.4.2 Remote Access Dial-In User Service Another security enhancement under discussion for wireless LAN is the use of RADIUS. RADIUS is a widely deployed protocol standardized by IETF for network access authentication, authorization, and accounting (AAA) (Rigney et al. 1997). The main advantage of RADIUS is that it is simple and easy to implement and to use and can fit into embedded devices like wireless LAN stations efficiently.

A RADIUS system, based on a distributed client-server architecture, provides secure access to networks and network services. Such a system consists of two main components: a RADIUS *server*, also known as an *authentication server*, and *client access protocol*. The RADIUS server in general resides at a central computer to serve all the client requests. The server authenticates a user against a password file, a network information service database, or a RADIUS database. A RADIUS client sends an authentication request via an independent communication protocol. One advantage of RADIUS architecture is that it simplifies the security process by separating security protocol from the communication protocols.

9.5 HiperLAN

High Performance Radio LAN (HiperLAN) is the European counterpart of IEEE 802.11 wireless LAN standards specified by the European Telecommunications Standard Institute as part of its Broadband Radio Access Network (BRAN) project. The initial version of the HiperLAN specifications was defined during the period of 1991 to 1996, and is known as *HiperLAN/1*. The new version, known as *HiperLAN/2*, was started soon after the initial HiperLAN specifications were completed and published in 2000. This section focuses on HiperLAN/2, and the term *HiperLAN* in general refers to HiperLAN/2 unless otherwise noted (Santamazia 2001).

9.5.1 HiperLAN Architecture and Topology

A hyperLAN network consists of a set of stations, called *mobile terminals* in the HiperLAN specifications, a set of access points (APs), and a connected wireline backbone network. HiperLAN supports the same two types of configurations as IEEE 802.11 wireless LAN: ad hoc and infrastructure, as shown in Fig. 9-5.

HiperLAN architecture includes three layers: physical, data link control, and convergence, as shown in Fig. 9-6. The physical layer and the data link control layer perform functions similar to their IEEE 802.11 counterparts.

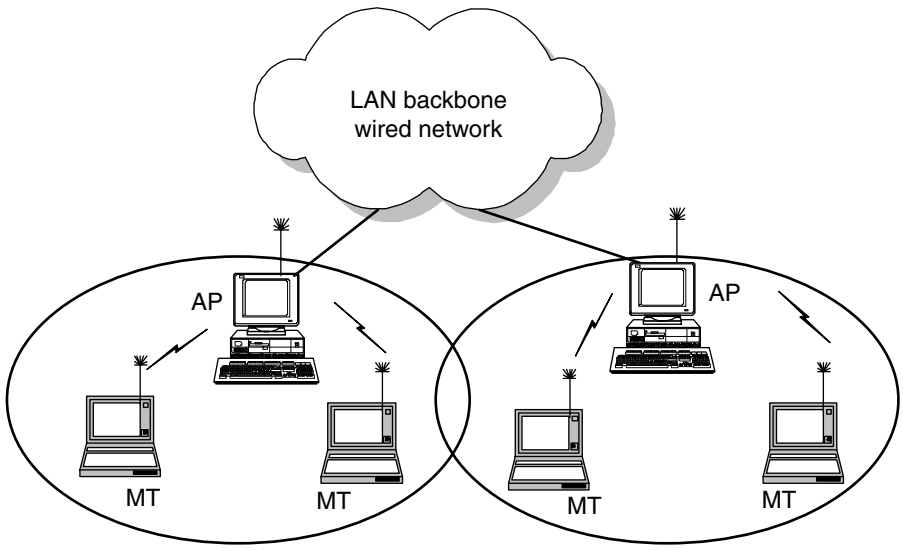


Figure 9-5
Configurations of the HiperLAN network.

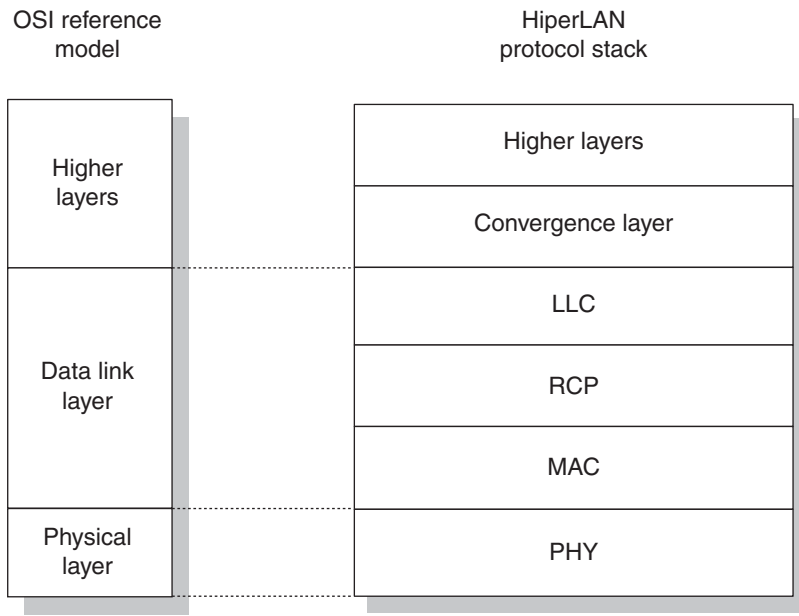


Figure 9-6
The protocol stack of HiperLAN.

Unique to HiperLAN is the convergence layer, which bears a close resemblance to the ATM convergence sublayer and is designed to provide a service access point to the higher layers.

9.5.2 Physical Layer

The physical layer specifies the radio frequency (RF) carrier channels, modulation scheme, channel spacing, and other aspects of radio links. HiperLAN devices operate in the frequency band of 5.15 to 5.3 GHz, with the maximum raw data rate reaching 54 Mbps. The frequency spectrum is divided into five carrier channels. While data transmission takes place on only one carrier channel, HiperLAN devices can operate on all five channels.

The physical layer maps MAC PDUs to physical PDUs and adds PHY signaling such as system parameters and headers required for RF signal synchronization. One factor contributing to the high data rate of HiperLAN is the signal modulation method that is based on the orthogonal frequency division multiplexing. OFDM modulation is combined with several subcarrier modulations and forward error correction to allow for various channel configurations with different data rates, ranging from 6 Mbps all the way up to 54 Mbps. The air interface of HiperLAN/2

is based on time division duplex (TDD) and dynamic time division multiple access (TDMA).

9.5.3 Data Link Control Layer

The data link control layer is responsible for the control of multiple access to the transmission medium, connection setup, and error detection, among other activities. The DLC layer consists of three sublayers: the Radio-Link Control Protocol (RCP), medium access control, and logical link control.

The RCP sublayer consists of three main functional blocks: DLC connection control, which is responsible for DLC connection setup and monitoring; radio resource control (RRC), which is responsible for radio resource management, channel selection, and channel monitoring; and association control, which deals with the association and reassociation between a mobile terminal and an access point.

The MAC sublayer is responsible for controlling the access by multiple mobile terminals to the shared radio-link medium. The air interface is based on TDD and dynamic TDMA, which allow for simultaneous two-way communications between an access point and a terminal.

The HiperLAN MAC frame is quite different from that of IEEE 802.11, which is based on the Ethernet MAC frame. The HiperLAN MAC frame consists of four parts: the broadcast channel, the down-link channel, the up-link channel, and the random access channel.

One attribute that sets HiperLAN2 apart is its supports for quality of service and traffic priority. It can offer different QoS for different connections and thus allows for a mix of different types of traffic, some real-time-sensitive (e.g., voice and video) and some real-time-insensitive (e.g., data application). HiperLAN assigns channel access priority dynamically to packets, based on the packet lifetime and user priority. For example, the shorter the packet lifetime gets, the higher the packet priority becomes.

9.5.4 Convergence Sublayer

HiperLAN defines a convergence sublayer between the data link control layer and the upper layers to provide a service access point to the higher layers. The main function of the convergence sublayer is to adapt service requests from the higher layers to the services offered by

the DLC and to convert the higher-layer packets into fixed-size DLC service data units.

The convergence sublayer is specific to each upper layer it supports. Currently there are two types of convergence sublayer that have been defined: cell-based and packet-based. The former supports ATM networks and their services while the latter interworks with wireline LAN technologies such as Ethernet-based IP.

9.5.5 Security

Like its IEEE 802.11 counterpart, the HiperLAN security scheme supports authentication and encryption. An access point and the connected mobile terminals (MTs) can authenticate each other to guard against unauthorized access to the HiperLAN network. Encryption can be applied to both control data and payload data. Control messages like the signaling messages for connection establishment, are encrypted to prevent any eavesdropping and masquerade attacks.

In addition, HiperLAN has a double ID system to restrict unauthorized access to network resources. In HiperLAN, each communicating node that can be an access point or a mobile terminal is given a HiperLAN ID (HID) and a node ID (NID). The combination of these two IDs uniquely identifies any station. All nodes with the same HID can communicate with each other, but stations from other networks cannot, which prevents unauthorized intrusion.

REVIEW QUESTIONS

1. What physical medium choices are specified for wireless LAN in IEEE 802.11 and which is covered in this chapter?
2. The IEEE 802.11 standards specify two radio transmission schemes, FHSS and DSSS. They both are of the spread spectrum scheme. Describe the motivation for using the spread spectrum transmission scheme for wireless LAN.
3. Describe the components and configuration of a wireless LAN. Describe the functions of an access point in such a wireless LAN.
4. Describe the main differences and similarities between the original IEEE 802.11 and IEEE 802.11b standards in terms of physical layer characteristics.

5. Describe the main differences between the IEEE 802.11a and IEEE 802.11b wireless LANs and explain why the devices of the two standards cannot work together.
6. Describe how the IEEE 802.11 MAC protocol works and how it addresses the hidden node problem. Also compare the IEEE 802.11 MAC protocol against the IEEE 802.3 (Ethernet) MAC protocol.
7. Describe how a station gets connected to an IEEE 802.11 wireless LAN when it is first powered up and how the power-saving mode works in an IEEE 802.11 device.
8. Describe the ad hoc wireless configuration, its application, and the ways the stations in such a network are connected to each other.
9. Describe the services provided by WEP to the wireless LAN and how the IEEE 802.1X specifications complement WEP.
10. Compare HiperLAN architecture with IEEE 802.11a architecture and describe the major similarities and differences.

REFERENCES

Bates, R. 2001. *Wireless Broadband Handbook*. New York: McGraw-Hill.

Burr, A. 2001. *Modulation and Coding for Wireless Communications*. Reading, MA: Addison-Wesley.

IEEE 802.11. 1999a. "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications." IEEE 802.11, Part II. Web site: www.ieee.org.

IEEE. 1999b. "Supplement to 802.11–1999, Wireless LAN MAC and PHY Specifications: Higher Speed Physical layer (PHY) Extension in the 2.4-GHz Band." IEEE 802.11b. Web site: www.ieee.org.

IEEE. 2001. "Part-Based Network Access Control." IEEE 802.1x. Web site: www.ieee.org.

O'Hara, B., and A. Petrick. 1999. *The IEEE 802.11 Handbook: A Designer's Companion*. IEEE Press.

Perkins, C. 2000. *Ad Hoc Networking*. Reading, MA: Addison-Wesley.

Prem, E. 2000. "Wireless Local Area Networks." White paper. Web site: www.cis.ohio-state.edu.

Rigney, C., and Rubens, A. 1997. "Remote Authentication Dial In User Service (RADIUS)." IETF RFC 2138. Web site: www.ietf.org.

Santamaria, A., ed. 2001. *Wireless LAN Standards and Applications*. Norwood, MA: Artech House.

Stavroulakis, P. 2001. *Wireless Local Loops: Theory and Applications.* New York: John Wiley & Sons.

Webb, W. 1998. *Introduction to Wireless Local Loop.* Norwood, MA: Artech House.

Zuo, Z. 1999. "In-Building Wireless LANs." White paper. Web site: www.cis.ohio-state.edu.

CHAPTER 10

LMDS, MMDS, and Wireless Broadband Access

10.1 Introduction

This chapter discusses three fixed wireless broadband access technologies, i.e., local multipoint distribution service (LMDS), multichannel multipoint distribution service (MMDS), and non-line-of-sight broadband wireless access (NLOS BWA). All three operate in the licensed radio frequency bands, though NLOS BWA may also operates in unlicensed RF bands.

LMDS is a broadband wireless local access technology. The term *multipoint* means that it is a point-to-multipoint broadcast technology. LMDS operates in the licensed RF range of 28 to 38 GHz. In the United States, this block of frequencies became commercially available only in 1998.

MMDS is another broadcast broadband wireless access technology that operates in two licensed RF bands: 2.15 to 2.16 GHz for wireless cable television service and 2.5 to 2.686 GHz for educational TV broadcasting.

NLOS BWA is a new generation of wireless networks that operates either on the MMDS frequency bands or the unlicensed industrial, scientific, and medical (ISM) band located between 2.40 and 2.49 GHz. NLOS BWA focuses on the non-line-of-sight and high-bandwidth wireless access. One of its main goals is to broaden the market reach of broadband fixed wireless access so it is not constrained by terrain and other limiting conditions in order to reach the massive residential access market. Unlike IEEE 802.11 wireless LAN described in Chap. 9, however, NLOS BWA is not yet standardized in its current form.

10.2 Local Multipoint Distribution Service (LMDS)

In the term *local multipoint distribution service, local* refers to the signal transmission range limit, and, as already noted, *multipoint* means that it is a point-to-multipoint, broadcast technology. Operating in the range of 28 to 38 GHz, it is intended to provide an alternative to wireline-based solutions to the broadband "last mile" access problem such as cable modem and xDSL. LMDS systems have the advantages of easy deployment, low up-front cost, and relative simplicity of system development (Bates 2000).

Since LMDS systems are susceptible to problems relating to rain and their line-of-sight (LOS) requirement, reliability is a main concern.

Another limitation is that the customer premise equipment (CPE) needed for LMDS systems in general is more bulky and expensive than that needed for MMDS systems.

10.2.1 LMDS System Components

An LMDS system consists of a base station and a set of CPEs connected to the base station via wireless links, as shown in Fig. 10-1, which depicts the configuration of an LMDS system deployed in a metro area to serve business customers (Tipparaju 1999).

10.2.1.1 LMDS Base Station A base station can be either omnidirectional or sectorized. An omnidirectional base station can serve users in

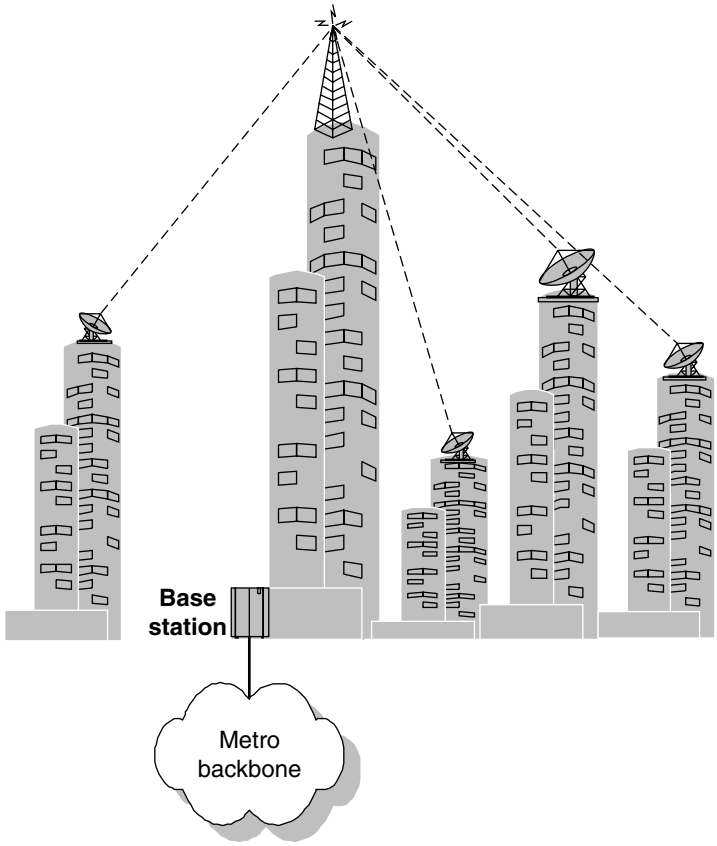


Figure 10-1
Configuration of an LMDS system.

all directions within the coverage area while a sectorized base station beams out signals at an angle and thus serves only those users that fall into the coverage area of that direction.

An LMDS base station consists of an outdoor antenna and indoor hub equipment. In general, it performs the following functions:

- Radio signal transmission and reception
- Modulation and demodulation that implements a modulation method
- Interfacing with the wireline public data network such as a central office or cable network head end
- Network management of both the base station and CPE

The outdoor LMDS base station antenna transmits and receives the wireless signals and needs to be sufficiently tall to reach the entire coverage area because of the system's line-of-sight requirement.

Interfacing with the wireline public data network involves interfaces such as T1, DS3, OC3, or higher-capacity network interface modules. Some base stations also have data switching capabilities.

The network management system is responsible for managing the whole LMDS system, including CPE and the base station itself. The more recent LMDS management systems use SNMP or Common Object Request Brokered Architecture-based (CORBA-based) management protocols.

10.2.1.2 LMDS CPE LMDS CPE products from different CPE vendors are all different, but they generally consist of a user outdoor antenna and indoor network interface units.

The user antennas—which can be as small as a dinner plate—are positioned outdoors on user roofs because of the line-of-sight requirement. A CPE antenna can be quickly installed, has low incremental costs once the basic infrastructure is in place, and involves very low maintenance.

The indoor CPE can be an Ethernet router or switch, an ATM switch, or simply a network interface card if the customer is a business user, or a set-top box or home network Ethernet hub if the customer is a residential or home office user.

Earlier LMDS systems required extremely fine alignment between base station transmission and the CPE antenna. With advanced forwarding error correction (FEC) technologies, however, the newer transceiver allows for a greater margin of error.

10.2.2 LMDS System Design

LMDS as a technology is not standardized, and thus vendors of LMDS systems are free to choose their own modulation scheme, multiplexing method, and frequency channel spacing.

10.2.2.1 LDMS Cell Design Several key factors affect the design of an LDMS cell. One is *size*. A larger cell requires a more powerful base station antenna with a higher transmission tower. Reliability decreases as the size of the cell increases because the signals weaken with distance. The projected traffic capacity determines the maximum number of users a cell can support.

Another cell design issue is *link budge*, which is the process of estimating the maximum distance a user can be reached within a cell while still achieving an acceptable level of QoS.

10.2.2.2 Modulation Techniques Due to the fact that LMDS involves fixed user stations, the dense modulation method can be used to achieve higher data throughput. While LMDS systems until now have used modulation schemes like QPSK and 16-quadrature amplitude modulation (16-QAM), more advanced modulation methods are being considered for the newer LMDS systems. These include 64-QAM, orthogonal frequency division multiplexing (OFDM), and some variants of OFDM.

10.2.2.3 Multiplexing Scheme In the downstream direction, the choices of multiplexing schemes include code division multiple access, frequency multiple access (FDMA), time division multiple access, or variants of these three schemes. FDMA allows a relatively fixed bandwidth and thus is more suitable for applications where the bandwidth requirement changes slowly over time. Most LMDS systems use TDMA for upstream links.

10.2.2.4 Frequency Reuse and Spacing Wireless channel spacing, the frequency band forming an individual wireless channel, includes 112, 56, 28, 14, 7, and 3.5 MHz. Regions such as the United States, Europe, and Japan adopt different channel spacings for different applications.

Reuse of frequencies between the cells can substantially increase the data bandwidth. A cellular pattern—hexagonal or rectangular, for example—allows for multiple frequencies within a section with minimal interference to the neighboring sector. A sector is made up of a number of cells. One often-used technique for frequency reuse is the maximization of isolation between adjacent sectors by polarization. Another technique

is maximizing directivity of the cell antennas while minimizing cross polarization and multipath effects.

10.2.3 Digital Audio Video Council LMDS Specifications

Various industry forums and standards bodies have developed recommendations and specifications for LDMS to encourage interoperability between different vendors' LMDS systems. One such recommendation is by the international industry alliance Digital Audio Video Council (DAVIC).

The DAVIC specifications require that ATM cells be used to carry the payload data in both downstream and upstream directions. In the downstream direction, ATM cells are used to carry user data in MPEG2. A packet framing format is defined to concatenate several cells into packets to be transmitted on the downstream channel. Control information bytes and the packet header format are also defined for LMDS services. For example, seven ATM cells (53 bytes each) plus three control bytes form two 187-byte packets to carry MPEG2 data.

In the upstream direction, LMDS uses TDMA for data transmission. A time slot is allocated to a particular user, and the user can transmit data only during the designated time slot. The user data is carried in AMT cells along with parity check and control bytes.

A MAC protocol similar to the one used in Ethernet allocates resource to user stations. As just described, the LMDS data frames are encapsulated in ATM cells, and each frame in the downstream direction takes two time slots to transmit. In the upstream direction, there are three types of time slots: polling, contention, and user. The polling slot is used for base stations to poll user stations, and the contention slot is used to resolve collisions of multiple user transmissions. The algorithm for resolving a collision is the same as the Ethernet MAC algorithm: each user station waits for a random amount of time before the next try.

10.3 Multichannel Multipoint Distribution Service (MMDS)

MMDS, by definition, is a wireless service rather than a specific technology. This section discusses MMDS concepts, MMDS system components, and physical layer specifications.

10.3.1 Introduction

Multichannel multipoint distribution service, often referred to as *wireless cable* or *wireless DSL*, is an alternative to broadband wireline access technologies. It owes its origin to two other services: multipoint distribution system (MDS) and instructional television fixed service (ITFS). MDS operates in the 2.15- to 2.16-GHz band for wireless cable television service and ITFS operates in the 2.5- to 2.686-GHz band for educational TV broadcasting (as its name suggests). MMDS in its current form was developed in 1983 when the U.S. FCC combined frequencies allocated for MDS and IFTS and renamed it "multichannel multipoint distribution service" (Hybrid 2000).

MMDS operates in the frequency ranges of 2.1 to 2.2 GHz and 2.5 to 2.7 GHz. In the late 1990s, the frequencies were allowed to be used for digital data and voice services. In 1998, the FCC approved two-way digital transmission over the MMDS frequencies. Those frequencies were auctioned in 1999 in the United States, and currently the majority of the commercially available MMDS frequency bands are owned by a small number of service providers such as Sprint and MCI/Worldcom.

Currently MMDS primarily provides two types of service: broadcast TV and broadband digital data access. It is intended to provide an alternative low-cost solution to the "last mile" access problem. The cost of MMDS network equipment is said to be lower than that of LDMS and is also easier to install and provision.

Efforts are underway in various standardization bodies to standardize the wireless broadband access technologies. One of the primary goals of the IEEE 802.16 committees is the development of fixed broadband wireless access standards. Various IEEE 802.16 committees were formed to advance a standard for the development of broadband wireless systems in the licensed frequency spectrum from 2 to 11 GHz, and one section of the standard deals with the MMDS, which operates in the 2.5-GHz band in North America and the 3.5-GHz band in many other countries. In Europe, similar efforts are underway for the broadband wireless access known as *Broadband Radio Access Network*. In Canada, it is referred to as *Local Multipoint Communication Systems*.

10.3.2 System Components

Architecture-wise, an MMDS system looks rather similar to an LMDS system, and is relatively simple in technology. A generic MMDS system, as shown in Fig. 10-2, consists of a base station and a set of CPE antennas,

as described below. MMDS is a point-to-multipoint system, and current generation of MMDS requires line of sight, i.e., no obstruction between a transmitter and a receiver. The MMDS system shown in Fig. 10-2 is a super cell architecture, where a single tower covers the whole serving area.

10.3.2.1 Base Station An MMDS base station, which is also called a *fixed wireless hub*, consists of an antenna and the communication hub equipment. The hub equipment normally consists of a router or switch that interfaces with the central office switching equipment or backbone routers. The base station can have different heights, and there is a tradeoff between the coverage area and height of the base station antenna. The higher the antenna tower is, the wider the coverage area. But the cost of building high towers is also higher. On average, a super cell architecture covers an area between 30 and 40 mi in radius. So many metropolitan areas restrict the maximum height of communication antenna towers. For example, in Seattle, Washington, the towers can be no higher than 35 ft.

10.3.2.2 CPE The customer premise equipment generally consists of a CPE antenna, a transceiver unit, and wireless desktop modem.

The CPE antenna and transceiver unit in most cases are rooftop-mounted and are connected to the desktop unit via a coaxial cable to

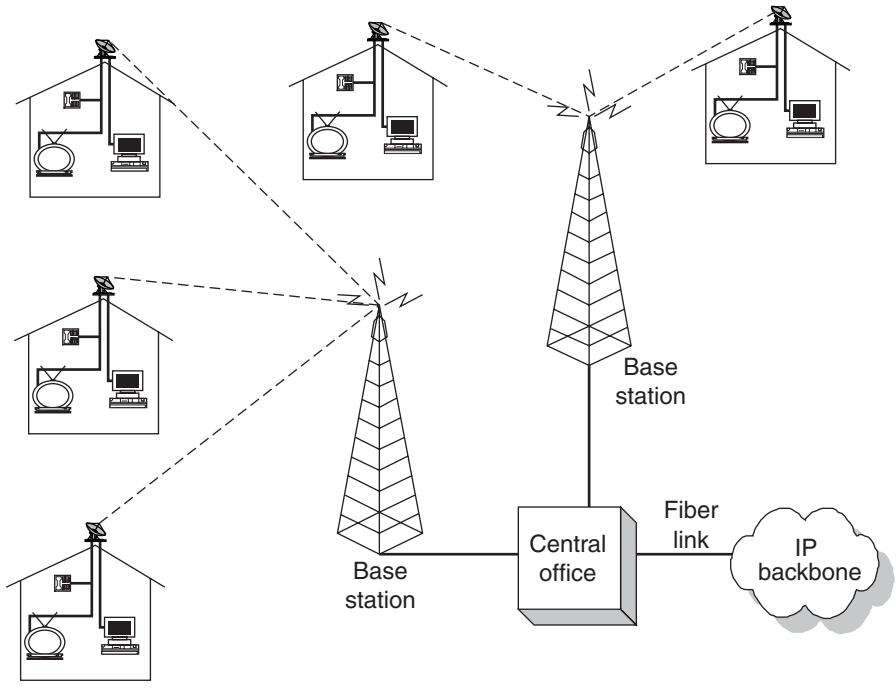


Figure 10-2
MMDS system architecture.

provide a two-way signal path and power to the transceiver. Many CPE antennas are capable of transmitting and receiving signals for two-way services. The transceiver is responsible for converting between radio signals and application signals such as analogue TV channel frequencies or digital signals.

A wireless desktop modem is connected to the CPE antenna via a coaxial cable and is responsible for detecting and selecting traffic destined for the CPE. The modulation techniques used at a wireless desktop modem range from 64 QAM to 16 QAM to QPSK.

10.3.2.3 Physical Layer The MMDS physical layer IEEE 802.16 specification deals with modulation technique, forward error correction, the channel multiplexing scheme, and the target raw data throughput (IEEE 2001).

For an MMDS system, the downstream capacity can reach 30 Mbps of raw data throughput and the upstream capacity can reach 200 Kbps of raw data throughput.

The IEEE 802.16 MMDS uses Reed Solomon coding to implement forward error correction (FEC). This type of coding adds 16 additional parity bytes to each 188-byte segment of raw data and can correct up to eight errors per segment. For downstream modulation, MMDS uses either 64 QAM or QPSK. 64 QAM can produce a spectral efficiency of 5 bits/Hz. With QPSK, the downstream data is broadcast in time division multiplexing format and has a spectral efficiency of 1.6 bits/Hz. MMDS makes use of the MPEG2 video compression standards for video data transmission for broadcast TV services.

A number of multiplexing schemes are specified for MMDS systems. One is time division multiple access with frequency division duplex (FDD). This is the most common choice for MMDS systems. Downstream transmission from the base station antenna to the user premise antennas is broadcast to all users, and only those with the designated decoding capability can accept the signals.

Upstream transmission uses a frequency different from that used for downstream transmission. Each user is dynamically assigned a time slot to transmit data in the upstream direction.

Two variants of the above TDMA/FDD technology are FDMA/FDD and TDMA/TDD. FDMA/FDD is the same as TDMA/FDD except that frequency rather than time is slotted for upstream traffic from different users. This scheme requires a wider frequency band to accommodate multiple users and is more expensive. In TDMA/TDD, downstream and upstream data are transmitted on the same frequency. Upstream data from multiple users are transmitted on the same frequency with a designated time slot for each user.

CDMA/FDD is yet another multiplex scheme, which assigns to each user a unique code while all users use the same frequency band. However, there is an associated limitation on the instantaneous user bandwidth.

10.3.3 Comparison Between MMDS and LMDS Systems

MMDS, as a fixed wireless technology, can use denser modulation schemes than LMDS to achieve higher data throughput without loss of quality. Weather conditions such as rain and fog have no effect on MMDS systems, and they are simpler to develop and thus less expensive than LMDS systems. The large amount of bandwidth MMDS systems can provide (up to 30 Mbps) make them a viable alternative to cable modem and xDSL for small businesses, home offices, telecommuting, and other types of applications that require bandwidths between T1 and T3.

There are two main limitations to MMDS systems: Currently all MMDS systems are line-of-sight—that is, if an MMDS receiver cannot "see" an MMDS transmitter, the link is impaired. Also, only a limited spectrum is allocated for MMDS—2.1 to 2.2 GHz and 2.5 to 2.7 GHz in the United States—which means the available bandwidth is limited.

Compared to LMDS systems, MMDS systems operate on lower frequencies, have a greater coverage range, and require less powerful signals. For MMDS, CPE needs in particular are less complicated and cheaper than they are for LMDS, giving MMDS a wider addressable market. Table 10-1 provides a summary comparison between the two.

10.4 Non-Line-of-Sight Broadband Wireless Access

This section describes a new generation of fixed broadband wireless access (BWA) technology that focuses on overcoming the line-of-sight limitation of MMDS systems, providing a much higher bandwidth, and allowing greater mobility of use within customer premises. This new generation of BWA operates in the MMDS frequency bands or the unlicensed ISM frequency bands between 2 to 3 GHz.

Non-line-of-sight, i.e., removal of the line-of-sight requirement, is important for achieving a large-scale deployment of BWA systems. The

TABLE 10-1

Comparison Between MMDS and LMDS

	LMDS	MMDS
Frequency bands	20–40 GHz	2–3 GHz
Line of sight	Yes	Yes
Rain fade of signal	Yes	No
Cost	CPE cost relatively high	CPE cost relatively low
Base station antenna	Small	Larger than that of LMDS
Modulation	QPSK, 16 QAM, and 64 QAM for downstream; QPSK for upstream	16 QAM and 64 QAM for downstream; QPSK for upstream
Bandwidth/data throughput	Over 1-Gbps downstream bandwidth reported	Over 50-Mbps downstream bandwidth reported
Available frequency bands	Large frequency bands	Limited between 2 and 3 GHz.

line-of-sight requirement limits the deployment of wireless systems, particularly in population-dense metropolitan areas where either the height of antenna towers is limited or the nature of the local terrain makes it difficult to achieve line-of-sight transmission.

The development and deployment of NLOS BWA system are at an early stage. The BWA solutions so far are all proprietary, and proprietary solutions make it difficult to achieve economies of scale. However, the initial standardization efforts are already underway at standards organizations like IEEE and ETSI.

10.4.1 BWA System Configuration

All fixed wireless systems, whether LMDS, MMDS, or BWA, have similar architectures, which include base station transceiver system (BTS) components and a set of CPE components. One BTS along with the CPEs it serves forms a cell. In general, there are three types of cell structures for fixed wireless systems: super cell, macro cell, and micro cell.

A *super cell*, or *mega cell* as it is sometimes called, employs a tall BTS antenna to serve an area in radius of up to 25 mi. In this structure, the BTS antenna height is typically in excess of hundreds of feet, and outdoor rooftop CPE antennas are needed for line-of-sight connection. Each cell is relatively independent of other cells, and there is little frequency reuse between cells. The super cell architecture is more suitable

for line-of-sight technology such as MMDS. One drawback of this architecture is the difficulty it presents for scaling the network up.

Macro cell architecture is a common choice for BWA network deployment. It has a coverage area smaller than that of a super cell but larger than that of a micro cell, with overlapping reuse of frequencies between neighboring cells.

Micro cell architecture is similar to macro cell architecture, except that its coverage area is much smaller—up to a radius of 1 mi. The antenna towers are much lower, typically somewhere between 20 to 40 ft high. The CPE antennas of this architecture can be indoors, and omnidirectional.

A BWA system, as shown in Fig. 10-3, consists of multiple cells, each with a BTS serving a set of end customer CPEs. LOS is not a requirement in the system depicted. The BTS antenna does not need to "see" directly all the CPE antennas even though some are located indoors. The frequencies are reused between cells, and each cell covers an area up to 5 mi in radius.

10.4.2 Non-Line-of-Sight Technologies

The design of a BWA system needs to overcome a number of factors to achieve non-line-of-sight effectiveness (Iospan 2001; Greenstein et al. 1999). They include the following:

Multipath. This refers to the condition where multiple copies of the signals sent by the transmitter reach the receivers via different paths and at slightly different times through reflection off local terrain such as buildings or hills.

Rayleigh fading. This refers to the weakening of the received signals due to the multipath effect, reflection, and other interference. There are three distinct types of fading within a given channel, all of which are generically called Rayleigh fading:

- Frequency-selective fading, where the signal fades at a certain frequency within a given channel
- Space-selective fading, where the signal fades at different locations
- Time-selective fading, where the signal fades at the receiver with time

The other factors contributing to the line-of-sight limitation include

- *Distance.* The greater the distance between a receiver and the transmitter, the worse transmissions are.

- *Season.* Leaves on trees in summer hinder transmission, although leafless trees in winter help.
- *Customer premise antenna height.* The lower the antenna height, the less effective transmissions are.
- *CPE antenna beamwidth.* The broader the beam, the more spread out the signals, and the less effective transmissions are.

There is no single silver bullet to achieve NLOS. A combination of technologies used together are required to overcome the above factors. They include advanced modulation technologies, smart antenna, frequency reuse, and layer-2 error correction algorithms, among others (Shtrom 2001; Iospan 2001).

10.4.2.1 Modulation Technologies One approach to achieving NLOS is the use of a new generation of wireless modulation techniques that

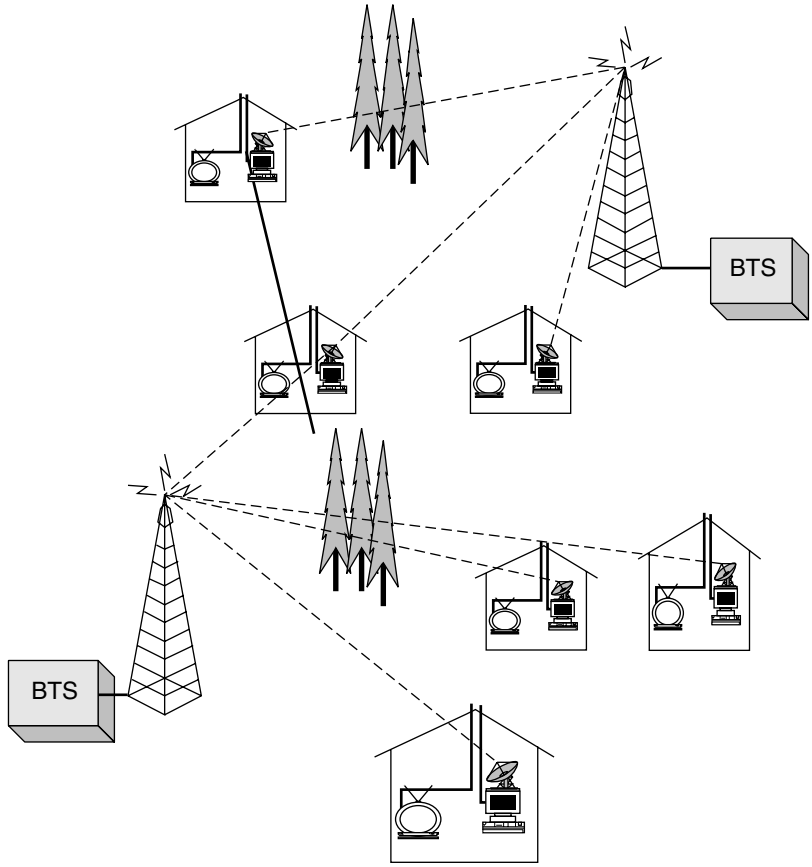


Figure 10-3
An example macro cell architecture in an NLOS BWA system.

aim to increase wireless bandwidth while enhancing security. Prominent among them is orthogonal frequency division multiplexing (OFDM).

OFDM is a modulation method that spreads digital information over a spectrum. Like CDMA, it uses the spread spectrum concept. However, OFDM uses spread spectrum in the frequency domain while CDMA uses it in the time domain. OFDM has been chosen as the transmission method for the European radio and TV standards. It is also being used for a new generation of wireless networks, xDSL, and a new generation of land-based transmission equipment.

The basic idea of OFDM is not very complicated: It uses multiple carriers (i.e., frequencies), which are called *subcarriers*, to multiplex the original data stream into multiple parallel data streams, each of which is modulated with a different frequency. The subcarriers have an orthogonal relationship to each other that enables the receiver to recover the received signals. At each particular frequency, the received signal is evaluated and recovered by the receiver, while all other signals are treated as zero and ignored.

There are a number of advantages associated with OFDM. For one, it provides diversity to prevent frequency-selective fading caused by multipath as described above. In addition, it averages the interference from neighboring cells by using different basic carrier permutations between users in different cells. Also, it enables higher data throughput via dense modulation schemes such as 64 QAM and 256 QAM and allows the use of smaller and more powerful CPE units with indoor omnidirectional antennas.

10.4.2.2 Smart Antennas "Smart antennas" are an emerging technology being actively pursued for use in NLOS broadband wireless access. The term refers to the use of multiple antennas at both transmitting and receiving ends and a set of associated techniques in wireless channel coding, modulation, and signal processing to enhance throughput and the data rate. Smart antenna techniques can be used for both downstream and upstream links and at both the BTS and the CPE antenna.

BWA is a good fit for the application of smart antenna techniques, foremost because BWA is fixed wireless and both the CPE antenna and modem are fixed in location for a BWA network. This enables the use of multiple antennas and associated techniques with CPE. For example, the CPE in a BWA network is line-powered (versus battery-powered as in mobile handsets), so that the additional power needed by a smart antenna is not a constraint.

Smart antennas have a set of associated techniques that increase the data rate and throughput of BWA systems by achieving high spectrum efficiency. Those techniques include the following:

Spatial multiplexing. Spatial division multiplexing uses multiple antennas at both ends of a wireless link to transmit multiple data streams. The original stream of digital symbols is split into multiple streams, one per antenna. These substreams are modulated and transmitted, all in the same radio channel, but at a lower rate. The streams are separated using the spatial signature of each transmit antenna. The receiver then merges the multiple streams to yield a high-rate data stream.

Spatial division multiplexing. Also called *channel reuse,* this technique exploits the beam-forming, directional antennas of wireless systems to support more than one user in the same frequency channel.

Channel estimation. Smart antennas need accurate channel knowledge to achieve interference avoidance. For example, on the receiving side, the channel knowledge can be directly obtained from the information (called *training sequence*) embedded by the transmitter.

10.4.2.3 Adaptive Antenna Array An adaptive antenna array is one of the key technologies used in BWA systems to overcome line-of-sight by expanding coverage and improving signal quality. An antenna array consists of N identical antenna elements arranged in a particular geometry so that the geometry of the array determines the amount of coverage in a given spatial region. For example, one widely used antenna array type is uniform linear array.

The key idea behind adaptive antenna array is that an antenna array of N elements can transmit or receive N beans simultaneously. The individual beams can be appropriately combined to increase throughput and reduce interference. This application of adaptive antenna array is also called *beam-forming. Adaptiveness* refers to the fact that the phases and amplitudes of the currents that excite each antenna element can be electronically adjusted to maximize the gain of the signals in a certain direction. Due to the reciprocal nature of antennas, the adaptive antenna array can be used at both the transmitter and receiver ends, and thus at both BWA BTS and CPE.

10.4.2.4 Forwarding Error Correction Advanced FEC schemes are used to overcome LOS. FEC can be achieved at the media access control layer to handle transmission error. Methods have been defined for error checking and error handling such as retransmission protocols in case of error to allow normal operations under degraded link conditions.

TABLE 10-2

Comparison Between NLOS BWA and MMDS

	NLOS BWA	MMDS
Modulation techniques	Wideband OFDM, OFDM, MIMO OFDM	64 QAM, 16 QAM, QPSK
BTS antenna	Antenna array	Single antenna
Deployment architectures	Macro cell in most cases that serves an area around 2–7 mi in radius	Macro or super cell in most cases that serve an area of up to 35 mi in radius.
Frequency bands	2–3 GHz, including MMDS and unlicensed bands	MMDS bands between 2–3 GHz
CPE equipment	Indoor antenna and smaller wireless modem	Outdoor antenna and small wireless modem
Multiple access technology	TDMA, CDMA	TDMA, CDMA
Data rate	Vary greatly from vendor to vendor, due to the proprietary nature	Max of 25–30 Mbps
Antenna restriction	Non-line-of-sight	Line-of-sight

MIMO: Multiple-input multiple-output.

10.4.3 Comparison Between BWA and MMDS systems

BWA is a fast-evolving technology that has the potential to become a viable alternative solution to the last-mile problem.

A comparison of BWA with MMDS, as listed in Table 10-2, summarizes the characteristics of NLOS BWA. In essence, NLOS BWA and MMDS share many of the same characteristics, such as modulation techniques and frequency band. The key differences lie in the use of smart antennas and multiplexing technologies by BWA systems to achieve high bandwidth and NLOS.

REVIEW QUESTIONS

1. Describe the frequency range of LMDS systems and their components. Discuss how an LMDS system can be plugged into a wireline backbone network.
2. Describe the portions of the LMDS system the DAVIC LMDS is intended to standardize and the target applications behind the DAVIC specifications.

3. Describe the frequency range of MMDS systems and explain why *MMDS* refers to a service rather than to a specific technology.
4. Compare LMDS and MMDS systems, describing the advantages and disadvantages of each.
5. Describe the main components of MMDS systems in terms of the base station and CPE.
6. Discuss the issues related to LOS and the motivations for overcoming LOS in a new generation of BWA systems.
7. Describe the main approaches for a BWA system to achieve a higher data rate than those of the existing MMDS and LMDS systems.
8. Describe the concept of adaptive antenna array. Discuss how it can be used to achieve higher bandwidth and as an interference cancellation method.
9. Describe the technologies employed in the new generation of BWA systems to achieve NLOS.

REFERENCES

Bates, R. 2000. "Broadband Telecommunications Handbook." New York: McGraw-Hill.

Greenstein, L., Ghassemzadeh, S., Erceg, V., and Michelson, D. 1999. "Ricean K-factors in narrowband fixed wireless channels." WPMC '99 Conference proceedings, Amsterdam. Sept.

Hybrid, Inc. 2000. "An introduction to fixed broadband wireless technology." White paper. Web site: www.hybrid.com.

IEEE. 2001. "Air Interface for Fixed Broadband Wireless Access Systems." IEEE 802.16. Web site: www.ieee.org.

Iospan Wireless, Inc. 2001. "Fixed broadband wireless access: state of the art, challenges and future directions." White paper. Web site: www.iospanwireless.com.

Shtrom, V. 2001. "Designing broadband wireless access systems for pure non-LOS environments." Iospan Wireless, Inc. White paper. Web site: www.iospanwireless.com.

Tipparaju, V. 1999. "Local multipoint distribution service (LMDS)." White paper. Web site: www.cis.ohio-state.edu.

CHAPTER 11

Wireless Personal Area Networks

11.1 Introduction

Wireless personal area networks (WPANs) are a type of access network. This section first defines WPANs, provides a brief historical background, and then discusses the characteristics of WPAN technology.

11.1.1 WPAN Defined

WPAN is a term defined in IEEE 802.15 to refer to the Bluetooth network (IEEE 2002). But in this chapter, the term is used to generically refer to a category of wireless access network of which Bluetooth is an important member. In effect, the focus of this chapter will be on Bluetooth. WPAN is a short-range wireless network that can be considered "the access network of access networks." It covers very short distances and connects personal computing devices such as PDAs, computers, and printers, among others. It is at the very fringe of the network hierarchy target to provide a "last-yard" solution and needs to connect to traditional access networks like LANs or residential access networks.

Technology-wise, WPAN is not that much different from wireless LAN described in Chap. 9, overlapping in some areas while being complementary in others. They both use radio as their transmission medium and operate in the same unlicensed ISM band (2.4 GHz). The major differences between them lie in their network coverage ranges and target applications. A WPAN normally operates in a very small, confined area like a home, a small office, or a warehouse. It addresses not the last-mile issue of the typical access network, but the "last-yard" issue of connecting user devices. Its target applications tend to focus on replacing cable wiring and connecting mobile devices without a central control.

In summary, WPAN in general tends to have the following characteristics:

- A mobile operating environment
- Problems with interference from other appliances
- Peer-to-peer communication and configuration rather than a configuration with centralized control

In addition to Bluetooth, this chapter also discusses two other similar WPAN technologies that target the similar applications: HomeRF and Digital Enhanced Cordless Telecommunications (DECT).

11.1.2 Main Issues of WPAN

WPAN technologies focus on the physical and media access control layer of the OSI network reference model. At the physical layer, both Bluetooth and HomeRF (although not DECT) use the unlicensed 2.4-GHz frequency band.

One important issue for the WPAN physical layer is dealing with heavy interference from appliances like cordless phones and microwave appliances that operate in the same part of the frequency spectrum. For this reason, WPAN technologies tend to use the frequency hopping spread spectrum as their transmission method because frequency hopping at a high hopping rate can avoid the channels that produce heavy interference.

One target application of WPANs is to connect all communications devices like desktop personal computers (PCs), laptop portables, cordless phones, and next-generation intelligent home appliances that involve both data and voice interfacing. This is handled by the MAC layer, which is responsible for allocating radio resources and controlling access to shared resources, using different MAC technologies, for example, TDMA for voice and CSMA/CA for data.

The WPAN configuration is peer-to-peer centric with a dynamically changing network configuration in a small group, called the *piconet* in Bluetooth. Thus the system design is centered around the ad hoc type of configuration and peer-to-peer communications, and deals with issues such as clock synchronization, link setup, multiplexing scheme, and power conservation scheme.

Another focus of WPAN is security. Given that devices are constantly coming into and leaving a piconet, the security mechanism is designed to accommodate the mobility of the user devices and the needs of peer-to-peer authentication and authorization.

11.1.3 WPAN Application Examples

Some examples of applications will help further describe WPAN. Some specific applications that have been mentioned include the following:

- *Briefcase trick.* A laptop computer can connect to a network as the owner walks into a hotel or conference center while the laptop is still in its owner's briefcase, since a radio link does not require line of sight. When taken out of the briefcase, the laptop is already connected to the Internet.

- *Data transfer between devices.* Seamless connectivity between user devices like home PCs, laptops, PDAs, and mobile phones is desirable. There are numerous possibilities for innovative applications here. Simple examples start with the automatic synchronization of schedules between a laptop and a PDA. The automatic downloading and synchronization of a contact list between a mobile phone and a PC is another example.
- *Mobile display.* PC monitors can be made mobile and positioned wherever it is convenient for them to be while they are connected to base units via radio linkage.
- *Wireless headset or hand-free devices.* These allow for hands-free access to mobile phones, audio devices, and other services.

11.2 Bluetooth Architecture and Protocol Stack

Bluetooth has the potential of becoming a widely deployed WPAN technology. As already noted, it is used synonymously with IEEE 802.15 WPAN throughout this chapter. This section provides an overview of Bluetooth architecture.

11.2.1 Brief History of Bluetooth

The Bluetooth concept, named after Harald Blaatand "Bluetooth" II, king of Denmark during the years 940–981, was first developed in 1994 by the Ericsson corporation of Sweden. The goal of Bluetooth technology is to create a standard, universally accepted wireless network that uses short-range radio links to connect portable and fixed electronic devices, replacing cable wiring. The devices covered under WPAN, also called *Bluetooth devices,* include laptop computers, desktop computers, printers, mobile phones, and appliances located no more than a few meters apart.

A group of influential companies that included IBM, Ericsson, Intel, Nokia, and Toshiba formed a Special Interest Group (SIG) in February 1998 to begin developing the open specifications for such a short-range wireless network. The Bluetooth 1.0 specifications became available in 1999 (Bluetooth SIG 1999).

The IEEE 802.15 Study Group (SG) was formed in March 1999 with the goal of developing the physical layer and wireless medium access control specifications of WPAN. The only proposal submitted and accepted was by the Bluetooth SIG. So IEEE 802.15 adopted the Bluetooth SIG Bluetooth 1.0 specifications as the basis for its WPAN specifications, with some cosmetic enhancement. Since then, Bluetooth SIG standardization efforts have been merged into the IEEE 802.15 SG efforts.

11.2.2 Protocol Stack of WPAN and Architecture Overview

WPAN covers the bottom two layers of the OSI network reference model, i.e., the physical layer and the data link layer. The mapping between the WPAN protocol stack and the OSI network reference model is shown in Fig. 11-1. The WPAN protocol stack consists of four layers: radio, baseband, link management protocol (LMP), and host control protocol [or Logical Control and Adaptation Protocol (L2CAP)] (Kansal 2001; Bray and Sturman 2000).

The radio layer, corresponding to the physical layer, deals with wireless transmission medium issues like frequency spectrum, modulation scheme, encoding method, etc. The remaining layers deal with the data link layer functions, including medium access control, transmission error detection and correction, and security. Specifically, the baseband and link management protocol layers are responsible for establishing and controlling links between Bluetooth devices. The bottom three layers generally are implemented in the hardware and firmware. The host control is present only when the L2CAP resides in the software.

11.2.3 Wireless Medium—Radio Physical Layer

WPAN or Bluetooth devices operate in the unlicensed 2.4-GHz ISM band, the same band where IEEE 802.11b operates. Because this frequency band is crowded and there can be strong interference from other appliances in the home or small office environment using it, a fast FHSS scheme and short data packets are used to minimize the interference. The number of frequency hops for FHSS is specified at $2402 + k$ MHz, where k can be $0, 1, \ldots, 78$. The nominal hop rate is 1600 hops/s.

The modulation scheme adopted by Bluetooth is Gaussian prefiltered binary phase shift key (PSK). The specified operating power level for

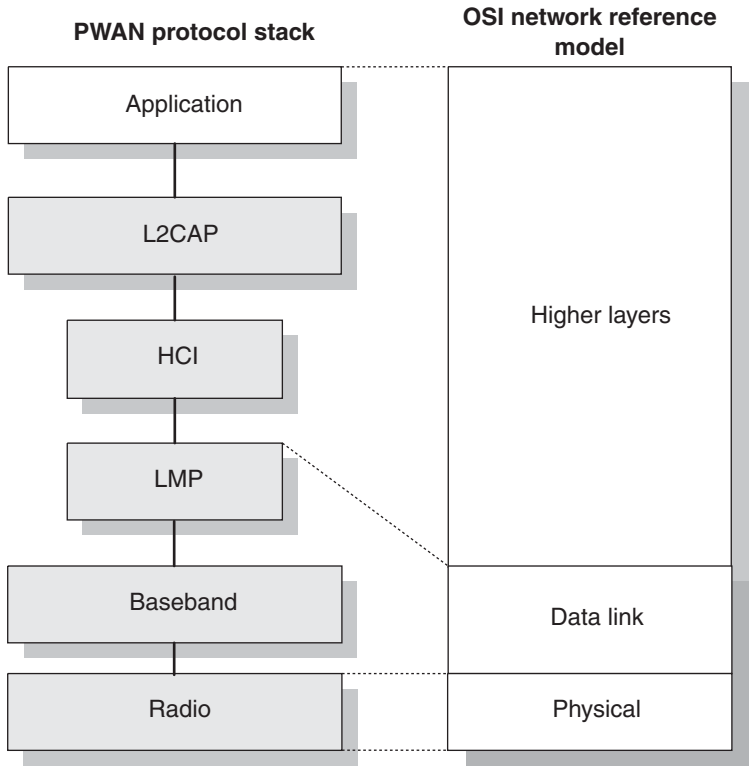


Figure 11-1
The Bluetooth protocol stack.

Bluetooth devices is dependent on the transmission distance. For a distance of 10 m, the transmission power is limited to 10 dBm and for a distance of 100 m to 20 dBm.

11.2.4 Baseband Layer

The baseband layer performs some of the data link layer functions, and its primary responsibility is controlling the radio links. Specifically it performs the following functions:

- *Provisioning of frequency hop sequence.*
- *Clock synchronization of Bluetooth devices.*
- *Packet forming on the wireless link.*
- *Radio connection establishment.* The WPAN specifications provide for two types of connections: (1) synchronous connection-oriented links, used for transfer of synchronous data like voice and stream data, and (2) asynchronous connectionless links, typically used for transfer of asynchronous data such as files and email.

- *Device address discovery.* An inquiry procedure is defined for a device to discover the addresses of other devices in proximity.
- *Encryption key and link key generation for security purpose.*
- *Error correction for packets.* The error correction method to be used depends on the type of packets. The packet types differ in their data capacity and error correction overhead.

The Bluetooth standards specify five types of channels for carrying different types of data: control information, link management information, user synchronous data, user asynchronous data, and isosyschronous data.

11.2.5 Link Manager Protocol

The Bluetooth protocols specify an LMP mainly for medium access control and radio link management. Specifically, it performs the following three main functions:

- Link configuration
- Piconet management
- Data link layer security management

The LMP defines a procedure to attach and detach Bluetooth devices from Bluetooth networks, which are very small networks known as *piconets.* The LMP is responsible for establishing asynchronous connectionless and synchronous connection-oriented links. In addition, it controls the procedure for a device to transition from one power mode to another mainly for power saving. In addition to the active state, a device can be in one of the low-power states called *hold, sniff,* and *park* (discussed in Sec. 11.3.2.3).

11.2.6 Logical Link Control and Adaptation Protocol

L2CAP is a protocol that provides interface for and services to higher-layer applications and to the baseband and link management protocol of the lower layer. Specifically, L2CAP performs the following three functions:

Multiplexing. This allows multiple applications to use a link between two devices simultaneously.

Packet segmentation and reassembly. This fits the packets from the application layer into the packets of the baseband layer by breaking large

packets into smaller ones or vice versa, depending on the direction of traffic flow. L2CAP can accept packet sizes of up 64 K, while the baseband packet payload size can be no larger than 2745 bits.

QoS. The L2CAP layer also allows applications to set parameters such as link peak bandwidth, latency, and delay variation. L2CAP provides the QoS depending on the link conditions.

11.2.7 Host Controller Interface

A host controller interface (HCI) interconnects a Bluetooth hardware module and a host machine like a PC. For many devices, the Bluetooth enabling module may come in the form of a hardware add-on card—for example, as a PCI card or universal service bus (USB) adapter for host devices like PCs or laptops. The add-on cards generally implement the lower layers including radio, baseband, and LMP. In order for the host device and the Bluetooth add-on hardware module to communicate over a physical bus like USB, an HCI driver is required on the host device and a host controller interface is required on the Bluetooth hardware card.

The HCI driver on a host device is responsible for formatting the data from the application on the host to be accepted by the HCI on the Bluetooth hardware module. The HCI on the Bluetooth hardware module can receive the data from the host device to be sent out over a radio link to the Bluetooth hardware device on the target device.

11.2.8 Application Layer

The application layer in the Bluetooth architecture refers to the applications that can directly communicate with L2CAP and that may run over TCP/IP or WAP (Wireless Application Protocol). Also the applications may run PPP for a dial-up connection to the Internet, FTP for file transfer, or other application-specific protocols.

11.3 Bluetooth Configuration and Operation

A Bluetooth network is unique in its configuration and operations.

11.3.1 Configuration and Components of a WPAN System

A Bluetooth network is always configured as an ad hoc configuration where there is no central control point. It features peer-to-peer connections between devices. In contrast, traditional wireless networks such as Personal Communications Service (PCS) and Global System for Mobile Communications (GSM) have a central control point and work in a hierarchical fashion—for example, a set of base station transceivers is controlled by a base station while a group of base stations is controlled by a mobile switching center (MSC).

A Bluetooth network has two types of configuration: piconet and scatternet. A piconet (the prefix *pico* means "very small") consists of a set of Bluetooth devices connected to the same channel that is identified by its unique hop sequence, as shown in Fig. 11-2. One device serves as the master, which is usually the device that initiated the connection. Up to seven other devices can be connected to the master device in the active state within a piconet, and many more can be connected in the low-power parked state. The master device is responsible for controlling channel sharing by communicating to a line manager at each device via the LMP.

A scatternet is made up of a set of piconets. It is a group of independent piconet networks in the same space but communicating over different channels. As shown in Fig. 11-3, there is no coordination between the piconets in the same scatternet.

11.3.2 Bluetooth System Operations

The normal operations of a Bluetooth system consist of two main phases: initial connection setup phase and data transmission phase (Miller 2001).

11.3.2.1 Initial Connection Setup Assume that a businessman walks into a hotel, and the Bluetooth-enabled laptop in his briefcase automatically connects to the Bluetooth network in the hotel to check his email.

INQUIRY AND INQUIRY RESPONSE The laptop automatically initiates an inquiry to find out an access point within its range. The inquiry message

can be addressed either to a generic address called the general inquiry access code (GIAC) or to a service-specific address called the dedicated inquiry access code (DIAC) to access services such as printers, email, etc. The inquiry message is repeated at 16 frequencies, which is known as the *inquiry hop sequence* or *train*.

A device that is willing to respond to an inquiry enters inquiry scan to listen to the hop frequency of its choice. When an inquiry message is received, an inquiry response message containing the responding device's address is sent to the inquiring device. Since more than one device may respond, the response messages are sent by the responding devices at intervals of a randomly chosen number of seconds, which minimizes the chance of collisions caused by multiple devices sending responses simultaneously.

PAGING AND PAGING RESPONSE The inquiring device—the businessman's laptop computer in this case—does not acknowledge receipt of the response messages. Instead, it decides on a responding device to

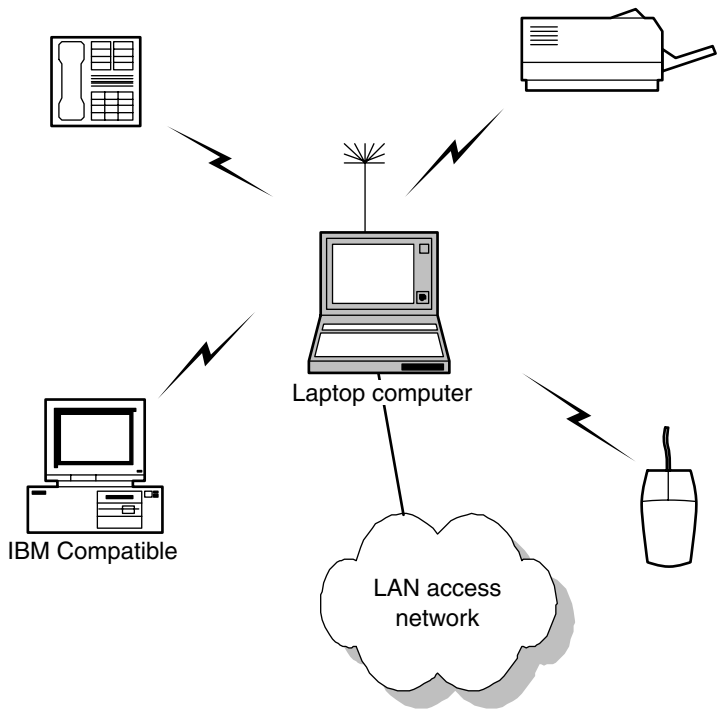


Figure 11-2
A Bluetooth piconet example.

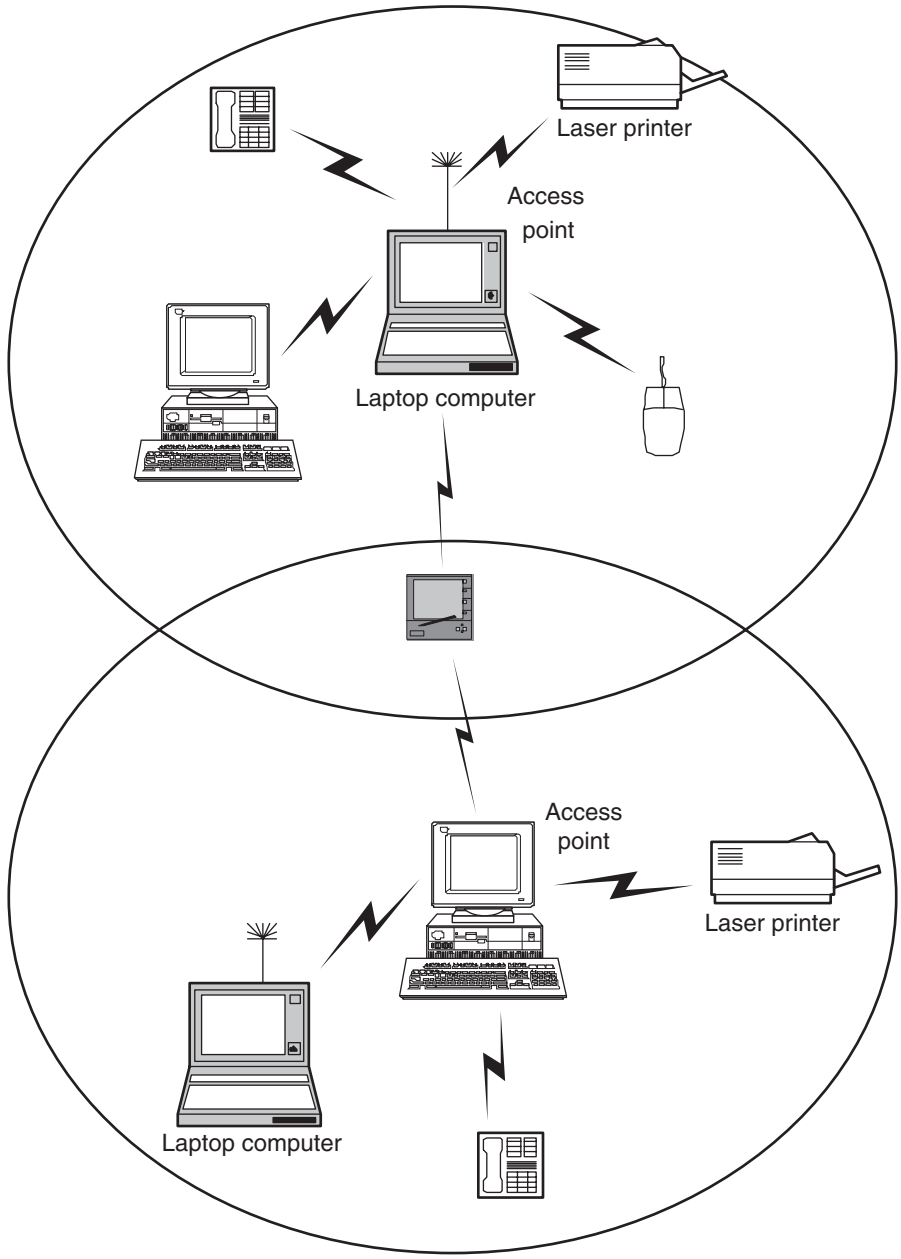


Figure 11-3
Illustration of Bluetooth scatternet concept.

connect to. It uses the clocking and frequency hopping sequence information included as part of the inquiry response packet to initiate a page request to start a procedure known as *paging* to synchronize its clock, frequency hop phase, and other parameters with those of the responding device.

The device that initiates the paging process serves at first as the piconet master. That device periodically enters into page scan to let other devices connect to it. When a connection is established, the paging device becomes the master of this connection. Then the slave device sends a request to the new device to switch their master/slave roles to let the new device join the piconet.

This piconet master device periodically pages all devices in the same piconet, the paging being triggered either by a new device entering the piconet, an old link becoming unavailable, or the passing of a certain amount of time. The master device sends a page message to all the slave devices on the piconet over the page hop sequence. The page hop sequence consists of 32 frequencies, divided into two trains with 16 frequencies each. The master device first tries train A, which is an estimate of the frequency hopping sequence that the scan device listens to, leaving a number of time slots for each slave device to respond. If train A fails to provide responses from all slave devices, the master device tries train B.

Each slave device, when in the page scan state, listens for a page message addressed to a designated address known as the device access code (DAC). Once it receives the page message, it sends back a page response message containing its ID packet with its DAC and the other configuration parameters. Upon receiving a response message, the master device sends its frequency hopping sequence (FHS) packet message to the slave. Included in the page message is the timing information of the master clock and an active ID number of the slave device assigned by the master. The slave device uses the FHS message to determine the channel access code for the piconet it has just entered. It then calculates its clock offset based on the master clock that will be used for communicating within this piconet. The master then sends a POLL message to the slave and the slave device can respond with any message to the POLL message to conclude the paging procedure, as well as the initial connection setup.

With this paging process completed, the laptop computer in the briefcase is now connected to a piconet of the hotel, and is ready to enter the data transmission phase of operation.

11.3.2.2 Data Transmission The previous step has connected the laptop to a hotel piconet. To check email or use any other service, the laptop must then establish a link or connection to another device that runs an application server in order to use the desired service. The three basic steps for accomplishing data transmission include connection setup, service discovery, and link setup and data transmission.

CONNECTION ESTABLISHMENT Say the laptop is checking for email. It must first establish a connection at the data link layer between itself and a remote device that runs an email server. This is accomplished via the exchange of Link Management Protocol (LMP) messages in a three-step procedure:

1. The local device sends an LMP host connect request message to other devices on the same piconet in a peer-to-peer fashion using the baseband packets. The request message contains the information on the type of service the local device is interested in.
2. The responding device that intends to provide the service sends an LMP accepted message to the requesting device. A responding device may decide against responding positively due to the lack of the requested resource or service.
3. The local device chooses the first responding device with the appropriate response. A layer-2 connection is established between the two devices.

SERVICE DISCOVERY Due to the dynamic nature of the ad hoc configuration of Bluetooth piconets, the inquiring device must go through a service discovery process to find the service needed, in this case email. The layer-2 connection established in the previous step can be used for the service discovery purpose.

The Bluetooth specifications define a Service Discovery Protocol (SDP) for this purpose. A device willing to have its application service discovered by another device must run an SDP server, and a device that intends to discover and use the service provided by another device must run an SDP client, one client per service. One device can run only one SDP server, and multiple SDP clients can run simultaneously on a device.

LINK SETUP AND DATA TRANSMIT Once the previous two steps have established a connection with a remote device and discovered the desired service on the device, access to the service is achieved via an L2CAP layer channel, or a *link* in Bluetooth terminology.

An application that needs a synchronous connection-oriented (SCO) link uses a baseband connection to carry the data. An application that needs an asynchronous connectionless (ACL) link uses L2CAP. An L2CAP channel is equivalent to a network layer connection or a transport layer session. Such a channel is set up via a two-way handshake procedure: A local device sends a connection request, and the remote device responds with a connect-request response message. In addition, L2CAP also allows the two devices to exchange service parameters such as maximum payload, device type, timeout limit, and QoS parameters, among others.

L2CAP is designed to be "thin" and efficient. It does not have any data error check capacity, relying instead on the baseband layer for data security, integrity, and in-order delivery of packets. L2CAP interfaces with the protocol and applications of higher layers to convert application packets into lower-layer format and vice versa in the other direction.

Once an L2CAP channel is set up, the local device can start transmitting data. In this case, the businessman can check his email messages. An L2CAP channel is per transaction-based, and once the local device is finished with the service, it uses a two-way handshake procedure to terminate the channel.

The Bluetooth specifications define two additional high-level protocols to help run applications over the Bluetooth stack and help L2CAP interface protocols like PPP, TCP/IP, and WAP. *RFCOMM* is an emulation of the serial port over wireless links to support the legacy serial port-based devices, and the *Telephony Control Specification* (TCS), defines the call control and signaling for voice service in a Bluetooth network. The name *RFCOMM* is derived from Radio Frequency Comm serial port.

11.3.2.3 Operation Modes A Bluetooth device can be in one of the following mutually exclusive operation modes:

Active mode. A Bluetooth device must be in active mode in order to transmit data. In this mode, master and slaves transmit data in alternating slots. The master transmits in even-numbered slots and slaves use odd-numbered slots. This is the most power-intensive operation mode. The Bluetooth specifications define certain optimization mechanisms for power saving, one of which is to let a slave device go to sleep until it is informed by the master to transmit data.

Sniff mode. This is a low-power mode in which the activities of a listening device are reduced. The device listens for transmissions only at fixed intervals decided by the LMP of the master device.

Hold mode. In this mode, a slave device suspends its activities on the ACL link, while the synchronous connection-oriented links may still be supported. Other activities such as scanning, paging, inquiring, and attending another piconet are still carried out. After a fixed duration negotiated between the slave device and the master device, the slave device comes out of hold mode to become active or to enter another operation mode.

Park mode. This is a very low-power mode in which a slave device engages in few activities. It gives up its active member address (assigned to it when it first became a member of the piconet) in exchange for a parked member address. In this mode, the slave device stays synchronized with the master device on the clock via a beacon channel. The slave device only monitors the broadcast messages sent over the beacon channel at fixed intervals decided by the master device. In addition to conserving power on the slave device, the park mode allows the master device to have more than seven slave devices on the piconet (the active member address is a 3-bit number). A slave device in hold mode can request to "wake up" or to be awakened by the master device.

11.4 Bluetooth Security

Bluetooth security centers on link layer security with a focus on authentication. The security considerations for Bluetooth networks include support for the peer-to-peer and omnipresent operating environment (Bray and Sturman 2000).

11.4.1 Bluetooth Security Architecture

Bluetooth mainly relies on link level security mechanisms to achieve secure communications. Beyond this, the higher application layer can further enhance security via its own security mechanisms.

The FHSS transmission scheme provides a level of security in itself. Instead of transmitting data over one single frequency, fast frequency hopping requires that the receiver be well synchronized to access the transmitted data.

Bluetooth defines two levels of security for devices and service, and each device is classified into one of them upon first entering a piconet:

- Trusted
- Untrusted

A trusted device has unrestricted access to other devices and services. A device has to be authenticated before it becomes a trusted device. Upon entering the network for the first time, a device is considered unknown and therefore is labeled untrusted. An untrusted device has no fixed relationship to the piconet, and its access to services is limited.

Bluetooth defines three modes where security is enforced:

- *Security mode 1:* No security is enforced; this is a nonsecure mode.
- *Security mode 2:* Security is enforced at the service level. In this mode, a device does not enforce security or initiate security procedures before a channel is set up at the L2CAP layer. This mode allows security be enforced at the higher application layer with a different, flexible security policy.
- *Security mode 3:* Security is enforced at the link level. In this mode, a device initiates security procedures before link setup on the LMP layer.

Bluetooth focuses on the security mode 3, where security is enforced at the link level. The security procedures include authentication, authorization, and encryption. The building blocks of the Bluetooth security mechanism include a set of keys and the procedures used for authentication and authorization.

11.4.2 Bluetooth Security Parameters

The Bluetooth security mechanisms include the following set of keys and security procedures for implementing authorization, authentication, and encryption:

Device address. Each Bluetooth device has a unique public address (BD_ADDR) that serves as an ID for the device and is part of the authentication scheme.

Personal Identification Number (PIN). This allows a user to log onto a device and can be fixed or selected by the user. The fixed PIN code allows

for automatic login, which can be useful in certain situation like at home. The length of a PIN can be anything between 1 to 16 octets, though generally it is four to six digits.

Link keys. Bluetooth defines four types of link keys that are used for different purposes. All link keys are 128 bits long, generated with either a E21 or E22 algorithm. The input to the key generator is a 128-bit random number that ensures the randomness of the keys. The four types are

- *Unit keys.* These are generated at the device installation time, stored in the system memory, and rarely changed.
- *Combination keys.* These are derived from two communicating device's addresses and a 128-bit random number.
- *Initialization keys.* These are generated when two devices attempt to communicate with each other for the very first time during the link initialization process, and are derived from each device's PIN code and address plus a 128-bit random number.
- *Master keys.* These are temporary link keys generated by the device serving as the master in the piconet. A master key is used for the master device to communicate with the slave devices on the same piconet. It is derived from a 128-bit random number and the current link key.

Secret keys. There are also two different secret keys used for authentication and encryption. An *authentication key* is generated at the time of system initialization. An *encryption key* is derived from the current link key, a 96-bit ciphering offset number (COF), and a 128-bit random number. It is changed each time encryption is applied.

11.4.3 Link Security Initialization Procedure

Bluetooth enforces link level security. The following steps are carried out before a layer-2 link is established between two devices:

- An initialization key is generated.
- An authentication procedure is carried out, as described below.
- A link key is generated for the link to be established at both devices.
- The link key is exchanged between the two devices.
- An encryption key is generated at each unit for data transmission.

11.4.4 Authentication Process

The authentication procedure is based on a challenge-response scheme. When a device is accessed to provide a service, the accessed device, which is referred to as the *verifier,* issues a challenge to the requesting device, which is referred to as the *claimant.* The claimant then sends a response to the challenge. The response contains the unique device address and the link key shared between the two devices. After authentication, encryption may be used to encrypt the data before transmission.

The challenge-response scheme checks the claimant's knowledge of a secret key through a two-move protocol using symmetric secret keys. At a high level, it works as follows:

Step 1. The verifier sends a random number along with an authentication code as part of the challenge.

Step 2. The device to be authenticated calculates a response value, using its secret key, its unique device address number, and the received random number, and then sends the response to the claimant.

Step 3. The verifier also calculates the response value based on the secret key, the device address, and the generated random number. The response value matching that calculated by the verifier indicates the knowledge of the secret key on the claimant's part.

The challenge-response scheme is said to be symmetric because both sides know the secret key. The scheme described above works one-way. However, a bidirectional authentication may be preferred in some peer-to-peer communications where two devices provide service to each other.

11.4.5 Security Challenges for Bluetooth Systems

Overall, Bluetooth's level of security is considered by many experts to be sufficient for small-scale applications. However, some challenges remain for the Bluetooth systems:

- *Access control at the link setup time only.* The access check is asymmetric but data flow is often bidirectional. Access to service in the direction that is not authenticated may expose the system's resources to unintended parties.

- *The static nature of the way some link keys are generated as described in Sec. 11.4.2.* One example is the fixed device address that is used as an input to link key generation.

11.5 HomeRF

There are two alternative WPAN technologies known as *HomeRF* and *Digital Enhanced Cordless Telecommunications*. Like Bluetooth, they target short-range wireless applications like residential home networking.

11.5.1 HomeRF Overview

Home networking with mobility became a very attractive application as PCs penetrated deeply into the home market and Internet connections became common in the late 1990s. To address this need, a group of information technology (IT) and computer industry vendors like IBM, HP, Compaq, and Intel formed the HomeRF Working Group in 1997, with the goal of producing interoperable wireless voice and data networking within the home environment affordable for the massive consumer market (HomeRF 2001).

HomeRF targets next-generation, innovative applications that take advantage of the increasingly intelligent devices in the home environment. Some examples include

- *Enhanced PC combined with phone service.* Caller ID from a cordless phone can be sent to a PC via a HomeRF wireless link to look up more information on the caller from a database on the PC. The PC can translate a voice sentence like "Call Joe" into a cell phone number.
- *Collaborative working environment at home.* The HomeRF wireless link enables file transfer between home PCs, sharing of a printer among multiple home PCs, and multiplayer game playing, among others.
- *Mobile display appliance.* The HomeRF wireless link allows for the mobile display of information available on a PC database while the PC base unit itself remains at a fixed location. Examples would be the display of a cooking recipe in the kitchen or at a dining table or reference checking via the Web at a group study session.

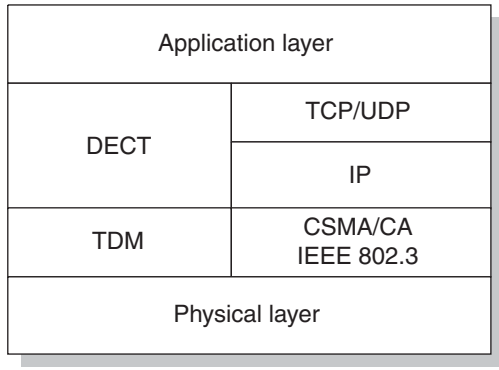


Figure 11-4
HomeRF network stack. (Xylink, 2001.)

Like wireless LAN and Bluetooth architecture, HomeRF addresses the bottom two layers of the OSI network reference model—the physical and data link layers. As shown in Fig. 11-4, the uniform physical layer supports multiple services (which is different from DECT architecture, as will be explained in Sec. 11.6). To support both data and voice services, two distinct MAC methods have been adopted for media access control and radio resource allocation. Voice service makes use of the DECT standard, while data service is based on the CSMA scheme used in Ethernet.

11.5.2 HomeRF Physical Layer

The HomeRF physical layer, like that of Bluetooth and IEEE 802.11b, operates in the unlicensed 2.4-GHz frequency band to support data, voice, and multimedia applications. The wireless transmission scheme it uses is FHSS at a hopping rate of 50 to 100 hops/s.

The initial HomeRF specification targets a 1.6-Mbps raw data rate, with a typical coverage range of 150 ft indoors. The second version of HomeRF specifications, completed in late 2001, targets a raw data rate of up to 10 Mbps, while the next version will target 20 Mbps (HomeRF 2001).

11.5.3 HomeRF MAC Layer

The MAC layer of HomeRF, as indicated in Fig. 11-4, supports both voice and data services with two distinct components: the data part and the voice part. HomeRF uses a variant of the CSMA scheme found in Ethernet: CSMA/CA, the same scheme used in the IEEE 802.11b MAC

layer. CSMA/CA supports data and multimedia services while TDMA is used for delivery of voice traffic.

11.5.4 HomeRF Network Configuration and Operations

A HomeRF system consists of networked home PCs and home mobile computing devices like laptops and PDAs, and can support up to 127 nodes. The HomeRF system components can be classified into four types of digital devices that operate in the home network environment:

- *Connection point (CP).* A CP can be a separate device like a home hub or an integral part of a home PC. A connection point interconnects multiple devices like a phone, a PC, and a networked appliance.
- *Isochronous node (I node).* An I node is a voice-centric digital device such as a cordless phone.
- *Asynchronous node (A node).* An A node is a data-centric device such as a PC, a laptop, or a personal digital assistant (PDA).
- *Combined asynchronous-isochronous node.* This is a device that has both voice and data capabilities, like an IP phone.

A HomeRF network like the IEEE 802.11 network supports both central-controlled and peer-to-peer ad hoc configurations. In a centrally controlled configuration, a connection point functions as the network server, the interface to the outside network, and the control point for the connected devices, very much like a typical LAN configuration.

In a peer-to-peer ad hoc configuration, a HomeRF network behaves like other ad hoc configured wireless networks such as Bluetooth. Outside devices go through registration with a local wireless network when first entering it, discover the services offered by other devices on the network, and establish peer-to-peer connections to utilize those services.

HomeRF uses a well-established shared-key encryption algorithm for data privacy and authentication.

11.6 DECT

Digital Enhanced Cordless Telecommunications is a European digital radio access network standard for wireless communications in environments like the home or a small office.

11.6.1 DECT Overview

DECT originated in the European standardization efforts relating to digital cordless phones in the late 1980s. In the mid-1990s, DECT shifted it focus to become a wireless access technology under the auspices of ETSI. The emphasis was on making DECT an access technology integrated with the GSM network. Starting in the late 1990s, DECT was fitted with data capability to support fast-growing Internet applications. DECT is dubbed as a "versatile technology" because of its long history and the many facets of its technology and standards. This section focuses on the latest incarnation of the DECT specifications.

The DECT standards, like those of HomeRF and IEEE 802.11, focus on the bottom two layers of the OSI network reference model—the physical layer and the media access control layer—as shown in Fig. 11-5.

A DECT network supports only the central-controlled configuration. A typical DECT network consists of a base station and a set of mobile terminals connected to and controlled by a base station. Devices like a LAN bridge with a DECT air interface can serve as the base station. The base station also has an interworking unit (IWU), as shown in Fig. 11-5, which is responsible for translating data link layer frame formats between the two types of networks supporting the data and voice services, i.e., it is responsible for converting frame formats from DECT DLC to CSMA/CA (DECT Forum 1997).

11.6.2 DECT Physical Layer

The physical layer of DECT operates in the licensed 1.88- to 1.9-GHz frequency band (the same frequency band used in the GSM network), and has 10 carriers or wireless channels at its disposal. Each carrier contains TDM frames of 24 time slots that in turn provide 12 duplex channels with 12 channels for sending and 12 channels for receiving. A data rate of 552 Kbps is achieved by combining 23 time slots in a unidirectional data transfer.

11.6.3 DECT MAC Layer

The DECT MAC layer is responsible for the effective allocation of physical resources, the multiplexing of multiple users, and the signaling of data onto shared physical resources. DECT devices use the time division

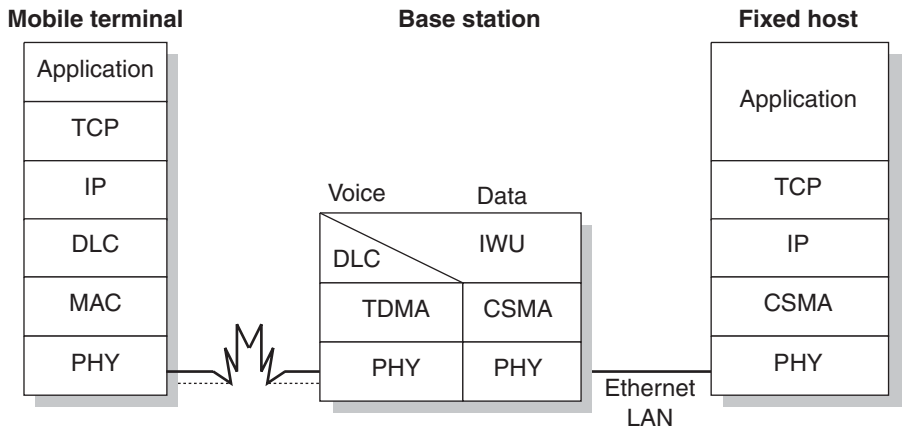


Figure 11-5
The protocol stack of DECT network. (DECT Forum 1997.)

multiple access (TDMA) control scheme at their MAC layer and can support a large number of devices connected into a network. In addition, DECT defines a dynamic channel selection/dynamic channel allocation (DCS/DCA) scheme to guarantee that the best-quality channel is selected for use at any given moment.

The DECT MAC layer also supports data service by adopting the same CSMA/CA method found in IEEE 802.11 and HomeRF for media access control. In addition, the DECT specifications include a data link control sublayer responsible for flow control, frame sequencing and segmentation, and reassembly of IP packets.

REVIEW QUESTIONS

1. Describe the relationship between the Bluetooth specifications and the IEEE 802.15 WPAN specifications.
2. Describe the main functions of the radio and baseband layers of a Bluetooth device.
3. Describe the functions of the Link Management Protocol and compare it against the equivalent layer of the OSI network reference model.
4. Describe the medium access control and modulation technique used in Bluetooth devices.
5. Explain what a piconet is and what a scatternet is, and the relationship between the two. Discuss the reasons for using a scatternet configuration for a WPAN system.

: **6.** Describe the paging process in the initial connection setup when a Bluetooth device first enters a piconet.

: **7.** In which mode of operation does a Bluetooth device use the least amount of power? In addition to power saving, what is the other advantage of the park mode of operation?

: **8.** Describe the security framework of Bluetooth networks, along with their authentication and encryption schemes.

: **9.** Describe briefly HomeRF technology and compare it with IEEE 802.15/Bluetooth technology in terms of the physical and the MAC layers.

10. Describe briefly DEC technology and compare it with IEEE 802.15/Bluetooth technology.

REFERENCES

Bluetooth SIG. 1999. "Specification of Bluetooth System. Version 1.0." Web site: www.bluetooth.com.

Bray, J., and Sturman, C. 2000. *Bluetooth: Connect Without Cables.* Englewood Cliffs, NJ: Prentice Hall PTR.

DECT Forum. 1997. "DECT—the Standard Explained." White paper. Web site: www.dectweb.com.

HomeRF Working Group. 2001. "Wireless Networking Choices for the Broadband Internet Home." White paper. Web site: www.homerf.org.

IEEE. 2002. "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Personal Area Networks (WPANs.)" IEEE 802.15.1. Web site: www.ieee.org.

Kansal, A. 2001. "Bluetooth tutorial." On-line tutorial. Web site: www.ee.iitb.ernet.in/uma/.

Miller, M. 2001. *Discovering Bluetooth.* New York: McGraw-Hill.

CHAPTER 12

Infrared Communications and Free Space Optics

12.1 Introduction

The term *free space optics* (FSO) refers to the technology that delivers high-speed data services using infrared light to transmit optical signals through free space. Other terms have also been used to refer to the same technology such as *wireless optics* and *fiberless optics*.

12.1.1 Fundamentals of Infrared Communications

The technology of laser beams being transmitted through free space has not been used for practical communication applications until very recently. In recent years, with the advances in fiber optical technology and strong market demand for broadband access solutions, interest in infrared-based communications has resurged and FSO wireless networks are being explored as an alternative last-mile broadband access technology. The practical applications are still at an early stage of development.

The laser beam used for FSO wireless networks is known as *infrared radiation* (IR) because it uses a specific region of the electromagnetic radiation spectrum that has wavelengths longer than those of visible light but shorter than those of radio waves. IR frequencies are higher than those of microwave but lower than those of visible light.

In terms of wavelength, the IR spectrum is divided into three regions. The wavelengths are measured in nanometer (nm) where 1 nm = 10^{-9} m. The first region is the *near IR band*, which consists of wavelengths from approximately 750 to 1300 nm. The *middle IR region* consists of wavelengths from 1300 to 3000 nm, while the *far IR region* refers to wavelengths from 3000 to 14,000 nm. FSO wireless networks in general use the near IR region and lower part of the middle IR region, or wavelengths ranging from 750 to 1500 nm.

The principle of FSO communications systems is the same as that of fiber optical systems: digital information is conveyed from point A to point B with the flashing of infrared light. The key difference is that the optical signals (light flashes) are transmitted in the atmosphere or in free space in an FSO system while they are carried over a fiber cable in a fiber optical system.

12.1.2 Three Infrared Communications Systems

This chapter covers three types of infrared communications systems or standards out on the market: IEEE 802.11 infrared-based LANs, free

space optics wireless networks, and IrDA (which stands for *Infrared Data Association*.

Infrared LAN is like IEEE 802.11a or IEEE 802.11b LAN, the only difference being that the transmission medium is infrared light rather than radio at the physical layer (IEEE 1999). The MAC layer is the same as in IEEE 802.11 LAN. There are very few products of this type out on the market. The term *FSO wireless networks* refers to the high-speed data networks that are being currently developed or have been deployed in recent years as an alternative last-mile broadband access solution. IrDA is an infrared-based communications system between appliances and computers for very short distances. It is widely used in appliances such as video cameras and computers.

12.2 Infrared Wireless LAN

The IEEE 802.11 standards specify the three physical layers for wireless LAN: the frequency hopping spread spectrum, the direct sequence spread spectrum (DSSS), and infrared. This section describes infrared wireless LAN as specified in IEEE 802.11. This "standard" version of infrared wireless LAN has yet to see a massive market following.

12.2.1 Physical Layer

The infrared physical layer uses infrared light as the transmission medium to carry data. The light flashes represent the digital 1 and 0 signals and convey the information from point A to point B.

12.2.1.1 Infrared Transmission The infrared of the IEEE 802.11 specification is also known as *baseband infrared* because it uses near-visible light in the 850- to 950-nm range. This is the same spectral range used by common consumer devices such as infrared remote controls and the communication technology like IrDA devices. The target data rate is 1 Mbps with an optional 2 Mbps (IEEE 1999).

There are two basic types of devices in an infrared-based LAN: infrared light emitters and infrared light receivers. The types of emitters and receivers determine the data rate and coverage distance among other properties. There are two types of emitters: light-emitting diodes and laser. LED, widely used now for newer types of TV and computer displays, produces a light source comprising of a band of frequencies,

normally between 25 and 100 mm. LED light is more spread out and less condensed and inexpensive than laser, although the data rate achieved via LED is also much lower than that of laser. LED is adequate for the specified wavelength and the targeted data rate of the IEEE 802.11 infrared LAN system.

The IR LAN physical layer is designed to operate only in indoor environments because of its physical characteristics. IR radiation does not pass through walls, and its signals are greatly weakened passing through most exterior windows. Therefore an IR system is most effective when confined within a compact interior space like a single, big conference room or a classroom.

The IR LAN uses both direct and diffuse infrared transmission modes. An IR system can transmit in one of two modes: direct and diffuse. In direct mode, the light emitter and the receiver directly face each other and must "see" each other for transmission. Thus line-of-sight is required. In diffuse mode, the emitter and receiver do not need to see each other directly. Signals emitted from the emitter are reflected off a surface like a wall or ceiling and then reach the receiver. The direct and diffuse modes are shown in Fig. 12-1 (Kim et al. 1998).

The normal operating range of an IEEE 802.11 IR LAN system is about 10 m, but it can reach 20 m if the receiver is more sensitive and the emitter more powerful.

The modulation method of the IEEE 802.11 IR LAN is pulse position modulation with 16 positions (16 PPM) to achieve the data rate of 1 Mbps. The 2-Mbps version uses a 4-PPM modulation method (IEEE 1999).

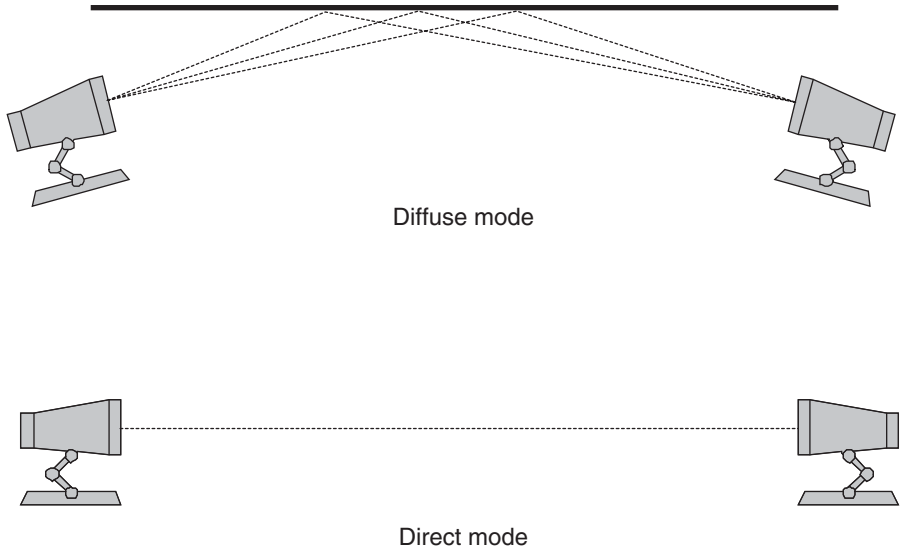


Figure 12-1
Direct and diffuse transmission mode.

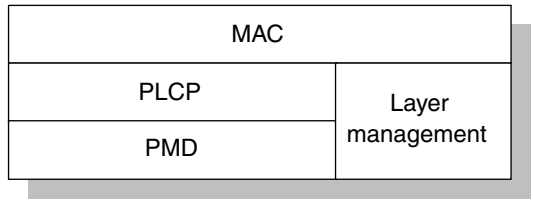


Figure 12-2
Infrared LAN protocol stack.

12.2.1.2 Infrared LAN Physical Layer Architecture The infrared LAN physical layer (PHY) consists of three functional entities, as shown in Fig. 12-2: the PMD function, the Physical Layer Convergence Protocol (PLCP) function, and the layer management function. Together they perform the conversion between optical signals and electronic signals and provide PHY service to the MAC layer via a well-defined service access point.

The PMD sublayer provides support for functions such as infrared clear channel assessment (CCA), signal transmission, and signal reception that can be used by the MAC sublayer via PLCP to send and receive data between two stations.

The PLCP allows the MAC sublayer to operate with minimum dependence on the PMD sublayer by simplifying the PHY service interface to the MAC sublayer. Physical layer-specific services are converted to more general services to the upper layer. The functions performed by the PLCP sublayer include infrared signal modulation and data encoding using the 16-PPM modulation method, data synchronization, clock recovery, and data rate regulation. The PLCP also provides data transmission and reception procedures and the procedure for CCA operations.

The PHY layer management provides management service of the local PHY elements to the MAC sublayer management entities.

12.2.2 Infrared LAN Security

An infrared communications system has the unique advantage of very secure communications. As will be described later, in practice it is very difficult, if not impossible, to intercept or alter the data sent between a transmitting station and a receiving station without alerting the receiving party about the communication.

12.2.3 MAC Layer of Infrared LAN

The upper layers of infrared LAN, from the MAC layer up, are the same as those for IEEE 802.11 wireless LAN. IEEE 802.11 defines the MAC layer

for the wireless network; the layers above the MAC layer are beyond the scope of wireless LAN.

IEEE 802.11 adopts CSCMA/CA, a modified version of CSMA/CD. CSMA/CA uses a four-way hand-shake protocol to ensure that there is only one station transmitting data at one time. For more details, see Chap. 9.

12.3 Free Space Optics System Architecture

Free space optics is a new generation of infrared wireless network technology developed to address the issue of broadband access networks. Yes, it is a type of light transmission-based system. But it is different from the IEEE 802.11 wireless system in that it provides nonstandard solutions to achieve very high data transmission rates. It is being developed by a group of small startups, which formed an industry forum called FSO Appliance as the technology's advocacy group early in the year 2002 (FSO Alliance 2002). The technology is new, and FSO development and deployment are still in their early stages.

12.3.1 Characteristics of FSO Systems

The principle of FSO is same as that of infrared LAN in that it uses light beams traveling through free space as optical links. The free space links behave similarly to fiber optic links: The output of the optical transmitter, the optical signals, are put into very focused beams of light. The difference is that FSO signals are sent across free space, while fiber optic signals are sent through strands of optical fiber.

FSO systems operate in the unlicensed infrared band, their wavelengths ranging from 750 to 1600 nm, which are in the near IR band and lower portion of the middle IR band. Most of the reported commercial FSO systems operate on two wavelength bands: 750 to 900 nm and 1500 to 1600 nm, largely because these two bands have better transmission quality through the atmosphere. In contrast, the IEEE 802.11 IR wireless LANs operate only in the wavelengths from 750 to 850 nm (Willbrand and Ghuman 2001).

FSO systems exclusively use directed infrared light that requires line-of-sight, while IEEE 802.11 IR wireless LAN operates with diffuse infrared light that does not require line-of-sight.

FSO systems aim to achieve extremely high data rates. The data rates of deployed systems reportedly have ranges from T1 (1.5 Mbps) all the way up to 10 Gbps, while the data rates in lab environments reportedly have reached as high as 60 Gbps.

The reported maximum FSO link distances vary by vendor, ranging from a few hundred meters to a few kilometers. In general, the tradeoff is the longer the transmission distance, the more powerful the light emitter has to be and more accurate the receivers must be, which leads to higher costs for users. The maximum effective distance between two stations is often determined by the fixed locations of two buildings and depends on parameters such as the laser power level, the forwarding error correction algorithm, the desired data bandwidth, and the visibility of the location.

FSO systems to date do not use wavelength multiplexing and thus only support one-light, one-optical channels. As WDM technology matures and cost-effective off-the-shelf components become available, it will be cost-justifiable for a FSO system to have WDM capability, and so to scale up an FSO wireless network.

FSO is characterized by a few features that are absent from IEEE 802.11 infrared wireless LAN and that help account for the longer transmission distances and much higher data bandwidths. Those features include advanced modulation techniques, multiple transmission channels, and shorter wavelength bands.

FSO systems take advantage of the newer generation of modulation techniques, such as orthogonal frequency division multiplexing (OFDM), and encoding schemes developed for fiber optical communications in recent years to maximize data throughput over optical links.

Many FSO systems operate in the band of shorter wavelengths around 1300 and 1500 nm. The shorter-wavelength band, which can be thought of as flashing light at a faster rate, results in higher bandwidth but also requires that both transmitters and receivers operate more precisely. Recent advances in optical component technologies make that practical. Note that FSO systems are not standardized and thus their developers are free to choose the wavelengths suitable for the target applications.

FSO systems make use of intelligent error correction mechanisms and algorithms. The laser beam is very narrow, allowing very little divergence. The convergent beam makes interception very difficult but also makes alignment of the initial system difficult as well. Intelligent autotracking and alignment help combat visibility problems resulting from building swaying, weather changes, and changes in the environment.

FSO systems avoid the costly and time-consuming need to trench and lay fiber in densely populated metropolitan areas while providing a huge amount of bandwidth (on the order of Gigabits per second). The network buildout capital is small for an FSO system compared to a wireline network. FSO carriers can avoid heavy buildout by deploying laser terminals after customers have signed on as opposed to building a wired network before signing up customers. In general, the cost of an FSO system is lower than that of wired fiber cable because the transmission medium, the air, is free.

FSO systems operate in the completely unregulated frequency spectrum (in the United States, no license is required once above 600 GHz). This avoids the need to go through a long and costly licensing process. There is very little traffic in that spectrum currently, and thus little interference with other transmissions.

12.3.2 FSO Signal Attenuation

FSO has a set of unique challenges to overcome. It requires line-of-sight, meaning that the transmitter must "see" the receiver for the signal transmission to go through. Numerous elements can cause partial or complete loss of line-of-sight (also known as *optical path*) across an unprotected free space or through the air. The following factors affecting the visibility of light are all concerns in the design of FSO systems:

Fog. This is one of the most challenging issues for FSO systems, and can drastically reduce the visibility and thus the bandwidth and increase the bit error rate. This in turn can cause signal attenuation as high as 315 dB/km, according to some research, which is much worse attenuation than that experienced by microwaves in heavy rainfall. Blizzards, dense smoke, and heavy rainfall can also cause diminished visibility.

Atmospheric scintillation. This refers to the effect that causes "shimmering" on a roadway or a rooftop on a hot summer day and the "twinkle" of stars. It can cause transmission interruptions on free space optical links.

Natural and humanly caused problems. A number of natural or humanly caused things may cause transient obstructions that are hard to anticipate. A flock of birds or insects, low-fly helicopters, advertising blimp balloons, window washers, construction cranes, among other hard-to-anticipate events may cause transient interruptions to visibility.

Variability in visibility. Variability and unpredictability in visibility present special challenges to the engineering of a FSO system. Visibility may

vary from morning to afternoon, from one side to the other of the same building, from summer to winter, or from an airport to a downtown office building. This means that each system must be set up and tuned individually.

One way to deal with fog is to use higher transmitting power and to choose wavelengths such as 1550 nm that permit high transmitting power. An equally effective method is to limit the transmission distance to a short span—say, between 200 and 300 m.

One way to deal with scintillation is to use multiple lasers separated by very short distances to send the same information to the receiver. The theory is that, in traveling to the receiver, not all the parallel beams will encounter the same scintillation pockets.

One often suggested solution for the problems associated with variable or unpredictable variability and unpredictability of the visibility is to build intelligence into a FSO system so that it can adjust its transmission power level based on the level of visibility.

12.3.3 FSO System Components and Configuration

An FSO system typically consists of a transmitter, a receiver, and an interface point to the wireline network. Figure 12-3 shows a logical view of an FSO system that consists of two transceivers, a duplex free space

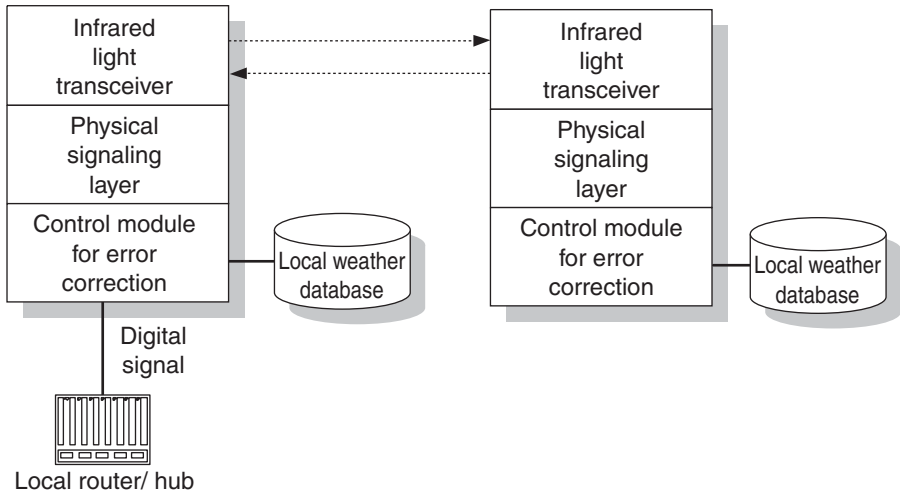


Figure 12-3
A logical view of an FSO system.

optical link, a physical signaling module, a control module, and a local routing module.

The infrared light transmitter is like any other optical transceiver, mainly consisting of a laser transmitter, a wavelength detector, and a signal encoder. The light beam coming out from the transmitter has a divergence angle in the range needed to account for sway of the building and scintillation. A receiver is an optical receiver that detects the optical signals and converts them into digital ones, as described in Chap. 6.

An FSO system equipped with dual transceivers can achieve duplex communications where a node can transmit and receive data at the same time. This effectively requires that dual FSO links be set up between the two communicating parties.

Each FSO system has a localized "weather" database that contains the parameter values for elements that could affect the visibility of this particular location. The database may include information on temperature, recorded visibility by season and time of day, precipitation, fog conditions and other parameters that may need to be taken into account for tuning the system.

An FSO system can be configured in three different ways: point-to-point, point-to-multipoint, or both. The point-to-point FSO system simply connects two buildings to allow for either simplex or duplex communications. This is ideal for a dedicated link between two sites. A point-to-multipoint FSO system is capable of transmitting data to multiple sites through the air. Point-to-multipoint technology provides support for mesh network topology and for bandwidth splitting among multiple sites.

An FSO system can be configured in various topologies. In a mesh topology, a site is connected to two or more other sites, mainly for the purpose of increasing network reliability. Alternatively, a "ring" topology allows an FSO site to become a ring node, or at least one hop away from the ring, so an alternative route can be readily available in case of link unavailability.

12.3.4 Security of FSO Systems

FSO systems, or any other infrared-based communications system for that matter, are inherently more secure than the radio frequency-based communications system. This is because an FSO system does not broadcast to everybody. An FSO system transmits a narrow, high-frequency beam of light to a specific destination. It is very difficult, if not impossible, to intercept the beamed signals without interrupting the original signals and alerting the intended receiver.

FSO light beams are difficult to detect. Laser beams are invisible to naked eyes and cannot be detected with regular spectrum analyzers or RF meters. This makes the eavesdropping or interception of the signals even harder.

12.4 FSO Applications and Deployment

This section describes several application scenarios of FSO systems and the issues to be considered in the deployment of an FSO system.

12.4.1 FSO Applications

An FSO system is a physical layer technology. It is protocol-independent and simply provides a transmission pipe that carries data in any protocol format. The reported applications supported by FSO systems include 100baseT Ethernet, FDDI, Gigabit Ethernet, OC3, and OC12. A network built on an FSO pipe can carry ATM, IP, frame relay, and other types of traffic that the adopted data link layer supports (Clark et al. 2001).

An FSO system can be easily attached to an existing fiber optical backbone network. As an access network, it can be integrated with little effort at the transmission layer of such a network, especially when both sides operate at the same wavelength.

12.4.1.1 Enterprise LAN Connectivity The most common applications of FSO systems have been for campus-wide connectivity or enterprise markets, because those applications require high-speed connections over short distances and are not constrained by the high degree of service availability requirement often seen for telecom markets. An FSO system can serve as the only data pipe to connect to a subnet located in two buildings or as a substitute for an existing low-speed connection while the existing pipe serves as a backup. Figure 12-4 shows one example where an FSO wireless optical link connects two campus buildings.

12.4.1.2 Metro Access Network Another emerging application is metro access networks, where an FSO system interconnects multiple business buildings in a downtown area where trenching and laying fiber cable are either prohibitively expensive or impossible due to regulations

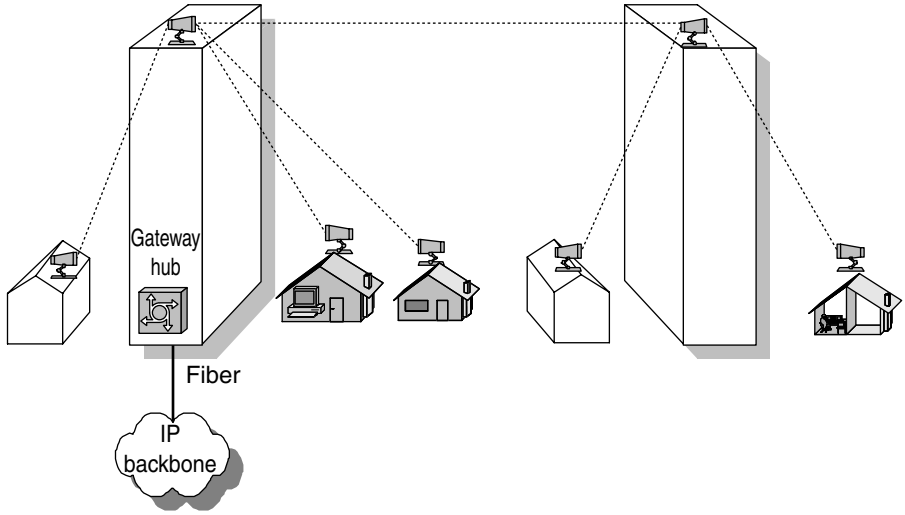


Figure 12-4
A free space optical system example.

or other reasons. An FSO system can be used as a primary broadband access network, as shown in Fig. 12-4. One other deployment scenario is to add an FSO system as a transparent overlay to an existing network like LDMS to boost the link capacity or as a backup to a primary network.

12.4.1.3 Emergency Backup Communications Systems An FSO system is an ideal candidate as an emergency backup network to provide network survivability in emergency situations where the existing wireline communication infrastructures have been devastated, as in situations like the 2001 Sept. 11 terrorist attack in New York City or a large fire. An FSO system can also be installed for special events such as large gatherings or at construction sites where optical links are needed on a temporary basis. They are useful in these cases because FSO systems are easy to set up on short notice.

12.4.2 FSO System Deployment Issues

The main issues for the deployment of FSO systems include system reliability and laser safety.

12.4.2.1 FSO System Reliability One of the key issues confronting FSO systems is their reliability or lack of it, which results from its inherent requirement of line-of-sight and numerous factors that may affect line-of-sight.

There are currently two general approaches to tackling the reliability issue. One is through network topology. An FSO system can be configured in various network topologies: mesh, star, or ring. The mesh topology has the advantage of providing redundant routes. Traffic can be rerouted onto alternative free space links, away from the area that is affected by fog, scintillation, or other conditions lowering visibility.

Another approach is to use the FSO system in combination with another system to better network reliability. Microwave is complementary to FSO in some key aspects since it is sensitive to elements that have little adverse effects on a FSO system and vice versa. For example, heavy rain can severely affect a microwave system's performance but has little effect on an FSO system. On the other hand, a microwave system is insensitive to fog and atmosphere scintillation.

12.4.2.2 Laser Beam Safety The safety of lasers in open space is another major concern, for both consumers and the general public, in regard to the use of FSO. There are no standards for FSO systems and the compliance with safety regulations by each vendor's product can be vendor-specific. Many factors together determine whether a laser light is harmful, chief among them being the brightness of the light, or the power density, not the total output of a light. Laser lights can potentially be harmful as a result of their brightness or radiance.

FSO systems, although operating in the unlicensed frequency spectrum, need to comply with the various laser safety standards defined by the IEC (International Engineering Consortium) for the international communities and by ANSI for North America. Although these standards are not completely aligned, in general, laser beams are classified into safety categories based on beam intensity, distance from the laser, and wavelengths among other considerations.

A majority of FSO systems, according to the vendor-published literature, are said to be of the Class 1 or Class 1m category that are defined as "safe under reasonably foreseeable future" and deemed as "eye safe." Even so, it is still advised to use operational precautions such as labeling laser components, posting visible signs for the laser beam path, or setting up an FSO system such that the laser beam passes through an area inaccessible to humans.

In addition, the following safety features can be designed into FSO systems:

- *Intelligent power control.* Systems can monitor the beam path and shut down or reduce the power once it finds the beam blocked.
- *Intelligent beam control.* Systems can activate an alternative beam or reroute the traffic when an active beam is found to be blocked.

12.5 IrDA

IrDA (Infrared Data Association), an industrial consortium, is another type of infrared-based communications system that defines a set of standards for devices like video cameras and computers to communicate with each other through infrared light over very short distances. The IrDA standards are widely used in consumer devices and in future may play a more important role in the home networking market.

IrDA defines two sets of standards, IrDA Data and IrDA Control. IrDA Data targets infrared-based cordless communications between consumer devices. IrDA Control targets communications between peripheral devices like mouses and host devices like computers (IrDA 2000). The IrDA standards are under consideration to be adopted as ISO standards. This section provides a high-level overview of the IrDA standards and systems.

12.5.1 Overview of IrDA Data Standard

The IrDA Data standard was originally defined in 1994 for the interoperable two-way cordless infrared light transmission data port for electronic devices like palm PCs, printers, phones, pagers, toys, and other mobile devices (IrDA 2001). It consists of a set of mandatory protocols that covers the bottom two layers of the OSI network reference model and a set of optional protocols that covers the functions of transport and the application layer. The mandatory protocol set includes the following:

- The Physical signaling layer that supports two-way communications with data rates ranging from 9 Kbps to 4 Mbps and transmission distances from 1 to 2 m
- The IR Link Access Protocol (IrLAP) that establishes device-to-device connections for reliable, ordered, data transfer and provides the device discovery procedure
- The IR Link Management Protocol (IrLMP) and Information Access Service (IAS) that provides for the multiplexing of multiple channels over an IrLAP connection and a protocol for service discovery

The set of optional protocols include the following:

- *Tiny TP,* which covers flow control on IrLMP connections with an optional segmentation and reassembly service

- *IrCOMM,* which covers serial and parallel COM port emulation services for legacy COM applications like printers and modems
- *IrOBEX,* which covers object exchange services similar to Hypertext Transfer Protocol (HTTP)
- *IrDA lite,* which covers the mechanism for code compression while maintaining compatibility with the full functionality
- *IrTran-P,* which covers an image exchange protocol for video devices like video cameras
- *IrMC,* which covers support for communications between mobile devices like cell phones and other communications devices to exchange information like calendar data
- *IrLAN,* which covers devices to access local area networks

12.5.2 Overview of IrDA Control Standard

IrDA control is the more recent of the two IrDA standards and targets cordless communications between cordless peripherals including keyboards, mouses, game pads, pointing devices, and host devices like PCs, laptops, and digital television sets. It supports bidirectional two-way communications with sophisticated control functions (IrDA 1998).

The IrDA control standard consists of the following three layers, analogous to the three wireless protocol stacks described in Chap. 9 that correspond to the bottom two layers of the OSI network reference model:

- Physical
- Media access control
- Logical link control (LLC)

The physical layer uses directed infrared light for point-to-point communication between a peripheral and a host device with a transmission distance between 5 and 10 m that is equivalent to current unidirectional infrared remote control. The data is encoded with a 16-pulse sequence multiplied by a 1.5-MHz subcarrier with a maximum data rate of 75 Kbps. The cyclic redundancy check is used to detect data transmission errors.

The MAC layer allows a host device to communicate with up to eight multiple peripheral devices at a time. The LLC layer provides data sequencing and retransmission when errors are detected.

REVIEW QUESTIONS

1. Describe the characteristics of infrared communications systems in terms of their advantages and disadvantages.
2. Describe the differences between IEEE 802.11 Infrared LAN systems and radio-based IEEE 802.11 LAN systems.
3. Describe the functional components of the IEEE 802.11 IR LAN physical layer and the functions of PDM and PLCP.
4. Describe the differences between free space optics and IEEE 802.11 IR LAN systems.
5. Describe the main factors that account for the longer transmission distances and much higher data rates of FSO systems compared to IEEE 802.11 IR LAN systems.
6. Explain why FSO as a physical layer technology is said to be protocol-independent and describe what an FSO system protocol stack might look like when applied in a LAN environment.
7. Describe one of the likely application scenarios of FSO systems and the main issues an FSO system must contend with in that scenario.
8. Describe the factors that may affect line-of-sight for an FSO system and the methods FSO systems have adopted in addressing reliability issues.
9. Describe the intended applications of IrDA Control and IrDA Data standards.

REFERENCES

Clark, G., Willebrand, H., and Achour, M. 2001. "Hybrid Free Space Optical/Microwave Communications Networks." White paper. LightPointe Inc. Web site: www.lightpointe.com.

FSO Alliance. 2002. "FSO General Information." White paper. Web site: www.fsoalliance.com.

IEEE 1999. "Part 11: Wireless LAN Medium Access Control and Physical Layer (PHY) Specifications." IEEE 802.11, 1999 ed. Web site: www.ieee.org.

IrDA. 1998. "IrDA Control Specification. Version 1.0." Infrared Data Association. Web site: www.irda.org.

IrDA. 2000. "Technical Summary of IrDA Data and IrDA Control." Web site: www.irda.org.

IrDA. 2001. "Serial Infrared Physical Layer Specification. Version 1.4." Infrared Data Association. Web site: www.irda.org.

Kim, I., Stieger, R., et al. 1998. "Wireless Optical Transmission of Fast Ethernet, FDDI, ATM and ESCON Protocol Data Using TerraLink Laser Communications System." *Optical Engineering.* Vol. 37, No. 12.

Willbrand, H., and Ghuman, B. 2001. *Free Space Optics: Enabling Optical Connectivity in Today's Network.* Indianapolis, IN: Sams Publishing.

CHAPTER 13

Satellite Packet Broadband Networks

13.1 Satellite Communications Basics

Communications satellites have less than half a century of history. Over the years, three different types of satellites have emerged that use two different types of links to communicate between satellites and earth stations and between two satellites. This section gives a brief history of satellite communications and describes their unique characteristics.

13.1.1 Brief History

In 1962, a few years after the first satellite Sputnik I was launched on October 4, 1957, by the Soviet Union, the United States launched communications satellites Telstar and Relay, which were aimed at carrying intercontinental telephone circuits. That event helped usher in the "global village" age. The International Telecommunications Satellite Organization (INTELSAT) was created via an international agreement to pool resources for the development and use of satellites. By the last count, INTELSAT has over 110 nations as members.

In the early 1970s, domestic communications satellites were launched in North America and Europe for transmission of TV signals and telephone calls over long distances. Examples include the ANIK communications satellite launched by Canada, and the first of the COMSTAR series launched by AT&T and COMSAT by the United States. Very quickly, movie channels and superstations via satellite became available to most Americans.

Also starting in the 1970s, a new type of communications satellites was launched: a maritime satellite providing mobile service to maritime customers. In 1979, the United Nation's International Maritime Organization helped establish the International Maritime Satellite Organization (INMARSAT). INMARSAT initially leased satellite transponders and later launched its own satellites, like INMARSAT III in 1990. The next natural extension to this satellite mobile service was the mobile service to land mobile users via satellite. In the early 1990s, satellite constellations such as Iridium were built to provide worldwide mobile telephony services.

The next major development in satellite communications was direct broadcast to customer premises. Satellite TV programming into homes gained market acceptance first, followed by Satellite Internet to home in the late 1990s and early 2000s. These recent developments were driven by

several factors. First, the growth of the Internet created a huge demand for raw bandwidth and an economically efficient way to provide broadband access to customer premises. Satellite broadcast also provides a viable, alternative, solution to the last-mile problem. The coming of the satellite Internet was also due to advances in several key satellite technologies, including advanced power systems, high-gain antenna, and new types of satellites.

Low earth orbit (LEO) and medium earth orbit (MEO) satellites were another development that stimulated the development of broadband satellite networks. The satellites developed and launched before the 1980s were mostly of the GEO type, located in orbits high above the earth. Advances in technology meant that LEO and MEO satellites could be located in orbits much closer to the earth and so can transmit data with much shorter time delays (Whalen 1998).

13.1.2 Three Types of Communications Satellite

Satellites are classified as either passive or active according to how they work. A passive satellite reflects or passes received radio signals back to the earth. It serves only as a mirror. In contrast, an active satellite acts as a repeater: It amplifies the signals received and then retransmits them back to the earth.

A satellite circles around the earth in a fixed orbit. The angle between the equatorial plane of the earth and the orbital plane of the satellite is known as the *angle of inclination*. If a satellite's orbital plane coincides with that of the equatorial plane, the angle of inclination is zero. Otherwise, the orbital plane is inclined at whatever angle it maintains in relation to the equator.

Communications satellites are also classified into three generic categories based on the distance of their orbits from the earth: LEO and MEO—already mentioned—and geosynchronous/geostationary earth orbit (GEO) (Network Computing 2001).

13.1.2.1 GEO Satellites Genosynchronous earth orbit satellites are the farthest away from the earth. It was discovered in 1945 by a scientist and author Arthur C. Clarke that an orbit with a distance of 22,000 mi above the equator has a unique characteristic: The gravitational pull toward the earth exactly matches the centrifugal force pulling the satellite away from the earth. As a result, a satellite in this orbital location

remains stationary in relation to the earth, hovering motionlessly over it. This kind of orbit is also known as a *geostationary* or *Clarke belt* orbit.

The majority of the early satellites were of the GEO type. There are several key advantages associated with GEO satellites. First, they have very large "footprints," covering large areas potentially as large as nations and continents. Appropriately positioned, three GEO satellites are sufficient to cover the entire earth.

Since only a small number of satellites are involved in a GEO satellite network, its topology is relatively simple to manage and maintain. Because each satellite covers a large area, the chance for handoff from one satellite to another is small. This means the delay from a satellite to an earth station is predictable and less variable. In addition, there is less routing complexity and jitter in the network. And the limited number of satellites in a network means intersatellite links are simple.

On the other hand, there are a number of disadvantages associated with GEO satellites, which have significant implications for the choice of packet over satellite architecture. One of them is the greater power needed for longer-distance transmission. Another relates to the longer transmission delays due to the longer transmission distances. Also, the earth station equipment needed for GEO satellites—CPE dishes, for example—tends to be bulky.

13.1.2.2 MEO Satellites MEO satellites orbit below the "Clarke belt," in a range of distances approximately 2000 to 12,000 mi above the earth. They have less transmission latency, ranging from 0.06 to 0.14 s for a round trip between the satellite and the earth. A MEO network needs more satellites than a GEO system to cover the entire earth. Up to this point, MEO satellites have played a complementary rather than primary role in broadband communications.

13.1.2.3 LEO Satellite LEO satellites orbit at distances from 500 to 1000 mi above the earth. They offer some very attractive features for broadband communications, the most important of which involves much shorter transmission delays than GEO or MEO satellites because of the short transmission distances between an earth station and a LEO satellite. This makes LEO satellites a favorite choice for real-time-sensitive application like voice over IP and multimedia streaming. Also CPE equipment like customer premises dishes tend to be smaller and more versatile.

On the other hand, the short orbit distances of LEO satellites present a set of challenges of their own that are not trivial to overcome. The first

challenge is that a large number of LEO satellites in low altitude are needed to achieve continuous global coverage, and because of the short orbit distance, each satellite only has a very short in-view time window, approximately 15 min to half an hour, for each earth station. This requires a much more complicated network topology and frequent, more complicated hand-off from one satellite to another, which in turn poses a nontrivial challenge for the management, maintenance, and control of a network requiring a large number of satellites.

Another challenge is the large jitter associated with LEO satellites. Transmission jitter is the variation of delay rather than delay itself. A LEO network has higher jitter than a GEO or an MEO network, resulting from its frequent and complicated intersatellite handoff. Jitter poses a great challenge to the design of real-time applications such as interactive gaming and VoIP.

The characteristics of GEO, MEO, and LEO satellites are summarized in Table 13-1.

13.1.3 Satellite Network Transmission Medium

There are two types of transmission media used between satellites and earth stations and between two satellites: radio frequency and free space optics (i.e., laser light). Radio frequency is the standard transmission media, especially for the satellite-earth links (WTECH 1998).

TABLE 13-1

Comparisons Between GEO, MEO, and LEO Satellites

Characteristics	LEO	MEO	GEO
Satellite life span	5 years	5–10 years	10 years
Launch expense	Least expensive	Moderately expensive	Most expensive
Constellation to cover the entire earth	50–60 satellites	28–35 satellites	3–5 satellites
Distance from earth	500–1000 mi	8000 mi	22,000 mi
Suitable applications	Real-time-sensitive applications	Broadcast and some real-time-sensitive applications	Broadcast
In-view time span per revolution	15–30 min	2–4 h	24 h; stationary to an earth station
Latency time of round trip	50 ms	Around 150 ms	500 ms

TABLE 13-2

Radio Frequency Bands for Satellite Communications

Frequency band	Frequency range, GHz
L band	0.5–1.5
C band	3.6–7.025
X band	7.25–8.4
Ku band	10.7–14.5
Ka band	17.3–31.0

Satellite communications use a radio frequency spectrum known as the super-high-frequency (SHF) band, which extends from 3 to 30 GHz. This band is sometimes referred to as the *centrimetric band,* because the wavelength of SHF signals ranges from 1 to 10 cm. More precisely, satellites mostly use three distinct subbands of the SHF band known as the *C-band,* the *Ku-band,* and the *Ka-band,* as shown in Table 13-2.

Early satellites used the lower Ku-band frequencies, and the Ku-band frequencies are now fully utilized at certain orbital positions. This, coupled with the increasing demand for higher bandwidth and the potential for highly compact earth terminals, has fueled interest in the higher Ka-band frequencies in recent years.

Free space optics, also commonly referred to as *laser beam,* is a new type of transmission medium that in recent years has been used mostly in intersatellite links. It is predicted that FSO will be the dominant choice for intersatellite communications in the near future because of its high bandwidth, the compact transceiver it uses, and other advantages.

13.1.4 Characteristics and Challenges of Satellite Communications

Broadband satellite networks have several unique characteristics compared to their earth-based counterparts. A satellite has large footprints that can potentially span nations and continents and service hard-to-reach corners of the earth. For example, as already noted, three GEO satellites can cover the entire globe. This makes satellite networks a particularly favorite choice for broadband communications in rural and far-reach areas.

A satellite network is unidirectional in nature: A satellite link carries traffic in only one direction. Two independent links, a downlink (from satellite to earth) and an uplink (from earth to satellite), are needed to carry traffic bidirectionally.

Satellite networks involve point-to-multipoint topology, with one satellite broadcasting to all the earth stations in its footprint. This topology is a natural fit for last-mile access networks that require a point-to-multipoint network structure with large bandwidth for the downlink direction and relatively small bandwidth for the uplink direction.

Satellite communications can provide a large amount of the bandwidth. In the early days of satellites, satellite networks had huge bandwidth advantages over their earth-based counterparts. For example, 36 trans-Atlantic phone circuits cost between $30 million and $50 million in the 1950s. In comparison, a satellite could carry over 1000 circuits at a much lower cost. Although the bandwidth advantage of satellites dissipated as optical fiber became a common transmission medium, the bandwidth of the currently reported broadband satellites has reached the OC48 rate (2.5 Gbps) and is closing in on the OC192 rate (10 Gbps). The projection is that a new generation of broadband satellites will reach 10s of Gbps in the not-so-distant future.

Satellites are reliable in the sense that they are free of the effects created by the earth's environment. For example, optical network outages are more often caused by human error such as cable cuts and accidents than by equipment failure. Satellites are reliable to the extent that their equipment is reliable. Once in orbit, humans do not touch them.

13.1.5 Satellite Communications Standards

Several standards organizations are active in the standardization process for packet over satellite. ATM Forum has been involved in defining interoperable interfaces for ATM over satellite broadband architecture since the late 1990s. IETF is active in defining extensions to the classic TCP protocol to support IP over satellite. The Telecommunications Industry Association (TIA) and ITU-T are working on interoperable standards to interconnect satellites with terrestrial networks.

13.2 Components of Satellite Broadband Networks

A satellite broadband network consists of four principal components: satellites, earth stations, transmission links (intersatellite and satellite-earth links), and ground CPE devices, as shown in Fig. 13-1.

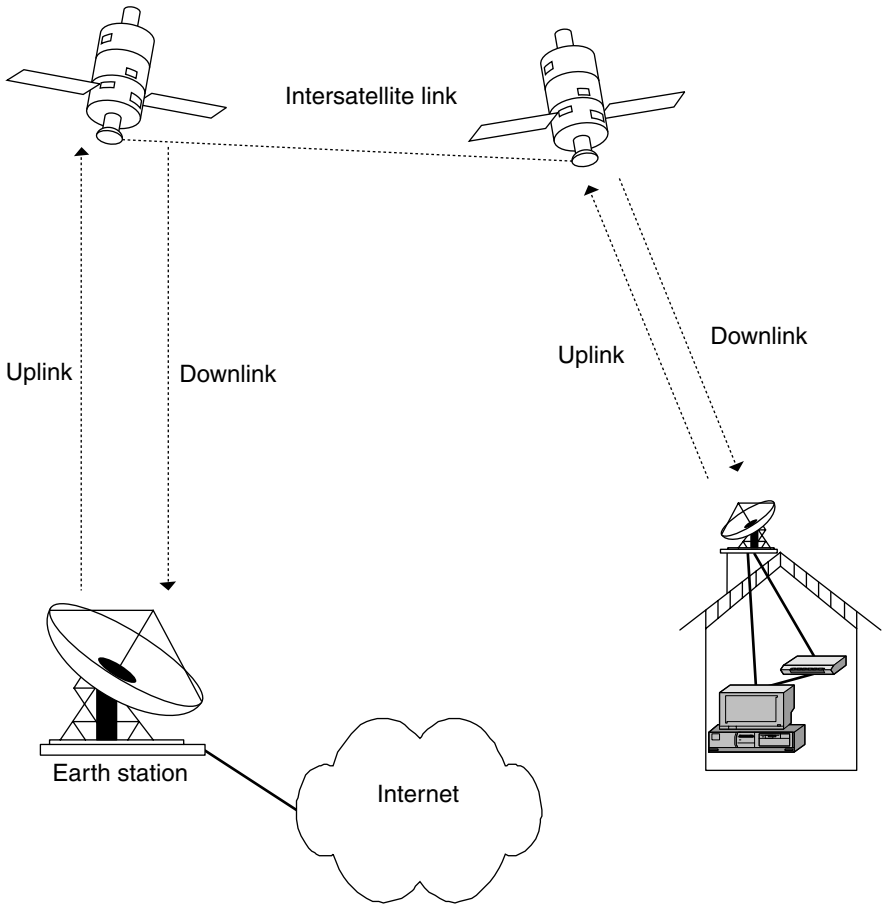


Figure 13-1
Components of a satellite network.

13.2.1 Satellite

A satellite is a very complicated system with thousands of parts. But at a functional level, a communications satellite consists of three main components of general interest: a power system, an antenna system with one or more transponders, and an onboard processing unit.

13.2.1.1 Power System The power system of a satellite is a key component that determines its life span and capacity. The more transmission power a satellite provides, the higher the bandwidth it can have, and the smaller and less costly the ground user terminal and antenna it enables.

A power system is normally a regenerable power source consisting of a solar cell, a battery, and power conditioning electronics. The solar cell provides the power for normal operations, while the battery provides the

backup power. The power conditioning electronics provides the control and maintenance functions for the power system.

Solar cell technology has greatly improved in efficiency over the past few decades and today provides much higher operating power than in earlier years. For example, early satellites had power as little as 1 W. Today's efficient solar cell systems and advanced battery technologies now can provide operating power in excess of 300 W.

13.2.1.2 Antenna and Transponder The satellite antenna is responsible for transmitting and receiving radio signals. The number of built-in transponders on an antenna determines the data throughput of the satellite. The large apertures (the area that can transmit and receive signals) combined with small beamwidths (the focus on signals) reduce the need for a large size of antenna.

Two kinds of antenna are in common use for communication satellites: reflector and phased array. Reflector antennas, shaped similarly to ground satellite dishes are a common type of antenna. Phased array antennas, shaped somewhat like flattened-out honey cones, are a newer type of antenna.

A transponder is a device attached to the satellite antenna that is responsible for emitting the signals going out and processing the received signals. Specifically, the signals received through an antenna on a fixed frequency are sent to the transponder, which in turn filters out background noise and converts the signals to another fixed frequency to avoid interfering with weaker incoming signals. On the transmission side, a transponder broadcasts the amplified signals to the reachable earth stations.

The development of the new generation of intelligent, high-gain antennas, coupled with the development of precise aperture for transmitting and receiving signals has led to the development of very small dishes that can be massively produced at a low cost. This allows the earth antennas to be much smaller than before and the satellite antennas to be much more efficient, less power consuming, and capable of transmitting data at much higher rates.

13.2.1.3 Onboard Processing Unit The onboard satellite processing unit is the "brain" of the satellite and is becoming more and more sophisticated, able to provide router and switch capabilities for broadband satellites. In general, such a unit performs the following three functions:

- *Control functions.* These include altitude control, power management, telemetry, and tracking control.

- *Antenna control and beam forming.* A phased array antenna requires a large number of radiating elements to control a large number of independently steerable beams, one element per beam. The control logic requires a considerable amount of processing power.
- *Broadband switching and routing.* The unit that does this is similar to the router or switches of a terrestrial broadband network, and can perform either IP routing or ATM switching onboard.

13.2.2 Earth Station

An earth station is the interconnecting point between a satellite and the terrestrial network. For a cable TV satellite network, it is the headend that receives the broadcast TV signals and relays the signals to the distribution centers. For a satellite broadband data network, as shown in Fig. 13-1, an earth station is connected to a LAN or a metro area network.

Earth stations, also known as *earth terminals,* in their early days were located far away from actual users to ensure minimal interference with satellite operation. Stations were then connected to terrestrial networks via high-speed links like fiber or coaxial cable.

An earth station in general consists of antennas, transceivers, telemetry equipment, and interfaces to the terrestrial network. It has a highly directional, high-gain antenna that is capable of transmitting and receiving signals at the same time. The earth station's receiver is designed to overcome downlink power loss or weak signals with strong noise. A common technique for this is to have a specially designed preamplifier mounted behind the antenna.

The transmitter of an earth station is specially designed to generate strong signals for transmission to satellites. The powerful signals coupled with the high-gain, highly directional antenna ensure that the signals are strong enough to reach satellites and, once received, are still strong enough to be decoded.

Each earth station is equipped with a router or a switch to forward the traffic to the indicated destination via the connected terrestrial network.

13.2.3 Satellite Transmission Links

There are two types of transmission links in a satellite broadband network: intersatellite and earth-satellite, as shown in Fig. 13-1.

13.2.3.1 Intersatellite Links Intersatellite links interconnect a constellation of satellites and relay data. They traditionally have been based on radio frequency technology, but FSO is becoming a viable alternative with more and more deployments in recent years. Although still at an early stage of deployment, FSO is expected to play a primary role in future intersatellite communication.

FSO intersatellites have several key advantages over their RF counterparts. In ideal circumstance, the optical link can be 10 times more efficient than the radio frequency link. High link efficiency then can be translated into a smaller size for the satellite, less power consumption, and much higher data rates comparable to those of optical fiber. An optical link requires a much narrower beamwidth than an RF link because a laser beam is more precise, and thus the telescope aperture for receiving an optical signal can be smaller than the antenna aperture needed for radio frequency.

13.2.3.2 Space-Terrestrial Links Space-terrestrial links are unidirectional. Broadcast satellites like those for cable TV provide only downlinks, or links from satellites to earth stations. For broadband Internet satellites, uplinks (from the earth stations to satellites) are provided as well, to allow the earth stations to send data to the satellites. Normally downlinks use much higher frequencies with wider frequency bands than uplinks because downlinks are designed to provide much higher data bandwidths than uplinks.

Radio frequencies for satellite operations are regulated and licensed by national or regional governing bodies and allocated by ITU-R for satellite operators worldwide. The uplinks and downlinks are independently licensed and allocated. For example, in the United Kingdom, the uplink bands for commercial satellite earth station equipment can be only one of the following:

- Proprietary uplink (Ku-bands) at 14.0–14.25 GHz
- Shared Ku-bands at 14.25 to 14.5 GHz
- Ka-band uplink transmissions in the range 29.5 to 30 GHz
- Shared Ka-bands at 27.5 to 29.5 GHz

In addition, only permanent earth stations are allowed to transmit in the C-band (5.725 to 5.85 GHz and 5.85 to 7.075 GHz) within the United Kingdom.

Radio frequency has been and will continue to be the primary transmission medium for the satellite-earth links. Although the use of optical

links is being explored, FSO suffers from major limitations relating to atmospheric and weather interference.

13.2.4 Customer Premises Equipment

Satellite equipment entering customer premises is a relatively recent development. Traditionally customer equipment has connected to satellite networks through earth stations. It was satellite TV systems like the direct broadcast system (DBS) that first beamed satellite signals directly into customer premises. Then the satellite Internet followed suit by allowing CPE to be directly connected to satellites in both downlink and uplink directions.

CPE includes satellite dishes, modems, and coaxial cable. The CPE dish has become smaller in size in recent years, with mass-produced phased array antennas. Satellite dishes need a clear view to south in North America since orbiting satellites are over the equatorial area. Trees and heavy rains can affect reception of the signals.

For the interactive Internet, two modems are needed, one for uplink and one for downlink, for broadband service. The modems are connected to the satellite dish via a coaxial cable.

13.3 Packet over Satellite

This section provides an overview of communications satellite network architectures that have evolved from early cable TV satellites to satellite long-haul models, to the Internet-to-home models. Two approaches are available for carrying packet over satellite networks: ATM over satellite and IP over satellite. The former was implemented in the second half of the 1990s, while the latter is a more recent development, very much in line with trends in terrestrial packet network technology.

13.3.1 Broadband Satellite Network Architectures

Broadband satellites evolved from broadcast TV satellites, to long-haul data transport models to the Internet-to-home models.

13.3.1.1 Satellite TV Broadcast Systems The early generations of cable TV broadcast satellites featured one-way broadcast communications: the TV signals were transmitted from the source to a satellite, which relayed the TV broadcast signals to the TV headend located thousands of miles away, as shown in the left half of Fig. 13-2. The headend then distributed TV programming to the distribution centers, which in turn distributed signals to the individual subscribers.

The next step of satellite TV was DBS, which directly beamed programming into individual subscriber premises, as shown on the right side of Fig. 13-2. The widely available DISH network in the United States is one example of DBS. Satellite TV, either broadcasting to headend or individual homes, is still one-way communication that only receives TV broadcast signals.

13.3.1.2 Satellite Long-Haul Transport Networks Satellite long-haul architecture is a relatively early model that evolved from cable TV satellite technology and leverages existing satellites to provide broadband data service.

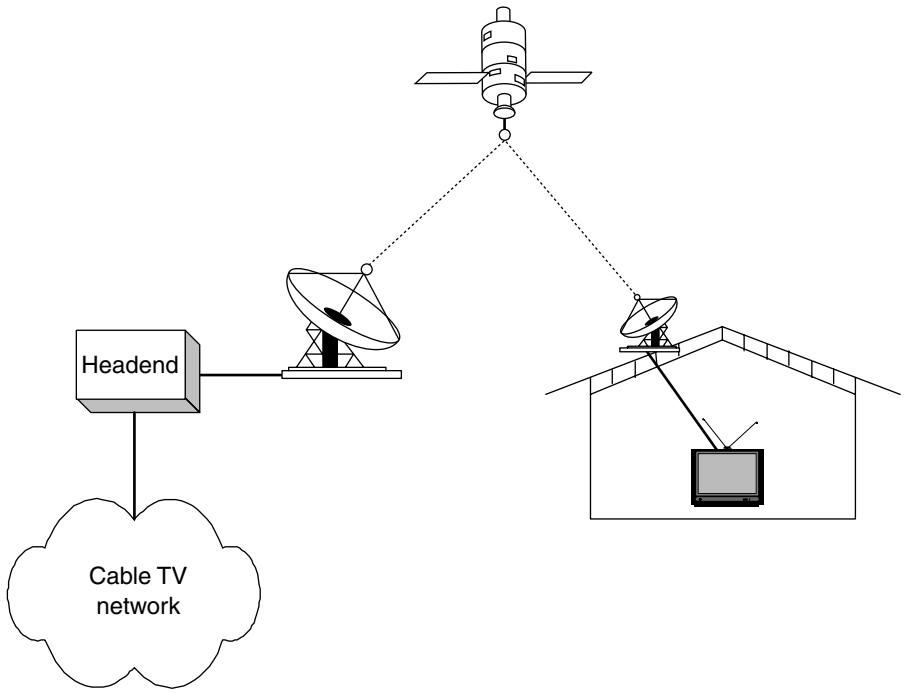


Figure 13-2
Cable TV broadcast satellite network.

In this architecture, a satellite acts as a long-haul transport network to carry broadband signals from one earth station to another, as shown in Fig. 13-3. In this respect, a long-haul satellite network is similar to a cable TV satellite broadcast network. End-user CPE is connected to an earth station, just like being connected to a point of presence, and the earth station is equipped with a router or ATM switch to forward the traffic to the connected terrestrial network.

Long-haul architecture also supports the uplink traffic by allowing users to send data to other networks in other parts of the world through satellite links. This uplink capability is the main difference between broadband satellite networks and satellite TV broadcast networks.

This network architecture has the advantage of being able to transport data over very long distances—across continents if necessary—without having to build any expensive ground infrastructure. For example, data

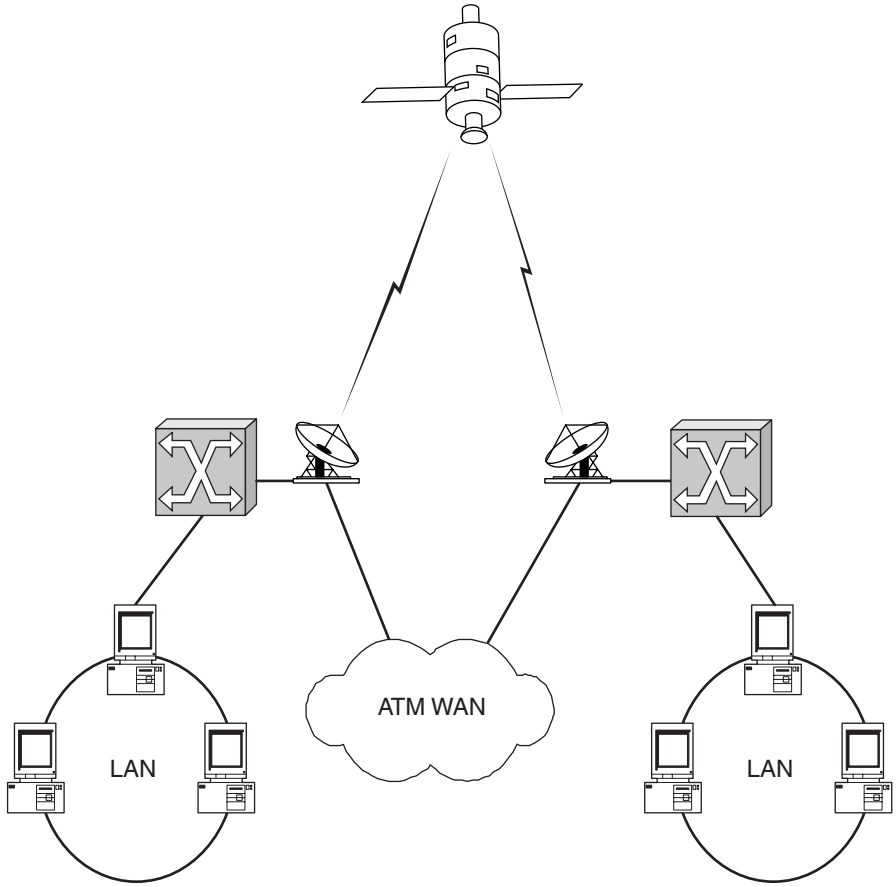


Figure 13-3
Satellite long-haul architecture.

services including frame relay and IP services can be provided by service providers such as Orion Network, PANAMSAT, and INTELSAT in the United States to ISPs located in Europe, Asia, and Latin America with no need of cables.

13.3.1.3 Broadband Satellite-to-Home Networks Broadband satellite-to-home networks are a logical next step in the evolution of broadband satellite architecture that supports two-way traffic and direct beaming of data to customer premises, as shown in Fig. 13-4. This architecture allows customers to send requests to the Internet and receive responses directly from satellites.

Satellites in this architecture provide solutions to both the last-mile and last-yard problems, in addition to being long-haul networks. A user request is transmitted from the customer antenna to a satellite, and the satellite forwards the request to an appropriate earth station. The earth station, which is connected to the Internet, retrieves the requested contents and sends the contents back to the satellite via an uplink to the satellite. The satellite then forwards the contents to the customer.

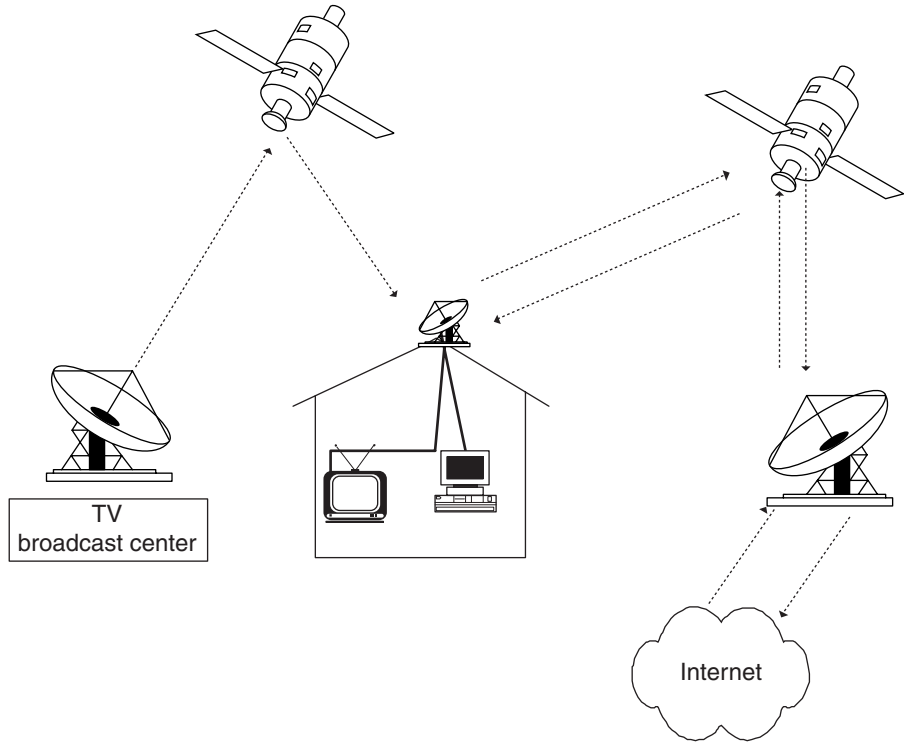


Figure 13-4
Satellite-to-home broadband architecture. (StarBand, 2000)

This architecture is particularly suitable for customers in rural areas and for corporate customers who need to interconnect multiple sites across continents to avoid prohibitively expensive wireline solutions.

Nonsatellite paths provide an alternative to user-to-satellite uplinks. Many earlier satellites that are used to provide broadband data service do not have uplink capabilities built in. Instead, a terrestrial return path via a dial-up connection like a PPP connection is used to send the user request to the destination network, and the large amount of data is downloaded via satellite downlinks.

GEO and LEO satellites are the favored choices for satellite-to-home architecture. GEO satellites are the choice for some service providers while others use LEO satellites either exclusively or LEO satellites combined with a GEO or MEO satellite. Some of the service providers around the world that either already offer or plan to offer this service include StarBand, Pegasus Express, Teledesic, and Teachyon.

13.3.2 ATM over Satellite

ATM over satellite is one of the two common approaches to packet over satellite. ATM is chosen mainly for two reasons: its fast hardware-based switching capability and its well-defined QoS schemes. By the end of the 1990s, ATM switching speeds implemented at hardware level could reach rates of OC3, OC12, and even higher. This compares well with the data rates of satellite links. QoS is a major consideration for broadband satellite networks given that satellite transmission errors can potentially be high.

An ATM over satellite architecture reference model has been defined through the joint efforts of ATM Forum, ITU and TIA, as shown in Fig. 13-5. The protocol stack of ATM over satellite has ATM entering the picture at the earth station, where interworking between the satellite and a terrestrial ATM network takes place (ITU-R 1998a; ITU-R 1998b).

The reference model and the associated proposed standards mainly focus on two aspects of ATM over satellite: a set of standard air interfaces, and the seamless interworking and interoperability between the satellites and the terrestrial networks via standard interworking functions. The standard air interfaces are intended for satellite systems ranging from personal communications systems to broadband systems.

Interoperability is achieved through an interworking function that joins a satellite network with a terrestrial ATM network. As shown in Fig. 13-5, the earth station features an ATM satellite interworking unit (ASIU) in the ATM over satellite architecture. The ASIU is responsible for

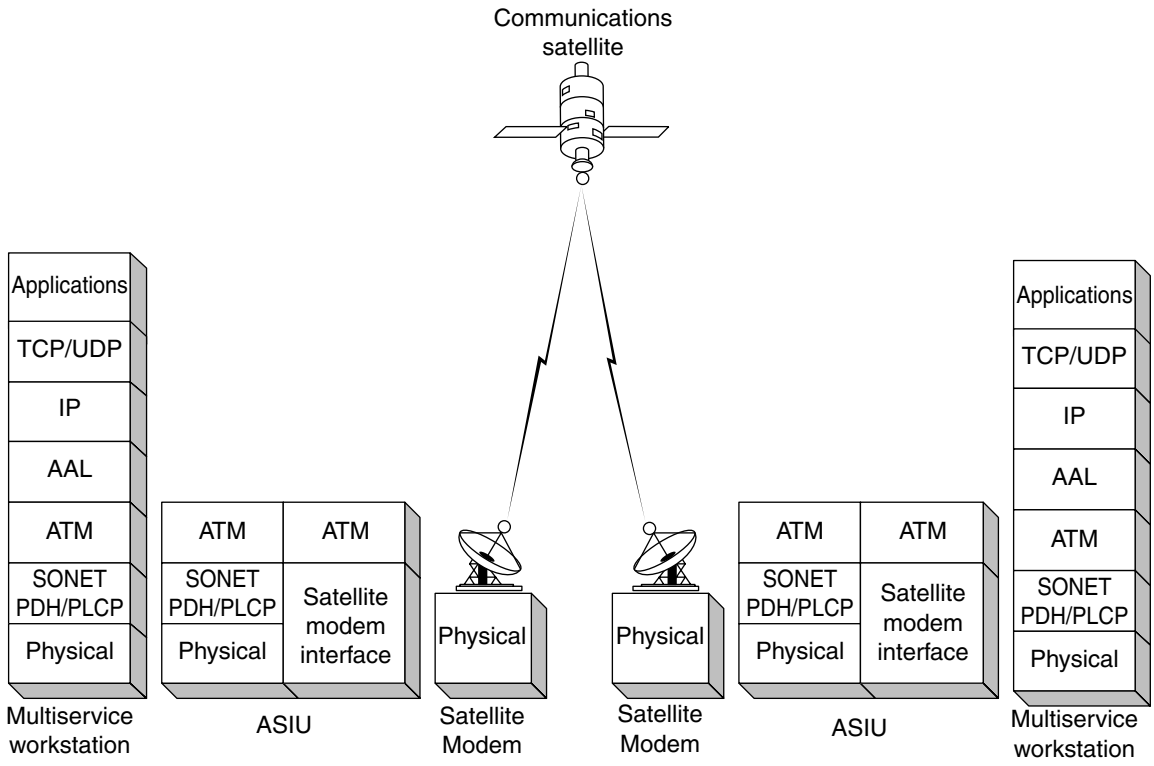


Figure 13-5 *ATM over satellite architecture. (WTECH, 1998)*

the management and control of system resources and the overall system administrative functions. Among the other major functions of the ASIU are dynamic bandwidth allocation, network access control, call monitoring, and system timing and synchronization.

Another key function ASIU performs is satellite link conditioning. Because of the inherently higher level of noise, a satellite link may have a higher bit error rate than a terrestrial optical link. The high BER may cause the dropping of ATM frames. ASIU corrects the bit errors in the ATM frames by means of a specially designed forward error correction module, and matches the BER of satellite with that of the terrestrial link.

One may note that the satellite architecture basically follows the reference model, as shown in Fig. 13-3. The satellite serves as a long-haul transmission network, connecting two ATM switches located far apart. Users are connected to the local ATM LAN, which is in turn connected to the satellite via an earth station.

13.3.3 TCP/IP over Satellite

TCP/IP over satellite is a more recent development than ATM over satellite, and the rise of interest in TCP/IP over satellite is in line with the packet technology evolution of terrestrial networks. The research and development efforts currently underway focus on a number of issues, of which three are of general interest: TCP/IP over satellite architecture, TCP protocol enhancements for applications to satellite networks, and IP routing for intersatellite links.

13.3.3.1 TCP/IP over Satellite Architecture LEO satellites rather than traditional GEO satellites are attracting more attention as the satellites of choice for a new generation of IP over satellite applications. This is primarily because LEO satellites have much shorter transmission delays—if managed well, the delays can be below the critical delay threshold of 250 ms for real-time applications, such as voice over IP and multimedia streaming. In contrast, the transmission latency of GEO satellites can reach 600 ms or more, despite the fact that the delay is more predictable and the delay and jitter of GEO satellites are better understood than those of LEO satellites (Wood 2000; Allman 2000).

Jitter is one of major challenges LEO satellites face when carrying IP traffic over satellites. The jitter comes from the complicated satellite network configuration and the need for frequent intersatellite traffic handoff. Various approaches have been proposed to address the jitter issue:

1. The use of large buffers at the earth station so the playback delay to users is constant.
2. The use of an inclined orbital pattern that angles off the equator. This should make the intersatellite traffic handoff more predictable, although it would add complexity to the system.
3. Smaller coverage for each LEO satellite. A large constellation with more satellites would have less jitter than a smaller constellation.

Satellite Internet to home is also being implemented without satellite uplinks. While the new generation of broadband satellites is intended to support satellite Internet to home with both satellite uplinks and downlinks, earlier satellites only supported downlinks. A currently prevalent implementation is to have a terrestrial connection like a PPP dial-up connection in place of a satellite uplink. In this architecture, a user request is sent to the destination via a terrestrial network, and a large amount of data is downloaded via a satellite downlink.

13.3.3.2 TCP Protocol Enhancements Some characteristics unique to satellite networks are not very suitable for the classical TCP protocol and raise the need to enhance the TCP protocol. First, satellite network traffic is asymmetric either because downlinks and uplinks have unbalanced data rates or because only downlinks are supported. Second, satellite link tends to have greater delays, more noise, and more jitter than the TCP protocol was originally designed to account for, and these conditions can potentially lead to abnormal behavior of the protocol (Allman et al. 1999).

The major issues caused by satellite-specific characteristics include packet retransmission and TCP window size shrinking. The window downsizing is based on the TCP protocol's "test water" algorithm. That algorithm sets the transmission window size (the rate of transmission) by testing how fast the network can respond. Because of satellite link delay, the window may get set to a size smaller than the actual capacity of the satellite link, so that the satellite's bandwidth is underutilized. A more serious issue is that if jitter exceeds the packet round trip time, TCP may interpret it as packet loss and start retransmission of the packet. This will result in considerable slowdown of user data transfer.

Research and development efforts are underway at IETF and other standards organizations to address the issues related to IP over satellite. For example, one proposed solution to the TCP retransmission problem is *spoofing*. Spoofing is a technique that provides a premature acknowledgment, that a TCP segment has been received. A "spoofing box" collects any duplicate acknowledgments to prevent confusion. Another modification to classic TCP that is under consideration is to have a larger window size to avoid the window size shrinking issue. Refer to IETF RFC 2760 for a detailed description of the issues relating to the proposed remedies for problems carrying IP over satellite (Allman, Dawkins et al., 2000).

13.3.3.3 Intersatellite Routing The choice of LEO satellites leads to the need to create and manage a large constellation of those satellites. The number of LEO satellites needed to ensure global coverage is in the range of 50 to 60. Traffic routing between satellites, or routing in the LEO constellation network, is an issue that is attracting increasing amounts of interest.

Intersatellite IP routing in a constellation network is driven by the need to support efficient IP multicast and IP QoS. Multicast allows a source to simultaneously send data to all users in a group. This requires either that multiple virtual "circuits" be established between the source and the multiple destinations, flooding messages to all users, or duplication of packets along the way. The bandwidth limitation of a satellite

network lies in the satellite-earth interface, as both terrestrial and constellation networks in themselves have large capacities. Intersatellite routing in a constellation network provides a more efficient use of the bandwidth, reducing the traffic on the satellite-earth interface. Another driving force behind the intersatellite routing is the need to support IP QoS. QoS via methods such as the DiffServ or IntServ models described in Chap. 18 will be supported in the satellite portion of the traffic flow in order to achieve end-to-end IP QoS.

REVIEW QUESTIONS

1. What are the differences between active and passive satellites?
2. Describe and compare LEO, MEO, and GEO satellites in terms of life span, orbit distance from the earth, transmission latency, etc.
3. Describe the two types of transmission media used in satellite broadband communications and compare the advantages and disadvantages of each.
4. Describe the main components of a satellite and the main functions the onboard processing unit performs.
5. Discuss the advantages of using free space optical links for intersatellite communications, and compare their use against that of radio frequency links. Describe the major hindrance to using optical links for the space-earth communications.
6. Describe the four distinct types of satellite communications systems—satellite broadcast to headend, satellite mobile voice network, satellite long-haul transport network, and satellite to home—in terms of the services supported by each type of network. Discuss especially whether two-way communications are supported, and whether the system reaches customer premises.
7. Compare the satellite long-haul network model and satellite to home model in terms of the data traffic each supports and whether satellites in the network extend their reach into customer premises.
8. Describe the ATM over satellite architecture and functions of the ASIU. Describe some of the reasons that ATM is chosen for broadband satellite networks.
9. Some early satellites are used to carry Internet to home traffic without a satellite uplink. Explain how a terrestrial connection can be used in place of a satellite uplink.

10. Describe the IP over satellite architecture and list the main issues with the TCP protocol caused by satellite link latency and jitter.
11. The satellite onboard processing unit is becoming more and more sophisticated to support packet routing in satellite constellation networks. Describe some of the reasons why constellation routing is needed.

REFERENCES

Allman, M., Dawkins, S., et al. 2000. "Ongoing TCP Research Related to Satellites." IETF RFC 2760. Web site: www.ietf.org.

Allman, M., Hayes, C., et al. 1999. "Enhancing TCP Over Satellite Channels Using Standard Mechanisms." IETF RFC 2488. Web site: www.ietf.org.

ITU-R. 1998a. "Availability Objectives for a HRDP When Used for the Transmission of B-ISDN ATM in the FSS." ITU-R draft Recommendation S.atm_av. Web site: www.itu.int/ITU-R/.

ITU-R. 1998b. "Performance for B-ISDN ATM via Satellite." ITU-R draft Recommendation S.atm. Web site: www.itu.int/ITU-R/.

Network Computing. 2001. "Networking in the 21st Century: The Sky Is the Limit." White paper. Web site: www.networkcomputing.com.

StarBand. 2000. "How Does Internet Satellite Operate?" White paper. Web site: www.starband.com.

Whalen, D. 1998. "Communications Satellites: Making the Global Village Possible." NASA white paper. Web site: www.hq.nasa.gov.

Wood, L. et al. 2000. "IP Routing Issues in Satellite Constellation Networks." *International Journal of Satellite Communications,* Vol. 18. No. 6.

WTECH. 1998. "Global Satellite Communications Technology and Systems." World Technology (WTECH) panel report for NASA. WTECH is a division of Loyola College. Web site: www.itri.loyola.edu.

CHAPTER 14

Passive Optical Networks

14.1 Introduction

Passive optical networks (PONs) are an optical broadband access technology that provide an optical last-mile solution. A PON is a point-to-multipoint optical network that uses passive optical components such as splitter, coupler, and splicer for out-of-plant components. An optical component is said to be *passive* if it does not require active power to function.

14.1.1 Brief History

In 1995, a group of vendors and telecom service providers, in an attempt to standardize PON access networks, formed the Full Service Access Network (FSAN) coalition. One goal of the FSAN coalition is to develop and standardize a cost-effective yet fast solution to create a "full service access network" that would extend emerging high-speed services, such as IP data, video, and 10/100 Ethernet, over fiber optics networks to residential and business customers worldwide (Spears 1999).

The FSAN coalition decided to adopt ATM over a passive optical network, known as ATM PON. In 1999, the ITU-T's Study Group 15 adopted the FSAN coalition's ATM PON specifications as standards G.983.1 (ITU-T, 1999).

An alternative, Ethernet PON technology, emerged while the ATM PON specifications were being finalized. The Ethernet PON efforts have been undertaken mainly by a group of startup companies and research institutes. Ethernet PON is gaining momentum as optical Ethernet (see Chap. 6) such as Gigabit Ethernet and 10 Gigabit Ethernet start taking hold in the metro network marketplace.

This chapter discusses both ATM and Ethernet PON.

14.1.2 PON Basics

PON technology has two essential characteristics that make it a good choice for a broadband access network:

- It supports point-to-multipoint architecture and allows multiple customers to share a single fiber facility. This is an essential requirement of an access network.
- A PON system does not require much maintenance because of the passive nature of outside plant components of a PON.

An access network is point-to-multipoint in nature, because it is prohibitively costly to have a point-to-point connection to each customer. So an important issue is how to build a fiber access network without having to dedicate one fiber to each customer. Passive equipment splits a single strand of fiber and allows the bandwidth to be shared without the use of more expensive "active" components like lasers. So multiple wavelengths, one per customer, can be combined onto a single fiber, which lowers the cost by spreading it among multiple customers.

A PON system uses passive optical components that require little maintenance and are relatively inexpensive compared to active components. This makes the initial network capital expenses and later maintenance costs sufficiently low, and make the PON model a viable last-mile solution (Eluminant 2000).

On the other hand, the fact that a PON does not regenerate optical signals also limits its reach. Without regeneration, light signals attenuate quickly. According to ITU G.983.1 (ITU-T 1999), PON systems have a theoretical distance limitation of 20 km.

14.1.3 PON Architecture

A PON network consists of three major components: an optical line terminator (OLT), an optical splitter and a combiner or a cascade of them, and a set of optical network units (ONUs) at the customer premises, as shown in Fig. 14-1 (LightReading 2000). This is a passive optical network because there are no active components between the OLT and ONUs.

14.1.3.1 Optical Line Terminal An OLT is a special-purpose switch located either at a service provider's central office (CO), a service provider's point of presence, or the headend of an optical cable network.

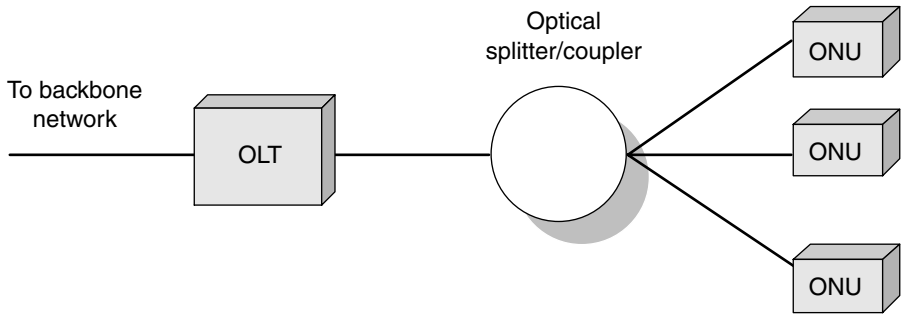


Figure 14-1
An example of PON architecture.

An OLT connects to one optical splitter and combiner or a cascade of them. The main functions of an OLT are as follows:

- It broadcasts the downstream data to the connected ONUs over a single fiber and sends traffic to a splitter that then splits and distribute the traffic to the designated ONUs at customer premises. The OLT either generates its own light signals or takes SONET signals (such as OC-12) from a colocated SONET cross-connect.
- It aggregates upstream traffic from multiple customer sites. It can use one of several multiplexing techniques such as TDM, CDM, or WDM to ensure that each transmission is sent back to the central office over one fiber strand. Upstream and downstream traffic use different frequencies on one fiber to avoid interference.
- It is responsible for interfacing metro backbone networks and performing network protocol conversions such as Ethernet to ATM or IP to ATM, as necessary.

14.1.3.2 Passive Optical Splitter and Coupler The passive splitter and coupler, the only passive components of a PON, are normally placed inside vaults in manholes, under the curbs of sidewalks, or in enforced outdoor cabinets near office parks. They are placed outside the controlled environments of locations like central offices or customer premises. Outside equipment is normally expensive to maintain, but passive optical components like splitters and couplers require little maintenance.

The main function of these passive components is to split optical signals going from an OLT to ONUs and combine them in the other direction. In the downstream direction, as the light broadcast from an OLT hits the splitter, it is deflected onto multiple fiber connections, depending on the splitter used. A splitter may branch from 2 to 32 or 64 branches.

14.1.3.3 Optical Network Unit PON systems terminates at so-called ONUs, also known as *optical network terminals* (ONTs), as shown in Fig. 14-1. An ONU takes in light that is sent from a passive splitter, converts it to specific types of bandwidth (such as 10/100-Mbit/s Ethernet, ATM, and T1 voice and data), and passes it on to enterprise routers, PBXs, switches, or residential homes. ONUs have an active optical component such as a laser or LED to send optical signals back to the central office at the command of the OLT. Given that the bandwidth allocated to an ONU is merely a branch of that of the OLT, an inexpensive LED can be a good choice. The general functions an ONU performs include the following:

- It converts optical signals to electrical signals and then to an application bandwidth for a user end terminal such as a PC, a TV, a home hub, enterprise routers, etc.
- It generates optical signals for upstream traffic.
- It interfaces last-yard access devices like a wireless LAN card or asynchronous digital subscriber loop (ADSL) line at customer premises.

ONUs can be installed directly in wiring closets or the data centers of customer premises. Alternatively, they can be placed outside plant locations, so that customers can hook into a PON from digital subscriber line (DSL) services. This approach gives customers the benefit of optical networking without needing new fiber going into their premises.

14.2 ATM PON

It was not until the mid-1990s that PONs was seriously considered for optical access. An ATM PON specification was proposed in 1998 and was adopted as ITU recommendation G.983.1 in 1999 (ITU-T, 1999).

14.2.1 ATM PON Architecture

An ATM PON carries all services in ATM cells, in both upstream and downstream directions. One main consideration for adopting ATM PON is the guaranteed bandwidth and well-defined QoS provided by ATM technology (see Chap. 3).

An ATM PON system, as described in Sec. 14.1, consists of an OTL, one or more optical splitter/couplers, and a set of ONTs, as shown in Fig. 14-2. One optical line terminal can serve up to 32 optical network terminals. The G.983.1 standard calls for a minimum reach of 20 km from an OLT to an ONT and an optical power budget consistent with a maximum split ratio of 32 (ITU-T, 1999).

An ATM PON can operate in either simplex or duplex mode. In simplex mode, a single wavelength, either 1510 or 1310 nm, is used for both upstream and downstream traffic. In duplex mode, simultaneous upstream and downstream transmission is achieved using independent wavelengths for each direction on a single fiber facility.

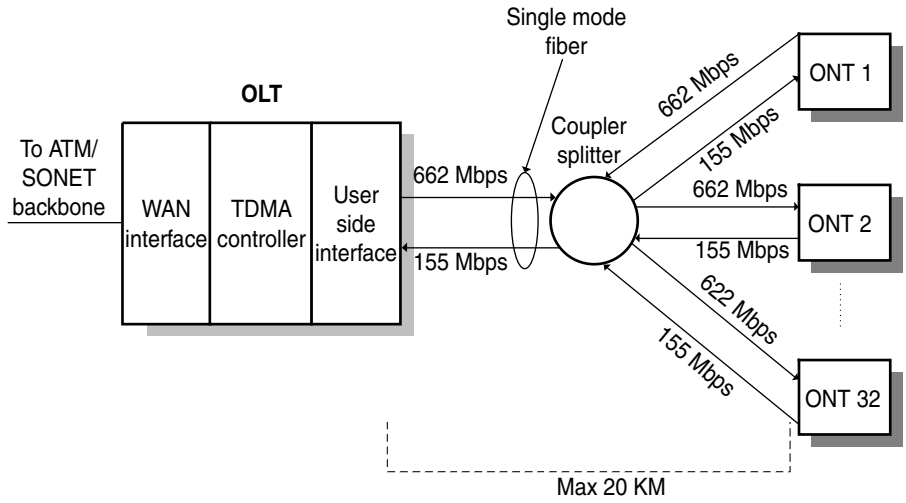


Figure 14-2
ATM PON architecture.

An ATM PON can be symmetric or asymmetric. An asymmetric PON has a downstream data rate of 622 Mbps and an upstream data rate of 155 Mbps while a symmetric PON has a data rate of 155 Mbps in both directions, as defined in G.983.1 (ITU-T, 1999).

In the downstream direction, data is broadcast in a continuous stream of ATM cells to all subtending ONTs. Only one transceiver is defined in an ATM-based OLT, and the optical signals from a transmitter are passively split into up to 32 ONTs. Each ONT identifies the cells addressed to it by inspecting each cell header. This is similar to cable modem's downstream operation, and the basic idea is illustrated in Fig. 14-3.

The ATM concept of virtual path identifier is used to set up a virtual path between an OLT and each subtending ONT. Although the downstream data stream is broadcast to all connected ONTs, an ONT only processes the cells with the VPI values that have been explicitly assigned to this ONT. For example, VPI = 4 is assigned to the bottom ONT and VPI = 3 is assigned to the top ONT, as shown in Fig. 14-3. This effectively establishes a point-to-point virtual path connection between an OLT and an ONT. A point-to-multipoint virtual path connection from an OLT to a selected set of ONTs can be established by assigning the same VPI to all selected ONTs.

The upstream data flow from an ONT to the OLT is different from the downstream broadcast, a key difference being that multiple ONTs share the same wavelength with different user data. ITU-T G.983.1 (ITU-T, 1999) specifies a TDMA for ATM PON to ensure that ATM cells from multiple ONTs are interleaved without interference from each other. Each

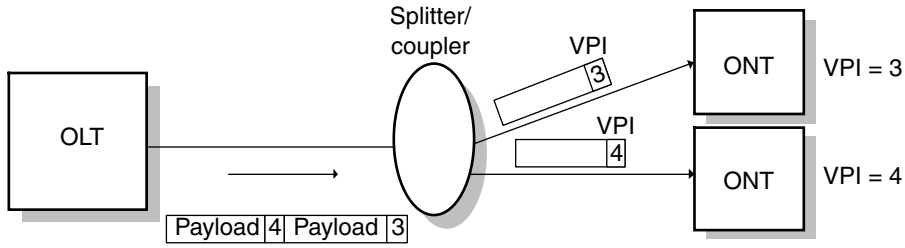


Figure 14-3
A VP connection example for downstream data. (ITU-T, 1999).

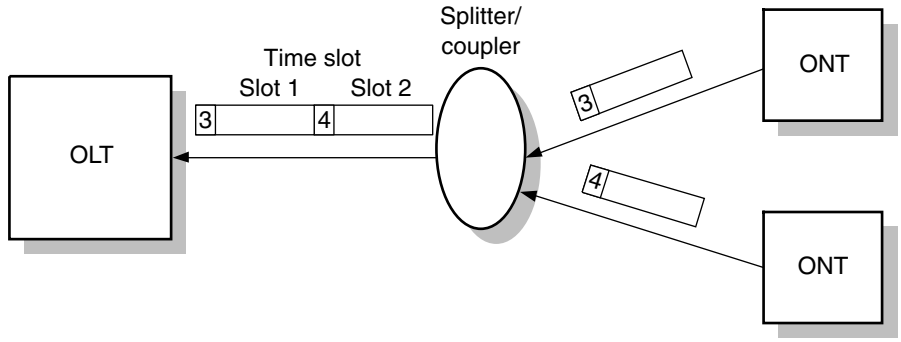


Figure 14-4
A time division access control example for up stream traffic in ATM PON.

ONT is assigned a time slot to transmit data in the upstream direction. An ONT transmits a 3-byte header followed by a 53-byte ATM cell at a line rate of 155 Mbps. Taking away the overhead cells, the actual user data payload capacity is around 147 Mbps with the basic idea of TDMA illustrated in Fig. 14-4 (Hogg 1999).

In addition to TDMA time slots for upstream access control, a minislot mechanism is defined in G.983.1 (ITU-T, 1999) to subdivide a time slot into multiple minislots to enable dynamic bandwidth control. Multiple ONTs can share a single time slot by using designated minislots.

The control of multiple accesses resides with the OLT that determines when each ONT can start transmitting upstream ATM cells. The OLT communicates the control information to the connected ONTs via downstream control cells. This way, the OLT can regulate the time and rate at which each ONT can send cells upstream. The ITU G983.1 recommendations (ITU-T, 1999) specify the details of the TDMA protocol to promote the interoperability between OLT and ONT products from different vendors.

14.2.2 Security and Privacy

Because the downstream data is broadcast from an OLT to all connected ONTs, there is always a possibility of eavesdropping on other downstream

signals when an ONT is located at a home or a business. The G.983.1 PON recommendation (ITU-T, 1999) specifies a simple form of data scrambling known as *churning* to protect the privacy of signals destined for an ONT. A byte-oriented encoding scheme based on a private key is exchanged between a given ONT and the OLT, and the key is updated on a periodic basis to ensure the security.

14.3 Ethernet PON

Ethernet PON, which uses Ethernet as opposed to ATM as the choice for the layer-2 technology of a PON system, is a fairly new and emerging technology. The standardization efforts are at an early stage, and the technical details have yet to be fully specified. Part of the standardization efforts for Ethernet PON is the forming of the "Ethernet in the First Mile" study group under the auspices of IEEE 802.3 (Kramer 2001; Pesavento and Kelsey 2001).

The main motivations for developing Ethernet PON are the following:

- As the large-scale deployment of optical Ethernet like Gigabit Ethernet and 10 Gigabit Ethernet spreads into metro backbone space, Ethernet PON is just a natural extension.
- Ethernet PON offers cost advantages over ATM/SONET-based access networks, at least in theory, because of the simplicity and ubiquity of Ethernet technology. Low cost is a key factor in the acceptance of any new broadband access network technology.

14.3.1 Ethernet PON Architecture

Ethernet PONs carry all services in variable-length Ethernet frames, each of which has a maximum length of 1518 bytes, in both upstream and downstream directions.

The overall architecture of Ethernet PON looks quite similar to that of ATM PON, as shown in Fig. 14-5: One OLT serves up to N ONUs. The line rates are vendor-specific at this stage of the technology, ranging from 622 Mbps to over 1 Gigabit. The minimum reach of the optical line between an OLT and an ONU is roughly the same as that of ATM PON, given the power budget and target splits at the passive splitter and coupler.

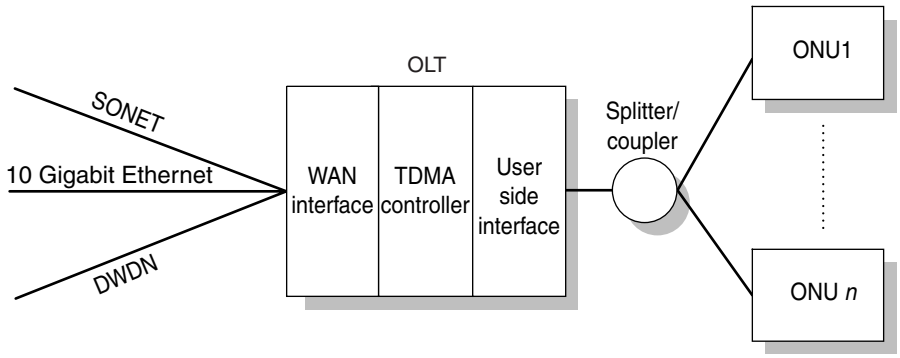


Figure 14-5
Ethernet PON architecture.

On the network side, an Ethernet PON OLT can be connected to a 10 Gigabit Ethernet network, a SONET network, or a DWDM network.

An Ethernet PON, like an ATM PON, can operate in either simplex or duplex mode. In simplex mode, a single wavelength, either 1510 or 1310 nm, is used for both upstream and downstream traffic. In duplex mode, simultaneous upstream and downstream transmission is achieved using independent wavelengths for each direction on a single fiber facility. Using more than two wavelengths to support services such as radio frequency video to carry cable TV programming has also been proposed.

In the downstream direction, data is broadcast in a continuous stream of variable-length Ethernet packets to all subtending ONUs. A MAC address-like scheme has been proposed to uniquely identify each ONU among all the connected ONUs. For a point-to-point connection, the OLT uses a unique address in each PON frame header to address a particular ONU. An ONU determines whether to accept a packet based on a match between the address in the frame header and its own address. For broadcast data, an address known to all subtending ONUs achieves the purpose.

It has been proposed that Ethernet PON also use a TDMA control mechanism. Each ONU is assigned a time slot to transmit data in the upstream direction. An upstream transmission frame contains one time slot for each connected ONU, and thus consists of N time slots where N is the number of subtending ONUs.

14.3.2 Security and Privacy

No security mechanism has yet been defined specifically for Ethernet PON. However, generic IP security mechanisms such as virtual LAN, IP Security (IPSec), and tunneling will be applicable.

14.3.3 Comparisons Between ATM PON and Ethernet PON

Both ATM and Ethernet PON technologies have advantages and disadvantages. ATM PON has a longer history and thus more mature standards. Among its chief advantages are QoS support and the built-in virtual path mechanism for distinguishing among ONUs. Its main disadvantages include its complexity and potentially high cost.

There is a momentum building up for Ethernet PON, largely because of its simplicity, its natural fit to LAN architecture, and the emerging optical Ethernet (Gigabit and 10 Gigabit Ethernet) for metro area networks. The potential low cost is another consideration, which is crucial to mass deployment of any technology in access network space.

Table 14-1 compares the two PON technologies.

14.4 Applications of Passive Optical Networks

Passive optical network technology is a promising last-mile solution for broadband access networks, because of its support for point-to-multipoint transmission, its large amount of bandwidth, and the low cost of the passive optical components and their maintenance. This section describes PON system configurations and different application scenarios.

14.4.1 PON Configurations

A PON system can be configured in a ring, a tree, or a bus topology. In fact, the configurations of the OLT and the optical splitter/coupler in a PON system are the same for all topologies; it is the ONUs and the optical splitter/coupler that can be configured in different topologies, as shown in Fig. 14-6.

In a ring configuration, the splitter/coupler and the ONUs form a ring with the splitter as the starting point of the ring. In a tree topology, as shown in the middle of Fig. 14-6, the splitter has a one-to-one connection to each ONU. The bus topology is similar to the traditional Ethernet configuration, as shown at the bottom of Fig. 14-6.

Table 14-1 A Simple Comparison Between ATM and Ethernet PON

	ATM PON	Ethernet PON
Technology complexity	ATM theoretically more complicated than Ethernet and more costly	Simple plug and play technology; cost presumably low
QoS	Excellent QoS support	No built-in QoS mechanism
Data rate	Maximum 622 Mbps downstream and 155 Mbps upstream	No standard bandwidth defined yet; but generally will be higher because of smaller overhead and newer fiber optics technologies
Approach to carrying IP traffic	IP over ATM	IP over Ethernet
Mechanism for distinguishing one ONU from another	ATM VPI	Ethernet MAC address-like mechanism
Multiple access control	TDMA	TDMA (proposal only)
WAN interface support	Compatible with SONET and ATM backbone	Compatible with optical Ethernet (Gigabit and 10 Gigabit Ethernet), SONET, and DWDM backbone
LAN interface support	Excellent for voice traffic, but need conversion from ATM cell to a local IP frame, most likely Ethernet	Natural fit for connecting to LANs, most of which are Ethernet-based; a challenge to provide high-quality voice service
Supported services	Voice, data, and video	Voice, data, and video
Standardization efforts	The first PON technology; standards defined	Standards effort at initial stage
Deployment	Initial field trial stage	Very initial trial stage

14.4.2 PON Application Scenarios

PON can extend the reach of fiber to different locations within the last mile to accommodate differing application scenarios. Depending on where an ONU of a PON system is located, there are four "fiber-to-the" (FTT*x*) scenarios:

- Fiber to the curb (FTTC), where an ONU resides at a roadside wire center
- Fiber to the building (FTTB), where an ONU resides inside a telecommunications closet of a multitenant business or residential building

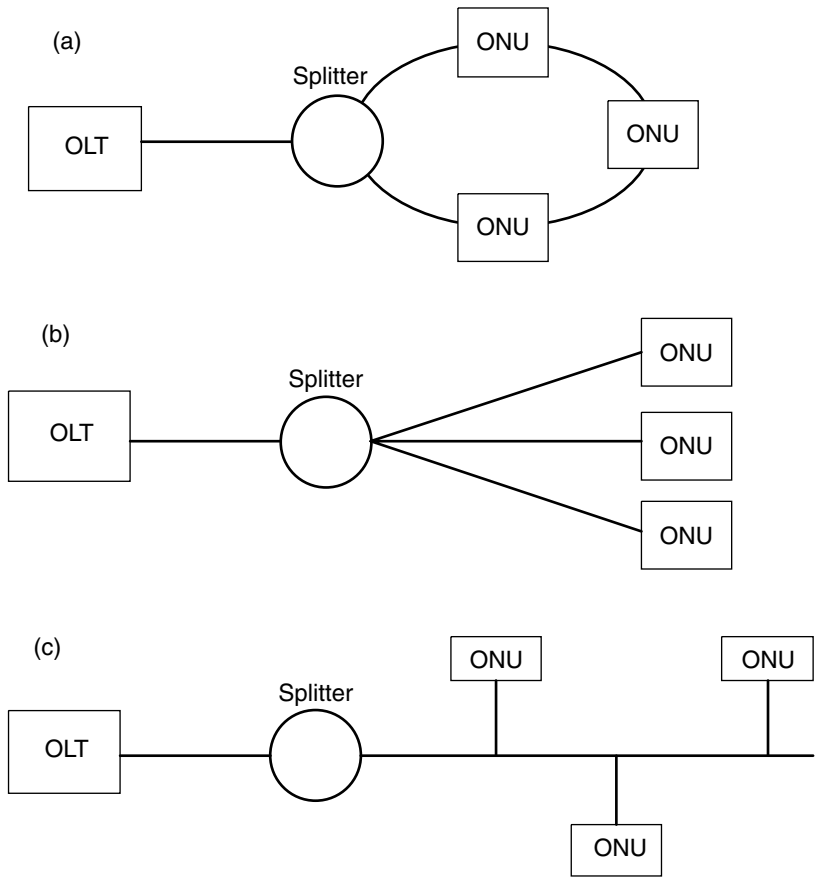


Figure 14-6
Three PON topologies.

- Fiber to the floor (FTTF), where an ONU is located at a wire cabinet of a building floor
- Fiber to the home (FTTH), where an ONU is located inside a residential customer's home

Fiber cable progressively inches closer to end customer premises equipment in the above application scenarios. Judging by the fact that it took a long period of time for digital loop carrier lines to replace analog lines in telephony access networks, fiber optical access network will also take some time to reach end consumers on a mass scale. It is economically more viable to deploy fiber in an incremental manner, starting from FTTB, FTTC, and moving toward FTTH.

14.4.3 PON Deployment Scenarios

A PON system can be deployed in what is called a *greenfield* situation, where there is no preexisting telecom infrastructure, or into an existing copper-based telephony access network.

14.4.3.1 Greenfield Deployment With the advances in fiber optical technologies and an emerging mass market, the cost of optical fiber has come down drastically in recent years. For new office buildings or residential areas, optical fiber compared to copper wire or coaxial cable, is becoming a very attractive network choice.

PON is an attractive option for the greenfield deployment because of its low maintenance cost and its ability to support multiple users sharing a single fiber from a central office. Without the constraints of existing infrastructure, the FTTH that has an ONU terminating at an individual residential home or a small office is fast becoming a viable option.

14.4.3.2 Colocation with Existing Copper Access Networks In the majority of cases, the existing telecom infrastructure needs to be taken into consideration in the deployment of a PON system.

One deployment scenario is to use a PON system as a feeder technology. In this scenario, a PON is used to upgrade the connection between a central hub point and an out-of-plant multiple service distribution point. Examples of central hub points include central offices in local exchange networks and headends in cable networks. Out-of-plant service distribution points can be integrated access devices or digital loop carrier terminals. ONUs can be colocated at out-of-plant service distribution points.

Another deployment scenario is to lay a PON system over an existing DLC telephony access network, as shown in the top half of Fig. 14-7. In this deployment scenario, the OLT is colocated with a remote digital terminal (RDT) at a central office. Many telephony digital access networks already have fiber installed, and PON makes use of that installed fiber facility. The ONUs are colocated with DLC access units, often found at the end of residential streets. PON optical signals terminate here, and user data is carried over the existing twisted pairs between the access unit and the premises of each customer. A second deployment scenario is illustrated in the bottom half of Fig. 14-7: A PON system terminates where a very high digital subscriber line (VDSL) system starts. Residential

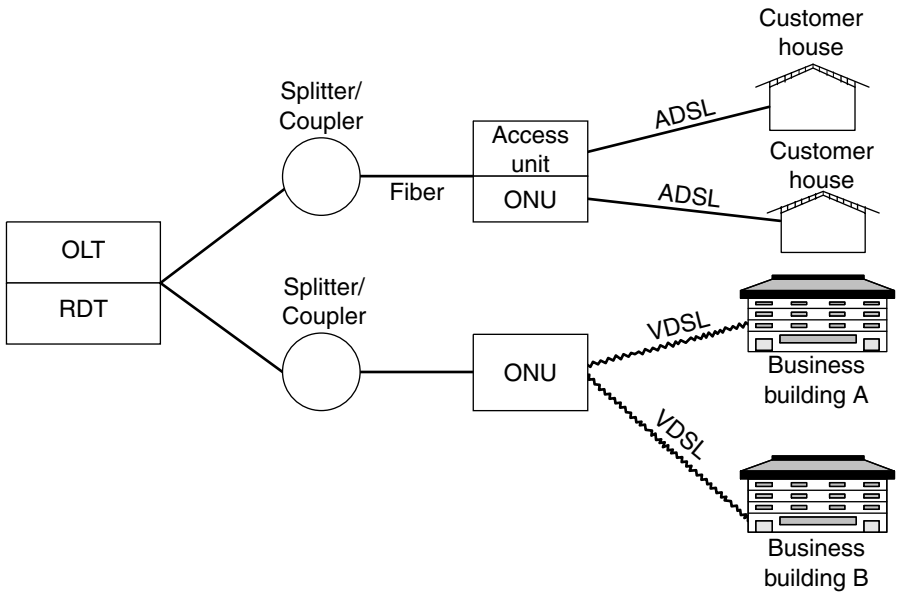


Figure 14-7
PON deployment scenario—overlay with DLC.

broadband technology over copper such as VDSL carries data to business customer premises.

REVIEW QUESTIONS

1. Describe what characterizes a PON system and the motivations for using PON as a last-mile solution.
2. Describe the three components of a PON system and the passive part of a PON system.
3. Describe the functions of an OLT, ONU, and optical splitter and where each component is normally located.
4. Describe how ATM PON distinguishes one ONT from another and the mechanism for multiple media access control.
5. Describe the asynchronous versus synchronous transmission modes and the simplex versus duplex operation modes of an ATM PON.
6. Describe the operations of ATM PON data transmission in both downstream and the upstream directions.
7. Discuss the motivations for the development of Ethernet PON and its advantages.

8. Describe the scheme proposed for multiple medium access control for Ethernet PON and discuss its advantages and disadvantages.
9. Discuss the three topologies of a PON system as described in this chapter and discuss the advantages of each.
10. Describe the four application scenarios of PON, dubbed FTT*x*, and discuss them in the context of the two deployment scenarios presented in this chapter.

REFERENCES

Eluminant. 2000. "Passive Optical Networks Tutorial." NEC Eluminant Technologies. White paper. Web site: www.eluminant.com.

Hogg, R. 1999. "ATM PON Maximizes Bandwidth to Homes and Business." *Lightwave*. August. Web site: www.pennet.com.

ITU-T. 1999. "ATM Passive Optical Network Specification." Recommendation G.983.1.

Kramer, G., et al. 2001. "Multiple Access Techniques for ePON." Presented at IEEE 802.3 "Ethernet in the First Mile" Study Group. Web site: www.ucdavis.edu.

LightReading. 2000. "A PON Primer." On-line tutorial. Web site: www.lightreading.com.

Spears, D., Ford, B., et al. 1999. "FSAN Initiative Propels Broadband Access Worldwide." *Lightwave*. Sept. Web site: www.pennet.com.

Pesavento, G., and Kelsey, M. 2001. "Ethernet in the First Mile." *Lightwave*. June. Web site: www.penet.com.

CHAPTER 15

Digital Subscriber Lines

15.1 Introduction to Local Loop

Digital subscriber line is a general term for a family of technologies that provides a broadband last-mile solution using existing copper wire-based local loop infrastructure. This section discusses existing local loop infrastructure to provide a context for the DSL technology.

The fundamental structure of telephone networks has remained the same for decades. The telephone access network, also known as the *local loop*, connects subscribers to the switching equipment of a local phone company, as shown in Fig. 15-1. There are three general configurations of local loops, also shown in Fig. 15-1: residential homes directly connected to a central office; residential homes connected to a remote digital terminal, which in turn is connected to a CO; and enterprise customers connected to a CO via a PBX (Cioffi, Silverman, and Starr 1999).

In the first local loop configuration shown at the top of Fig. 15-1, individual copper wires from residential homes terminate at an intermediate point, known as the *digital loop carrier* or *remote digital terminal*. An RDT performs two primary functions. First, it reduces the effective length of the copper running out of the central office, thus improving the reliability of the service. Second, it aggregates the traffic from individual homes onto a high-capacity transmission line such as T1, DS3, or SONET OC3, thus reducing the number of lines coming out of the central office. The overwhelming majority of installed bases still use copper wire between their central offices and RDTs.

The second local loop configuration, shown as the middle part of Fig. 15-1, directly connects a private branch exchange at an enterprise premise to a central office using four copper wires.

In the third local loop configuration, shown at the bottom of Fig. 15-1, copper wire phone lines come out of a central office as bundles. The local loop is divided into customer service areas (CSAs) that are serviced by remote terminals (RTs). A CSA is then further divided into smaller distribution areas each of which is serviced by a feeder distribution interface (FDI). An FDI can serve up to 500 phone lines. The FDI is the last point where the phone lines are still bundled. At the FDI, the phone lines are unbundled and each individual line goes to a residential home. This local loop configuration was originally designed to support Plain Old Telephony Service services only.

Copper wire accounts for the majority of the connections between the COs and residential homes or enterprises. More than 95 percent of local access loops consist of a single pair (two-wire circuit) of twisted wires for POTS.

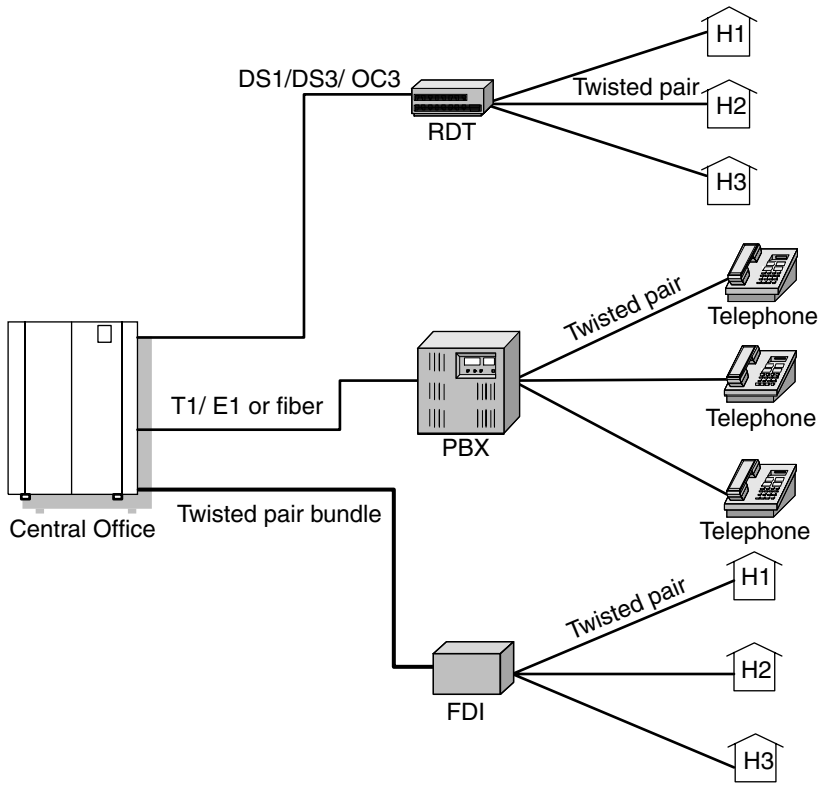


Figure 15-1
Copper wire-based local loop configuration.

Copper wire has a limited reach because signal dissipation makes the signal unusable beyond the specified distance range. Two methods widely used to extend the reach of copper wire are thick wire and loading coil. Everything else being equal, the thinner the copper wire, the shorter the distance the signal can travel. Thus thicker wires are used to extend the reach from central offices to homes. A loading coil is an inductor formed around an iron core in the shape of either a horseshoe or a cylinder that has the effect of improving the range and quality of a phone line. The twisted-pair copper wire can have an effective distance over 18,000 ft if loading coil is placed every 6000 ft on the line.

15.2 DSL Basics

It is estimated that by the end of the 20th century there was an installed base in the world of over 700 million lines of copper wire from homes

to central offices for POTS, with two-thirds of that located in the United States. It is this huge installed base of copper wire that DSL technology is developed to leverage.

This section provides an overview of DSL, describes how DSL technology works, and introduces the family of DSL technologies.

15.2.1 DSL Overview

The DSL concept was originally developed by telephone companies in the late 1980s to provide video-on-demand service in response to the efforts by cable companies to provide telephone service via coax cables. The technology was complex and costly at the time, and the envisioned video service market did not materialize. In second half of the 1990s, with the invention of discrete multitone (DMT) line code and the explosive growth of Internet traffic, tremendous pressure was put on the existing copper wire-based local access loops. DSL technology then refocused to provide Internet-based data services (DSL Forum 1997; DSL Forum 2000).

In a nutshell, DSL technology can be summarized as follows:

- DSL operates on the existing twisted pair copper wire infrastructure to achieve much higher data rates up to 7 Mbps, more than a 100-fold increase over the data rate achieved by POTS lines. This enables service providers to provide fast access to the Internet at far less cost than would be incurred by laying fiber.
- DSL leaves the existing POTS undisturbed while providing residential users with broadband access to the Internet.
- DSL is a family of technologies that allows flexible bandwidth allocation based on user demand. It can be symmetric or asymmetric. Symmetric DSL provides equal bandwidth in both upstream and downstream directions, while asymmetric DSL provides higher bandwidth in the downstream direction than in the upstream direction.
- DSL, unlike cable modem technology, offers dedicated bandwidth, and the number of simultaneous users does not affect user access speed because users do not share the resources.
- Unlike a dial-up connection, a DSL connection is "always on."

The international standardization efforts on DSL started in late 1990s, and the ITU G.99x.x series of specifications (ITU-T, 1999a, 1999b, 2001a, 2001b, 2001c) are dedicated to DSL technology.

15.2.2 Key Concepts of DSL Technology

The key concepts of DSL technology consist of three parts: the use of high-frequency bands for digital signal transmission, the approaches to a set of issues caused by the use of high-frequency bands, and the adoption of the dense modulation technique (Rhodes et al. 2001; Paradyne 1998).

The use of the high-frequency spectrum to achieve high data rates creates a set of problems. It is the way that DSL addresses that set of issues which defines the characteristics of DSL technology. The main problems include the following:

- Crosstalk
- Impulse noise
- Radio noise
- Faster signal attenuation
- Bridge taps

15.2.2.1 DSL and High-Frequency Bands The fundamental idea of DSL technology is to use high-frequency electric pulses to increase data transmission rates over copper wire. Note that the rate of the electric pulse traveling on the copper wire ultimately determines the data transmission rate. The low rate of 56 Kbps of POTS line is due to the fact that POTS lines are constrained to 43 KHz of frequency.

DSL uses the frequency band above that of POTS, starting from 25 KHz and extending all the way to over 2 MHz. The specific spectrum band used can vary from one DSL vendor to another. For an asymmetric DSL system such as asynchronous DSL (ADSL), different frequency spectrums are used for upstream transmission (from a user to the central office) and downstream transmission (from the central office to a user). Typically, the upstream frequency band is from 25 to 200 KHz and the downstream band is from 200 KHz to 1.1 MHz. There is a frequency guard band between the POTS spectrum and the DSL spectrum to avoid interference between them. This is how DSL achieves high bandwidth without disturbing traditional telephone service.

15.2.2.2 Signal Modulation of DSL Another cornerstone of DSL technology is modulation technology. Frequency modulation, in its simplest form, can be viewed as a process of "shaping" the electromagnetic wave into a form suitable to carry digital signals (0s and 1s). Thus the modulation technique plays a key role in determining the data rate

of a carrier. One modulation technique used in the DSL physical layer is discrete multitone. A DMT system transmits data on multiple subcarriers in a manner similar to the orthogonal frequency division multiplexing technique used in advanced wireless systems. A DMT modulator takes N inputs of data symbols in parallel and transmits the symbols on N subcarriers (Paradyne 1998).

15.2.2.3 DSL Approach to Crosstalk *Crosstalk* refers to interference affecting intended signals produced by unintended signals. It is caused by the electromagnetic radiation from the wires bundled together. There are two basic types of crosstalk, both of which appear at the receiver end as interfering noise: near-end and far-end.

Near-end crosstalk appears when a transmitter interferes with a receiver located on the same end of cable. Far-end crosstalk occurs when the transmitter interferes with a receiver on the opposite end of the transmission line. Normally far-end crosstalk interference is less of an issue than near-end crosstalk interference because the noise is substantially weakened after it traverses the whole length of the line.

DSL technology tackles the crosstalk problem from two perspectives: the dense modulation technique and spectrum allocation. As mentioned above, a DMT modulator takes N inputs of data symbols in parallel and transmits them on N subcarriers. The data transmission rate on each subcarrier is $1/N$ of the original data transmission rate. The symbols transmitted through N subcarriers have an orthogonal relationship so that the interference which occurs between symbols is minimal.

Spectrum allocation is another way to reduce the effect of crosstalk. Some DSL systems use different frequency spectrums to transmit and receive signals. This eliminates near-end crosstalk because a transmitter and receiver adjacent to each other operate at different frequency bands. Far-end crosstalk is less of an issue to begin with and is further reduced because when separate frequency bands are used they reduce the chance of interference occurring.

15.2.2.4 DSL Approach to Impulse Noise Interference that is short in duration but large in magnitude is known as *impulse noise*. Impulse noise can be caused by lightning or power surges produced by things like the startup of a motor engine. Impulse noise, like crosstalk, can cause interference and increase the data transmission error rate.

DSL technology combines an interleaving technique with a signal coding scheme to correct data errors caused by impulse noise. Interleaving is a process of rearranging the data in such a way that data bits located

contiguously in time are placed apart in the transmission and then put back together at the receiving end. Interleaving combined with a digital signal coding scheme can spread these errors in time and thus lessen their impact. A burst of errors can cause more of a problem for a receiver than the same number of errors spread out over time.

15.2.2.5 DSL Approach to Radio Noise Radio noise is the interference caused by a wireless source. Copper phone lines act as antennas and pick up undesired signals. The common sources of radio noise are AM and FM radios since their spectrums overlap the DSL spectrum. This is much less of an issue for POTS systems due to their low frequency bands.

There are several techniques at the disposal of DSL to deal with radio noise. One is the use of adaptive radio frequency cancellation filters to filter out the interfering radio noise. Another is to use dynamic bit allocation to turn off subcarriers near the frequencies of interference.

15.2.2.6 Faster Signal Attenuation High-frequency signals attenuate faster than low-frequency signals much like a car burning gas faster at a high speed than at a lower speed. One way to overcome the faster signal attenuation is to use a wire with less resistance so the signals can travel longer distance. Thicker wires normally have less resistance than the thin wires.

15.2.2.7 Bridge Taps *Bridge taps* refers to the sections of wire that do not lie in the direction of the communication path, such as opening and closing cable slices or the section of wire that is connected to the loop at one end but not terminated at the other end. In a DSL system, when transmitted signals arrive at a bridge tap, the signal divides. While part of the signal energy continues on to the receiving end, part of it is diverted to the unterminated end. The signals are delayed and multiple versions of the same signal, or *self-interference*, are created, very much like the case of the multipath effect in wireless systems.

In addition to using techniques such as a frequency-domain equalizer to overcome the interference caused by bridge taps, a straightforward solution is simply to remove the bridge taps. But this seemingly trivial task can be time-consuming and costly.

15.2.2.8 Other Issues for DSL DSL systems run on the existing copper wire networks, but not all existing copper wires can support DSL. There are some constraints for the deployment of DSL system on an existing local loop, which include loading coils, transmission distances, and wire conditions.

To extend the reach of copper twisted wires, a widely used technique by local phone companies is to insert loading coil into the wire. Loading coil must be removed in order for the DSL modem to work. However, once the coil is removed, the reach of the wire may fall short of the required distance and rewiring may be required.

The length of the local loop is another consideration. The maximum reach of most DSL systems is 18,000 ft or less. In general, the higher the data transmission rate, the shorter the transmission distance of the DSL system. To extend the distance, a repeater can be used on the DSL line. The condition of an existing copper wire is yet another consideration for DSL deployment. The poor quality of a line may make it unsuitable for DSL use.

15.2.3 DSL Family

A DSL system is mainly characterized from the following dimensions: data throughput, transmission distance, whether the transmission is symmetric, and the target application. The data throughput of a DSL system is largely determined by the transmission frequency band used on the copper wire and by modulation techniques. A DSL system is said to be symmetric if the downstream and upstream transmission rates are the same. Some DSL systems are developed for business applications, some for residential broadband access, and others for both (Goralski 1998).

15.2.3.1 Asymmetric Digital Subscriber Line ADSL is one of the most widely deployed types of DSL technology. The data transmission rates for downstream and upstream are different, the downstream rate ranging between 1.5 and 6 Mbps and the upstream rate up to 640 Kbps. ADSLs are targeted for both residential and small business customers.

15.2.3.2 DSL Lite or G.Lite Commonly known as *G.lite*, this type of DSL is a simplified version of ADSL. It has a lower data rate than ADSL and does not require splitting at the customer end of the POTS line. Splitting is a process of splitting one incoming signal into different signals according to the different transmission frequencies, as will be explained shortly. G.Lite, officially ITU-T standard G.992.2 (ITU-T, 1999b), provides a data rate of up to 1.544 Mbps in the downstream direction and from 128 to 512 Kbps in the upstream direction. Transmission

distances for G.Lite can reach 18,000 ft (about $5^1/_2$ mi) on 24-gauge copper wire. It is expected that G.Lite will enjoy very large-scale deployment in the residential market.

15.2.3.3 High Bit Rate DSL High bit rate DSL (HDSL) is an earlier version of DSL and is symmetric in the transmission data rates for both downstream and upstream directions. It is a mature technology that has been in use for years, with a large installed base before ADSL came on the scene.

HDSL devices can transmit data over a single copper twisted line at a transmission rate up to 768 Kbps in both directions or over two twisted pairs at a T1-equivalent transmission rate of 1.5 Mbps. At these speeds, HDSL service can reach up to 12,000 ft from the central office to the service point to the customers. The service distance can be extended to 18,000 ft for a data transmission rate of 384 Kbps. The target applications of HDSL are for small businesses that require symmetric communication capabilities such as enterprise intranets, high-volume email, electronic commerce, and videoconferencing.

A major drawback of HDSL is that it does not support POTS overlay. A separate pair of copper wires is needed for voice.

15.2.3.4 Symmetric DSL Symmetric DSL (SDSL) is very similar to HDSL, with a symmetric data transmission rate of 1.544 Mbps for North America and 2.048 Mbps for Europe in either direction on a duplex line. Part of the targeted market for both HDSL and SDSL is the replacement of existing T1/E1 service.

15.2.3.5 Very High Data Rate DSL Very high data rate DSL (VDSL) is an evolving technology that aims to provide much higher data transmission rates over relatively short distances. The envisioned data rates are between 51 and 55 Mbps for downstream and 1.6 and 2.3 Mbps upstream over a distance of 1000 ft. VDSL is viewed as future technology for video applications that require connection to a fiber loop.

15.2.3.6 Other Variants There are some other DSL variants that are either vendor-specific or in the early stages of development. They include Unidirectional DSL (UDSL), Rate-Adaptive DSL (RADSL), ISDN DSL (IDSL), and Consumer DSL (CDSL).

The various DSL systems now in general use are summarized in Table 15-1.

TABLE 15-1

Summary of DSL Variants

DSL type	Data transmission rate	Distance limit	Main applications
IDSL	128 Kbps in both directions	18,000 ft on 24-gauge wire	Similar to ISDN BRI but no voice service
CDSL	1 Mbps downstream; less upstream	18,000 ft on 24-gauge wire	Splitterless home and small office data service
G.Lite	1.544 to 6 Mbps downstream	18,000 ft on 24-gauge wire	Splitterless DSL; simplified ADSL
HDSL	1.544 Mbps duplex on 2 twisted pair lines; 2.04 Mbps duplex on 3 twisted pair lines	12,000 ft on 24-gauge wire	T1/E1 service replacement
SDSL	1.544 Mbps duplex (North America); 2.04 Mbps (Europe) on a single duplex line downstream	12,000 ft on 24-gauge wire	T1/E1 service replacement
ADSL	1.544 to 6.1 Mbps downstream; up to 640 kbps upstream	1.54 Mbps at 18,000 ft; 6.312 Mbps at 12,000 ft	Residential and small business Internet access and multimedia services
RADSL	640 Kbps to 2.2 Mbps downstream; 272 Kbps to 1.88 Mbps upstream	…	Internet access service for residential and small enterprise customers
VDSL	12.9 to 52.8 Mbps downstream; 1.6 Mbps to 2.3 Mbps upstream	4500 ft at 12.96 Mbps	Connections to fiber-based networks

15.3 DSL Network Components

A DSL network overlays the existing copper wire access loop, consisting of the network side components and customer premise components, as shown in Fig. 15-2. A DSL access multiplexer (DSLAM) is the main component on the network side.

15.3.1 Digital Subscriber Line Access Multiplexer

A DSLAM normally resides at the central office and is the cornerstone of a DSL system. The functions of the DSLAM have evolved since the

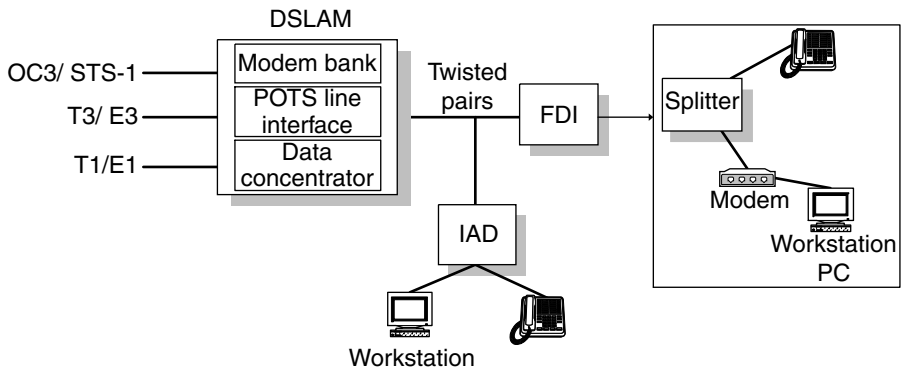


Figure 15-2
DSL network components.

beginning of the deployment of DSL systems in the mid-1990s. A typical DSLAM has the following main components, as shown in Fig. 15-2:

- *Modem bank*. There is one modem per provisioned customer to perform data transmission and reception, conversion between digital and analog signals, and signal modulation.
- *POTS line interface*. The interface for the traditional phone service using analog signals if the fan-out point is an FDI or using digital signals if the fan-out point is a DLC.
- *Data traffic concentrator*. This module aggregates the application data traffic from the connected DSL users onto a high capacity link. The application data can be ATM cell, frame relay cell, IP packets in Ethernet frames, or another format. A DSLAM, depending on the specific configuration, may perform the conversion from one data format to another, such as Ethernet frame to ATM cell, for example.
- *Backbone network interface*. These transmission interfaces for the DSLAM include T1/E1, T3/E3, and OC3/STS-1 links connected to a backbone network device such as a router or a central office switch.

15.3.2 DSL Customer Premises Equipment

There are two types of the customer premise equipment (CPE): residential and enterprise, as shown in Fig. 15-2. Residential CPE includes a POTS splitter and DSL modem, while enterprise CPE includes an integrated access device (IAD).

15.3.2.1 POTS Splitter A POTS splitter is a passive three-way terminal that splits the signals on the copper wire into POTS signals and data

signals at the customer premises. The top of the right-hand side of Fig. 15-2 shows a typical splitter that forks an incoming POTS line into two lines, one going to the modem and the other to the POTS phone. The POTS splitter is an optional device. Some DSL performs the splitting function at the central office end of the POTS line and is thus called *splitterless DSL*.

15.3.2.2 DSL Modem A DSL modem connects a user computer to the network via a phone line. A modem consists of a receiver and transmitter, an analog-to-digital converter, a digital-to-analog converter, a modulator, and a memory module.

Similarly to traditional modem, an ADSL modem is also designed to have internal and external models. The internal modem is usually based on the PCI bus for one-person use and is less costly. The external modem usually has an Ethernet interface that can be connected to a computer on the one side and to an Ethernet hub/switch on the other side.

15.3.2.3 Integrated Access Device IAD is a new generation of customer premises equipment that integrates multiple functions into one device and serves as the interface between DSL network side equipment like the DSLAM and the customer's voice and data equipment. An IAD may include the following components:

- A DSL modem
- A data switching/routing device such as an enterprise router or hub if the IAD is for enterprise use
- A packetizer that converts the voice into data packets (cell, IP packets, frame relay frame, etc.)
- A traffic scheduler that prioritizes traffic based on the type of traffic such as voice and data, real time versus nonreal time
- A voice switching module at the customer premise, which is a traditional PBX for telephony services if the IAD is at an enterprise premise.

15.4 DSL Data and Voice Services

DSL is a physical layer technology that allows both data and voice services. One promising service is voice over DSL (VoDSL). This section discusses an ATM-based DSL data service model and then the architecture, transport layer, and signaling of VoDSL service.

15.4.1 DSL Data Service

DSL provides physical connectivity to the high-speed Internet via a central office switch or router. A DSL overlay data connection model, recommended by DSL Forum and shown in Fig. 15-3, has IP over PPP over ATM over DSL (DSL Forum 2000).

This model uses ATM virtual connection to provide data transport service between customer premises and the central office. Other choices of transport service include X.25, frame relay, and Ethernet. According to various estimates by market researchers, ATM has the largest installed base in the deployed DSL systems.

With this model, a user first sets up a PPP dial-up session with the Internet service provider, with a dynamic assigned IP address. Then the user data, such as a request for a Web page, is first converted into an IP packet and the IP packet is encapsulated inside a PPP session to identify the user session. Then the encapsulated session is encapsulated inside AAL5 cells, which are then sent to the central office on an ATM virtual circuit encapsulated inside DSL frames carried on the copper wire between the customer's home and the central office switch.

PPP over ATM is defined in IETF recommendation RFC 2364, and is designed to work on any point-to-point interface to send data over a point-to-point line (Gross 1998). Such interfaces include telephone, leased,

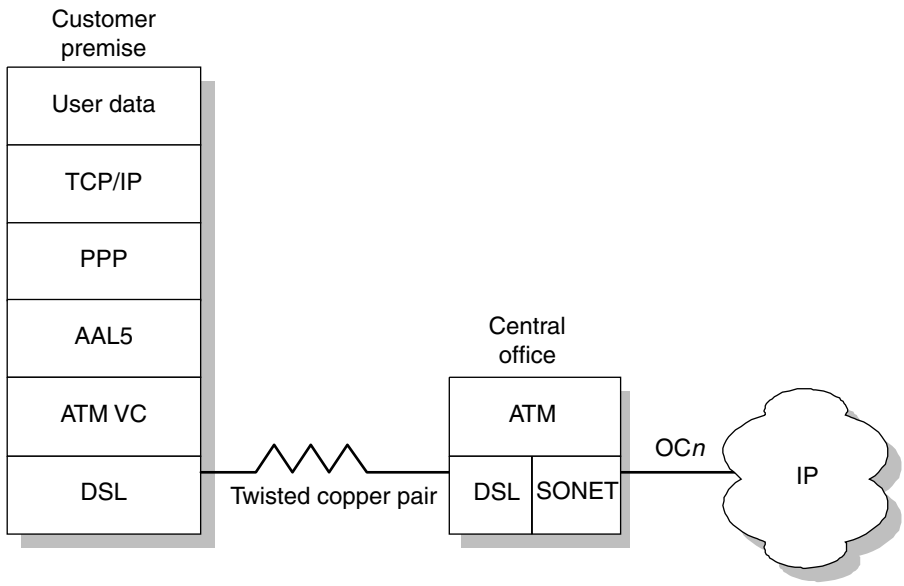


Figure 15-3
An ATM-based DSL data service model.

dedicated, or direct lines, and may use point-to-point channels or virtual circuits of multiplexed interfaces such as ISDN. The ATM adaptation layer is responsible for inserting higher-layer information into the cells to be transported over an ATM virtual connection. AAL5 is most often used for connectionless Internet traffic with minimal overhead.

15.4.2 Voice over DSL Service

Voice over DSL service allows the service provider to offer multiple voice lines via a single copper wire pair while still providing broadband data access. It requires additional network components, transport service for voice channels, and a call signaling mechanism.

15.4.2.1 Additional Network Elements A DSL network needs additional components in order to support voice services. In addition to the DSLAM and IAD as described above, a media gateway is needed to interconnect a DSL network with a public switched telephone network (PSTN), as shown in Fig. 15-4. In this architecture, the DSLAM has the capability of separating voice traffic from data traffic and then sending the data traffic to a packet network and voice traffic to a media gateway.

The medium gateway serves as the interface between the traditional circuit-switched telephony network and the DSL network. The functions it performs include voice compression, echo cancellation, and depacketization of the voice packets if so desired by the local exchange (class 5 switch).

15.4.2.2 VoDSL Transport Services There are three options for the transport layers of VoDSL service: IP, ATM, and frame relay. The overwhelming majority of installed DSL systems use ATM transport, which has been adopted by DSL Forum as its de facto standard. The QoS that ATM provides is a main motivation for adopting it as the transport layer for VoDSL. ATM provides connection-oriented services, and a connection, once set up, can remain in effect either permanently or for the duration of a call. The QoS parameters can be set at the time of the connection setup.

ATM virtual channel connections are used to carry multiple voice channels, one VC per voice channel, in place of the traditional POTS channel. More advanced schemes, though not widely deployed yet, multiplex multiple voice channels onto a single VCC. The voice is packetized, carried over the DSL link, and routed through an ATM network.

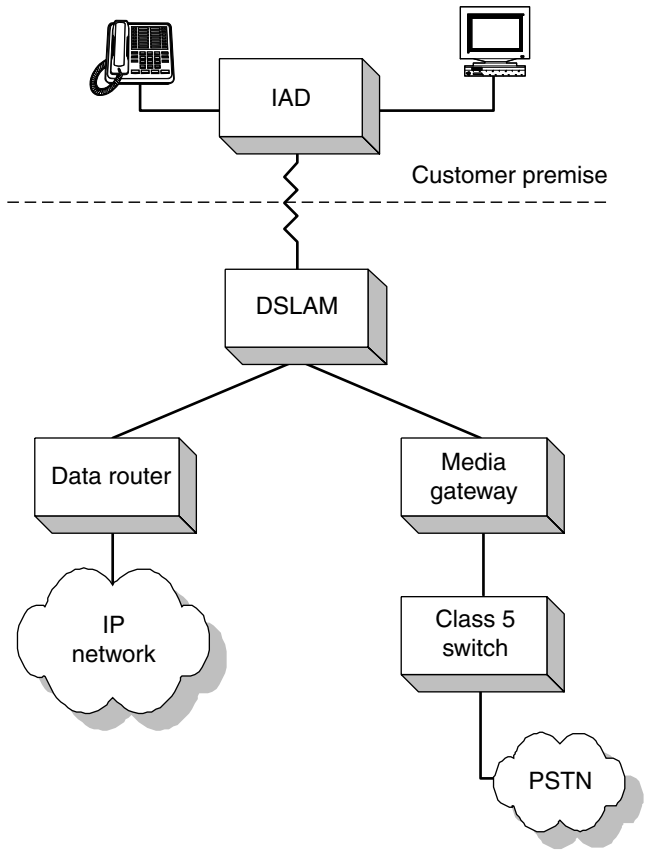


Figure 15-4
A high-level view of VoDSL architecture.

The earlier versions of the DSL system used ATM AAL1 to carry voice between an IAD and a DSLAM, a model known as time division multiplexing over ATM. More recently, DSL Forum adopted ATM AAL2 as the transport mechanism for voice service over DSL. AAL2 uses packet-interleaved multiplexing and is more efficient because it allows the network to allocate bandwidth dynamically on the DSL service. If no voice call is in session, all bandwidth can be allocated to data service. AAL2 also allows for silence suppression that can save bandwidth by up to 50 percent in some studies. AAL2 allows for the detection of silence, and no packets are sent for the silent moment.

While the earlier versions of VoDSL use ATM permanent virtual circuit, the more recent trend is to use switched virtual circuit for voice service since it provides more flexible and efficient bandwidth allocation.

15.4.2.3 VoDSL Signaling The first generation of VoDSL systems uses a channel associated signaling method that imbeds call signaling information inside the packets that carry the voices. This is also called *in-band signaling* since the signaling information and calls share the same communications channel.

It has been proposed that an out-of-band signaling scheme, which separates signal information from the payload data into separate communications channels, be adopted for VoDSL service. The candidates for out-of-band signaling schemes include H.248 for the access side of VoDSL at customer premises.

REVIEW QUESTIONS

1. Describe the three local loop configurations and the functions of a remote digital terminal.
2. Describe what DSL technology was originally developed for and what factors motivated its refocus.
3. Characterize DSL technology in terms of transmission medium, target applications, and its relation to POTS.
4. Discuss the main constraints on the data transmission rate over POTS lines and the method for overcoming those constraints.
5. Describe the two types of crosstalks—impulse noise and radio noise—caused by the use of high frequencies in DSL transmission and DSL's approaches to dealing with them.
6. Describe the differences between ADSL, G.Lite, and VDSL in terms of data transmission rates, transmission distances, and target applications.
7. Describe the components of a DSL network and the functionality of a DSLAM and POTS splitter.
8. Discuss the DSL Forum-recommended DSL data service model and the functions performed by the PPP.
9. Describe the functions of an IAD and where it is normally located, at customer premises or the central office.
10. Briefly describe how a DSL network offers multiple voice channels over a single copper wire while still providing high-bandwidth data service.

REFERENCES

Cioffi, J., Silverman, P., and Starr, T. 1999. *Understanding Digital Subscriber Line Technology.* Englewood Cliffs, NJ: Prentice Hall.

DSL Forum. 1997. "ADSL Forum System Reference Model." ADSL Forum TR-001. Web site: www.adsl.com.

DSL Forum. 2000. "ADSL Tutorial." White paper. Web site: www.adsl.com.

Goralski, W. 1998. *ADSL and DSL Technologies.* New York: McGraw-Hill.

Gross, G., et al. 1998. "PPP over AALS." IETF RFC 2364. Web site: www.ietf.org.

ITU-T. 1999a. "Asymmetrical Digital Subscriber Line (ADSL) Transceiver." Recommendation G.992.1. Web site: www.itu.int/ITU-T/.

ITU-T. 1999b. "Splitterless Asymmetrical Digital Subscriber Line (ADSL) Transceiver." Recommendation G.992.2. Web site: www.itu.int/ITU-T/.

ITU-T. 2001a. "Very-High-Speed Digital Subscriber Line Foundation." Recommendation G.993.1. Web site: www.itu.int/ITU-T/.

ITU-T. 2001b. Handshake Procedures for Digital Subscriber Line (DSL) Transceiver." Recommendation G.994.1. Web site: www.itu.int/ITU-T/.

ITU-T. 2001c. "Overview of Digital Subscriber Line (DSL) Recommendations." Recommendation G.995.1. Web site: www.itu.int/ITU-T/.

Paradyne. 1998. "The DSL Source Book." Paradyne Co. Web site: www.paradyne.com.

Rhodes, R., Pugel, M., Litwin, L., and Richardson, J. 2001. "ADSL Technology Explained, Part 1: physical layer." *Communication Systems Design.* April. Web site: www.commsdesign.com.

CHAPTER 16

Packet Cable Networks

16.1 Introduction

The packet broadband cable network is built on the existing broadcast cable TV (CATV) (also known as *community antenna TV*) networks. An introduction to CATV networks provides a context for this chapter.

16.1.1 Brief History of Cable Networks

The first cable television appeared in the late 1940s. At the time, the fledging TV industry provided off-air broadcast signals only to population-dense metropolitan areas. Cable television systems provided an alternative for those areas that had poor TV signal reception because of obstructions or long distances from signal transmitters. In some areas like New York City, multiple signal reflections and shadows cast by tall buildings also made good-quality off-air reception difficult. (Often at the time it was the local TV equipment retailers who constructed antennas in their communities and strung coax cables from the antennas to their communities with the intention of opening up new markets for TV sets, which is why cable TV became known as *community antenna TV*.)

Initially the sole purpose of cable television was to deliver local TV programming to areas the off-air broadcast signals could not reach or where the reception was poor. By the late 1960s, according to W. Ciciora, nearly all of the areas of the U.S. and Canada that could benefit from a cable TV had been served. Growth in the cable industry basically stopped (Ciciora et al., 1998). Then, in the mid-1970s, the new technology of satellite delivery of broadcast signals to cable systems brought a new life to the cable television industry. Channels were added to carry programming from other areas of dedicated programming sources via satellite delivery systems. New value-added services such as specialty channels and pay-per-view channels brought their own life to cable television.

In the late 1980s, cable TV operators started deploying optical fiber to upgrade the older cable systems and to build new systems. Among the other new technologies deployed was a new generation of amplifiers. Together, optical fiber and the new amplifier technology brought about the birth of hybrid fiber coax (HFC) cable networks.

In the 1990s, the explosive growth of Internet data services and CATV systems led to exploring them as alternatives to the traditional telephony access networks in regard to the last-mile solution. Cable networks started to move toward becoming general communications networks, rather than just a single-service broadcast TV.

One recent change taking place in the cable TV industry is the transition from analog TV to digital TV, leading toward the digital TV age. Broadcasters have been mandated by the U.S. federal government to switch to all-digital TV broadcasting by the middle of the first decade of this new century.

Table 16-1 gives a summary of cable TV's history.

16.1.2 CATV Network Basics

This section discusses CATV network topology, then describes the basics of analog and digital CATV.

16.1.2.1 CATV Cable System Topology A typical CATV network is configured as a tree, as shown in Fig. 16-1. This was the configuration commonly seen before cable modem arrived on the scene. CATV was solely for one-way broadcast of TV programming. At the center of a cable network is a headend, which is connected to a set of trunk cables. Each trunk cable is connected to a set of feeder cables, and each feeder cable is connected to a set of drop cables, which directly go into customer premises (Ciciora et al., 1998).

TABLE 16-1
Summary of Cable Network TV History

Year	Event
1941	The black-white TV technical standards known as NTSC* emerged.
1948	Ed Parson of Oregon built the first CATV system with twin-lead transmission wire strung from housetop to housetop. In 1950, a coax cable-based cable network was built in Pennsylvania.
1953	The black-and-white TV standard was modified to support color television.
Late 1980s	HFC cable networks started to emerge.
Early 1990s	Two-way cable systems and cable modems start to be widely deployed.
1998	Digital broadcasting starts in the top U.S. TV markets.
1999	Data over Cable Service Interface Specifications (DOCSIS) standards adopted.
2000	IP-based packet cable specification emerged.
2006	The analog spectrum will be returned to the FCC for auction in the United States.

*NTSC: National Television Standards Committee.

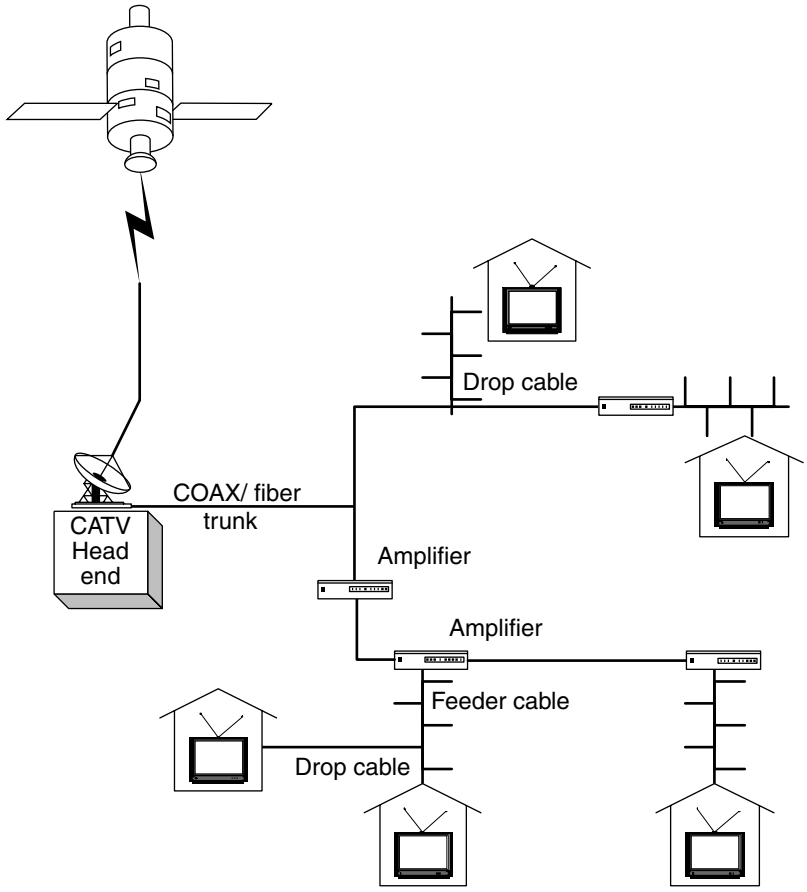


Figure 16-1
HFC CATV architecture.

HEADEND A *headend* is the operation center of a CATV cable access network. It is connected to many distribution nodes via trunk cables, which can be made of coax cable or fiber. The main functions a headend performs include the following:

- Receiving broadcast signals from satellite or microwave dishes
- Mixing local or recorded TV programming
- Assigning channel frequencies to all signals destined for cable distribution

TRUNK CABLES Coming out of the headend are trunk cables that connect the headend to a set of distribution points. Traditionally, coax cable has been used for trunk cable, but starting in the late 1980s, optical fiber

began to replace it. If a trunk cable goes beyond the prescribed effective distance ranges, the signals attenuate to a level where it is difficult to maintain their quality, requiring the use of analog amplifiers. On average, amplifiers are placed at 2000-ft intervals along the cable trunk. While a cascade of amplifiers can extend the reach of a cable trunk, they also introduce additional noise that causes signal distortion. Only a finite number—30 to 40—of amplifiers can be cascaded, which is a limitation on CATV systems.

FEEDER CABLES Branching out from trunk cables are many feeder cables, also known as *distribution cables*, that connect the trunk cables to a set of distribution points. A feeder cable is responsible for serving a local neighborhood, which ranges from 500 to 2000 residential homes or offices. To avoid excessive signal attenuation and noise, the feeder cable is limited in effective reach and typically can have a maximum of two amplifiers. The homes located beyond the maximum reach of the feeder cable cannot be served by cable TV services.

DROP CABLE Along feeder cables are periodic taps that connects them to drop cables. Drop cables, limited in length to approximately 150 ft, directly enter customer premises. Connected to the drop cables inside customer premises are customer devices such as TV sets, set top boxes, and VCRs.

16.1.2.2 Analog CATV Basics Frequency allocation is the same whether the analog TV signals are transmitted over cable or through the air. The cable spectrum is divided into 6-MHz channels, and the signals of one program are sent over one channel. Out of 6 MHz of bandwidth, only 4.5 MHz is actual transmission capacity; the remaining 1.5 MHz is used as a guard band to prevent adjacent channels from interfering with each other.

The cable transmission frequency spectrum is the same as the broadcast frequency, which starts from 55.25 MHz and goes up close to 1 GHz. Channel allocation in the United States for cable systems is different from that for broadcast TV. For broadcast TV in the United States, the FCC originally allocated parts of the very high frequency (VHF) spectrum to twelve channels. The channels were in two separate frequency blocks to avoid interfering with the existing radio stations:

Channels 2 to 6: 54 to 88 MHz

Channels 7 to 13: 174 to 216 MHz

Channels 14 to 69 were established later at frequencies between 470 and 812 MHz.

In contrast to broadcast TV, with CATV there is no interference with the existing services, and thus it was allocated a continuous single frequency block. All the newer TV sets have a tuner allowing them to tune to either broadcast TV channels or cable TV channels. The CATV frequencies are as follows:

Channel 2: 55.25 MHz

Channel 3: 61.25 MHz

Channel 4: 67.25 MHz

Channel 5: 73.25 MHz

Channel 6: 79.25 MHz

……

Channel 90: 594.25 MHz

Early cable systems had about 200 MHz in total bandwidth, supporting thirty-three channels (6 MHz each). Later, as programming options grew and technologies progressed, cable bandwidth was increased to 300, 400, 500, and 550 MHz, supporting over 90 channels.

CATV is a point-to-multipoint broadcast system, using the frequency division multiplexing technique to multiplex many channels onto a single coax cable. Scrambling is used to allow subscribers to view only the channels they have subscribed to. *Scrambling* means that a signal which is slightly offset from the channel signal is inserted to interfere with the picture. A set top box in the customer premises descrambles or restores the original channel signals.

16.1.2.3 Digital CATV Basics Digital CATV, like digital broadcast TV, uses digital streams (0s and 1s) to represent pictures and sound, instead of using amplitude and frequency like analog TV. Using the same 6-MHz channel, digital CATV signals can carry much more information than analog signals, so resolution and levels of detail of the images transmitted by digital TV are much higher than those transmitted by analog TV.

Digital TV uses a scheme known as MPEG-2 to compress and encode information into each TV channel. MPEG-2 is already the industry standard for DVD videos and some of the satellite TV broadcast systems. While 6 MHz of bandwidth supports only one analog channel, MPEG-2 allows digital TV to support up to ten channels on the same bandwidth.

With 550 MHz of bandwidth, close to 1000 channels can be supported on digital TV. Instead of using scrambling, digital CATV uses encryption to encrypt all signals, and only a receiving device with the proper key can decrypt the signals.

A disadvantage of digital TV is that in regard to quality it is an all-or-nothing proposition, unlike analog TV, where quality fades gradually. Users get good reception or they get no reception, with nothing in between.

16.2 Data Cable Network—Cable Modem Systems

Using cable networks becomes a logical choice when a solution is sought to the issue of broadband access of the last mile in order to accommodate the explosive growth of the Internet. A cable network is a broadband access network with a bandwidth of 6 MHz per channel, a broadband bandwidth in the access network by today's standards.

However, the original cable system architecture was never intended to be a general-purpose, two-way communications network. Its primary goal was to deliver high-bandwidth video signals to residential homes. In order to provide two-way Internet data services, substantial changes to the existing architecture have been required. This is where cable modem technology comes in.

In North America, cable modem technologies are standardized around DOCSIS (data over cable service interface specification). The rest of this section describes the history of DOCSIS and its European counterpart, discusses its key concepts, and then describes its architecture, its physical layer, its MAC layer, and its higher layer services.

16.2.1 Brief History of DOCSIS

The standardization efforts for cable data networks first started with the IEEE 802.14 Cable TV MAC and PHY Protocol Working Group in 1994. The IEEE 802.14 proposal selected ATM as the data link layer protocol and offered multiple choices for the physical layer (IEEE 802.14 Working Group, 1995). (IEEE 802.14 Working Group has ceased its activities since 2000.)

In response to the slow progress in public standardization efforts and market pressure, starting in late 1995, North America cable network operators started their own standardization efforts for data cable networks by forming a joint research and development consortium known as CableLabs. The goal was to develop a new generation of cable networks that provided general telecommunications services, including data over cable service. This new generation of cable networks is usually known as *cable modem networks*. The standard, known as DOCSIS 1.0, was published in 1996 (CableLabs 1996). A new version, DOCSIS 1.1, was published in 1998. The DOCSIS specifications were then accepted by the international standardization body ITU-T and became ITU Recommendation J.112 (ITU-T 1998). DOCSIS uses IP over Ethernet as its data transport technology, and cable operators in North America began its wide deployment in the late 1990s.

In Europe, there are two competing proposed specifications supported by two large consortia. One is based on North America's DOCSIS with the addition of an 8-MHz downstream channel. This specification is known as *EuroDOCSIS* for its close resemblance to DOCSIS, and is backed by the European Cable Modem Consortium. The second proposed specification, backed by the Digital Audio Video Council interoperability consortium, is known as *digital video broadcasting* (DVB). DOCSIS uses IP over Ethernet as its data transport layer technology, while DVB uses ATM.

16.2.2 Cable Modem Network Configuration

Cable modem systems are built upon existing CATV infrastructure, but with substantial additional capabilities. The main differences between the two, as shown in Fig. 16-2 and Fig. 16-1 (CableLabs 2002a; Microsoft 1999), include the following:

- Cable modem networks allow two-way communications instead of only one-way.
- Cable modem networks add fiber links to interconnect regional access networks to IP backbone networks.
- Cable modem networks add data routing/switching capabilities (CMTS) at the headend.
- A cable modem termination system (CMTS) is added at the customer end.

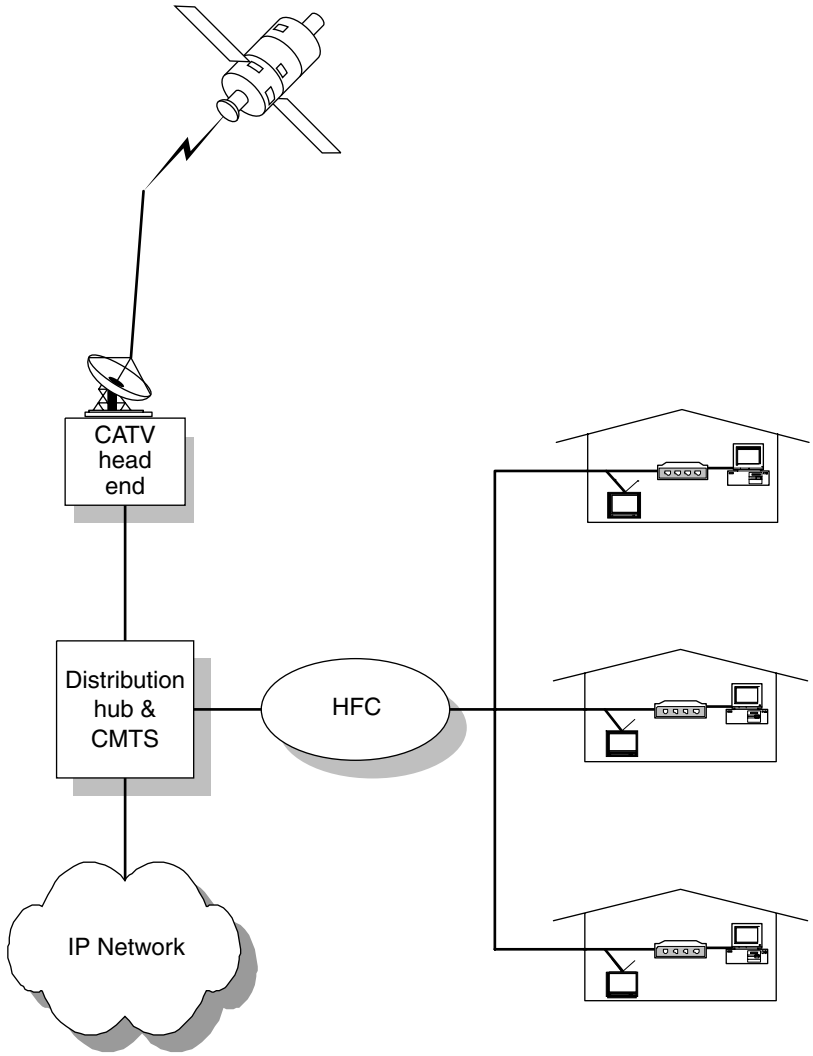


Figure 16-2
Cable modem network architecture.

The most important added capabilities of cable modem networks have to do with the CMTS and cable modems (CMs), as described below.

16.2.2.1 Headend The headend of DOCSIS cable modem networks are very similar to those of CATV networks: They receive the signals for TV programming from satellites and transmit the signals to the distribution hubs. This portion of the networks is still one-way communication without any upstream channels.

16.2.2.2 Distribution Hub and Cable Modem Termination System

The distribution hub and CMTS are new additions to conventional CATV networks intended to support Internet data services. The distribution hub is the interface point between the regional network and the cable plants and is located close to the neighborhood it serves. Each hub serves between 500 to 1000 homes or small offices. The main components found inside a hub include the following:

- The cable modem termination system
- An IP switch or router
- Local cashing servers

The IP switch or router inside a hub plays the similar role to that of a LAN router or switch. It concentrates and routes the data traffic that originates from customers' homes or offices and sends the data to the regional network. The local cashing server manages the contents cached at the local hub.

A CMTS interfaces the headend and the customer cable modems via cable plants. Each CMTS unit provides a dedicated large amount of downstream bandwidth ranging anywhere from 30 to over 100 Mbps that is shared by many users. The upstream bandwidth per CMTS ranges from 2 to 10 Mbps. The major functions of a CMTS include the following:

- Controlling the bandwidth allocation for the data traffic to each modem and enforcing the bandwidth allocation policy
- Assigning a time slot to each cable modem for transmitting upstream messages
- Enforcing QoS policies such as traffic shaping and policing, and packet classification based on QoS classes

16.2.2.3 Cable Modem

CM is a key component of cable modem networks. Located at the customer premises, as shown in Fig. 16-3, it connects the customer premises data devices such as PCs or enterprise Ethernet switches to the CMTS at the connected hub. The splitter separates the Internet data channel from normal CATV programming and passes the TV programming to the set-top box and the Internet data to the cable modem. A cable modem has four major components:

- A tuner that receives the modulated digital signals and passes them to the demodulator to convert them back to the original digital signals

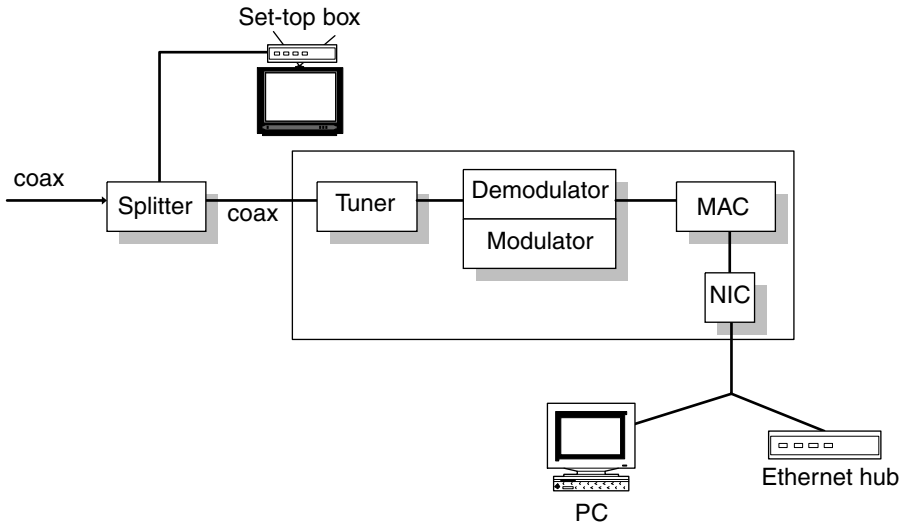


Figure 16-3
Cable modem components and customer premises equipment configuration.

- A modulator/demodulator that turns downstream channels' radio-frequency signals into simpler signals and passes them to an analog-to-digital (A/D) converter
- A media access controller that implements the network protocols and access control policies such as Ethernet's carrier sense multiple access with collision detection (CSMA/CD) control algorithm
- A network interface card that connects different types of user data devices to the cable data network

16.2.3 DOCSIS Architecture

DOCSIS version 1.0 specifies the MAC and IP layer interfaces between the cable modems at customers' premises and the CMTS at the distribution hub. The interface specifications define an end-to-end protocol stack, as shown in Fig. 16-4 (CableLabs 1996). DOCSIS version 1.1 adds specifications for quality of service at the MAC layer interface (CableLabs 2002b).

The following application scenario of sending Web data requests from a user data device like a PC helps illustrate how the DOCSIS system works:

1. A user PC generates an IP packet encapsulated in an Ethernet frame and sends the frame to the cable modem at the desktop.
2. The modem's DOCSIS layer encapsulates the Ethernet frame in a DOCSIS MAC frame and then acts as a bridge, forwarding the

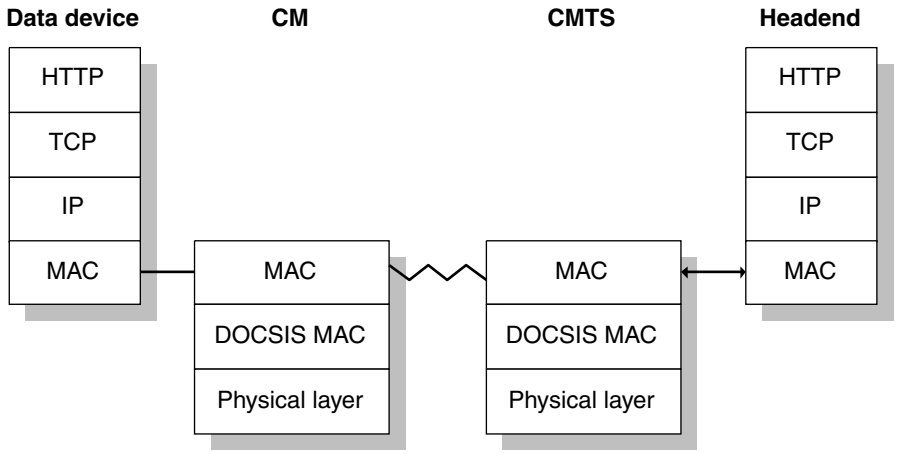


Figure 16-4
End-to-end protocol stacks of DOCSIS architecture. (CableLabs, 2002a.)

DOCSIS MAC frame to the cable modem termination system at the distribution hub.

3. The CMTS extracts the Ethernet frame from the DOCSIS MAC frame and forwards it to the router at the CMTS or headend. If the requested content is cached at one of the CMTS caching servers, the content is returned to the user in a downstream channel. Otherwise, the CMTS router routes the request to the headend router or the connected IP network.

The communications between two customer premises served by the same CMTS must pass through the CMTS. The CMTS functions as an Ethernet LAN for the neighborhood it serves.

16.2.3.1 Physical Layer of Cable Modem Systems The physical layer of a cable modem network consists mainly of the specification of modulation techniques for downstream and upstream channels, video and audio compression schemes, and the security encryption scheme:

- Downstream data channel rates range from 20 Mbps (16 QAM) to 40 Mbps (256 QAM), with a typical configuration of 30 Mbps (64 QAM) in 6-MHz channels.
- Upstream data channel rates range from 320 kbps (QPSK) to 10.24 Mbps (16 QAM).
- MPEG-2 is used for digital video compression and encoding.
- Dolby audio AC-3 is used for audio compression.
- The U.S. data encryption standards (DESs) are used for security.

16.2.3.2 DOCSIS MAC Layer Cable modem networks use shared media, meaning multiple users share a coax cable with a fixed amount of bandwidth. The media access control is a key issue addressed by DOCSIS standards. Cable modem networks are asymmetric, with a high data transmission rate in the downstream direction and very limited bandwidth in the upstream direction. The MAC specifications address the access control issues for the upstream and downstream directions (CableLabs 2002b).

DOCSIS UPSTREAM MAC SCHEDULING The upstream MSC scheme is more complicated than downstream MAC because there is more contention for resources in the upstream direction. The MAC scheme requires that each customer cable modem first seek permission to send data. The controlling CMTS uses an admission control algorithm to determine whether or not a request can be granted. Once a request is granted, the CMTS issues a service identifier (SID) to the requesting modem, and monitors and controls the data flow from the modem. The SID functions more or less like the token of a Token Ring network. In addition, it also specifies the service parameters associated with a granted data flow. The modem can send the data only if it complies with the SID classification. A SID can be created, deleted, and modified by the controlling CMTS at any time.

There are four operational modes, each defining a way and a class of QoS for a cable modem to send data upstream:

- *Unsolicited grants.* When a request is accepted in this mode, the CMTS schedules a fixed-size bandwidth and the modem does not need to contend for the channel. It is similar to the constant bit rate service of ATM.
- *Real-time polling.* The CMTS reserves some time slots and periodically unicasts request polls to the corresponding modem. If the modem answers the polls, the time slots are assigned to the modem.
- *Committed information rate.* In this mode, the CMTS forces the cable modems to use contention-based requests. Each time a modem intends to send a packet, it makes a request first.
- *Tiered best effort.* In this mode, best-effort service combined with the layer-2 priority mechanism is offered to the requesting modem. Eight priority levels are currently defined in DOCSIS 1.1.

DOCSIS DOWNSTREAM MAC The downstream MAC is simpler because there is much less contention for resources. A separate channel is assigned to a class of services and is designated to a requesting cable modem.

16.2.3.3 DOCSIS Security Mechanism Cable is a medium shared among many users in the same neighborhood, and security and privacy are very important concerns. The main goal of the DOCSIS security mechanism is to provide data transport security. Data transport security provides users with data privacy and prevents the unauthorized access to network services.

Data privacy means that only the intended recipient can have access to the data contents of a transmission. Data privacy in DOCSIS cable modem networks has two parts: encrypting traffic flows between cable modems and the controlling CMTS, and using an authenticated client-server key management protocol to control the distribution of encryption key information to client cable modems.

The encryption protocol consists of the following components:

- Specification of the DOCSIS MAC frame format for carrying the encrypted packet data
- A set of supported data encryption algorithm and authentication algorithm
- Rules for applying the cryptographic algorithms to the packet data encapsulated in the DOCSIS MAC frames

DOCSIS currently uses the cipher block chaining (CBC) mode of the US. data encryption standards (DESs) to encrypt the data flow between a cable modem and the controlling CMTS. The security mechanism leaves room for more advanced encryption algorithms if so desired in the future.

All modem clients are authenticated and authorized by the controlling CMTS before the requested service is granted, in order to prevent service theft or unauthorized access to network service. Cable modems use the DOCSIS key management protocol to obtain authorization and encryption key information from the CMTS and to support periodic reauthorization and key refresh. The DOCSIS key management protocol uses X.509 digital certificates (ITU-T 2000), RSA public key encryption, and triple DES to secure key exchange between a CM and the controlling CMTS.

16.3 Multiservice Cable Networks—Packet Cable Networks

Packet cable networks are an ongoing effort by the cable industry to evolve the cable modem network into the next-generation general-purpose

communications network. The goal of packet cable networks is to provide multiple services such as voice over IP over cable and multimedia applications. This section provides an overview of the architecture developed by CableLabs, known as *PacketCable*. The term *packet cable network* will be used throughout this section to refer to the multiservice, general-purpose, and IP-based cable network described in the PacketCable specification. The section starts with a brief introduction to the standardization efforts in this area, before proceeding to a discussion of packet cable architecture.

16.3.1 Introduction

The PacketCable specifications are based on the DOCSIS specifications. Completed in December 1999, the PacketCable 1.0 specifications define the protocols and network interfaces to support the voice over IP service over a DOCSIS cable network (CableLabs 1999). PacketCable 1.1, completed in December 2000, adds capabilities to support lifeline service delivery (CableLabs 2000b, 2000c). As of this writing, eight of the PacketCable specifications have been ratified by ITU-T as global cable standards; the global version of these standards is termed *IPCablecom*. The European standards body ETSI is also considering the adoption of IPCablecom standards.

The packet cable network builds on the infrastructure of the physical layer and the MAC layer of DOCSIS cable modem networks. The PacketCable specifications focus on the higher-layer specifications, with an emphasis on real-time-sensitive services.

16.3.2 Packet Cable Network Architecture

The packet cable network overlays the DOCSIS architecture, with an emphasis on the call signaling and customer premise equipment (CableLabs 1998). The architectural goals of the PacketCable network include the following:

- Supporting multiple real-time-sensitive and non-real-time services such as telephony, fax, and Internet access
- Being scalable to support millions of subscribers
- Supporting both primary and secondary residential telephone line services
- Building on the DOCSIS-specified HFC physical and IP-based transport networks

- Providing quality of services equal to or better than PSTN telephony services [e.g., one-way delay for local IP access no more than 45 ms, less than 1 percent of call blocking rate during the high day busy hours (HDBH), and less than 0.25 frame slips per second due to the unsynchronized clock or packet loss]

Figure 16-5, based on the PacketCable 1.0 and 1.1 architecture framework (CableLabs, 2000), presents the overall framework and scope of the packet cable network. The main functional components encompassed in the architectural framework include the following:

- :Multimedia terminal adapter (MTA)
- HFC access network
- Cable modem termination system
- Call manager server (CMS)
- Media gateway
- Media gateway controller
- Signaling gateway
- Operation support system (OSS) servers

The above functional components of the reference architecture can be categorized in four functional groups: DOCSIS cable modem network, call servers, PSTN interface gateways, and OSS servers.

16.3.2.1 DOCSIS Cable Modem Network The PacketCable architecture adds additional functions or functional components to the DOCSIS cable modem network. MTA is an added functional component and additional functions are added to the CMTS.

MULTIMEDIA TERMINAL ADAPTER An MTA is a client device located at the customer premises. An MTA may have a DOCSIS cable modem embedded inside it. Functionally, an MTA is an interface device that on the one hand provides an interface to the other customer premises equipment like telephones, PCs, and cable TV. On the other hand, it provides an interface to the other packet cable network elements via the HFC access network. The functions an MTA performs include the following:

- Call control functions such as call signaling with the call management server and QoS signaling with CMS and CMTS
- Security functions such as ensuring the authentication, integrity, and confidentiality of a message between the MTA and other packet cable network elements

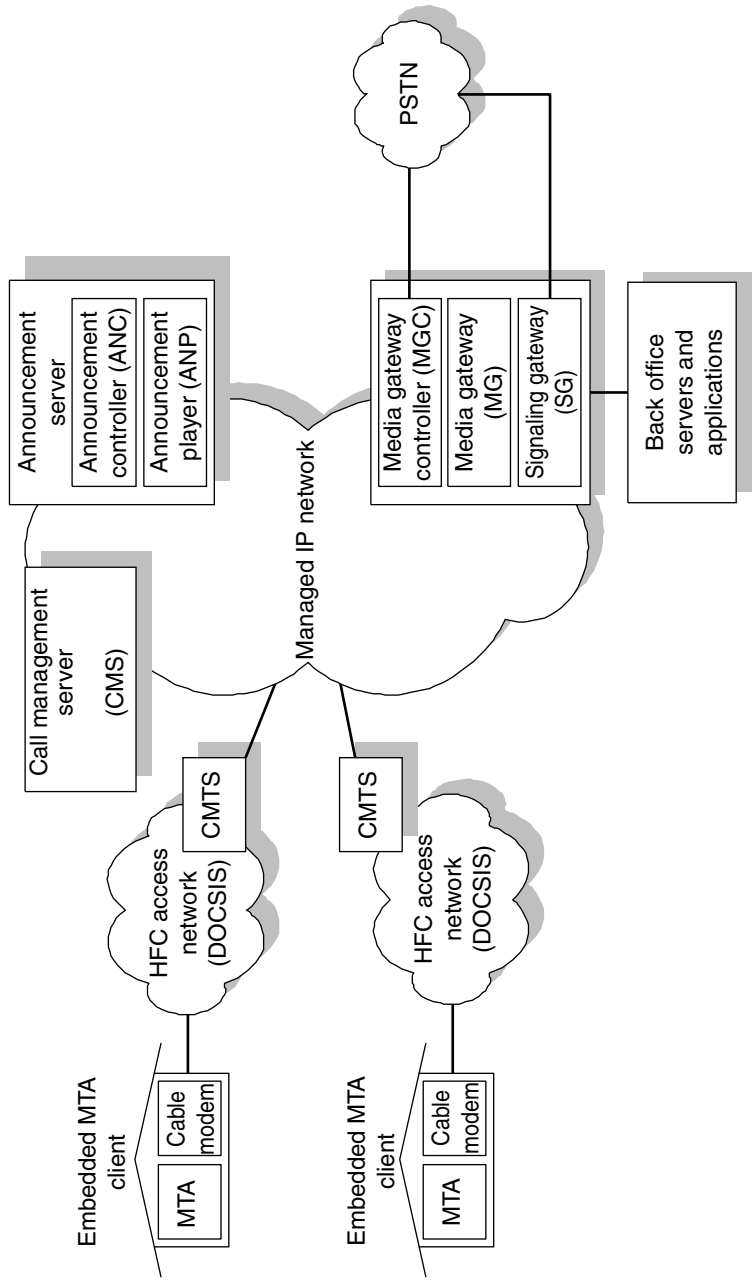


Figure 16-5 PacketCable reference architecture. (CableLabs, 2000.)

- Media functions such as G.711 codec, encoding/decoding media streams, and mapping media streams to MAC services of DOCSIS access networks
- Analog phone line interfacing such as audio tone, voice transport, DTMF, caller ID signaling, and voice mail indication

HFC ACCESS NETWORK This is a DOCSIS hybrid fiber/coax access network that provides physical layer and transport layer services to the packet cable network. The access network is a bidirectional, shared medium network.

CMTS The cable modem termination system in the packet cable network plays a role similar to that of DSLAM in a DSL network. A CMTS can be located either at the cable headend or a distribution hub.

A CMTS in a packet cable network has call-related responsibilities in addition to the functions it performs for the cable modem network. It provides data switching and forwarding functions as well as the radio frequency interface to and from the cable modem in the cable modem network. Specifically, the CMTS is responsible for appropriately classifying, prioritizing, flow-controlling, queuing, scheduling, and shaping all of the traffic flows between cable data subscribers and the data switching equipment like routers at the headend. In addition, the CMTS performs the following functions:

- Managing the IP packet's tag of service (TOS) field for real-time-sensitive application like telephony services
- Converting QoS parameters from the IP backbone network to the DOCSIS QoS parameters
- Recording the resource usage for billing purposes

16.3.2.2 Call Servers Call-related servers include calling the management server and a set of call-related resource servers.

The call management server plays a role similar to that of soft-switch and provides call control and call signaling services. The call control functions include the following:

- Digit analysis and call routing
- Call state machine management
- Call feature implementation
- Call treatment handling

One kind of call-related resource server is an announcement server that controls and plays call announcements for either call handling (e.g., "The number you dialed is invalid") or call features (e.g., voice mail message service).

16.3.2.3 PSTN Interface Gateways The gateways are mainly for interfacing the PSTN network to support real-time applications like telephony services and multimedia services.

PSTN gateways provide gateways from the packet cable network to the PSTN network for telephony services. Functionally, a PSTN gateway consists of three components: media gateway, media gateway controller (MGC), and a signaling gateway.

MEDIA GATEWAY CONTROLLER A media gateway controller is a complicated functional component that performs many functions, including the following:

- It instructs a controlled media gateway to connect/disconnect a bearer path, to generate events for the detection of in-band signaling, and to apply a resource such as tone.
- It makes call routing decisions that involve both PSTN and packet cable networks.

MEDIA GATEWAY A media gateway provides physical connectivity between a PSTN network and a packet cable network and manages physical resources in the form of bearer paths between a PSTN and a packet cable network, as instructed by the MGC. It is also responsible for detecting and generating events for the MGC such as in-band signaling.

Primarily, a media gateway provides the mapping functions between a packet cable network and a PSTN network. It is equipped with two sides of network interfaces. On the packet cable network side, it has an IP network interface to receive and send IP packets and to process the IP protocol stack including IP and UTP/TCP. On the PSTN network side, it supports TDM digital interfaces such as T1/E1, T3/E3, OC3, etc. In addition, its transport services also include the translation between conferencing endpoints and other terminal types and between audio and video codecs.

SIGNALING GATEWAY The primary function of the signaling gateway is to translate between the call signaling messages of a circuit-switched PSTN network and SIP call signaling messages. Session Initiation Protocol (see Chap. 23) has been adopted as the call signaling protocol for

IP-based packet cable networks. Specifically, the signaling gateway functions include the following:

- Terminating signaling links such as SS7 links from PSTN and generating the appropriate SIP signaling messages for a packet cable network
- Mapping the address from a PSTN to the address in a packet cable network

16.3.2.4 OSS Components The OSS components include various servers for backend office support functions. The servers may include domain name servers, system log servers, record keeping servers, billing servers, subscriber provisioning servers, and trouble ticket servers, among others.

16.3.3 QoS Issues

The PacketCable architecture uses the IP over Ethernet over cable adopted in DOCSIS architecture as the foundation for voice, data, and multimedia services. It also uses the IP QoS schemes RSVP and DiffServ (see Chap. 18, on IP QoS architecture and protocols) to address QoS issues of cable IP networks such as loss of packets, latency, and jitter. In addition, multiprotocol label switching (MPLS) (see Chap. 17), is being explored to provide traffic engineering and VPN services over packet cable networks.

Besides the latency and jitter of cable IP networks, the PacketCable architecture also has the issue of reliability to contend with. POTS phone systems have evolved into such a state over the years that 5-nines uptime (99.999 percent of uptime) is required. Cable networks draw their power from the electrical utility companies that in most areas do not offer 5-nines reliability. Thus either the headend or CMTS in packet cable networks need to provide fault tolerance capability via equipment redundancy and power backup facilities to meet at least the "lifeline service" requirement.

REVIEW QUESTIONS

1. Describe the original motivations for developing CATV and how satellite technology has transformed the cable network industry.

2. Describe the CATV configuration and the functions performed by the headend and the distribution hub.
3. Compare and explain the differences between broadcast TV and CATV in terms of frequency allocations.
4. Explain the motivations for the development of the DOCSS standards and the main goal of the cable modem network standard.
5. Compare Fig. 16-1 with Fig. 16-2 and describe the main differences between a CATV network and a cable modem network in terms of network configuration.
6. Describe the components and functions of a CMTS and the cable modem at a customer's premises in a cable modem network.
7. Describe the additional functions that a CMTS performs in a packet cable network.
8. Describe the four categories of components in the packet cable reference architecture and compare them to the DOCSIS architecture as shown in Fig. 16-2.
9. Describe the media gateway in a packet cable network and the functions it performs.
10. Describe the kind of mechanisms adopted in the packet cable architecture to address the IPO QoS issues.

REFERENCES

CableLabs. 1996. "Data Over Cable Interface Specifications: Cable Modem Termination System—Network Side Interface Specification." Cable Labs DOCSIS 1.0. SP-CMTS-NSII01-960702. Web site: www.cablemodem.com.

CableLabs. 1998. "What Is PacketCable?". White paper. Web site: www.packetcable.com.

CableLabs. 1999. "PacketCable™ 1.0 Architecture Framework." PacketCable Lab Technical Report PKT-TR-ARCH-V01-991201. Web site: www.packetcable.com.

CableLabs. 2000a. "PacketCable 1.2 Architecture Framework—Technical Report." PKT-TR-Arch1.2-v01-001229. Web site: www.packetcable.com.

CableLabs. 2000b. "PacketCable™ 1.2 Architecture Framework." Packet-Cable Lab Technical Report PKT-TR-ARCH1.2-V01-001229. Web site: www.packetcable.com.

CableLabs. 2000c. "VoIP Availability and Reliability Model for the PacketCable™ Architecture." PacketCable Lab Technical Report: PKT-TR-VoIPAR-V01-001128. Web site: www.packetcable.com.

CableLabs. 2002a. "DOCSIS 2.0.—Data over Cable Service Interface Specification: Cable Modem to Customer Premises Equipment Interface Specification." SP-CMCI-107-020301. Web site: www.packetcable.com.

CableLabs. 2002b. "DOCSIS Cable Modem to Customer Premise Equipment Interface Specification." SP-MCI-101-020301. Web site: www.packetcable.com.

Ciciora, W., Farmer, James J., and Large, D. 1998. *Modern Cable Television Technology: Video, Voice and Data Communication.* San Francisco: Morgan Kaufmann.

IEEE 802.14 Working Group. 1995. "Cable-TV Functional Requirements and Evaluation Criteria." IEEE 802.14 Standard Proposal. Web site: www.ieee.org.

ITU-T. 1998. "Transmission systems for Interactive Cable Television Services." ITU-T Recommendation J.112. Web site: www.itu.int/ITU-T.

ITU-T. 2002a. "IPCablecom Signaling Transport Protocol." ITU-T Recommendation J.165. Web site: www.itu.int/ITU-T.

ITU-T. 2002b. "IPCablecom Trunking Gateway Control Protocol (TGCP)." ITU-T Recommendation J.171. Web site: www.itu.int/ITU-T.

ITU-T. 2000. "The Directory: Public-Key and Attribute Certificate Frameworks." ITU-T Recommendation X.509. Web site: www.itu.int/ITU-T.

Microsoft. 1999. "Cable Architecture." Microsoft white paper. Web site: www.microsoft.com.

PART 4

Next-Generation IP Networks

The IP network of a decade from now probably will no longer be granddad's IP network any more. The best-effort, free-for-all, hop-by-hop routing IP network is being replaced by a QoS-based, more manageable IP. The new IP network infrastructure is starting to take shape as the Internet moves toward becoming a generic, multiservice communications network. This part of the book introduces three key components of the new IP network infrastructure: multiprotocol label switching, IP QoS architecture, and QoS policy provisioning, as shown in Fig. P4-1.

MPLS represents a fundamental extension to the original IP network, with two major changes to the existing infrastructure. First, it adds controllability to IP networks. An IP network is much like a "free-for-all" highway without traffic control. All the traffic can be crammed onto the highway at once and each router along the way tries its best to get the traffic through without any guarantee of doing so. MPLS marks "lanes" with labels for the IP highway, and each packet flow has to follow a predefined lane or path. MPLS reduces the randomness and adds controllability to the old IP network. Second, MPLS adds switching capability to routing-based IP networks. Traditional IP networks have every router along the way examine the destination address inside a packet and determine the next hop. In a switched network, each switch routes the traffic from the input port into a predetermined output port without examining the contents of each packet. The benefits of this change include speedup of the network traffic and scalability of networks.

The IP QoS architecture represents another major extension to the original IP network. IP networks as originally defined provide "best-effort" services without any QoS guarantee. This is no longer satisfactory for real-time-sensitive applications like IP telephony and multimedia services that the next-generation IP network is called upon to support.

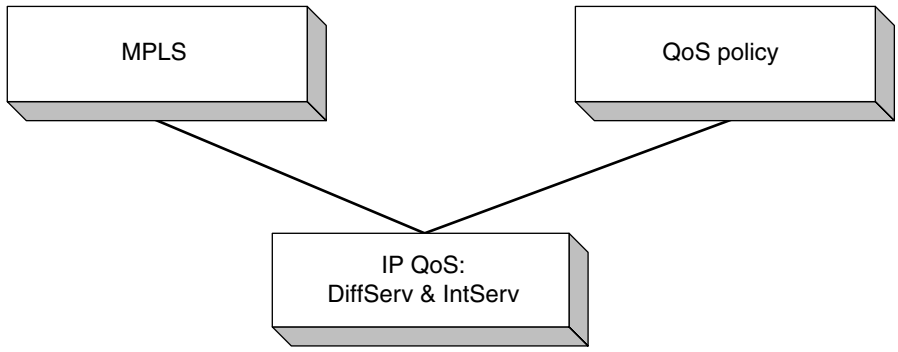


Figure P4-1
The components of new IP infrastructure.

Part IV of this book introduces two of the most commonly deployed IP QoS models: DiffServ model and IntServ model.

QoS policy is the last piece of the end-to-end QoS puzzle. QoS policies address the issues of how a user requests service with a QoS guarantee, how a user request is mapped to a network resource allocation policy, and how a policy is carried out by network devices. Part IV discusses the architectural framework of QoS policy provisioning.

CHAPTER 17

Multiprotocol Label Switching Networks

17.1 MPLS Basics

17.1.1 Introduction

IP networks were initially designed with network survivability in a decentralized networking environment as the central goal. Thus the Internet infrastructures and protocols were intended from the very beginning for this purpose. As the Internet is evolving into a general-purpose communications network, the new realities require the development of new Internet infrastructure to support real-time-sensitive and multimedia applications such as voice over IP and video conference calls.

MPLS is a key component of the new Internet infrastructure and represents a fundamental extension to the original IP-based Internet with two changes to the existing infrastructure.

First, it adds controllability to IP networks. As already noted, an IP network is much like a "free-for-all" highway without traffic control, to use the analogy of a highway system. All the traffic can be crammed onto the highway at once, and each router along the way tries its best to get the traffic through without any guarantee of succeeding. MPLS marks "lanes" with labels for the IP highway, and each packet flow has to follow a predefined lane or path. Once the "lanes" are marked, a set of traffic parameters can be associated with each lane to guarantee the service delivery. It reduces randomness and adds controllability to the IP network.

Second, MPLS adds switching capability to the routing-based IP network. The traditional Internet structure has every router along the way examine the destination address inside a packet and determine the next hop. In a switched network, each switch routes the traffic from the input port to a predetermined output port without examining the contents of each packet. This is also called *route once and switch many times*, since the packet contents are examined only at the entry of the MPLS network to determine a proper "lane" for the packet. The benefits of this change include speedup of network traffic and network scalability.

In another analogy, if an IP address is the street address of the packet mail, MPLS provides the ZIP code. MPLS attaches a label to a package indicating that this is first-class, high-priority mail, which also speeds up delivery.

Toward the same goals of adding controllability and "route once and switch many times" capability to IP networks, multiple proprietary

approaches were developed from the early 1990s to the mid-1990s, including the following:

- Multiple Protocols over ATM (MPoA) (ATM Forum)
- Tag switching (Cisco)
- Cell switching router (CSR) (Toshiba)
- IP navigator (Lucent)
- Aggregate route-based IP switching (ARIS) (IBM)
- IP switching (Ipsilon)

The MPLS standardization efforts started in the mid-1990s and picked up steam toward the end of the decade upon the realization that, for such a fundamental change, any proprietary approach would not achieve wide adoption without an industry-agreed-upon standard. Another driving force behind the MPLS standardization has been the deployment of high-density WDM optical networks on a large scale, which has required a traffic control mechanism to support IP traffic directly over a vast amount of bandwidth.

17.1.2 MPLS Concepts

This section describes the basic concepts that form the foundation of MPLS technology.

17.1.2.1 MPLS Label An MPLS label is "a short fixed length physically contiguous identifier which is used to identify a forwarding equivalency class (FEC), usually of local significance" (Rosen et al., 2001). A label is an extra field embedded in each IP packet header between the access layer and the IP layer (or between layer-2 and layer-3). The label is locally significant, meaning that it is meaningful only for a link between two nodes.

An MPLS label allows a router or switch to find out the next hop to which the packet is to be forwarded and the operation to be performed on the label stack before forwarding the packet to the next hop, such as pop, swap, replace the label. A label has the following fields, as shown in Fig. 17-1.

- *The label field (20 bits).* This carries the actual value of the MPLS label.
- *The experimental field (exp) (3 bits).* This can be used to indicate the class-of-service (CoS) the packet belongs to. It affects the queuing and discarding priority of the packet.

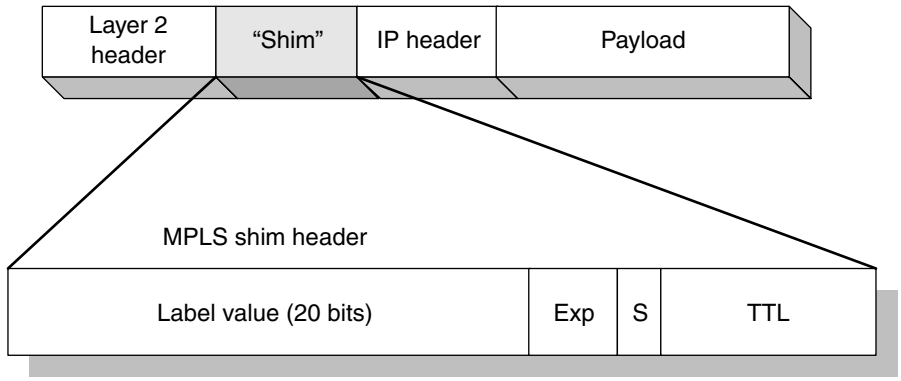


Figure 17-1
MPLS label structure.

- *The stack (s) field (1 bit).* This is intended to support a hierarchical label stack, which will be described shortly.
- *Time-to-live field (8 bits).* This provides the conventional IP TTL function to prevent endless looping of packets in an MPLS network, a function similar to the TTL field in the IP packet header. Each time a packet is forwarded in an MPLS network, the TTL value decreases by 1. When the TTL value reaches zero before the packet reaches the destination, the packet is discarded.

A packet with an MPLS label is said to be a *labeled packet,* and, conversely, a packet without a label is called an *unlabeled packet.*

A label with 20 bits in length can have a label value up to 2^{20}, with some label values reserved for special purposes. For example, the label value 0 represents the *IPv4 Explicit NULL Label,* indicating that the forwarding of the labeled packet must be based on the IPv4 header. The label value 2 represents the *IPv6 Explicit NULL Label,* indicating that the forwarding of the packet must be based on the IPv6 header.

17.1.2.2 Label Stack A labeled packet is a set of labels organized as a last-in-first-out stack inserted into an IP packet header. The stack field of the label allows multiple labels to be stacked in a packet. Label stacking serves two primary purposes: First, it allows an MPLS network to scale indefinitely in a hierarchical fashion; second, it supports MPLS tunneling, as will be described in Sec. 17.5 on MPLS applications. The topmost label on a label stack indicates the actions to be taken at the router. The labels are numbered inside out—from the bottom of the stack to the top.

17.1.2.3 Forwarding Equivalency Class An FEC is a set of packets that have the same traffic characteristics and are forwarded in the

same manner. For example, all packets in the same FEC follow the same label-switched path (LSP), i.e., the same route in an MPLS network, with the same priority and the same label. A packet is assigned an FEC at the entry point of an MPLS network. An example of an FEC is a set of unicast packets that share the same IP address prefix. Another example is a set of multicast packets with the same source and destination IP address.

17.1.2.4 Label-Switched Path A label-switched path, as shown in Fig. 17-2, is a route that starts from an ingress node and ends at an egress node of an MPLS network. An LSP consists of a number of segments across the MPLS network, each of which connects two label-switched routers (LSRs). Two LSPs are shown in Fig. 17-2: One consists of segments connecting S1, S2, S3, and S4, and the other is made up of segments connecting S1, S2, S3, S5, and S6. At a high level, an LSP may be viewed as a virtual circuit in an ATM or a frame relay network. The mechanical details of LSP will be described shortly.

17.1.2.5 Label-Switched Router and Label-Switched Edge Router A node in a MPLS network is termed a *label-switched router.* An LSR can be an ingress LSR, egress LSR, or intermediate LSR, as shown in Fig. 17-2. An LSR is responsible for setting up a label-switched path, label swapping, and maintenance of the forward information base (FIB). An ingress or egress LSR is also known as a *label-switched edge router* (LER). An LSP can be set up either statically via an operator's manual operation or dynamically via signaling messages, as will be discussed shortly. *Label swapping* is an operation to replace an incoming label with an outgoing label. A label is only significant locally between adjacent nodes, and as a packet is forwarded to the next hop, a label meaningful to the next node is swapped for the incoming one. An FIB is a database that contains the mapping between labels.

The point at which a packet enters the MPLS network is called an *ingress LSR,* and where a packet leaves the MPLS network is called an *egress LSR.* An LSR connects to a non-MPLS router on one side and to an intermediate LSR on the other. An ingress or egress LSR is also responsible for the following:

- Binding a user traffic stream to an FEC and an FEC to an LSP.
- Internetworking with heterogeneous networking technologies. An LSR may play the roles of both LER and LSR. In other words, it may serve as an LSR for some LSPs and as an LER for other LSPs.

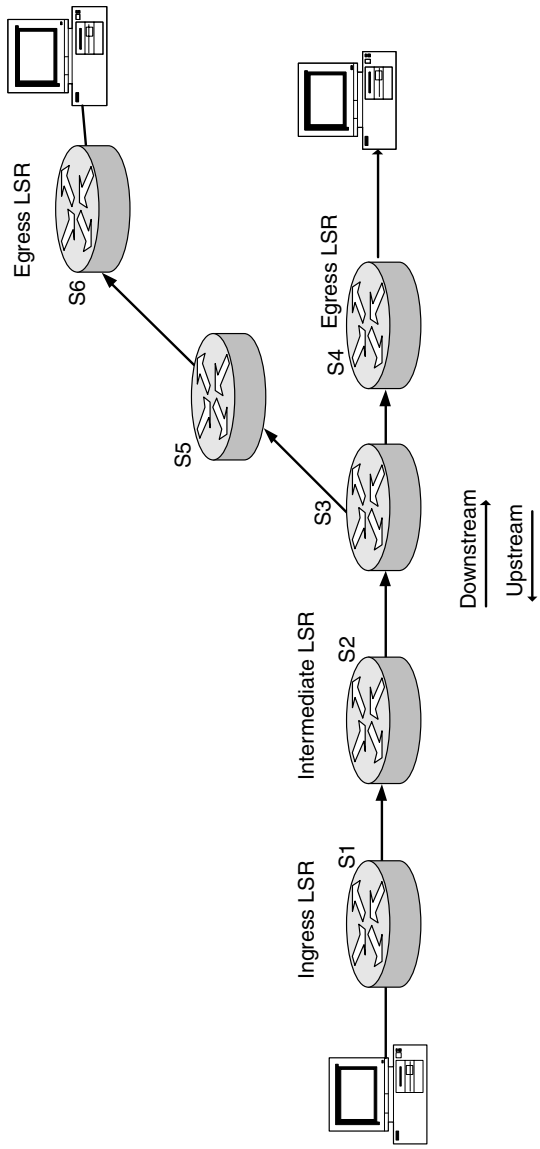


Figure 17-2 MPLS network components.

17.2 MPLS Routing

MPLS follows a key principle of the new generation of IP routers by separating the routing and control functions from the forwarding function. The routing and control functions refer to the calculation and determination of an LSP, while the forwarding function refers to the dynamic operations of forwarding packets along the defined LSP from node to node. In contrast, traditional IP networks intertwine the two and perform them at the time of packet forwarding (Black 2000; Harnedy 2001).

The separation allows label-switched paths or routes in an MPLS network to be calculated and set up before the packet forwarding operation. In a sense, the switched label path setup can be performed off-line if so desired while the forwarding of packets along the path is a run-time operation. Throughout this chapter, the term *MPLS routing* and the term *LSP setup* or LSP *provisioning* will be used interchangeably.

This section focuses on MPLS routing and control while Sec. 17.3 deals with the MPLS forwarding operations. This section first discusses the basic concepts of MPLS routing to provide a background, and then focuses on the four commonly available MPLS routing methods:

- Manual LSP setup
- Hop-by-hop label distribution using LDP
- Label distribution using Label Distribution Protocol—constraint restricted (LDP-CR)
- Traffic engineering- (TE-) based LSP setup

17.2.1 MPLS Routing Concepts

There are four general tasks involved in determining an LSP, although not all steps are required for every LSP. Depending on the method used, some LSPs require all four steps, while others require fewer.

1. *Initiating the LSP setup process.* This step initiates the process of an LSP setup, which can be initiated either by an ingress LER or an egress LER.
2. *Determining the next hop of the LSP.* This step may require discovering the network topology.
3. *Label assignment and mapping.* This step assigns a label and maps it to a forwarding equivalency class.

4. *Label distribution.* This step distributes an assigned label to the next hop on the defined LSP.

An LSP can be set up in either the downstream or upstream direction, depending on the direction a label is sent from one LSR to another. If LSR A sends a label to LSR B, A is said to be the upstream node of LSR B and B is said to be the downstream LSR of A.

17.2.1.1 Two Modes of Label Distribution There are two modes for setting up a LSP: downstream-on-demand and downstream unsolicited. In the downstream-on-demand mode, a downstream LSR requests a label from the upstream LSR for a particular FEC. In the downstream unsolicited mode, an upstream LSR distributes a label-to-FEC binding to a downstream LSR that has not made an explicit request for the binding. This is normally used in the network topology-driven LSP setup approach, where the position of an LSR in the network determines whether it receives a label and is part of an LSP.

17.2.1.2 Explicitly Routed LSP and Topology-based LSP There are two general approaches to LSP setup: topology-based and explicitly routed LSP. Thus, there are two types of LSPs: explicitly routed LSP and topology-based LSP. For explicitly routed LSP, the path information is either partially or completely determined at the ingress LER where the LSP setup process is initiated. The labels are sent to the downstream LSRs in the downstream-on-demand mode, and a label-request packet from a downstream LSR to an upstream LSR carries the path information for the downstream LSRs that the LSP traverses.

On the other hand, the topology-based LSP is set up in a hop-by-hop fashion in the sense that each LSP finds its downstream and upstream LSR for an LSP using its local routing table.

17.2.1.3 Label Binding to an FEC Binding a label to an FEC means to determine the type of traffic that will travel on the label switched path. An FEC is a logical grouping of packets that share the same characteristics such as destination or source address prefix and traffic characteristics (e.g., QoS parameters, class-of-service, etc). This is a service policy decision specific to each service provider. Two general approaches to the label binding are specified in RFC 3031 (Rosen et al., 2001):

- *Independent approach.* Each LSR, upon recognizing a particular FEC, makes an independent decision to bind a label to an FEC based on the policies and other considerations specified by the service provider and to distribute the binding to the associated LSRs.

- *Ordered approach.* An LSR binds a label to an FEC only if it is the egress LSR for that FEC or if it has already received a label binding for the FEC from its next hop. This approach must be used if an FEC is associated with a set of traffic properties such as QoS parameters.

Binding a label to an FEC may or may not be one-to-one. A label can be bound to one FEC or a set of FECs. In other words, multiple FECs can share the same LSP. In the case of binding one label to multiple FECs, the FECs are aggregated into another FEC, the members of which are other FECs. The multiple labels can also be bound to a single FEC.

17.2.1.4 Forwarding Information Base A forwarding information base (FIB) contains information such as the labels and the associated bindings at each local LSR. The MPLS FIB uses the syntax of the SNMP management information base (MIB). Examples of FIB contents include the following:

- *The next-hop label forwarding entry (NHLFE).* This consists of the next hop for the labeled packet.
- *The incoming label map (ILM).* This indicates that the action is to map each incoming packet's label to a set of NHLFE labels.
- *The FEC to NHLFE map (FTN).* This indicates that the action is to label and forward unlabeled packets. Each FEC is mapped to a set of NHLFEs.
- *Label swapping.* This field indicates the action is to combine all the above three actions: forwarding, ILM and FTN. The exact action depends on whether the packet is labeled.

17.2.2 Manual LSP Setup

Manual LSP setup is similar to setting up an end-to-end PVC in an ATM network: Manually set up one leg of a path at a time. This is done by a network-level management system. The general steps of manual LSP setup include the following:

1. Starting from the ingress node, a label is automatically generated or manually created for the immediate downstream LSR, and then assigned to an FEC. The FIB at the ingress node is updated to reflect the change.
2. At the immediate downstream LSR, a new label and label binding pair is generated as an outgoing label for the incoming label from the ingress node. The local FIB is updated to store the mapping

from the incoming label to the outgoing label. In the same manner, an LSP is set up at each downstream LSR.

3. At the egress LER, only the incoming label is created.

Manual LSP setup is a time-consuming and potentially errorprone process, but useful for configuring a test environment or for getting around the problem with other methods of LSP setup. It can be useful when there are special considerations for a particular LSP that cannot be accommodated by any other provisioning methods.

17.2.3 LDP and Hop-by-Hop-Based LSP Setup

A Label Distribution Protocol is defined in IETF RFC 3036 (Andersson et al., 2001) for discovering peer LSR nodes and establishing a session specifically for LSP setup.

LDP allows an LSR to distribute labels to its LSR peers. When an LSR assigns a label to an FEC, it uses LDP to distribute the label and its FEC binding. LDP is an integral part of MPLS, and has the following characteristics:

- It provides a mechanism for a LSR to discover peer LSRs and establish a connection for the purpose of LSP setup.
- It defines a set of messages for peer LSR discovery and for label distribution. The messages include DISCOVERY, ADJACENCY, LABEL ADVERTISEMENT, and NOTIFICATION.
- It provides reliable message delivery using TCP.

LDP supports both unsolicited downstream and downstream-on-demand label distribution modes and both independent and ordered label binding approaches. However, pure LDP-based LSP setup does not support traffic engineering or QoS parameters.

LDP supports network topology-based, hop-by-hop LSP setup. Using LDP, each LSR finds its downstream and upstream nodes for an LSP using information from its local routing table. An LSP is set up based on a network prefix. Each LSR assigns a label to this network prefix and distributes it to its upstream neighbor.

The general steps of hop-by-hop LSP setup using LDP include the following:

1. The LDP route advertisement mechanism builds up a network topology data at each local LSR.

2. LDP then establishes sessions with LSR peers as it discovers their adjacency. This is accomplished with UDP, and at the same time the router's routing table is updated with the adjacency information.
3. LDP then distributes labels for each route to each peer with which it has established an LSP provisioning session using TCP if the provisioning mode is downstream unsolicited.
4. Alternatively, the local LSR requests labels using TCP from the next-hop LSR peer for each entry in its routing table if the provisioning mode is downstream-on-demand.

17.2.4 LDP-CR and Constraint-Based LSP Setup

LDP-CR (constraint restricted) is an extension to the LDP with traffic engineering capability. As with LDP, LDP-CR supports both strict and loose explicitly routed LSPs. It uses UDP for discovering MPLS peers and TCP for control, management, and label requests and distribution (Jamoussi 2002).

LDP-CR extends the standard Interior Gateway Protocol (IGP) to allow link metrics, such as the total link bandwidth, the available link bandwidth, and link "colors" (i.e., CoS-based resources), to be exchanged between routers. This information is then used to calculate the traffic-engineered LSP based on an heuristic algorithm.

The LDP-CR traffic engineering extension to the LDP feature sets are fairly extensive:

- *QoS and traffic parameters.* LDP-CR provides the ability to define per-hop behavior based on the traffic engineering parameters such as data rates, link bandwidth, and weight factor given to those parameters.
- *Support for both strictly and loosely explicit routing.*
- *Path preemption.* With this ability, a LSR can set priority to either allow or prohibit preemption of one LSP by another LSP.
- *Path reoptimization.* This ability allows an LSR to reroute the loosely routed LSPs based on traffic pattern changes.
- *LSP pinning.* This ability allows an LSP to be "nailed down," preventing any change to it.
- *Failure notification and failure recovery.* This allows an LSR to notify the ingress LSR of the failure of an LSP setup process.

- *Failure recovery.* This allows the operator to define mapping policies to reestablish an LSP for automatic failure recovery at each node of a failed LSP.

The steps of LSP setup using LDP-CR are similar to those for LDP, as shown in Fig. 17-3:

1. First, before the LSP setup process is initiated, it is already determined at the ingress node S1, based on policy or traffic engineering parameters, that the LSP consists of ingress node S1, intermediate nodes S2 and S3, and egress node S4.
2. The ingress LSR first generates a label and label binding for the LSP. It then initiates the setup process by sending to the next hop LSR a LABEL_REQQUEST message containing the explicit S2, S3, and S4 route and the traffic engineering parameters. S1 also reserves the resource required for the LSP, as specified in the traffic engineering parameters.
3. S2, upon receiving the LABEL_REQUEST message, generates a label and a label binding, reserves the resource as required for the LSP, and forwards the message to the next hop LSR S3, once it realizes it itself is not the egress node of the LSP. S3 performs the same routines and forwards the message to S4.
4. S4, upon recognizing that it is the egress node, makes the reservation of required resources, generates a new label for the LSP, and then builds a LABEL_MAPPING message containing the details of the agreed-upon traffic parameters for the LSP. S4 sends the message back to S3.
5. In a similar manner, the LABEL_MAPPING message is forwarded upstream from S3, to S2 and finally to S1. Each node notes down the agreed-upon traffic parameters and adds a new label for the next leg of the path.

An LDP-CR-based LSP setup results in so-called *hard states* of the LSP, meaning that all the information about the paths are exchanged at the initial path setup time and no additional information exchange is required between peer LSRs until the LSP is torn down. When the operator decides that an LSP is no longer needed, messages must be exchanged to notify all nodes of the path that the path can be destroyed and the allocated resources reallocated.

One challenge for the LDP-CR approach to LSP setup is the heavy cleanup after a provisioning TCP session fails. All LSPs established with a particular TCP session must be torn down if the TCP session fails. The

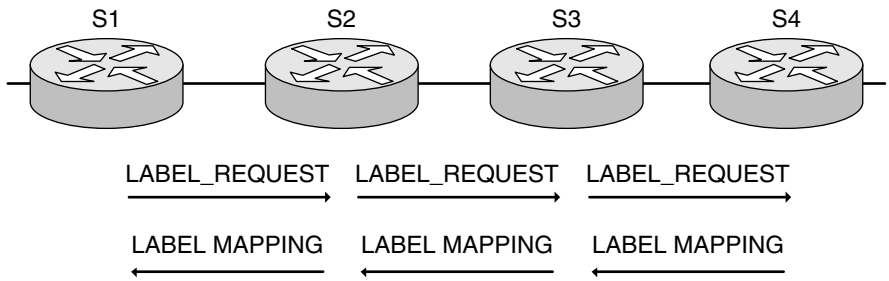


Figure 17-3
Illustration of LSP provisioning using LDP-CR.

impact can potentially be very significant if the number of LSPs set up with a TCP session is very large.

17.2.5 LSP Setup Using RSVP—TE

RSVP provides an application with the ability to reserve resources like bandwidth and other QoS parameters on a traditional "best-effort" IP network. RSVP-traffic engineering (TE) is an extension to RSVP with traffic engineering capabilities specifically for MPLS (Awduche 2001).

RSVP-TE-based LSP setup assumes that the destination LSR of the LSP is already known before the LSP setup process is initiated. The label assignment and distribution proceed backward from the egress LSR toward the ingress LSR, as illustrated in Fig. 17-4. The general steps can be summarized as follows:

1. Before the provisioning process is initiated, it is already determined based on QoS policy or traffic engineering parameters, that the path consists of the ingress node S1, intermediate nodes S2 and S3, and egress node S4.
2. The ingress LSR initiates the LSP setup process by sending a path request message to the destination or egress LSR.
3. In response to the path request message, the egress LSR first assigns a label for the segment of the path between the local LSR and S3 and distributes the label with the associated FEC binding to LSR S3.
4. LSR S3 in turn assigns a label value (label = 30) for the segment of the path between LSR S3 and LSR S2. In the same fashion, LSR S2 assigns and distributes the label to the ingress LSR.
5. The ingress LSR, after receiving its label assignment and binding, issues a RESV_conf message to confirm the LSP reservation, which completes the LSP setup.

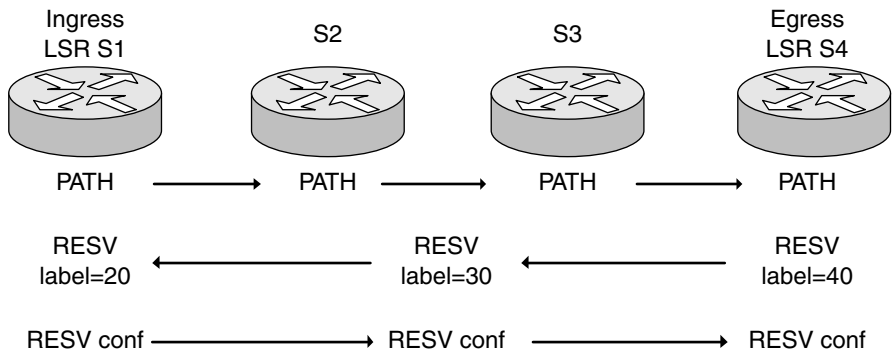


Figure 17-4
Illustration of LSP setup with RSVP-TE.

An LSP established using RSVP-TE is an explicitly routed LSP, generally for the traffic engineering purpose, to allow traffic flows to be routed away from the shortest path calculated by the routers using the conventional shortest-path routing algorithm.

RSVP-TE-based LSP setup, in contrast to LDP-CR-based LSP setup, results in so-called *soft states* of the LSPs, meaning that after the path is set up, refresh messages must be periodically exchanged between the peers to reaffirm the provisioned state of the LSP. Otherwise, a maintenance timer senses the path is dormant and causes the path state information to be deleted and allocated resources to be reclaimed. The advantage of this self-cleaning approach is that all dormant and expired resources can be reclaimed. The downside is that the overhead incurred from the refresh messages can be considerable, especially when the path involves many LSR nodes and the number of LSPs is large. Efforts are underway at IETF to reduce the number of refresh messages needed and thus alleviate the overhead.

An LSP setup as just described is unidirectional. In order to have bidirectional communications, two LSPs need to be set up. One way of setting up a bidirectional LSP is that the egress LSR, upon receiving an LSP request message, sends a response as well as an LSP request in the opposite direction.

17.2 MPLS Forwarding

This section describes the operations of forwarding packets along an LSP that has been set up using one of the methods described above. Again, one key concept of MPLS is the separation of the packet routing and control functions from the packet forwarding function. The separation

makes the packet forwarding relatively simple, because the routing decision is already made at the time when an LSP is set up and all that needs to be done is to forward the packet along the LSP (Rosen et al. 2001).

This section describes three basic types of operations involved in forwarding a packet:

- *Label binding.* Assigning an unlabeled packet with a label and pushing the label onto the label stack of the packet.
- *Label swapping.* Replacing the label from an incoming packet with a label for sending the packet out to the next leg along an LSP.
- *Label popping.* Deleting a label before sending a packet to a non-MPLS node.

The type of operation to perform at each node very much depends on whether the node is an ingress, egress, or intermediate LSR node in the MPLS network.

17.3.1 Label Binding at Ingress LER

The ingress LER, at the edge of an MPLS network, is responsible for generating a label for each unlabeled packet coming in from a non-MPLS router and then binding the label to a label switched path. This process is known as *label binding.*

The procedure of label binding, the process of associating user traffic to a specific user class of service, has the following general steps:

1. Processing the packet header information to decide whether or not the packet is labeled. If the packet is already labeled, then going to the label swapping operation described in Sec. 17.3.2. Otherwise, proceeding further to step 2.
2. Classifying an incoming packet stream into categories, associating the packets to an FEC, and then assigning the associated label.
3. Finding the LSP associated with the label from the FIB and forwarding the packet to the next node along the LSP.

How to associate a packet stream of a user to a FEC is beyond the MPLS forwarding operation itself. It is a QoS policy issue and depends on the fact that a user with a particular service level agreement (SLA) has already been assigned an FEC either through SLA provisioning or other means. Another assumption is that the ingress LSR can recognize the user traffic once it arrives, using IP layer information such as source

and destination IP addresses, even in cases where the IP address is dynamically assigned.

The default label is applied to those packet flows without QoS guarantees. All traffic going into an MPLS network must be assigned a label. The packet flows from those users without an SLA are assigned a default label and thus the default FEC and are treated at an MPLS node with the best-effort service of traditional IP networks instead of guaranteed QoS.

An example, shown in Fig. 17-5, will help illustrate the label binding operation. The ingress LSR receives a packet with the destination IP address starting with 60.50, performs one table lookup, and finds FEC 15 defined on the destination IP address prefix 60.50. In addition, the table lookup also yields a label value 40, an action to perform on the label, and an outgoing port number 1. The label value is assigned into the packet MPLS header. The push action means to push the label onto the label stack of the packet. Then the packet is sent to the next hop along the LSP from the designated port 1.

17.3.2 Label Swapping at Intermediate LSR

An intermediate LSR simply performs a label swapping. Label swapping is the process of swapping the incoming label for an outgoing label along the assigned LSP. The mapping of an incoming label to an outgoing label is stored in the FIB.

Continuing the example in Fig. 17-5, Fig. 17-6 shows the operation on a packet at an intermediate LSR. When the packet arrives at the local port 2 with label 40, a table lookup yields an outgoing port 10 and an outgoing label 8. The action SWAP means replacing the incoming label value 40 with label value 8. Then the packet is sent out via port 10 to the next hop along the LSP.

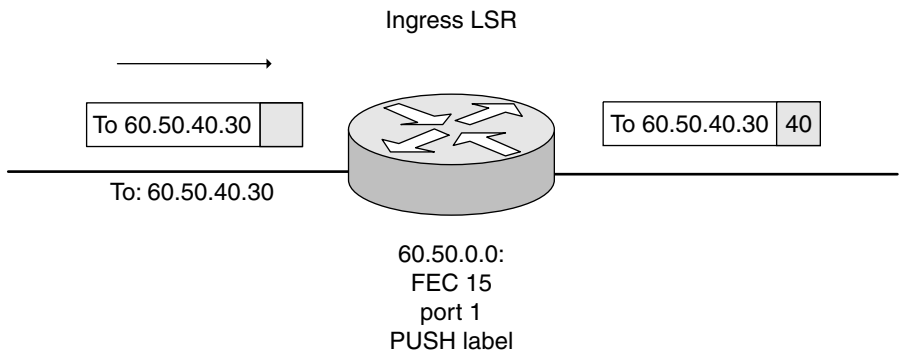


Figure 17-5
Illustration of label binding operation.

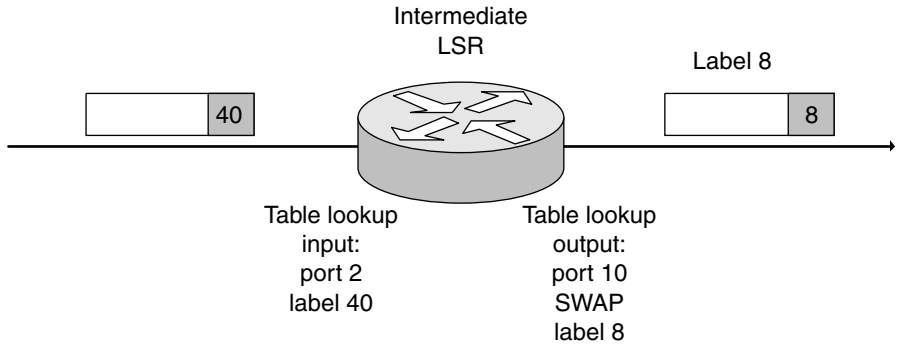


Figure 17-6
Example of a label swapping operation.

One challenge at an intermediate LSR is to deal with the label merging issue. In the case of multiple labels binding to one FEC, packets of the same FEC may come into an LSR with different labels but go out with the same label. The capability of merging multiple incoming labels into one outgoing label is known as *label merging*, and an LSR with such a capability is said to be *label merging-capable*. Not all LSRs are merging-capable. For example, the MPLS-capable ATM switches do not support label merging because each LSP is mapped to an individual virtual circuit. For nonmerging labels, packets of the same FEC that come from different interfaces are forwarded with different labels.

17.3.3 Label Popping at Egress LER

The egress LSR, at the edge of an MPLS network, is the last stop of the LSP where a labeled packet is converted to an unlabeled one, and the "shim header" of the label is cleared out. In other words, the egress LSR must undo what the ingress LSR did on the packet. There are two primary responsibilities of an egress LSR:

- Examining the label and following the action specified in the table, which is to pop the label off the label stack. This in effect cleans the label.
- Determining the next hop in a non-MPLS fashion, i.e., by conventional hop-by-hop routing.

The two distinct tasks require two table lookups, one for the FIB table lookup and one for the routing table, if the egress node is to behave the same way as other LSR nodes. The first table lookup, using the label value from the incoming packet, yields the label action pop, i.e., cleaning the label stack. Then the lookup of the routing table, using the destination IP address, yields the outgoing port and the next-hop IP address.

Penultimate hop popping (PHP) is a scheme developed to alleviate the performance penalty caused by the two table lookups at an egress LSR. A penultimate LSR is an LSR that is immediately before the egress LSR along a LSP. The PHP requires that the label action at the penultimate LSR be provisioned as a pop instead of a swap, as it normally would be. So when a labeled packet arrives at the penultimate LSR, the label is cleaned and the egress LSR only needs to perform one table lookup to find the outgoing port and the next hop. This approach puts an extra burden on LSP routing and may not always be feasible.

A variant of the PHP is to have the penultimate LSR behave the same way as other regular intermediate LSRs, i.e., swapping the label instead of popping it. The label value the penultimate node assigns for the last leg of the LSP has a special meaning: An explicit null value 0 signals to the egress LSR that it can ignore the FIB table and perform only the routing table lookup.

17.4 Generalized MPLS

Generalized MPLS (GMPLS) is the extended version of MPLS intended to support IP traffic directly over optical networks. As optical networks experience rapid growth, it becomes attractive to simplify the network layers by bypassing other layers like ATM or SONET and directly carry IP traffic over fiber networks. To that end, MPLS is being extended to accommodate optical switching as well as other networking technologies like TDM and SONET.

Generalized MPLS is a proposed standard still under study at IETF, and this section provides an overview of its basic concepts (Mannie 2002).

17.4.1 Generalized MPLS Label and Label-Switched Path

GMPLS extends the concept of label and label-switched path. In regular MPLS, a label is a number of up to 32 bits that identifies a traffic flow. The label itself does not necessarily imply a relationship to bandwidth allocation or quality of service for the traffic flow.

Generalized MPLS extends the label concept to anything that is sufficient to identify a traffic flow type in order to encompass other technologies within the label-switched scheme (Banerjee et al. 2001; Mannie

et al. 2002). Many other technology-specific identifiers, other than abstract numbers, can serve as labels:

- *Whole fiber label.* This label identifies one fiber out of a bundle of potentially many fibers between two label-switched nodes for data flow. The interpretation of the fiber number is a local matter between the two nodes connected by the fiber link.
- *Wavelength or lambda label.* This label identifies a wavelength out of a number of wavelengths that result from wave division multiplexing. A wavelength is also known as *lambda.*
- *Waveband label.* This label identifies a waveband, a grouping of consecutive wavelengths to be switched together.
- *Timeslot label.* This label identifies a time slot of a fiber link resulting from time division multiplexing. A TDM label value is sufficient to specify the allocated time slots.
- *SONET/SDH label.* This label is represented as a sequence of five numbers, known as S, U, K, L, and M, which selects branches of the SONET/SDH TDM hierarchies at increasingly fine levels of details.
- *Packet label.* This is the conventional label (32 bits) representing a packet flow between two LSRs.

The proposed generalized label format is a bite array of variable length. The length of each label is determined by the type of label. For example, the wavelength label has a length of 32 bits, indicating the fiber, lambda, or port being used from the sender's perspective.

17.4.2 GMPLS Signaling and Routing for LSP Setup

The LDP, LDP-CR, and RSVP-TE-based LSP setup procedures as described above remain the same for GMPLS. But there are a number significant extensions to the regular MPLS, including the concept of directional LSP, out-of-band signaling, label set, and explicit label control.

17.4.2.1 Bidirectional LSP A significant extension of GMPLS is that each label-switched path is bidirectional rather than unidirectional. An LSP of conventional MPLS, set up as described in the earlier sections, is unidirectional. In order to have bidirectional communications, two independent LSPs need to be set up.

GMPLS achieves bidirectional LSPs by extending the RSVP-TV and LDP-CR signaling messages. The basic idea is that, at the time a leg of forward LSP is set up at each LSR, the LSR also negotiates with the upstream LSR to set up the same leg of LSP in the reverse direction. When the forward LSP is set up, an LSP of the reverse direction is established as well. Optical links are normally bidirectional, and there is a need to model the label-switched path as bidirectional links.

17.4.2.2 Out-of-band Signaling Another significant extension is the out-band signaling of GMPLS. LSPs of conventional MPLS are set up using in-band signaling, meaning that the signal message and the data payload travel the same path. For optical networks, there is a need to have a separate signaling channel for the label setup. One reason is that a separate signal channel separates the data plane from the control plane so that the optical switch does not need to understand the signaling protocols for setting up a label-switched path. This simplifies the technology. Another reason for the extension is to save bandwidth. In optical networks, the finest granularity of an optical link is a wavelength or a time slot of time division multiplexing on the optical link. In either case, it is wasteful to use a whole optical channel for a signaling message. The signaling message can follow a separate, lower-speed channel.

The out-of-band signaling approach is not without its own challenges. The key issues that need to be addressed include the following:

- *Routing calculation.* The routing function of GMPLS needs to calculate two routes, one for the signaling path itself and one for the data path on which user payload data flows.
- *Data interface identification.* Since the signaling messages do not travel on the same path where data packets flow, they need a way to indicate the data interface for setting up a LSP.

The first issue is addressed by extending the existing routing protocol such as Optimal Shortest Path First (OSPF). The second issue is tackled by extending the IP addressing scheme in the label to include a link ID, interface ID, or port interface.

17.4.2.3 Label Set GMPLS introduces the concept of label set to provide more control for setting up an LSP. An upstream LSR can include a label set, a group of labels, in its signaling request to restrict the downstream LSR's choice of labels for the link in between. The downstream LSR must select a label from the given label set or the LSP

setup fails. This restriction is necessary because an optical LSR may have restrictions in dealing with wavelengths or wavebands. For example, an optical LSR may be unable to convert one wavelength to another, or only be able to receive a subset of the wavelengths that can be switched by the neighboring LSRs.

17.4.2.4 Explicit Label Control GMPLS enhances the concept of explicit route of regular MPLS and introduces the concept of *explicit label control*. The basic idea here is to allow the ingress LSR to specify the labels to be used on one, some, or all of the explicitly routed links for a label-switched path. This is needed because the ingress LSR may desire that the given wavelength be used along the whole path or part of the path to avoid distortion of optical signals or for other reasons.

17.5 MPLS Applications

MPLS, an important extension to the original IP protocol, is an "enabling" technology that provides an infrastructure to enable a variety of services, especially those that are real-time-sensitive and have specific resource requirements. Prominent among the MPLS-enabled services are VPN, traffic engineering, and voice over MPLS (Davie and Rekhter 2000).

17.5.1 VPN

Virtual private networks, which provide the intranet and extranet services to enterprise customers, is a prime application for MPLS. MPLS provides two key components of VPN: a tunneling mechanism and data security. See Chap. 21 for more details on VPN.

There are three basic components of a VPN (Gleeson 2000):

- Opaque transport, also known as *tunnels*, that carry customer data between VPN sites and encapsulate the customer data regardless of the protocol the customer uses on the VPN
- QoS capability for bandwidth and latency guarantees to meet business requirements
- Data security on VPN to prevent misdirection, modification, or snooping of the customer data

17.5.1.1 LSP for VPN Tunneling and Data Security A key VPN component is the tunneling mechanism that connects the multiple VPN sites, and separates customer data from public data while oblivious to the contents of the customer data and protocol at the customer site to carry the data on the VPN.

Label-switched paths are ideal to serve as VPN tunnels on a public IP network. An LSP automatically separates one user's data flow from other traffic by labeling each customer's data packets uniquely at the ingress LSR of an LSP and by admitting only the data belonging to a particular customer into the LSP tunnel (Gleeson et al. 2000; Brittain and Farrel 2000). More specifically, MPLS provides the following VPN mechanisms:

- Multiple user protocols on MPLS VPN can be encapsulated by the tunnel ingress LSR when the data is first admitted into the tunnel since the intermediate LSRs do not look into the user data packets.
- Multiplexing of different VPN data flows onto shared public backbone links is achieved by using separate LSP tunnels, one for each VPN.
- The VPN QoS requirement can be assured by reserving resources at the time the LSP is set up. An LSP can be associated with traffic engineering parameters such as bandwidth and QoS parameters because MPLS supports QoS schemes like Differentiated Services (DiffServ), Integrated Services (Intserv), and RSVP.
- LSP tunnels provide inherent data protection by encapsulating the user data inside a tunnel. Also, since intermediate LSRs do not need to look into the user data packets, the user data can be encrypted without any extra burden on the network in forwarding it.

17.5.1.2 The Benefits of MPLS VPN There are several advantages associated with MPLS-based VPNs. Foremost is economic efficiency. Because MPLS VPNs are built on public IP networks, their cost can be much lower than some deployed alternatives such as leased lines, ATM, or frame relay-based VPNs.

MPLS VPNs are scalable with an increasing number of sites and members within each site. This is possible because labels can be stacked to build a hierarchical VPN encompassing a large number of sites and a large number of nodes within each site. MPLS VPNs support an "any-to-any" connection model among multiple sites of a customer within a VPN without actually building a fully meshed network. LSP tunnels

are built over IP networks, and the tunnels can reach wherever there is IP connectivity.

17.5.2 Traffic Engineering

Another major application of MPLS is traffic engineering. Traditional IP networks route traffic based on a shortest-path algorithm. This can result in traffic congestion by directing all traffic between an ingress node and an egress node through the same links. Traffic engineering optimizes network resource utilization by directing traffic flows along the engineered paths and allows a high degree of control over the paths that user data packets travel.

MPLS with the LSP and MPLS routing mechanisms provides the following TE capabilities (IEC 2001):

- The ability to set up an explicit route or LSP starting from the source. This is the key to any traffic engineering.
- The ability to compute a path with a variety of constraints such as resource allocation and administrative policy taken into consideration by allowing the ingress node to obtain the knowledge via a protocol like LDP-CR.
- The ability to distribute the routing constraints associated with a link throughout the network once the path is computed. LDP, LDP-CR, and RSVP-TE of MPLS all provide this ability (Awduche et al. 2001; Brittain 2000a).
- The ability to reserve network resources according to the link constraints.

TE capabilities have many uses in IP networks. For example, LSPs can be set up to achieve traffic redirection. One kind of traffic (e.g., BGP traffic) can be directed to follow one LSP while another type of traffic (e.g., IGP traffic) follows a different LSP or normal hop-by-hop routes. LSP can be set up to avoid the congested area, utilize the underutilized network resources, and achieve load balancing.

17.5.3 Voice over MPLS Networks

MPLS provides the traffic engineering and QoS capabilities that in turn address some of the main issues facing voice over traditional IP networks such as jitter and delay.

There are three different approaches to carrying voice over MPLS networks (Kankkunen et al. 2000):

- *Permanently set up LSPs as dedicated circuits for voice traffic.* This is similar to ATM PVC or circuit emulation services.
- *Set up LSPs in the manner of a "switched virtual circuit" for voice traffic.* An LSP is set up for each individual call and torn down when the call is completed.
- *Transport voice sample packets over MPLS without IP headers.* This approach can complement the other two to reduce the delay in encoding the IP packet headers and to reduce the amount of data to be carried over a network.

The voice-over-MPLS application is still at an early stage of development. One main reason for the relative lack of attention to it is that the bottleneck for the voice over IP application is at the local loop or access network while the initial focus of MPLS is on the network core. As MPLS moves toward the network edge, and as QoS is applied at the edge, the voice-over-MPLS application will eventually mature.

17.5.4 MPLS and IP QoS

MPLS itself does not provide a QoS mechanism but is designed to go hand-in-hand with standard IP QoS mechanisms to provide QoS to the traditional best-effort IP networks.

Currently there are two standard QoS schemes for IP networks: DiffServ and IntServ/RSVP. MPLS provides the basic infrastructure to work with either of these schemes to provide QoS for broadband IP packet networks, although the MPLS-DiffServ combination is a more common choice because of the simplicity and scalability of the solution (Faucheur 2002). Details on DiffServ and IntServ can be found in Chap. 18.

One common approach to the support of DiffServ within MPLS is to use a small number of service classes that are easy to manage and provision, despite the fact that DiffServ can support a maximum of 64 classes of service. The three proposed classes of services are the following:

- High-priority, low-latency premium class—Gold Service
- Guaranteed-delivery mission-critical class—Silver Service
- Low-priority best-effort class—Bronze Service

REVIEW QUESTIONS

1. Describe the fundamental issues MPLS is intended to address and the reasons for the standardization of MPLS.
2. Describe the fundamental aspects of the original IP network the MPLS is designed to change.
3. A fundamental concept of the new generation of IP networks is the separation of the routing function from the forwarding operation. Discuss the advantages and significance of this separation.
4. Describe the basic concepts of label, label-switched path, label binding, and forwarding equivalence class.
5. Define the MPLS routing function and the main tasks involved in an LSP setup.
6. Describe the differences between an explicitly routed LSP and a topology-based LSP and between a downstream-on demand LSP setup and a downstream unsolicited LSP setup.
7. Discuss the characteristics of LDP and the main LDP-CR extensions to the LDP. Then describe LDP-based hop-by-hop routing and LDP-CR-based routing.
8. Compare LDP-CR-based routing with RSVP-TE-based routing and describe the main differences between the two.
9. Describe three major MPLS forwarding operations—label binding, label swapping, and label popping—and where along an LSP each operation is performed.
10. Discuss the reasons for the development of GMPLS and how it extends the label concept.
11. Describe how MPLS LSP provides the tunneling and security mechanisms for the VPN application and some of the advantages of MPLS VPN.
12. Describe how MPLS routing mechanisms and LSPs provide traffic engineering capabilities. Also describe applications of those capabilities.
13. Describe the advantages MPLS has for voice over IP applications and discuss some of the reasons why the voice over MPLS application has not achieved large-scale deployment.

REFERENCES

Andersson, L., Doolan, P., Feldman, N. et al. 2001. "LDP Specification." IETF RFC 3036. Web site: www.ietf.org.

Awduche, D., Berger, L., Gan, D. et al. 2001. "RSVP-TE: Extensions to RSVP for LSP Tunnels." IETF RFC 3209. Web site: www.ietf.org.

Banerjee, A., Drake, J., et al. 2001. "Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements." *IEEE Communications*, Vol. 39, No. 1, pp. 144–150.

Black, U. 2000. *MPLS and Label Switching Networks*. Englewood Cliffs, NJ: Prentice Hall PTR.

Brittain, P., and Farrel, A. 2000a. "MPLS Traffic Engineering: A Choice of Signaling Protocols." Data Connection Ltd. White paper. Web site: www.dataconnection.com.

Brittain, P., and Farrel, A. 2000b. "MPLS Virtual Private Networks." Data Connection Ltd. White paper. Web site: www.dataconnection.com.

Davie, B., and Rekhter, Y. 2000. *MPLS: Technology and Applications*. San Francisco, CA: Morgan Kaufmann.

Faucheur, F. (ed.). 2002. "Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering." IETF Internet Draft. draft-ietf-tewg-diff-te-proto-01.txt. Web site: www.ietf.org.

Gleeson, B., Lin, A., et al. 2000. "A Framework for IP Based Virtual Private Networks." IETF RFC 2764. Web site: www.ietf.org.

Harnedy, S. 2001. *MPLS Primer: An Introduction to Multiprotocol Label Switching*. Englewood Cliffs, NJ: Prentice Hall PTR.

IEC. 2001. "A Comparison of Multiprotocol Label Switching (MPLS) Traffic-Engineering Initiatives." White paper. Web site: www.iec.org.

Jamoussi, B., et al. 2002. "Constraint-Based LSP Setup Using LDP." IETF RFC 3212. Web site: www.ietf.org.

Kankkunen, A., et al. 2001. "VoIP over MPLS Framework." IETF Internet Draft. draft-kankkunen-vompls-fw-01.txt. Web site: www.ietf.org.

Mannie, E., Ashwood-Smith, P., Awduche D., et al. 2002. "Generalized Multi-Protocol Label Switching (GMPLS) Architecture." IETF Internet Draft. draft-ietf-ccamp-gmpls-architecture-02.text. Web site: www.ietf.org.

Rosen, E., Tappan, D., Fedorkow, G., et al. 2001. "MPLS Label Stack Encoding." IETF RFC 3032. Web site: www.ietf.org.

CHAPTER 18

IP QoS Architectures and Protocols

18.1 Introduction

This section describes the three dimensions of the IP QoS concept and provides a historical perspective on early IP QoS.

18.1.1 Basic QoS Concepts

Quality of service is an elusive term that has been used by many people not necessarily to mean the same thing. QoS can have different connotations and a different emphasis for different types of network technologies. For practical purposes, there are three fundamental dimensions of QoS in regard to IP networks: data delivery delay, data throughput, and resource availability.

The first dimension, data delivery delay, deals with the issue of how fast data can be delivered from end to end. The goal of this QoS dimension is to minimize the delay and delay variation. The data delivery delay dimension has the following components:

Latency. This is the time between one node sending a packet and another node receiving it. The latency in an IP network consists of several pieces:

- Transmission delay or propagation delay: the time spent on the transmission link.
- Processing delay: the time needed to encode and collect samples into IP packets for transmission.
- Queuing delay: the time a packet spends in queues at various nodes.
- Network delay: the time spent for the transmission of packets across an entire network. This delay is a function of the physical medium and the protocols used to transmit the data.

Jitter. Jitter is the variable arrival times of packets, caused by the fact that packets do not all traverse the network at the same speed, and often not over the same route. Removing jitter requires collecting packets and holding them long enough to allow the slowest packets to arrive and be played in the correct sequence. This causes additional delay.

The second dimension, data throughput, deals with the amount of data that can be delivered to a customer for a given time interval. It is measured with bandwidth and packet loss:

Bandwidth. This measures the data transmission capacity in Mbps (megabits per second) or Gbps (gigabits per second). Bandwidth represents the engineered theoretical maximum capacity of a connection.

Packet loss rate. This measures the number of packets lost due to network congestion or outage during the transmission as a percentage of the total number of packets transmitted averaged over a certain period of time. The intervals can be hourly, biweekly, monthly, etc.

The third dimension, resource availability, addresses the issue of the amount of resources available and for how long. IP networks are shared resources, and shared networks may not be able to satisfy all the requirements of all the applications at the same time, due to limited network resources or outage.

The term *QoS* sometime is used synonymously with the term *class-of-service*. Overall, QoS is a more generic, encompassing term than CoS. The concept of CoS refers to the idea that services can be categorized into separate classes, which can be treated differently. The capability of differentiating different classes of service is an important QoS mechanism.

A related term is *service level agreement,* which is a formal, negotiated agreement between a service provider and a customer. A set of measurable QoS parameters normally form the core of an SLA, and may include data bandwidth, bit error rate, and transmission delay.

18.1.2 Historical Perspective on QoS

QoS was not much of an issue until IP networks came along and real-time-sensitive applications such as voice over IP were targeted for those data networks. QoS is not an issue for circuit-switched networks because the 64/56-Kbps data rate is built into their transmission and switching equipment.

There are two general approaches to QoS in data networks: the layer-2 (the data link layer) approach and the layer-3 (the network layer) approach. They are not alternative approaches but differ in scope. The scope of layer-2 QoS covers homogeneous networks such as ATM and frame relay where the QoS mechanisms are built into the data link layer protocol. They provide virtual circuits as a foundation for QoS control. Since the circuits are "virtual," the QoS mechanisms are needed for ensuring service delivery. The ATM standards define a very rich QoS infrastructure by supporting traffic agreements and traffic control and connection admission control. Frame relay provides a simpler yet rich and practical set of QoS mechanisms such as committed information rate, congestion notification, and frame relay fragmentation.

On the other hand, the scope of IP QoS is concerned with internetworking across multiple networks of different types, for example, host to Ethernet network, ATM network, Ethernet network, and then another

host. So there is an issue of mapping IP-level QoS into layer 2-QoS when an IP packet enters a particular network.

The QoS was not much of a concern when IP networks were first designed. The IP packet header has a type-of-service field that provides the initial IP network class of service mechanism. The three precedence bits are used to classify packets up to eight categories of services. Packets of lower precedence can be dropped in favor of packets of higher precedence in case of network congestion. This, plus other schemes, constituted the early attempts to achieve QoS on IP networks. Neither the precedence scheme nor the other schemes are widely implemented by network equipment vendors.

IP QoS became a more prominent issue after the commercialization of the Internet in the United States, i.e., after the transition of the NSFnet from being supervised by the government agency the National Science Foundation to being supervised by the commercial Internet service providers in the early 1990s. As broadband IP packet networks begin to emerge as multiservice, all-purpose communications platforms in late 1990s, the issue of QoS becomes more pressing than ever. Serious efforts were devoted to the area, resulting in two QoS models for IP packet network: IntServ and DiffServ.

The two QoS models take two different approaches to the QoS issues. IntServ more or less attempts to simulate the "virtual circuit" of ATM or frame relay on layer-3 of a network by setting up an end-to-end route with fixed QoS parameters. DiffServ achieves QoS by defining several common classes of service with associated queue priorities and drop precedence on a per-hop basis. End-to-end QoS is achieved by the fact that all DiffServ-compliant nodes along a path respect the commonly defined classes of service (QoS Forum 1999).

18.2 RSVP and IntServ QoS Model

Resource ReserVation Protocol (RSVP) is a signaling protocol for resource reservation setup and control. IntServ is an IETF QoS model that uses RSVP to provide QoS on packet networks like IP networks.

18.2.1 RSVP

RSVP represents the biggest departure from the traditional best-effort IP networks by providing mechanisms to set up a path that is the closest thing to an end-to-end virtual circuit.

RSVP operates over the IP layer or TCP/UDP layer. As a result, there are two types of RSVP protocols depending on how the RSVP protocol data is carried in an IP packet. *Native RSVP* encapsulates the RSVP header and message inside the IP header. *UDP-encapsulated RSVP* has its header contained in a UDP datagram.

RSVP is receiver-based protocol. The resources reservation is performed backwards from the receiver to the sender with the RESV message. It was designed this way to accommodate large, heterogeneous, multicast receiver groups (Zhang et al. 1997).

18.2.1.1 RSVP Messages RSVP defines two main messages used for the purposes of path setup and resource reservation: PATH message and RESV message. PATH is a path request message, while RESV is a reservation request message. Each message may have one or both of the following components:

- *Tspec:* traffic specifications that specify the QoS parameters, including upper and lower bounds of desired bandwidth and jitter
- *Rspec:* reservation specifications that specify the resource QoS for reservations

18.2.1.2 RSVP Protocol Operations RSVP provides a mechanism to reserve the resource along a predefined route or path, although calculating the route itself is beyond the scope of RSVP. So RSVP operates on the assumption that an end-to-end route is already determined using some administrative policy or routing algorithms such as OSPF. The mechanism has five steps:

Step 1. An application at the sending host sends a PATH message with TSpec to a destination host along a predetermined path for the unicast mode.

Step 2. Each RSVP-capable intermediate node in the downstream direction (i.e., the direction toward the destination node) establishes a "path state" that includes the address of the next hop in the upstream direction. For example, node R2 remembers the address of the upstream adjacent node R1, as shown in Fig. 18-1.

Step 3. An application on the receiver host, upon receiving the PATH message, builds and sends an RESV message in the reverse direction. The RESV message has two parts: the received Tspec and a reservation specification (Rspec) that specifies the requested service parameters such as peak rate, packet rate, etc.

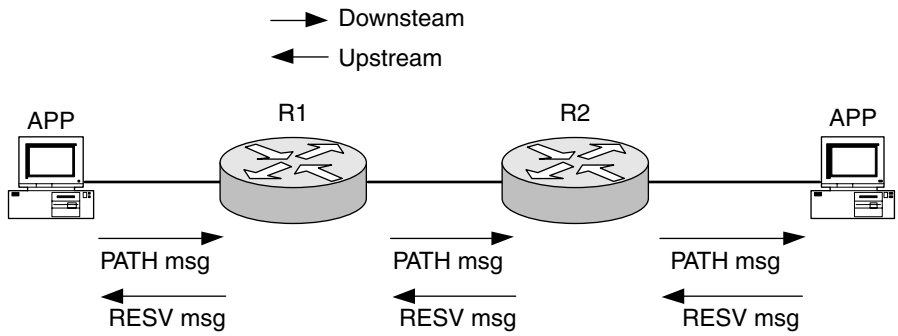


Figure 18-1
RSVP operation example.

Step 4. Each RSVP-capable node in the upstream direction (i.e., the direction toward the sender) along the route, upon receiving the RESV message, uses the admission control process to authenticate the request and allocate requested resources. If either authentication fails or the resource request cannot be satisfied, the node sends an error message to the receiving host. Otherwise, the node sends the RESV message upstream to the next node.

Step 5. The last node along the route, upon receiving the RESV message, sends a confirmation message back to the receiver to confirm the route.

The resource reservation state resulting from the RSVP operations of each node is "soft," meaning that they need to be refreshed periodically. Otherwise, it is deemed no longer valid and the reservation is deleted from the node.

18.2.2 IntServ QoS Model

The Integrated Services QoS model defines two types of service and a set of service parameters for a path reserved via RSVP. The two types of service of IntServ are distinguished by the guarantee of the service parameters (Braden, Clark, and Shenker 1994):

- *Guaranteed service.* This ensures the availability of bandwidth as specified in reservation message and provides firm bounds on end-to-end queuing delays by combining the parameters from all the nodes along the RSVP route. This service more or less simulates a virtual circuit for the real-time-sensitive applications.
- *Controlled load service.* This is equivalent to a best-effort service for the applications that are more tolerant of delay and packet drops.

IntServ defines a set of service parameters that can be included in the Tspec and Rspec of an RESV message for both guaranteed and controlled load services, which are similar to ATM CoS-like definitions of services:

Token rate. This is the continually sustainable bandwidth required for a packet flow. In other words, it is the average data rate to be achieved.

Token bucket depth. This is the amount of data that is allowed to exceed the average data rate for a specified short period of time.

Peak rate. This is the maximum data rate at which an application can send data.

Minimum policed size. This is the smallest packet size in bytes generated by the sending application.

Maximum packet size. This is the maximum allowed packet size in bytes. Any packets exceeding this size cannot receive QoS-controlled treatment.

A crude analogy will help summarize the RSVP-based IntServ QoS model: It is like reserving a fast lane on a superhighway from coast to coast, where every leg of the route needs to be reserved individually. For every leg of the route, guards are sent out on a periodic basis to make sure that the reservation sign has not been torn down. Once the lane is reserved all the way to the destination, vehicles of the same traffic flow can follow the reserved path all the way to the destination without any hindrance (Wroclawski 1997).

The IntServ QoS model with RSVP provides the highest level of IP QoS guarantee. It allows an application to specify QoS with a fine level of granularity and with the best guarantees of service delivery possible on an IP network. However, the end-to-end QoS guarantee of the IntServ model also brings about some major limitations, including the following:

- Each node must maintain the provisioning state information on a per-route basis, and this adversely affects the scalability of the model. It becomes more difficult to maintain a large amount of provisioning state information as the network grows very large.
- The periodic refreshing of the provisioning state incurs an overhead that becomes much worse as the network grows large.
- The complexity of the provisioning system is also a factor affecting its deployment.

18.3 DiffServ QoS Model

18.3.1 Introduction

Differentiated Services is an alternative packet QoS model that is much simpler and more widely deployed than the IntServ QoS model. It is a hop-by-hop QoS model in contrast to the end-to-end QoS behavior of the IntServ model. The DiffServ model is characterized by a level of QoS control and scalability that is coarser than that of the IntServ QoS model. In addition, DiffServ defines a building-block approach to QoS with a set of building blocks to be used at each node. End-to-end QoS is achieved via the fact that all the nodes respect the DiffServ fields of each packet (Blake et al. 1998).

DiffServ is based on the differentiated service field of the original IP header. Remember that the IP header has a 1-byte type-of-service field that was originally intended for service-type information but was never widely used due to a lack of consensus on the definition. The ToS byte has been redefined as the DS (differentiated service) field, which consists of two components as shown in Fig. 18-2.

The first component occupies the first 6 bits that are used to encode a maximum of 64 differentiated service codepoints (DSCPs), or classes of services. The second part is currently unused (CU). A DS value replaces the ToS value if ToS is present at the DiffServ ingress node (Nichols et al. 1998).

A DiffServ-capable node is known as a *DS node* and a DS node at the edge of a DS network can be either an *ingress DS node* or *egress DS node*, depending on the direction of the traffic. A set of contiguous DS nodes forms a *DS domain* that has the same service provisioning policies and common definitions of classes of services, as shown in Fig. 18-3. Multiple DS domains form a *DS region* or *DS network* (Cisco 2001).

18.3.2 DiffServ Per-Hop Behavior

First, an analogy will provide a high-level view of DiffServ QoS model behavior. The DiffServ model provides a service like a toll pass that specifies the privilege level for each packet flow. All vehicles of the same traffic flow have the same pass. At each intersection, a "road master" decides which queue a vehicle should follow according to the privilege level of the pass. The higher the privilege level of the pass, the higher the priority of the queue and shorter the wait will be.

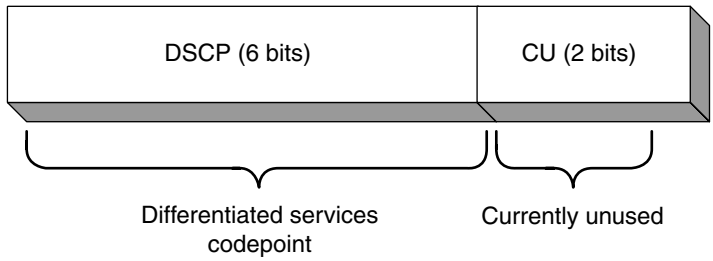


Figure 18-2
DS field of DiffServ QoS model.

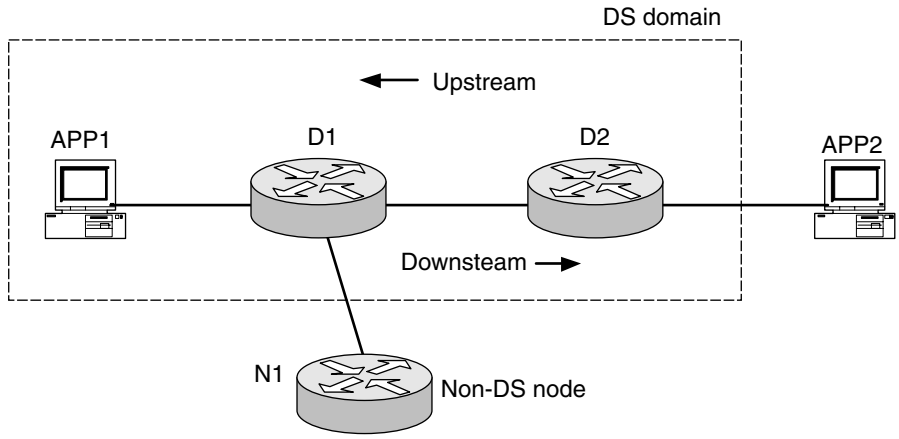


Figure 18-3
A DS domain example.

A key idea of the DiffServ QoS model is that a packet's forwarding decision is localized, on a per-hop basis, and decided by the local road master, as in the above example. Packets entering each DS node are marked with a class of service or per-hop behavior (PHB) to indicate to the next downstream DS node how this packet should be handled. Packets of the same microflow (between the same two applications) do not have to have the same PHB end-to-end. In other words, a packet marked with one class of service is subject to remarking as it traverses the DS network.

The DiffServ QoS model provides a small number of PHBs, which can be viewed as classes of service, as discussed above. Although the 6-bit DSCP part of the DS field can encode up to 64 different classes of service or PHBs, for practical purposes the number of DSCPs used in a single DS network is much smaller.

Two classes of service have been standardized and thus are respected by all DS nodes, in addition to the best-effort service of conventional IP networks:

Expedited forward: The packets of this PHB are not affected by other PHB traffic and the packet rate is guaranteed above the specified

value. Packets of this PHB have higher priority than other traffic to ensure low packet loss, small delay, and minimum jitter. The edge node of a DS network ensures that only the appropriate number of packets of this class is admitted into the DS network so that the resource can be guaranteed. The expedited forward (EF) class of service is intended for applications with a low tolerance for delay and jitter, such as voice over IP.

Assured forwarding: This is a set of classes of services defined for assured forwarding (AF) that provides statistically guaranteed services. Each class of service or PHB is defined by two QoS dimensions, a priority queue *x* and a drop preference *y*. The priority queue number of a packet determines the order in which the packet is sent out. The packets belonging to the same microflow with the same priority queue number but different drop preferences must be forwarded in the same order they arrived, unless the packet is dropped.

Across different DS domains, the PHB of one domain has to be mapped to a PHB of another domain via service level agreement.

Each DS node in a DS network only needs to remember the state information on a per-PHB basis, or class of service, which is relatively small. There is no need to provision anything on a per-session basis. This is why the DS QoS model can scale up to large networks with large numbers of nodes.

18.3.3 DiffServ QoS Model Components

DiffServ QoS control behavior is localized at each DS node. The functions performed at a DS node include classifying each packet, assigning it with a class of service or PHB according to the service level agreement, enforcing the traffic, and performing other processing before sending it to the next downstream DS node.

The DiffServ QoS model consists of the five functional components at a DS node, as shown in Fig. 18-4 and defined in RFC 2457 (Blake et al., 1998):

Packet classifier. Any IP packet, after entering a DS node via a network interface card, first sees the classifier. The classifier is a DiffServ functional module that examines the IP header of the packets, identifies the source of the packet flow, retrieves the service level agreement from the policy server for this source, and assigns a class of service to the packet, according to the traffic profile of the packet. If this is an ingress node, the classifier divides incoming traffic into aggregates, or group of packets, that share the same traffic profile.

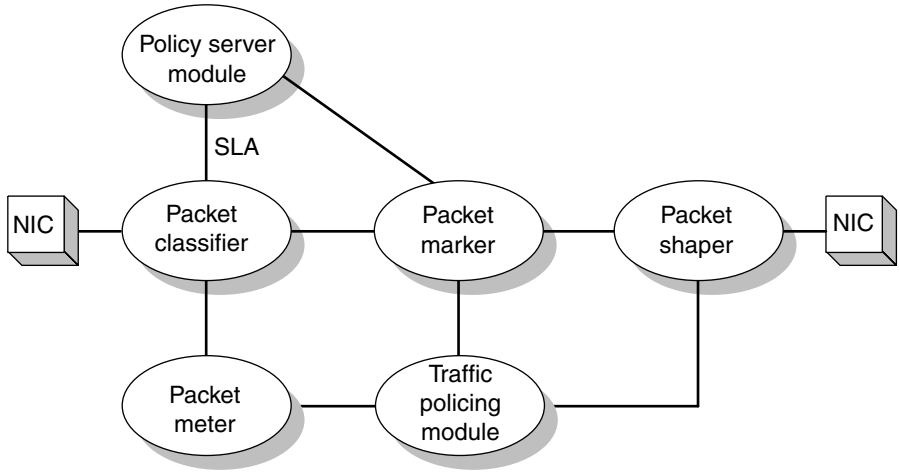


Figure 18-4 Functional components of a DiffServ QoS model.

Packet meter. This functional component then checks all the packets going through the classifier to monitor the traffic rate and reports the packet rate to the packet marker and shaper.

Packet marker. This functional component, according to the classification of the packet, marks the DSCP subfield of the DS field of the packet header. The DSCP value specifies the PHB of the packet within the same DS domain.

Traffic policing module. At an ingress node, this functional component polices the packet aggregates according to the traffic conditioning agreement (TCA). If an aggregate is out of profile it is either sent to the packet shaper for dropping or remarked with a different PHB so the packet flow conforms to the traffic profile.

Packet shaper. This functional component, according to the monitored packet rate, controls the forwarding rate of the packets so that the flow does not exceed the specified rate and no congestion occurs. If necessary, the shaper can drop packets, based on the traffic profile. Or if there is a separate packet dropper module, it performs the packet dropping. Finally the packet is forwarded to the NIC and sent to the next DS node.

18.4 QoS of Access Network-Subnet Bandwidth Management

DiffServ and IntServ were more or less developed with the core IP network in mind. One missing piece of the end-to-end QoS puzzle is QoS

for the access packet network. The Subnet Bandwidth Management (SBM) protocol has been developed to extend the IntServ QoS model to Ethernet LAN. Ethernet, the dominant LAN technology, either in a shared medium form or its switched form, provides a service analogous to the best-effort IP service, in which variable delays can affect real-time applications.

Parallel efforts at both IEEE and IETF were undertaken to extend the QoS model to Ethernet LAN. The IEEE 802.1Q and 802.1D define the standards that allow Ethernet switches to classify frames in order to expedite delivery of time-critical traffic (IEEE 1998a and 1998b). The IETF Integrated Services over Specific Link Layers Working Group has defined a mapping of the upper-layer QoS protocols and services to a QoS model of layer-2 technologies like Ethernet. The SBM for shared or switched 802 LANs such as Ethernet is the main result of those efforts.

This section provides an overview of SBM functional components and SBM's high-level protocol operations for bandwidth reservation on LAN.

18.4.1 SBM Components

SBM is a signaling protocol that allows a QoS resource manager to communicate with a set of SBM clients to reserve resources at a local host and to map the higher-layer QoS parameter to the layer 2-LAN context.

An SBM-capable LAN consists of a designated SBM (DSBM) and a set of SBM clients, as shown in Fig. 18-5 (Yavatkar et al. 2000). A DSBM is a protocol entity that resides in a LAN device such as a router and is responsible for managing the resource on a LAN segment. There is one DSBM per LAN segment. Thus, a managed segment is a segment with a DSBM present and responsible for admission control over requests for resource reservation. A DSBM client is a functional entity that is responsible for requesting services and making reservations on behalf of the host it resides on for applications running on the host.

There might be more than one SBM entity on a single segment but only one of the entities can be a DSBM. An election algorithm is defined for multiple SBMs to elect a DSBM. One DSBM can preside over multiple LAN segments if some SBM-transparent device connects multiple segments.

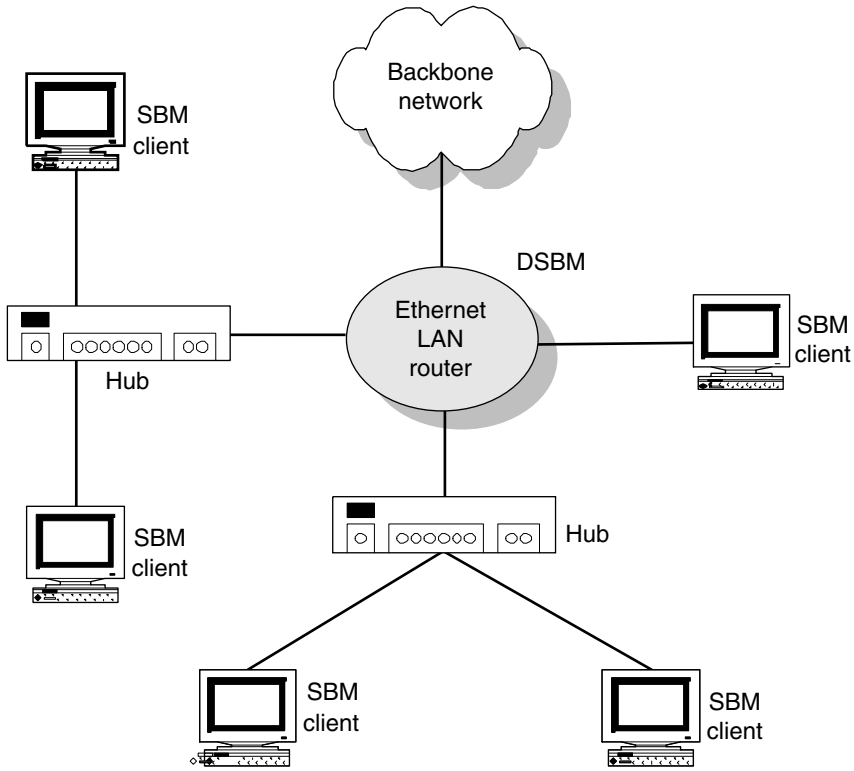


Figure 18-5
An overview of SBM entities.

18.4.2 SBM Protocol Operations

There are mainly two types of SBM protocol operations: initialization and admission control. The initialization involves both DSBM (i.e., the SBM manager) and SBM clients. They do the following:

DSBM initialization: The DSBM collects information on the available resource such as the amount of bandwidth that can be reserved, from each of its managed segments. Usually this information is statically configured in the device like a LAN router.

SBM client initialization: Each client in the managed segment searches for the existence of a DSBM. If there are no DSBMs found, the client itself might participate in an election to become a DSBM for that segment.

Admission control: The DSBM, as the manager of the segment, ensures that the segment-wide resources are allocated according to the designed rules. There are two cases: An SBM client receives a resource reservation request and a DSBM issues a resource reservation request itself.

If the client receives an RSVP PATH message, it forwards the request to its DSBM instead of the destination address. The DSBM modifies the message, builds or adds to a PATH state its own MAC and IP address, then forwards the request to its destination address. The DSBM updates the path state map and the available resource state for the segment it manages. If the requested resources are not available, the DSBM issues a RESVError to the requester. If there is a cascade of managed segments in the domain, the PATH message propagates through the DSBM of each segment and PATH states are updated at each DSBM.

When a DSBM client wishes to issue a resource reservation message RESV in response to the PATH message, it sends the RESV message to the DSBM's address that is found in the received PATH message. Again, in case of a cascade of managed segments, the RESV message must pass through each DSBM successfully in order to reach its destination.

REVIEW QUESTIONS

1. Describe the three dimensions of the packet network QoS concept: data delivery delay, data throughput, and available resource. Also discuss the relationships between the three dimensions.
2. Explain why QoS is not much of an issue for circuit-switched networks and why the issue of QoS became prominent after the commercialization of the Internet.
3. Explain why the reservation state resulting from the RSVP reservation operation at a node is "soft." What is the consequence if the soft reservation state is not refreshed?
4. RSVP is said to be a receiver-based protocol. Explain why this is the case by considering the directions (i.e., downstream or upstream) the RESV and PATH messages are passed from node to node.
5. The DiffServ QoS model is characterized by per-hop behavior. Explain what a PHB is and how it is used in a DiffServ network.
6. Describe how end-to-end QoS is achieved with the DiffServ model, given that the DiffServ model has QoS control only on a per-node basis.
7. Describe the two standardized PHBs or classes of service and explain how the QoS of the EF class of service is guaranteed in a DiffServ network.

8. Describe the functional components of the DiffServ QoS model, i.e., the functions performed by the packet classifier, the traffic policing module, the packet marker, the packet meter, and the packet shaper.
9. The packet classifier has the responsibility of associating a microflow or packets of the same application with a PHB or a class of service. Explain how the packet classifier makes this decision.
10. Compare and contrast the IntServ and DiffServ QoS models. One features end-to-end resource reservation, and the other features localized PHB. Describe the advantages and disadvantages of each.
11. Describe the motivations behind the development of the Subnet Bandwidth Management protocol and the relationships between the two types of protocol entities, designated SBM and SBM clients.

REFERENCES

Blake, S., Black, D., et al. 1998. "An Architecture for Differentiated Services." IETF RFC 2475. Web site: www.ietf.org.

Braden, R., Clark, D., and Shenker, S. 1994. "Integrated Services in the Internet Architecture: An Overview." IETF RFC 1633. Web site: www.ietf.org.

Cisco Systems. 2001. "DiffServ-the Scalable End-to-End QoS Model." White paper. Web site: www.cisco.com.

IEEE. 1998a. "Common Specifications—Media Access Control (MAC) Bridge." IEEE 802.1D. Web site: www.ieee.org.

IEEE. 1998b. "Virtual Bridged Local Area Networks." IEEE 802.1Q. Web site: www.ieee.org.

Nichols, K., Blake, S., et al. 1998. "Definition of the Differentiated Service Field (DS field) in the IPv4 and IPv6 Headers." IETF 2474. Web site: www.ietf.org.

QoS Forum. 1999. "QoS Protocol and Architecture." White paper. Web site: www.qosforum.com.

Wroclawski, J. 1997. "The Use of RSVP with IETF Integrated Services." IETF RFC 2210. Web site: www.ietf.org.

Yavatkar, R., Hoffman, D., et al. 2000. "Subnet Bandwidth Manager (SBM): A Protocol for RSVP-Based Admission Control over IEEE 802-Style Networks." IETF RFC 2814. Web site: www.ietf.org.

Zhang, L., Braden, R., et al. 1997. "Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification." IETF RFC 2205. Web site: www.ietf.org.

CHAPTER 19

QoS Policy and Common Open Policy Service Protocol

19.1 Introduction

The last piece of the end-to-end QoS puzzle is QoS policy. QoS policies address the issue of how users request a service with QoS guarantee and how the user requests are mapped to the network resource allocation request to be implemented in a network.

Efforts to define a standard QoS policy architecture framework and the protocols for passing policy data between concerned parties started in the mid-1990s at IETF, and the standards themselves started emerging in the late 1990s. This was soon followed by their implementation, although their deployment is at the early stage.

The QoS policy architectural framework consists of three layers with three logical entities and a set of protocols for communications between the entities, as shown in Fig. 19-1 (Chan et al. 2001; QoS Forum 1999). Parallel to the three layers are the network layers to indicate where each policy entity might reside physically, though the architecture itself does not mandate any physical implementation.

At the top level is a policy repository (PR) that contains the policy in the form of policy rules, and an authentication server that authenticates the user requesting service. The policy repository and the associated authentication server are at the network level, serving the whole network of a service provider.

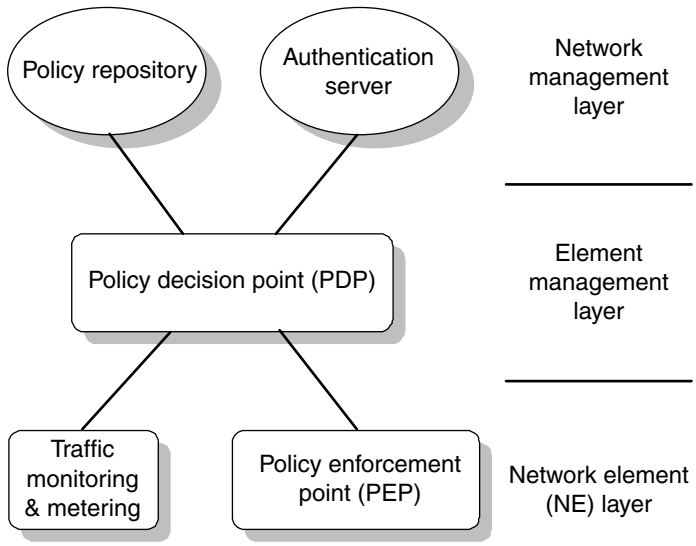


Figure 19-1
QoS policy architectural framework.

The policy decision point (PDP) is where the policy decisions are made. It interfaces multiple policy enforcement points (PEPs) to enforce the policies and interfaces the policy repository to retrieve the policy. A PDP is normally located at a network element management level that is responsible for one or more PEPs.

A PEP carries out the policy decisions made by the PDP at a network device, taking into consideration the current traffic conditions at the device such as available bandwidth and length of the queue. In addition, a PEP reports the local traffic conditions to the controlling PDP to help the PDP make a policy decision. The metering and monitoring function at a device continuously checks the device traffic conditions, such as whether policy constraints are being violated, and static information like device addresses and dynamic information like available bandwidth.

19.2 Policy Repository and Policy Information Model

At the core of a policy repository is a policy information model and a protocol for other entities to access the policy repository. The policy repository in essence is no more than a directory-based information base for storing policy data. The interface to the policy repository uses the standard Lightweight Directory Access Protocol (LDAP). But first, a flow of the policy provisioning process provides a context.

19.2.1 Policy Provisioning Process

The user business requirements are turned into a repository policy schema and policy objects at the policy repository, as shown in Fig. 19-2. The user first inputs the business goals and objectives via a user interface, and the output of the input system is a set of the policy descriptions expressed in a language such as the policy framework definition language (PFDL).

The next step is to translate the policy descriptions into a set of policy schema and objects. The policy descriptions are still in high-level, imprecise terms while the policy schema and objects are one step closer to a representation that is realizable at a network device. The translation is normally an automatic process that can be done by an off-the-shelf policy translator.

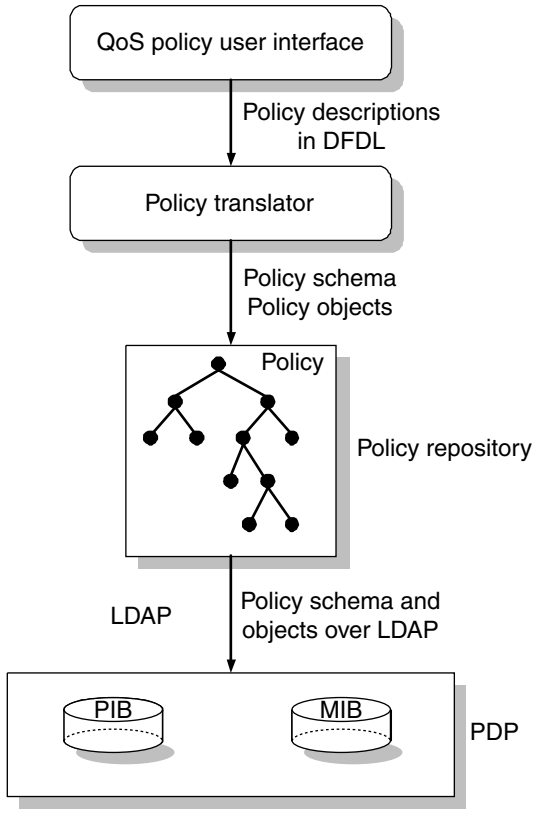


Figure 19-2
QoS policy repository overview.

The following step is to store and organize the policy schema and policy objects in the policy repository. This is essentially an issue of an information retrieval system. IETF QoS policy working groups take advantages of the widely used LDAP and associated directory structure to represent and store the QoS policy schema and objects. The policy schema is represented as an LDAP schema and policy objects as LDAP objects as defined in IETF RFC 2251 (Wahl et al. 1997).

The above steps, as part of the QoS provisioning process, are accomplished in an off-line fashion, before the requested service is rendered. The policy schema and objects stored in the LDAP directory-based information system are retrieved by the PDP.

The PDP retrieves the policy schema and objects via LDAP. This is normally performed at the time the requested service is rendered, and the retrieval needs to be efficient with minimal delay to support real-time applications. The LDAP-based directory structure enables tree-based search, which is efficient for data retrieval.

19.2.2 Policy Repository

The policy repository contains a set of policy classes that are organized as an LDAP directory information tree (DIT), which is based on the data model specified in ITU standard X.500 (ITU-T 2001). Each tree node represents a policy class with a set of attributes. The combination of the attributes of a tree node is called a *relative distinguished name* (RDN), which must be unique among all its sibling tree nodes (all nodes at the same level with the same parent node). The concatenation of the relative distinguished names of the entries from the root of the tree to a node forms the *distinguished name* (DN) of the node, which uniquely identifies the node in the entire tree.

A set of mandatory and optional attributes for a DIT node that forms an RDN is defined in the LDAP data model. When translated into the QoS policy-specific domain, they include the following:

- *objectClass:* a string name identifying a policy class
- *creatorName (cn):* a string for the distinguished name of the user who inserted this entry into the policy repository
- *attributeType:* the type of attributes that are defined in a policy schema

A policy repository has a directory server that responds to the queries and requests from clients and searches for policy rules and objects, and administrates and maintains the repository.

19.2.3 Policy Repository Interface

LDAP provides an interface for policy repository clients such as policy decision points to access and retrieve the policy rules and objects. LDAP provides the following operations for repository clients:

- *Bind:* This operation allows a client such as a PDP to send a connection request to the server and to be authenticated by the repository server.
- *Unbind:* This operation allows a client such as a PDP to terminate an LDAP session with the server.
- *Search* or *query:* This operation allows a client to search and retrieve a policy entry in the repository.
- *Add, delete or modify:* This operation allows a client to add a new entry, or delete or modify an existing entry in the repository. The

client performing this type of operation is likely to be a QoS policy provisioning entity like the QoS policy user interface shown in Fig. 19-2.

- *Abandon:* This operation allows a client such as PDP to request the repository server to abandon an outstanding operation.

19.2.4 Policy Core Information Model

A core policy information model has been standardized by IETF in RFC 3060 (Boore et al., 2001) to provide a foundation and basic vocabulary for representing QoS policy information in a vendor-independent and device-independent fashion. At the core of the information model is a set of QoS policy classes, as described below:

- *PolicyGroup:* A container class to hold all policy classes associated with a business objective or an organization. It is an administrative decision to define the scope of a PolicyGroup class. A PolicyGroup class may recursively contain other PolicyGroup classes.
- *PolicyRule:* A policy decision in the form of "If *condition* then *action*." A condition in a PolicyRule can be a set of conditions combined with logical operator such as AND, OR, or NOT. A policy action specified in a PolicyRule can be performed if and only if the PolicyRule condition evaluates to TRUE. The conditions and actions associated with a policy rule are modeled with subclasses of the classes PolicyCondition and PolicyAction. A policy rule can be viewed as consisting of one or more PolicyConditon, PolicyAction, and PolicyTimePeriodCondition classes.
- *PolicyCondition.* An abstract class representing a set of policy conditions that evaluate to a final TRUE or FALS value. A set of conditions ANDed together are said to be in *conjunctive normal form* (CNF); those that are ORed together are said to be in *disjunctive normal form* (DNF). This class inherits all its properties from the Policy class.
- *PolicyTimePeriodCondition.* A class representing the time period during which an associated policy rule is active or inactive.
- *VendorPolicyCondition.* A class providing an equipment vendor or service provider with the flexibility to define the policy conditions specific to the vendor.
- *PolicyAction.* A class representing one or more operations to be executed at a network device that will affect the network's traffic condition.

- *VendorPolicyAction.* A class providing an equipment vendor or service provider with the flexibility to define policy actions specific to the vendor.

19.3 Common Open Policy Service Protocol

This section discusses the policy decision making process and the interface between the PDP, also known as the *policy server,* and the PEP, also known as the *policy client,* as defined in IETF RFC 3748 (Durham, 2000). The interface is the Common Open Policy Service (COPS) protocol.

19.3.1 COPS Protocol Model

COPS is a standard protocol defined by IETF in RFC 2748 (Durham, 2000) for communication between a policy server and a policy client, as shown in Fig. 19-3.

A PEP, a policy client, normally resides at a network device such as a router or switch and is responsible for carrying out the policy decisions it obtains from a policy server. A policy server can be a local or remote PDP. A PEP can have a local PDP (LPDP), which is a subset of the functionality of a remote PDP. The local PDP kicks in to make a policy decision in place of the remote PDP when the connection between the PEP and the remote PDP is lost.

Specifically, the responsibilities of a PEP include the following:

- Establishing a TCP connection to a policy server or PDP with a client-open message. The PEP attempts to connect to the backup server in case the connection to the primary server is not successful. The PDP can have multiple types of clients and needs to establish one connection per client. In establishing a client connection, it may negotiate security parameters with the server depending on the type of client.
- Generating and sending to the policy server a policy decision request when the need for a policy decision arises, such as the arrival of a packet flow or resource reservation request.
- Carrying out a policy decision and reporting the result of the action back to the policy server. The decision can be about the admission control, resource allocation, or message forwarding.

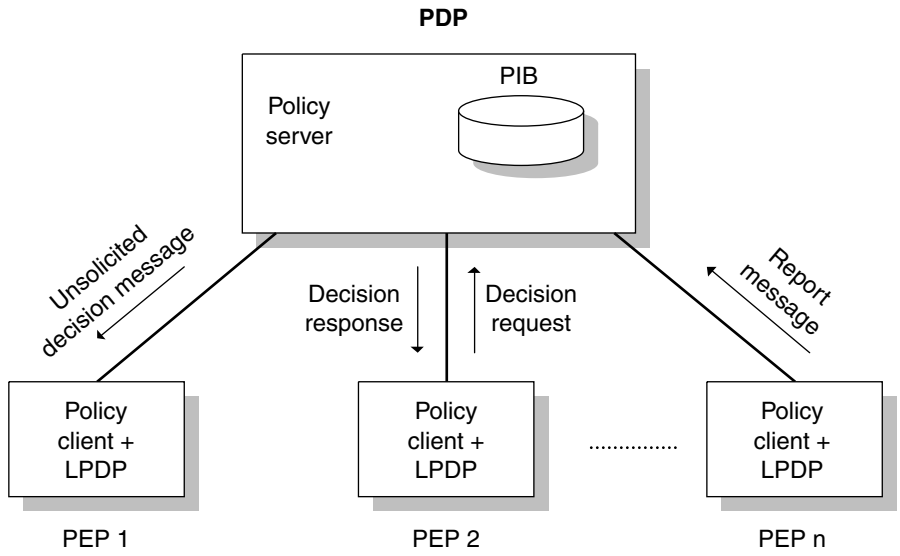


Figure 19-3
A COPS protocol overview and COPS message exchanges.

- Reporting the client information to the server for the purpose of accounting or state monitoring.
- Maintaining the state of the requests it creates. Once a decision request is created, it takes on a life of its own and goes through the life cycle of various states. For example, after processing a flow of packets, the PEP may send a delete request message to the PDP to remove the state associated with the request and stop any further decision making at the PDP on that request.
- Monitoring and maintaining the connection to the policy server. When the connection is found lost, it attempts to reestablish another connection to the server or a different connection to an alternative server.

PDP, also called remote policy server, is mainly responsible for making a QoS policy decision in response to requests from a policy client. Among other things, the responsibilities of the policy server include the following:

- Authenticating clients and granting client connections.
- Interfacing to the policy repository via LDAP and retrieving the policy rules from the policy repository for making policy decisions.
- Making the policy decisions in response to policy decision requests from policy clients, taking into consideration factors such as type of request, the current request state, and the context

of this request. In some cases, after considering the history of new events, the policy server may reverse its decision and send the PEP another decision message. Thus the server can send its decisions asynchronously.

- Maintaining the states of the client requests at the server side. It may ask all clients to resynchronize the client states in cases such as server reinitialization.

COPS distinguishes between three request types that originate from a PEP and need to be considered by the PDP:

1. *Admission control request.* If a packet flow arrives at a PEP, the PEP asks the PDP for a admission control decision on the flow.
2. *Resource allocation request.* The PEP requests the PDP for a decision on whether, how, and when to reserve local resources for the request.
3. *Forwarding request.* The PEP asks the PDP how to modify a request and forward it to the other network devices.

The COPS protocol is state-based in two respects. First, a request state is shared between a policy server and a policy client, and both the server and the client remember the current state of a request. Second, the state of one request may affect the way the policy server responds to a new request from the client.

19.3.2 COPS Protocol Operations

COPS supports the following three types of the message exchange between a policy server and a policy client, as shown in Fig. 19-3:

- *Request-response message exchange.* There is one response message from the PDP to the PEP for each request message from the PEP.
- *Unsolicited message from a PEP to the PDP or vice versa.*
- *Maintenance message like the keep-alive message.*

19.3.2.1 Request-Response Operation COPS defines three pairs of request-response messages between a PDP and a PEP:

- *Request-decision message pair* for a client to request and for a server to convey a policy decision
- *Client open-client accept message pair* for a policy client PEP to connect to a policy server PDP

- *Synchronize state-synchronize state complete message* pair for a policy server to synchronize the request state

The request-decision message pair is the most complicated, and includes the following basic steps:

- A PEP, at a service rendering time such as the arrival of new packets, sends a decision request to the PDP to authenticate the user and to decide on what action to take on the user traffic.
- The policy server, upon receiving a request from the PEP, makes a decision based on a number of factors, such as previous state of the request for the same customer, resource availability, the type of client making the request, request context, etc.
- The server sends the decision message to the PEP and updates the state of the request.

The request message from a client to a server may include parameters such as the following:

- Client type, such as COPS client, TCP client, etc.
- Client handle object, which indicates the state of request on the client side.
- Context object, which specifies the type of event that triggered the request message. For example, it can be admission control, resource allocation, or forwarding request.
- LPDP decision object, which represents the decisions made at the local PDP.

The decision message from a policy server to a client includes the associated client handle and one or more decision objects grouped by context object contained in the request message. The decision message from a PDP may include parameters like the following:

- Client handle object, which indicates the current state of the request from the PDP's perspective.
- Decision object, which represents the policy decision made by the server. The decision object contains a command to perform and decision data. The command can be install, remove, or NULL. The decision data can be client-specific, request state data, or others.

To avoid deadlock, the client starts a timer after issuing a request message. The PEP is responsible for deleting the outstanding client handle at the server.

The client-open and client-accept message exchange is the second pair of request-response messages used by a policy client establishing a connection to the policy server. The PEP initiates a client-open request message to attempt connecting to the PDP, upon the completion of its initialization. In response to the client-open message, the PDP sends either a client-accept message to accept the client connection or a client-close message with the appropriate error code to reject the client connection attempt.

The third pair of request-response messages is a synchronize state request from the policy server to a client and the synchronize state complete response from the client to the server. The synchronize state message is a server request to the client to resend its state information. The client performs the state synchronization by reissuing queries of the specified client type for the existing state in the PEP. Once the synchronization is completed, the client PEP sends a response message synchronize state complete to the PDP. In case of error, the PEP sends a delete request state message in place of synchronize state complete message to delete the state for the client handle specified in the request message.

The fourth pair of message exchanges between a policy client PEP and a policy server PDP is the keep-alive message. A keep-alive message is generated by the PEP randomly between 1.4 and 3.4 seconds of the interval specified in the keep-alive timer parameter contained in the client-accept message when the client connection is established. The PDP echoes a keep-alive message back to the PEP. The keep-alive messages are exchanged, even in the absence of any other messages, for the purpose of fault tolerance, as will be discussed shortly.

19.3.2.2 Unsolicited Messages This is a unilateral message from a policy client to a policy server, as shown between the PDP and the right-hand side PEP in Fig. 19-3. A PEP can unilaterally send a report state message to the PDP under two cases:

- After receiving a decision message from the PDP, the PEP needs to report the success or failure in applying the decision about access control, packet forwarding, or resource allocation. The PDP can also solicit a report state message from the PEP by setting the solicited message flag in the decision message sent from the PDP to the PEP.
- The PEP can send an unsolicited report state message on a periodic basis to report client-specific information for accounting and state monitoring purposes.

The PDP can also send an unsolicited decision message to the PEP to replace a previous decision or a piece of installed configuration data in response to a new event. The PEP is expected to honor the new decision.

19.3.3 COPS Fault Tolerance and Security

The COPS protocol is used in distributed client-server networking environments, and fault tolerance and security are major concerns.

COPS message-level reliability is achieved via the use of TCP connections between the policy server and policy clients. The connections remain in effect as long as the parties on both sides are operational.

On top of TCP connections that guarantee message delivery, the COPS protocol specifies keep-alive messages to monitor the connections between PEPs and PDPs, as described earlier. After a timeout period expires without receiving an echoed keep-alive message from a PDP, the PEP initiates reconnection to the primary PDP or attempts to connect to the backup PDP. While disconnected, the PEP relies on the local PDP for policy decision making. Once a connection is reestablished, the PEP notifies the PDP of any deleted state or new events that took place when the connection was lost.

COPS provides message-level security. Message-level security such as authentication, message integrity, and replay protection is achieved by authenticating and securing a message channel between the PEP and the PDP using an existing security protocol such as IPSC or Transport Layer Security (TLS) protocol. The IPSC authentication header is used for the validation of a connection and the IPSC encapsulation security payload is used to provide both validation and message secrecy. TLS is used for both connection-level validation and privacy.

Another COPS message-level security mechanism is the use of an integrity object in all the COPS messages exchanged between a policy server and its policy clients. The integrity message provides sequence numbers to avoid replay attacks. The policy server and each client choose the initial sequence numbers for each other, and those numbers are then incremented with each subsequent message sent over the connection. The initial sequence numbers are chosen in such way that they are monotonically increasing and never repeat for a particular key.

19.3.4 Policy Information Base

A policy information base (PIB) in essence is an information base that contains metadata or policy provisioning classes. For those familiar with SNMP, the PIB of COPS plays the same role as the SNMP MIB. For all practical purposes, a PIB does the following:

- It provides a common vocabulary for a policy server and policy clients to communicate with each other and exchange policy information.
- It allows only a predefined set of operations to be performed.
- It is defined using the formal specification language ASN.1 (ITU-T 1988)

A PIB consists of a set of provisioning classes, and each policy provisioning class contains a set of attributes. A PIB is conceptually organized as a tree where a branch is a provisioning class and the leaves are the attributes. Each provisioning class, as well as each class attribute, is uniquely identified by an object identifier.

There are three types of operations defined on PIB:

- Adding to or deprecating a provisioning class from an existing PIB
- Adding to or deprecating an attribute from an existing provisioning class in a PIB
- Extending or augmenting an existing provisioning class in one PIB with a new provisioning class defined in another PIB

The instances of provisioning classes are encoded into the COPS messages as described in the preceding section and exchanged between the PDP and the PEP.

A note on the difference between the policy rules in the policy repository and the provisioning classes in a PIB. The policy rules are declarative statements of business objectives. They are represented in a vendor- and device-independent fashion. They represent the static, relatively high-level knowledge that needs to be implemented in dynamic states and vendor-specific devices. The PIB provisioning classes are the mechanism for representing the dynamic and vendor-specific policy data to implement the policy rules at a device. There is no one-to-one correspondence between a policy rule and a provisioning class in the PIB.

19.4 AN END-TO-END QOS EXAMPLE

An end-to-end QoS example will thread together all the elements of quality of service discussed in this chapter and the elements of MPLS and QoS models such as DiffServ discussed in Chaps. 17 and 18 (Rajan et al. 1999).

19.4.1 Scenario

Assume a customer signs up for a telecommuting service that guarantees an agreed-upon amount of bandwidth for the fixed hours, say, from 8 a.m. to 6 p.m., on a daily basis during the weekdays.

The service is realized through an MPLS- and QoS-enabled packet broadband network and can be self-provisioned by the customer via a Web-enabled service provisioning system. The hypothetical packet broadband network consists of the following components:

- A Web-based service provisioning with QoS capability
- A directory-enabled policy repository
- A policy decision server or policy decision point
- A set of MPLS-enabled routers, each with a policy client
- DiffServ supported by all nodes

19.4.2 End-to-end QoS process

The customer's subscription to the telecommuting service triggers the whole QoS provisioning process. For illustration purposes, the various possible error conditions are not considered.

Step 1: Policy rule provisioning. The subscription with an SLA that the customer entered from the Web-based service provisioning system first goes to a directory-enabled QoS policy system that maps the SLA into a set of business rules, and then stores the rules in the policy repository along with the customer data. The SLA specifies the maximum and minimum guaranteed bandwidth and the downtime limit.

Step 2: Policy decision provisioning. A QoS policy decision system, serving as the policy decision points of COPS, receives a notification from the Web-based service provisioning system of the customer's subscription of the telecommuting service. The PDP queries for and retrieves the business rules for this subscriber via LDAP from the policy repository. The policy rules are mapped into the router-specific and implementation-specific PIB objects. The policy decision is to set up a MPLS label-switched path for this customer for the designated hours with DiffServ's expedited forward class of service.

Step 3: MPLS LSP setup. The PDP sends a decision message to the PEP at an edge router that will first see the customer's packet flow. The decision message contains a command to install the LSP with a specification of ingress node, all the intermediate nodes, and the egress node for the LSP. The ingress node where the PEP resides initiates the LSP setup process using the LDP-CR.

Step 4. QoS provisioning at PEP. The PDP then downloads all the PIB object instances to the PEP at the ingress router, which in turn convert them into the internal implementation-specific data structure.

At the service rendering time, when the user starts sending the traffic, the ingress node recognizes the customer packet flow and sends the traffic onto the provisioned LSP for the customer.

REVIEW QUESTIONS

1. QoS policy converts a user service request into a network resource request to be implemented at a network device. Discuss why QoS policy is the last piece of the end-to-end QoS puzzle from this perspective.
2. Describe the IETF QoS architecture in terms of three major modules: policy repository, policy decision point, and policy enforcement point. Discuss the relationships between them.
3. Describe the QoS policy repository provisioning process and the entities involved in the process such as policy provisioning user interface, policy translators, and policy repository.
4. Describe the structure of the policy repository and the interface it provides for clients to access the repository.
5. Describe three types of request-response message exchanges between a policy server and a policy client: decision request and decision response, client open and client accept, synchronize state and synchronize state complete messages.
6. Describe the types of actions that may be contained in a decision message from a policy server to a policy client.
7. Describe the different message-level security mechanisms used in the COPS message exchanges and how COPS guards against replay attack.

REFERENCES

Boore, B., Ellesson, E., et al. 2001. "Policy Core Information Model Specification (v1)." IETF RFC 3060. Web site: www.ietf.org.

Chan, K., Seligson, J., et al. 2001. "COPS Usage for Policy Provisioning." IETF RFC 3084. Web site: www.ietf.org.

Durham, D. (Ed.). 2000. "Common Open Policy Service (COPS) Protocol." IETF RFC 2748. Web site: www.ietf.org.

ITU-T. 2001. "Information Technology—Open Systems Interconnection—The Directory: Overview of concepts, models, and services." Recommendation X.500. Web site: www.itu.int/ITU-T.

ITU-T. 1988. "Specification of Abstract Syntax Notation One (ASN.1)." Recommendation X.208. Web site: www.itu.int/ITU-T.

QoS Forum. 1999. "Introduction to QoS Policies." White paper. Web site: www.qosforum.com.

Rajan, R., et al. 1999. "A Policy Framework for Integrated and Differentiated Service in the Internet." *IEEE Network*, Vol. 13, No. 5, pp. 36–41.

Wahl, M., Howes, T., and Kille, S. 1997. "Lightweight Directory Access Protocol (v3): Attributes Syntax Definitions." IETF RFC 2251. Web site: www.ietf.org.

PART 5

Packet Broadband Network Services

Part V of this book moves up along the network protocol layers to the application and service layer and introduces three types of packet broadband services: storage area network (SAN), VPN, and voice over IP and multimedia applications.

Storage area network is a broadband networking technology as well as a service. On the packet broadband network infrastructure, SAN provides connectivity between storage devices and network servers that consume data and thus enables a host of related services: remote data backup, data recovery, data mirroring, etc. SAN has become a critical broadband service in recent years as a result of the growing trend that large amounts of data and data storage devices are increasingly being networked and shared via the Internet.

VPN is a packet broadband service with great potential. A packet broadband virtual private network is built on shared public packet network facilities to deliver services that, in appearance, are as reliable and secure as the services delivered by a dedicated private network. There are two types of broadband VPN service: access VPN and internetworking VPN. Access VPN connects mobile or remote users like telecommuters to an enterprise LAN. Internetworking VPNs are built on top of public IP networks to connect geographically distributed corporate LANs to support services like intranets and extranets.

Multimedia and voice over IP applications are built on broadband IP networks to provide revenue-rich voice and multimedia applications. H.323 and SIP are the two most common technologies for providing these services over broadband IP networks.

CHAPTER 20

Storage Area Networks

20.1 Introduction

Storage area networks are among the primary applications of packet broadband networks. A SAN itself is a high-speed special-purpose network that interconnects different kinds of data storage devices to provide data storage service to networked users. Typically, a storage area network is part of an overall communications network for an enterprise. A strong push is underway to extend the SAN solution beyond LANs to remote locations for backup and archival storage, using wide-area network carrier technologies such as ATM or SONET.

The typical operations SANs support include disk mirroring, backup and restore, archival and retrieval of archived data, data migration from one storage device to another, and the sharing of data among different servers in a network.

20.1.1 Evolution of Storage Solutions

Data storage solutions have evolved along with computer networking technologies, with three distinct phases of development (Khattar 1999):

Server-attached storage. Server-based data storage was the first generation storage solution in the early 1970s to 1980s. The storage was part of a general-purpose computer server. Data access was specific to the computing platform, operating system, file system, and database system being used.

Network-attached storage. Network-based storage was the second-generation data storage solution to be developed. It involves a dedicated file server on the network serving all the workstations and hosts connected to the network. The dedicated file server communicates with other hosts via a LAN-specific protocol such as Ethernet. The network-wide disk storage has a logical partitioning to accommodate different applications and purposes. Storage and data sharing is limited to the LAN where the file server resides. Coupling between the file server and data storage is required. This is the most prevalent scheme in use today.

Storage area networks. Storage area networks are the next step in data storage solution. SAN is characterized by the separation of data server and data storage. The storage functions are distributed over a high-speed broadband network and a storage device located far away from the data server. SAN features high-performance, dedicated, network connectivity

over optical fiber in most cases. SAN also features capabilities such as device pooling, high availability, and centralized management.

20.1.2 Motivations for Storage Area Networks

The move from network file systems to storage area networks has been motivated mainly by the need to support the emerging Internet-based business models. Specifically, the new business model places new requirements on the data storage systems:

- They must be scalable to handle increasingly larger amounts of data in a real-time fashion. The systems must be able to adapt to rapidly changing demands and requirements of data storage.
- They must provide reliable and flexible access to data. Making important business decisions and transactions depends on the availability and timely response of the storage systems, even in the face of natural disasters like earthquakes, floods, hurricanes, or in case of human-produced disasters like terrorist attacks.
- They must be able to transfer sensitive business data securely. This is a key prerequisite for the acceptance of a new generation of SANs as they migrate from private-lease facilities to public packet broadband networks.
- They must be able to support heterogeneous IT environments and globalization as business operating environments become truly global.
- They must be high performance to support an amount of data transfer that was unimaginable just a few years ago.
- They must offer centralized management of data storage even as the storage functions are distributed.

20.1.3 SAN Standards

SAN standards play a key role in the wide acceptance and effective development and deployment of SAN technologies. Storage area networks straddle multiple industries like computer, IT, communications networks, storage devices, and transmission equipment, among others. Thus the organizations involved in the SAN standardization effort are numerous. While some organizations focus on SAN solution advocacy,

others are active in defining the SAN architecture and market requirements and developing SAN equipment interoperability specifications. Some key SAN standards organizations include the following:

Storage Network Industry Association (SNIA). This is an international computer system industry forum of developers, system integrators, and IT professionals that focuses on advocating and promoting storage networking technology and solutions.

Fibre Channel Association (FCA). FCA is a nonprofit organization whose mission is to nurture and facilitate the development of markets for fibre channel products. It sponsors education programs, monitors standards activity, and promotes interoperability among member products.

Fibre Channel Community (FC/C). FC/C (not to be confused with FCC, the Federal Communications Commission) is another nonprofit organization that, in addition to providing marketing support, uses venues like trade shows and exhibits to promote fibre channel technology and the use of SANs. It consists of over 100 computer industry-related companies, and focuses on the migration of storage attachment to network-attached storage in networking environments.

The formal standards organizations that play leading roles in defining the SAN standards include ANSI and IETF. ANSI was responsible for defining the Fibre Channel (FC) standards, with two ANSI groups, X3T10 and X3T11, as major players in developing the earlier SAN standards. X3T10 was responsible for the SCSI standards, which deal with the interface to storage devices for LANs. X3T11 was responsible for X3T11 fibre channel standards, the most widely deployed SAN standards to date (ANSI 1996a, 1996b, 1997, 1999). IETF is active in the development of next-generation SANs that will provide SAN solutions on IP networks.

The current storage area network solutions are largely LAN-based, accommodating enterprise customers. An important trend currently underway is to expand SANs beyond LANs and achieve scalability to the global level.

20.2 Storage Area Network Basics

This section describes the components and management aspects of a typical SAN as deployed today that is largely LAN-based and enterprise-specific.

20.2.1 Storage Area Network Components

SAN is a network that connects a set of distributed storage devices to a set of data servers. Thus, a SAN consists of three major components: storage devices, one or more data servers, and the interconnection and interface protocol between the storage devices and the data servers, as shown in Fig. 20-1 (Khattar 1999; Tang 1997).

20.2.1.1 Storage devices Since the first disk drive became commercially available in 1956, storage devices have evolved along with computer and communications technologies. Based on some estimates, between 1970 and 1990, the storage device capacity measured by number of bits per unit of area grew by an averaged annual rate of around 25 percent. Between 1990 and 1997, the annual growth rate of storage density was more than 60 percent. Since late 1997, storage density has almost doubled

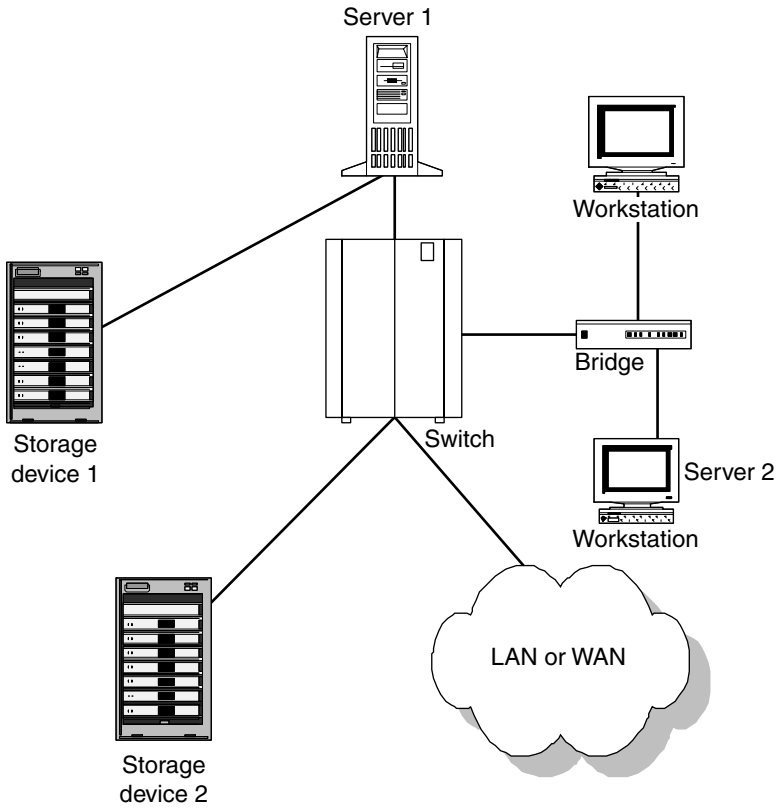


Figure 20-1
Storage area network components.

every year and is approaching 200 gigabits per square inch, a spectacular achievement in a little more than 35 years.

Storage devices come in various configurations to support network server-attached, network-attached, and SAN applications. Examples of the configurations include single standalone disk, disk array, and tape library. Stand-alone disks are large, single disks that can be accessed by a single server or shared among several servers. One draw back of this configuration is lack of redundancy.

An alternative scheme is redundant array of independent (or inexpensive) disks (RAID), developed to increase the performance and reliability of data storage by spreading data across multiple drives. RAID is an assembly of disk drives, or a disk array that operates as a single storage unit. In general, the drives can be any storage system with random data access, such as magnetic hard drives, optical storage, magnetic tapes, etc. Different levels of redundancy of RAID have been defined by the National Storage Industry Consortium (NSIC). In a redundant mode, if one or a few disk drives fail, they can normally be exchanged without interruption of normal system operation. Thus, disk arrays can ensure that no data is lost if one disk drive in the array fails.

A tape library is another type of storage device that generally consists of little more than one or more tape drives, together with some kind of autoloader capable of inserting the required tape cartridges under software control. Tape libraries are designed to automate the network backup process with automatic tape changes, thus offering the ability to protect large amounts of data.

20.2.1.2 Server A data server can be a mainframe or a regular computer that has the interface ability of accessing the data for read, write, and query operations. Data servers have been closely associated with operating systems since the beginning of computers. The commonly available servers on the market include Windows NT, mainframe-based OS/360, and different Unix-based servers.

20.2.1.3 SAN Connection Models There are two basic modes of connections between storage devices and servers: direct and switched, as shown in Fig. 20-1.

A direct connection can be established between a server and a storage device, such as the connection between storage device 1 and server 1 in Fig. 20-1. A variant of this connection mode is the combination of a storage device and a server within a system, also shown in Fig. 20-1. This

mode of connection is suitable for early SANs where the SAN is limited to a LAN in a network-attached storage model.

The switched mode of storage-server connection allows any-to-any connections within a SAN and allows the SAN to go beyond a geographical area and interconnect to other SANs via wide-area networks. For example, in Fig. 20-1 server 2 can reach storage device 1 via a SAN switch.

20.2.1.4 SAN Interconnecting Devices A SAN interconnecting device is responsible for connecting the storage device, servers, and other networking devices. Devices of this type include the SAN router, hub, SAN switch, and SAN gateway, as shown in Fig. 20-1:

SAN bridge. A SAN bridge facilitates the communications between a LAN and a SAN segment and between networks with different protocols. An example is the fiber connection (FICON) bridge that allows an enterprise system connection (ESCON) to be tunneled over the Fibre Channel protocol.

SAN hub. A SAN hub is a device that allows a fixed number of storage devices and servers to be directly connected to it, and thus to each other via the hub. The fibre channel hub, a common type of SAN hub, allows up to 126 nodes to be connected to it. Compared to a SAN switch, a SAN hub is less versatile, but it is a cost-effective solution.

SAN switch. A SAN switch is like any other switch that switches storage data traffic as opposed to other traffic. A SAN switch enables a server to access any other storage device that is also connected to the switch, as opposed to only the one it is directly connected to. A SAN switch can take on a variety of forms, and is often technology-specific. A SAN hub can be viewed as a simplified version of a SAN switch. Examples of full-blown SAN switches include fibre channel switches and Fibre Channel over IP (FCIP) gateways, as will be illustrated later.

SAN gateway. A SAN gateway is a device that interconnects two or more different networks or devices with the optional capability of performing protocol conversion. A gateway is often used to connect a local SAN to a WAN, thus extending the SAN across the WAN.

20.2.1.5 Data Storage Interface A storage interface consists of a protocol for a server to communicate to the storage device on the SAN, with both the server and the storage device supporting the protocol and a set of physical connection parameters. Several interfaces have been defined over the years. The common ones include SCSI, ESCON, FICON, and HPSS, as explained below.

SCSI Small Computer System Interface (SCSI) is a set of device access commands originally developed by Apple Computer; later they became the ANSI standard storage device interface (ANSI 1996c). SCSI allows servers to access storage and other peripheral devices such as disk drives, tape drives, CD-ROM drives, printers, and scanners with fast parallel interface. SCSI ports are found universally in most of the personal computers today and are supported by all major operating systems. For example, a widely deployed version of SCSI, called *Ultra-2 SCSI,* uses a 16-bit bus and can transfer data at up to 80 megabytes/s. SCSI allows up to 7 or 15 devices (depending on the bus width) to be connected to a single SCSI port in a daisy-chain fashion.

The original SCSI, now known as *SCSI-1,* evolved into SCSI-2, known as *plain SCSI* as it became widely supported. The latest version is SCSI-3, which consists of a set of primary commands and additional specialized command sets to meet the needs of specific device types. The SCSI-3 command sets are used not only for the SCSI-3 parallel interface but also for additional parallel and serial protocols, including Fibre Channel, Serial Bus Protocol (SBP), and Serial Storage Protocol (SSP). The latest SCSI-3 standard increases the maximum burst rate from 80 to 160 Mbps by operating at the full-clock rate rather than the half-clock rate.

ESCON *Enterprise systems connection* is a generic name for a network storage system developed by IBM and other vendors. The key component is an interface specification responsible for interconnecting IBM S/390 computers with each other and with attached storage, locally attached workstations, and other devices using optical fiber technology and dynamically modifiable switches called ESCON Directors. ESCON's fiber optic cabling can extend this local-to-the-mainframe network up to 60 km (over 37 mi) with chained Directors. The data rate on the link itself can reach a maximum of 200 Mbps. Vendor enhancements may provide additional distance and higher throughput.

FICON Fiber connection is a newer generation of storage device interface defined by IBM to succeed the ESCON interface. FICON is a high-speed input/output (I/O) interface for mainframe computer connections to storage devices. With the high-speed FICON I/O capacity achieved through the combination of a new architecture and faster physical link rates, the FICON interface can be eight times as efficient as ESCON.

FICON use the ANSI Fibre Channel—Physical and Signaling Interface (FC-PH) standard, which specifies a signal, cabling, and bidirectional link

transmission rate of 100 Mbps at distances of up to 20 km (ANSI 1999). A bridge feature enables the support of existing ESCON control units and the support of full-duplex data transfers.

HPSS The High-Performance Storage System (HPSS) is a new, open system interface designed to manage petabits of data produced and used by supercomputers. HPSS is an open system based on the IEEE Mass Storage Reference Model, Version 5, the established design guide for very large-scale storage systems (IEEE 1994). All computer and storage nodes may be attached directly to the network so that data is transferred by the most direct route, at network speeds, without interruption.

HPSS is a major development project, which began in 1993 as a Cooperative Research and Development Agreement (CRADA) between the U.S. government and the computer and storage network industry. It is the result of a collaborative effort by leading U.S. government supercomputer laboratories and industry to address the urgent need for high-end storage requirements. HPSS is offered commercially by IBM Global Government Industry, with new releases available on a continuing basis.

20.2.2 SAN Management

SAN management is a combination of management functions in a number of areas including network management, application management, storage management, and software distribution management.

The Simple Network Management Protocol (SNMP), originally designed for enterprise and data network management and the most widely deployed management protocol, has been widely adopted as the SAN management protocol. Many SAN-specific management information bases have been defined and adopted as IETF standards. The MIBs defined so far have focused on the configuration and provisioning of SAN components.

Enterprise Storage Resource Management (ESRM) is another set of industry-consortium-type SAN management standards that covers several areas. One area is the real-time monitoring of all the components of a SAN. Another is storage management, including operations such as backup, restore, archive, retrieve, and space usage management. Still another area is storage resource management, which includes capacity management, data and device migration management, and storage assets management.

20.2.3 SAN Applications

SAN is intended to support a wide range of applications and services that are becoming a necessity in Internet-based communications. The services include data protection and disaster recovery, long-distance mirroring, data sharing, and remote backup.

DATA PROTECTION AND DISASTER RECOVERY The traditional data recovery methods such as recovery from tape are no longer satisfactory for the current Internet e-business models. The new requirements of high availability of data and systems mean that duplicate systems and data must be ready at any moment to take over in the event of an active system failure.

There are several methods for achieving data protection via redundant copies. They include storage mirroring, remote cluster storage, peer-to-peer remote copy, extended remote copy, and concurrent copy.

DATA SHARING A SAN, as a special-purpose network, separates data storage from the server and centralizes the data storage. In so doing, SANs allow data to be shared among multiple servers with minimal impact on system performance, and, depending on whether the server platforms are the same, the data sharing can be homogeneous or heterogeneous.

LONG-DISTANCE MIRRORING This service transparently replicates data no matter how it is organized or where it is located. A SAN provides the flexibility to place mirrored disks in another building close by, or at a remote location across the country.

REMOTE DATA VAULTING AND BACKUP A SAN as a dedicated network allows data vaulting and backup to go beyond the traditional near-line or off-line schemes. Data backup can be done remotely by specialized data storage service providers, and can be done dynamically based on the need of the customer in a nonintrusive manner without affecting the local LAN. Thus, this is also known as *LAN-less* or *server-free backup*.

20.3 Introduction to Fibre Channel

Fibre Channel is an integrated set of SAN standards developed by ANSI that defines a layered architecture for SAN networks, a set of SAN topologies, and a set of services SANs provide to users (ANSI 1997, 1999).

One essential concept of storage area networks is the separation of data servers from data storage, so the interconnection of servers and data storage is a key issue to be addressed by any SAN solution. Fibre Channel brings networking capabilities to server-storage connectivity.

20.3.1 Overview of Fibre Channel Architecture

Several FC concepts are important to understanding Fibre Channel architecture. The concept of *channel* refers to a direct or switched point-to-point connection between the communicating devices on a SAN network. Such a channel can transport data at a very high speed.

The *switch* connecting storage devices is called a *fabric* in FC terminology. A *link* in a SAN consists of two unidirectional fibers transmitting in opposite directions and their associated transmitters and receivers. Each *fiber* is attached to the transmitter of a port at one end and the receiver of a port at the other end.

A fibre channel *port* is a physical interface on a data network node where a channel can either originate or terminate. A port can be one of several types. A port on a node device such as a PC or a workstation is called an *N_Port,* while a port on a FC fabric is known as an *F_Port* and a port with arbitrated loop capabilities is known as an *L_port.* These types are obviously not mutually exclusive and combining a port with loop capabilities on a node results in an *NL_port.*

The Fibre Channel standards, in the spirit of the OSI network reference model, define a five-layer SAN network architecture, FC-0 through FC-4, although the Fibre Channel layers are not aligned with the OSI network layers on a layer-by-layer basis. The functions and the approximate mapping to the OSI layers are shown in Fig. 20-2 (ANSI 1996b; ANCI 1999; InterOperability Lab 1998).

20.3.2 FC-0 and FC-1—Physical Layer

The FC-0 and FC-1 layers map to the physical layer of the OSI network reference model. FC-0 at the bottom of the functional hierarchy of Fibre Channel defines the physical link of a SAN, like the physical layer of the OSI model, including the fiber, fiber connectors, transmitter, receiver, optical controller, etc. The standards allow a range of different kinds of optical links to be used to meet different performance and price efficiency requirements.

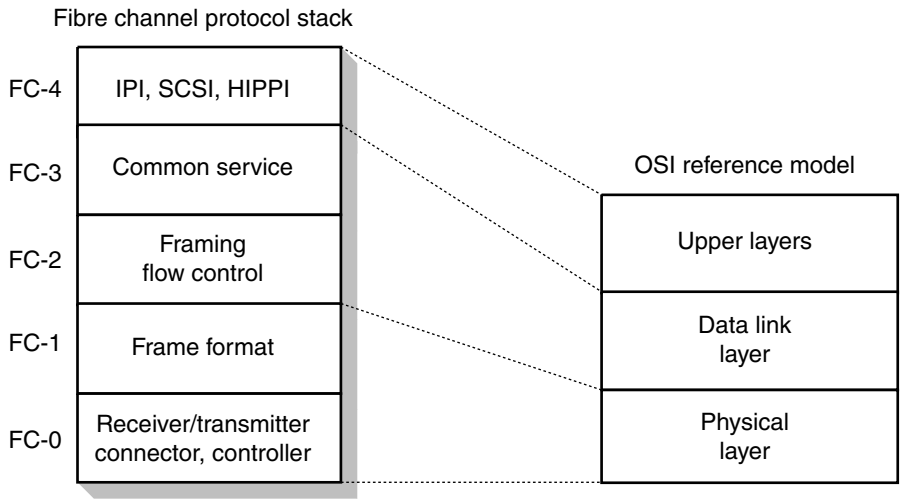


Figure 20-2
Fibre Channel network layers.

FC-1 defines other parts of the physical layer function: information encoding and decoding algorithms, special characters, and error control. It uses 8B/10B character encoding, i.e., for every 8 bits of data, 2 extra control bits are added for transmission on the wire.

20.3.3 FC-2—Data Link Layer

The FC-2 layer performs the functions of the data link layer of the OSI network reference model: signaling protocol, framing rules, frame exchange and sequence management, flow control, and frame segmentation and reassembly. The FC-2 layer essentially provides a transport pipe between two ports.

20.3.3.1 FC Frame The FC frame, like the frame of Ethernet, defines the basic building blocks of an FC connection at the data link layer with the mechanisms for flow control and sequence management. An FC frame consists of four fields, as shown in Fig. 20-3:

- *Frame flag*, at the start and end of a frame, 4 bytes each. The flag delineates one frame from another.
- *Frame header*, a 24-byte field that specifies the source and destination addresses, frame sequence number, frame type, sequence ID, and exchange ID, as shown in Fig. 20-3.

- *Data fields,* up to 2112 bytes that include a 2048-byte payload and an optional 64-byte header.
- *Cyclic redundancy check,* a 4-byte field used to detect transmission errors in the header.

20.3.3.2 Frame Sequence and Exchange FC introduces two new concepts at layer-2: sequence and exchange. A sequence is a set of one or more related frames transmitted unidirectionally from one N_port to another N_port. Each frame within a sequence is uniquely identified by a Seq_ID in frame header, as shown in Fig. 20-3. The idea is to have a whole block of data, bundled in a set of frames, treated as a single unit. For example, error recovery of the upper protocol can be performed at the sequence boundaries.

An *exchange,* in FC terminology, refers to an even larger block of data, consisting of one or more sequences involved in the same data transfer operation. Obviously these sequences must be sequential, and only one can be active at a given time. An exchange is identified by an exchange ID in the frame header, as shown in Fig. 20-3.

FC effectively provides a data containment hierarchy, with a frame as basic unit. With this hierarchy, increasingly larger blocks of data can be transferred as a single unit between two FC nodes, with the exchange at the top of the hierarchy.

In addition, the Fibre Channel specifications define a set of protocols for FC ports to exchange configuration parameters in setting up operating environments or for providing service to the upper layers.

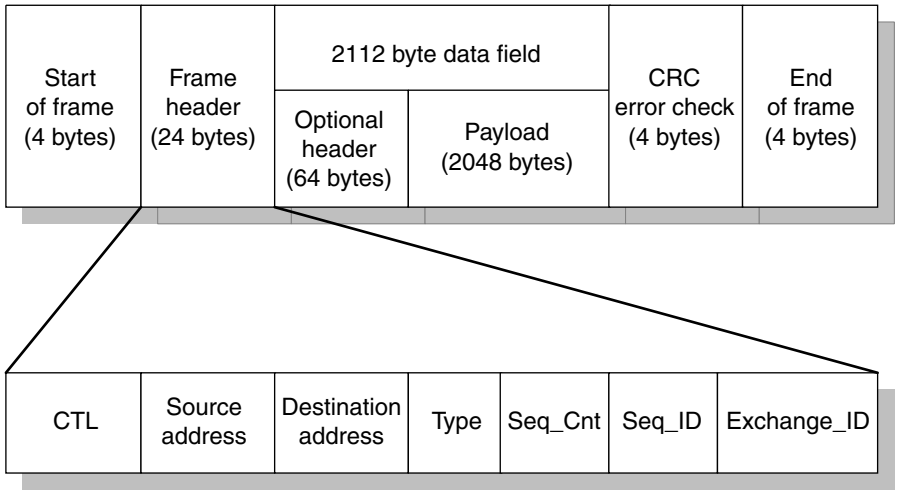


Figure 20-3 The FC frame structure.

20.3.3.3 Flow Control of FC The FC-2 layer also provides flow control mechanisms at the data link layer to regulate the frame flow between two N_ports or between an N_port and an FC fabric to prevent the overflow of frames. There are different flow control mechanisms to accommodate different classes of services.

End-to-end flow control, designed to support class 1 service, regulates the frame flow between N_ports via the acknowledgment of receipt of data frames by the destination N_port. When the receiver buffers overflow, a frame with a "busy" indication is sent to the sender port. The sender port then slows down the pace of transmitting frames.

The second flow control approach is called *buffer-to-buffer flow control,* and is used between an N_port and an F_port, or between two N_points in a point-to-point topology. Two ports establish a credit count to indicate the available receiving buffer at the time the initial parameter exchange is established between the two ports. When a receiver port has free buffers available, it signals to the transmitting port to send frames by issuing a Receiver_Ready primitive. This flow control is designed to support the class 3 service, as described below.

20.3.4 FC-3 Layer

The FC-3 layer complements the FC-2 layer by providing a set of special service features on top of the framework provided by the FC-2 layer. Those features include the following:

- *Stripping.* This feature allows an FC system to use multiple N_ports in parallel across multiple links to transmit a large block of data.
- *Hunt group.* This feature provides the ability for more than one port to respond to the same alias address so that, if one port is busy, another port can respond. This reduces the chance of rejecting a call.
- *Multicast.* This feature allows a port to deliver a single transmission to multiple destination ports. The destination ports can be all N_ports on a fabric or a subset of the N_ports on a fabric.

Note that the FC-3 layer provides the service features that normally belong to the network layer but it does not provides a full-blown network layer function.

20.3.5 FC-4—Upper Layers of FC

The FC-4 layer supports the multiple protocols of the upper layers, such as the network, session, presentation, and application layers, by specifying the rules for mapping between the upper-layer protocols. This allows Fibre Channel to carry data from a range of applications and networks using different protocols.

The following network and channel protocols are either already supported or in consideration to be supported:

- Small Computer System Interface
- Intelligent Peripheral Interface (IPI)
- High Performance Parallel Interface (HPPI) Framing Protocol
- Internet Protocol
- ATM Adaptation Layer 5 for data traffic
- Fibre Channel Link Encapsulation (FC-LE)
- Single Byte Command Code Set Mapping (SBCCS)
- IEEE 802.2 Logical Link Control

20.3.6 Three Classes of Service

FC defines three classes of services to accommodate different types of application traffic and to differentiate the services offered to customers.

Class 1 service provides a dedicated connection with a guarantee of maximum bandwidth between two N_ports. A connection, once established, is retained and maintained by the FC fabric until the session is finished. It guarantees in-order delivery of frames. Each frame that is sent from the source port, either individually or as a member of a sequence, is acknowledged by the receiving N_port. This class of service is best suited for sustained, high-throughput transactions.

Class 2 service provides frame-switched, connectionless service that allows bandwidth to be shared by multiplexing frames from multiple sources onto the same channel. The frames arriving at the destination port may not be in the same order as they were sent. The receiving port uses information such as sequence count, sequence ID, and exchange ID to recover the original order of the frames. Like class 1 service, the receiving port acknowledges each frame it receives. If the receiving

port does not have a free buffer available due to a congestion condition, it sends a Busy message to the source port and the sender tries again.

Class 3 service is same as class 2 service except that frame delivery is not acknowledged by the receiving port. This is similar to the IP datagram service that provides fast and high throughput with little overhead but without a QoS guarantee. Class 3 services are more suitable for those applications that can tolerate frame loss to a certain degree but are highly real-time-sensitive.

20.3.7 FC Network Topologies

Fibre channel systems support three different interconnecting schemes between ports: switching, hub, and loop. As a result, Fibre Channel allows for three topologies: point-to-point, arbitrated loop, and switched.

20.3.7.1 Point-to-Point Topology A point-to-point topology is the simplest of all. It is composed of only two directly connected fibre channel devices that are capable of bidirectional communications. The transmitter of one device goes to the receiver of the other device and vice versa in the other direction, without any interconnecting device in between. The link is dedicated for the two ports on the two connected nodes. This is well suited for providing class 1 services, i.e., dedicated connections with guaranteed bandwidth. Two FC nodes must use the same data rate at all times in this configuration.

The application examples of the point-to-point topology include one storage device connected to a server, as shown in Fig. 20-4, one workstation directly connected to another workstation, and a file server connected to a disk array.

20.3.7.2 Fabric-Switched Topology Also called *Fibre Channel fabric topology* or *cross-point topology,* this Fibre Channel system configuration

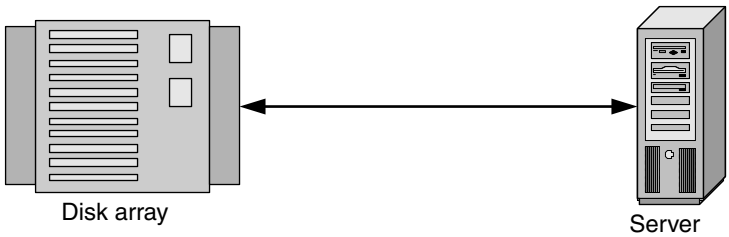


Figure 20-4
Fibre Channel point-to-point topology.

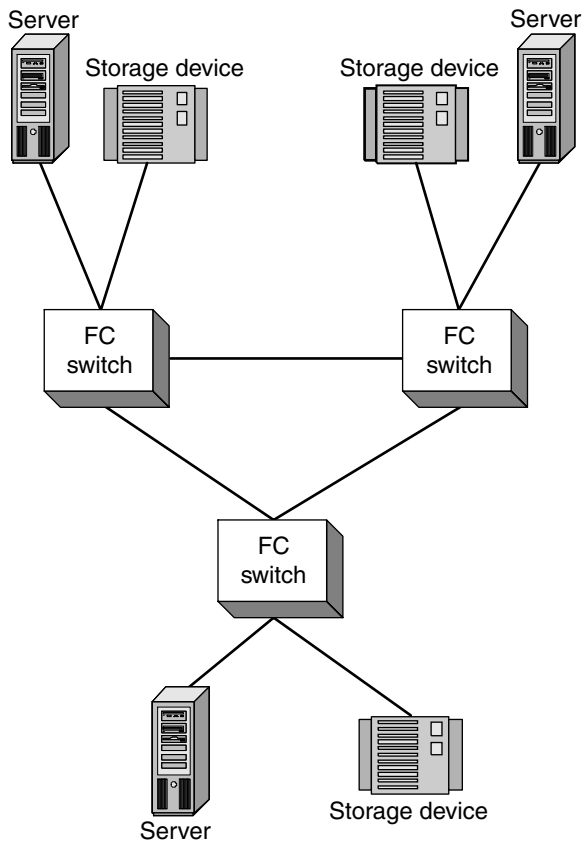


Figure 20-5
Fabric-switched FC system configuration.

allows for a large number of FC devices to be interconnected via FC switches. The key components in this system configuration are FC switches, as shown in Fig. 20-5. At a high level, each FC switch consists of at least F_ports (incoming and outgoing ports), a switching fabric that cross-connects the ports, and software support of FC protocols.

The main responsibility of an FC switch, which functions very much like a regular network switch, is to route data frames to their destination. The switch uses the destination address in each frame header to send the frame out via the appropriate F_port toward its destination. Switched FC systems can be cascaded into multiple layers of FC systems, to achieve any-to-any connections in a large interconnected network. The downside of this configuration is that the management tasks such as configuration, security, and performance monitoring all become more complicated.

20.3.7.3 Arbitrated Loop The arbitrated loop configuration of FC has become the de facto "standard" configuration of FC systems in recent years (ANSI 1996a) and thus deserves a more detailed description. It is termed *arbitrated* because each node in the configuration goes through an arbitration process to gain the right to send data. As shown in Fig. 20-6, a simple arbitrated loop configuration consists of a set of FC nodes known as *loop nodes* and a set of unidirectional links that connect the loop nodes into a one-way loop. A maximum of 127 nodes can be connected into a loop.

An arbitrated loop has two phases: initialization and normal operation. A loop must go through an initialization phase before entering the normal operation phase. The following major steps constitute the initialization process:

1. Each port transmits a loop initialization primitive (LIP) upon its power on. This also happens when a port detects a link failure. The LIP will trigger all the other nodes to transmit their LIP messages.
2. The loop nodes select a loop master. Then each port continuously sends a Loop Initialization Select Master message until a loop master is selected. If there is an FC switch or hub device connected to the loop, the switch or hub becomes the loop master. Otherwise, the port with the lowest port ID becomes the loop master.
3. Then the loop master dynamically negotiates with all the loop nodes and coordinates the process where each port gets to select a 1-byte loop physical address. Then all the ports report their loop physical addresses back to the loop master. At this point, the loop master notifies each port that the loop is operational. The 1-byte length of the loop physical address limits the number of nodes on the loop to 127.

An FC system in a loop configuration operates as follows: At any given time, only one loop node can be transmitting data. When a device on the loop desires to transmit data, it first arbitrates to win control of the loop, which it accomplishes by transmitting the arbitrate (ARB) primitive signaling message, including its own arbitrated loop physical address in the signaling message. If more than one node on the loop is arbitrating, the loop physical addresses embedded in the signaling messages are compared. When an arbitrating node receives another node's ARB message, it compares two loop physical addresses. The ARB message with smaller physical address value is forwarded, and the ARB message with higher loop address is blocked. An arbitrating node, once it gets back the ARB

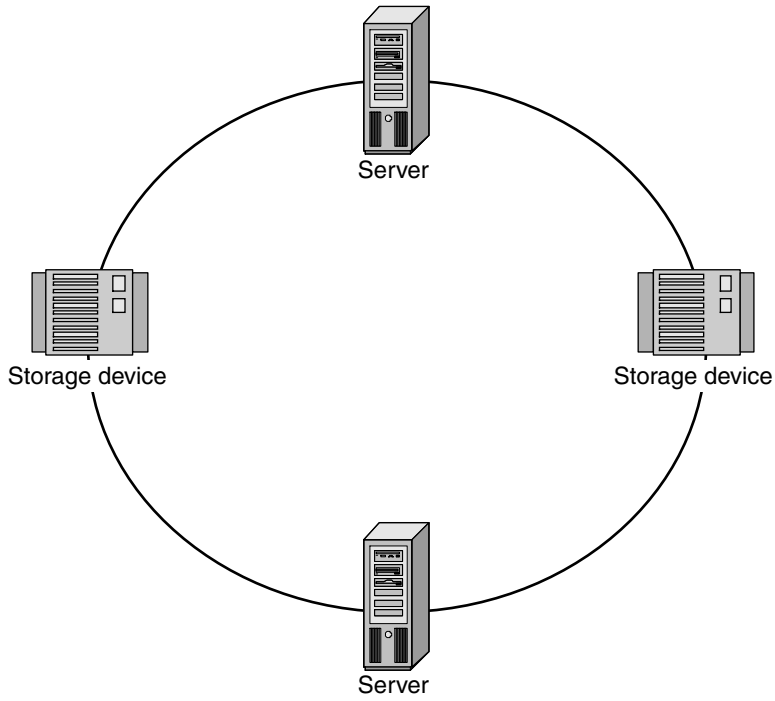


Figure 20-6
Arbitrated loop configuration of an FC system.

message it sent out, gains control over the loop. This way, the loop node with lowest loop address value will gain control over all the other arbitrating nodes.

Once a node takes control, it establishes a point-to-point connection to the destination node on the loop using the login procedure. The entire bandwidth of the link is at the disposal of the loop node with control. There is no limit on how long a loop node can retain control of the loop to accommodate the large bulk data transfer of the FC system. However, there is an optional access fairness algorithm that prohibits a loop node from arbitrating again before all the other nodes have had a chance to arbitrate.

There are several advantages associated with the loop topology that make it such a popular choice of FC system configuration. The topology has a simple implementation and a low-cost solution that can accommodate a relatively large number of devices and can accommodate diverse types of loop nodes ranging from storage devices to servers to FC hubs and switches. In addition, this topology supports virtual point-to-point connection that connects any two nodes on the loop. The extensibility of the loop topology allows a loop to be connected to another loop, or a switched configuration.

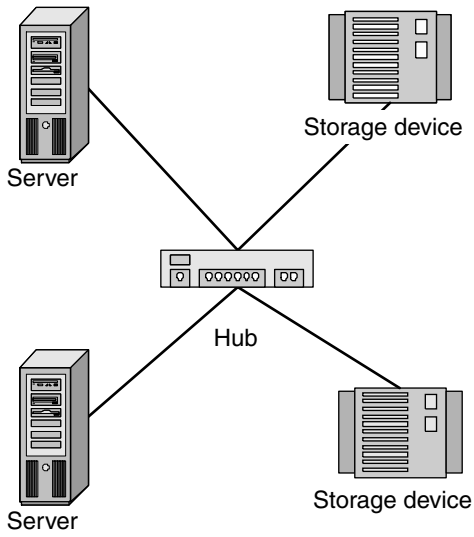


Figure 20-7
A hub-based FC system configuration.

The downside of this topology is the lack of fault tolerance. As in any other loop topology, should any link in the loop fail, the communications between all the loop ports are interrupted.

A hub-based configuration of a FC system can be viewed as a mix of switched topology and arbitrated loop. As shown in Fig. 20-7, a hub-based configuration has a set of FC nodes, each node is connected to a central FC hub, and it is through the hub that messages and data are sent from one node to another.

20.4 SAN over IP

SAN over IP networks are a recent trend in SAN technology development. This section first discusses the motivations for migrating SAN over IP-based MAN and WAN and then describes two approaches to SAN over IP: Fibre Channel over IP and iSCSI.

20.4.1 Introduction to IP SANs

The limitations of the current SAN solutions and the new requirements of the Internet-based e-business model are mainly responsible for the development of SAN over IP solutions.

The FC SAN systems and most of the SAN solutions until very recently have been largely LAN-based. Both storage devices and servers are connected to the same LAN. Fibre Channel SANs are limited to a distance of about 10 km, normally confined within a single LAN. This is a severe limiting factor for applications such as business-continuance and disaster recovery.

The current FC SAN solutions primarily cater to large enterprises that can afford a dedicated network like FC SAN. The solutions center on large campus and data centers. However, few solutions are available and affordable for small and mid-size businesses that are scattered around geographically, and often located far away from the data centers.

On the other hand, with the rise of e-business models, network-stored data has become mission-critical in nature, with the amount of data needing storage experiencing exponential growth. New solutions that can take advantage of ubiquitous IP networks and fast-deployment optical backbone networks are in high demand (Brocade 2001).

There are two approaches to SAN over IP currently under IETF consideration for standards status. One is the Internet Small Computer System Interface (iSCSI), which simply extends well-tested LAN-based SCSI to optical IP MAN and WAN. The other method, Fibre Channel over IP (FCIP), translates Fibre Channel control codes and data into IP packets for transmission between geographically distant Fibre Channel SANs.

20.4.2 iSCSI Approach

iSCSI is a new IP-based storage networking standard for linking data storage facilities, developed by IETF (Satran et al. 2002). By carrying SCSI commands over IP networks, iSCSI is used to facilitate data transfers over backbone IPs and to manage storage over long distances. Because of the ubiquity of IP networks, iSCSI can be used to transmit data over LANs or WANs and can enable location-independent data storage and retrieval.

20.4.2.1 iSCSI SAN Protocol Model iSCSI is the result of overlaying SCSI over TCP/IP, two widely deployed protocols, as shown in Fig. 20-8 (Satran et al. 2002). The top three layers, i.e., host application, OS, and SCSI, can be collectively viewed as the application layer in the OSI network reference model. These three layers are no different from network-attached storage (NAS) applications such as shared network data storage, where

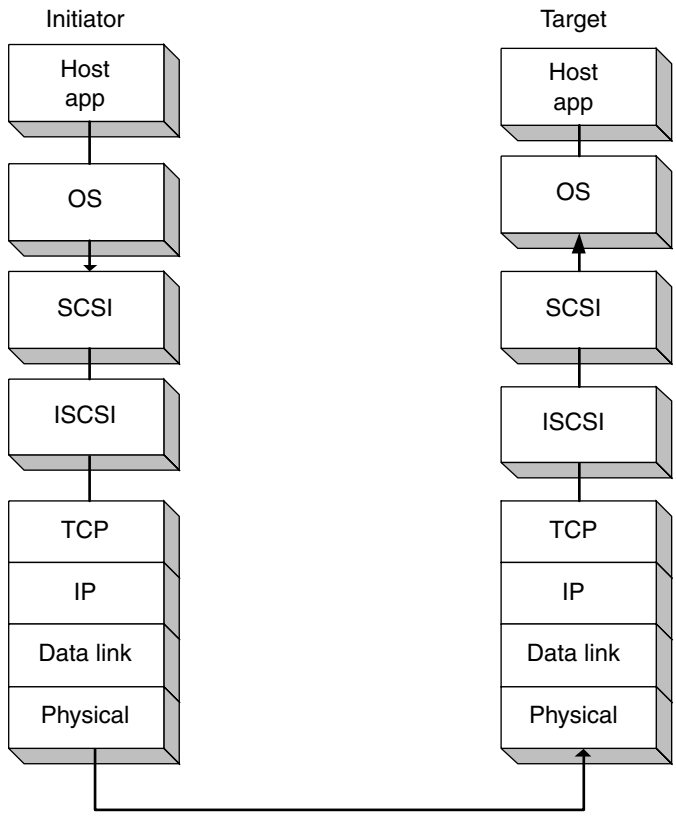


Figure 20-8
iSCSI over IP.

SCSI commands are issued. From the TCP layer on down, it is the same as an IP network. The middle layer sandwiched between the two, iSCSI, is the key component of the proposed iSCSI standard, which performs the following major functions:

- It encapsulates SCSI commands into a TCP packet.
- It provides an authenticated log-in mechanism for security.
- It performs a data integrity check.

An SCSI command is encapsulated in a TCP packet as shown in Fig. 20-9. The encapsulation is not as simple it may seem. For example, TCP/IP is byte-stream-oriented while SCSI data is command-oriented. A command is transmitted as a single unit. The iSCSI layer must manage the boundaries of the command using a buffering mechanism. While the SCSI commands assume the high bandwidth available and the commands

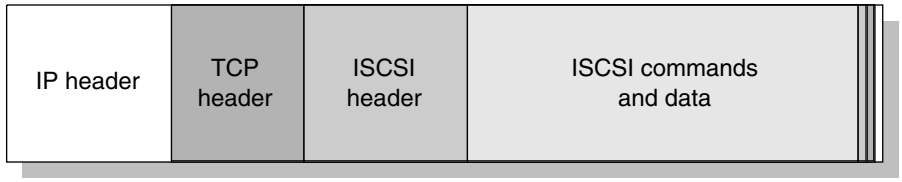


Figure 20-9
SCSI command encapsulation in a TCP packet.

can arrive as a unit, the congestion in IP networks certainly poses a challenge to the SCSI operation.

An iSCSI system that initiates an SCSI command, normally a server, is known as an *initiator,* and an iSCSI system that receives the command, such as a storage device, is known as a *target,* as shown in Fig. 20-8. Initiators initiate and set up an iSCSI session, which consists of a group of TCP connections that connect an initiator to a target. A session is defined by a session ID that has both an initiator and a target part. TCP connections can be added or removed from a session.

iSCSI supports ordered command delivery within a session via commands numbering. A command number is unique within a session. In addition to the ordered command delivery, the command numbering can also be used for flow control. Commands sent from an initiator must be acknowledged by a target, except for those marked for immediate delivery. The responses from a target to an initiator are numbered as well.

Each iSCSI session goes through two phases: log-in and normal operation or full-feature. The operations that happen during the iSCSI log-in phase include TCP connection establishment, authentication of the parties, session parameter negotiation, and security association protocol initiation. The full-feature phase immediately following the log-in phase allows the full set of SCSI commands to be exchanged between the initiator and the target.

20.4.2.2 iSCSI System Configuration and Operations An iSCSI can connect multiple storage devices that are located far apart via an IP network. A simple iSCSI system configuration is depicted in Fig. 20-10 (Net Convergence 2001), where an iSCSI server is located inside an Ethernet LAN and has established a session with a storage device located at a remote Ethernet LAN. The iSCSI server can communicate to any other iSCSI SAN storage device via an IP network.

The example of a remote file backup application will help illustrate how an iSCSI works. When an end user or application at the iSCSI

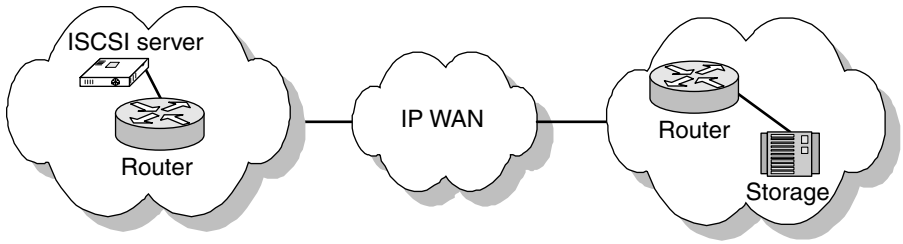


Figure 20-10
An iSCSI system configuration.

server sends a request (e.g., retrieving a huge file), the operating system generates the appropriate SCSI commands and data request, which then go through encapsulation and, if necessary, encryption procedures. A packet header at the TCP layer is added before the resulting IP packets are transmitted over an Ethernet connection to the LAN router. The IP packet is sent from the LAN router to the backbone router and then routed to the terminating LAN router. When the packet is received at the iSCSI storage device, it is decrypted (if it was encrypted before transmission), and disassembled. The SCSI commands and request are separated out. The extracted SCSI commands are sent on to the SCSI controller, and from there to the SCSI storage device. Because iSCSI is bidirectional, the protocol can also be used to return data in response to the original request.

The iSCSI approach to SAN over IP marries two of the most widely deployed protocols, SCSI and TCP/IP, with a focus on providing remote access to the large installed base of SCSI storage devices via fast backbone IP networks to support applications like remote data backup. The iSCSI specification is currently in the final stage before being adopted as an IETF standard (Rajagopal et al. 2001) and has support of influential IP network equipment vendors like Cisco and SCSI storage device vendors like IBM.

20.4.3 FCIP

A second approach to SAN over IP is known as *Fibre Channel over IP.* This encapsulates Fibre Channel frames in IP packets for transport over Gigabit Ethernet, 10 Gigabit Ethernet, SONET, or ATM MAN or WAN. This approach is also known as *Fibre Channel tunneling* or *storage tunneling,* because the broadband IP network essentially provides a tunnel to transfer FC frames between SANs across wide-area networks.

20.4.3.1 FCIP Protocol Model An FCIP-based network in general has three parts: an IP network to provide transport service, two FC systems connected through the IP network, and an FCIP interworking module responsible for the interworking between the IP network and the FC system. The FCIP interworking module consists of two parts, as shown in Fig. 20-11: an FCIP link endpoint and an FCIP control and service module (Rajagopal et al. 2001; San Valley 2001).

FCIP LINK AND LINK ENDPOINT An FCIP link, like a virtual circuit, is a connection provided by the FCIP protocol to connect two geographically and logically separate Fibre Channel systems to form a single FC network, using an IP network as the transport network. An FCIP link contains one or more TCP connections.

The FCIP link endpoint at the end of an FCIP link is a key component with the intelligence for the interworking between an FC network and an IP network. An FCIP link endpoint is created at the time a TCP connection is established. Contained in the FCIP link endpoint is an FCIP data engine. The FCIP data engine is a transparent data translation point between a Fibre Channel entity and an IP network entity. It is a

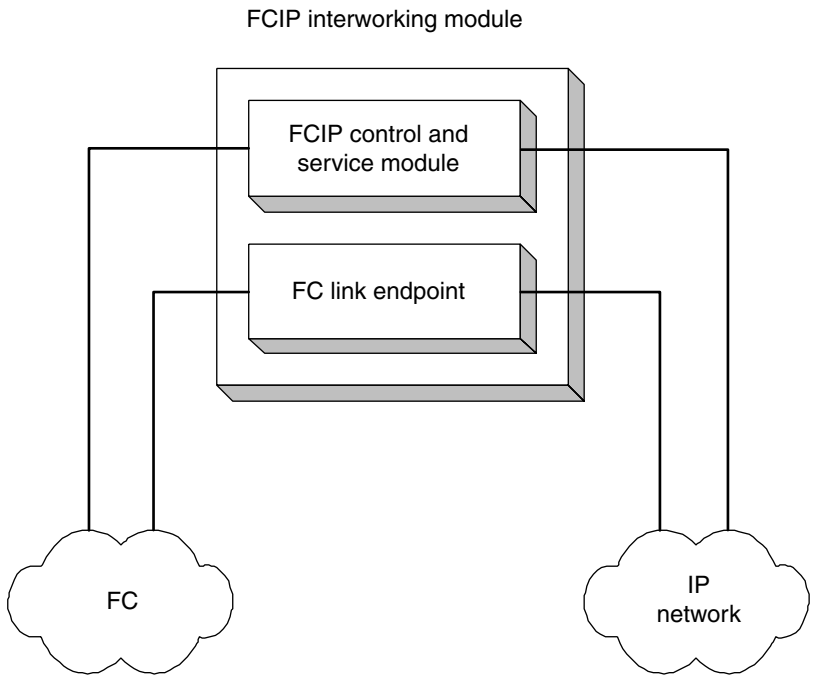


Figure 20-11
A functional view of FCIP architecture.

pair of FCIP link endpoints that communicate to each other over one or more TCP connections to join two isolated FC systems to form a single FC system. The data engine performs two main functions:

- It encapsulates FC frames into TCP packets to be sent to the other FC system over an IP network at the sending end and extracts FC frames from the received TCP packets at the receiving end.
- It detects data transmission errors and performs simple error recovery.

FCIP CONTROL AND SERVICE MODULE This module is mainly responsible for mapping between IP network features and FC network features such as addressing, flow control, and security.

Flow control is one important aspect of tunneling an FC frame through an IP network. Both the FC and IP networks have their own flow control mechanisms, which work independently in their own domains. An FC network uses a credit-based flow control while TCP uses a window-based mechanism. FCIP flow control management is needed for the following two scenarios: (1) Line speeds mismatch between the FC and IP interfaces and either the FC network or the IP network encounters congestion; (2) mapping the constraints the flow control method of one network imposes to the constraints of another is not a trivial task and requires substantial work in future.

Security is another concern when overlaying an FC system over a public IP network. The attack may take one of the several forms:

- An unauthorized FC element gains access to the resources of or monitors and manipulates Fibre Channel traffic flowing over physical media used by the IP network.
- Unauthorized FCIP encapsulated frames are injected into a TCP connection.
- The payload of encapsulated FCIP frames may be altered.
- Unauthorized agents masquerade as valid FCIP elements and interrupt normal operations.

Prevention or detection of such attacks is a generic security issue not unique to FCIP. The draft FCIP standard requires that FCIP entities implement IP security mechanisms such as IPSec and the Internet Key Exchange (IKE) protocol. IPSec in tunnel mode provides data integrity and confidentiality. For more details on IPSec, see Chap. 21 on VPN. The IKE protocol provides mechanisms to establish and maintain a secure connection between two FCIP peers of an FCIP link.

20.4.3.2 FCIP System and Operations An FCIP SAN system generally requires an FCIP gateway that performs the encapsulation of frames into the IP packet and mapping between the FC network and the IP network parameters. An FCIP gateway connects an FC network to an IP network, as shown in Fig. 20-12. One FCIP gateway bridges an IP network and a SAN of FC arbitrated loop configuration. Another FCIP gateway connects a hub-based FC SAN to an IP backbone network. Together, the FCIP gateways connect together two geographically distributed FC SANs via an IP backbone network.

An example of a user sending backup data using SAN over IP will illustrate the basic steps of the FCIP SAN system as illustrated in Fig. 20-12. Assume that a user at the FC SAN server shown on the left-hand side of Fig. 20-12 sends a stream of FC frames to the left-hand side FCIP gateway. Here are the basic steps of transferring the FC frames over the IP network to a server of another FC SAN shown on the right-hand side of the figure across a wide-area IP network:

Step 1: Endpoint Initialization. The FCIP service and control module generates an endpoint link ID and performs tasks such as encryption of data as required by the IP network. The QoS parameters supported by the IP network are defined and set at this point.

Step 2: Data encapsulation. Then each FC frame is encapsulated into one or more TCP packets at the FCIP link endpoint.

Step 3: Packet forwarding. The FCIP gateway determines the route and outgoing port as a normal IP router would do. Then the FCIP link endpoint forwards IP packets like any other IP packets. The intermediate IP network nodes are not aware of the FC frames embedded in the packet.

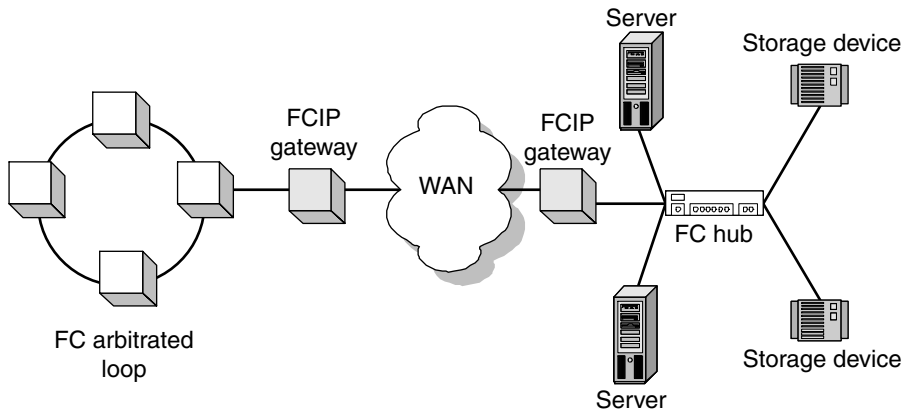


Figure 20-12
An FCIP SAN system.

Step 4: FC frame extraction. When the packet arrives at the terminating FCIP gateway shown on the right-hand side of Fig. 20-12, the link layer and IP layer headers are peeled off and the payload data is passed onto the TCP layer. The FC frame is identified at this layer, and is extracted and passed to the control and service module of the gateway.

Step 5: FC frame processing. The control and service module processes the extracted frames, performing any mapping, and then forwards the frames to the FC hub, which routes them to the appropriate destination server without knowing the frames came off an IP network.

In summary, FCIP is intended to bridge geographically distributed SANs across WANs to form a single SAN, while iSCSI presents the potential to enable wide-area access storage networks for remote access and backup applications. They complement each other in providing high-speed universal access to data that is stored in a single center or in virtually distributed storage networks.

REVIEW QUESTIONS

1. Data storage solutions have gone through an evolution of three stages, with storage area networks being the most recent. Describe the characteristics of each stage.
2. The storage area network solution promotes data sharing by sharing data storage. Describe how this solution is related to the Internet-based e-business model and discuss the motivations behind the sharing of network storage.
3. What management protocol is most commonly used for SANs?
4. Fibre Channel defines a five-layer network model. Describe the mapping between the FC layers and the OSI network layer.
5. What are the two flow control mechanisms of Fibre Channel defined at layer FC-2? Which class of service is each flow control mechanism used with?
6. Describe the sequence and exchange concept of the FC-2 layer. What is the main purpose of sequence and exchange mechanisms?
7. What are the three topologies Fibre Channel systems support? Discuss how interconnecting devices such as a hub and a fibre channel switch define a particular topology.
8. Describe the arbitrated loop configuration and explain the reasons behind its large-scale deployment.

9. What are the management functions and protocol that have been adopted for FC SANs?
10. What are the reasons for extending SAN over IP networks, and what are the two main approaches for doing so?
11. What are the two basic approaches to extending SAN over IP networks. In what way do they complement each other? In what way do they compete against each other?
12. Describe the FCIP approach to SAN over IP networks and discuss the main advantages of this approach.
13. Describe the iSCSI approach to SAN over IP networks and discuss the main advantages of this approach.
14. Briefly discuss the challenges involved in encapsulating SCSI commands inside a TCP packet.

REFERENCES

ANSI. 1996a. "Fibre Channel—Arbitrated Loop (FC-AL)." ANSI X3.272. Web site: www.ansi.org.

ANSI. 1996b. "Fibre Channel—Fabric Generic Requirements (FC-FG)." ANSI X3.289. Web site: www.ansi.org.

ANSI. 1996c. "SCSI-3 Fibre Channel Protocol (FCP)." ANSI X3.269. Web site: www.ansi.org.

ANSI. 1997. "Fibre Channel 2nd Generation (FC-PH-2) (formerly FC-EP)." ANSI X3.287. Web site: www.ansi.org.

ANSI. 1999. "Fibre Channel Physical and Signaling Interface (FC-PH)." ANSI X3.230. Web site: www.ansi.org.

Brocade. 2001. "Connecting SANs over metropolitan and Wide Area Networks." White paper. Web site: www.brocade.com.

IEEE. 1994. "Reference Model for Open Storage Systems Interconnection—Mass Storage System Reference Model Version 5." IEEE Storage System Standards Working Group. Web site: www.sswg.org/public_documents.html.

InterOperability Lab. 1998. "Fibre Channel—Tutorials and Resources." University of New Hampshire document. Web site: www.iol.unh.edu/training/fc/fc_tutorial.html.

Khattar, R., Murphy M., et al. 1999. *Introduction to Storage Area Network*. IBM Redbook. Web site: www.redbooks.ibm.com.

Net Convergence Inc. 2001. "Introduction to iSCSI." White paper. Web site: www.netconvergence.com.

Rajagopal, M., Bhagwat, R. et al. 2001. "Fibre Channel over TCP/IP (FCIP)." IETF-IPS draft document. Web site: www.ietf.org.

San Valley. 2001. "Interconnecting Fibre Channel SANs over Optical and IP Infrastructures." White paper. Web site: www.sanvalley.com.

Satran, J., Smith, D., et al. 2002. "iSCSI." IETF draft document. IETF-ips-iSCSI-09. Web site: www.ietf.org.

Tang, D. 1997. "Storage Area Networking." White paper. Web site: www.gadzoox.com.

CHAPTER 21

Packet Broadband VPN

21.1 Introduction

A virtual private network is a network built on shared public networking facilities to deliver services that, in appearance, are as reliable and secure as the services provided by dedicated private networks. VPN has emerged in recent years as one of the data network services with great revenue potential.

21.1.1 Brief History of VPN

The driving force behind the development of VPN technologies has been the desire for economically efficient private networks to connect the geographically distributed locations of many organizations in such a way that the organizations can have control over the services provided over their private networks. Mirroring the evolution of communications networks, VPN has evolved from pure voice services to data services to integrated voice and data services.

In regard to voice services, VPN had its origin in the telephone volume discount outbound calling plan used by corporate customers for calls between different branches. Then it evolved into the leased line service, where a leased telephone line connects two branch locations of a corporation. These are point-to-point lines that can be in two-wire or four-wire configurations. Eventually leased lines were used to connect PBXs at various locations. These dedicated lines between the distributed locations of organizations effectively created private networks. Furthermore, sophisticated numbering and dialing plans have been developed for such private networks.

The next major development in private-network voice services was the centrex service. It is the responsibility of each organization using a leased-line private network to manage the services and maintain the PBX equipment, often at a substantial cost. The telephone companies came up with the idea of building private networks using public telephony networks, which would provide corporations with the same abilities as private networks but relieve them of the burden of managing the services and maintaining the equipment. The term *virtual private network* was born because private networks are no longer built on dedicated point-to-point lines. Instead, they are built on switched public telephony networks.

Data VPN evolved along a similar path. Dedicated, leased lines were used to build private data networks connecting the multiple locations of corporations. The technologies for building such private networks for data services included X.25, frame relay, and ATM. Then virtual private networks

run over public switched data networks were introduced to relieve companies of the burden of maintaining and managing their private networks.

Another driving force behind data VPN has been the new business models evolving around the Internet. As business models have shifted toward e-business on a global scale, not only big corporations need to connect multiple locations, but many businesses forming partner relationships on short- or long-term bases have the need for private data network services.

The next step is the development of VPN on top of broadband IP networks. In recent years, a tremendous amount of effort has been poured into the development of broadband IP-based VPN that can carry both data and voice services and can be built on a moment's notice. A set of protocols has been developed for the broadband VPN that includes Point-to-Point Tunneling Protocol (PPTP), Layer-2 Tunneling Protocol (L2TP), and MPLS.

21.1.2 Broadband VPN Components

A broadband VPN consists of two types of elements: a tunneling mechanism and a set of security mechanisms.

21.1.2.1 Tunneling Mechanisms Tunneling mechanisms or tunnels are a type of virtual connection between two endpoints that provide secure communications over an open public network like the Internet.

There are two general classes of tunnels: end-to-end and node-to-node. The end-to-end tunneling mechanism extends a tunnel from an end user's terminal like a PC to a server on a corporate LAN on the other end of the VPN. In this scenario, the VPN devices at the two ends of the connection establish the tunnels and handle the encryption and decryption of data passed between the two points. This type of tunnel is very useful for applications such as telecommuting and remote log-in by mobile users.

The second type of tunnel, node-to-node, extends tunnels to the edges of networks, as opposed to end-user devices. This type of tunnel is useful for interconnecting two distributed networks such as two LANs. In this configuration, authentication, encryption, and decryption of user data take place at the security gateway that resides at the edge of a LAN interfacing a public network like the Internet.

The tunneling mechanism is a VPN-specific technology that has become the focus of the research and development efforts for the past few years. Other functional components of the VPN largely come from general computing and networking technologies as described below.

21.1.2.2 Encryption and Decryption An integral component of a public network-based VPN is the encryption and decryption mechanism, which can be located at a security gateway or a client device such as a user PC. Encryption ensures the confidentiality of data as it travels over a public network such as the Internet. In addition, encryption is also used as a tool to achieve user authenticity and user data integrity. The encryption mechanism used in VPN is the same as those used in other networking scenarios where data encryption is required. Likewise, the key issues with regard to the encryption of VPN—distributing the encryption key to users and deciding the encryption key size—are the same issues involving encryption in general.

21.1.2.3 Access Control and Authentication Access control via authentication is another important piece of VPN. Authentication is used to identify users or user processes that attempt to gain access to a VPN system. It provides the proof of identity to the other communicating party in a transaction that involves two parties. This is vital for a VPN to be able to support applications such as online banking and other business transactions.

Again, VPN authentication is no different from authentication in general networking scenarios. The fundamental issue of authentication is that an entity to be authenticated must prove its identity by showing the knowledge of a secret. The primary means of authentication is a password. A secret password known to a single user or a small group of users is secure as long as it remains known only to the intended users.

Another part of authentication is the exchange of secret information between the communicating parties and the exchange of information with a trusted third party, if necessary. The methods for secret key information exchange can be divided into three categories: symmetric, asymmetric, and zero knowledge.

The symmetric authentication approach has the two communicating parties share a common secret authentication key or share a secret authentication key with a trusted third party. The secret key is used to encrypt data, like stamping the data with the sender's identity. The receiver party must have the key to decrypt the data and recover the data contents sent over the VPN tunnel.

Asymmetric authentication is based on public key algorithms. The sender proves its identity by demonstrating its knowledge of a secret signature key. Then the signature key is verified by anyone knowing the sender's public verification key. The public verification key is obtained via a trusted third party or other mutually agreed-upon means.

Zero-knowledge-based authentication assumes that the communicating parties know nothing about each other's secret keys. All validations and authentication are performed via a third party, a trusted certificate authority (CA), which provides each party with a private digital certificate and is responsible for authenticating each user.

21.1.3 VPN Applications

There are two general categories of VPN applications: remote access and interconnecting distributed LANs that form an intranet or extranet.

21.1.3.1 VPN for Remote Access A VPN allows a remote VPN client at home or in a hotel room to connect to a corporate network in a secure manner. This VPN application allows a remote or mobile user to telecommute. In the remote access architecture, as shown on the left-hand side of Fig. 21-2, a remote client first calls into a local ISP server connected to a public network such as the Internet. Then the VPN client establishes a connection to the VPN server on the company LAN. Once the connection is established, the remote client can communicate with the company LAN via a secure tunnel over the public network as easily as if the client resided on the LAN itself.

21.1.3.2 VPN for Intranet or Extranet Another application of VPN is to interconnect two or more geographically distributed corporate LANs to form an intranet or extranet. Figure 21-1 shows a VPN connecting two corporate LANs, a LAN at site A and a LAN at site B, to form a single corporate intranet that functions as a single LAN to the users on both LANs. In this configuration, the routers on the edges of both LANs have both a VPN client and server.

The extranet application is very similar to the intranet application. The only difference is that a VPN interconnects the LANs of different corporations that are in a business relationship and have the need to communicate on either a long-term or short-term basis.

21.2 PPTP, L2TP, and Layer-2 VPN

Layer-2 VPN technologies are the first generation of standard technologies to build VPNs over public IP networks. *Layer-2 VPN* refers to a set of technologies that allow a VPN to be built over layer-2 (the data link layer)

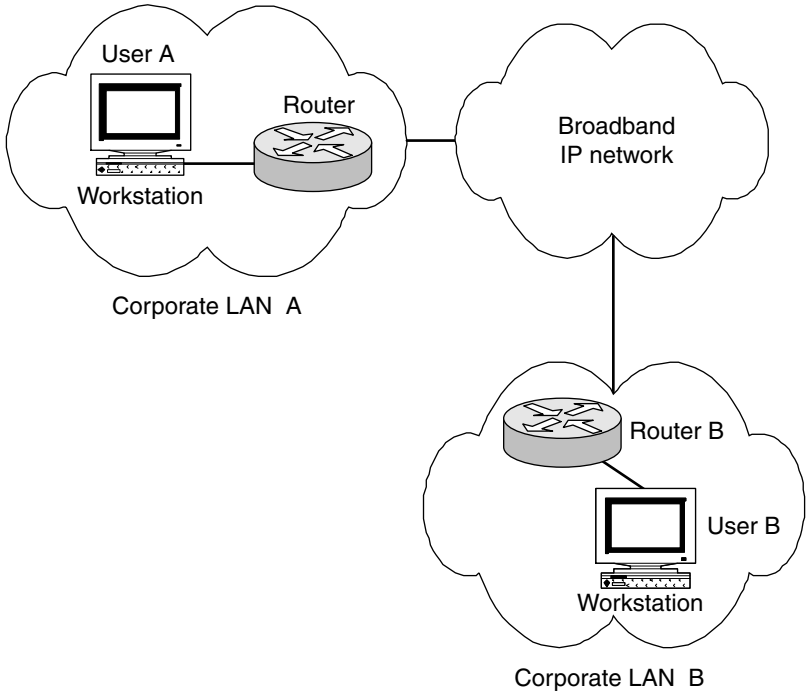


Figure 21-1
Illustration of VPN applications.

and that are independent of layer-3 technologies. In other words, the VPN data can be carried by any layer-3 protocols, such as IP and IPX.

There are three major layer-2 VPN technologies: Point-to-Point Tunneling Protocol (PPTP), Layer 2 Forwarding (L2F), and Layer-2 Tunneling Protocol (L2TP). PPTP is widely deployed in the IT industry using a large NT installed base with support from Microsoft. L2F is championed by data networking equipment vendors like Cisco and Nortel. L2TP is a combination of the two. This section will focus on PPTP and L2TP, because L2F is very similar to PPTP in architecture and system components.

All three layer-2 VPN protocols are based on PPP. PPP, as defined in IETF RFC1661 (Simpson, 1994), defines an encapsulation mechanism for transporting multiprotocol packets across layer-2 (L2) point-to-point links. For a detailed discussion of PPP, see Chap. 4.

21.2.1 PPTP-Based VPN

PPTP is built on top of PPP, a data link layer protocol that is used as a dial-up connection protocol for an end user to connect to the Internet.

PPTP, like PPP, operates at layer-2 (i.e., the data link layer) and supports the end user-to-network connection (Hamzel et al. 1999).

21.2.1.1 PPTP VPN System Configuration PPTP provides end-to-end VPN tunneling, from an end-user device to a PPTP server at the network side. As shown in Fig. 21-2, a PPTP VPN system consists of a PPTP server, a set of client devices, PPTP tunnels, and a set security mechanisms used over the PPTP tunnels and embedded in other components (Microsoft 1997).

PPTP SERVER A PPTP server, also called a *PPTP network server,* residing at the network side, is responsible for handling the server side of the PPTP protocol and is the terminating point of a PPTP tunnel. The responsibilities of a PPTP server include authenticating users, granting user session requests, decrypting user data, and regulating traffic flow to avoid congestion. The PPTP server is not physically restricted to any particular kind of network device; it can be on a LAN router, a LAN server computer, or a WAN device.

PPTP CLIENT A PPTP client is the PPTP tunnel originating point and is responsible for encrypting the user data, initiating a user session, and other PPTP client-side functions. A PPTP client generally resides at an end-user device such as a PC. For example, Microsoft NT supports the client side of the PPTP protocol. A PPTP tunnel extends from a user end device to the PPTP server at the network, as shown by the configuration in Fig. 21-2. An alternative configuration is to have the PPTP client reside at a network access server (NAS) if the user end device does not support the client side of the PPTP. In this case, the end user connects to the NAS via means such as a dial-up connection, an ISDN line, or an

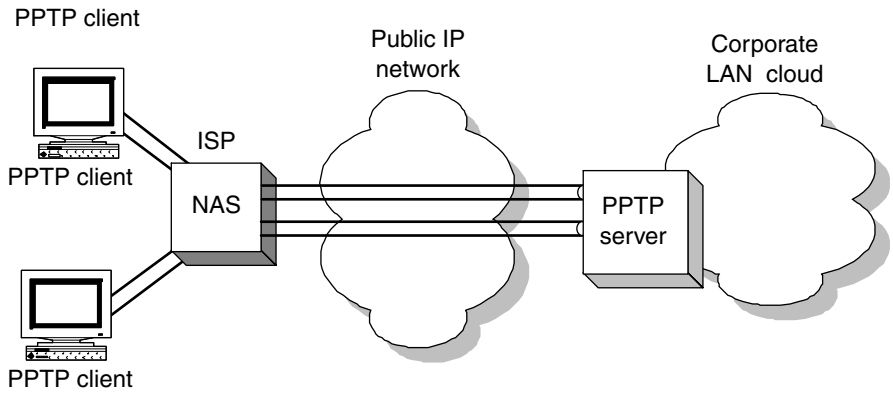


Figure 21-2
Configuration of a PPTP VPN system.

ADSL line, and then establishes a PPTP session. The user data is encapsulated in PPP packets and the PPP packets are then encapsulated in PPTP packets at the NAS.

The PPTP server and client functions may be combined, and the combined server and client may reside at one network device. This configuration is useful where a PPTP VPN is used to connect two geographically distributed corporate LANs.

PPTP TUNNELS At the center of PPTP technology is its tunneling mechanism. A PPTP tunnel is a virtual connection between a PPTP client and a PPTP server. The virtual connection is effectively a secure TCP connection with the data and PPTP control information encrypted.

PPTP uses three other protocols to establish PPTP tunnels: PPP to carry the user data, with the user data encapsulated inside PPP frames; Generic Routing Encapsulation (GRE) to encapsulate the PPP packets to be carried on the IP network; and the IP protocol to carry the GRE packets. As shown in Fig. 21-3, the user data is carried inside PPP packets, and a PPP packet is encapsulated inside an IP packet. Note that the data in the grayed layers is encrypted.

GRE is an IETF standard that defines a mechanism for encapsulating arbitrary types of packets within an arbitrary transport protocol. PPTP uses an extended version of GRE as defined in RFC 1702 (Hanks et al., 1994). GRE, which operates in the network layer, provides a way to encapsulate the layer-3 protocols such as IPX, AppleTalk, and DECnet for IP networks.

Multiple PPTP sessions can be multiplexed onto a single PPTP tunnel. A key inside the GRE header indicates which session a particular PPP

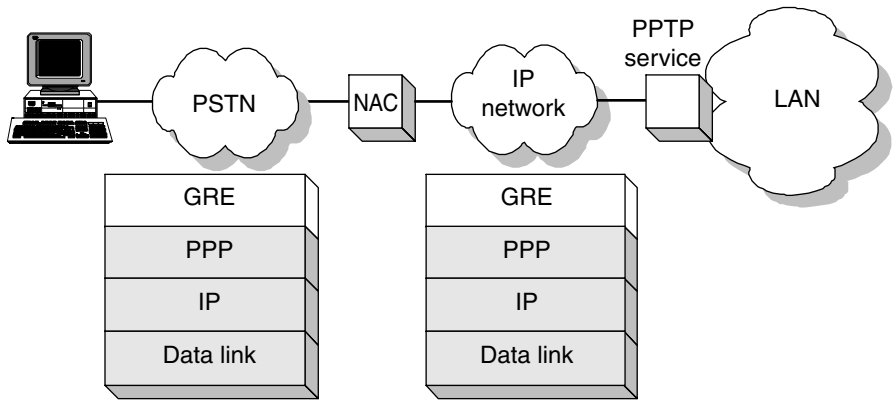


Figure 21-3
PPTP tunneling.

packet belongs to. This allows the PPP packets of different sessions to be multiplexed onto a single tunnel.

ENCRYPTION PPTP does not specify a particular encryption algorithm. In the Windows NT environment, Microsoft Point-to-Point Encryption (MPPE) is available, which can encrypt PPP packets on a user PC, so that the data is secure from the user PC to the PPTP remote-access server.

AUTHENTICATION PPTP uses PPP's Password Authentication Protocol (PAP) and Challenge Handshake Authentication Protocol (CHAP) as its authentication mechanism. A variant of CHAP developed by Microsoft, known as MS-CHAP, is also widely implemented in user devices. Full authentication and accounting of each connection may be done through a RADIUS client or locally.

21.2.1.2 PPTP Protocol Operations PPTP protocol operations consist of two phases: session initiation and data transmission, which comes after the session initiation phase.

PPTP tunnel and session initiation phase. This consists of two steps. First, a PPP client uses the PPP protocol to establish a connection to an ISP NAS via a phone line or ISDN line. The user data is also encrypted with a PPP encryption mechanism. Next, the client uses the established PPP connection to establish a PPTP tunnel to connect the PPTP client to the PPTP server; this is called *PPTP control connection establishment.* The client establishes a TCP connection to the PPTP server to exchange a set of PPTP control messages. The control messages are transmitted in control packets in a TCP datagram, which contains a PPP header, a TCP header, a PPTP control message, and an appropriate trailer.

PPTP data transmission. Once a PPTP tunnel and a user session are established, user IP data packets are encapsulated in GRE packets, which in turn are encapsulated in PPP packets that are transmitted between the client and the PPTP server. At the PPTP server, the GRE header is stripped and the PPP data block is decrypted before the data is sent to the destination node on the LAN.

21.2.2 L2TP VPN

The Layer-2 Tunneling Protocol is an extension of the widely deployed Point-to-Point Tunneling Protocol, with some features from L2F, another tunneling protocol supported by data equipment vendors like Cisco.

L2TP shares the same basic framework with PPTP, and uses PPP as the carrier protocol to carry user data. However, in comparison, L2TP is distinguished from its predecessors by the following features:

- It supports a wider range of layer-3 protocols, including non-IP protocols such as IPX and AppleTalk, as well as any WAN backbone technology, including frame relay, ATM, X.25, and SONET.
- It has a simplified tunnel setup process with UDP for control messaging instead of a TCP connection.
- It has enhanced security features that incorporate the IPSec security framework.

21.2.2.1 L2TP VPN System Configuration An L2TP VPN looks very much like a PPTP VPN in configuration, as shown in Fig. 21-4. An L2TP VPN consists of an L2TP access concentrator (LAC), an L2TP network server (LNS), an L2TP tunnel, and a set of security mechanisms (Townsley 1999).

L2TP ACCESS CONCENTRATOR An L2TP Access Concentrator is an L2TP tunnel endpoint that originates or terminates an L2TP call and is a peer to the L2TP network server. Generally, a LAC can be an enterprise router, a LAN gateway, or an ISP network access server at a point of presence. As shown in Fig. 21-4, a LAC, in addition to being a peer to the L2TP server, also interfaces a client remote system with a dial-up, ISDN, or ADSL connection. Although a LAC may connect to a remote user end system via a network access server, often the LAC and NAS functions are combined into one device.

L2TP NETWORK SERVER (LNS) An L2TP Network Server resides at the other side of an L2TP tunnel and acts as a peer to the LAC. An LNS is the logical termination point of a PPP tunnel that is tunneled from the remote system via the LAC. Physically, an LNS can be an edge router or switch at either an enterprise LAN or a service provider network.

L2TP TUNNEL An L2TP tunnel looks almost identical to a PPP tunnel, with a small difference in the encapsulation headers. Like a PPTP tunnel, an L2TP tunnel is a virtual connection between a remote end-user system and an LNS. L2TP data tunneling begins with a PPP payload. L2TP encapsulates the PPP payload with a PPP header and an L2TP header, and this results in an L2TP-encapsulated packet. Then in the L2TP protocol, a UDP packet is used to encapsulate the L2TP packet.

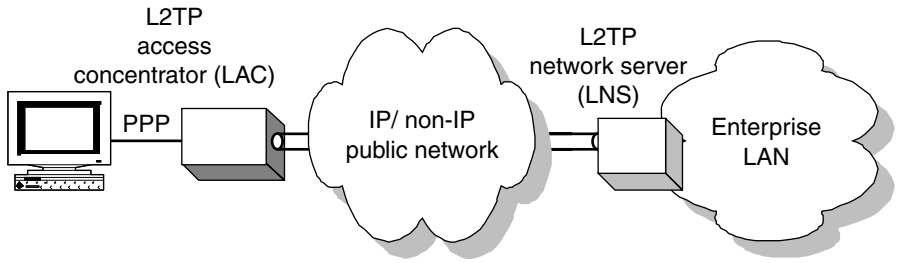


Figure 21-4
An L2TP VPN system configuration.

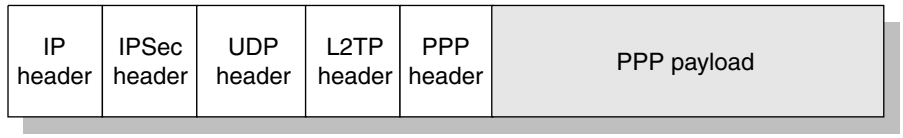


Figure 21-5
L2TP packet encapsulation.

Like PPTP, L2TP uses three protocols to establish an L2TP tunnel. It uses PPP to carry the user data with the user data encapsulated inside PPP frames. It uses L2TP packets to encapsulate the PPP packet and then uses the UDP packet to encapsulate the L2TP packet to be carried over an IP network. As shown in Fig. 21-5, the user data is carried inside PPP packets, and the PPP packets are encapsulated inside an L2TP packet, which in turn is encapsulated inside a UDP packet, which in turn is encapsulated inside an IP packet. Note that the grayed portion of packets is encrypted.

L2TP SECURITY The most important improvement of L2TP is the enhanced VPN security achieved through the use of the IPSec open security framework, which is described in detail in Sec. 21-3. The IPSec framework provides encryption, authentication, and access control mechanisms at the IP layer. As shown in Fig. 21-5, L2TP provides a means for the L2TP packet to be encapsulated in an IPSec packet.

21.2.2.2 L2TP Protocol Operations Assuming the scenario where a telecommuter initiates an L2TP VPN connection to a corporate intranet to work at home, the following steps illustrate the main thrust of the L2TP VPN operations:

1. The remote end-user system initiates a PPP connection via dial-up to the ISP's NAS, which also has the LAC function combined into the same device. The steps of setting up the PPP connection include PPP LCP negotiation and the exchange of CHAP challenge and response messages.

2. The ISP's NAS and LAC initiates an L2TP connection to the LNS at the corporate gateway router. An L2TN tunnel is setup between the end-user system at home and the corporate gateway. A UDP session, as opposed to a TCP session, is used to set up the L2TP tunnel. As part of the tunnel setup, IPSec security parameters such as encryption and authentication algorithms are agreed upon between the LAC and the LNS. Then LNS uses the authentication algorithm to authenticate the connection.
3. The remote end-user system uses the L2TP tunnel that has been established in the previous step to set up a PPP session. The user data is sent via the PPP frames encapsulated inside the L2TP tunnel. The PPP data is encrypted either at the remote end-user system, if it is equipped with the feature, or at the LAC.

This is a two-stage connection process, i.e., a user sets up a layer-2 connection to an access concentrator (e.g., modem bank, ADSL DSLAM, etc.), and the concentrator then tunnels individual PPP frames to the LNS. This two-stage process has the distinct advantage of allowing the actual processing of PPP packets to be divorced from the termination of the layer-2 circuit, which allow some users to avoid long-distance toll charges since the user connection terminates at a (local) circuit concentrator. From a user's perspective, there is no functional difference between the user directly terminating at the LNS or via an L2TP tunnel.

21.2.2.3 Comparisons Between PPTP and L2TP L2TP and PPTP have their unique characteristics as well as their similarities, which are summarized in Table 21-1. They have PPP in common as their user data carrier protocol. They are more suitable for dial-up VPN applications than other VPN alternatives such as MPLS-based VPN, although they can also be used for internetworking VPN or site-to-site connectivity applications. However, L2TP distinguishes itself by taking advantage of the newly available IPSec security framework, which much more greatly enhanced security features than its predecessor.

21.3 IPSec and Layer-3 VPN

IPSec, short for *IP Security,* is an open standard security framework that can be used with other tunneling protocols like L2TP to provide user

TABLE 21-1

A Simple Comparison Between PPTP and L2TP

	PPTP	L2TP
Carrier protocol for user data	PPP	PPP
Supported layer-2 and layer-3 protocols	IP, IPX	IP, ATM, frame relay, X.25
Tunnel setup messaging protocol	Over TCP	Over UDP
Encapsulation	IP over GRE over PPP	IP over L2TP over PPP
Security authentication and encryption	PPP, CHAP*	IPSec
Encryption	None	IPSec

*CHAP: challenge handshake authentication protocol.

authentication and data confidentiality. When used with L2TP, IPSec secures the data while L2TP provides the tunnel. IPSec can also be used alone to build a layer-3 VPN. In this case, IPSec provides both layer-3 tunneling and data security. This section describes the security components of IPSec and layer-3 tunneling for a layer-3 VPN.

One reason for developing the layer-3 VPN is the fact that the layer-2 VPNs built with frame relay or PPTP do not offer any encryption or authentication, and the integrity of packets that have been transmitted across an open network cannot be verified.

21.3.1 IPSec Overview

IP itself did not have security measures built into it when it was first conceived. IPSec is designed to solve the security problem with the current version of IPv4, and is mandatory for all IPv6 implementations. The typical security issues on an IP network include loss of confidentiality, loss of data integrity, and identity masquerade.

The first concern is lack of confidentiality. The Internet is wide open, and any unauthorized party can see the data passing through a particular node if the data is not encrypted. An untrusted party can read the source and destination addresses of each packet and also the payload contents of the IP packet. In regard to the loss of data integrity, classical IP does not have any guard against the modification of contents of an IP packet that is passing through a node.

Identify masquerade means an untrusted party may intercept the source and destination addresses of IP packets, masquerade as the source, and send out the data packets with malicious intention.

IPSec is an open standard framework for IP network security that encompasses many components. The main components include the following:

- IPSec specifies a set of protocols and defines an IPSec header that provides the infrastructure for carrying authentication and encapsulation information.
- It specifies a set of procedures for exchanging security keys, negotiating security parameters, and setting up a secure IP tunnel between two points across a public IP network.
- It specifies a set of standard cryptographic algorithms to provide user authentication, data integrity, and confidentiality of services over the framework.

Correspondingly, there are four categories of IETF IPSec standards that resulted from many years of effort at IETF, as shown in the Appendix of this chapter.

IPSec enables security at the edge of the network, and there are three major advantages associated with this approach. The most obvious is that the data is secured end-to-end across the network, from remote branch offices or subscribers all the way to a home gateway, a firewall, or a secure server at a corporate LAN. Another advantage is that IPSec allows enterprise customers to exercise total control over security implementation and policy. They do not have to depend on untrusted third parties to implement part of their security architecture. Finally, this solution scales well since the heavy processing requirements for encryption and authentication are distributed among many devices at the edge of the network.

21.3.2 IPSec Components

The key components of IPSec include an Authentication Header (AH) protocol, and Encapsulation Security Payload (ESP) protocol, and a Security Association (SA) protocol. AH provides the data security. ESP provides the layer-3 data encapsulation and encryption. SA provides a procedure for two parties to negotiate security parameters. In addition, IPSec specifies Internet Key Exchange as an authentication protocol to secure the IPSec tunnels between communicating parties.

21.3.2.1 Authentication Header and Authentication Header Protocol One IPSec header is the Authentication Header, which is inserted at the head of an IP datagram, as shown in Fig. 21-6. AH provides

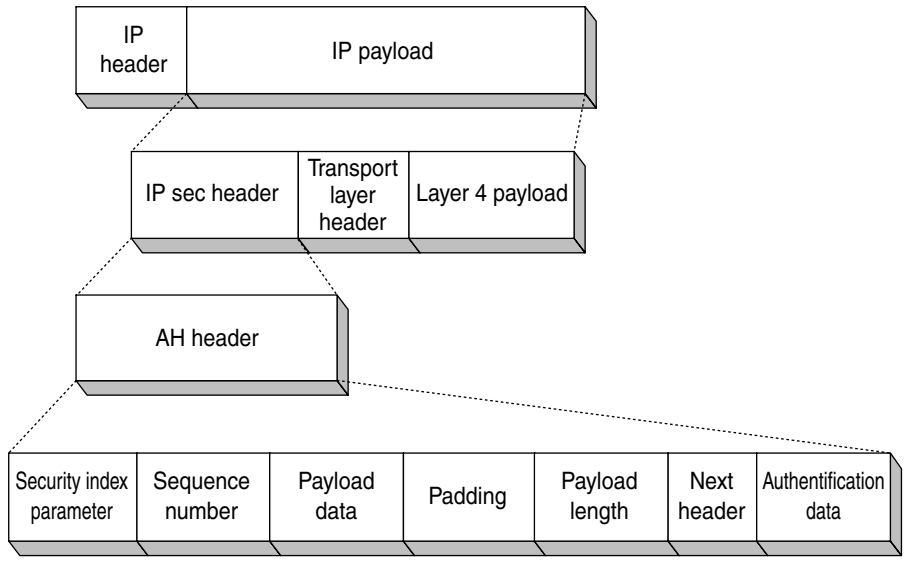


Figure 21-6
The IPSec headers.

integrity and authenticity for IP packet data, including the fields in the IP packet header. An AH protocol is defined to achieve the desired data authentication.

At a high level, the AH protocol works as follows: It first computes a cryptographic authentication function over the IP datagram using a secret authentication key in the computation. The receiver verifies the authentication data upon reception. The fields that must be changed in transition such as the time-to-live field of IPv4 or the hop limit of IPv6, are excluded from the authentication calculation for performance efficiency.

21.3.2.2 Encapsulating Security Payload The IPSec Encapsulating Security Payload protocol provides security services such as confidentiality, integrity, and authenticity to the IP datagram not provided by the AH.

The ESP protocol builds an ESP packet by encapsulation and encryption. It encapsulates either an entire IP datagram or only the upper-layer protocol data, encrypts the ESP contents, and then appends a new cleartext IP header to the newly encrypted encapsulation security payload to form an ESP packet. The source and destination addresses of the packet are encapsulated and encrypted along with rest of the IP datagram, making it impossible to breach the confidentiality of the data. The cleartext IP header is used to carry the protected data through the untrusted portion of a backbone IP network. At the receiving end, the encrypted ESP packet data is decrypted and reverted back to the original IP datagram.

An ESP packet consists of an ESP header, an ESP payload, and an ESP footer. The ESP packet format is designed to keep as much information as possible confidential by encapsulating the information in the ESP packet.

The ESP header consists of the Security Parameter Index (SPI) and sequence number fields. The SPI, a 32-bit field together with the IP destination address, identifies a security association (SA). An SA is an association between two communicating parties with a set of security parameters that are negotiated at the time the association is established. More details on SA will be provided shortly. The SPI is an arbitrary value chosen at the time the security association is established. The sequence number, a 32-bit field, indicates the packet's position in a sequence. This number is incremented every time a packet is sent. If replay protection is enabled, the sequence number is not allowed to cycle. When all the sequence numbers have been exhausted, a new SA must be established with a new security key and new start of sequence numbers.

The payload data, which is encrypted, is the actual data carried by the IP packet. The type of data is indicated by the next header field in the ESP footer.

The ESP footer has padding, next header, and authentication data fields. The padding field, containing 0 to 255 randomly generated bytes, may be used for a number of reasons, among them to conceal the actual length of the packet, or to ensure that the encrypted data terminates at a 4-byte or 8-byte data boundary. The next header field identifies the type of data, which is specified by the Internet Assigned Number Authority (IANA). The authentication data field is optional, and is included only if the authentication service is turned on for the SA. It contains an integrity check value generated by the authentication algorithm agreed upon at the time the SA is established. The length of the field depends on the algorithm used.

Either AH or ESP is used in the majority of applications, although they can be used independently or together. For both protocols, IPSec does not mandate a particular security algorithm to use, but instead provides an open framework for implementing industry-standard algorithms.

21.3.2.3 IPSec Security Association An IPSec Security Association is a relationship between two communicating nodes that must be established to negotiate and manage the IPSec security parameters. The parameters that must be agreed upon include the type of security service to be provided such as encryption, authentication, and integrity. The security algorithm to be used is also decided at SA setup time.

An IPSec SA is unidirectional, meaning that for each pair of communicating nodes A and B, there are at least two security connections, one from A to B and one from B to A. Each SA is uniquely identified by a randomly generated SPI combined with the destination address. Once an IPSec SA is negotiated and established between two end systems, the SA information is stored at each end system and is consulted each time a packet is sent.

21.3.2.4 IPSec Internet Key Exchange Internet Key Exchange is an integral part of the IPSec security framework that provides mechanisms to authenticate and secure an IPSec tunnel between two communicating systems. The IKE procedure involves two general steps for establishing a Security Association: authentication and key exchanges.

First, both nodes have to be authenticated to each other. IKE does not mandate a specific authentication method, and the communicating nodes can negotiate and agree on an authentication protocol.

Then the communicating parties must exchange a session key in order to encrypt the IKE tunnel. The Diffie-Hellman protocol (Rescorla 1999) is used to agree on a common session key. At the end of the key exchange, a secure tunnel with a Security Association has been established.

21.3.3 IPSec Tunneling

IPSec also provides a tunneling mechanism to build layer-3 VPN, in addition to providing security service to layer-2 tunneling protocols like L2TP. There are two types of IPSec tunneling: IPSec ESP and IPSec AH, based on the encapsulation header used. An IPSec tunnel, like a layer-2 tunnel, is a virtual data path on which encapsulated data travels between two communicating parties. IPSec tunneling operates at the network layer (layer-3), meaning the entire IP packet is encapsulated and secured for transfer via one of the IPSec security protocols (Goodman 2001).

21.3.3.1 IPSec ESP Tunneling IPSec ESP tunneling is achieved via a new tunneling header that encapsulates the ESP packet, as shown in the top part of Fig. 21-7. The original IP header placed before the transport (TCP/UDP) header carries the source and destination addresses. The tunnel header, combined with the ESP header, also known as the *outer IP header*, carries the addresses of the security gateways of both the source and the destination. The original IP packet plus the ESP trailer is encrypted to provide data confidentiality, while the ESP header and the encrypted

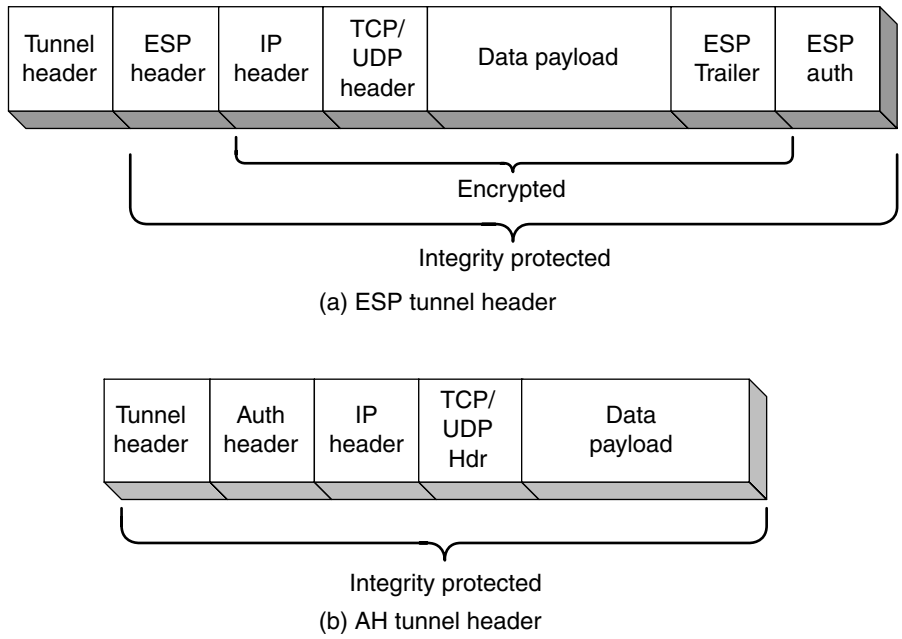


Figure 21-7
IPSec tunneling headers.

original IP header are integrity-protected. The information in the new IP header is sufficient to route the packet from the source to the destination security gateway, where the ultimate destination is extracted.

21.3.3.2 IPSec AH Tunneling An AH tunnel looks like an ESP tunnel except for the header. The entire new packet, including the tunnel header and the original IP packet, is protected for integrity, as shown in the bottom half of Fig. 21-7, while the tunnel header is not integrity-protected for ESP tunneling. No encryption at all is provided for the packet. ESP and AH can be combined to provide tunneling that includes both integrity for the entire packet and confidentiality for the original IP packet.

21.3.4 Two IPSec Operation Modes

IPSec supports two modes of operations: *transport* and *tunnel*. In the transport mode, only the IP payload is encrypted. This provides less stringent security and has the advantage of adding fewer security-related bytes to each IP packet. At any intermediate node, the source and destination IP addresses are not encrypted, and thus can be looked into. But the

higher-layer protocol headers such as TCP or UDP along with the payload data are encrypted. In this mode, encryption must be done at the two end host systems instead of at the edge routers. This mode of operation is more suitable for cases where the exposure of source and destination addresses is not an issue and only the confidentiality of data contents need to be ensured.

The tunnel mode is a more secure mode of operation. In this mode, the entire IP datagram is encrypted, and it becomes the payload of a new IP packet. The network edge router, rather than the end host system, encrypts the IP packets and forwards them along the IPSec tunnel. The destination router decrypts the original IP datagram and forwards it on to the end host system. One advantage of the tunnel mode is that the end systems do not need to be modified to implement IPSec. All that can be detected by any potential attacker is the tunnel endpoint.

In summary, the transport mode is less secure but involves less overhead and is less of a burden on the network, and must be implemented by both end host systems. In contrast, the tunnel mode provides more security, and is implemented at the edge routers without any change to the end host system, but it incurs more overhead overall.

21.4 MPLS VPN

One prime application of MPLS technology is VPN. This section provides a broad overview of MPLS-based VPN, starting with the reasons for developing MPLS VPN, and then describing the MPLS VPN system configuration, security mechanisms, and the considerations for MPLS VPN system deployment.

21.4.1 Reasons for Developing MPLS VPN

As the MPLS technology matures, MPLS VPN deployment starts to become attractive since it provides certain benefits that other VPN technologies may not have or cannot provide efficiently.

First and foremost, MPLS is a natural fit for VPN applications. The goal of a VPN is straightforward: to build a virtual network that behaves like an extension of a private corporate network on a shared network infrastructure. The result is that geographically distributed private LANs are connected together just like one single LAN. MPLS VPN is

simple to implement and economically efficient to build. Once an MPLS-enabled IP network is in place, building a VPN on top of that network is trivial because each label-switched path is a VPN tunnel by nature and is secure because of the way the LSP operates (Pepelnjak and Guichard 2002).

MPLS VPN offers the flexibility for implementation at layer-2 or layer-3, satisfying different customer needs. Layer-2 MPLS VPN provides a secure tunnel for various types of layer-2 traffic such as Ethernet, ATM, and frame relay, while layer-3 MPLS VPN offers an efficient, secure tunnel to connect dispersed IP networks. MPLS VPN provides the benefits of two worlds. MPLS VPN provides a platform for the service provider to offer many revenue-generating services such as intranet and extranet services over a single IP network.

21.4.2 Two Types of MPLS VPNs

There are two general approaches to building a VPN on an MPLS-enabled network: the overlay model and the peer model. The resulting VPNs are also known as *layer-2 VPN* and *layer-3 VPN,* respectively. Layer-3 MPLS VPN is built exclusively on IP networks, while layer-2 VPN uses MPLS LSPs to provide VPN services across different types of networks (e.g., ATM, frame relay, etc.).

21.4.2.1 Layer-3 MPLS VPN Layer-3 MPLS VPN, also known as the *peer VPN* model or *2547 VPN* (named after RFC 2547) had an early start in MPLS VPN. In essence, as stated in RFC 2547 (Rosen and Rekhter, 2001), "MPLS is used for forwarding packets over the backbone and BGP (Border Gateway Protocol) is used for distributing routes over the backbone" for setting up a VPN.

A layer-3 MPLS VPN network consists of multiple MPLS-enabled networks, as shown in Fig. 21-8. There are at least two customer networks that are connected by a public provider network. Each customer edge (CE) router is a peer of the provider edge (PE) router and provides the PE router with the route information of the customer edge private network. The PE router responsible for establishing a label-switched path across the public provider network is capable of storing a private routing table on a per-customer VPN connection basis.

A new VPN-IPv4 address is proposed in RFC 2547 (Rosen and Rekhter, 2001) to allow the routes of one MPLS VPN to be distinguished from the address of another VPN and to allow the different VPNs to use

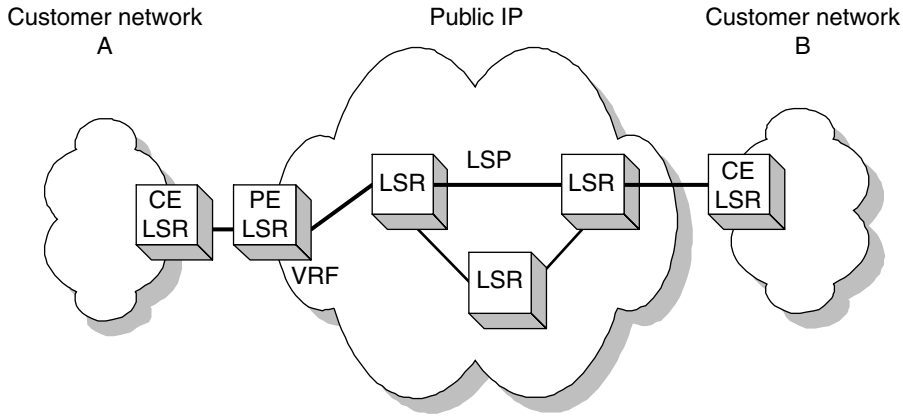


Figure 21-8
A layer-3 MPLS VPN.

overlapping addresses. This proposed VPN-IP address has a router distinguisher field with 64 bits that effectively become the customer identifier.

Another key component of the peer-model MPLS VPN is a VPN Routing and Forwarding (VRF) table located at the provider router. A VRF effectively specifies an MPLS-based virtual router that has a Routing and Forwarding table for each VPN to segregate and identify individual VPN customers. A VPN customer can be created or added by dynamically creating or destroying a VRF table entry. Each VRF can ensure a unique address space for a customer. Every VPN has a logically independent routing domain. This enhances the service provider's ability to offer a fully flexible virtual router service without physical per-VPN routers.

The layer-3 MPLS VPN also extends the Border Gateway Protocol to piggyback the label information on the traditional routing information fields. It propagates reachability information of a VPN and valid recipient information to other members of the same VPN via the BGP extension.

21.4.2.2 Layer-2 MPLS VPN Layer-2 MPLS VPN, also known as *overlay model VPN*, is based on point-to-point tunnels on MPLS backbone networks. The goal is to extend the layer-2 VPN services across IP backbone networks to other types of networks such as ATM, frame relay, and X.25. At a high level, a label-switched path functions like a layer-2 tunnel that can carry various forms of layer-2 traffic. As shown in Fig. 21-9, a layer-2 MPLS VPN connects an Ethernet LAN to an ATM-based LAN via a public IP using an LSP tunnel (Tomsu and Wieser 2002).

The key component of the layer-2 MPLS VPN is the MPLS tunnel, which can handle various forms of layer-2 traffic. The tunnel must have a point-to-point encapsulation mechanism for Ethernet, frame relay,

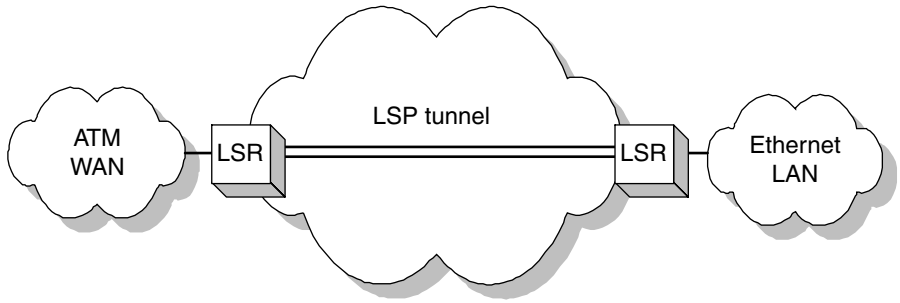


Figure 21-9
A layer-2 MPLS VPN.

ATM, TDM, and PPP/HDLC traffic. Although the standards on the encapsulation mechanism are still being finalized, it appears that a consensus is emerging around the Martini draft, a standard draft named after its author Luca Martini that defines an encapsulation mechanism (Martini et al., 2001).

21.4.3 MPLS VPN Security

The layer-3 MPLS VPN model in itself offers inherent data privacy by means of "data separation." That is, the data traffic belonging to different VPNs is always handled in a different context by using separate VRFs in the provider edge routers and by using the MPLS label-stacking functionality. The result of this data separation is that traffic from one VPN cannot be injected deliberately or accidentally into another VPN and that traffic belonging to one VPN cannot be deliberately or accidentally received in another VPN.

There are two main challenges as far as MPLS VPN security is concerned:

1. First, the portion of an end-to-end route before the first label-switched edge router is not secure. That is, the data is not secured between the customer edge router and the provider edge router if an LSP starts at the PE router.
2. Second, part of the network belongs to a third party, and a backbone network might consist of different parts owned by different parties. Moreover, a VPN on the same backbone network in some cases is not restrained to one single domain. Security across different domains and across networks of different service providers remains a challenge.

Various proposals have been made to address the above MPLS VPN security concerns. For example, one proposal is to integrate CE-CE IPsec with layer-3 MPLS VPNs through the use of a "tunnel endpoint discovery" technique. This would allow each CE to determine dynamically, for a given destination address, the proper remote tunnel endpoint for data encryption. Incorporating IPSec into provider network and using a multi-party Security Association to address the concern of multifragments of a VPN have also been suggested. Rather than encrypting traffic at the customer premises, security functions can be implemented at the edge of the backbone network. This would provide data privacy across the shared backbone network and would secure data entry into that network, making it difficult to transmit unauthorized traffic into the network.

Appendix. IPSec-Related IETF RFCs

Category	Title	RFC
IPSec general	Security architecture for the Internet Protocol	RFC 2401
	IP security document roadmap	RFC 2411
	ICMP security failure messages	RFC 2521
	Security model with tunnel-mode for NAT domain	RFC 2709
	Framework for IP-based virtual private network	RFC 2764
IPSec header	Encapsulation Security Payload	RFC 2406
	IP authentication header	RFC 2402
Procedures for security parameter negotiation and key exchanges	Internet IP security domain of interpretation for ISAKMP	RFC 2407
	Internet Security Association and Key Management Protocol (ISAKMP)	RFC 2408
	Internet Key Exchange	RFC 2409
	OAKLEY key determination protocol	RFC 2412
	PF_KEY key management API, v2	RFC 2367
	Photuris: session-key management protocol	RFC 2522
	Requirement for Kerberized Internet negotiation of keys	RFC 3129
Cryptographic algorithms	ESP DES-CBC cipher algorithm with explicit IV	RFC 2405
	ESP CBC-mode cipher algorithm	RFC 2451
	HMAC: key-hashing for message authentication	RFC 2104
	Test cases for HMAC-MD5 and HMAC-SHA-1	RFC 2202
	Use of HMAC-MD5-96 within ESP and AH	RFC 2403

Category	Title	RFC
Cryptographic algorithms	Use of HMAC-SHA-1-96 within ESP and AH	RFC 2404
	Use of HMAC-RIPEMD-160-96 within ESP and AH	RFC 2857
	UNULL encryption algorithm and its use with IPSec	RFC 2410
	IP authentication using keyed MD5	RFC 1828
	ESP DES-CBC transform	RFC 1829
	HMAC-MD5 IP authentication with replay prevention	RFC 2085
	IP Payload Compression protocol	RFC 2393
	IP payload compression using DEFLATE	RFC 2394
	IP payload compression using LZS	RFC 2395

Source: All references published by IETF. Web site: www.ietf.org.

REVIEW QUESTIONS

1. What does the term *virtual* in the virtual private network refer to? Describe the main differences between early telephony VPN and broadband VPN.
2. Describe the functional components of a VPN system and then briefly discuss the applications VPNs are intended to support.
3. Describe the characteristics common to all layer-2 VPN protocols. Discuss the advantages and disadvantages of layer-2 VPN tunneling.
4. Describe the tunneling mechanisms and security features such as encryption and authentication of PPTP VPN. Briefly describe the two phases of PPTP operations.
5. What were the reasons behind the development of L2TP? Describe L2TP tunneling and how it is different from PPTP tunneling.
6. Compare L2TP with PPTP, describing the features common to them and the main differences between them.
7. Describe the reasons for developing IPSec and the main components of the IPSec framework.
8. Authentication Headers and Encapsulating Security Payloads are two main IPSec security mechanisms and protocols for providing security to IP packet data. Describe the service each provides and why the majority of implementations only deploy one of them instead of both.

9. IPSec has two operation modes: transport and tunnel. Describe the main differences between the two and how each works.
10. IPSec, used independently, provides an alternative way for building a layer-3 VPN. Describe the tunneling mechanism the IPSec framework provides.
11. IPSec can also complement layer-2 VPN technologies such as L2TP. Describe at a high level how the combination of IPSec and L2TP works.
12. Compare MPLS layer-2 VPN with MPLS layer-3 VPN, describing the main differences between them. Then discuss the advantages and disadvantages of each.
13. Describe the main security mechanism that is built into the layer-3 MPLS VPN solution. Also, describe the general security concerns in regard to MPLS VPN.

REFERENCES

Goodman, J. 2001. "What Is IPSec Tunneling?" Microsoft white paper. Web site: www.microsoft.com.

Hamzeh, K., Pall, G., et al. 1999. "Point-to-Point Tunneling Protocol." IETF RFC 2637. Web site: www.ietf.org.

Hanks, S., Li, T., et al. 1994. "Generic Routing Encapsulation over Ipv4 Networks." IETF RFC 1702. Web site: www.ietf.org.

Microsoft. 1997. "Understanding Point-to-Point Tunneling Protocol." White paper. Web site: http://msdn.microsoft.com/library.

PepeInjak, I., and Guichard, J. 2002. *MPLS and VPN Architecture*. Indianapolis, IN: Cisco Press.

Rosen, E., and Rekhter, Y. 2001. "BGP/MPLS VPNs." IETF RFC 2547bis. Web site: www.ietf.org.

Simpson, W. 1994. "The Point-to-Point Protocol (PPP)." IETF RFC 1661. Web site: www.ietf.org.

Tomsu, P., and Wieser, G. 2002. *MPLS-Based VPNs: Designing Advanced Virtual Private Networks*. Englewood Cliffs, NJ: Prentice Hall PRT.

Townsley, W., Valencia, A. et al. 1999. "Layer Two Tunneling Protocol `L2TP.'" IETF RFC 2661. Web site: www.ietf.org.

Martini, L., et al. 2001. "Transport of Layer 2 Frames over MPLS." IETF draft document draft-martini-l2circuit-trans-mpls-08.txt. Web site: www.ietf.org.

Rescorla, E. 1999. "Diffie-Hellman Key Agreement Method." IETF RFC 2631. Web site: www.ietf.org.

CHAPTER 22

The H.323 System and Broadband Multimedia Applications

22.1 Introduction

H.323 in its broad sense is an umbrella recommendation by ITU-T that sets standards for multimedia communications over packet-based local area networks such as Ethernet, fast Ethernet, Token Ring, and other technologies that do not provide guaranteed QoS.

The series of ITU-T recommendations under the H.323 umbrella can be divided into the following four categories:

H.323 architecture and guideline

- *H.323.* Describes overall architecture, operations, and procedures of H.323 systems (ITU-T 2000c)
- *H.221.* Provides guidelines for implementing the H.323 series recommendations (ITU-T 2002a)

Control and call signaling protocols

- *H.225.0.* Specifies messages for call control including signaling, registration, and admissions, and packetization/synchronization of media streams (ITU-T 2000a)
- *H.245.* Specifies messages for opening and closing channels for media streams, and other commands, requests, and indications (ITU-T 2001)

Services supported on H.323 systems

- *H.450.x.* Define procedures and protocols for providing telephone-like services on H.323 systems (ITU-T 1998a)
- *H.235.* Defines a security framework used to provide authentication, encryption, and integrity for H.323 systems (ITU-T 2000d)
- *H.332.* Provides large-scale or loosely coupled conferencing service using H.323 systems (ITU-T 1998b)

Audio and video codec for H.323 systems

- *H.261.* Defines video codec for audiovisual services at P $\times$ 64 Kbps (ITU-T 1993b)
- *H.263.* Specifies a new video codec for video over POTS (ITU-T 1998c)
- *G.711.* Defines Audio codec: 3.1 KHz at 48, 56, and 64 Kbps, i.e., the normal telephony codec (ITU-T 1988b)
- *G.722.* Defines audio codec: 7 KHz at 48, 56, and 64 Kbps (ITU-T 1988a)

- *G.723.1.* Defines audio codec: for 5.3 and 6.3 Kbps modes (ITU-T 1996c)
- *G.728.* Defines audio codec: 3.1 KHz at 16 Kbps (ITU-T 1992)
- *G.729.* Defines audio codec: 8 kbps audio codec (ITU-T 1996a)

The H.323 recommendation has gone through three major revisions since its initial approval in 1996. H.323 v2 was approved in 1998 with enhancements focusing on the VoIP applications. The major changes of this version included the fast connect capability that uses a minimal call signaling message to allow quick establishment of a connection. H.323 v3 was approved in 1999, with changes focusing on interworking with PSTN and scaling up to large-scale network deployment. The latest version, H.323 v4 was approved in the end of 2001, with the changes focusing on scalability of the H.323 system and end-user services like call intrusion detection and service via HTTP.

22.1.1 H.323 System Overview

An H.323 system consists of a set of network elements and interfaces between the elements, as shown in Fig. 22-1. There are four types of network elements defined by H.323: H.323 terminals, gatekeepers, gateways, and multipoint control units (MCUs) (H.323 Forum 2001; Kumar et al. 2001).

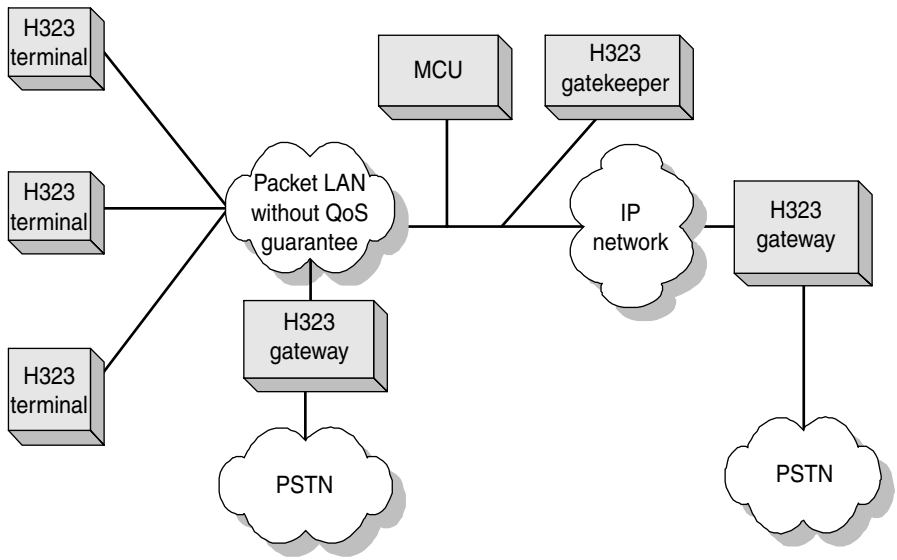


Figure 22-1
An overview of H.323 system configuration.

An H.323 gateway is the interface point between an H.323 system and a non-H.323 system such as a PSTN network. The main responsibilities of such an interface point include translations between H.323 conference endpoints and other terminal types, translations between transmission formats, and translations between communications protocols.

An H.323 gatekeeper is the intelligence center of an H.323 system that acts like a virtual switch. Among other things, it performs address translation from LAN aliases for terminals and gateways to IP addresses. The other important functions of a gatekeeper include bandwidth management and call routing.

An H.323 MCU supports multimedia conferences between three or more endpoints. An H.323 MCU consists of a multipoint controller (MC) and zero or more multipoint processors (MPs). The MC handles negotiations between all terminals to determine common video and audio capabilities and controls and allocates conference resources.

An H.323 terminal is an end-user device that supports and presents to the end user a multimedia service such as voice call, data services, and video conferencing.

An H.323 system is built on top of a packet LAN such as Ethernet, Token Ring, and fast Ethernet. The H.323 system components are connected to each other via a LAN, as shown in Fig. 22-1. While the gateway is connected to an outside PSTN network, the other three H.323 network elements generally function within the same LAN.

An H.323 system can be configured to support a variety of services, including desktop videoconferencing, collaborative computing, network gaming, distant multimedia learning, interactive shopping, Internet telephony and video telephony, and other multimedia applications.

21.1.2 H.323 Terminals

An H.323 terminal is a client endpoint on the LAN that provides real-time, two-way communications capabilities. An H.323 terminal is a complicated device with many functions, and it can be a PC, an IP phone, or a standalone device running H.323 multimedia services. All H.323 terminals must support audio (voice) communications, while the support for video and data communications is optional.

An H.323 terminal has, as shown in Fig. 22-2, six functional components to provide the multimedia services: a system control unit, an H.225.0 layer, a network interface, an audio codec unit, a video codec unit, and user data applications. The last two components are optional,

while the first four components provide the mandatory functions of H.323 terminals.

22.1.2.1 System Control Unit An H.323 terminal may have three types of signaling and control channels: control, call signaling, and registration, admission, and status (RAS) signaling. The system control unit of an H.323 terminal makes use of the three types of channels to perform the following control and signaling functions:

- The H.245 control function, which is responsible for media control at the terminal.
- The H.225.0 RAS control function, which is responsible for controlling RAS signaling. H.323 terminals need to communicate with a gatekeeper, if one is present, to perform registration, admissions, and bandwidth management. This is achieved by using H.225.0 RAS messages. A terminal has a RAS channel connecting to the gatekeeper, which is independent of the call signaling channel and the H.245 control channel.
- The H.225.0 call control function, which uses an H.225.0 call signaling message to establish a connection between two H.323 endpoints.

22.1.2.2 H.225.0 Layer The function of the H.225.0 layer is to manage the logical channels required of an H.323 terminal. For example, one

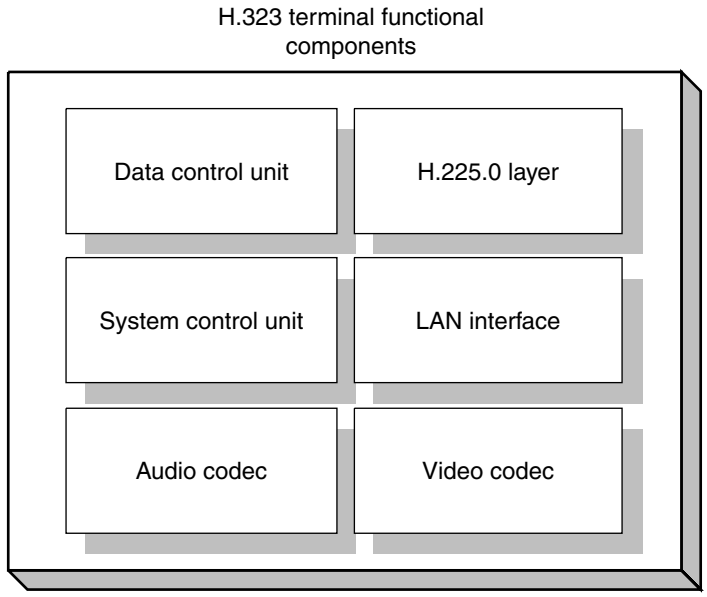


Figure 22-2 *The functional components of an H.323 terminal.*

such task is to manage channel number assignment and each logical channel, which is identified by a logical channel number (LCN) in the range of 0 to 65535 and is associated with a transport connection.

22.1.2.3 Audio Codec An H.323 terminal is equipped with an audio codec capable of mixing audio signals from multiple audio channels and handling various transmission bit rates seen in applications ranging from IP telephony to multipoint conferencing.

It is mandatory that all H.323 terminals have the ability to encode and decode speech. The primary codec standard is specified in G.711, while other codec standards as specified in G.722, G.728, G.729, and G.723.1, are optionally supported. The codec capability and specification are made known between two terminals using H.245 capability exchange messages. An H.323 terminal may optionally support more than one audio channel. In the case of multiple channels, the terminal is able to mix the channels to present a composite audio stream to the end user.

22.1.2.4 Video Codec Video-related capabilities are optional for H.323 terminals, as specified in H.323, version 2. The capabilities include video codec, video error correction, video conferencing, and the ability to operate in various modes.

The H.323 video codec unit that provides the ability to encode and decode video signals should support multiple video formats. While the Quarter Common Intermediate Format (QCIF) is the primary format, as specified in H.261 and H.263, other video codec and video formats are supported optionally as well. The video capabilities are exchanges between two H.323 terminals using H.245 capability exchange messages.

An H.323 terminal should support more than one video channel if video conferencing is supported. There are two modes of video operation: symmetric and asymmetric. In a symmetric mode, the transmitting and receiving operations use the same bit rate, picture format, and algorithms. In an asymmetric mode, while the decoder receives video signals at one bit rate, picture format, and algorithm defined during the initial capability exchange between the two terminals, the encoder is free to transmit at a different rate, with a different picture format, and with a different algorithm.

22.1.2.5 Data Control Unit In addition to video and audio controls for audio and video applications, an H.323 terminal can optionally have a

data control unit for one or more data channels for data applications such as email and World Wide Web browsing. Data operations standards as specified in T.120 (ITU-T 1996b) are the default basis for interoperability between an H.323 terminal and a non-H.323 terminal like H.324 or an H.320 terminal. The T.120 data channel may be bidirectional or unidirectional, depending on the type of application.

A data call is set up in the same way other calls are set up. That is, after the capability exchange between the terminals, a bidirectional logical channel is opened for a T.120 connection of a data call. In case the terminal intends to create a conference that includes audio, video, and T.120 data, it is necessary to establish an H.245 control channel before the T.120 data connection is made.

Like video and audio capabilities, the data capabilities are negotiated and made known between an H.323 terminal and other terminals via H.245 capability messages.

22.1.2.6 Packet Network Interface An H.323 terminal has an interface to the underlying packet network. The interface may vary according to the network technology (e.g., Ethernet, Token Ring, etc.) used, and therefore this interface is not standardized, but is left for vendor-specific implementation. However, function-wise, the interface should provide the following capabilities:

- Reliable end-to-end connection (such as TCP) for H.245 control signals, data channels, and call signaling channels.
- Unreliable end-to-end connection (such as UDP) for audio channels, video channels, and the RAS channel.
- Multiple operation modes such as duplex or simplex, unicast or multicast, depending on the application, the capabilities of the terminals, and the configuration of the network.

22.1.3 H.323 Gateway

An H.323 media gateway is the joining point between a packet network and a circuit-switched network like PSTN. In general, the gateway is the interface/gateway between IP and PSTN networks. It provides for real-time, two-way communication between H.323 terminals on an IP network and other types of terminals (e.g. analog PSTN phone, other ITU terminals). The primary function of the H.323 gateway is to interface

with the outside PSTN. The gateway is an optional component in an H.323 system and is not needed if the system does not interface with a PSTN. In essence, an H.323 gateway performs the mapping between an H.323 system and a PSTN at the following different levels:

- The translation between different transmission formats—for example, from H.225.0 and H.221 to the transmission format used at PSTN
- The translation between different communications protocols and protocol messages such as H.225 calling signaling and Q.931 signaling protocols and messages
- The translation between different video, audio, and data formats

As shown in Fig. 22-3, the gateway is equipped with the knowledge of two protocol stacks, a set of conversion functions, and a call control manager.

On the side facing the H.323 network, the gateway talks the following five protocols: (1) the Real-Time Transport Protocol (RTP) for transferring real-time messages on a LAN; (2) the Real-Time Transport Control Protocol (RTCP) for control and monitoring real-time data transfer; (3) the H.245 Control Signaling protocol for exchange terminal capabilities; (4) H.225.0 Call Signaling for call setup and release; and (5) H.225 RAS for registration with a gatekeeper.

On the side facing PSTN, the gateway talks the PSTN signaling protocols, which include the widely deployed SS7 (ITU-T 1993c) and ISDN signaling protocols, and receives the data in a TDM transmission format such as T1/DS1, T3/DS3, OC3, etc.

In between is a set of conversion functions that translates H.323 protocol messages to the PSTN protocol message or vice versa for call setup and release and converts between the media formats of different networks.

22.1.4 H.323 Gatekeeper

A gatekeeper is an optional component of an H.323 system and can be physically implemented in different ways: embedded in a gateway, distributed among terminals, or coexisting with MCUs, MCs, or even non-H.323 network devices.

At a functional level, an H.323 gatekeeper consists of three layers, as shown in Fig. 22-4 (Cisco 2001). At the core of the gatekeeper is its control and signaling functions, which include H.225 RAS, H.225 Call Signaling, and H.245 Control Signaling. Some of the control and signaling functions

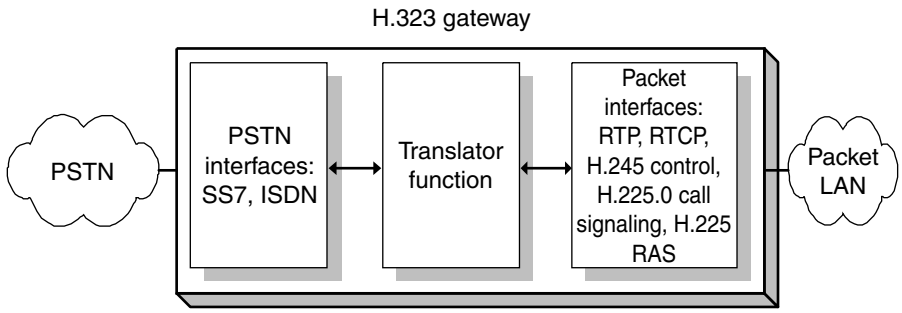


Figure 22-3
A configuration of an H.323 gateway.

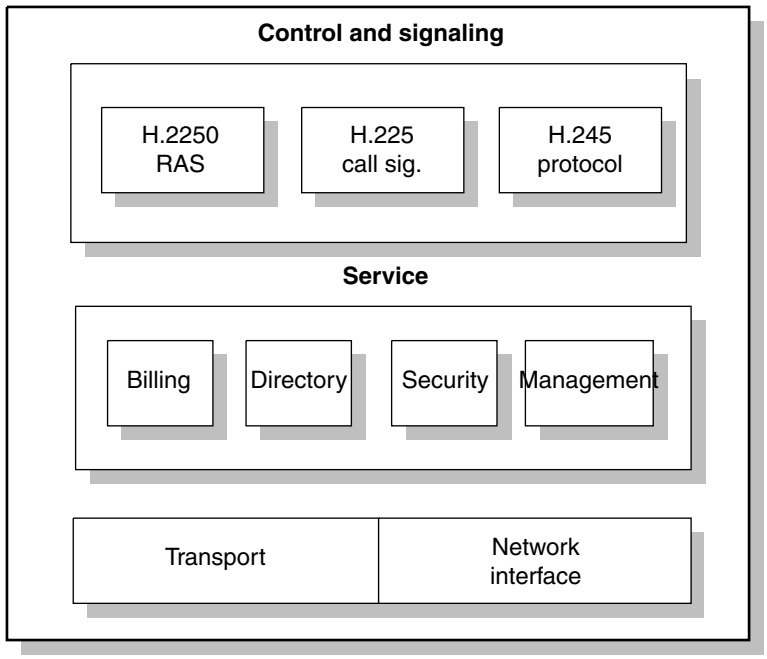


Figure 22-4
A configuration of an H.323 gatekeeper.

are mandatory while others are optional, as will be explained shortly. The middle functional layer provides a set of services that include billing, directory, security, and zone management. The bottom layer provides transport service and a network interface that includes TCP, RTP, and physical IP network interface cards.

H.323 endpoints are organized and grouped together into zones for administrative convenience. Each zone has one gatekeeper that is responsible for managing all the endpoints in the zone. The zone concept is similar to the domain name server (DNS) on a LAN that provides service

to a specific area. H.323 zones are set up normally according to geographic area or administrative division.

The primary function of a gatekeeper is the call control, involving the H.323 endpoints and providing the following mandatory services to all the endpoints in the zone:

- *Address translation.* This converts an alias address to a transport address using a translation table. An alias address, presenting a logical address such as *John@companyA.com,* must be translated into a transport address to locate the entity represented by the alias.
- *Admission control.* The gatekeeper is responsible for authorizing network access at an H.323 terminal. A network access request can be rejected for a variety of reasons—for example, call authorization failure. The admission control criteria are not standardized and are left to the choice of the implementers.
- *Bandwidth control.* The gatekeeper is responsible for controlling a number of H.323 terminals that may access the network at the same time to maintain the desired level of performance. The gatekeeper can reject a call from a terminal because of bandwidth limitations. A bandwidth request from an H.323 terminal can be granted or rejected on various grounds such as availability of bandwidth on the network, allocation policy, etc. The criteria for determining bandwidth availability are not standardized.

A gatekeeper may optionally provide several other services, most related to call control:

- *Call control signaling.* The gatekeeper can choose to serve as an intermediary to process and complete call signaling for a terminal. A gatekeeper may direct two terminals to connect the call signaling channel directly to each other without going through the gateway if the gateway is not equipped to handle the H.225.0 call control signals.
- *Call authorization.* The gatekeeper may authorize or reject a call from a terminal because of call authorization failure. The reasons for rejecting a call may include, but are not limited to, restricted access to or from a particular terminal or gateway, or restricted access during a particular period of time. The specific criteria for call rejection are not subject to standardization.
- *Call management.* The gatekeeper is responsible for managing the overall states of terminals involved in a call. It maintains a list of ongoing H.323 calls and the busy terminals that are involved in the

calls. This information is necessary for the gatekeeper to perform bandwidth management.

Other services that have been mentioned for the gatekeeper but not standardized include bandwidth reservation for terminals, directory services, and security services.

The long list of optional service of the gatekeeper indicates that the implementation and capabilities of H.323 gatekeepers can vary widely. It can be a very simple system with basic endpoint maintenance capabilities, or a full-blown system with a call agent, service features, and zone management functions.

22.1.5 Multipoint Control Unit

An H.323 multipoint control unit, an optional component of H.323 systems, supports multipoint conferencing between three or more endpoints. An MCU is required in the centralized configuration of a conference call where all the terminals send audio, video, data, and control streams to the MCU in a point-to-point connection. The MCU centrally manages the conference using H.245 control functions. An MCU is not required in a decentralized multipoint conference where the participating H.323 terminals multicast audio and video signals to each other without sending the data to an MCU.

An MCU consists of a multipoint controller and zero or more multipoint processors. Figure 22-5 shows one example of a configuration where the MC and MPs are encapsulated in one independent MCU.

22.1.5.1 Multipoint Controller An MC provides control functions to support conferences between multiple endpoints. One of its primary functions is to carry out the capability exchanges with the endpoints participating in a multipoint conference. The MC first sends a capability set to each endpoint that indicates the mode of operation such as centralized or decentralized. The MC may change the capability set already sent to the endpoints as a result of terminals leaving or joining the conference or for other reasons. The choice of conference mode is determined after a terminal has established a connection with the MCU using the H.245 signaling channel.

22.1.5.2 Multipoint Processor A multipoint processor receives audio, video, and/or data streams from terminals or other endpoints in a multipoint conference and processes the streams before returning them to the

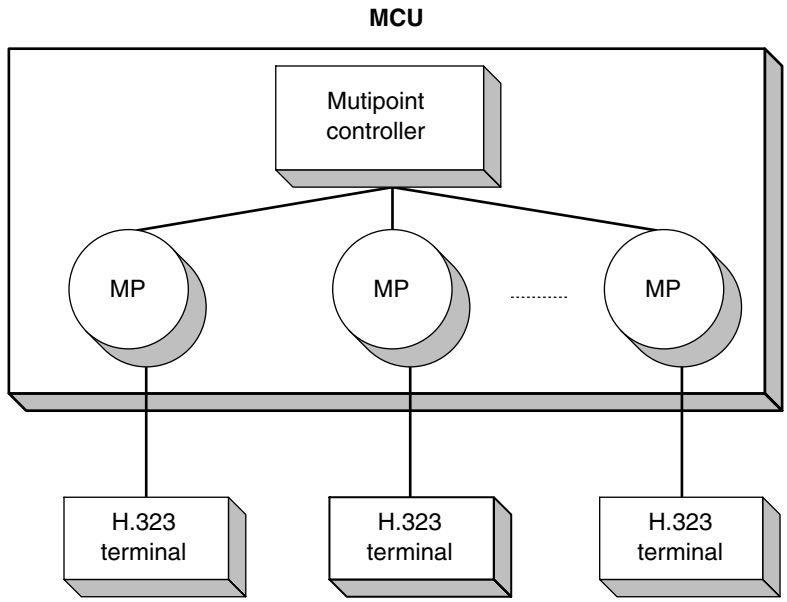


Figure 22-5
Configuration of an MCU.

endpoints. In the case of a video stream, an MP processes the video according to the video codec specified in ITU H.261 (ITU-T 1993b). In addition, an MP also performs video switching or video mixing. Video switching means selecting the video that the MP outputs to the different terminals. For example, in a video conference call, when a different speaker speaks up, the video needs to be switched to the new speaker. The switching can be triggered by the detection of an event such as a speaker change or an explicit H.245 control message. Video mixing is the process of formatting more than one video input stream into a single stream that the MP outputs to the terminals. For example, multiple picture sources can be combined into an array of pictures in the video output.

An MP should be able to perform audio mixing, switching, or a combination of the two. Audio mixing is a process of decoding the input audio streams to linear signals, performing a linear combination of the signals, and recording the result in the appropriate audio format. The MP may eliminate or attenuate some input signals to reduce noise.

22.1.6 H.323 Addressing Scheme

The H.323 addressing scheme provides a means to address an H.323 entity. There are three ways an entity can be addressed in an H.323 network:

network address, transport layer service access point (TSAP) identifier, and alias address.

22.1.6.1 Network Address A network address uniquely identifies an H.323 entity on the network. In certain cases, multiple entities may share a network address. For example, when an H.323 terminal has a colocated MCU, the MCU is not uniquely addressable to the outside. The network address is dependent on the network environment and network technology. For example, the address for an IP-based network is different from that for a PSTN network.

22.1.6.2 Transport-Layer Service Access Point Identifier For each network address, an H.323 entity may have several transport-layer service access point identifiers. The TSAP identifier allows multiplexing of several channels associated with the same network address.

Some TSAP identifiers are statically defined and well known to all endpoints in an H.323 system, such as the call signaling channel TSAP identifier, the RAS channel of each gatekeeper, and the multicast address called the *discovery multicast address*. Some TSAP identifiers are dynamically defined. For example, the TSAP identifiers for the H.245 control channels, audio channels, video channels, and data channels are assigned dynamically during a call.

22.1.6.3 Alias Address In addition to the unique network address and TSAP identifiers, an H.323 endpoint may have one or more alias addresses associated with it. An alias address may represent the endpoint or the conference associated with the endpoint. An alias address should be unique within a zone. However, a gatekeeper and multipoint controller cannot have alias addresses, because the alias address mainly allows the outside entity to address an endpoint within an H.323 system and the gatekeeper and MC are both not addressable by outside entities.

22.2 H.323 Protocol Stack

The H.323 protocol stack centers on the grayed two layers shown in Fig. 22-6, which are either specifically defined or adopted for H.323 systems: three application layer protocols and two real-time protocols. The application layer protocols are the H.225.0 Call Signaling protocol, H.225 RAS, and the H.245 Control Protocol that make use of IETF's Real-Time Protocol

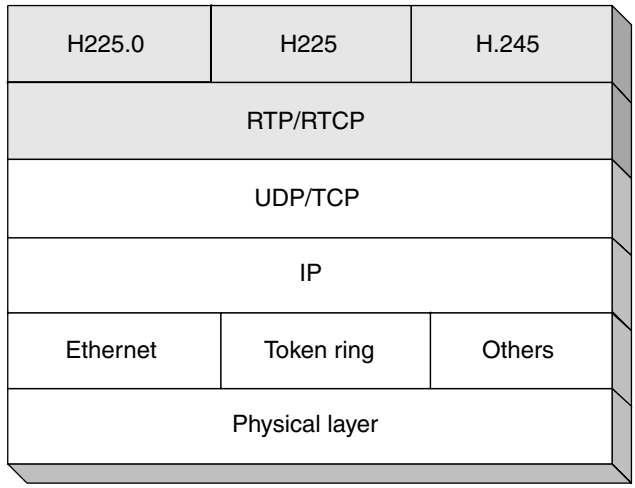


Figure 22-6
The H.323 protocol stack.

(RTP) and Real-Time Control Protocol (RTCP). Though RTP/RTCP is not defined exclusively for H.323 systems, H.323 systems adopted RTP/RTCP early on. The data link layers can be any of the LAN protocols such as Ethernet, Token Ring, and others that do not provide a QoS guarantee.

22.2.1 RTP and RTCP

IP networks were initially conceived with an emphasis on their ability to deliver messages in spite of network failure, little consideration was given to real-time responsiveness. RTP and RTCP are specifically designed to complement the classic IP networks and to support real-time-sensitive applications on IP networks.

22.2.1.1 Real-Time Protocol Real-Time Protocol (RTP), version 2, as defined in IETF RFC 1889 (Schulzrinne et al. 1996), is a real-time transport protocol that provides end-to-end delivery services to support real-time-sensitive applications such as interactive audio and video, VoIP, and multimedia applications.

The services RTP provides include payload type identification, sequence numbering, and time stamping. RTP operates at the transport layer but does not provide all of the functionality that is typically provided by a transport protocol. In practices, RTP typically runs on top of UDP to utilize its multiplexing and checksum services. Theoretically, other transport protocols like TCP can also carry RTP, but that is not a very common practice.

RTP is tailored for real-time applications. A key component of RTP is the timing information in the RTP header that is necessary to synchronize and display audio and video data and to determine whether packets have been lost or have arrived out of order. In addition, the header specifies the payload type, thus allowing multiple types of data and data compression. RTP provides support for auxiliary profile and payload format specifications. As an example, a payload format might specify what type of audio or video encoding is carried in the RTP packet.

An *RTP session* is an association between a pair of source and destination transport addresses (one network address plus a pair of ports for RTP and RTCP) set up by an application. In a multimedia session, each medium is carried in a separate RTP session. For example, audio and video would travel on separate RTP sessions and a recipient could selectively accept a particular medium.

RTP does not provide any mechanisms to ensure timely delivery, quality-of-service, or in-order packet delivery. It makes no assumptions about the underlying network. If an application requires such guarantees, RTP must be accompanied by other mechanisms such as RSVP and TCP to support resource reservation and to provide reliable service. The companion protocol RTCP does provide information on RTP session performance such as quality of service of the RTP session.

22.2.1.2 Real-Time Control Protocol RTP provides the services to support real time application while the Real-Time Control Protocol is a companion protocol that monitors the performance of RTP sessions. Specifically RTCP provides the following four services:

Performance monitoring. RTCP's primary function is to provide information to an application regarding the quality of data distribution. Each RTCP packet contains sender and/or receiver reports that report statistics useful to the application. These statistics include the number of packets sent, the number of packets lost, interarrival jitter, etc., that can be important for an application to know. For example, the sender may modify its transmission rate based on this feedback; receivers can determine whether problems are local, regional, or global; network managers may use information in the RTCP packets to evaluate the performance of their networks for RTP applications or multicast distribution.

RTP source identification. RTCP carries a transport-level identifier for an RTP source, known as the *canonical name* (CNAME). This CNAME is used to keep track of the participants in an RTP session. Receivers use

the CNAME to associate multiple data streams from a given participant in a set of related RTP sessions, e.g., to synchronize audio and video streams.

RTCP transmission interval adjustment. To prevent control traffic from overwhelming a network and to allow RTP to scale up to a large number of session participants, RTCP has the ability to limit the control traffic to at most 5 percent of the overall session traffic. This is achieved by adjusting the rate at which RTCP packets are sent as a function of the number of RTP session participants.

Session control data multicasting. As an optional function, RTCP can be used as a convenient method for multicasting a minimal amount of information to all session participants. For example, RTCP might multicast a notification from one session to the displays of all other sessions.

22.2.2 H.225.0 RAS

H.225.0 RAS (ITU-T 2000a) is a protocol that allows an H.323 terminal to register with a gatekeeper, request bandwidth, and report the status at the system startup time. The terminal first establishes a RAS channel to the gatekeeper prior to the establishment of the H.245 control channel and the H.255 call signaling channel. RAS sends connection request, admission request, and registration and deregistration messages:

- Generic connection request messages allow a terminal to discover a gatekeeper and register with it. The messages include gatekeeper request (GRQ), gatekeeper confirmation (GCF), and gatekeeper reject (GRJ).
- Admission request messages allow a gatekeeper to control admissions into the zone.
- Registration and deregistration messages allow a terminal to register with or leave a gatekeeper. A registration request (RRQ) message is for a terminal to register with a gatekeeper. A registration confirmation (RCF) message is for a gatekeeper to confirm the registration of a terminal. A registration reject (RRJ) message is for a gatekeeper to reject registration from a terminal. A unregistration request (URQ) message is for a terminal to deregistrate from the indicated gatekeeper. An unregistation confirmation (UCF) message is for a gatekeeper to confirm the deregistration of a terminal from the gatekeeper.

22.2.3 H.245 Control Protocol

H.245 defines a control protocol for an H.323 terminal and a gatekeeper to exchange information on multimedia capabilities (ITU-T 2001). The capabilities describe the audio, video, data, and multipoint conference abilities of the terminal. In addition, other items like communication mode (e.g., direct or indirect call routing) and initial bandwidth allocations are also negotiated between the gatekeeper and a terminal using the control channel. The control messages are defined as request, response, command, and indication. Examples of such messages include terminal capability, request mode, communication mode, conference request and response, and logical channel signaling.

In the absence of a gatekeeper or when choosing the direct communication mode, two H.323 terminals or endpoints use H.245 messages to exchange information about each other's capabilities and bandwidth information.

22.2.4 H.225.0 Call Signaling Protocol

The H.225.0 Call Signaling protocol is based on the ITU Q931 Call Signaling protocol (ITU-T, 1993a), and can be viewed as a subset of the latter. This is why Q931 is listed as a call signaling protocol instead of H.225 in some H.323 literature. This call signaling function uses H.225 call signaling messages to establish a connection between two H.323 endpoints. The H.255 protocol defines the message and procedures used to establish and take down a call, to request bandwidth changes, and to control the status during the call within an H.323 system.

The H.225 Call Control protocol is used either between a gatekeeper and an H.323 terminal or between two H.323 terminals. If the gatekeeper has call control abilities and it is decided between the gatekeeper and a registered terminal that call routing will come through the gatekeeper, then H.255 call signaling messages are exchanged between the gatekeeper and the terminal.

In the case of two endpoints communicating directly instead of through a gatekeeper, the calling party directly seeks out the called party instead of relying on a gatekeeper to locate the called party. Once the called party is found, a call signaling channel is established for call control and the channel remains in effect for the duration of the call. For example, the calling endpoint sends a call signaling SETUP message to the well-known signaling channel TSAP identifier of the called endpoint to

request the setup of a direct connection between the two endpoints. The called endpoint responds with a CALL PROCEEDING message to indicate that it has initiated the establishment of the requested connection. Optionally, the called endpoint may send an ALERTING message that more or less says, "I am aware of your request and am still working on it." The called endpoints send a CONNECT message that contains an H.245 control channel transport address to indicate that the requested connection has been accepted and the calling party can proceed to the next phase of call signaling.

22.2.5 H.323 System Internal Interfaces

It is the interfaces that connect together the components of an H.323 system and make them a functional system. Each interface includes a protocol stack and the associated messages, as described above.

There are three distinct sets of interfaces in a homogeneous H.323 system, and the corresponding protocol stacks are shown in Fig. 22-7. An H.323 system is homogeneous if it does not interface any outside network like a PSTN via an H.323 gateway. The interfaces include the gatekeeper-endpoint interface, the gatekeeper-gatekeeper interface, and the endpoint-endpoint interface.

22.2.6 Interfaces Between an H.323 Gateway and Other Network Components

H.323 gateways let an H.323 system go beyond a LAN and extend its reach to WANs. A gateway is an interconnecting point between an H.323 system and a circuit-switched PSTN, and has several interfaces to other network elements as shown in Fig. 22-8.

An H.323 gateway interfaces an outside circuit-switched network like PSTN for calls that originate from the H.323 system and are destined to the PSTN or vice versa. It receives from and sends to the PSTN signaling messages such as Q931 and Q2931 that are understood by the PSTN. The gateway also translates transmission formats between the H.323 system and the PSTN for voice, data, and video.

A gateway interfaces a gatekeeper just like an H.323 terminal would. A gateway appears to a gatekeeper and an H.323 terminal as just another endpoint for calls that come from a PSTN. The discussion above of the

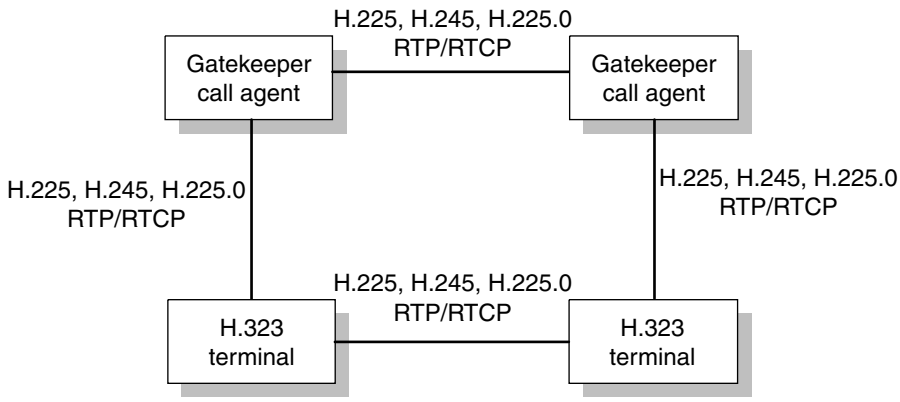


Figure 22-7
Interfaces of a homogeneous H.323 system.

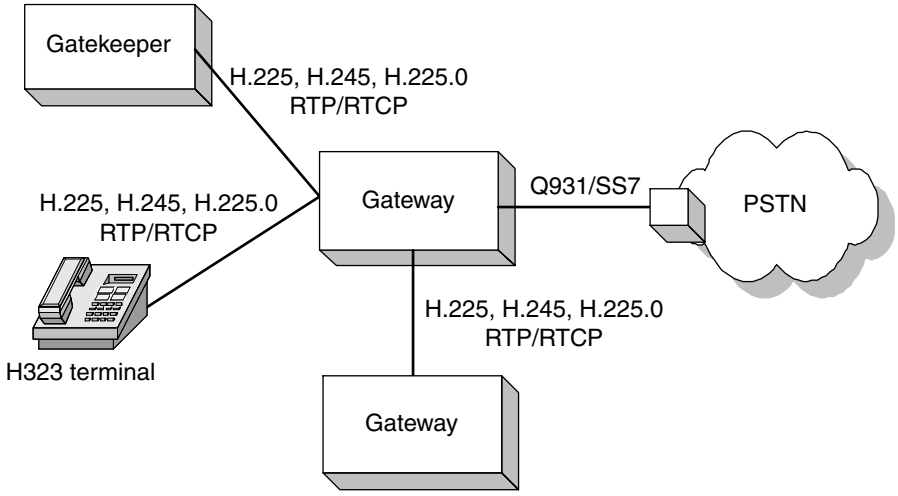


Figure 22-8
H.323 gateway interfaces.

interface between a gatekeeper and an endpoint applies here. That is, H.225.0 RAS, H.245 control messages, and H.225 Call Signaling messages are exchanged between an H.323 gateway and a gatekeeper, as shown in Fig. 22-8.

A gateway also interfaces other endpoints of an H.323 system. This includes two cases: gateway-gateway interfaces and gateway-H.323 terminal interfaces. For an H.323 gateway that is directly connected to another gateway on the same network, one gateway appears just as another H.323 endpoint to the other gateway and vice versa. To an H.323 terminal, the gateway appears as a peer terminal and communicates with H.225.0 RAS, H.245 control messages, and H.225 Call Signaling messages.

22.3 H.323 System Operation and Deployment

An H.323 system operation example will help thread together the H.323 system components described above to present a high-level view of how an H.323 system works. Then two deployment scenarios of H.323 systems will indicate where H.323 systems may be used.

22.3.1 H.323 System Operation Example

This example demonstrates how the pieces of an H.323 system come together and work as a functional system. The example is a multimedia conference call on a homogeneous H.323 system—that is, the call originates and terminates on the same H.323 system. There are two phases of the conference call: system initialization and multimedia calls.

System initialization involves four steps:

1. Gatekeeper discovery. An endpoint such as an H.323 terminal first needs to locate a gatekeeper and register with it before the H.323 system is initialized and configured. There are two kinds of gatekeeper discoveries: static and dynamic. The static gatekeeper discovery, also known as the manual discovery, associates an endpoint with a gatekeeper in a predetermined manner. The gatekeeper is determined either a priori or a well-known designed gatekeeper is used. The association can be performed manually by the operator configuring the system.

2. Endpoint registration. The next step is endpoint registration with the chosen gatekeeper. The registration is part of the endpoint's configuration process in which the endpoint joins a zone and informs the gatekeeper of its transport address and alias address.

3. Endpoint identification. For call setup, one endpoint, like an H.323 terminal, needs to know another endpoint's contact information, such as its signaling channel address and RAS channel address, from the second endpoint's alias addresses. It is practically impossible for an endpoint to statically maintain the contact information of other endpoints. Instead, an endpoint discovers the other endpoints in a similar way to how it discovers a gatekeeper: by sending a location request message.

4. Admission and bandwidth allocation. Before a terminal sets up a communication channel with another endpoint in the network, the terminal requests permission to access the network with an admission request message. The message contains information about call type, call

model, call services, and bandwidth. The call type lets the gatekeeper determine the bandwidth requirement. Examples of call types include point-to-point calls and multipoint conference calls.

Multimedia conference calls are intra-zone calls where the originating terminals and target terminals are all located in the same zone. They involve the following operations:

1. The originating terminal initiates a conference call to the gatekeeper via an H.245 message. Three channels are established between the originating H.323 terminal and the gatekeeper: the control channel, the RAS channel, and the signaling channel using TCP.
2. The gatekeeper connects the originating terminal to the requested terminals via H.245 control messages carried over an RTP channel.
3. All terminals send the audio and video to the MCU. Among the tasks the MCU performs are mixing audio and video streams, video and audio stream control, and compression and decompression of the audio and video data.
4. As directed by the gatekeeper, the MCU sends the audio and video data to the target terminals.
5. A terminal can leave the conference call in the middle of the conference by sending a DISCONNECT message.

22.3.2 H.323 System Applications

There are two general application scenarios for H.323 systems: LAN and WAN. A common application of an H.323 system relates to corporate LANs. In fact, H.323 was initially designed as a LAN application and later extended to WAN and other scenarios. In a LAN scenario, the gatekeeper function and even the MCU functions can be implemented as software functions on the LAN server or even on desktop PCs.

H.323 deployment in WAN is commonly coupled with interworking with PSTN. Parallel to building an independent H.323 system in a WAN environment is the scenario where H.323 gatekeeper and gateway functions are added to the existing softswitch and gateway to make them H.323-capable. This option is often chosen for its low cost and incremental build-out approach.

Although H.323 systems were originally designed for multimedia applications, VoIP is the dominant application so far. This is in part due to the fact that H.323 system deployment is still in an early stage and in part because the availability of access network bandwidth or the lack of

it is a key constraining factor in the deployment of packet broadband applications. With advances in broadband access network technologies and the increasing central processing unit (CPU) power of desktop computers, it is expected that packet broadband multimedia applications will see increasingly larger-scale deployment in the not so distant future.

REVIEW QUESTIONS

1. Discuss the reasons for the development of multimedia broadband applications over IP networks.
2. Describe the scope of the H.323 recommendation and the major changes in each major revision since its initial approval in 1996.
3. Describe the main components of an H.323 system and the interfaces between them.
4. Discuss the concept of H.323 endpoint. In its practical implementation, what can be an endpoint? Under what circumstances can an H.323 gateway become an endpoint?
5. Describe the configuration of an H.323 gateway and its main functions. There are two levels of interface between a gateway and an H.323 terminal: signaling and transport. Describe the major protocols for each level of interface.
6. An H.323 gatekeeper can take on a wide range of responsibilities, some mandatory and some optional. What are the mandatory functions of a gatekeeper, based on H.323 v4?
7. Describe the responsibilities and components of an H.323 MCU. What are the differences between a centralized and decentralized MCU?
8. Describe the mandatory and optional responsibilities of an H.323 gatekeeper. What is an H.323 zone and how is a zone associated with a gatekeeper?
9. Describe what can be an H.323 terminal and what minimal functions such a terminal must be able to carry out?
10. Provide a high-level description of a two-way VoIP call from an H.323 terminal to another H.323 terminal on the same H.323 system.
11. Describe briefly what RTP and RTCP are designed for and how they are used in an H.323 system.

12. Describe the H.323 protocol stack, the main functions of the H.245 Control Protocol, and how that protocol is used between a gatekeeper and an H.323 terminal.
13. Describe the main functions of H.225.0 RAS and how it is used between two H.323 terminals in the absence of a gatekeeper.
14. Describe the H.225 Call Signaling protocol and its relationship with the Q.931 Call Signaling protocol.
15. H.323 was initially designed for multimedia applications on packet LANs. It was then extended to go beyond LANs and to interwork with PSTNs. What are the major extensions to H.323 for it to work with PSTNs?

REFERENCES

Cisco. 2001. "Cisco gatekeeper external interface reference, version 3." Cisco Systems Inc. White paper. Web site: www.cisco.com.

H323 Forum. 2001. "H.323—A primer on the H.323 series standard." White paper. Web site: www.h323forum.com/papers/primer.

Handley, M., Schulzrinne, H., et al. 1999. "SIP: Session Initiation Protocol." IETF RFC 2543. Web site: www.ietf.org.

ITU-T. 1988a. "7 kHz Audio-Coding within 64 kbit/s." Recommendation G.722. Web site: www.itu.int/itu-t.

ITU-T. 1988b. "Pulse Code Modulation (PCM) of Voice Frequencies." Recommendation G.711. Web site: www.itu.int/itu-t.

ITU-T. 1992. "Coding of Speech at 16 kbit/s Using Low-delay Code Excited Linear Prediction." Recommendation G.728. Web site: www.itu.int/itu-t.

ITU-T. 1993a. "Digital subscriber signaling system No. 1 (DSS1)—ISDN User-Network Interface Layer 3 Specification for Basic Call Control." Recommendation Q.931. Web site: www.itu.int/itu-t.

ITU-T. 1993b. "Video Codec for Audiovisual Services at p × 64 kbit/s." Recommendation H.261. Web site: www.itu.int/itu-t.

ITU-T. 1993c. "Introduction to CCITT Signalling System No. 7." Recommendation 701. Web site: www.itu.int/itu-t.

ITU-T. 1996a. "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)." Recommendation G.729. Web sie: www.itu.int/itu-t.

ITU-T. 1996b. "Data Protcols for Multimedia Conferencing." Recommendation T.120. Web site: www.itu.int/itu-t.

ITU-T. 1996c. "Speech Coders: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s." Recommendation G.723.1. Web site: www.itu.int/itu-t.

ITU-T. 1998a. "Generic Functional Protocol for the Support of Supplementary Services in H.323." Recommendation H.450.1. Web site: www.itu.int/itu-t.

ITU-T. 1998b. "H.323 Extended for Loosely Coupled Conferences." Recommendation 332. Web site: www.itu.int/itu-t.

ITU-T. 1998c. "Video Coding for Low Bit Rate Communication." Recommendation H.263. Web site: www.itu.int/itu-t.

ITU-T. 1999. "Narrow-band Visual Telephone Systems and Terminal Equipment." Recommendation H.320. Web site: www.itu.int/itu-t.

ITU-T. 2000a. "Call Signaling Protocols and Media Stream Packetization for Packet-Based Multimedia Communication Systems." Recommendation 225.0. Web site: www.itu.int/itu-t.

ITU-T. 2000b. "Gateway Control Protocol." Recommendation 248. Web site: www.itu.int/itu-t.

ITU-T. 2000c. "Packet-Based Multimedia Communications Systems." Recommendation H.323. Web site: www.itu.int/itu-t.

ITU-T. 2000d. "Security and Encryption for H-Series (H.323 and other H.245-based) Multimedia Terminals." Recommendation H.235. Web site: www.itu.int/itu-t.

ITU-T. 2001. "Communication Procedures—Control Protocol for Multimedia Communications." Recommendation H.245. Web site: www.itu.int/itu-t.

ITU-T. 2002a. "Implementors' Guide for H.320 Recommendation Series." Recommendation H.221. Web site: www.itu.int/itu-t.

ITU-T. 2002b. "Terminal for Low Bit-Rate Multimedia Communication." Recommendation H.324. Web site: www.itu.int/itu-t.

Kumar, V., Korpi, M., and Sengodan, S. 2001. *IP Telephony with H.323: Architecture for Unified Networks and Integrated Services.* New York: John Wiley & Sons.

Schulzrinne, H., Casner, S., et al. 1996. "RTP: A Transport Protocol for Real-Time Applications." IETF RFC 1889. Web site: www.ietf.org.

CHAPTER 23

SIP and VoIP

23.1 SIP Protocol Basics

The basic building blocks of the Session Initiation Protocol include the protocol entities, the SIP address format, the SIP client and server relationships, and the protocol message exchanges and operations.

23.1.1 Brief History

SIP, an IETF standard, is an ASCII-based application layer control/signaling protocol for creating, modifying, maintaining, and terminating sessions with one or more participating terminals on an IP network. A session, in contrast to a connection in circuit-switched telephone networks, consists of a set of data streams that flow from a sender to one or more receivers. The data stream can be carried over a reliable TCP or unreliable UDP layer. SIP is an alternative protocol and architecture to H.323 for providing multimedia applications over IP networks.

SIP is designed to provide signaling and session management capabilities on packet networks. Signaling allows call information to be carried across network boundaries. Session management provides the ability to control the attributes of an end-to-end call.

Since the draft standard of SIP as RFC 2543 (Handley et al. 1999) was first published, additional service features have been added on a continuous basis. A revision to the original SIP was completed in mid-year 2002 (Rosenberg, Schulzrinne et al., 2002)

SIP is considered "Internet friendly" in many respects. Its simplicity and exclusive focus on the Internet have helped the protocol achieve wide acceptance in VoIP application despite the fact that it is a relatively late comer compared with the H.323 standards. The fact that the SIP client is incorporated into Microsoft Windows XP since late 2001 bears witness to the scope of its acceptance.

23.1.2 Overview of SIP Protocol

First, SIP is a peer-to-peer as well as a client-server protocol. The peers in a session are called *user agents* (UAs). A user agent can function in the following two roles:

- User agent client (UAC)—a client application that initiates a SIP request

- User agent server (UAS)—a server application that contacts the user when a SIP request is received and returns a response on behalf of the user

Typically, a SIP endpoint is capable of functioning as both a UAC and a UAS, but serves only as one or the other during each transaction. Whether the endpoint functions as a UAC or a UAS depends on the UA that initiated the request.

SIP *transaction* and *session* are other two key concepts. A SIP transaction consists of all the messages exchanged between a SIP client and a SIP server in a single session. "A SIP transaction occurs between a client and a server and comprises all messages from the first request sent from the client to the server up to a final (non-1xx) response sent from the server to the client. A transaction is identified by the CSeq sequence number within a single call leg" (Rosenberg et al., 2002).

A SIP session is a set of data streams flowing from one or more senders to one or more receivers. "A multimedia session is a set of multimedia senders and receivers and the data streams flowing from senders to receivers. A multimedia conference is an example of a multimedia session" (Handley and Jacobson, 1998). A session is identified by a session identifier in general, and, if SDP is used, a session is defined by the concatenation of the user name, session ID, network type, address type, and address elements in the origin field.

SIP provides a set of core capabilities to support VoIP and multimedia applications (Camarillo 2001; Rosenberg et al., 2002):

- It establishes a session between an originating and target endpoint (i.e., the calling and called parties) if it determines that a call can be completed. SIP also supports midcall changes, such as the addition of another endpoint to a conference call or changing a media characteristic or codec.
- It determines the location of a target endpoint. SIP supports address resolution, name mapping, and call redirection using the address resolution service from a location server.
- It determines the media capabilities of a target endpoint using the Session Description Protocol (SDP). SIP determines the "lowest level" of common services between the endpoints.
- It determines the availability of a target endpoint. If a call cannot be completed because the target endpoint is unavailable, SIP determines whether the called party is already on the phone or did not answer in the allotted number of rings. It

then returns a message indicating why the target endpoint was unavailable.

- It handles the transfer and termination of calls. SIP supports the transfer of calls from one endpoint to another. During a call transfer, SIP simply establishes a session between the transferee and a new endpoint (specified by the transferring party) and terminates the session between the transferee and the transferring party. At the end of a call, SIP terminates the sessions between all the parties.

23.1.3 SIP Address Format

SIP defines a SIP Universal Resource Locator (SIP URL) that is based on the World Wide Web (WWW) URL with extensions that can accommodate a variety of addresses such as host name, port, Web URL, and email address, among others. A SIP URL is included in every message to indicate the originator, current destination, and final recipient of a SIP request. The general format of the SIP URL consists of three major parts: user information, host port information, and Universal Request Identifier (URI) parameters, as shown in Fig. 23-1 (Rosenberg et al., 2002; Rosenberg and Schulzrinne 2002a).

The user information part identifies the user involved in the SIP request, and an optional password can be associated with the user information part. User information can be a user name, a telephone number, or some combination of the two. The host port part identifies a host name, and a port associated with the host. It can be a simple host name, an IPv4 address, an IPv6 address, a domain name, or some combination of these.

The URI parameter part gives a great deal of flexibility to specify a wide range of parameters. They include the network transport layer protocol parameter such as UDP, TCP, and SCTP; additional user parameters such as IP address, phone, or other users; SIP request type; additional host address information; and other types of parameters.

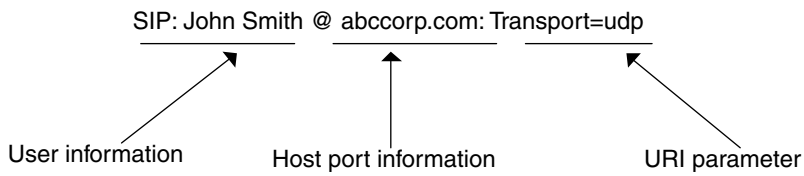


Figure 23-1
SIP address example.

23.1.4 SIP Server

A SIP server is a software system responsible for serving the requests from SIP clients by providing the requested services to the requesting clients. There are three different types of SIP servers, as described below, largely distinguished by the functions each performs: proxy, redirect, and registrar.

23.1.4.1 Proxy SIP Server A proxy server is an intermediate server that receives a SIP request from a client and then either directly forwards it or regenerates it and then sends it to the appropriate servers on the client's behalf. Basically, a proxy server receives SIP messages and forwards them to the next SIP server or a user agent in the network. Proxy servers can provide functions such as authentication, authorization, network access control, routing, reliable request retransmission, and security.

A proxy server can be implemented as *stateful* or *stateless*. A stateless proxy does not maintain any state information regarding a request once the request is forwarded. It forwards each request it receives to a downstream server or user agent and each response it receives back to the client without any processing. A stateful proxy, on the other hand, maintains state machines to keep track of the states of both incoming requests and outgoing requests. Some proxy servers must be stateful. The choice of a stateless or stateful proxy really depends on the desired services. Some services require state information. For those that do not, a stateless proxy with little overhead can scale very well in a large network.

The cases where a stateful proxy server is required include, for example, multiple-point conference calls, where the proxy server needs a forking capability and must therefore be stateful to maintain the state information of each leg of the call. A proxy server must also be stateful if it accepts stateful transport layer connections like a TCP connection or sends a request to multiple destinations (Rosenberg and Schulzrinne 2000a).

23.1.4.2 Redirect SIP Server A redirect server provides a client with information about the next hop or hops that a message should take to allow the client to contact the next-hop server or UAS directly. A redirect server does not issue any SIP requests of its own.

23.1.4.3 Registrar SIP Server A registrar server processes requests from SIP clients for the registration of their current locations. Registrar servers are often colocated with redirect or proxy servers.

23.1.5 SIP Client

A SIP client is a user agent client (UAC) that implements client-side SIP protocol capabilities. Physically, a SIP client can be a computer, in a multimedia telephone set or an IP phone set. A SIP client is responsible for initiating a request to a server.

23.1.6 SIP Protocol Operation

The SIP protocol operations center on simple request-response message exchanges where a SIP client or UAC sends a request to a server or a UAS and the server responds with a response message to the client.

SIP supports request forking where a server can fork a request into multiple clients either in parallel or sequential fashion to support applications such as conference calls or presence services, as will be explained later.

23.1.7 SIP Messages

SIP is a relatively simple protocol, like many Internet protocols, with a small number of messages. A message, termed a *method* in the SIP specification, is either a request issued by a client or a response to a request. There are a total of six mandatory request messages and a few extension request messages. There is only one generic response message.

23.1.7.1 SIP Request Messages The SIP request message has a generic format, as shown in Fig. 23-2, with three fields: request_line, message header, and message body. The message header, in turn, consists of three headers: general header, request header, and entity header.

The *request-line* part contains generic information about the request, indicating the type of request message such as INVITE, ACK, or CANCEL, as described below, the SIP version, and the request URI. A request URI is a SIP URL as described above or a generic URI as defined in IETF RFC 2396 (Berners-Lee et al., 1998).

The *general-header* part contains the information generic to both request and response messages. It may include fields such as call sequence number, call ID, call info, encryption method specification, timestamp, the *to* address of the call, the *from* address of the call, and the path that the request has taken so far (*via* field), among others. Note that not every message has every field.

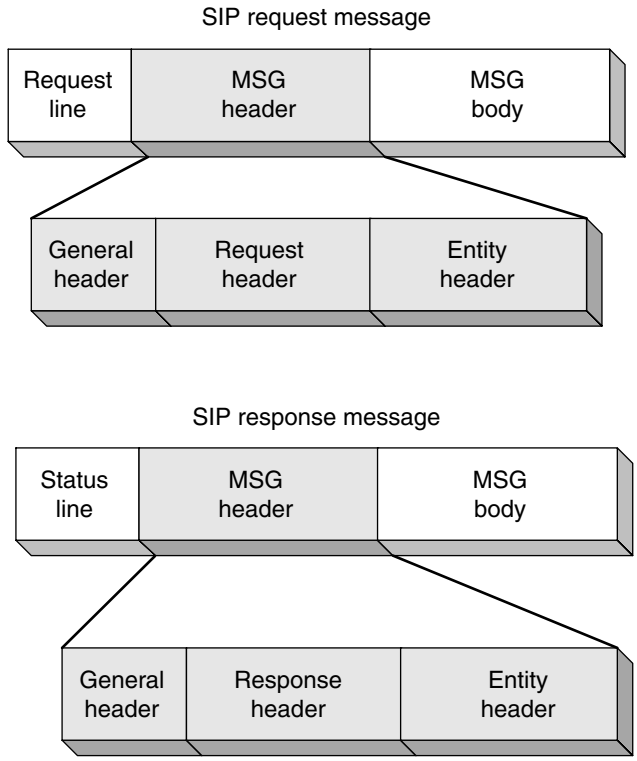


Figure 23-2
Generic structure of a SIP request message.

The *request-header* part allows the client to pass additional request-specific information to the server. It contains such fields as *priority,* which indicates the urgency of this request, *alert-info,* which provides an alternative call announcement in place of a default ring tone, *response-key,* which suggests the encryption key the called party should use in its response, and *subject,* which indicates the nature of the call, among others.

The *entity-header* part contains the control information or metadata about the message body if one is present. It contains fields such as *content-disposition,* which indicates how to interpret the message body, *content-length,* which indicates the length of the message body field, *content-encoding,* which indicates whether the message body has been encoded, and *content-type,* which indicates the message content type.

The *message-body* part contains the contents of the message, whose format depends on the message type, which is indicated in the entity header of the message header. Some types of request message, such as CANCEL, do not have a message body. The session description of the

Session Description Protocol is contained in the message body if one is present. Other message content types include free text, HTML page, and media-specific contents such as audio and video data.

SIP request messages can be grouped into two general categories. The first category is call setup and call takedown request messages for a client to make a request and for a server to respond to the request:

- *INVITE.* A caller, which can be a UAC or SIP server, issues an INVITE message to invite the called party to participate in a SIP session. The message may contain the address of the called party, the media to be used in the call, and other parameters.
- *ACK.* This message allows a client to confirm that it has received the final response to an INVITE request.
- *BYE.* This allows a client to indicate to the server that it intends to release the call leg.
- *CANCEL.* This message allows a client to cancel an outstanding request such as an INVITE or an ACK request.

Registration-related messages are the second type of SIP request messages, and allow a client to register with a server for address translation service or to issue a generic query request:

- *REGISTER.* This allows a client to bind the address in a request message to one or more URIs where the client can be reached.
- *OPTIONS.* This allows a client to query the server about its capabilities.

23.1.7.2 SIP Response Message The SIP response message allows a server to respond to a request from a client. A response message takes the format shown in the bottom half of Fig. 23-2. It is almost identical to the request message format except for the *response-header* part. The *response-header* contains the fields that include the status of the request, the reason for the status, the *from* address, the *to* address, call ID, the information about the server issuing the response, etc. Note that a response travels the same route as the request.

23.2 SIP System Architecture

The SIP system architecture consists of a set of SIP system components and a set of interfaces between the components that are built

upon the SIP protocol. A SIP system, as described in this chapter, is a broad term referring to a SIP-based packet broadband multimedia application system.

23.2.1 SIP System Protocol Stack Overview

A SIP system in general is concerned with application layer and transport layer protocols, as shown in Fig. 23-3 (Sinnereich and Johnston 2001). The application layer protocols include SDP, SIP, MEGACO/H.248, and Real-Time Stream Protocol (RTSP). SIP and SDP are used for SIP system signaling. MEGACO/H.248 (ITU-T 2002) supports interworking with PSTN media gateways. RTSP is used to provide multimedia session control for applications such as multimedia conference calls. The transport layer protocols include RTP and RTCP (Schulzrinne et al. 1996), discussed in Chap. 22. Since all of these protocols except for SDP and RTSP have been discussed, we turn to those next.

23.2.1.1 Session Description Protocol SDP, an IETF standard defined in RFC 2327 (Handley and Jacobson, 1998), is a protocol that

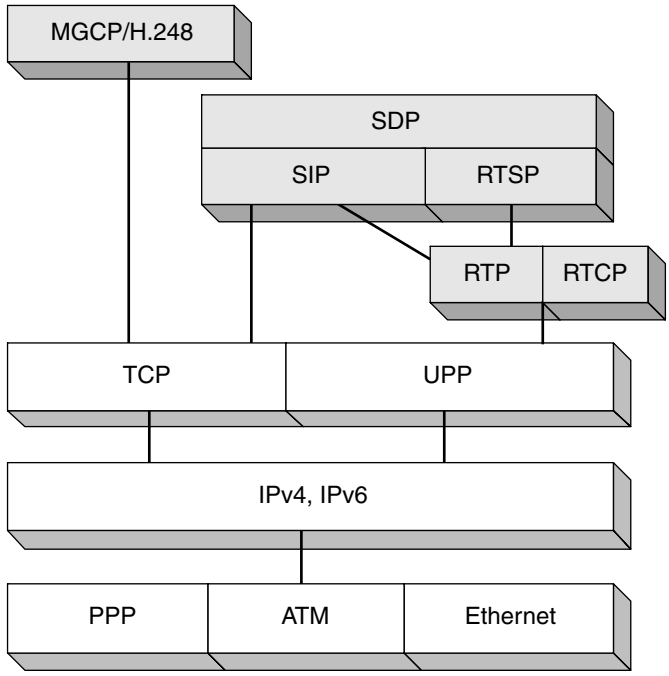


Figure 23-3
SIP protocol stack.

allows a client to announce the existence of a multimedia session to other clients and enables other clients to join in a session such as a multimedia conference call. The information SDP communicates to clients includes the session name and purpose, the time period during which the session is active, the type of media used for the session and other information such as address, port, and media format needed for clients to receive calls.

The SDP messages are encapsulated inside the SIP message body. Each SIP message contains zero or more SDP messages, and each SDP message can contain only one session description, although SDP allows the descriptions of multiple sessions to be concatenated into one SDP message.

23.2.1.2 Real-Time Stream Protocol RTSP is an application layer protocol used in conjunction with SIP in a SIP system to support multimedia applications. It operates over RTP over UDP over IP, as shown in Fig. 23-3. As defined in RFC 2326 (Schulzrinne et al., 1998), RTSP provides control of multiple data delivery sessions and allows a client to choose delivery channels based on UDP, TCP, or IP-multicast.

RTSP provides three types of operations to users: invite a media server to join a multimedia session, interface a media server to retrieve media data, and add additional media to an active presentation session. Syntax-wise, RTSP is designed to look like HTTP.

RTSP is designed to deliver real-time data content in the form of streaming. Packet data streaming breaks the data into packets of sizes that are based on the bandwidth available between a client and a server. RTSP supports playback of the data in real-time fashion. For example, when enough packets have been received by the client, the client applications can be playing one packet, decompressing another, and downloading the third. Applications of this type include real-time audio or video streaming that allows a user to play back the media data (e.g., listening to a song in MPEG 3 format) without waiting for the completion of downloading the entire file. Both live data feeds and stored clips can be the sources of media data.

23.2.2 SIP System Configuration

A SIP system, which is much more generic than SIP as a protocol, consists of three layers, as shown in Fig. 23-4: media and device, server and service, and application service.

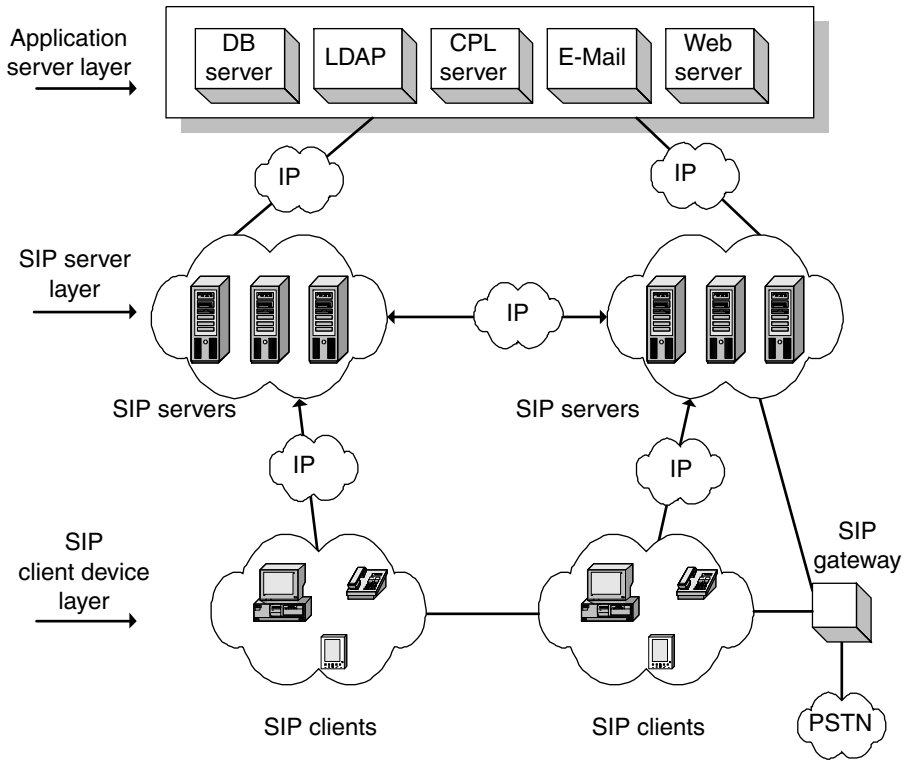


Figure 23-4 Three-layer SIP system architecture.

The media and device layer at the bottom consists of the SIP endpoints such as SIP client terminal and gateway that directly deal with the transmission media as well as interfacing the upper layer. The SIP server layer provides call control signaling and SIP services to the SIP clients. The application server layer provides application-specific services to the SIP server layer.

A SIP system is a multilayer client-server system. A SIP server provides service to the SIP clients while it itself is a client of the application servers that provide application services.

A SIP system is distributed. The SIP server intelligence is distributed among different SIP servers. Application intelligence is distributed among multiple application servers. System intelligence is distributed to the edge of network rather than concentrated at the center.

23.2.2.1 SIP Server A SIP server is a SIP protocol entity that carries out the SIP protocol operations. A SIP server is responsible for serving

requests from SIP clients and providing the requested services to the requesting clients. For practical purposes, a SIP server can be configured in various ways. For example, it can reside at a networked computer, or be colocated with softswitch call agent, a gateway controller, or an intelligent box that implements call control.

23.2.2.2 SIP Client Devices A SIP client device implements the client side of SIP protocol capabilities. Physically, SIP client devices include the following:

- A SIP phone that can normally act as both a UAS and a UAC
- A PC that has SIP client capability as well as phone capability, also known as *softphones*.
- A SIP gateway that provides call control and translation functions between a SIP call and a non-SIP system call, such as an H.323 system call or a circuit-switched PSTN call

23.2.2.3 SIP Gateway Analogous to an H.323 gateway, a SIP gateway provides translation functions between a SIP system and a non-SIP system. Given a large variety of non-SIP systems, there is no standard set of translation capabilities. A set of mapping functions is often implemented between a SIP system and a PSTN network. In addition, a gateway is also responsible for call setup and clearing on both the SIP side and the circuit-switched network side.

A SIP gateway provides two levels of translation services: signaling translation and transport-level translation. The signaling translation service includes mapping between SIP messages and SS7, Channel Associated Signaling, and primary rate interface (PRI) signaling messages that are commonly seen in circuit-switched PSTN networks.

A SIP gateway has two sides of network interfaces. On the IP network side, it has an IP network interface to receive and send IP packets and to process the whole IP protocol stack including IP, UTP/TCP, and RTP/RTCP. On the PSTN network side, it supports TDM digital interfaces such as T1/E1, T3/E3, OC3, etc. In addition, the transport services also include translation between conferencing endpoints and other terminal types and between audio and video codecs of a SIP system and a non-SIP system.

23.2.2.4 Application Server The application servers fall outside the scope of the SIP protocol itself but provide application-specific services to a SIP server. The scope of the application server layer has

expanded along with the increasing number of the applications a SIP system can support.

A *name locator server* is capable of translating one type of address to another type. Examples would include translation from email to a phone number, from phone to URL address, etc.

A *location server* is a data storage server used by a SIP redirect or proxy server to obtain information about a called party's location. A location server may be colocated with a SIP server. A location server is analogous to a service control point (SCP) of an Intelligent Network, where an 800 number is translated into a regular phone number.

Call Processing Language (CPL) server allows users to create simple Internet-based telephony services such as call waiting, call forwarding, and free phone service. A CPL server is an execution environment that can execute the services created with CPL.

The other types of application server with which a SIP server can interact include LDAP servers, database application servers, or XML servers. These provide services such as directory, authentication, email, Web access, and billing.

23.2.3 SIP System Operation Example

An example will help illustrate how a SIP system works. This example is derived from the example provided in RFC 3261 (Rosenberg et al., 2002). Assume that a SIP system is implemented on a corporate LAN. The scenario is that a user A calls a user B at the office, and user B happens to be in a conference room at the time of the call, as shown in Fig. 23-5.

1. After user A dials for user B, the SIP UAC at user A's PC sends an INVITE request message to a SIP proxy server on the LAN. User B is identified by the email address in the message, and the INVITE message initiates a SIP session.

2. The proxy server sends a request to the location server to get the detailed address of the called party. The location server sends back a response with the current address of user B. The location server is either manually configured for the proxy server or can be dynamically discovered by the SIP server at the system initialization time.

3. The proxy server initiates another INVITE message with the IP phone number as the *to* address in the message header on behalf of user A.

4. The IP phone in the conference room, after user B picks up the phone, sends a response back to the proxy server. The proxy server then

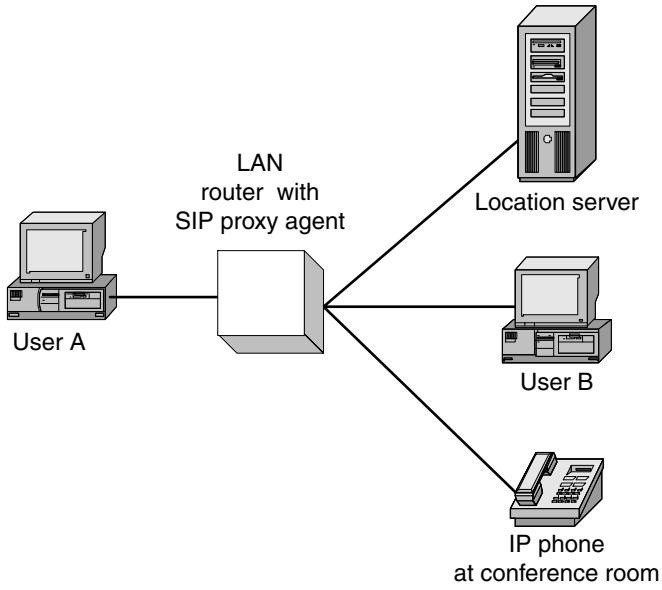


Figure 23-5
SIP system operation example.

generates and sends a response to user A that contains the OK status of the request.

5. User A then sends an ACK message to acknowledge receipt of the response to the INVITE message and confirms the UPD and RTP ports to be used to carry the phone conversation. The call setup is then completed after this message.

6. The IP packets containing the compressed phone conversation are transmitted between user A's PC and user B's IP phone, using RTP packets over UDP over IP.

7. The UAC at user B's phone issues a BYE request to the server after user B hangs up the phone when the conversation is finished. The proxy server then issues a BYE request on behalf of the SIP client to the calling party. The UAC at user A's PC stops transmitting any data to the destination indicated in the BYE message.

If the SIP server is of the redirect server type rather than a proxy server, the main difference in the operations will be at step 3. In this case, the redirect server passes the conference room IP phone number, the current location of user B, to user A instead of issuing an INVITE message on user A's behalf. User A's UAC initiates a second INVITE message directly to the IP phone.

23.3 SIP Applications

As the Internet penetrates ever more deeply into daily life, it has been predicted that SIP will soon begin to be deployed on a large scale. This section first provides a set of SIP application scenarios each of which describes a field such as wireless or packet cable where SIP systems are being deployed. Then it describes a set of SIP-supported services. A simple comparison between H.323 and SIP will help summarize the characteristics of SIP.

23.3.1 SIP Applications

A wide range of SIP deployments falls into four general areas: enterprise network, PSTN, packet cable, and 3G wireless.

23.3.1.1 SIP for Enterprise IP Telephony A SIP system in an enterprise LAN is one of the most common SIP system deployment scenarios. In this case, the SIP servers and SIP clients, and application servers such as location servers, are all located on the LAN.

The ability to interwork with PSTN is one important issue for SIP in a LAN environment. The SIP proxy server has been extended to support the interworking function, and the extension has resulted in the *outbound proxy server*, which is currently under consideration for the new version of the SIP standard. An outbound proxy server is responsible for routing all outbound requests from a domain such as a LAN to a SIP gateway, which then further routes the call into a PSTN network. In addition, the outbound proxy server may also manage corporate firewalls and policy enforcement and provide a dial plan.

Thus, the next generation of LAN routers and switches will be equipped with the capabilities of SIP gateways in order to interface PSTN networks at both the signaling and transport levels.

23.3.1.2 SIP in Packet Cable Networks SIP is the designated call signaling and control protocol in the packet cable architecture framework 1.2 specifications for the broadband multimedia packet cable networks currently under development. It is expected that SIP-based packet cable networks will be in deployment in not so distant future.

Extensions to SIP are underway within IETF to make SIP more suitable for packet cable applications. One extension is support for distributed call control (DCS), which is a key element for providing large-scale residential

IP telephony services over a cable network. One result of the SIP extensions is the DCS-proxy with additional SIP message headers to identify and distribute the distributed call state information.

23.3.1.3 SIP in 3G Mobile Networks SIP is the designated call signaling protocol by 3G Partner Project (3GPP), a global consortium charged with the task of developing the detailed 3G wireless technical specifications. SIP will be used for call signaling between mobile terminals and networks and between network call nodes. It will also be used for all IP multimedia call signaling on 3G wireless networks (Schulzrinne and Wedlund 2001).

SIP extensions to support 3G mobile applications are under consideration at IETF. One key extension is the support for the call state control function (CSCF), the call agent in wireless networks. That extension includes the following three components:

- *Proxy CSCF:* A SIP entity that is the first point of contact in a visited network and is responsible for locating the user's home network and providing translation, security, and authorization services.
- *Serving CSCF:* A SIP entity responsible for controlling sessions; it acts as registrar and triggers and executes services. The serving CSCF will access and consult the user's profile before rendering a service. Physically, it can be located in the home or visited network.
- *Interrogating CSCF:* A SIP entity that is the first point of contact in the home network. It assigns the serving CSCF and forwards SIP requests.

In addition, following 3GPP's decision to adopt SIP as its signaling protocol, SIP was selected as the platform upon which a mobile instant message (IM) service is offered for 3G mobile deployments. Some examples of this type of service include traffic news to driver's handset, delivery tracking, stock monitoring, sales-force tracking, and taxi services.

23.3.2 SIP Supported Services

Many extensions have been made to the original SIP in recent years, mainly to support a wide range of multimedia and VoIP services. Four general categories of service are under consideration for SIP support: common telephony, IN, Internet, and multimedia. Some of services have been standardized while others are still at the trial stage.

For SIP VoIP applications, a SIP system is expected to support a common set of Internet telephony services. The Telecommunications Industry Association is in the process of identifying and recommending a set of Internet telephony services much like the standard set of PSTN services identified by Bellcore. The set of Internet telephony services includes those that are widely used and commonly expected from a PSTN network:

- Call forwarding
- Call transfer
- Caller ID
- Three-way calls
- Call waiting
- Camp on
- Do-not-disturb
- Call hold and call return
- Business PBX and centrex services

Efforts are underway to have SIP systems support another category of telephony services, known as intelligent network (IN) services, that include free phone (800 number calls), calling card, and debt card call services.

Multimedia services such as multimedia conferencing are another category of services expected of the broadband packet network these days. Complicated media control is beyond the scope of the SIP protocol itself, but SIP provides a signaling mechanism to support multimedia applications. Each medium is a separate session, and multiple sessions can be combined to make a multimedia conferencing session. A SIP system will have to reply on other protocols for conference control functions such as the election of a conference chair and floor control. However, currently there are no standard Internet protocols for conference control yet.

SIP systems are perceived to have the advantage of supporting the "native" Internet services such as email, text-based chat sessions, and instant messaging. Internet IM combined with telephony service is a new possibility for SIP. Internet IM is associated with a "buddy list," and the presence or absence of a member of the list can be indicated. If a user is online and available to receive a message, an indicator displays this information to other users who have subscribed to that user's presence information. By clicking on the name of the user, an instant message can be sent in near real time. With SIP, a session consisting of any form

of communication can be set up. So it is possible to promote an IM session to a telephone call, a whiteboard meeting, or a video session at the click of a button. This is an efficient business communications tool for applications such as setting up conferences based on the availability of attendees.

23.3.3 Comparison Between H.323 and SIP

SIP and H.323 represent two different approaches to developing VoIP and multimedia applications over IP networks. The comparisons between the two approaches, as listed in Table 23-1, serve to sum up the defining characteristics of each (Packetizer 2001).

H.323 started out as a general framework targeting multimedia applications on packet LANs, which were then extended beyond LAN boundaries.

TABLE 23-1

Comparisons Between H.323 and SIP Systems

	H.323	SIP
Relationships between signaling entities	Master-slave	Client-server, peer-to-peer, although master-slave is as an supported extension
Addressing scheme	E.164 numbering, URL, H.323 identifier	SIP URL, an augmented URL that may include many types of address
Signaling transport	Transport-neutral, UDP or TCP	UDP and TCP
Media transport protocol	RTP over UDP over IP	RTP over UDP over IP
Interworking with PSTN	Easier since the call signaling protocol H.225 based on ITU Q931	No relation to the established PSTN signaling protocols
Message encoding	Binary	Text
Message syntax definition language	ASN.1	Augmented Backus-Naur form (ABNF)
Type of services supported	Packet telephony services, multimedia services	Internet telephony services; multimedia services; Internet services
Application areas	Integrated LAN, interworking with PSTN	Integrated LAN, 3G wireless, packet cable networks, and interworking with PSTN

There is a complete set of specifications, ranging from codec, to call signaling, to conference control. H.323 has ended up being quite comprehensive, as well as complicated. In contrast, SIP started out as a simple protocol to set up and end generic sessions on the Internet. It almost seems that it "happens" to be applicable to VoIP and multimedia applications. Then many extensions were made to the original SIP to accommodate various other needs such as telephony applications and wireless applications. Overall, SIP, just like many other Internet protocols, is simpler and less encompassing itself while relying on many other Internet protocols to become a complete functional system.

23.3.4 Billing for SIP Systems

Traditionally, the raw data from which billing data is derived, called *call detailed records* (CDR) is based on call models. The better defined the call model, the easier to extract the CDR data. The H.323 call model is largely based on the Q931 call model, and thus detailed records can be generated with little difficulty. On the other hand, SIP represents a quite different call model, a distributed one that does not fit well into the established patterns of call models. In order for packet broadband applications supported by SIP systems to become a sustainable business model, a set of new standard billing models need to be established. This can hardly be a quick process given the complexity of the issues.

REVIEW QUESTIONS

1. Describe the three layers of the generic SIP architecture as discussed in this chapter. The SIP protocol itself actually only covers two out of the three layers. What are those two layers?
2. Describe three different types of SIP servers and when each type is used.
3. SIP is considered very "Internet friendly." Discuss why.
4. Describe the main components of the SIP URL or addressing scheme. List the types of addresses that can be in a SIP URL. If a user is identified by a telephone number, where will the phone number be in the SIP URL?
5. Compare the SIP request and response messages and describe the parts they have in common and the main differences between them.

6. Describe the main functions of SDP, how it is used with SIP, and how it is encapsulated inside a SIP message.
7. Describe what constitutes a SIP endpoint and under what circumstances a SIP gateway becomes a SIP endpoint.
8. Describe the types of service a SIP system can support. Explain why the majority of SIP system implementations focus on VoIP applications and telephony services.
9. Describe how RTSP can be used with SIP in a SIP system on a corporate LAN and the type of service RTSP supports.
10. Based on your understanding of SIP, discuss the advantages of using SIP as a call signaling protocol between a mobile terminal and a base station in a wireless network.
11. Draw a few comparisons between an H.323 and a SIP system and describe what applications each system is better suited for.

REFERENCES

Berners-Lee, T., Fielding, R. et al. 1998. "Uniform Resource Identifier (URI): Generic Syntax." IETF RFC 2396. Web site: www.ietf.org.

Camarillo, G. 2001. *SIP Demystified*. New York: McGraw-Hill.

Handley, M., Schulzrinne, H. et al. 2001. "SIP: Session Initiation Protocol." IETF. RFC 2543bis. Web site: www.ietf.org.

Handley, M., and Jacobson, V. 1998. "SDP: Session Description Protocol." IETF RFC 2327. Web site: www.ietf.org.

Handley, M., Schulzrinne, H., Schooler, E., et al. 1999. "SIP: Session Initiation Protocol." IETF RFC 2543. Web site: www.org.ietf.

ITU-T. 2002. "Gateway Control Protocol." Recommendation H.248. Web site: www.itu.int/ITU-T/.

Ong, L., and Yoakum, J. 2002. "An Introduction to the Stream Control Transmission Protocol (SCTP)." IETF RFC 3286. Web site: www.ietf.org.

Packetizer. 2001. "H.323 versus SIP: a Comparison." White paper. Web site: www.packetizer.com.

Rosenberg, J., Schulzrinne, H., et al. 2002. "SIP: Session Initiation Protocol." IETF RFC 3261. Web site: www.ietf.org.

Rosenberg, J., and Schulzrinne, H. 2002a. "Session Initiation Protocol (SIP): Locating SIP Servers." IETF RFC 3263.

Schulzrinne, H. et al. 1996. "RTP: A Transport Protocol for Real-Time Applications." IETF RFC 1889. Web site: www.ietf.org.

Schulzrinne, H., Casner, S., et al. 1996. "RTP: A Transport Protocol for Real-Time Applications." IETF RFC 1999. Web site: www.ietf.org.

Schulzrinne, H., Rao, A., and Lanphier, R. 1998. "Real Time Streaming Protocol (RTSP)." IETF RFC 2326. Web site: www.ietf.org.

Schulzrinne. H., and Wedlund. E. 2001. "Application Layer Mobility Using SIP." *Mobile Computing and Communications Review,* Vol. 1, No. 1.

Sinnereich, H., and Johnston, A. 2001. *Internet Communications Using SIP: Delivering VoIP and Multimedia Service with Session Initiation Protocol.* New York: John Wiley & Sons.

Stewart, R., Xie, Q., et al. 2000. "Stream Control Transmission Protocol." IETF RFC 2960. Web site: www.ietf.org.