PART

4

Network Layer

Objectives

The network layer is responsible for the source-to-destination delivery of a packet, possibly across multiple networks (links). Whereas the data link layer oversees the delivery of the packet between two systems on the same network (links), the network layer ensures that each packet gets from its point of origin to its final destination.

The network layer is responsible for the delivery of individual packets from the source to the destination host.

The network layer adds a header that includes the logical addresses of the sender and receiver to the packet coming from the upper layer. If a packet travels through the Internet, we need this addressing system to help distinguish the source and destination.

When independent networks or links are connected together to create an internetwork, routers or switches route packets to their final destination. One of the functions of the network layer is to provide a routing mechanism.

In Part 4 of the book, we first discuss logical addressing (referred to as IP addressing in the Internet). We then discuss the main as well as some auxiliary protocols that are responsible for controlling the delivery of a packet from its source to its destination.

Part 4 of the book is devoted to the network layer and the services provided by this layer.

Chapters

This part consists of four chapters: Chapters 19 to 22.

Chapter 19

Chapter 19 discusses logical or IP addressing. We first discuss the historical classful addressing. We then describe the new classless addressing designed to alleviate some problems inherent in classful addressing. The completely new addressing system, IPv6, which may become prevalent in the near future, is also discussed.

Chapter 20

Chapter 20 is devoted to the main protocol at the network layer that supervises and controls the delivery of packets from the source to destination. This protocol is called the Internet Protocol or IP.

Chapter 21

Chapter 21 is devoted to some auxiliary protocols defined at the network layer, that help the IP protocol do its job. These protocols perform address mapping (logical to physical or vice versa), error reporting, and facilitate multicast delivery.

Chapter 22

Delivery and routing of packets in the Internet is a very delicate and important issue. We devote Chapter 22 to this matter. We first discuss the mechanism of delivery and routing. We then briefly discuss some unicast and multicast routing protocols used in the Internet today.

CHAPTER 19

Network Layer: Logical Addressing

As we discussed in Chapter 2, communication at the network layer is host-to-host (computer-to-computer); a computer somewhere in the world needs to communicate with another computer somewhere else in the world. Usually, computers communicate through the Internet. The packet transmitted by the sending computer may pass through several LANs or WANs before reaching the destination computer.

For this level of communication, we need a global addressing scheme; we called this logical addressing in Chapter 2. Today, we use the term **IP address** to mean a logical address in the network layer of the TCP/IP protocol suite.

The Internet addresses are 32 bits in length; this gives us a maximum of 2^{32} addresses. These addresses are referred to as IPv4 (IP version 4) addresses or simply IP addresses if there is no confusion.

The need for more addresses, in addition to other concerns about the IP layer, motivated a new design of the IP layer called the new generation of IP or IPv6 (IP version 6). In this version, the Internet uses 128-bit addresses that give much greater flexibility in address allocation. These addresses are referred to as IPv6 (IP version 6) addresses.

In this chapter, we first discuss IPv4 addresses, which are currently being used in the Internet. We then discuss the IPv6 addresses, which may become dominant in the future.

19.1 IPv4 ADDRESSES

An **IPv4 address** is a 32-bit address that *uniquely* and *universally* defines the connection of a device (for example, a computer or a router) to the Internet.

An IPv4 address is 32 bits long.

IPv4 addresses are unique. They are unique in the sense that each address defines one, and only one, connection to the Internet. Two devices on the Internet can never have the same address at the same time. We will see later that, by using some strategies, an address may be assigned to a device for a time period and then taken away and assigned to another device.

On the other hand, if a device operating at the network layer has m connections to the Internet, it needs to have m addresses. We will see later that a router is such a device.

The IPv4 addresses are universal in the sense that the addressing system must be accepted by any host that wants to be connected to the Internet.

The IPv4 addresses are unique and universal.

Address Space

A protocol such as IPv4 that defines addresses has an **address space.** An address space is the total number of addresses used by the protocol. If a protocol uses N bits to define an address, the address space is 2^N because each bit can have two different values (0 or 1) and N bits can have 2^N values.

IPv4 uses 32-bit addresses, which means that the address space is 2^{32} or 4,294,967,296 (more than 4 billion). This means that, theoretically, if there were no restrictions, more than 4 billion devices could be connected to the Internet. We will see shortly that the actual number is much less because of the restrictions imposed on the addresses.

The address space of IPv4 is 2^{32} or 4,294,967,296.

Notations

There are two prevalent notations to show an IPv4 address: **binary notation** and **dotted-decimal notation.**

Binary Notation

In binary notation, the IPv4 address is displayed as 32 bits. Each octet is often referred to as a byte. So it is common to hear an IPv4 address referred to as a 32-bit address or a 4-byte address. The following is an example of an IPv4 address in binary notation:

01110101 10010101 00011101 00000010

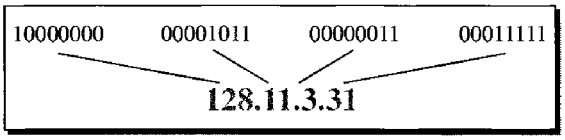
Dotted-Decimal Notation

To make the IPv4 address more compact and easier to read, Internet addresses are usually written in decimal form with a decimal point (dot) separating the bytes. The following is the **dotted-decimal notation** of the above address:

117.149.29.2

Figure 19.1 shows an IPv4 address in both binary and dotted-decimal notation. Note that because each byte (octet) is 8 bits, each number in dotted-decimal notation is a value ranging from 0 to 255.

Figure 19.1 *Dotted-decimal notation and binary notation for an IPv4 address*



Numbering systems are reviewed in Appendix B.

Example 19.1

Change the following IPv4 addresses from binary notation to dotted-decimal notation.

a. 10000001 00001011 00001011 11101111

b. 11000001 10000011 00011011 11111111

Solution

We replace each group of 8 bits with its equivalent decimal number (see Appendix B) and add dots for separation.

a. 129.11.11.239

b. 193.131.27.255

Example 19.2

Change the following IPv4 addresses from dotted-decimal notation to binary notation.

a. 111.56.45.78

b. 221.34.7.82

Solution

We replace each decimal number with its binary equivalent (see Appendix B).

a. 01101111 00111000 00101101 01001110

b. 11011101 00100010 00000111 01010010

Example 19.3

Find the error, if any, in the following IPv4 addresses.

a. 111.56.045.78

b. 221.34.7.8.20

c. 75.45.301.14

d. 11100010.23.14.67

Solution

a. There must be no leading zero (045).

b. There can be no more than four numbers in an IPv4 address.

c. Each number needs to be less than or equal to 255 (301 is outside this range).

d. A mixture of binary notation and dotted-decimal notation is not allowed.

Classful Addressing

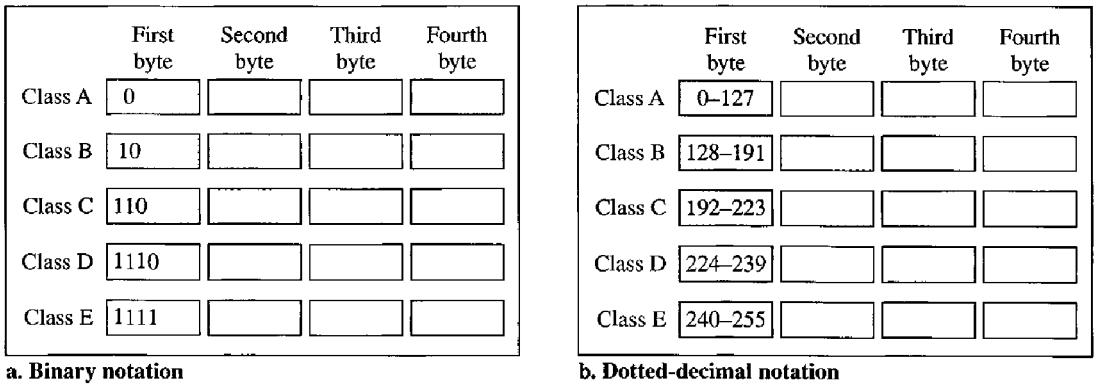
IPv4 addressing, at its inception, used the concept of classes. This architecture is called **classful addressing.** Although this scheme is becoming obsolete, we briefly discuss it here to show the rationale behind classless addressing.

In classful addressing, the address space is divided into five classes: A, B, C, D, and E. Each class occupies some part of the address space.

In classful addressing, the address space is divided into five classes: A, B, C, D, and E.

We can find the class of an address when given the address in binary notation or dotted-decimal notation. If the address is given in binary notation, the first few bits can immediately tell us the class of the address. If the address is given in decimal-dotted notation, the first byte defines the class. Both methods are shown in Figure 19.2.

Figure 19.2 *Finding the classes in binary and dotted-decimal notation*



a. Binary notation **b. Dotted-decimal notation**

Example 19.4

Find the class of each address.

a. **00**000001 00001011 00001011 11101111
b. **11**000001 10000011 00011011 11111111
c. **14**.23.120.8
d. **252**.5.15.111

Solution

a. The first bit is 0. This is a **class A address.**
b. The first 2 bits are 1; the third bit is 0. This is a **class C address.**
c. The first byte is 14 (between 0 and 127); the class is A.
d. The first byte is 252 (between 240 and 255); the class is E.

Classes and Blocks

One problem with classful addressing is that each class is divided into a fixed number of blocks with each block having a fixed size as shown in Table 19.1.

Table 19.1 *Number of blocks and block size in classful IPv4 addressing*

Class	*Number of Blocks*	*Block Size*	*Application*
A	128	16,777,216	Unicast
B	16,384	65,536	Unicast
C	2,097,152	256	Unicast
D	1	268,435,456	Multicast
E	1	268,435,456	Reserved

Let us examine the table. Previously, when an organization requested a block of addresses, it was granted one in class A, B, or C. **Class A addresses** were designed for large organizations with a large number of attached hosts or routers. **Class B addresses** were designed for midsize organizations with tens of thousands of attached hosts or routers. **Class C addresses** were designed for small organizations with a small number of attached hosts or routers.

We can see the flaw in this design. A block in class A address is too large for almost any organization. This means most of the addresses in class A were wasted and were not used. A block in class B is also very large, probably too large for many of the organizations that received a class B block. A block in class C is probably too small for many organizations. **Class D addresses** were designed for multicasting as we will see in a later chapter. Each address in this class is used to define one group of hosts on the Internet. The Internet authorities wrongly predicted a need for 268,435,456 groups. This never happened and many addresses were wasted here too. And lastly, the **class E addresses** were reserved for future use; only a few were used, resulting in another waste of addresses.

In classful addressing, a large part of the available addresses were wasted.

Netid and Hostid

In classful addressing, an IP address in class A, B, or C is divided into **netid** and **hostid.** These parts are of varying lengths, depending on the class of the address. Figure 19.2 shows some netid and hostid bytes. The netid is in color, the hostid is in white. Note that the concept does not apply to classes D and E.

In class A, one byte defines the netid and three bytes define the hostid. In class B, two bytes define the netid and two bytes define the hostid. In class C, three bytes define the netid and one byte defines the hostid.

Mask

Although the length of the netid and hostid (in bits) is predetermined in classful addressing, we can also use a **mask** (also called the **default mask**), a 32-bit number made of

contiguous 1s followed by contiguous 0s. The masks for classes A, B, and C are shown in Table 19.2. The concept does not apply to classes D and E.

Table 19.2 *Default masks for classful addressing*

Class	*Binary*	*Dotted-Decimal*	*CIDR*
A	11111111 00000000 00000000 00000000	255.0.0.0	/8
B	11111111 11111111 00000000 00000000	255.255.0.0	/16
C	11111111 11111111 11111111 00000000	255.255.255.0	/24

The mask can help us to find the netid and the hostid. For example, the mask for a class A address has eight 1s, which means the first 8 bits of any address in class A define the netid; the next 24 bits define the hostid.

The last column of Table 19.2 shows the mask in the form /*n* where *n* can be 8, 16, or 24 in classful addressing. This notation is also called slash notation or **Classless Interdomain Routing (CIDR)** notation. The notation is used in classless addressing, which we will discuss later. We introduce it here because it can also be applied to classful addressing. We will show later that classful addressing is a special case of classless addressing.

Subnetting

During the era of classful addressing, **subnetting** was introduced. If an organization was granted a large block in class A or B, it could divide the addresses into several contiguous groups and assign each group to smaller networks (called **subnets**) or, in rare cases, share part of the addresses with neighbors. Subnetting increases the number of 1s in the mask, as we will see later when we discuss classless addressing.

Supernetting

The time came when most of the class A and class B addresses were depleted; however, there was still a huge demand for midsize blocks. The size of a class C block with a maximum number of 256 addresses did not satisfy the needs of most organizations. Even a midsize organization needed more addresses. One solution was **supernetting.** In supernetting, an organization can combine several class C blocks to create a larger range of addresses. In other words, several networks are combined to create a supernetwork or a **supernet.** An organization can apply for a set of class C blocks instead of just one. For example, an organization that needs 1000 addresses can be granted four contiguous class C blocks. The organization can then use these addresses to create one supernetwork. Supernetting decreases the number of 1s in the mask. For example, if an organization is given four class C addresses, the mask changes from /24 to /22. We will see that classless addressing eliminated the need for supernetting.

Address Depletion

The flaws in classful addressing scheme combined with the fast growth of the Internet led to the near depletion of the available addresses. Yet the number of devices on the Internet is much less than the 2^{32} address space. We have run out of class A and B addresses, and

a class C block is too small for most midsize organizations. One solution that has alleviated the problem is the idea of classless addressing.

Classful addressing, which is almost obsolete, is replaced with classless addressing.

Classless Addressing

To overcome address depletion and give more organizations access to the Internet, **classless addressing** was designed and implemented. In this scheme, there are no classes, but the addresses are still granted in blocks.

Address Blocks

In classless addressing, when an entity, small or large, needs to be connected to the Internet, it is granted a block (range) of addresses. The size of the block (the number of addresses) varies based on the nature and size of the entity. For example, a household may be given only two addresses; a large organization may be given thousands of addresses. An ISP, as the Internet service provider, may be given thousands or hundreds of thousands based on the number of customers it may serve.

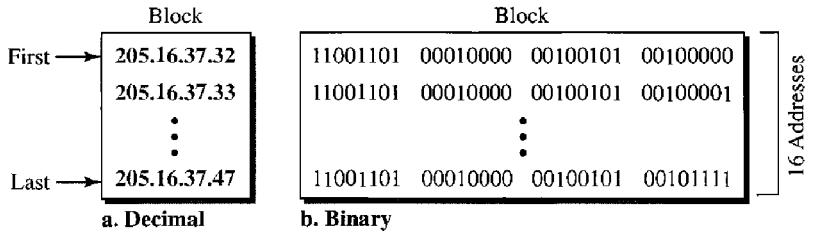
Restriction To simplify the handling of addresses, the Internet authorities impose three restrictions on classless address blocks:

1. The addresses in a block must be contiguous, one after another.
2. The number of addresses in a block must be a power of 2 (1, 2, 4, 8, . . .).
3. The first address must be evenly divisible by the number of addresses.

Example 19.5

Figure 19.3 shows a block of addresses, in both binary and dotted-decimal notation, granted to a small business that needs 16 addresses.

Figure 19.3 *A block of 16 addresses granted to a small organization*



We can see that the restrictions are applied to this block. The addresses are contiguous. The number of addresses is a power of 2 ($16 = 2^4$), and the first address is divisible by 16. The first address, when converted to a decimal number, is 3,440,387,360, which when divided by 16 results in 215,024,210. In Appendix B, we show how to find the decimal value of an IP address.

Mask

A better way to define a block of addresses is to select any address in the block and the mask. As we discussed before, a mask is a 32-bit number in which the n leftmost bits are 1s and the $32 - n$ rightmost bits are 0s. However, in classless addressing the mask for a block can take any value from 0 to 32. It is very convenient to give just the value of n preceded by a slash (CIDR notation).

In IPv4 addressing, a block of addresses can be defined as
x.y.z.t /*n*
in which x.y.z.t defines one of the addresses and the /*n* defines the mask.

The address and the /n notation completely define the whole block (the first address, the last address, and the number of addresses).

First Address The first address in the block can be found by setting the $32 - n$ rightmost bits in the binary notation of the address to 0s.

The first address in the block can be found by setting the rightmost $32 - n$ bits to 0s.

Example 19.6

A block of addresses is granted to a small organization. We know that one of the addresses is 205.16.37.39/28. What is the first address in the block?

Solution

The binary representation of the given address is 11001101 00010000 00100101 00100111. If we set 32 − 28 rightmost bits to 0, we get 11001101 0001000 00100101 0010000 or 205.16.37.32. This is actually the block shown in Figure 19.3.

Last Address The last address in the block can be found by setting the $32 - n$ rightmost bits in the binary notation of the address to 1s.

The last address in the block can be found by setting the rightmost $32 - n$ bits to 1s.

Example 19.7

Find the last address for the block in Example 19.6.

Solution

The binary representation of the given address is 11001101 00010000 00100101 00100111. If we set 32 − 28 rightmost bits to 1, we get 11001101 00010000 00100101 00101111 or 205.16.37.47. This is actually the block shown in Figure 19.3.

Number of Addresses The number of addresses in the block is the difference between the last and first address. It can easily be found using the formula 2^{32-n}.

The number of addresses in the block can be found by using the formula 2^{32-n}.

Example 19.8

Find the number of addresses in Example 19.6.

Solution

The value of n is 28, which means that number of addresses is 2^{32-28} or 16.

Example 19.9

Another way to find the first address, the last address, and the number of addresses is to represent the mask as a 32-bit binary (or 8-digit hexadecimal) number. This is particularly useful when we are writing a program to find these pieces of information. In Example 19.5 the /28 can be represented as 11111111 11111111 11111111 11110000 (twenty-eight 1s and four 0s). Find

a. The first address
b. The last address
c. The number of addresses

Solution

a. The first address can be found by ANDing the given addresses with the mask. ANDing here is done bit by bit. The result of ANDing 2 bits is 1 if both bits are 1s; the result is 0 otherwise.

Address:	11001101	00010000	00100101	00100111
Mask:	**11111111**	**11111111**	**11111111**	**11110000**
First address:	11001101	00010000	00100101	00100000

b. The last address can be found by ORing the given addresses with the complement of the mask. ORing here is done bit by bit. The result of ORing 2 bits is 0 if both bits are 0s; the result is 1 otherwise. The complement of a number is found by changing each 1 to 0 and each 0 to 1.

Address:	11001101	00010000	00100101	00100111
Mask complement:	**00000000**	**00000000**	**00000000**	**00001111**
Last address:	11001101	00010000	00100101	00101111

c. The number of addresses can be found by complementing the mask, interpreting it as a decimal number, and adding 1 to it.

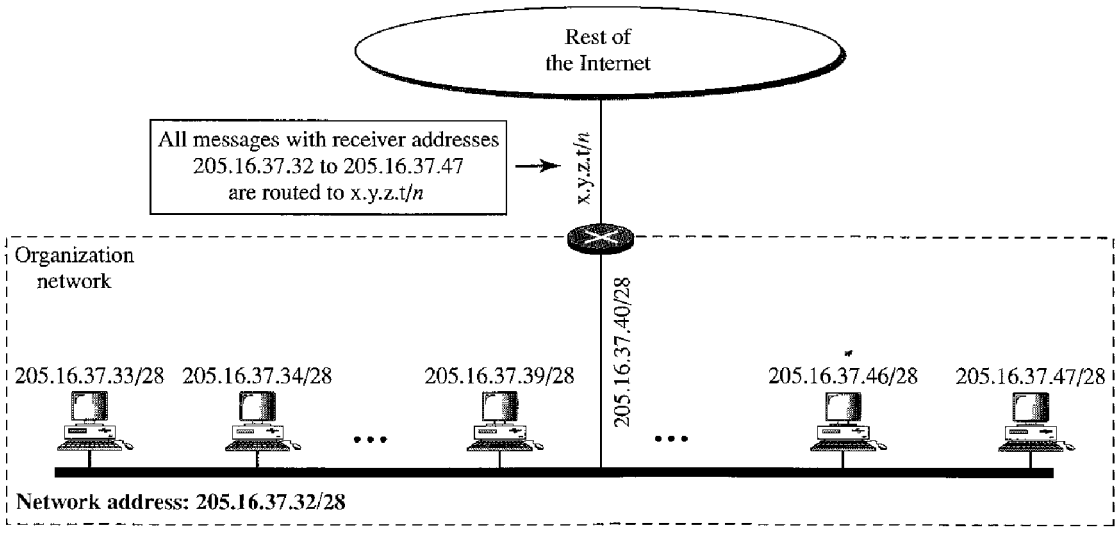
Mask complement:	**00000000 00000000 00000000 00001111**
Number of addresses:	15 + 1 = 16

Network Addresses

A very important concept in IP addressing is the **network address.** When an organization is given a block of addresses, the organization is free to allocate the addresses to the devices that need to be connected to the Internet. The first address in the class, however, is normally (not always) treated as a special address. The first address is called the network address and defines the organization network. It defines the organization itself to the rest of the world. In a later chapter we will see that the first address is the one that is used by routers to direct the message sent to the organization from the outside.

Figure 19.4 shows an organization that is granted a 16-address block.

Figure 19.4 *A network configuration for the block 205.16.37.32/28*



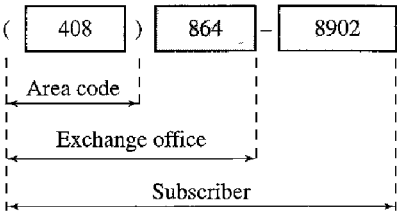
The organization network is connected to the Internet via a router. The router has two addresses. One belongs to the granted block; the other belongs to the network that is at the other side of the router. We call the second address x.y.z.t/*n* because we do not know anything about the network it is connected to at the other side. All messages destined for addresses in the organization block (205.16.37.32 to 205.16.37.47) are sent, directly or indirectly, to x.y.z.t/*n*. We say directly or indirectly because we do not know the structure of the network to which the other side of the router is connected.

The first address in a block is normally not assigned to any device; it is used as the network address that represents the organization to the rest of the world.

Hierarchy

IP addresses, like other addresses or identifiers we encounter these days, have levels of hierarchy. For example, a telephone network in North America has three levels of hierarchy. The leftmost three digits define the area code, the next three digits define the exchange, the last four digits define the connection of the local loop to the central office. Figure 19.5 shows the structure of a hierarchical telephone number.

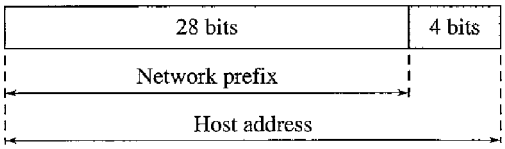
Figure 19.5 *Hierarchy in a telephone network in North America*



Two-Level Hierarchy: No Subnetting

An IP address can define only two levels of hierarchy when not subnetted. The n leftmost bits of the address x.y.z.t/n define the network (organization network); the $32 - n$ rightmost bits define the particular host (computer or router) to the network. The two common terms are prefix and suffix. The part of the address that defines the network is called the **prefix;** the part that defines the host is called the **suffix.** Figure 19.6 shows the hierarchical structure of an IPv4 address.

Figure 19.6 *Two levels of hierarchy in an IPv4 address*



The prefix is common to all addresses in the network; the suffix changes from one device to another.

Each address in the block can be considered as a two-level hierarchical structure: the leftmost n bits (prefix) define the network; the rightmost $32 - n$ bits define the host.

Three-Levels of Hierarchy: Subnetting

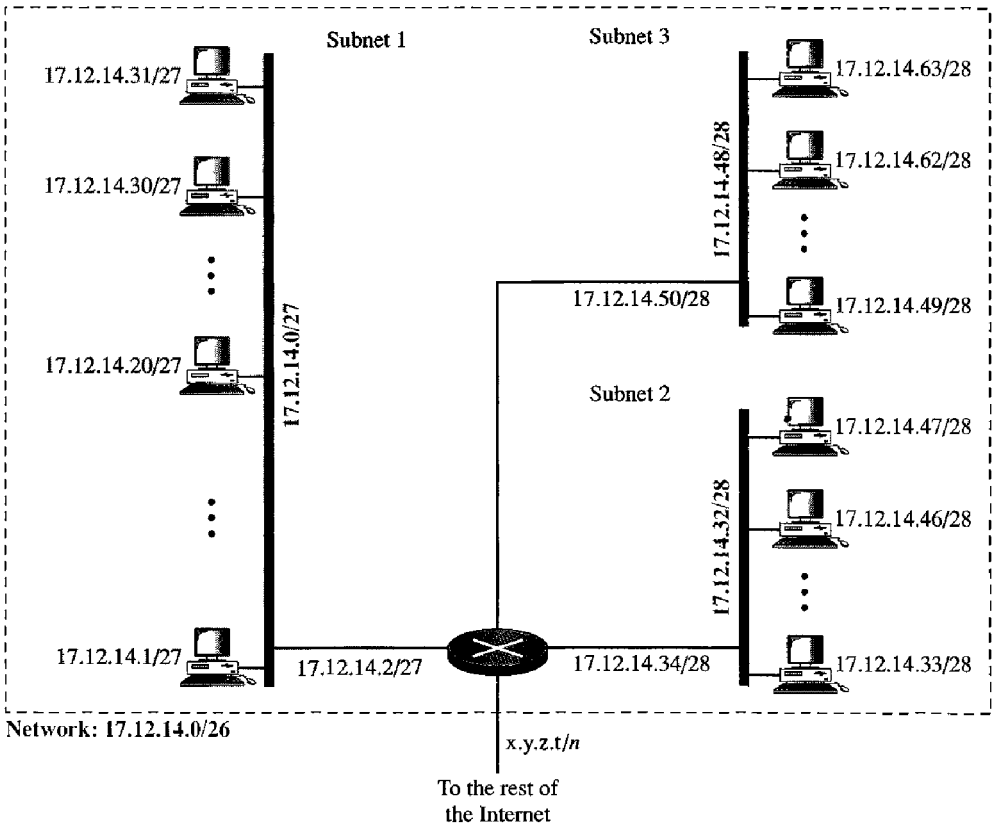
An organization that is granted a large block of addresses may want to create clusters of networks (called subnets) and divide the addresses between the different subnets. The rest of the world still sees the organization as one entity; however, internally there are several subnets. All messages are sent to the router address that connects the organization to the rest of the Internet; the router routes the message to the appropriate subnets. The organization, however, needs to create small subblocks of addresses, each assigned to specific subnets. The organization has its own mask; each subnet must also have its own.

As an example, suppose an organization is given the block 17.12.40.0/26, which contains 64 addresses. The organization has three offices and needs to divide the addresses into three subblocks of 32, 16, and 16 addresses. We can find the new masks by using the following arguments:

1. Suppose the mask for the first subnet is n1, then 2^{32-n1} must be 32, which means that n1 = 27.
2. Suppose the mask for the second subnet is n2, then 2^{32-n2} must be 16, which means that n2 = 28.
3. Suppose the mask for the third subnet is n3, then 2^{32-n3} must be 16, which means that n3 = 28.

This means that we have the masks 27, 28, 28 with the organization mask being 26. Figure 19.7 shows one configuration for the above scenario.

Figure 19.7 *Configuration and addresses in a subnetted network*



Let us check to see if we can find the subnet addresses from one of the addresses in the subnet.

a. In subnet 1, the address 17.12.14.29/27 can give us the subnet address if we use the mask /27 because

Host: 00010001 00001100 00001110 00011101
Mask: /27
Subnet: 00010001 00001100 00001110 00000000 ➡ (17.12.14.0)

b. In subnet 2, the address 17.12.14.45/28 can give us the subnet address if we use the mask /28 because

Host: 00010001 00001100 00001110 00101101
Mask: /28
Subnet: 00010001 00001100 00001110 00100000 ➡ (17.12.14.32)

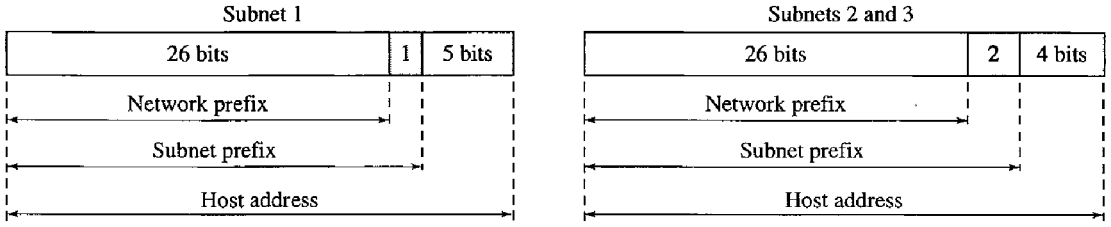
c. In subnet 3, the address 17.12.14.50/28 can give us the subnet address if we use the mask /28 because

Host: 00010001 00001100 00001110 00110010
Mask: /28
Subnet: 00010001 00001100 00001110 00110000 ➡ (17.12.14.48)

Note that applying the mask of the network, /26, to any of the addresses gives us the network address 17.12.14.0/26. We leave this proof to the reader.

We can say that through subnetting, we have three levels of hierarchy. Note that in our example, the subnet prefix length can differ for the subnets as shown in Figure 19.8.

Figure 19.8 *Three-level hierarchy in an IPv4 address*



More Levels of Hierarchy

The structure of classless addressing does not restrict the number of hierarchical levels. An organization can divide the granted block of addresses into subblocks. Each subblock can in turn be divided into smaller subblocks. And so on. One example of this is seen in the ISPs. A national ISP can divide a granted large block into smaller blocks and assign each of them to a regional ISP. A regional ISP can divide the block received from the national ISP into smaller blocks and assign each one to a local ISP. A local ISP can divide the block received from the regional ISP into smaller blocks and assign each one to a different organization. Finally, an organization can divide the received block and make several subnets out of it.

Address Allocation

The next issue in classless addressing is address allocation. How are the blocks allocated? The ultimate responsibility of address allocation is given to a global authority called the *Internet Corporation for Assigned Names and Addresses* (ICANN). However, ICANN does not normally allocate addresses to individual organizations. It assigns a large block of addresses to an ISP. Each ISP, in turn, divides its assigned block into smaller subblocks and grants the subblocks to its customers. In other words, an ISP receives one large block to be distributed to its Internet users. This is called **address aggregation:** many blocks of addresses are aggregated in one block and granted to one ISP.

Example 19.10

An ISP is granted a block of addresses starting with 190.100.0.0/16 (65,536 addresses). The ISP needs to distribute these addresses to three groups of customers as follows:

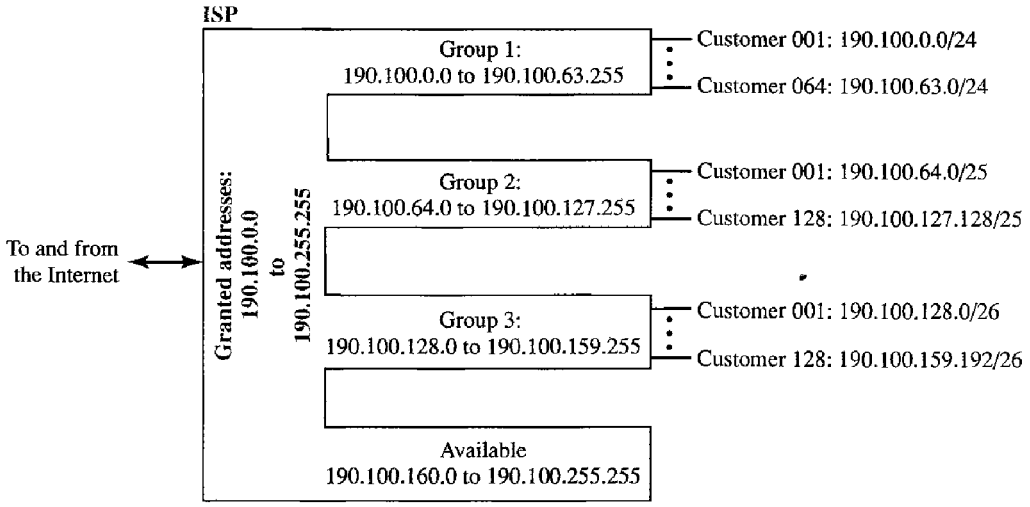
a. The first group has 64 customers; each needs 256 addresses.
b. The second group has 128 customers; each needs 128 addresses.
c. The third group has 128 customers; each needs 64 addresses.

Design the subblocks and find out how many addresses are still available after these allocations.

Solution

Figure 19.9 shows the situation.

Figure 19.9 *An example of address allocation and distribution by an ISP*



1. **Group 1**

   For this group, each customer needs 256 addresses. This means that 8 ($\log_2 256$) bits are needed to define each host. The prefix length is then $32 - 8 = 24$. The addresses are

   1st Customer: 190.100.0.0/24 190.100.0.255/24
   2nd Customer: 190.100.1.0/24 190.100.1.255/24
   . . .
   64th Customer: 190.100.63.0/24 190.100.63.255/24
   Total = 64 × 256 = 16,384

2. **Group 2**

   For this group, each customer needs 128 addresses. This means that 7 ($\log_2 128$) bits are needed to define each host. The prefix length is then $32 - 7 = 25$. The addresses are

   1st Customer: 190.100.64.0/25 190.100.64.127/25
   2nd Customer: 190.100.64.128/25 190.100.64.255/25
   . . .
   128th Customer: 190.100.127.128/25 190.100.127.255/25
   Total = 128 × 128 = 16,384

3. **Group 3**

   For this group, each customer needs 64 addresses. This means that 6 ($\log_2 64$) bits are needed to each host. The prefix length is then $32 - 6 = 26$. The addresses are

1st Customer: *190.100.128.0/26* *190.100.128.63/26*
2nd Customer: *190.100.128.64/26* *190.100.128.127/26*
. . .
128th Customer: *190.100.159.192/26* *190.100.159.255/26*
Total = $128 \times 64 = 8192$

Number of granted addresses to the ISP: 65,536
Number of allocated addresses by the ISP: 40,960
Number of available addresses: 24,576

Network Address Translation (NAT)

The number of home users and small businesses that want to use the Internet is ever increasing. In the beginning, a user was connected to the Internet with a dial-up line, which means that she was connected for a specific period of time. An ISP with a block of addresses could dynamically assign an address to this user. An address was given to a user when it was needed. But the situation is different today. Home users and small businesses can be connected by an ADSL line or cable modem. In addition, many are not happy with one address; many have created small networks with several hosts and need an IP address for each host. With the shortage of addresses, this is a serious problem.

A quick solution to this problem is called **network address translation (NAT).** NAT enables a user to have a large set of addresses internally and one address, or a small set of addresses, externally. The traffic inside can use the large set; the traffic outside, the small set.

To separate the addresses used inside the home or business and the ones used for the Internet, the Internet authorities have reserved three sets of addresses as private addresses, shown in Table 19.3.

Table 19.3 *Addresses for private networks*

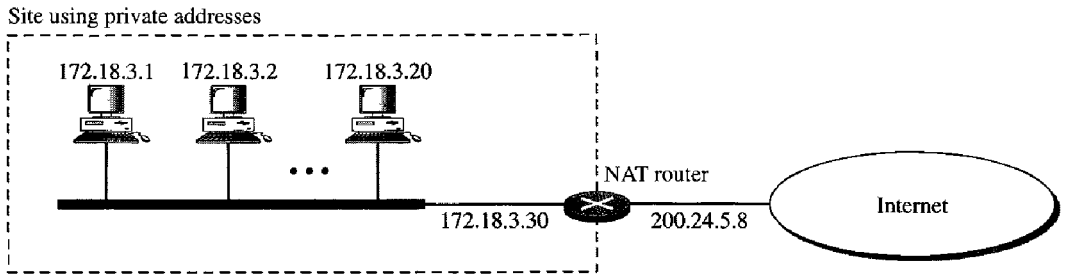
Range			*Total*
10.0.0.0	to	10.255.255.255	2^{24}
172.16.0.0	to	172.31.255.255	2^{20}
192.168.0.0	to	192.168.255.255	2^{16}

Any organization can use an address out of this set without permission from the Internet authorities. Everyone knows that these reserved addresses are for private networks. They are unique inside the organization, but they are not unique globally. No router will forward a packet that has one of these addresses as the destination address.

The site must have only one single connection to the global Internet through a router that runs the NAT software. Figure 19.10 shows a simple implementation of NAT.

As Figure 19.10 shows, the private network uses private addresses. The router that connects the network to the global address uses one private address and one global address. The private network is transparent to the rest of the Internet; the rest of the Internet sees only the NAT router with the address 200.24.5.8.

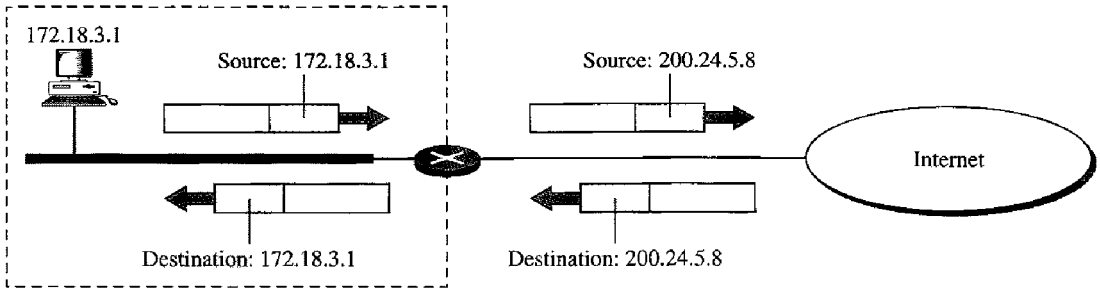
Figure 19.10 *A NAT implementation*



Address Translation

All the outgoing packets go through the NAT router, which replaces the *source address* in the packet with the global NAT address. All incoming packets also pass through the NAT router, which replaces the *destination address* in the packet (the NAT router global address) with the appropriate private address. Figure 19.11 shows an example of address translation.

Figure 19.11 *Addresses in a NAT*

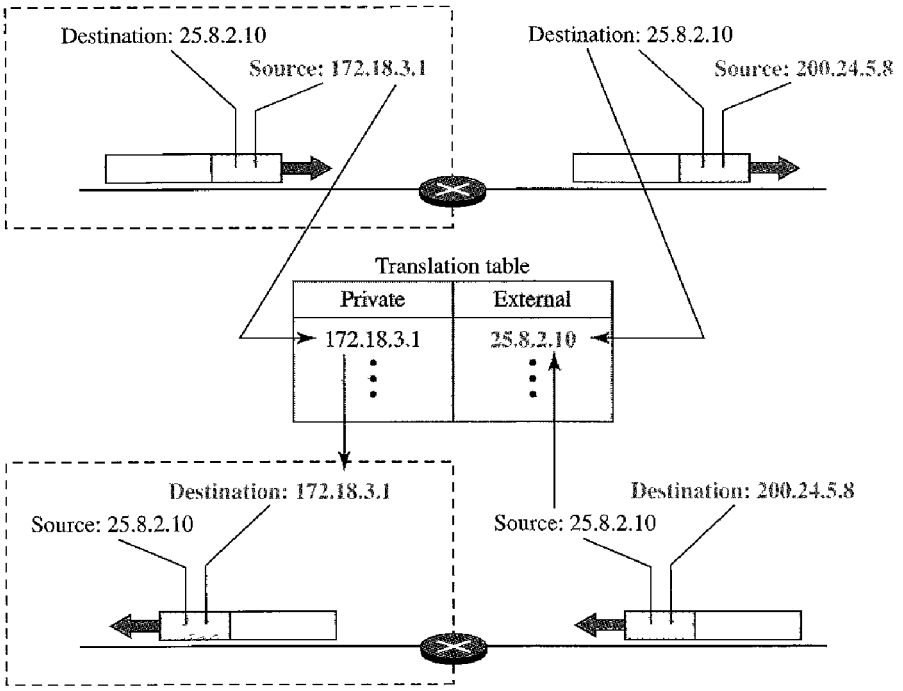


Translation Table

The reader may have noticed that translating the source addresses for outgoing packets is straightforward. But how does the NAT router know the destination address for a packet coming from the Internet? There may be tens or hundreds of private IP addresses, each belonging to one specific host. The problem is solved if the NAT router has a translation table.

Using One IP Address In its simplest form, a translation table has only two columns: the private address and the external address (destination address of the packet). When the router translates the source address of the outgoing packet, it also makes note of the destination address—where the packet is going. When the response comes back from the destination, the router uses the source address of the packet (as the external address) to find the private address of the packet. Figure 19.12 shows the idea. Note that the addresses that are changed (translated) are shown in color.

Figure 19.12 *NAT address translation*



In this strategy, communication must always be initiated by the private network. The NAT mechanism described requires that the private network start the communication. As we will see, NAT is used mostly by ISPs which assign one single address to a customer. The customer, however, may be a member of a private network that has many private addresses. In this case, communication with the Internet is always initiated from the customer site, using a client program such as HTTP, TELNET, or FTP to access the corresponding server program. For example, when e-mail that originates from a non-customer site is received by the ISP e-mail server, the e-mail is stored in the mailbox of the customer until retrieved. A private network cannot run a server program for clients outside of its network if it is using NAT technology.

Using a Pool of IP Addresses Since the NAT router has only one global address, only one private network host can access the same external host. To remove this restriction, the NAT router uses a pool of global addresses. For example, instead of using only one global address (200.24.5.8), the NAT router can use four addresses (200.24.5.8, 200.24.5.9, 200.24.5.10, and 200.24.5.11). In this case, four private network hosts can communicate with the same external host at the same time because each pair of addresses defines a connection. However, there are still some drawbacks. In this example, no more than four connections can be made to the same destination. Also, no private-network host can access two external server programs (e.g., HTTP and FTP) at the same time.

Using Both IP Addresses and Port Numbers To allow a many-to-many relationship between private-network hosts and external server programs, we need more information in the translation table. For example, suppose two hosts with addresses 172.18.3.1 and 172.18.3.2 inside a private network need to access the HTTP server on external host

25.8.3.2. If the translation table has five columns, instead of two, that include the source and destination port numbers of the transport layer protocol, the ambiguity is eliminated. We discuss port numbers in Chapter 23. Table 19.4 shows an example of such a table.

Table 19.4 *Five-column translation table*

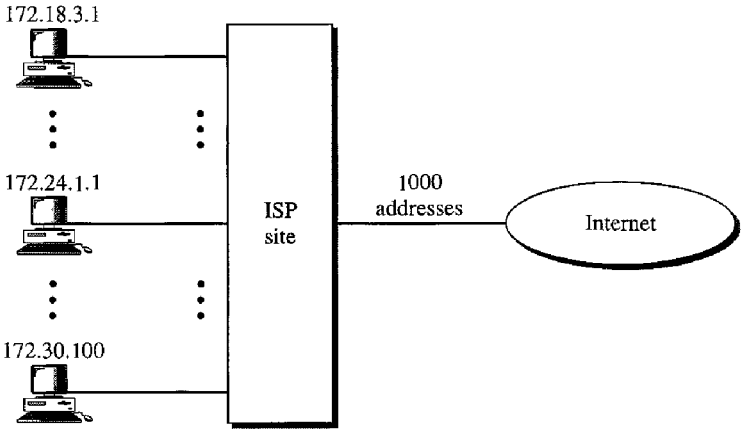
Private Address	*Private Port*	*External Address*	*External Port*	*Transport Protocol*
172.18.3.1	1400	25.8.3.2	80	TCP
172.18.3.2	1401	25.8.3.2	80	TCP
...	...	...	...	...

Note that when the response from HTTP comes back, the combination of source address (25.8.3.2) and destination port number (1400) defines the private network host to which the response should be directed. Note also that for this translation to work, the temporary port numbers (1400 and 1401) must be unique.

NAT and ISP

An ISP that serves dial-up customers can use NAT technology to conserve addresses. For example, suppose an ISP is granted 1000 addresses, but has 100,000 customers. Each of the customers is assigned a private network address. The ISP translates each of the 100,000 source addresses in outgoing packets to one of the 1000 global addresses; it translates the global destination address in incoming packets to the corresponding private address. Figure 19.13 shows this concept.

Figure 19.13 *An ISP and NAT*



19.2 IPv6 ADDRESSES

Despite all short-term solutions, such as classless addressing, Dynamic Host Configuration Protocol (DHCP), discussed in Chapter 21, and NAT, address depletion is still a long-term problem for the Internet. This and other problems in the IP protocol itself,

such as lack of accommodation for real-time audio and video transmission, and encryption and authentication of data for some applications, have been the motivation for IPv6. In this section, we compare the address structure of IPv6 to IPv4. In Chapter 20, we discuss both protocols.

Structure

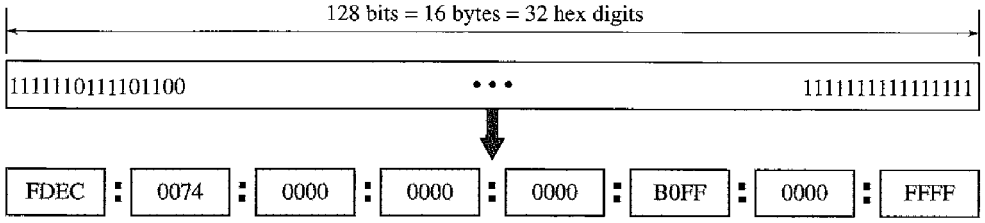
An **IPv6 address** consists of 16 bytes (octets); it is 128 bits long.

An IPv6 address is 128 bits long.

Hexadecimal Colon Notation

To make addresses more readable, IPv6 specifies **hexadecimal colon notation.** In this notation, 128 bits is divided into eight sections, each 2 bytes in length. Two bytes in hexadecimal notation requires four hexadecimal digits. Therefore, the address consists of 32 hexadecimal digits, with every four digits separated by a colon, as shown in Figure 19.14.

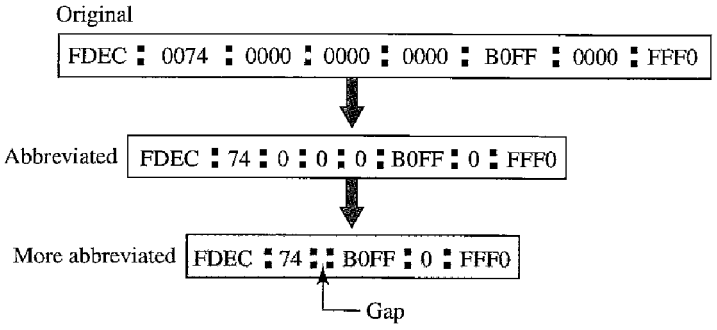
Figure 19.14 *IPv6 address in binary and hexadecimal colon notation*



Abbreviation

Although the IP address, even in hexadecimal format, is very long, many of the digits are zeros. In this case, we can abbreviate the address. The leading zeros of a section (four digits between two colons) can be omitted. Only the leading zeros can be dropped, not the trailing zeros (see Figure 19.15).

Figure 19.15 *Abbreviated IPv6 addresses*



Using this form of **abbreviation,** 0074 can be written as 74, 000F as F, and 0000 as 0. Note that 3210 cannot be abbreviated. Further abbreviations are possible if there are consecutive sections consisting of zeros only. We can remove the zeros altogether and replace them with a double semicolon. Note that this type of abbreviation is allowed only once per address. If there are two runs of zero sections, only one of them can be abbreviated. Reexpansion of the abbreviated address is very simple: Align the unabbreviated portions and insert zeros to get the original expanded address.

Example 19.11

Expand the address 0:15::1:12:1213 to its original.

Solution

We first need to align the left side of the double colon to the left of the original pattern and the right side of the double colon to the right of the original pattern to find how many 0s we need to replace the double colon.

```
XXXX:XXXX:XXXX:XXXX:XXXX:XXXX:XXXX:XXXX
 0:  15:              :   1:  12:1213
```

This means that the original address is

0000:0015:0000:0000:0000:0001:0012:1213

Address Space

IPv6 has a much larger address space; 2^{128} addresses are available. The designers of IPv6 divided the address into several categories. A few leftmost bits, called the *type prefix,* in each address define its category. The type prefix is variable in length, but it is designed such that no code is identical to the first part of any other code. In this way, there is no ambiguity; when an address is given, the type prefix can easily be determined. Table 19.5 shows the prefix for each type of address. The third column shows the fraction of each type of address relative to the whole address space.

Table 19.5 *Type prefixes for IPv6 addresses*

Type Prefix	*Type*	*Fraction*
0000 0000	Reserved	1/256
0000 0001	Unassigned	1/256
0000 001	ISO network addresses	1/128
0000 010	IPX (Novell) network addresses	1/128
0000 011	Unassigned	1/128
0000 1	Unassigned	1/32
0001	Reserved	1/16
001	Reserved	1/8
010	**Provider-based unicast addresses**	**1/8**

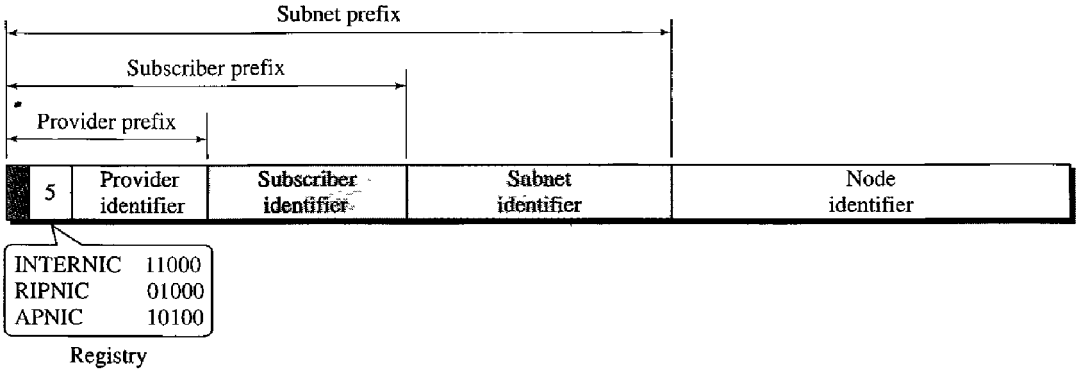
Table 19.5 *Type prefixes for IPv6 addresses (continued)*

Type Prefix	*Type*	*Fraction*
011	Unassigned	1/8
100	Geographic-based unicast addresses	1/8
101	Unassigned	1/8
110	Unassigned	1/8
1110	Unassigned	1/16
1111 0	Unassigned	1/32
1111 10	Unassigned	1/64
1111 110	Unassigned	1/128
1111 1110 0	Unassigned	1/512
1111 1110 10	Link local addresses	1/1024
1111 1110 11	Site local addresses	1/1024
1111 1111	Multicast addresses	1/256

Unicast Addresses

A **unicast address** defines a single computer. The packet sent to a unicast address must be delivered to that specific computer. IPv6 defines two types of unicast addresses: geographically based and provider-based. We discuss the second type here; the first type is left for future definition. The provider-based address is generally used by a normal host as a unicast address. The address format is shown in Figure 19.16.

Figure 19.16 *Prefixes for provider-based unicast address*



Fields for the provider-based address are as follows:

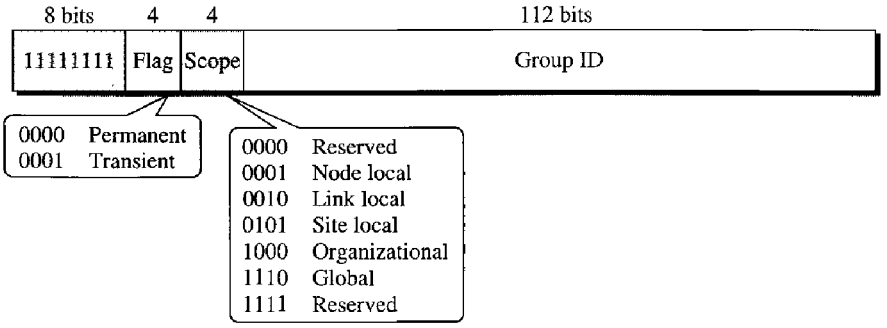
- **Type identifier.** This 3-bit field defines the address as a provider-based address.
- **Registry identifier.** This 5-bit field indicates the agency that has registered the address. Currently three registry centers have been defined. INTERNIC (code 11000) is the center for North America; RIPNIC (code 01000) is the center for European registration; and APNIC (code 10100) is for Asian and Pacific countries.

- **Provider identifier.** This variable-length field identifies the provider for Internet access (such as an ISP). A 16-bit length is recommended for this field.
- **Subscriber identifier.** When an organization subscribes to the Internet through a provider, it is assigned a subscriber identification. A 24-bit length is recommended for this field.
- **Subnet identifier.** Each subscriber can have many different subnetworks, and each subnetwork can have an identifier. The subnet identifier defines a specific subnetwork under the territory of the subscriber. A 32-bit length is recommended for this field.
- **Node identifier.** The last field defines the identity of the node connected to a subnet. A length of 48 bits is recommended for this field to make it compatible with the 48-bit link (physical) address used by Ethernet. In the future, this link address will probably be the same as the node physical address.

Multicast Addresses

Multicast addresses are used to define a group of hosts instead of just one. A packet sent to a **multicast address** must be delivered to each member of the group. Figure 19.17 shows the format of a multicast address.

Figure 19.17 *Multicast address in IPv6*



The second field is a flag that defines the group address as either permanent or transient. A permanent group address is defined by the Internet authorities and can be accessed at all times. A transient group address, on the other hand, is used only temporarily. Systems engaged in a teleconference, for example, can use a transient group address. The third field defines the scope of the group address. Many different scopes have been defined, as shown in Figure 19.17.

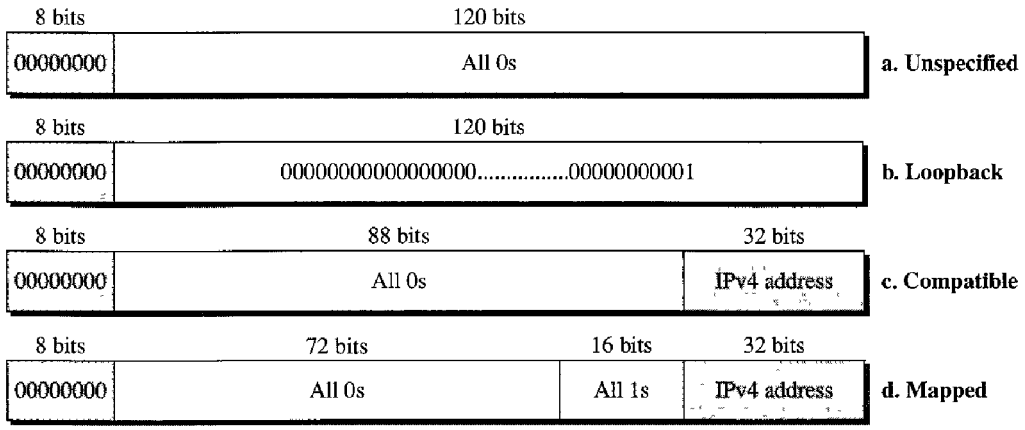
Anycast Addresses

IPv6 also defines anycast addresses. An **anycast address,** like a multicast address, also defines a group of nodes. However, a packet destined for an anycast address is delivered to only one of the members of the anycast group, the nearest one (the one with the shortest route). Although the definition of an anycast address is still debatable, one possible use is to assign an anycast address to all routers of an ISP that covers a large logical area in the Internet. The routers outside the ISP deliver a packet destined for the ISP to the nearest ISP router. No block is assigned for anycast addresses.

Reserved Addresses

Another category in the address space is the **reserved address.** These addresses start with eight 0s (type prefix is 0000 0000). A few subcategories are defined in this category, as shown in Figure 19.18.

Figure 19.18 *Reserved addresses in IPv6*

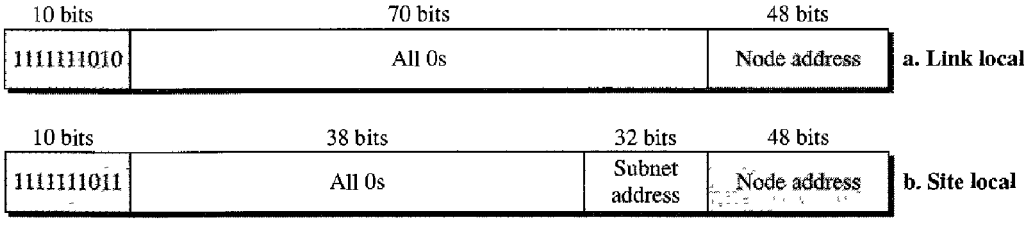


An **unspecified address** is used when a host does not know its own address and sends an inquiry to find its address. A **loopback address** is used by a host to test itself without going into the network. A **compatible address** is used during the transition from IPv4 to IPv6 (see Chapter 20). It is used when a computer using IPv6 wants to send a message to another computer using IPv6, but the message needs to pass through a part of the network that still operates in IPv4. A **mapped address** is also used during transition. However, it is used when a computer that has migrated to IPv6 wants to send a packet to a computer still using IPv4.

Local Addresses

These addresses are used when an organization wants to use IPv6 protocol without being connected to the global Internet. In other words, they provide addressing for private networks. Nobody outside the organization can send a message to the nodes using these addresses. Two types of addresses are defined for this purpose, as shown in Figure 19.19.

Figure 19.19 *Local addresses in IPv6*



A **link local address** is used in an isolated subnet; a **site local address** is used in an isolated site with several subnets.

19.3 RECOMMENDED READING

For more details about subjects discussed in this chapter, we recommend the following books and sites. The items in brackets [. . .] refer to the reference list at the end of the text.

Books

IPv4 addresses are discussed in Chapters 4 and 5 of [For06], Chapter 3 of [Ste94], Section 4.1 of [PD03], Chapter 18 of [Sta04], and Section 5.6 of [Tan03]. IPv6 addresses are discussed in Section 27.1 of [For06] and Chapter 8 of [Los04]. A good discussion of NAT can be found in [Dut01].

Sites

- www.ietf.org/rfc.html Information about RFCs

RFCs

A discussion of IPv4 addresses can be found in most of the RFCs related to the IPv4 protocol:

760, 781, 791, 815, 1025, 1063, 1071, 1141, 1190, 1191, 1624, 2113

A discussion of IPv6 addresses can be found in most of the RFCs related to IPv6 protocol:

1365, 1550, 1678, 1680, 1682, 1683, 1686, 1688, 1726, 1752, 1826, 1883, 1884, 1886, 1887, 1955, 2080, 2373, 2452, 2463, 2465, 2466, 2472, 2492, 2545, 2590

A discussion of NAT can be found in

1361, 2663, 2694

19.4 KEY TERMS

address aggregation
address space
anycast address
binary notation
class A address
class B address
class C address
class D address
class E address
classful addressing
classless addressing
classless interdomain routing (CIDR)

compatible address
dotted-decimal notation
default mask
hexadecimal colon notation
hostid
IP address
IPv4 address
IPv6 address
link local address
mapped address
mask
multicast address
netid
network address
network address translation (NAT)
prefix
reserved address
site local address
subnet
subnet mask
subnetting
suffix
supernet
supernet mask
supernetting
unicast address
unspecified address

19.5 SUMMARY

- At the network layer, a global identification system that uniquely identifies every host and router is necessary for delivery of a packet from host to host.
- An IPv4 address is 32 bits long and uniquely and universally defines a host or router on the Internet.
- In classful addressing, the portion of the IP address that identifies the network is called the netid.
- In classful addressing, the portion of the IP address that identifies the host or router on the network is called the hostid.
- An IP address defines a device's connection to a network.
- There are five classes in IPv4 addresses. Classes A, B, and C differ in the number of hosts allowed per network. Class D is for multicasting and Class E is reserved.
- The class of an address is easily determined by examination of the first byte.
- Addresses in classes A, B, or C are mostly used for unicast communication.
- Addresses in class D are used for multicast communication.
- Subnetting divides one large network into several smaller ones, adding an intermediate level of hierarchy in IP addressing.
- Supernetting combines several networks into one large one.
- In classless addressing, we can divide the address space into variable-length blocks.
- There are three restrictions in classless addressing:
  a. The number of addresses needs to be a power of 2.
  b. The mask needs to be included in the address to define the block.
  c. The starting address must be divisible by the number of addresses in the block.
- The mask in classless addressing is expressed as the prefix length (/*n*) in CIDR notation.

- To find the first address in a block, we set the rightmost $32 - n$ bits to 0.
- To find the number of addresses in the block, we calculate 2^{32-n}, where n is the prefix length.
- To find the last address in the block, we set the rightmost $32 - n$ bits to 0.
- Subnetting increases the value of n.
- The global authority for address allocation is ICANN. ICANN normally grants large blocks of addresses to ISPs, which in turn grant small subblocks to individual customers.
- IPv6 addresses use hexadecimal colon notation with abbreviation methods available.
- There are three types of addresses in IPv6: unicast, anycast, and multicast.
- In an IPv6 address, the variable type prefix field defines the address type or purpose.

19.6 PRACTICE SET

Review Questions

1. What is the number of bits in an IPv4 address? What is the number of bits in an IPv6 address?
2. What is dotted decimal notation in IPv4 addressing? What is the number of bytes in an IPv4 address represented in dotted decimal notation? What is hexadecimal notation in IPv6 addressing? What is the number of digits in an IPv6 address represented in hexadecimal notation?
3. What are the differences between classful addressing and classless addressing in IPv4?
4. List the classes in classful addressing and define the application of each class (unicast, multicast, broadcast, or reserve).
5. Explain why most of the addresses in class A are wasted. Explain why a medium-size or large-size corporation does not want a block of class C addresses.
6. What is a mask in IPv4 addressing? What is a default mask in IPv4 addressing?
7. What is the network address in a block of addresses? How can we find the network address if one of the addresses in a block is given?
8. Briefly define subnetting and supernetting. How do the subnet mask and supernet mask differ from a default mask in classful addressing?
9. How can we distinguish a multicast address in IPv4 addressing? How can we do so in IPv6 addressing?
10. What is NAT? How can NAT help in address depletion?

Exercises

11. What is the address space in each of the following systems?
    a. A system with 8-bit addresses
    b. A system with 16-bit addresses
    c. A system with 64-bit addresses

12. An address space has a total of 1024 addresses. How many bits are needed to represent an address?
13. An address space uses the three symbols 0, 1, and 2 to represent addresses. If each address is made of 10 symbols, how many addresses are available in this system?
14. Change the following IP addresses from dotted-decimal notation to binary notation.
    a. 114.34.2.8
    b. 129.14.6.8
    c. 208.34.54.12
    d. 238.34.2.1
15. Change the following IP addresses from binary notation to dotted-decimal notation.
    a. 01111111 11110000 01100111 01111101
    b. 10101111 11000000 11111000 00011101
    c. 11011111 10110000 00011111 01011101
    d. 11101111 11110111 11000111 00011101
16. Find the class of the following IP addresses.
    a. 208.34.54.12
    b. 238.34.2.1
    c. 114.34.2.8
    d. 129.14.6.8
17. Find the class of the following IP addresses.
    a. 11110111 11110011 10000111 11011101
    b. 10101111 11000000 11110000 00011101
    c. 11011111 10110000 00011111 01011101
    d. 11101111 11110111 11000111 00011101
18. Find the netid and the hostid of the following IP addresses.
    a. 114.34.2.8
    b. 132.56.8.6
    c. 208.34.54.12
19. In a block of addresses, we know the IP address of one host is 25.34.12.56/16. What are the first address (network address) and the last address (limited broadcast address) in this block?
20. In a block of addresses, we know the IP address of one host is 182.44.82.16/26. What are the first address (network address) and the last address in this block?
21. An organization is granted the block 16.0.0.0/8. The administrator wants to create 500 fixed-length subnets.
    a. Find the subnet mask.
    b. Find the number of addresses in each subnet.
    c. Find the first and last addresses in subnet 1.
    d. Find the first and last addresses in subnet 500.

22. An organization is granted the block 130.56.0.0/16. The administrator wants to create 1024 subnets.
    a. Find the subnet mask.
    b. Find the number of addresses in each subnet.
    c. Find the first and last addresses in subnet 1.
    d. Find the first and last addresses in subnet 1024.
23. An organization is granted the block 211.17.180.0/24. The administrator wants to create 32 subnets.
    a. Find the subnet mask.
    b. Find the number of addresses in each subnet.
    c. Find the first and last addresses in subnet 1.
    d. Find the first and last addresses in subnet 32.
24. Write the following masks in slash notation (*/n*).
    a. 255.255.255.0
    b. 255.0.0.0
    c. 255.255.224.0
    d. 255.255.240.0
25. Find the range of addresses in the following blocks.
    a. 123.56.77.32/29
    b. 200.17.21.128/27
    c. 17.34.16.0/23
    d. 180.34.64.64/30
26. An ISP is granted a block of addresses starting with 150.80.0.0/16. The ISP wants to distribute these blocks to 2600 customers as follows.
    a. The first group has 200 medium-size businesses; each needs 128 addresses.
    b. The second group has 400 small businesses; each needs 16 addresses.
    c. The third group has 2000 households; each needs 4 addresses.

    Design the subblocks and give the slash notation for each subblock. Find out how many addresses are still available after these allocations.
27. An ISP is granted a block of addresses starting with 120.60.4.0/22. The ISP wants to distribute these blocks to 100 organizations with each organization receiving just eight addresses. Design the subblocks and give the slash notation for each subblock. Find out how many addresses are still available after these allocations.
28. An ISP has a block of 1024 addresses. It needs to divide the addresses among 1024 customers. Does it need subnetting? Explain your answer.
29. Show the shortest form of the following addresses.
    a. 2340:1ABC:119A:A000:0000:0000:0000:0000
    b. 0000:00AA:0000:0000:0000:0000:119A:A231
    c. 2340:0000:0000:0000:0000:119A:A001:0000
    d. 0000:0000:0000:2340:0000:0000:0000:0000

30. Show the original (unabbreviated) form of the following addresses.
    a. 0::0
    b. 0:AA::0
    c. 0:1234::3
    d. 123::1:2
31. What is the type of each of the following addresses?
    a. FE80::12
    b. FEC0::24A2
    c. FF02::0
    d. 0::01
32. What is the type of each of the following addresses?
    a. 0::0
    b. 0::FFFF:0:0
    c. 582F:1234::2222
    d. 4821::14:22
    e. 54EF::A234:2
33. Show the provider prefix (in hexadecimal colon notation) of an address assigned to a subscriber if it is registered in the United States with ABC1 as the provider identification.
34. Show in hexadecimal colon notation the IPv6 address
    a. Compatible to the IPv4 address 129.6.12.34
    b. Mapped to the IPv4 address 129.6.12.34
35. Show in hexadecimal colon notation
    a. The link local address in which the node identifier is 0::123/48
    b. The site local address in which the node identifier is 0::123/48
36. Show in hexadecimal colon notation the permanent multicast address used in a link local scope.
37. A host has the address 581E:1456:2314:ABCD::1211. If the node identification is 48 bits, find the address of the subnet to which the host is attached.
38. A site with 200 subnets has the class B address of 132.45.0.0. The site recently migrated to IPv6 with the subscriber prefix 581E:1456:2314::ABCD/80. Design the subnets and define the subnet addresses, using a subnet identifier of 32 bits.

Research Activities

39. Find the block of addresses assigned to your organization or institution.
40. If you are using an ISP to connect from your home to the Internet, find the name of the ISP and the block of addresses assigned to it.
41. Some people argue that we can consider the whole address space as one single block in which each range of addresses is a subblock to this single block. Elaborate on this idea. What happens to subnetting if we accept this concept?
42. Is your school or organization using a classful address? If so, find out the class of the address.

CHAPTER 20

Network Layer: Internet Protocol

In the Internet model, the main network protocol is the **Internet Protocol (IP).** In this chapter, we first discuss internetworking and issues related to the network layer protocol in general.

We then discuss the current version of the Internet Protocol, version 4, or IPv4. This leads us to the next generation of this protocol, or IPv6, which may become the dominant protocol in the near future.

Finally, we discuss the transition strategies from IPv4 to IPv6. Some readers may note the absence of IPv5. IPv5 is an experimental protocol, based mostly on the OSI model that never materialized.

20.1 INTERNETWORKING

The physical and data link layers of a network operate locally. These two layers are jointly responsible for data delivery on the network from one node to the next, as shown in Figure 20.1.

This internetwork is made of five networks: four LANs and one WAN. If host A needs to send a data packet to host D, the packet needs to go first from A to R1 (a switch or router), then from R1 to R3, and finally from R3 to host D. We say that the data packet passes through three links. In each link, two physical and two data link layers are involved.

However, there is a big problem here. When data arrive at interface f1 of R1, how does R1 know that interface f3 is the outgoing interface? There is no provision in the data link (or physical) layer to help R1 make the right decision. The frame does not carry any routing information either. The frame contains the MAC address of A as the source and the MAC address of R1 as the destination. For a LAN or a WAN, delivery means carrying the frame through one link, and not beyond.

Need for Network Layer

To solve the problem of delivery through several links, the network layer (or the internetwork layer, as it is sometimes called) was designed. The network layer is responsible for host-to-host delivery and for routing the packets through the routers or switches. Figure 20.2 shows the same internetwork with a network layer added.

Figure 20.1 *Links between two hosts*

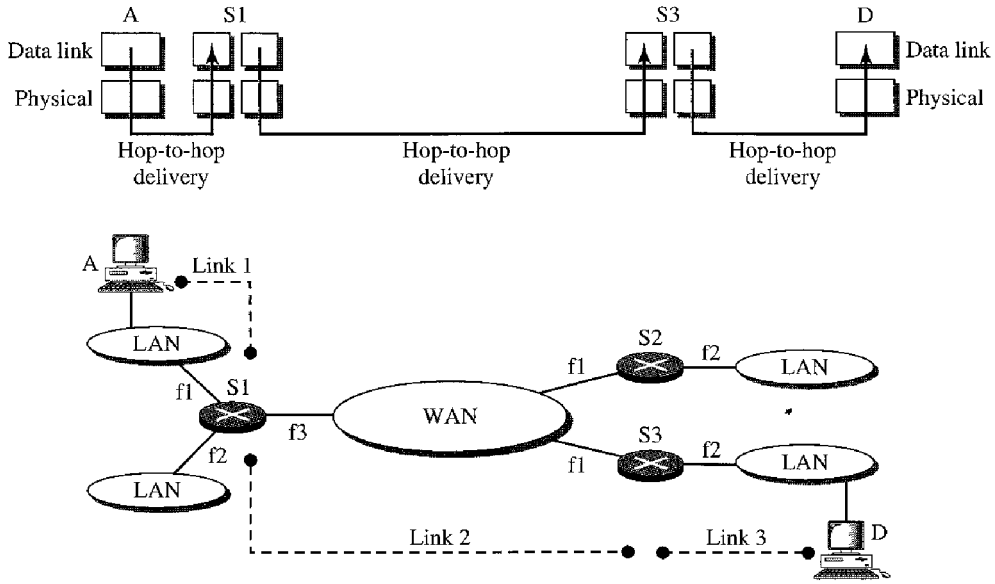


Figure 20.2 *Network layer in an internetwork*

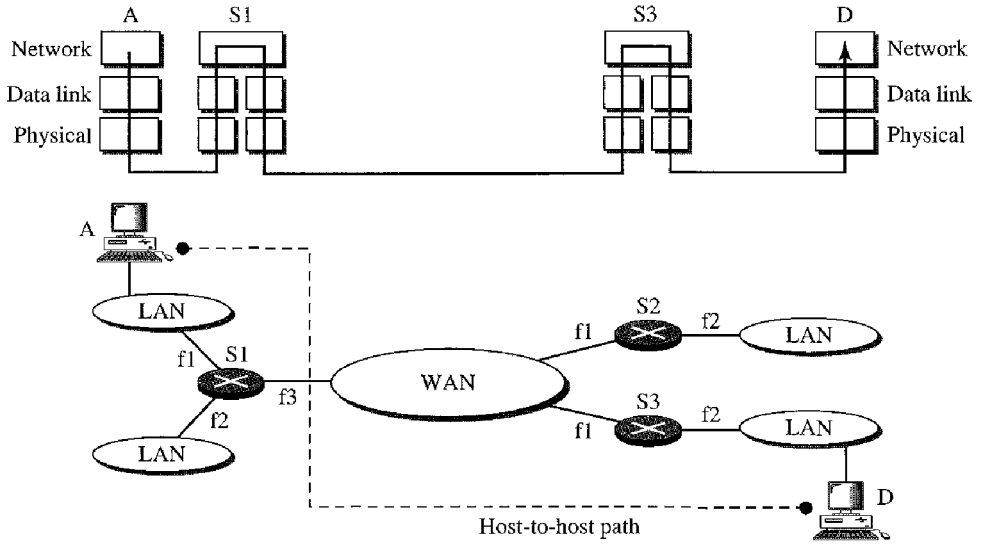
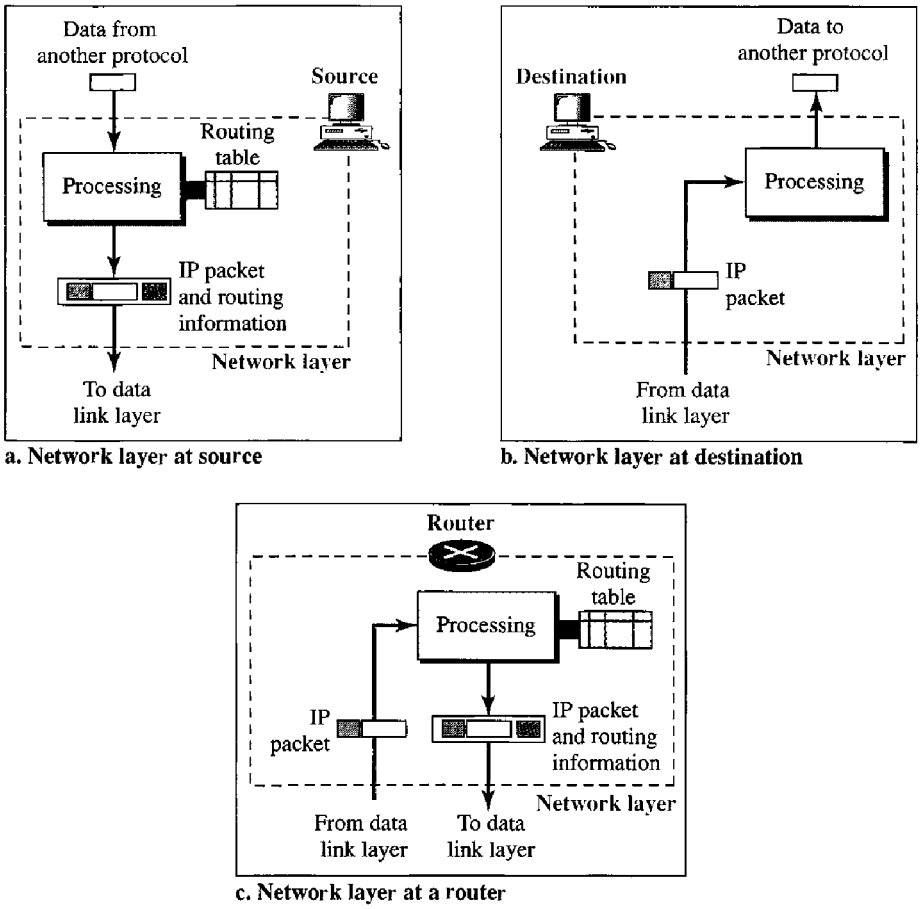


Figure 20.3 shows the general idea of the functionality of the network layer at a source, at a router, and at the destination. The network layer at the source is responsible for creating a packet from the data coming from another protocol (such as a transport layer protocol or a routing protocol). The header of the packet contains, among other information, the logical addresses of the source and destination. The network layer is responsible for checking its routing table to find the routing information (such as the outgoing interface of the packet or the physical address of the next node). If the packet is too large, the packet is fragmented (fragmentation is discussed later in this chapter).

Figure 20.3 *Network layer at the source, router, and destination*



The network layer at the switch or router is responsible for routing the packet. When a packet arrives, the router or switch consults its routing table and finds the interface from which the packet must be sent. The packet, after some changes in the header, with the routing information is passed to the data link layer again.

The network layer at the destination is responsible for address verification; it makes sure that the destination address on the packet is the same as the address of the host. If the packet is a fragment, the network layer waits until all fragments have arrived, and then reassembles them and delivers the reassembled packet to the transport layer.

Internet as a Datagram Network

The Internet, at the network layer, is a packet-switched network. We discussed switching in Chapter 8. We said that, in general, switching can be divided into three broad categories: circuit switching, packet switching, and message switching. Packet switching uses either the virtual circuit approach or the datagram approach.

The Internet has chosen the datagram approach to switching in the network layer. It uses the universal addresses defined in the network layer to route packets from the source to the destination.

Switching at the network layer in the Internet uses the datagram approach to packet switching.

Internet as a Connectionless Network

Delivery of a packet can be accomplished by using either a connection-oriented or a connectionless network service. In a **connection-oriented service,** the source first makes a connection with the destination before sending a packet. When the connection is established, a sequence of packets from the same source to the same destination can be sent one after another. In this case, there is a relationship between packets. They are sent on the same path in sequential order. A packet is logically connected to the packet traveling before it and to the packet traveling after it. When all packets of a message have been delivered, the connection is terminated.

In a connection-oriented protocol, the decision about the route of a sequence of packets with the same source and destination addresses can be made only once, when the connection is established. Switches do not recalculate the route for each individual packet. This type of service is used in a virtual-circuit approach to packet switching such as in Frame Relay and ATM.

In **connectionless service,** the network layer protocol treats each packet independently, with each packet having no relationship to any other packet. The packets in a message may or may not travel the same path to their destination. This type of service is used in the datagram approach to packet switching. The Internet has chosen this type of service at the network layer.

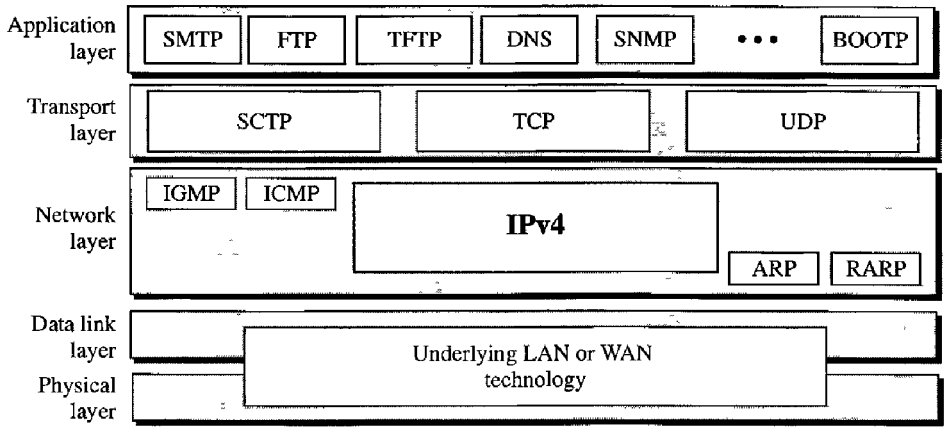
The reason for this decision is that the Internet is made of so many heterogeneous networks that it is almost impossible to create a connection from the source to the destination without knowing the nature of the networks in advance.

Communication at the network layer in the Internet is connectionless.

20.2 IPv4

The **Internet Protocol version 4 (IPv4)** is the delivery mechanism used by the TCP/IP protocols. Figure 20.4 shows the position of IPv4 in the suite.

Figure 20.4 *Position of IPv4 in TCP/IP protocol suite*



IPv4 is an unreliable and connectionless datagram protocol—a **best-effort delivery** service. The term *best-effort* means that IPv4 provides no error control or flow control (except for error detection on the header). IPv4 assumes the unreliability of the underlying layers and does its best to get a transmission through to its destination, but with no guarantees.

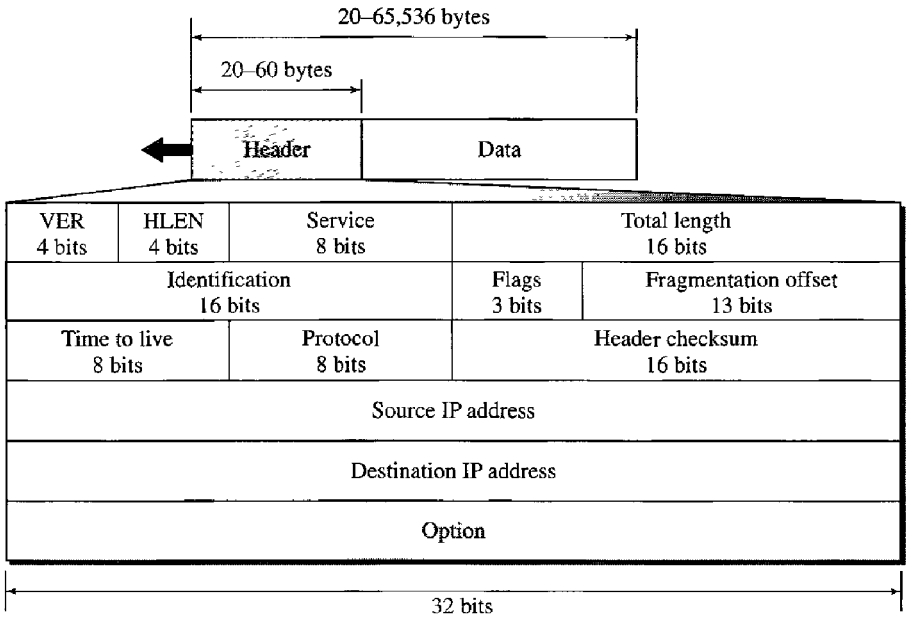
If reliability is important, IPv4 must be paired with a reliable protocol such as TCP. An example of a more commonly understood best-effort delivery service is the post office. The post office does its best to deliver the mail but does not always succeed. If an unregistered letter is lost, it is up to the sender or would-be recipient to discover the loss and rectify the problem. The post office itself does not keep track of every letter and cannot notify a sender of loss or damage.

IPv4 is also a connectionless protocol for a packet-switching network that uses the datagram approach (see Chapter 8). This means that each datagram is handled independently, and each datagram can follow a different route to the destination. This implies that datagrams sent by the same source to the same destination could arrive out of order. Also, some could be lost or corrupted during transmission. Again, IPv4 relies on a higher-level protocol to take care of all these problems.

Datagram

Packets in the IPv4 layer are called **datagrams.** Figure 20.5 shows the IPv4 datagram format.

Figure 20.5 *IPv4 datagram format*

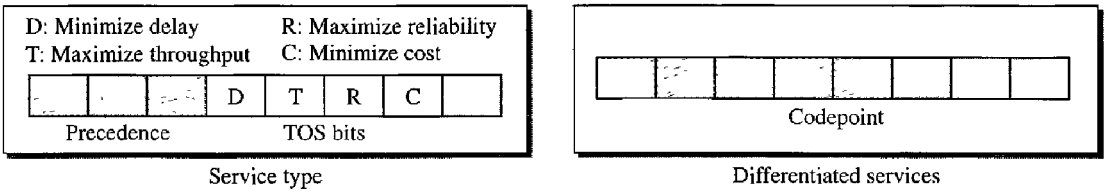


A datagram is a variable-length packet consisting of two parts: header and data. The header is 20 to 60 bytes in length and contains information essential to routing and

delivery. It is customary in TCP/IP to show the header in 4-byte sections. A brief description of each field is in order.

- **Version (VER).** This 4-bit field defines the version of the IPv4 protocol. Currently the version is 4. However, version 6 (or IPng) may totally replace version 4 in the future. This field tells the IPv4 software running in the processing machine that the datagram has the format of version 4. All fields must be interpreted as specified in the fourth version of the protocol. If the machine is using some other version of IPv4, the datagram is discarded rather than interpreted incorrectly.
- **Header length (HLEN).** This 4-bit field defines the total length of the datagram header in 4-byte words. This field is needed because the length of the header is variable (between 20 and 60 bytes). When there are no options, the header length is 20 bytes, and the value of this field is 5 ($5 \times 4 = 20$). When the option field is at its maximum size, the value of this field is 15 ($15 \times 4 = 60$).
- **Services.** IETF has changed the interpretation and name of this 8-bit field. This field, previously called **service type,** is now called **differentiated services.** We show both interpretations in Figure 20.6.

Figure 20.6 *Service type or differentiated services*



1. **Service Type**
   In this interpretation, the first 3 bits are called precedence bits. The next 4 bits are called **type of service (TOS)** bits, and the last bit is not used.
   a. **Precedence** is a 3-bit subfield ranging from 0 (000 in binary) to 7 (111 in binary). The precedence defines the priority of the datagram in issues such as congestion. If a router is congested and needs to discard some datagrams, those datagrams with lowest precedence are discarded first. Some datagrams in the Internet are more important than others. For example, a datagram used for network management is much more urgent and important than a datagram containing optional information for a group.

The precedence subfield was part of version 4, but never used.

   b. **TOS bits** is a 4-bit subfield with each bit having a special meaning. Although a bit can be either 0 or 1, one and only one of the bits can have the value of 1 in each datagram. The bit patterns and their interpretations are given in Table 20.1. With only 1 bit set at a time, we can have five different types of services.

Table 20.1 *Types of service*

TOS Bits	*Description*
0000	Normal (default)
0001	Minimize cost
0010	Maximize reliability
0100	Maximize throughput
1000	Minimize delay

Application programs can request a specific type of service. The defaults for some applications are shown in Table 20.2.

Table 20.2 *Default types of service*

Protocol	*TOS Bits*	*Description*
ICMP	0000	Normal
BOOTP	0000	Normal
NNTP	0001	Minimize cost
IGP	0010	Maximize reliability
SNMP	0010	Maximize reliability
TELNET	1000	Minimize delay
FTP (data)	0100	Maximize throughput
FTP (control)	1000	Minimize delay
TFTP	1000	Minimize delay
SMTP (command)	1000	Minimize delay
SMTP (data)	0100	Maximize throughput
DNS (UDP query)	1000	Minimize delay
DNS (TCP query)	0000	Normal
DNS (zone)	0100	Maximize throughput

It is clear from Table 20.2 that interactive activities, activities requiring immediate attention, and activities requiring immediate response need minimum delay. Those activities that send bulk data require maximum throughput. Management activities need maximum reliability. Background activities need minimum cost.

2. **Differentiated Services**
   In this interpretation, the first 6 bits make up the **codepoint** subfield, and the last 2 bits are not used. The codepoint subfield can be used in two different ways.
   a. When the 3 rightmost bits are 0s, the 3 leftmost bits are interpreted the same as the precedence bits in the service type interpretation. In other words, it is compatible with the old interpretation.

b. When the 3 rightmost bits are not all 0s, the 6 bits define 64 services based on the priority assignment by the Internet or local authorities according to Table 20.3. The first category contains 32 service types; the second and the third each contain 16. The first category (numbers 0, 2, 4, . . . , 62) is assigned by the Internet authorities (IETF). The second category (3, 7, 11, 15, . . . , 63) can be used by local authorities (organizations). The third category (1, 5, 9, . . . , 61) is temporary and can be used for experimental purposes. Note that the numbers are not contiguous. If they were, the first category would range from 0 to 31, the second from 32 to 47, and the third from 48 to 63. This would be incompatible with the TOS interpretation because XXX000 (which includes 0, 8, 16, 24, 32, 40, 48, and 56) would fall into all three categories. Instead, in this assignment method all these services belong to category 1. Note that these assignments have not yet been finalized.

Table 20.3 *Values for codepoints*

Category	*Codepoint*	*Assigning Authority*
1	XXXXX0	Internet
2	XXXX11	Local
3	XXXX01	Temporary or experimental

❑ **Total length.** This is a 16-bit field that defines the total length (header plus data) of the IPv4 datagram in bytes. To find the length of the data coming from the upper layer, subtract the header length from the total length. The header length can be found by multiplying the value in the HLEN field by 4.

$$\text{Length of data} = \text{total length} - \text{header length}$$

Since the field length is 16 bits, the total length of the IPv4 datagram is limited to 65,535 ($2^{16} - 1$) bytes, of which 20 to 60 bytes are the header and the rest is data from the upper layer.

The total length field defines the total length of the datagram including the header.

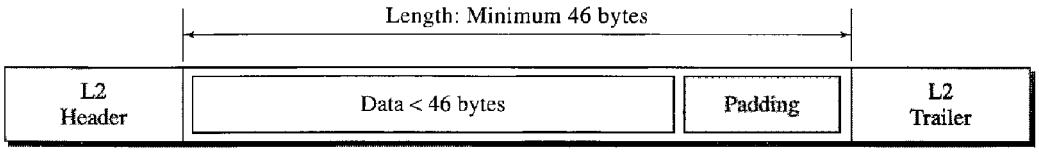
Though a size of 65,535 bytes might seem large, the size of the IPv4 datagram may increase in the near future as the underlying technologies allow even more throughput (greater bandwidth).

When we discuss fragmentation in the next section, we will see that some physical networks are not able to encapsulate a datagram of 65,535 bytes in their frames. The datagram must be fragmented to be able to pass through those networks.

One may ask why we need this field anyway. When a machine (router or host) receives a frame, it drops the header and the trailer, leaving the datagram. Why include an extra field that is not needed? The answer is that in many cases we really do not need the value in this field. However, there are occasions in which the

datagram is not the only thing encapsulated in a frame; it may be that padding has been added. For example, the Ethernet protocol has a minimum and maximum restriction on the size of data that can be encapsulated in a frame (46 to 1500 bytes). If the size of an IPv4 datagram is less than 46 bytes, some padding will be added to meet this requirement. In this case, when a machine decapsulates the datagram, it needs to check the total length field to determine how much is really data and how much is padding (see Figure 20.7).

Figure 20.7 *Encapsulation of a small datagram in an Ethernet frame*

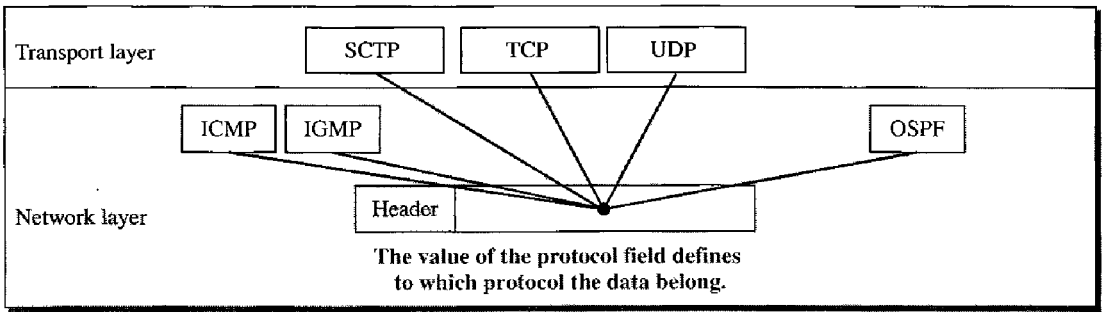


- **Identification.** This field is used in fragmentation (discussed in the next section).
- **Flags.** This field is used in fragmentation (discussed in the next section).
- **Fragmentation offset.** This field is used in fragmentation (discussed in the next section).
- **Time to live.** A datagram has a limited lifetime in its travel through an internet. This field was originally designed to hold a timestamp, which was decremented by each visited router. The datagram was discarded when the value became zero. However, for this scheme, all the machines must have synchronized clocks and must know how long it takes for a datagram to go from one machine to another. Today, this field is used mostly to control the maximum number of hops (routers) visited by the datagram. When a source host sends the datagram, it stores a number in this field. This value is approximately 2 times the maximum number of routes between any two hosts. Each router that processes the datagram decrements this number by 1. If this value, after being decremented, is zero, the router discards the datagram.

  This field is needed because routing tables in the Internet can become corrupted. A datagram may travel between two or more routers for a long time without ever getting delivered to the destination host. This field limits the lifetime of a datagram.

  Another use of this field is to intentionally limit the journey of the packet. For example, if the source wants to confine the packet to the local network, it can store 1 in this field. When the packet arrives at the first router, this value is decremented to 0, and the datagram is discarded.
- **Protocol.** This 8-bit field defines the higher-level protocol that uses the services of the IPv4 layer. An IPv4 datagram can encapsulate data from several higher-level protocols such as TCP, UDP, ICMP, and IGMP. This field specifies the final destination protocol to which the IPv4 datagram is delivered. In other words, since the IPv4 protocol carries data from different other protocols, the value of this field helps the receiving network layer know to which protocol the data belong (see Figure 20.8).

Figure 20.8 *Protocol field and encapsulated data*



The value of this field for each higher-level protocol is shown in Table 20.4.

Table 20.4 *Protocol values*

Value	*Protocol*
1	ICMP
2	IGMP
6	TCP
17	UDP
89	OSPF

- **Checksum.** The checksum concept and its calculation are discussed later in this chapter.
- **Source address.** This 32-bit field defines the IPv4 address of the source. This field must remain unchanged during the time the IPv4 datagram travels from the source host to the destination host.
- **Destination address.** This 32-bit field defines the IPv4 address of the destination. This field must remain unchanged during the time the IPv4 datagram travels from the source host to the destination host.

Example 20.1

An IPv4 packet has arrived with the first 8 bits as shown:

01000010

The receiver discards the packet. Why?

Solution

There is an error in this packet. The 4 leftmost bits (0100) show the version, which is correct. The next 4 bits (0010) show an invalid header length ($2 \times 4 = 8$). The minimum number of bytes in the header must be 20. The packet has been corrupted in transmission.

Example 20.2

In an IPv4 packet, the value of HLEN is 1000 in binary. How many bytes of options are being carried by this packet?

Solution

The HLEN value is 8, which means the total number of bytes in the header is 8×4, or 32 bytes. The first 20 bytes are the base header, the next 12 bytes are the options.

Example 20.3

In an IPv4 packet, the value of HLEN is 5, and the value of the total length field is 0x0028. How many bytes of data are being carried by this packet?

Solution

The HLEN value is 5, which means the total number of bytes in the header is 5×4, or 20 bytes (no options). The total length is 40 bytes, which means the packet is carrying 20 bytes of data (40 – 20).

Example 20.4

An IPv4 packet has arrived with the first few hexadecimal digits as shown.

0x45000028000100000102 . . .

How many hops can this packet travel before being dropped? The data belong to what upper-layer protocol?

Solution

To find the time-to-live field, we skip 8 bytes (16 hexadecimal digits). The time-to-live field is the ninth byte, which is 01. This means the packet can travel only one hop. The protocol field is the next byte (02), which means that the upper-layer protocol is IGMP (see Table 20.4).

Fragmentation

A datagram can travel through different networks. Each router decapsulates the IPv4 datagram from the frame it receives, processes it, and then encapsulates it in another frame. The format and size of the received frame depend on the protocol used by the physical network through which the frame has just traveled. The format and size of the sent frame depend on the protocol used by the physical network through which the frame is going to travel. For example, if a router connects a LAN to a WAN, it receives a frame in the LAN format and sends a frame in the WAN format.

Maximum Transfer Unit (MTU)

Each data link layer protocol has its own frame format in most protocols. One of the fields defined in the format is the maximum size of the data field. In other words, when a datagram is encapsulated in a frame, the total size of the datagram must be less than this maximum size, which is defined by the restrictions imposed by the hardware and software used in the network (see Figure 20.9).

The value of the MTU depends on the physical network protocol. Table 20.5 shows the values for some protocols.

Figure 20.9 *Maximum transfer unit (MTU)*

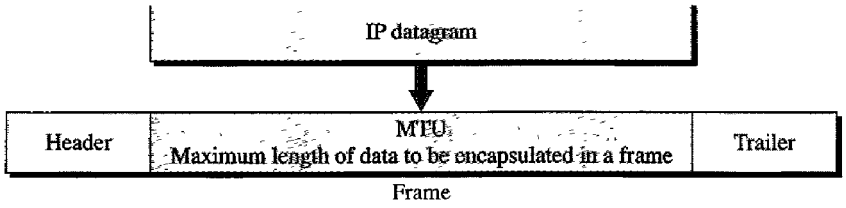


Table 20.5 *MTUs for some networks*

Protocol	*MTU*
Hyperchannel	65,535
Token Ring (16 Mbps)	17,914
Token Ring (4 Mbps)	4,464
FDDI	4,352
Ethernet	1,500
X.25	576
PPP	296

To make the IPv4 protocol independent of the physical network, the designers decided to make the maximum length of the IPv4 datagram equal to 65,535 bytes. This makes transmission more efficient if we use a protocol with an MTU of this size. However, for other physical networks, we must divide the datagram to make it possible to pass through these networks. This is called **fragmentation.**

The source usually does not fragment the IPv4 packet. The transport layer will instead segment the data into a size that can be accommodated by IPv4 and the data link layer in use.

When a datagram is fragmented, each fragment has its own header with most of the fields repeated, but with some changed. A fragmented datagram may itself be fragmented if it encounters a network with an even smaller MTU. In other words, a datagram can be fragmented several times before it reaches the final destination.

In IPv4, a datagram can be fragmented by the source host or any router in the path although there is a tendency to limit fragmentation only at the source. The reassembly of the datagram, however, is done only by the destination host because each fragment becomes an independent datagram. Whereas the fragmented datagram can travel through different routes, and we can never control or guarantee which route a fragmented datagram may take, all the fragments belonging to the same datagram should finally arrive at the destination host. So it is logical to do the reassembly at the final destination. An even stronger objection to reassembling packets during the transmission is the loss of efficiency it incurs.

When a datagram is fragmented, required parts of the header must be copied by all fragments. The option field may or may not be copied, as we will see in the next section. The host or router that fragments a datagram must change the values of three fields:

flags, fragmentation offset, and total length. The rest of the fields must be copied. Of course, the value of the checksum must be recalculated regardless of fragmentation.

Fields Related to Fragmentation

The fields that are related to fragmentation and reassembly of an IPv4 datagram are the identification, flags, and fragmentation offset fields.

- **Identification.** This 16-bit field identifies a datagram originating from the source host. The combination of the identification and source IPv4 address must uniquely define a datagram as it leaves the source host. To guarantee uniqueness, the IPv4 protocol uses a counter to label the datagrams. The counter is initialized to a positive number. When the IPv4 protocol sends a datagram, it copies the current value of the counter to the identification field and increments the counter by 1. As long as the counter is kept in the main memory, uniqueness is guaranteed. When a datagram is fragmented, the value in the identification field is copied to all fragments. In other words, all fragments have the same identification number, the same as the original datagram. The identification number helps the destination in reassembling the datagram. It knows that all fragments having the same identification value must be assembled into one datagram.
- **Flags.** This is a 3-bit field. The first bit is reserved. The second bit is called the *do not fragment* bit. If its value is 1, the machine must not fragment the datagram. If it cannot pass the datagram through any available physical network, it discards the datagram and sends an ICMP error message to the source host (see Chapter 21). If its value is 0, the datagram can be fragmented if necessary. The third bit is called the *more fragment* bit. If its value is 1, it means the datagram is not the last fragment; there are more fragments after this one. If its value is 0, it means this is the last or only fragment (see Figure 20.10).

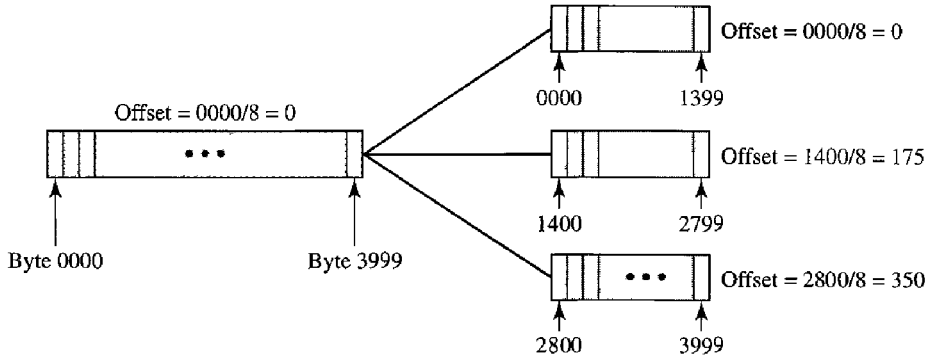
Figure 20.10 *Flags used in fragmentation*

- **Fragmentation offset.** This 13-bit field shows the relative position of this fragment with respect to the whole datagram. It is the offset of the data in the original datagram measured in units of 8 bytes. Figure 20.11 shows a datagram with a data size of 4000 bytes fragmented into three fragments.

  The bytes in the original datagram are numbered 0 to 3999. The first fragment carries bytes 0 to 1399. The offset for this datagram is 0/8 = 0. The second fragment carries bytes 1400 to 2799; the offset value for this fragment is 1400/8 = 175. Finally, the third fragment carries bytes 2800 to 3999. The offset value for this fragment is 2800/8 = 350.

  Remember that the value of the offset is measured in units of 8 bytes. This is done because the length of the offset field is only 13 bits and cannot represent a

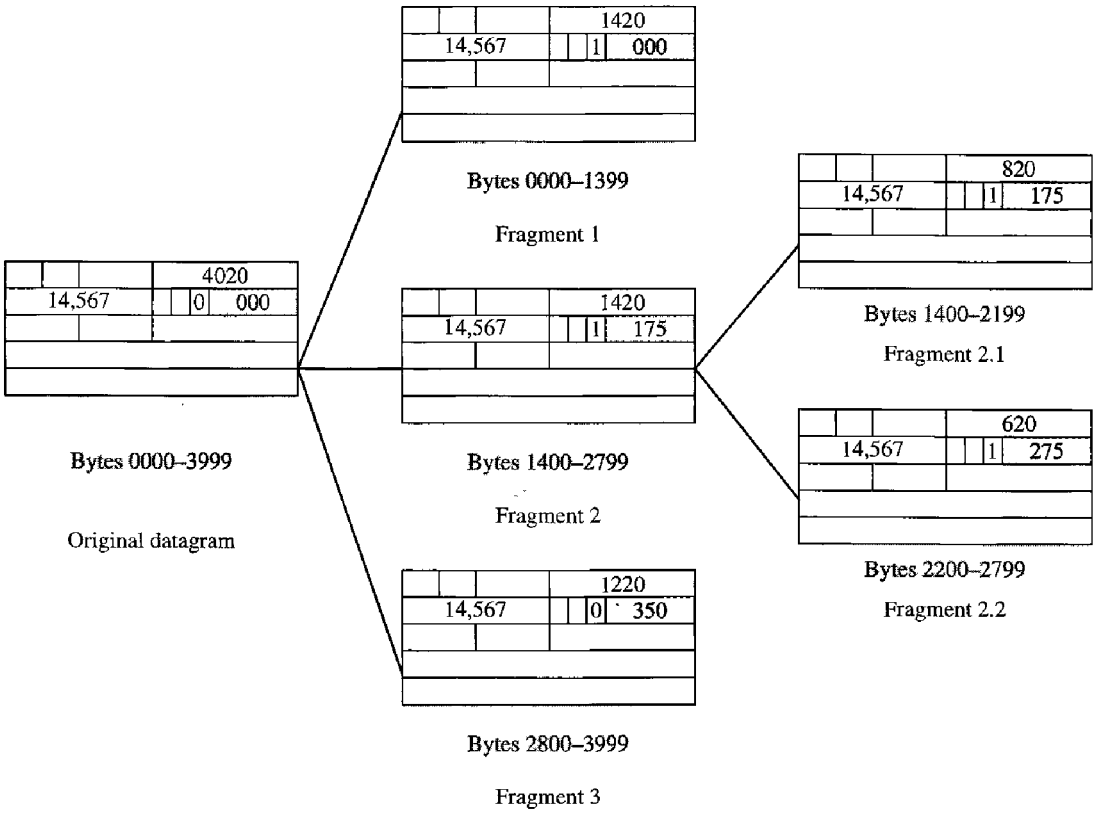
Figure 20.11 *Fragmentation example*



sequence of bytes greater than 8191. This forces hosts or routers that fragment datagrams to choose a fragment size so that the first byte number is divisible by 8.

Figure 20.12 shows an expanded view of the fragments in Figure 20.11. Notice the value of the identification field is the same in all fragments. Notice the value of the flags field with the *more* bit set for all fragments except the last. Also, the value of the offset field for each fragment is shown.

Figure 20.12 *Detailed fragmentation example*



The figure also shows what happens if a fragment itself is fragmented. In this case the value of the offset field is always relative to the original datagram. For example, in the figure, the second fragment is itself fragmented later to two fragments of 800 bytes and 600 bytes, but the offset shows the relative position of the fragments to the original data.

It is obvious that even if each fragment follows a different path and arrives out of order, the final destination host can reassemble the original datagram from the fragments received (if none of them is lost) by using the following strategy:

1. The first fragment has an offset field value of zero.
2. Divide the length of the first fragment by 8. The second fragment has an offset value equal to that result.
3. Divide the total length of the first and second fragments by 8. The third fragment has an offset value equal to that result.
4. Continue the process. The last fragment has a *more* bit value of 0.

Example 20.5

A packet has arrived with an *M* bit value of 0. Is this the first fragment, the last fragment, or a middle fragment? Do we know if the packet was fragmented?

Solution

If the *M* bit is 0, it means that there are no more fragments; the fragment is the last one. However, we cannot say if the original packet was fragmented or not. A nonfragmented packet is considered the last fragment.

Example 20.6

A packet has arrived with an *M* bit value of 1. Is this the first fragment, the last fragment, or a middle fragment? Do we know if the packet was fragmented?

Solution

If the *M* bit is 1, it means that there is at least one more fragment. This fragment can be the first one or a middle one, but not the last one. We don't know if it is the first one or a middle one; we need more information (the value of the fragmentation offset). See Example 20.7.

Example 20.7

A packet has arrived with an *M* bit value of 1 and a fragmentation offset value of 0. Is this the first fragment, the last fragment, or a middle fragment?

Solution

Because the *M* bit is 1, it is either the first fragment or a middle one. Because the offset value is 0, it is the first fragment.

Example 20.8

A packet has arrived in which the offset value is 100. What is the number of the first byte? Do we know the number of the last byte?

Solution

To find the number of the first byte, we multiply the offset value by 8. This means that the first byte number is 800. We cannot determine the number of the last byte unless we know the length of the data.

Example 20.9

A packet has arrived in which the offset value is 100, the value of HLEN is 5, and the value of the total length field is 100. What are the numbers of the first byte and the last byte?

Solution

The first byte number is $100 \times 8 = 800$. The total length is 100 bytes, and the header length is 20 bytes (5×4), which means that there are 80 bytes in this datagram. If the first byte number is 800, the last byte number must be 879.

Checksum

We discussed the general idea behind the checksum and how it is calculated in Chapter 10. The implementation of the checksum in the IPv4 packet follows the same principles. First, the value of the checksum field is set to 0. Then the entire header is divided into 16-bit sections and added together. The result (sum) is complemented and inserted into the checksum field.

The checksum in the IPv4 packet covers only the header, not the data. There are two good reasons for this. First, all higher-level protocols that encapsulate data in the IPv4 datagram have a checksum field that covers the whole packet. Therefore, the checksum for the IPv4 datagram does not have to check the encapsulated data. Second, the header of the IPv4 packet changes with each visited router, but the data do not. So the checksum includes only the part that has changed. If the data were included, each router must recalculate the checksum for the whole packet, which means an increase in processing time.

Example 20.10

Figure 20.13 shows an example of a checksum calculation for an IPv4 header without options. The header is divided into 16-bit sections. All the sections are added and the sum is complemented. The result is inserted in the checksum field.

Options

The header of the IPv4 datagram is made of two parts: a fixed part and a variable part. The fixed part is 20 bytes long and was discussed in the previous section. The variable part comprises the options that can be a maximum of 40 bytes.

Options, as the name implies, are not required for a datagram. They can be used for network testing and debugging. Although options are not a required part of the IPv4 header, option processing is required of the IPv4 software. This means that all implementations must be able to handle options if they are present in the header.

The detailed discussion of each option is beyond the scope of this book. We give the taxonomy of options in Figure 20.14 and briefly explain the purpose of each.

No Operation

A **no-operation option** is a 1-byte option used as a filler between options.

Figure 20.13 *Example of checksum calculation in IPv4*

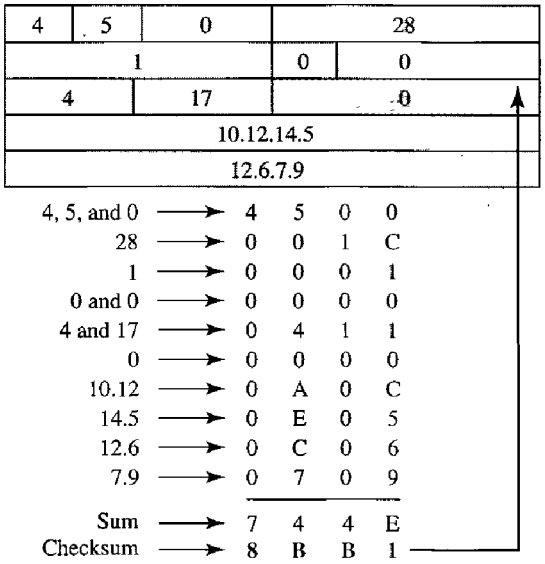
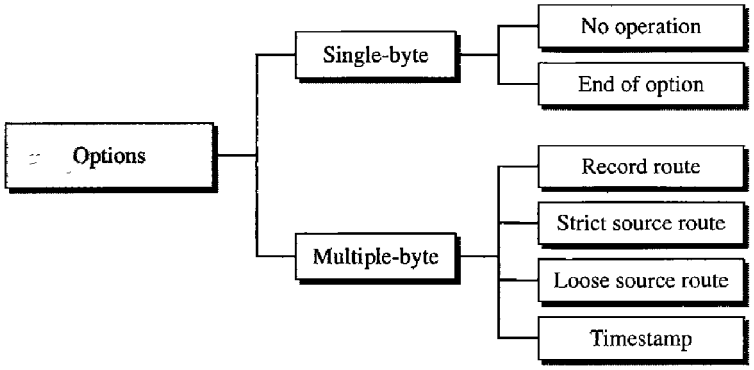


Figure 20.14 *Taxonomy of options in IPv4*



End of Option

An **end-of-option option** is a 1-byte option used for padding at the end of the option field. It, however, can only be used as the last option.

Record Route

A **record route option** is used to record the Internet routers that handle the datagram. It can list up to nine router addresses. It can be used for debugging and management purposes.

Strict Source Route

A **strict source route option** is used by the source to predetermine a route for the datagram as it travels through the Internet. Dictation of a route by the source can be useful

for several purposes. The sender can choose a route with a specific type of service, such as minimum delay or maximum throughput. Alternatively, it may choose a route that is safer or more reliable for the sender's purpose. For example, a sender can choose a route so that its datagram does not travel through a competitor's network.

If a datagram specifies a strict source route, all the routers defined in the option must be visited by the datagram. A router must not be visited if its IPv4 address is not listed in the datagram. If the datagram visits a router that is not on the list, the datagram is discarded and an error message is issued. If the datagram arrives at the destination and some of the entries were not visited, it will also be discarded and an error message issued.

Loose Source Route

A **loose source route option** is similar to the strict source route, but it is less rigid. Each router in the list must be visited, but the datagram can visit other routers as well.

Timestamp

A **timestamp option** is used to record the time of datagram processing by a router. The time is expressed in milliseconds from midnight, Universal time or Greenwich mean time. Knowing the time a datagram is processed can help users and managers track the behavior of the routers in the Internet. We can estimate the time it takes for a datagram to go from one router to another. We say *estimate* because, although all routers may use Universal time, their local clocks may not be synchronized.

20.3 IPv6

The network layer protocol in the TCP/IP protocol suite is currently IPv4 (Internetworking Protocol, version 4). IPv4 provides the host-to-host communication between systems in the Internet. Although IPv4 is well designed, data communication has evolved since the inception of IPv4 in the 1970s. IPv4 has some deficiencies (listed below) that make it unsuitable for the fast-growing Internet.

- Despite all short-term solutions, such as subnetting, classless addressing, and NAT, address depletion is still a long-term problem in the Internet.
- The Internet must accommodate real-time audio and video transmission. This type of transmission requires minimum delay strategies and reservation of resources not provided in the IPv4 design.
- The Internet must accommodate encryption and authentication of data for some applications. No encryption or authentication is provided by IPv4.

To overcome these deficiencies, **IPv6 (Internetworking Protocol, version 6),** also known as **IPng (Internetworking Protocol, next generation),** was proposed and is now a standard. In IPv6, the Internet protocol was extensively modified to accommodate the unforeseen growth of the Internet. The format and the length of the IP address were changed along with the packet format. Related protocols, such as ICMP, were also modified. Other protocols in the network layer, such as ARP, RARP, and IGMP, were

either deleted or included in the ICMPv6 protocol (see Chapter 21). Routing protocols, such as RIP and OSPF (see Chapter 22), were also slightly modified to accommodate these changes. Communications experts predict that IPv6 and its related protocols will soon replace the current IP version. In this section first we discuss IPv6. Then we explore the strategies used for the transition from version 4 to version 6.

The adoption of IPv6 has been slow. The reason is that the original motivation for its development, depletion of IPv4 addresses, has been remedied by short-term strategies such as classless addressing and NAT. However, the fast-spreading use of the Internet, and new services such as mobile IP, IP telephony, and IP-capable mobile telephony, may eventually require the total replacement of IPv4 with IPv6.

Advantages

The next-generation IP, or IPv6, has some advantages over IPv4 that can be summarized as follows:

- **Larger address space.** An IPv6 address is 128 bits long, as we discussed in Chapter 19. Compared with the 32-bit address of IPv4, this is a huge (2^{96}) increase in the address space.
- **Better header format.** IPv6 uses a new header format in which options are separated from the base header and inserted, when needed, between the base header and the upper-layer data. This simplifies and speeds up the routing process because most of the options do not need to be checked by routers.
- **New options.** IPv6 has new options to allow for additional functionalities.
- **Allowance for extension.** IPv6 is designed to allow the extension of the protocol if required by new technologies or applications.
- **Support for resource allocation.** In IPv6, the type-of-service field has been removed, but a mechanism (called *flow label*) has been added to enable the source to request special handling of the packet. This mechanism can be used to support traffic such as real-time audio and video.
- **Support for more security.** The encryption and authentication options in IPv6 provide confidentiality and integrity of the packet.

Packet Format

The IPv6 packet is shown in Figure 20.15. Each packet is composed of a mandatory base header followed by the payload. The payload consists of two parts: optional extension headers and data from an upper layer. The base header occupies 40 bytes, whereas the extension headers and data from the upper layer contain up to 65,535 bytes of information.

Base Header

Figure 20.16 shows the **base header** with its eight fields.

These fields are as follows:

- **Version.** This 4-bit field defines the version number of the IP. For IPv6, the value is 6.
- **Priority.** The 4-bit priority field defines the priority of the packet with respect to traffic congestion. We will discuss this field later.

Figure 20.15 *IPv6 datagram header and payload*

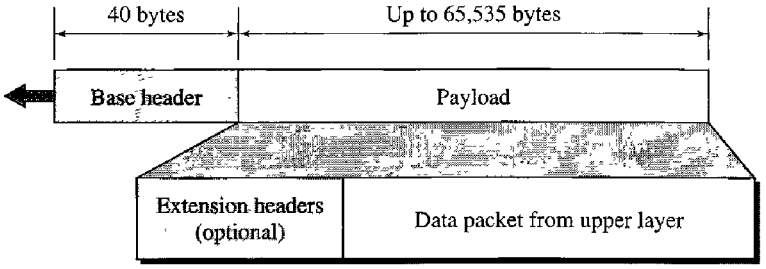
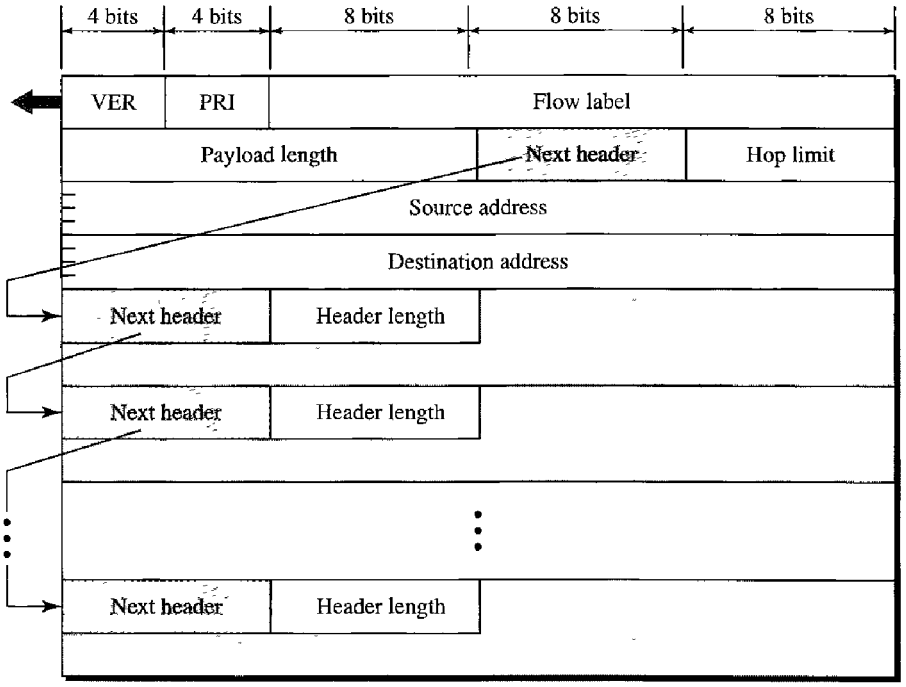


Figure 20.16 *Format of an IPv6 datagram*



- **Flow label.** The **flow label** is a 3-byte (24-bit) field that is designed to provide special handling for a particular flow of data. We will discuss this field later.
- **Payload length.** The 2-byte payload length field defines the length of the IP datagram excluding the base header.
- **Next header.** The **next header** is an 8-bit field defining the header that follows the base header in the datagram. The next header is either one of the optional extension headers used by IP or the header of an encapsulated packet such as UDP or TCP. Each extension header also contains this field. Table 20.6 shows the values of next headers. Note that this field in version 4 is called the *protocol*.
- **Hop limit.** This 8-bit **hop limit** field serves the same purpose as the TTL field in IPv4.
- **Source address.** The source address field is a 16-byte (128-bit) Internet address that identifies the original source of the datagram.

Table 20.6 *Next header codes for IPv6*

Code	*Next Header*
0	Hop-by-hop option
2	ICMP
6	TCP
17	UDP
43	Source routing
44	Fragmentation
50	Encrypted security payload
51	Authentication
59	Null (no next header)
60	Destination option

- **Destination address.** The destination address field is a 16-byte (128-bit) Internet address that usually identifies the final destination of the datagram. However, if source routing is used, this field contains the address of the next router.

Priority

The priority field of the IPv6 packet defines the priority of each packet with respect to other packets from the same source. For example, if one of two consecutive datagrams must be discarded due to congestion, the datagram with the lower **packet priority** will be discarded. IPv6 divides traffic into two broad categories: congestion-controlled and noncongestion-controlled.

Congestion-Controlled Traffic If a source adapts itself to traffic slowdown when there is congestion, the traffic is referred to as **congestion-controlled traffic.** For example, TCP, which uses the sliding window protocol, can easily respond to traffic. In congestion-controlled traffic, it is understood that packets may arrive delayed, lost, or out of order. Congestion-controlled data are assigned priorities from 0 to 7, as listed in Table 20.7. A priority of 0 is the lowest; a priority of 7 is the highest.

Table 20.7 *Priorities for congestion-controlled traffic*

Priority	*Meaning*
0	No specific traffic
1	Background data
2	Unattended data traffic
3	Reserved
4	Attended bulk data traffic
5	Reserved
6	Interactive traffic
7	Control traffic

The priority descriptions are as follows:

- **No specific traffic.** A priority of 0 is assigned to a packet when the process does not define a priority.
- **Background data.** This group (priority 1) defines data that are usually delivered in the background. Delivery of the news is a good example.
- **Unattended data traffic.** If the user is not waiting (attending) for the data to be received, the packet will be given a priority of 2. E-mail belongs to this group. The recipient of an e-mail does not know when a message has arrived. In addition, an e-mail is usually stored before it is forwarded. A little bit of delay is of little consequence.
- **Attended bulk data traffic.** A protocol that transfers data while the user is waiting (attending) to receive the data (possibly with delay) is given a priority of 4. FTP and HTTP belong to this group.
- **Interactive traffic.** Protocols such as TELNET that need user interaction are assigned the second-highest priority (6) in this group.
- **Control traffic.** Control traffic is given the highest priority (7). Routing protocols such as OSPF and RIP and management protocols such as SNMP have this priority.

Noncongestion-Controlled Traffic This refers to a type of traffic that expects minimum delay. Discarding of packets is not desirable. Retransmission in most cases is impossible. In other words, the source does not adapt itself to congestion. Real-time audio and video are examples of this type of traffic.

Priority numbers from 8 to 15 are assigned to **noncongestion-controlled traffic.** Although there are not yet any particular standard assignments for this type of data, the priorities are usually based on how much the quality of received data is affected by the discarding of packets. Data containing less redundancy (such as low-fidelity audio or video) can be given a higher priority (15). Data containing more redundancy (such as high-fidelity audio or video) are given a lower priority (8). See Table 20.8.

Table 20.8 *Priorities for noncongestion-controlled traffic*

Priority	*Meaning*
8	Data with greatest redundancy
. . .	. . .
15	Data with least redundancy

Flow Label

A sequence of packets, sent from a particular source to a particular destination, that needs special handling by routers is called a *flow* of packets. The combination of the source address and the value of the *flow label* uniquely defines a flow of packets.

To a router, a flow is a sequence of packets that share the same characteristics, such as traveling the same path, using the same resources, having the same kind of security, and so on. A router that supports the handling of flow labels has a flow label table. The table has an entry for each active flow label; each entry defines the services required by

the corresponding flow label. When the router receives a packet, it consults its flow label table to find the corresponding entry for the flow label value defined in the packet. It then provides the packet with the services mentioned in the entry. However, note that the flow label itself does not provide the information for the entries of the flow label table; the information is provided by other means such as the hop-by-hop options or other protocols.

In its simplest form, a flow label can be used to speed up the processing of a packet by a router. When a router receives a packet, instead of consulting the routing table and going through a routing algorithm to define the address of the next hop, it can easily look in a flow label table for the next hop.

In its more sophisticated form, a flow label can be used to support the transmission of real-time audio and video. Real-time audio or video, particularly in digital form, requires resources such as high bandwidth, large buffers, long processing time, and so on. A process can make a reservation for these resources beforehand to guarantee that real-time data will not be delayed due to a lack of resources. The use of real-time data and the reservation of these resources require other protocols such as Real-Time Protocol (RTP) and Resource Reservation Protocol (RSVP) in addition to IPv6.

To allow the effective use of flow labels, three rules have been defined:

1. The flow label is assigned to a packet by the source host. The label is a random number between 1 and $2^{24} - 1$. A source must not reuse a flow label for a new flow while the existing flow is still active.
2. If a host does not support the flow label, it sets this field to zero. If a router does not support the flow label, it simply ignores it.
3. All packets belonging to the same flow have the same source, same destination, same priority, and same options.

Comparison Between IPv4 and IPv6 Headers

Table 20.9 compares IPv4 and IPv6 headers.

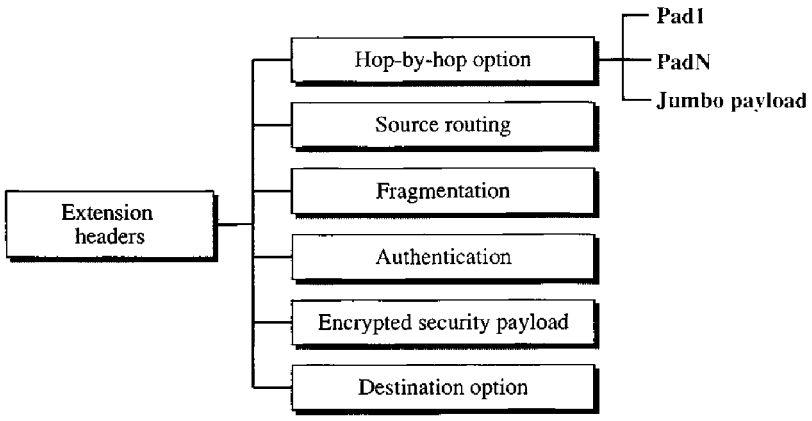
Table 20.9 *Comparison between IPv4 and IPv6 packet headers*

Comparison
1. The header length field is eliminated in IPv6 because the length of the header is fixed in this version.
2. The service type field is eliminated in IPv6. The priority and flow label fields together take over the function of the service type field.
3. The total length field is eliminated in IPv6 and replaced by the payload length field.
4. The identification, flag, and offset fields are eliminated from the base header in IPv6. They are included in the fragmentation extension header.
5. The TTL field is called hop limit in IPv6.
6. The protocol field is replaced by the next header field.
7. The header checksum is eliminated because the checksum is provided by upper-layer protocols; it is therefore not needed at this level.
8. The option fields in IPv4 are implemented as extension headers in IPv6.

Extension Headers

The length of the base header is fixed at 40 bytes. However, to give greater functionality to the IP datagram, the base header can be followed by up to six **extension headers.** Many of these headers are options in IPv4. Six types of extension headers have been defined, as shown in Figure 20.17.

Figure 20.17 *Extension header types*



Hop-by-Hop Option

The **hop-by-hop option** is used when the source needs to pass information to all routers visited by the datagram. So far, only three options have been defined: **Pad1, PadN,** and **jumbo payload.** The Pad1 option is 1 byte long and is designed for alignment purposes. PadN is similar in concept to Pad1. The difference is that PadN is used when 2 or more bytes is needed for alignment. The jumbo payload option is used to define a payload longer than 65,535 bytes.

Source Routing The source routing extension header combines the concepts of the strict source route and the loose source route options of IPv4.

Fragmentation

The concept of **fragmentation** is the same as that in IPv4. However, the place where fragmentation occurs differs. In IPv4, the source or a router is required to fragment if the size of the datagram is larger than the MTU of the network over which the datagram travels. In IPv6, only the original source can fragment. A source must use a **path MTU discovery technique** to find the smallest MTU supported by any network on the path. The source then fragments using this knowledge.

Authentication

The **authentication** extension header has a dual purpose: it validates the message sender and ensures the integrity of data. We discuss this extension header when we discuss network security in Chapter 31.

Encrypted Security Payload

The **encrypted security payload (ESP)** is an extension that provides confidentiality and guards against eavesdropping. We discuss this extension header in Chapter 31.

Destination Option The **destination option** is used when the source needs to pass information to the destination only. Intermediate routers are not permitted access to this information.

Comparison Between IPv4 Options and IPv6 Extension Headers

Table 20.10 compares the options in IPv4 with the extension headers in IPv6.

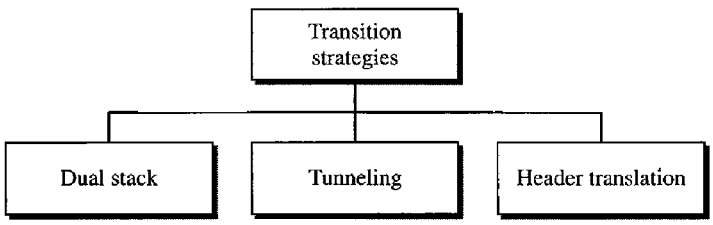
Table 20.10 *Comparison between IPv4 options and IPv6 extension headers*

Comparison
1. The no-operation and end-of-option options in IPv4 are replaced by Pad1 and PadN options in IPv6.
2. The record route option is not implemented in IPv6 because it was not used.
3. The timestamp option is not implemented because it was not used.
4. The source route option is called the source route extension header in IPv6.
5. The fragmentation fields in the base header section of IPv4 have moved to the fragmentation extension header in IPv6.
6. The authentication extension header is new in IPv6.
7. The encrypted security payload extension header is new in IPv6.

20.4 TRANSITION FROM IPv4 TO IPv6

Because of the huge number of systems on the Internet, the transition from IPv4 to IPv6 cannot happen suddenly. It takes a considerable amount of time before every system in the Internet can move from IPv4 to IPv6. The transition must be smooth to prevent any problems between IPv4 and IPv6 systems. Three strategies have been devised by the IETF to help the transition (see Figure 20.18).

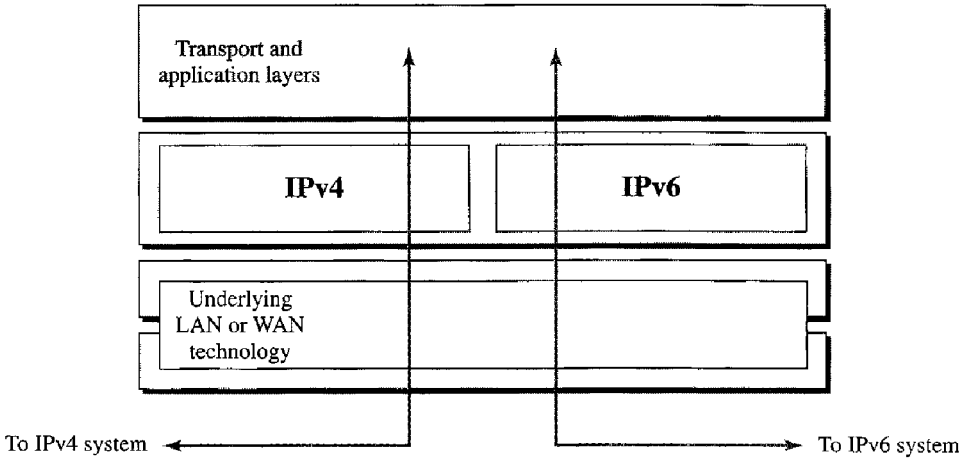
Figure 20.18 *Three transition strategies*



Dual Stack

It is recommended that all hosts, before migrating completely to version 6, have a **dual stack** of protocols. In other words, a station must run IPv4 and IPv6 simultaneously until all the Internet uses IPv6. See Figure 20.19 for the layout of a dual-stack configuration.

Figure 20.19 *Dual stack*

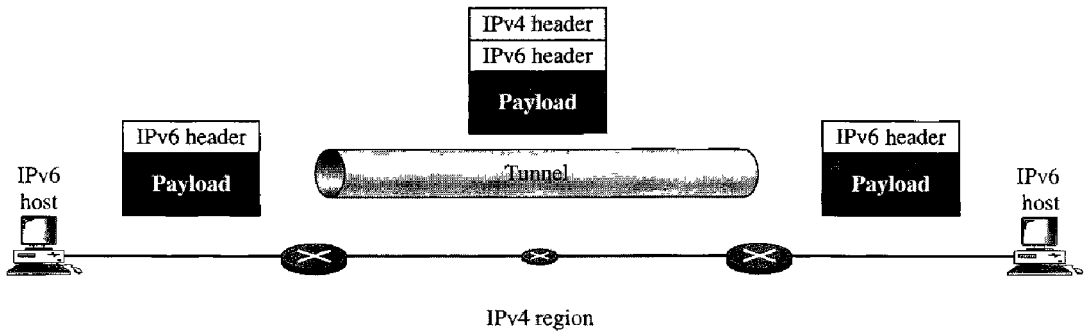


To determine which version to use when sending a packet to a destination, the source host queries the DNS. If the DNS returns an IPv4 address, the source host sends an IPv4 packet. If the DNS returns an IPv6 address, the source host sends an IPv6 packet.

Tunneling

Tunneling is a strategy used when two computers using IPv6 want to communicate with each other and the packet must pass through a region that uses IPv4. To pass through this region, the packet must have an IPv4 address. So the IPv6 packet is encapsulated in an IPv4 packet when it enters the region, and it leaves its capsule when it exits the region. It seems as if the IPv6 packet goes through a tunnel at one end and emerges at the other end. To make it clear that the IPv4 packet is carrying an IPv6 packet as data, the protocol value is set to 41. Tunneling is shown in Figure 20.20.

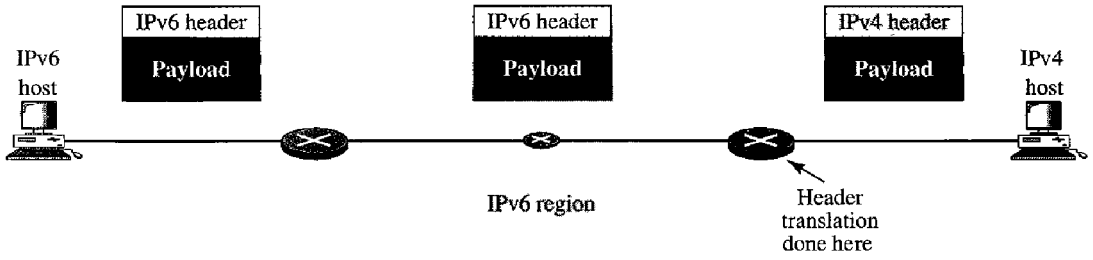
Figure 20.20 *Tunneling strategy*



Header Translation

Header translation is necessary when the majority of the Internet has moved to IPv6 but some systems still use IPv4. The sender wants to use IPv6, but the receiver does not understand IPv6. Tunneling does not work in this situation because the packet must be in the IPv4 format to be understood by the receiver. In this case, the header format must be totally changed through header translation. The header of the IPv6 packet is converted to an IPv4 header (see Figure 20.21).

Figure 20.21 *Header translation strategy*



Header translation uses the mapped address to translate an IPv6 address to an IPv4 address. Table 20.11 lists some rules used in transforming an IPv6 packet header to an IPv4 packet header.

Table 20.11 *Header translation*

Header Translation Procedure
1. The IPv6 mapped address is changed to an IPv4 address by extracting the rightmost 32 bits.
2. The value of the IPv6 priority field is discarded.
3. The type of service field in IPv4 is set to zero.
4. The checksum for IPv4 is calculated and inserted in the corresponding field.
5. The IPv6 flow label is ignored.
6. Compatible extension headers are converted to options and inserted in the IPv4 header. Some may have to be dropped.
7. The length of IPv4 header is calculated and inserted into the corresponding field.
8. The total length of the IPv4 packet is calculated and inserted in the corresponding field.

20.5 RECOMMENDED READING

For more details about subjects discussed in this chapter, we recommend the following books and sites. The items in brackets [. . .] refer to the reference list at the end of the text.

Books

IPv4 is discussed in Chapter 8 of [For06], Chapter 3 of [Ste94], Section 4.1 of [PD03], Chapter 18 of [Sta04], and Section 5.6 of [Tan03]. IPv6 is discussed in Chapter 27 of [For06] and [Los04].

Sites

- www.ietf.org/rfc.html Information about RFCs

RFCs

A discussion of IPv4 can be found in following RFCs:

760, 781, 791, 815, 1025, 1063, 1071, 1141, 1190, 1191, 1624, 2113

A discussion of IPv6 can be found in the following RFCs:

1365, 1550, 1678, 1680, 1682, 1683, 1686, 1688, 1726, 1752, 1826, 1883, 1884, 1886, 1887, 1955, 2080, 2373, 2452, 2463, 2465, 2466, 2472, 2492, 2545, 2590

20.6 KEY TERMS

authentication
base header
best-effort delivery
codepoint
connectionless service
connection-oriented service
datagram
destination address
destination option
differentiated services
dual stack
encrypted security payload (ESP)
end-of-option option
extension header
flow label
fragmentation
fragmentation offset
header length
header translation
hop limit
hop-by-hop option
Internet Protocol (IP)
Internet Protocol, next generation (IPng)
Internet Protocol version 4 (IPv4)
Internet Protocol version 6 (IPv6)
jumbo payload
loose source route option
maximum transfer unit (MTU)
next header
noncongestion-controlled traffic
no-operation option
packet priority
Pad1
PadN
path MTU discovery technique
precedence

record route option
service type
source address
strict source route option
time to live
timestamp option
tunneling
type of service (TOS)

20.7 SUMMARY

- IPv4 is an unreliable connectionless protocol responsible for source-to-destination delivery.
- Packets in the IPv4 layer are called datagrams. A datagram consists of a header (20 to 60 bytes) and data. The maximum length of a datagram is 65,535 bytes.
- The MTU is the maximum number of bytes that a data link protocol can encapsulate. MTUs vary from protocol to protocol.
- Fragmentation is the division of a datagram into smaller units to accommodate the MTU of a data link protocol.
- The IPv4 datagram header consists of a fixed, 20-byte section and a variable options section with a maximum of 40 bytes.
- The options section of the IPv4 header is used for network testing and debugging.
- The six IPv4 options each have a specific function. They are as follows: filler between options for alignment purposes, padding, recording the route the datagram takes, selection of a mandatory route by the sender, selection of certain routers that must be visited, and recording of processing times at routers.
- IPv6, the latest version of the Internet Protocol, has a 128-bit address space, a revised header format, new options, an allowance for extension, support for resource allocation, and increased security measures.
- An IPv6 datagram is composed of a base header and a payload.
- Extension headers add functionality to the IPv6 datagram.
- Three strategies used to handle the transition from version 4 to version 6 are dual stack, tunneling, and header translation.

20.8 PRACTICE SET

Review Questions

1. What is the difference between the delivery of a frame in the data link layer and the delivery of a packet in the network layer?
2. What is the difference between connectionless and connection-oriented services? Which type of service is provided by IPv4? Which type of service is provided by IPv6?
3. Define fragmentation and explain why the IPv4 and IPv6 protocols need to fragment some packets. Is there any difference between the two protocols in this matter?

4. Explain the procedure for checksum calculation and verification in the IPv4 protocol. What part of an IPv4 packet is covered in the checksum calculation? Why? Are options, if present, included in the calculation?
5. Explain the need for options in IPv4 and list the options mentioned in this chapter with a brief description of each.
6. Compare and contrast the fields in the main headers of IPv4 and IPv6. Make a table that shows the presence or absence of each field.
7. Both IPv4 and IPv6 assume that packets may have different priorities or precedences. Explain how each protocol handles this issue.
8. Compare and contrast the options in IPv4 and the extension headers in IPv6. Make a table that shows the presence or absence of each.
9. Explain the reason for the elimination of the checksum in the IPv6 header.
10. List three transition strategies to move from IPv4 to IPv6. Explain the difference between tunneling and dual stack strategies during the transition period. When is each strategy used?

Exercises

11. Which fields of the IPv4 header change from router to router?
12. Calculate the HLEN (in IPv4) value if the total length is 1200 bytes, 1176 of which is data from the upper layer.
13. Table 20.5 lists the MTUs for many different protocols. The MTUs range from 296 to 65,535. What would be the advantages of having a large MTU? What would be the advantages of having a small MTU?
14. Given a fragmented datagram (in IPv4) with an offset of 120, how can you determine the first and last byte numbers?
15. Can the value of the header length in an IPv4 packet be less than 5? When is it exactly 5?
16. The value of HLEN in an IPv4 datagram is 7. How many option bytes are present?
17. The size of the option field of an IPv4 datagram is 20 bytes. What is the value of HLEN? What is the value in binary?
18. The value of the total length field in an IPv4 datagram is 36, and the value of the header length field is 5. How many bytes of data is the packet carrying?
19. An IPv4 datagram is carrying 1024 bytes of data. If there is no option information, what is the value of the header length field? What is the value of the total length field?
20. A host is sending 100 datagrams to another host. If the identification number of the first datagram is 1024, what is the identification number of the last (in IPv4)?
21. An IPv4 datagram arrives with fragmentation offset of 0 and an *M* bit (*more* fragment bit) of 0. Is this a first fragment, middle fragment, or last fragment?
22. An IPv4 fragment has arrived with an offset value of 100. How many bytes of data were originally sent by the source before the data in this fragment?

23. An IPv4 datagram has arrived with the following information in the header (in hexadecimal):

0x45 00 00 54 00 03 58 50 20 06 00 00 7C 4E 03 02 B4 0E 0F 02

a. Is the packet corrupted?
b. Are there any options?
c. Is the packet fragmented?
d. What is the size of the data?
e. How many more routers can the packet travel to?
f. What is the identification number of the packet?
g. What is the type of service?

24. In an IPv4 datagram, the *M* bit is 0, the value of HLEN is 5, the value of total length is 200, and the offset value is 200. What is the number of the first byte and number of the last byte in this datagram? Is this the last fragment, the first fragment, or a middle fragment?

Research Activities

25. Find out why there are two security protocols (AH and ESP) in IPv6.

CHAPTER 21

Network Layer: Address Mapping, Error Reporting, and Multicasting

In Chapter 20 we discussed the Internet Protocol (IP) as the main protocol at the network layer. IP was designed as a best-effort delivery protocol, but it lacks some features such as flow control and error control. It is a host-to-host protocol using logical addressing. To make IP more responsive to some requirements in today's internetworking, we need the help of other protocols.

We need protocols to create a mapping between physical and logical addresses. IP packets use logical (host-to-host) addresses. These packets, however, need to be encapsulated in a frame, which needs physical addresses (node-to-node). We will see that a protocol called **ARP,** the **Address Resolution Protocol,** is designed for this purpose. We sometimes need reverse mapping—mapping a physical address to a logical address. For example, when booting a diskless network or leasing an IP address to a host. Three protocols are designed for this purpose: RARP, BOOTP, and DHCP.

Lack of flow and error control in the Internet Protocol has resulted in another protocol, ICMP, that provides alerts. It reports congestion and some types of errors in the network or destination host.

IP was originally designed for unicast delivery, one source to one destination. As the Internet has evolved, the need for multicast delivery, one source to many destinations, has increased tremendously. IGMP gives IP a multicast capability.

In this chapter, we discuss the protocols ARP, RARP, BOOTP, DHCP, and IGMP in some detail. We also discuss ICMPv6, which will be operational when IPv6 is operational. ICMPv6 combines ARP, ICMP, and IGMP in one protocol.

21.1 ADDRESS MAPPING

An internet is made of a combination of physical networks connected by internetworking devices such as routers. A packet starting from a source host may pass through several different physical networks before finally reaching the destination host. The hosts and routers are recognized at the network level by their logical (IP) addresses.

However, packets pass through physical networks to reach these hosts and routers. At the physical level, the hosts and routers are recognized by their physical addresses.

A **physical address** is a local address. Its jurisdiction is a local network. It must be unique locally, but is not necessarily unique universally. It is called a *physical* address because it is usually (but not always) implemented in hardware. An example of a physical address is the 48-bit MAC address in the Ethernet protocol, which is imprinted on the NIC installed in the host or router.

The physical address and the logical address are two different identifiers. We need both because a physical network such as Ethernet can have two different protocols at the network layer such as IP and IPX (Novell) at the same time. Likewise, a packet at a network layer such as IP may pass through different physical networks such as Ethernet and LocalTalk (Apple).

This means that delivery of a packet to a host or a router requires two levels of addressing: logical and physical. We need to be able to map a logical address to its corresponding physical address and vice versa. These can be done by using either static or dynamic mapping.

Static mapping involves in the creation of a table that associates a logical address with a physical address. This table is stored in each machine on the network. Each machine that knows, for example, the IP address of another machine but not its physical address can look it up in the table. This has some limitations because physical addresses may change in the following ways:

1. A machine could change its NIC, resulting in a new physical address.
2. In some LANs, such as LocalTalk, the physical address changes every time the computer is turned on.
3. A mobile computer can move from one physical network to another, resulting in a change in its physical address.

To implement these changes, a static mapping table must be updated periodically. This overhead could affect network performance.

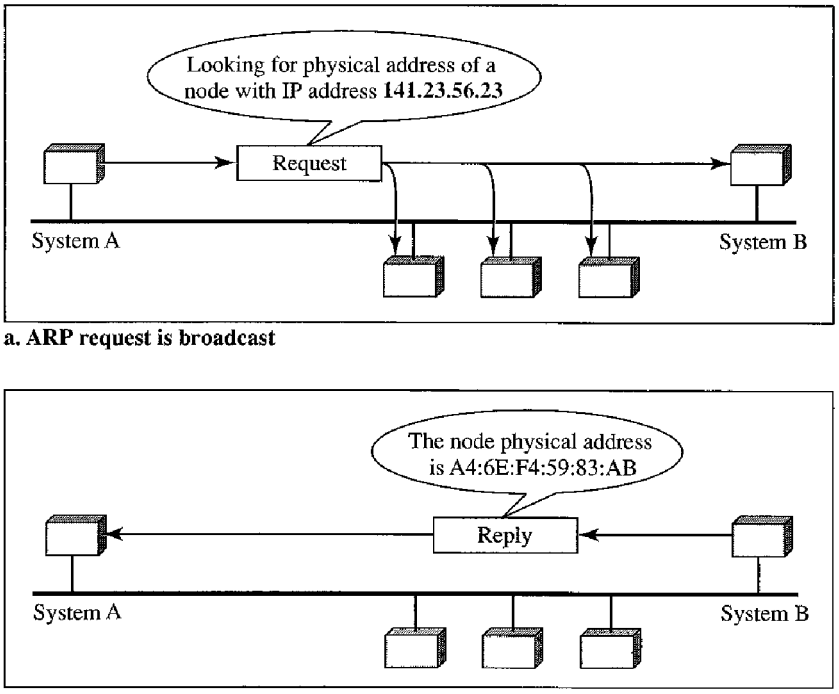
In **dynamic mapping** each time a machine knows one of the two addresses (logical or physical), it can use a protocol to find the other one.

Mapping Logical to Physical Address: ARP

Anytime a host or a router has an IP datagram to send to another host or router, it has the logical (IP) address of the receiver. The logical (IP) address is obtained from the DNS (see Chapter 25) if the sender is the host or it is found in a routing table (see Chapter 22) if the sender is a router. But the IP datagram must be encapsulated in a frame to be able to pass through the physical network. This means that the sender needs the physical address of the receiver. The host or the router sends an ARP query packet. The packet includes the physical and IP addresses of the sender and the IP address of the receiver. Because the sender does not know the physical address of the receiver, the query is broadcast over the network (see Figure 21.1).

Every host or router on the network receives and processes the ARP query packet, but only the intended recipient recognizes its IP address and sends back an ARP response packet. The response packet contains the recipient's IP and physical addresses. The packet is unicast directly to the inquirer by using the physical address received in the query packet.

Figure 21.1 *ARP operation*



In Figure 21.1a, the system on the left (A) has a packet that needs to be delivered to another system (B) with IP address 141.23.56.23. System A needs to pass the packet to its data link layer for the actual delivery, but it does not know the physical address of the recipient. It uses the services of ARP by asking the ARP protocol to send a broadcast ARP request packet to ask for the physical address of a system with an IP address of 141.23.56.23.

This packet is received by every system on the physical network, but only system B will answer it, as shown in Figure 21.1b. System B sends an ARP reply packet that includes its physical address. Now system A can send all the packets it has for this destination by using the physical address it received.

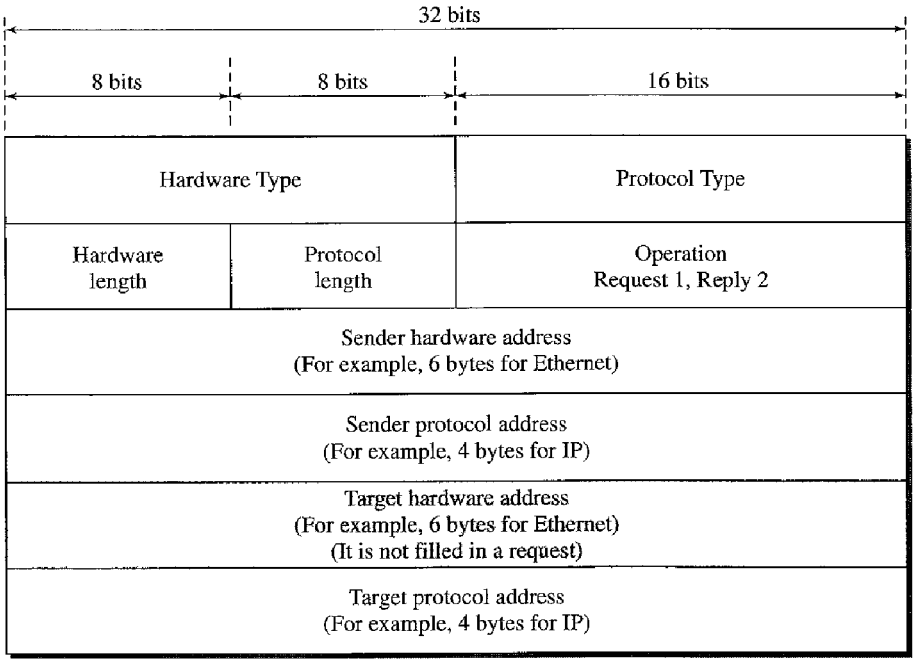
Cache Memory

Using ARP is inefficient if system A needs to broadcast an ARP request for each IP packet it needs to send to system B. It could have broadcast the IP packet itself. ARP can be useful if the ARP reply is cached (kept in cache memory for a while) because a system normally sends several packets to the same destination. A system that receives an ARP reply stores the mapping in the cache memory and keeps it for 20 to 30 minutes unless the space in the cache is exhausted. Before sending an ARP request, the system first checks its cache to see if it can find the mapping.

Packet Format

Figure 21.2 shows the format of an ARP packet.

Figure 21.2 *ARP packet*



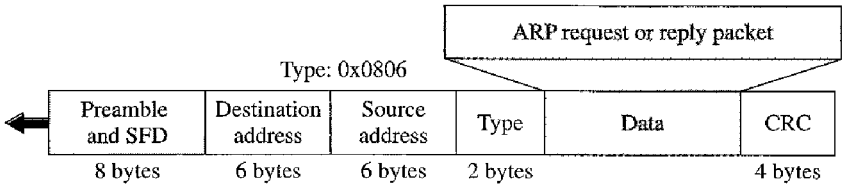
The fields are as follows:

- **Hardware type.** This is a 16-bit field defining the type of the network on which ARP is running. Each LAN has been assigned an integer based on its type. For example, Ethernet is given type 1. ARP can be used on any physical network.
- **Protocol type.** This is a 16-bit field defining the protocol. For example, the value of this field for the IPv4 protocol is 0800_{16}. ARP can be used with any higher-level protocol.
- **Hardware length.** This is an 8-bit field defining the length of the physical address in bytes. For example, for Ethernet the value is 6.
- **Protocol length.** This is an 8-bit field defining the length of the logical address in bytes. For example, for the IPv4 protocol the value is 4.
- **Operation.** This is a 16-bit field defining the type of packet. Two packet types are defined: ARP request (1) and ARP reply (2).
- **Sender hardware address.** This is a variable-length field defining the physical address of the sender. For example, for Ethernet this field is 6 bytes long.
- **Sender protocol address.** This is a variable-length field defining the logical (for example, IP) address of the sender. For the IP protocol, this field is 4 bytes long.
- **Target hardware address.** This is a variable-length field defining the physical address of the target. For example, for Ethernet this field is 6 bytes long. For an ARP request message, this field is all 0s because the sender does not know the physical address of the target.
- **Target protocol address.** This is a variable-length field defining the logical (for example, IP) address of the target. For the IPv4 protocol, this field is 4 bytes long.

Encapsulation

An ARP packet is encapsulated directly into a data link frame. For example, in Figure 21.3 an ARP packet is encapsulated in an Ethernet frame. Note that the type field indicates that the data carried by the frame are an ARP packet.

Figure 21.3 *Encapsulation of ARP packet*



Operation

Let us see how ARP functions on a typical internet. First we describe the steps involved. Then we discuss the four cases in which a host or router needs to use ARP. These are the steps involved in an ARP process:

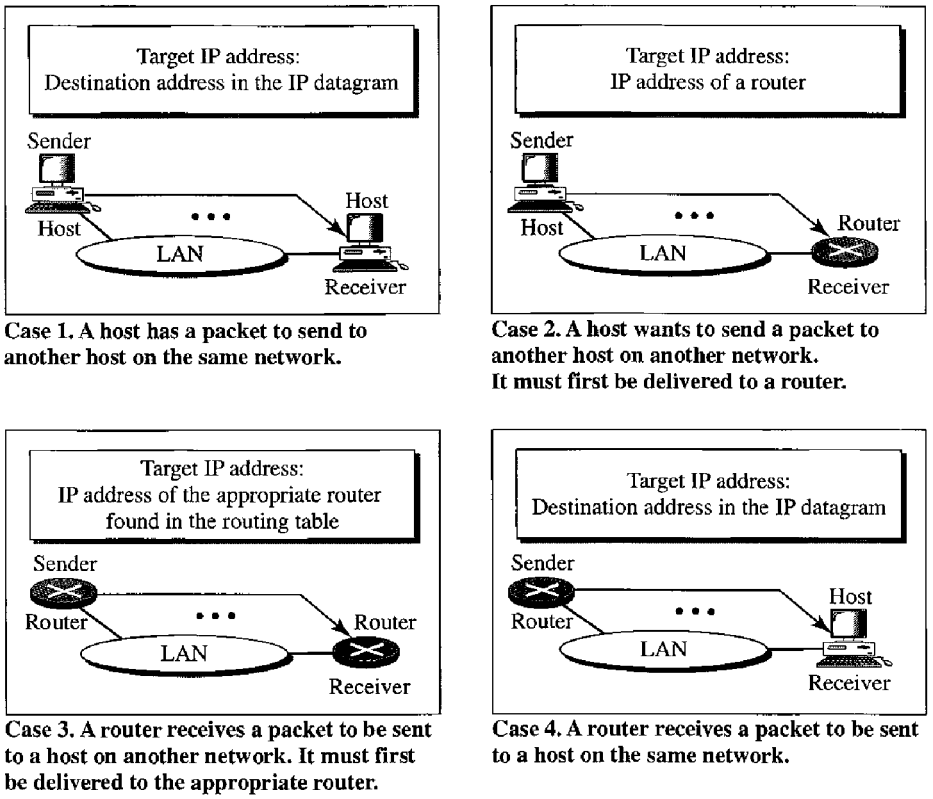
1. The sender knows the IP address of the target. We will see how the sender obtains this shortly.
2. IP asks ARP to create an ARP request message, filling in the sender physical address, the sender IP address, and the target IP address. The target physical address field is filled with 0s.
3. The message is passed to the data link layer where it is encapsulated in a frame by using the physical address of the sender as the source address and the physical broadcast address as the destination address.
4. Every host or router receives the frame. Because the frame contains a broadcast destination address, all stations remove the message and pass it to ARP. All machines except the one targeted drop the packet. The target machine recognizes its IP address.
5. The target machine replies with an ARP reply message that contains its physical address. The message is unicast.
6. The sender receives the reply message. It now knows the physical address of the target machine.
7. The IP datagram, which carries data for the target machine, is now encapsulated in a frame and is unicast to the destination.

Four Different Cases

The following are four different cases in which the services of ARP can be used (see Figure 21.4).

1. The sender is a host and wants to send a packet to another host on the same network. In this case, the logical address that must be mapped to a physical address is the destination IP address in the datagram header.

Figure 21.4 *Four cases using ARP*



Case 1. A host has a packet to send to another host on the same network.

Case 2. A host wants to send a packet to another host on another network. It must first be delivered to a router.

Case 3. A router receives a packet to be sent to a host on another network. It must first be delivered to the appropriate router.

Case 4. A router receives a packet to be sent to a host on the same network.

2. The sender is a host and wants to send a packet to another host on another network. In this case, the host looks at its routing table and finds the IP address of the next hop (router) for this destination. If it does not have a routing table, it looks for the IP address of the default router. The IP address of the router becomes the logical address that must be mapped to a physical address.
3. The sender is a router that has received a datagram destined for a host on another network. It checks its routing table and finds the IP address of the next router. The IP address of the next router becomes the logical address that must be mapped to a physical address.
4. The sender is a router that has received a datagram destined for a host on the same network. The destination IP address of the datagram becomes the logical address that must be mapped to a physical address.

An ARP request is broadcast; an ARP reply is unicast.

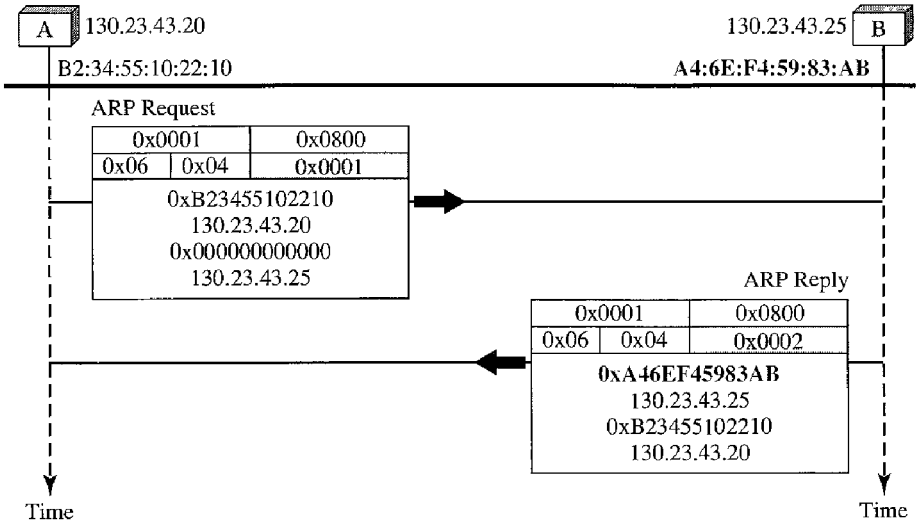
Example 21.1

A host with IP address 130.23.43.20 and physical address B2:34:55:10:22:10 has a packet to send to another host with IP address 130.23.43.25 and physical address A4:6E:F4:59:83:AB (which is unknown to the first host). The two hosts are on the same Ethernet network. Show the ARP request and reply packets encapsulated in Ethernet frames.

Solution

Figure 21.5 shows the ARP request and reply packets. Note that the ARP data field in this case is 28 bytes, and that the individual addresses do not fit in the 4-byte boundary. That is why we do not show the regular 4-byte boundaries for these addresses.

Figure 21.5 *Example 21.1, an ARP request and reply*

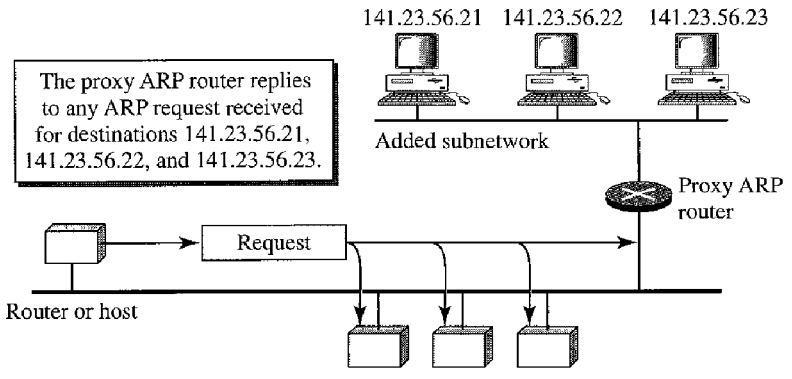


Proxy ARP

A technique called *proxy* ARP is used to create a subnetting effect. A **proxy ARP** is an ARP that acts on behalf of a set of hosts. Whenever a router running a proxy ARP receives an ARP request looking for the IP address of one of these hosts, the router sends an ARP reply announcing its own hardware (physical) address. After the router receives the actual IP packet, it sends the packet to the appropriate host or router.

Let us give an example. In Figure 21.6 the ARP installed on the right-hand host will answer only to an ARP request with a target IP address of 141.23.56.23.

Figure 21.6 *Proxy ARP*



However, the administrator may need to create a subnet without changing the whole system to recognize subnetted addresses. One solution is to add a router running a proxy ARP. In this case, the router acts on behalf of all the hosts installed on the subnet. When it receives an ARP request with a target IP address that matches the address of one of its protégés (141.23.56.21, 141.23.56.22, or 141.23.56.23), it sends an ARP reply and announces its hardware address as the target hardware address. When the router receives the IP packet, it sends the packet to the appropriate host.

Mapping Physical to Logical Address: RARP, BOOTP, and DHCP

There are occasions in which a host knows its physical address, but needs to know its logical address. This may happen in two cases:

1. A diskless station is just booted. The station can find its physical address by checking its interface, but it does not know its IP address.
2. An organization does not have enough IP addresses to assign to each station; it needs to assign IP addresses on demand. The station can send its physical address and ask for a short time lease.

RARP

Reverse Address Resolution Protocol (RARP) finds the logical address for a machine that knows only its physical address. Each host or router is assigned one or more logical (IP) addresses, which are unique and independent of the physical (hardware) address of the machine. To create an IP datagram, a host or a router needs to know its own IP address or addresses. The IP address of a machine is usually read from its configuration file stored on a disk file.

However, a diskless machine is usually booted from ROM, which has minimum booting information. The ROM is installed by the manufacturer. It cannot include the IP address because the IP addresses on a network are assigned by the network administrator.

The machine can get its physical address (by reading its NIC, for example), which is unique locally. It can then use the physical address to get the logical address by using the RARP protocol. A RARP request is created and broadcast on the local network. Another machine on the local network that knows all the IP addresses will respond with a RARP reply. The requesting machine must be running a RARP client program; the responding machine must be running a RARP server program.

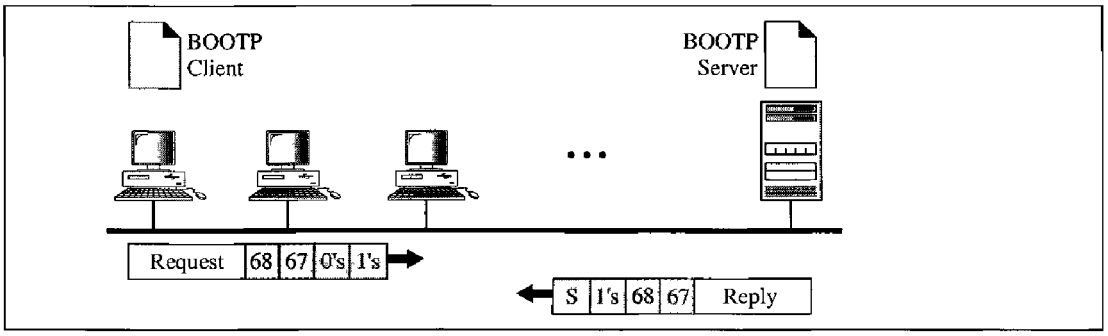
There is a serious problem with RARP: Broadcasting is done at the data link layer. The physical broadcast address, all 1s in the case of Ethernet, does not pass the boundaries of a network. This means that if an administrator has several networks or several subnets, it needs to assign a RARP server for each network or subnet. This is the reason that RARP is almost obsolete. Two protocols, BOOTP and DHCP, are replacing RARP.

BOOTP

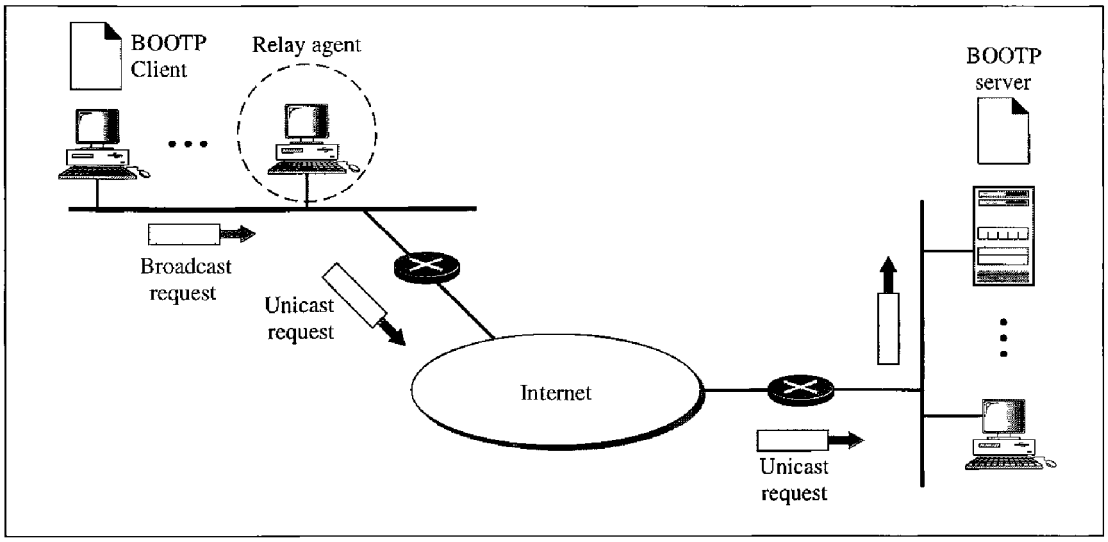
The **Bootstrap Protocol (BOOTP)** is a client/server protocol designed to provide physical address to logical address mapping. BOOTP is an application layer protocol. The administrator may put the client and the server on the same network or on different

networks, as shown in Figure 21.7. BOOTP messages are encapsulated in a UDP packet, and the UDP packet itself is encapsulated in an IP packet.

Figure 21.7 *BOOTP client and server on the same and different network*



a. Client and server on the same network



b. Client and server on different networks

The reader may ask how a client can send an IP datagram when it knows neither its own IP address (the source address) nor the server's IP address (the destination address). The client simply uses all 0s as the source address and all 1s as the destination address.

One of the advantages of BOOTP over RARP is that the client and server are application-layer processes. As in other application-layer processes, a client can be in one network and the server in another, separated by several other networks. However, there is one problem that must be solved. The BOOTP request is broadcast because the client does not know the IP address of the server. A broadcast IP datagram cannot pass through any router. To solve the problem, there is a need for an intermediary. One of the hosts (or a router that can be configured to operate at the application layer) can be used as a relay. The host in this case is called a **relay agent.** The relay agent knows the unicast address of a BOOTP server. When it receives this type of packet, it encapsulates the message in a unicast datagram and sends the request to the BOOTP server. The packet,

carrying a unicast destination address, is routed by any router and reaches the BOOTP server. The BOOTP server knows the message comes from a relay agent because one of the fields in the request message defines the IP address of the relay agent. The relay agent, after receiving the reply, sends it to the BOOTP client.

DHCP

BOOTP is not a **dynamic configuration protocol.** When a client requests its IP address, the BOOTP server consults a table that matches the physical address of the client with its IP address. This implies that the binding between the physical address and the IP address of the client already exists. The binding is predetermined.

However, what if a host moves from one physical network to another? What if a host wants a temporary IP address? BOOTP cannot handle these situations because the binding between the physical and IP addresses is static and fixed in a table until changed by the administrator. BOOTP is a static configuration protocol.

The **Dynamic Host Configuration Protocol (DHCP)** has been devised to provide static and dynamic address allocation that can be manual or automatic.

DHCP provides static and dynamic address allocation that can be manual or automatic.

Static Address Allocation In this capacity DHCP acts as BOOTP does. It is backward-compatible with BOOTP, which means a host running the BOOTP client can request a static address from a DHCP server. A DHCP server has a database that statically binds physical addresses to IP addresses.

Dynamic Address Allocation DHCP has a second database with a pool of available IP addresses. This second database makes DHCP dynamic. When a DHCP client requests a temporary IP address, the DHCP server goes to the pool of available (unused) IP addresses and assigns an IP address for a negotiable period of time.

When a DHCP client sends a request to a DHCP server, the server first checks its static database. If an entry with the requested physical address exists in the static database, the permanent IP address of the client is returned. On the other hand, if the entry does not exist in the static database, the server selects an IP address from the available pool, assigns the address to the client, and adds the entry to the dynamic database.

The dynamic aspect of DHCP is needed when a host moves from network to network or is connected and disconnected from a network (as is a subscriber to a service provider). DHCP provides temporary IP addresses for a limited time.

The addresses assigned from the pool are temporary addresses. The DHCP server issues a **lease** for a specific time. When the lease expires, the client must either stop using the IP address or renew the lease. The server has the option to agree or disagree with the renewal. If the server disagrees, the client stops using the address.

Manual and Automatic Configuration One major problem with the BOOTP protocol is that the table mapping the IP addresses to physical addresses needs to be manually configured. This means that every time there is a change in a physical or IP address, the administrator needs to manually enter the changes. DHCP, on the other hand, allows both manual and automatic configurations. Static addresses are created manually; dynamic addresses are created automatically.

21.2 ICMP

As discussed in Chapter 20, the IP provides unreliable and connectionless datagram delivery. It was designed this way to make efficient use of network resources. The IP protocol is a best-effort delivery service that delivers a datagram from its original source to its final destination. However, it has two deficiencies: lack of error control and lack of assistance mechanisms.

The IP protocol has no error-reporting or error-correcting mechanism. What happens if something goes wrong? What happens if a router must discard a datagram because it cannot find a router to the final destination, or because the time-to-live field has a zero value? What happens if the final destination host must discard all fragments of a datagram because it has not received all fragments within a predetermined time limit? These are examples of situations where an error has occurred and the IP protocol has no built-in mechanism to notify the original host.

The IP protocol also lacks a mechanism for host and management queries. A host sometimes needs to determine if a router or another host is alive. And sometimes a network administrator needs information from another host or router.

The **Internet Control Message Protocol (ICMP)** has been designed to compensate for the above two deficiencies. It is a companion to the IP protocol.

Types of Messages

ICMP messages are divided into two broad categories: **error-reporting messages** and **query messages.**

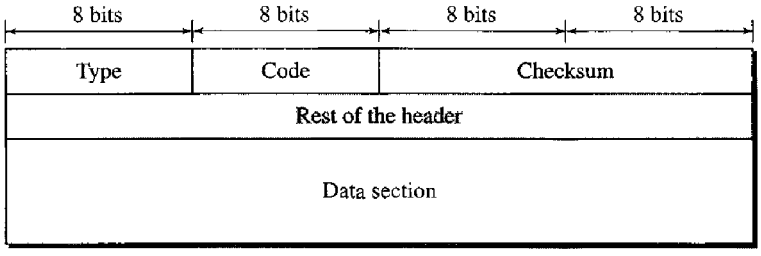
The error-reporting messages report problems that a router or a host (destination) may encounter when it processes an IP packet.

The query messages, which occur in pairs, help a host or a network manager get specific information from a router or another host. For example, nodes can discover their neighbors. Also, hosts can discover and learn about routers on their network, and routers can help a node redirect its messages.

Message Format

An ICMP message has an 8-byte header and a variable-size data section. Although the general format of the header is different for each message type, the first 4 bytes are common to all. As Figure 21.8 shows, the first field, ICMP type, defines the type of the

Figure 21.8 *General format of ICMP messages*



message. The code field specifies the reason for the particular message type. The last common field is the checksum field (to be discussed later in the chapter). The rest of the header is specific for each message type.

The data section in error messages carries information for finding the original packet that had the error. In query messages, the data section carries extra information based on the type of the query.

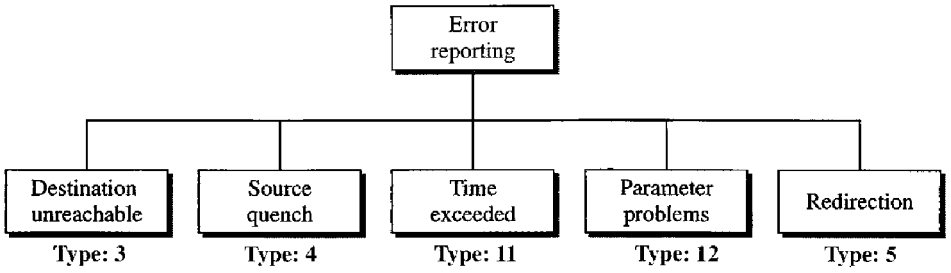
Error Reporting

One of the main responsibilities of ICMP is to report errors. Although technology has produced increasingly reliable transmission media, errors still exist and must be handled. IP, as discussed in Chapter 20, is an unreliable protocol. This means that error checking and error control are not a concern of IP. ICMP was designed, in part, to compensate for this shortcoming. However, ICMP does not correct errors—it simply reports them. Error correction is left to the higher-level protocols. Error messages are always sent to the original source because the only information available in the datagram about the route is the source and destination IP addresses. ICMP uses the source IP address to send the error message to the source (originator) of the datagram.

ICMP always reports error messages to the original source.

Five types of errors are handled: destination unreachable, source quench, time exceeded, parameter problems, and redirection (see Figure 21.9).

Figure 21.9 *Error-reporting messages*

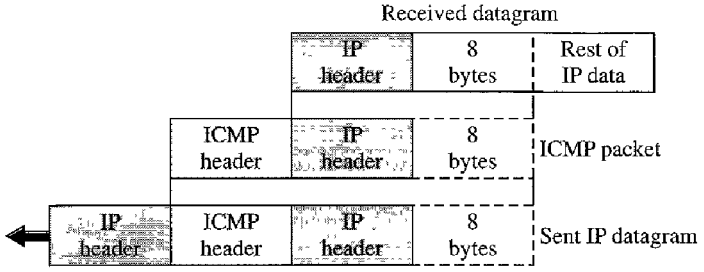


The following are important points about ICMP error messages:

- **No ICMP error message will be generated in response to a datagram carrying an ICMP error message.**
- **No ICMP error message will be generated for a fragmented datagram that is not the first fragment.**
- **No ICMP error message will be generated for a datagram having a multicast address.**
- **No ICMP error message will be generated for a datagram having a special address such as 127.0.0.0 or 0.0.0.0.**

Note that all error messages contain a data section that includes the IP header of the original datagram plus the first 8 bytes of data in that datagram. The original datagram header is added to give the original source, which receives the error message, information about the datagram itself. The 8 bytes of data are included because, as we will see in Chapter 23 on UDP and TCP protocols, the first 8 bytes provide information about the port numbers (UDP and TCP) and sequence number (TCP). This information is needed so the source can inform the protocols (TCP or UDP) about the error. ICMP forms an error packet, which is then encapsulated in an IP datagram (see Figure 21.10).

Figure 21.10 *Contents of data field for the error messages*



Destination Unreachable

When a router cannot route a datagram or a host cannot deliver a datagram, the datagram is discarded and the router or the host sends a **destination-unreachable message** back to the source host that initiated the datagram. Note that destination-unreachable messages can be created by either a router or the destination host.

Source Quench

The IP protocol is a connectionless protocol. There is no communication between the source host, which produces the datagram, the routers, which forward it, and the destination host, which processes it. One of the ramifications of this absence of communication is the lack of *flow control*. IP does not have a flow control mechanism embedded in the protocol. The lack of flow control can create a major problem in the operation of IP: congestion. The source host never knows if the routers or the destination host has been overwhelmed with datagrams. The source host never knows if it is producing datagrams faster than can be forwarded by routers or processed by the destination host.

The lack of flow control can create congestion in routers or the destination host. A router or a host has a limited-size queue (buffer) for incoming datagrams waiting to be forwarded (in the case of a router) or to be processed (in the case of a host). If the datagrams are received much faster than they can be forwarded or processed, the queue may overflow. In this case, the router or the host has no choice but to discard some of the datagrams. The **source-quench message** in ICMP was designed to add a kind of flow control to the IP. When a router or host discards a datagram due to congestion, it sends a source-quench message to the sender of the datagram. This message has two purposes. First, it informs the source that the datagram has been discarded. Second, it warns the source that there is congestion somewhere in the path and that the source should slow down (quench) the sending process.

Time Exceeded

The **time-exceeded message** is generated in two cases: As we see in Chapter 22, routers use routing tables to find the next hop (next router) that must receive the packet. If there are errors in one or more routing tables, a packet can travel in a loop or a cycle, going from one router to the next or visiting a series of routers endlessly. As we saw in Chapter 20, each datagram contains a field called *time to live* that controls this situation. When a datagram visits a router, the value of this field is decremented by 1. When the time-to-live value reaches 0, after decrementing, the router discards the datagram. However, when the datagram is discarded, a time-exceeded message must be sent by the router to the original source. Second, a time-exceeded message is also generated when not all fragments that make up a message arrive at the destination host within a certain time limit.

Parameter Problem

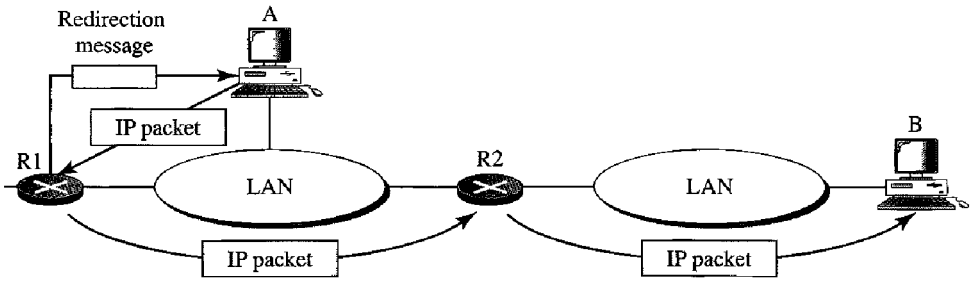
Any ambiguity in the header part of a datagram can create serious problems as the datagram travels through the Internet. If a router or the destination host discovers an ambiguous or missing value in any field of the datagram, it discards the datagram and sends a **parameter-problem message** back to the source.

Redirection

When a router needs to send a packet destined for another network, it must know the IP address of the next appropriate router. The same is true if the sender is a host. Both routers and hosts, then, must have a routing table to find the address of the router or the next router. Routers take part in the routing update process, as we will see in Chapter 22, and are supposed to be updated constantly. Routing is dynamic.

However, for efficiency, hosts do not take part in the routing update process because there are many more hosts in an internet than routers. Updating the routing tables of hosts dynamically produces unacceptable traffic. The hosts usually use static routing. When a host comes up, its routing table has a limited number of entries. It usually knows the IP address of only one router, the default router. For this reason, the host may send a datagram, which is destined for another network, to the wrong router. In this case, the router that receives the datagram will forward the datagram to the correct router. However, to update the routing table of the host, it sends a redirection message to the host. This concept of redirection is shown in Figure 21.11. Host A wants to send a datagram to host B.

Figure 21.11 *Redirection concept*



Router R2 is obviously the most efficient routing choice, but host A did not choose router R2. The datagram goes to R1 instead. Router R1, after consulting its table, finds that the packet should have gone to R2. It sends the packet to R2 and, at the same time, sends a redirection message to host A. Host A's routing table can now be updated.

Query

In addition to error reporting, ICMP can diagnose some network problems. This is accomplished through the query messages, a group of four different pairs of messages, as shown in Figure 21.12. In this type of ICMP message, a node sends a message that is answered in a specific format by the destination node. A query message is encapsulated in an IP packet, which in turn is encapsulated in a data link layer frame. However, in this case, no bytes of the original IP are included in the message, as shown in Figure 21.13.

Figure 21.12 *Query messages*

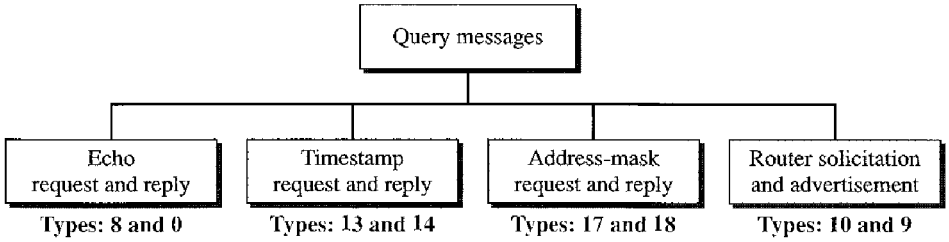
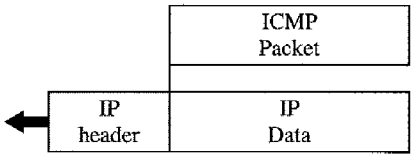


Figure 21.13 *Encapsulation of ICMP query messages*



Echo Request and Reply

The **echo-request** and **echo-reply messages** are designed for diagnostic purposes. Network managers and users utilize this pair of messages to identify network problems. The combination of echo-request and echo-reply messages determines whether two systems (hosts or routers) can communicate with each other. The echo-request and echo-reply messages can be used to determine if there is communication at the IP level. Because ICMP messages are encapsulated in IP datagrams, the receipt of an echo-reply message by the machine that sent the echo request is proof that the IP protocols in the sender and receiver are communicating with each other using the IP datagram. Also, it is proof that the intermediate routers are receiving, processing, and forwarding IP datagrams. Today, most systems provide a version of the *ping* command that can create a series (instead of just one) of echo-request and echo-reply messages, providing statistical information. We will see the use of this program at the end of the chapter.

Timestamp Request and Reply

Two machines (hosts or routers) can use the **timestamp request and timestamp reply messages** to determine the round-trip time needed for an IP datagram to travel between them. It can also be used to synchronize the clocks in two machines.

Address-Mask Request and Reply

A host may know its IP address, but it may not know the corresponding mask. For example, a host may know its IP address as 159.31.17.24, but it may not know that the corresponding mask is /24. To obtain its mask, a host sends an **address-mask-request message** to a router on the LAN. If the host knows the address of the router, it sends the request directly to the router. If it does not know, it broadcasts the message. The router receiving the address-mask-request message responds with an **address-mask-reply message,** providing the necessary mask for the host. This can be applied to its full IP address to get its subnet address.

Router Solicitation and Advertisement

As we discussed in the redirection message section, a host that wants to send data to a host on another network needs to know the address of routers connected to its own network. Also, the host must know if the routers are alive and functioning. The **router-solicitation and router-advertisement messages** can help in this situation. A host can broadcast (or multicast) a router-solicitation message. The router or routers that receive the solicitation message broadcast their routing information using the router-advertisement message. A router can also periodically send router-advertisement messages even if no host has solicited. Note that when a router sends out an advertisement, it announces not only its own presence but also the presence of all routers on the network of which it is aware.

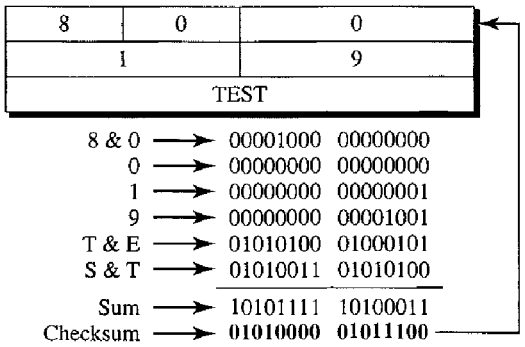
Checksum

In Chapter 10, we learned the concept and idea of the checksum. In ICMP the checksum is calculated over the entire message (header and data).

Example 21.2

Figure 21.14 shows an example of checksum calculation for a simple echo-request message. We randomly chose the identifier to be 1 and the sequence number to be 9. The message is divided

Figure 21.14 *Example of checksum calculation*



into 16-bit (2-byte) words. The words are added and the sum is complemented. Now the sender can put this value in the checksum field.

Debugging Tools

There are several tools that can be used in the Internet for debugging. We can determine the viability of a host or router. We can trace the route of a packet. We introduce two tools that use ICMP for debugging: *ping* and *traceroute*. We will introduce more tools in future chapters after we have discussed the corresponding protocols.

Ping

We can use the *ping* program to find if a host is alive and responding. We use *ping* here to see how it uses ICMP packets.

The source host sends ICMP echo-request messages (type: 8, code: 0); the destination, if alive, responds with ICMP echo-reply messages. The *ping* program sets the identifier field in the echo-request and echo-reply message and starts the sequence number from 0; this number is incremented by 1 each time a new message is sent. Note that *ping* can calculate the round-trip time. It inserts the sending time in the data section of the message. When the packet arrives, it subtracts the arrival time from the departure time to get the round-trip time (RTT).

Example 21.3

We use the *ping* program to test the server fhda.edu. The result is shown below:

```
$ ping fhda.edu
PING fhda.edu (153.18.8.1)  56 (84)  bytes of data.
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=0    ttl=62    time=1.91 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=1    ttl=62    time=2.04 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=2    ttl=62    time=1.90 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=3    ttl=62    time=1.97 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=4    ttl=62    time=1.93 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=5    ttl=62    time=2.00 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=6    ttl=62    time=1.94 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=7    ttl=62    time=1.94 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=8    ttl=62    time=1.97 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=9    ttl=62    time=1.89 ms
64 bytes from tiptoe.fhda.edu (153.18.8.1): icmp_seq=10   ttl=62    time=1.98 ms

--- fhda.edu ping statistics ---
11 packets transmitted, 11 received, 0% packet loss, time 10103ms
    rtt min/avg/max = 1.899/1.955/2.041 ms
```

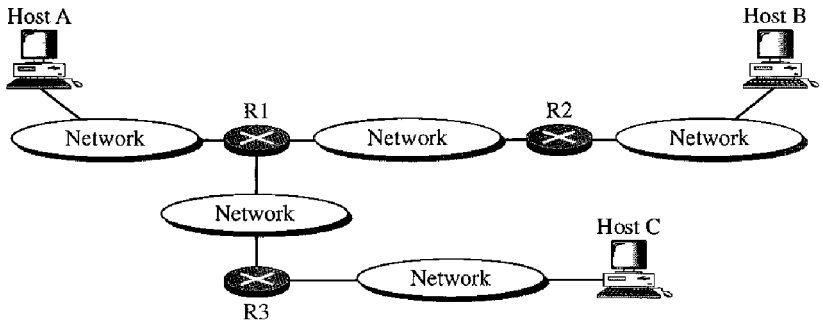
The *ping* program sends messages with sequence numbers starting from 0. For each probe it gives us the RTT time. The TTL (time to live) field in the IP datagram that encapsulates an ICMP message has been set to 62, which means the packet cannot travel more than 62 hops. At the beginning, *ping* defines the number of data bytes as 56 and the total number of bytes as 84. It is obvious that if we add 8 bytes of ICMP header and 20 bytes of IP header to 56, the result is 84. However,

note that in each probe *ping* defines the number of bytes as 64. This is the total number of bytes in the ICMP packet (56 + 8). The *ping* program continues to send messages, if we do not stop it by using the interrupt key (ctrl + c, for example). After it is interrupted, it prints the statistics of the probes. It tells us the number of packets sent, the number of packets received, the total time, and the RTT minimum, maximum, and average. Some systems may print more information.

Traceroute

The ***traceroute*** program in UNIX or ***tracert*** in Windows can be used to trace the route of a packet from the source to the destination. We have seen an application of the *traceroute* program to simulate the loose source route and strict source route options of an IP datagram in Chapter 20. We use this program in conjunction with ICMP packets in this chapter. The program elegantly uses two ICMP messages, time exceeded and destination unreachable, to find the route of a packet. This is a program at the application level that uses the services of UDP (see Chapter 23). Let us show the idea of the *traceroute* program by using Figure 21.15.

Figure 21.15 *The traceroute program operation*



Given the topology, we know that a packet from host A to host B travels through routers R1 and R2. However, most of the time, we are not aware of this topology. There could be several routes from A to B. The *traceroute* program uses the ICMP messages and the TTL (time to live) field in the IP packet to find the route.

1. The *traceroute* program uses the following steps to find the address of the router R1 and the round-trip time between host A and router R1.
   a. The *traceroute* application at host A sends a packet to destination B using UDP; the message is encapsulated in an IP packet with a TTL value of 1. The program notes the time the packet is sent.
   b. Router R1 receives the packet and decrements the value of TTL to 0. It then discards the packet (because TTL is 0). The router, however, sends a time-exceeded ICMP message (type: 11, code: 0) to show that the TTL value is 0 and the packet was discarded.
   c. The *traceroute* program receives the ICMP messages and uses the destination address of the IP packet encapsulating ICMP to find the IP address of router R1. The program also makes note of the time the packet has arrived. The difference between this time and the time at step a is the round-trip time.

The *traceroute* program repeats steps a to c three times to get a better average round-trip time. The first trip time may be much longer than the second or third because it takes time for the ARP program to find the physical address of router R1. For the second and third trips, ARP has the address in its cache.

2. The *traceroute* program repeats the previous steps to find the address of router R2 and the round-trip time between host A and router R2. However, in this step, the value of TTL is set to 2. So router R1 forwards the message, while router R2 discards it and sends a time-exceeded ICMP message.
3. The *traceroute* program repeats step 2 to find the address of host B and the round-trip time between host A and host B. When host B receives the packet, it decrements the value of TTL, but it does not discard the message since it has reached its final destination. How can an ICMP message be sent back to host A? The *traceroute* program uses a different strategy here. The destination port of the UDP packet is set to one that is not supported by the UDP protocol. When host B receives the packet, it cannot find an application program to accept the delivery. It discards the packet and sends an ICMP destination-unreachable message (type: 3, code: 3) to host A. Note that this situation does not happen at router R1 or R2 because a router does not check the UDP header. The *traceroute* program records the destination address of the arrived IP datagram and makes note of the round-trip time. Receiving the destination-unreachable message with a code value 3 is an indication that the whole route has been found and there is no need to send more packets.

Example 21.4

We use the *traceroute* program to find the route from the computer voyager.deanza.edu to the server fhda.edu. The following shows the result:

```
$ traceroute fhda.edu
traceroute to fhda.edu (153.18.8.1), 30 hops max, 38 byte packets
 1 Dcore.fhda.edu (153.18.31.254)  0.995 ms  0.899 ms  0.878 ms
 2 Dbackup.fhda.edu (153.18.251.4)  1.039 ms  1.064 ms  1.083 ms
 3 tiptoe.fhda.edu (153.18.8.1)  1.797 ms  1.642 ms  1.757 ms
```

The unnumbered line after the command shows that the destination is 153.18.8.1. The TTL value is 30 hops. The packet contains 38 bytes: 20 bytes of IP header, 8 bytes of UDP header, and 10 bytes of application data. The application data are used by *traceroute* to keep track of the packets.

The first line shows the first router visited. The router is named Dcore.fhda.edu with IP address 153.18.31.254. The first round-trip time was 0.995 ms, the second was 0.899 ms, and the third was 0.878 ms.

The second line shows the second router visited. The router is named Dbackup.fhda.edu with IP address 153.18.251.4. The three round-trip times are also shown.

The third line shows the destination host. We know that this is the destination host because there are no more lines. The destination host is the server fhda.edu, but it is named tiptoe.fhda.edu with the IP address 153.18.8.1. The three round-trip times are also shown.

Example 21.5

In this example, we trace a longer route, the route to xerox.com.

```
$ traceroute xerox.com
traceroute to xerox.com (13.1.64.93), 30 hops max, 38 byte packets
 1  Dcore.fhda.edu      (153.18.31.254)    0.622 ms   0.891 ms   0.875 ms
 2  Ddmz.fhda.edu       (153.18.251.40)    2.132 ms   2.266 ms   2.094 ms
 3  Cinic.fhda.edu      (153.18.253.126)   2.110 ms   2.145 ms   1.763 ms
 4  cenic.net           (137.164.32.140)   3.069 ms   2.875 ms   2.930 ms
 5  cenic.net           (137.164.22.31)    4.205 ms   4.870 ms   4.197 ms
    ....                ....               ...        ....       ...
14  snfc21.pbi.net      (151.164.191.49)   7.656 ms   7.129 ms   6.866 ms
15  sbcglobal.net       (151.164.243.58)   7.844 ms   7.545 ms   7.353 ms
16  pacbell.net         (209.232.138.114)  9.857 ms   9.535 ms   9.603 ms
17  209.233.48.223      (209.233.48.223)   10.634 ms  10.771 ms  10.592 ms
18  alpha.Xerox.COM     (13.1.64.93)       11.172 ms  11.048 ms  10.922 ms
```

Here there are 17 hops between source and destination. Note that some round-trip times look unusual. It could be that a router was too busy to process the packet immediately.

21.3 IGMP

The IP protocol can be involved in two types of communication: unicasting and multicasting. Unicasting is the communication between one sender and one receiver. It is a one-to-one communication. However, some processes sometimes need to send the same message to a large number of receivers simultaneously. This is called **multicasting,** which is a one-to-many communication. Multicasting has many applications. For example, multiple stockbrokers can simultaneously be informed of changes in a stock price, or travel agents can be informed of a plane cancellation. Some other applications include distance learning and video-on-demand.

The **Internet Group Management Protocol (IGMP)** is one of the necessary, but not sufficient (as we will see), protocols that is involved in multicasting. IGMP is a companion to the IP protocol.

Group Management

For multicasting in the Internet we need routers that are able to route multicast packets. The routing tables of these routers must be updated by using one of the multicasting routing protocols that we discuss in Chapter 22.

IGMP is not a multicasting routing protocol; it is a protocol that manages **group membership.** In any network, there are one or more multicast routers that distribute multicast packets to hosts or other routers. The IGMP protocol gives the **multicast routers** information about the membership status of hosts (routers) connected to the network.

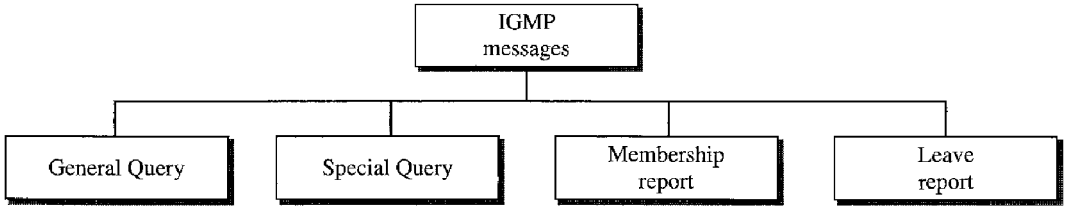
A multicast router may receive thousands of multicast packets every day for different groups. If a router has no knowledge about the membership status of the hosts, it must broadcast all these packets. This creates a lot of traffic and consumes bandwidth. A better solution is to keep a list of groups in the network for which there is at least one loyal member. IGMP helps the multicast router create and update this list.

IGMP is a group management protocol. It helps a multicast router create and update a list of loyal members related to each router interface.

IGMP Messages

IGMP has gone through two versions. We discuss IGMPv2, the current version. IGMPv2 has three types of **messages:** the **query,** the **membership report,** and the **leave report.** There are two types of **query messages: general** and **special** (see Figure 21.16).

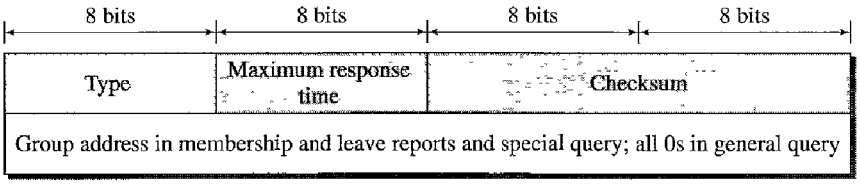
Figure 21.16 *IGMP message types*



Message Format

Figure 21.17 shows the format of an IGMP (version 2) message.

Figure 21.17 *IGMP message format*



- **Type.** This 8-bit field defines the type of message, as shown in Table 21.1. The value of the type is shown in both hexadecimal and binary notation.

Table 21.1 *IGMP type field*

Type	*Value*
General or special query	0x11 or 00010001
Membership report	0x16 or 00010110
Leave report	0x17 or 00010111

- **Maximum Response Time.** This 8-bit field defines the amount of time in which a query must be answered. The value is in tenths of a second; for example, if the

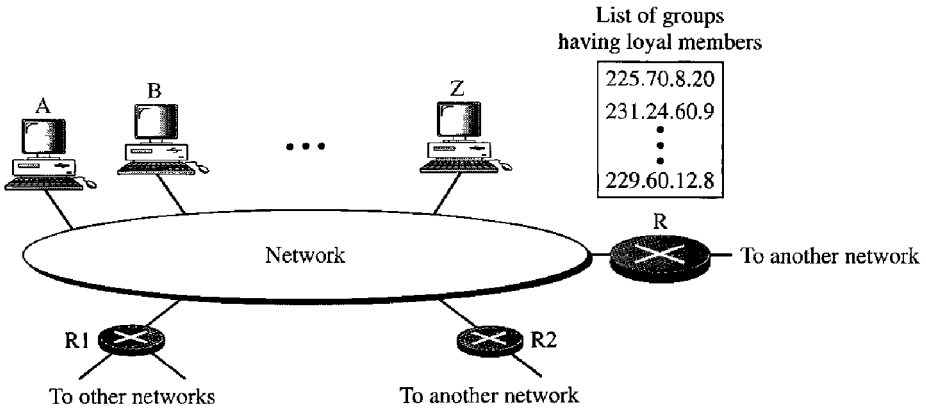
value is 100, it means 10 s. The value is nonzero in the query message; it is set to zero in the other two message types. We will see its use shortly.

- **Checksum.** This is a 16-bit field carrying the checksum. The checksum is calculated over the 8-byte message.
- **Group address.** The value of this field is 0 for a general query message. The value defines the groupid (multicast address of the group) in the special query, the membership report, and the leave report messages.

IGMP Operation

IGMP operates locally. A multicast router connected to a network has a list of multicast addresses of the groups with at least one loyal member in that network (see Figure 21.18).

Figure 21.18 *IGMP operation*



For each group, there is one router that has the duty of distributing the multicast packets destined for that group. This means that if there are three multicast routers connected to a network, their lists of **groupids** are mutually exclusive. For example, in Figure 21.18 only router R distributes packets with the multicast address of 225.70.8.20.

A host or multicast router can have membership in a group. When a host has membership, it means that one of its processes (an application program) receives multicast packets from some group. When a router has membership, it means that a network connected to one of its other interfaces receives these multicast packets. We say that the host or the router has an *interest* in the group. In both cases, the host and the router keep a list of groupids and relay their interest to the distributing router.

For example, in Figure 21.18, router R is the distributing router. There are two other multicast routers (R1 and R2) that, depending on the group list maintained by router R, could be the recipients of router R in this network. Routers R1 and R2 may be distributors for some of these groups in other networks, but not on this network.

Joining a Group

A host or a router can join a group. A host maintains a list of processes that have membership in a group. When a process wants to join a new group, it sends its request to the host.

The host adds the name of the process and the name of the requested group to its list. If this is the first entry for this particular group, the host sends a membership report message. If this is not the first entry, there is no need to send the membership report since the host is already a member of the group; it already receives multicast packets for this group.

The protocol requires that the membership report be sent twice, one after the other within a few moments. In this way, if the first one is lost or damaged, the second one replaces it.

In IGMP, a membership report is sent twice, one after the other.

Leaving a Group

When a host sees that no process is interested in a specific group, it sends a leave report. Similarly, when a router sees that none of the networks connected to its interfaces is interested in a specific group, it sends a leave report about that group.

However, when a multicast router receives a leave report, it cannot immediately purge that group from its list because the report comes from just one host or router; there may be other hosts or routers that are still interested in that group. To make sure, the router sends a special query message and inserts the groupid, or **multicast address,** related to the group. The router allows a specified time for any host or router to respond. If, during this time, no interest (membership report) is received, the router assumes that there are no loyal members in the network and purges the group from its list.

Monitoring Membership

A host or router can join a group by sending a membership report message. It can leave a group by sending a leave report message. However, sending these two types of reports is not enough. Consider the situation in which there is only one host interested in a group, but the host is shut down or removed from the system. The multicast router will never receive a leave report. How is this handled? The multicast router is responsible for monitoring all the hosts or routers in a LAN to see if they want to continue their membership in a group.

The router periodically (by default, every 125 s) sends a general query message. In this message, the group address field is set to 0.0.0.0. This means the query for membership continuation is for all groups in which a host is involved, not just one.

The general query message does not define a particular group.

The router expects an answer for each group in its group list; even new groups may respond. The query message has a maximum response time of 10 s (the value of the field is actually 100, but this is in tenths of a second). When a host or router receives the general query message, it responds with a membership report if it is interested in a group. However, if there is a common interest (two hosts, for example, are interested in the same group), only one response is sent for that group to prevent unnecessary traffic. This is called a delayed response. Note that the query message must be sent by only one

router (normally called the query router), also to prevent unnecessary traffic. We discuss this issue shortly.

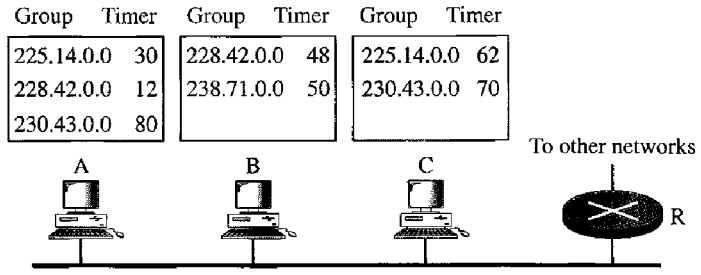
Delayed Response

To prevent unnecessary traffic, IGMP uses a **delayed response strategy.** When a host or router receives a query message, it does not respond immediately; it delays the response. Each host or router uses a random number to create a timer, which expires between 1 and 10 s. The expiration time can be in steps of 1 s or less. A timer is set for each group in the list. For example, the timer for the first group may expire in 2 s, but the timer for the third group may expire in 5 s. Each host or router waits until its timer has expired before sending a membership report message. During this waiting time, if the timer of another host or router, for the same group, expires earlier, that host or router sends a membership report. Because, as we will see shortly, the report is broadcast, the waiting host or router receives the report and knows that there is no need to send a duplicate report for this group; thus, the waiting station cancels its corresponding timer.

Example 21.6

Imagine there are three hosts in a network, as shown in Figure 21.19.

Figure 21.19 *Example 21.6*



A query message was received at time 0; the random delay time (in tenths of seconds) for each group is shown next to the group address. Show the sequence of report messages.

Solution

The events occur in this sequence:

a. **Time 12:** The timer for 228.42.0.0 in host A expires, and a membership report is sent, which is received by the router and every host including host B which cancels its timer for 228.42.0.0.

b. **Time 30:** The timer for 225.14.0.0 in host A expires, and a membership report is sent, which is received by the router and every host including host C which cancels its timer for 225.14.0.0.

c. **Time 50:** The timer for 238.71.0.0 in host B expires, and a membership report is sent, which is received by the router and every host.

d. **Time 70:** The timer for 230.43.0.0 in host C expires, and a membership report is sent, which is received by the router and every host including host A which cancels its timer for 230.43.0.0.

Note that if each host had sent a report for every group in its list, there would have been seven reports; with this strategy only four reports are sent.

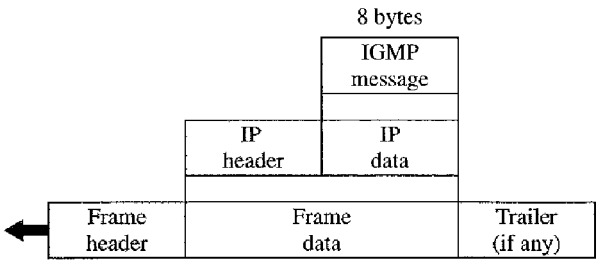
Query Router

Query messages may create a lot of responses. To prevent unnecessary traffic, IGMP designates one router as the **query router** for each network. Only this designated router sends the query message, and the other routers are passive (they receive responses and update their lists).

Encapsulation

The IGMP message is encapsulated in an IP datagram, which is itself encapsulated in a frame. See Figure 21.20.

Figure 21.20 *Encapsulation of IGMP packet*



Encapsulation at Network Layer

The value of the protocol field is 2 for the IGMP protocol. Every IP packet carrying this value in its protocol field has data delivered to the IGMP protocol. When the message is encapsulated in the IP datagram, the value of TTL must be 1. This is required because the domain of IGMP is the LAN. No IGMP message must travel beyond the LAN. A TTL value of 1 guarantees that the message does not leave the LAN since this value is decremented to 0 by the next router and, consequently, the packet is discarded. Table 21.2 shows the destination IP address for each type of message.

The IP packet that carries an IGMP packet has a value of 1 in its TTL field.

Table 21.2 *Destination IP addresses*

Type	*IP Destination Address*
Query	224.0.0.1 All systems on this subnet
Membership report	The multicast address of the group
Leave report	224.0.0.2 All routers on this subnet

A query message is multicast by using the multicast address 224.0.0.1 All hosts and all routers will receive the message. A membership report is multicast using a

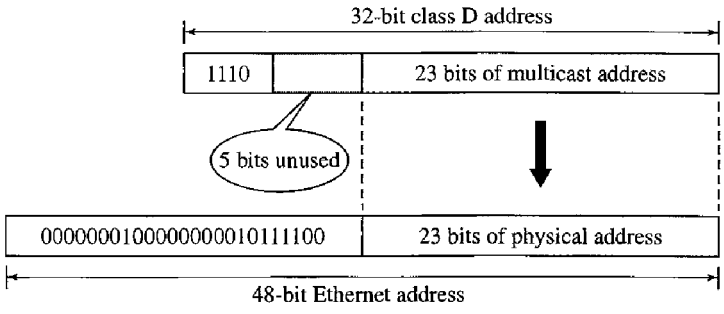
destination address equal to the multicast address being reported (groupid). Every station (host or router) that receives the packet can immediately determine (from the header) the group for which a report has been sent. As discussed previously, the timers for the corresponding unsent reports can then be canceled. Stations do not need to open the packet to find the groupid. This address is duplicated in a packet; it's part of the message itself and also a field in the IP header. The duplication prevents errors. A leave report message is multicast using the multicast address 224.0.0.2 (all routers on this subnet) so that routers receive this type of message. Hosts receive this message too, but disregard it.

Encapsulation at Data Link Layer

At the network layer, the IGMP message is encapsulated in an IP packet and is treated as an IP packet. However, because the IP packet has a multicast IP address, the ARP protocol cannot find the corresponding MAC (physical) address to forward the packet at the data link layer. What happens next depends on whether the underlying data link layer supports physical multicast addresses.

Physical Multicast Support Most LANs support physical multicast addressing. Ethernet is one of them. An Ethernet physical address (MAC address) is six octets (48 bits) long. If the first 25 bits in an Ethernet address are 00000001000000000010111100, this identifies a physical multicast address for the TCP/IP protocol. The remaining 23 bits can be used to define a group. To convert an IP multicast address into an Ethernet address, the multicast router extracts the least significant 23 bits of a class D IP address and inserts them into a multicast Ethernet physical address (see Figure 21.21).

Figure 21.21 *Mapping class D to Ethernet physical address*



However, the group identifier of a class D IP address is 28 bits long, which implies that 5 bits is not used. This means that 32 (2^5) multicast addresses at the IP level are mapped to a single multicast address. In other words, the mapping is many-to-one instead of one-to-one. If the 5 leftmost bits of the group identifier of a class D address are not all zeros, a host may receive packets that do not really belong to the group in which it is involved. For this reason, the host must check the IP address and discard any packets that do not belong to it.

Other LANs support the same concept but have different methods of mapping.

An Ethernet multicast physical address is in the range 01:00:5E:00:00:00 to 01:00:5E:7F:FF:FF.

Example 21.7

Change the multicast IP address 230.43.14.7 to an Ethernet multicast physical address.

Solution

We can do this in two steps:

a. We write the rightmost 23 bits of the IP address in hexadecimal. This can be done by changing the rightmost 3 bytes to hexadecimal and then subtracting 8 from the leftmost digit if it is greater than or equal to 8. In our example, the result is 2B:0E:07.

b. We add the result of part a to the starting Ethernet multicast address, which is 01:00:5E:00:00:00. The result is

01:00:5E:2B:0E:07

Example 21.8

Change the multicast IP address 238.212.24.9 to an Ethernet multicast address.

Solution

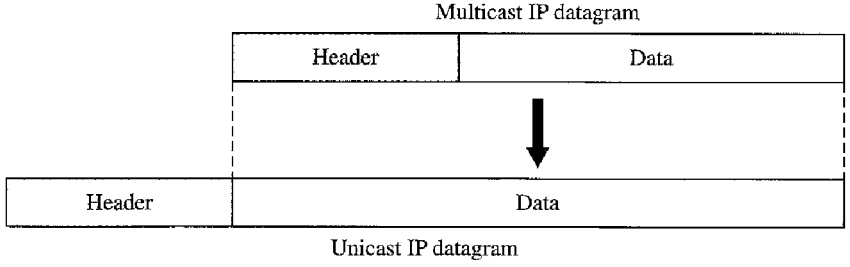
a. The rightmost 3 bytes in hexadecimal is D4:18:09. We need to subtract 8 from the leftmost digit, resulting in 54:18:09.

b. We add the result of part a to the Ethernet multicast starting address. The result is

01:00:5E:54:18:09

No Physical Multicast Support Most WANs do not support physical multicast addressing. To send a multicast packet through these networks, a process called *tunneling* is used. In **tunneling,** the multicast packet is encapsulated in a unicast packet and sent through the network, where it emerges from the other side as a multicast packet (see Figure 21.22).

Figure 21.22 *Tunneling*



Netstat Utility

The *netstat* utility can be used to find the multicast addresses supported by an interface.

Example 21.9

We use *netstat* with three options: -n, -r, and -a. The -n option gives the numeric versions of IP addresses, the -r option gives the routing table, and the -a option gives all addresses (unicast and multicast). Note that we show only the fields relative to our discussion. "Gateway" defines the router, "Iface" defines the interface.

```
$ netstat -nra
Kernel IP routing table
```

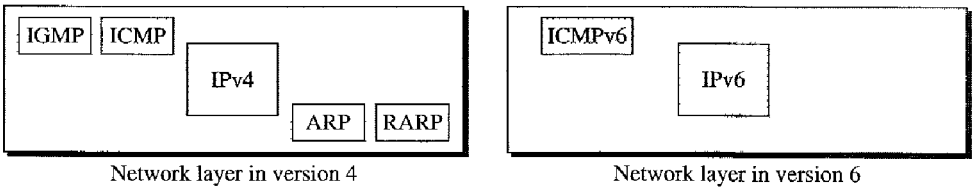
Destination	Gateway	Mask	Flags	Iface
153.18.16.0	0.0.0.0	255.255.240.0	U	eth0
169.254.0.0	0.0.0.0	255.255.0.0	U	eth0
127.0.0.0	0.0.0.0	255.0.0.0	U	lo
224.0.0.0	0.0.0.0	224.0.0.0	U	eth0
0.0.0.0	153.18.31.254	0.0.0.0	UG	eth0

Note that the multicast address is shown in color. Any packet with a multicast address from 224.0.0.0 to 239.255.255.255 is masked and delivered to the Ethernet interface.

21.4 ICMPv6

We discussed IPv6 in Chapter 20. Another protocol that has been modified in version 6 of the TCP/IP protocol suite is ICMP (ICMPv6). This new version follows the same strategy and purposes of version 4. ICMPv4 has been modified to make it more suitable for IPv6. In addition, some protocols that were independent in version 4 are now part of **Internetworking Control Message Protocol (ICMPv6).** Figure 21.23 compares the network layer of version 4 to version 6.

Figure 21.23 *Comparison of network layers in version 4 and version 6*



The ARP and IGMP protocols in version 4 are combined in ICMPv6. The RARP protocol is dropped from the suite because it was rarely used and BOOTP has the same functionality.

Just as in ICMPv4, we divide the ICMP messages into two categories. However, each category has more types of messages than before.

Error Reporting

As we saw in our discussion of version 4, one of the main responsibilities of ICMP is to report errors. Five types of errors are handled: destination unreachable, packet too big, time exceeded, parameter problems, and redirection. ICMPv6 forms an error packet,

which is then encapsulated in an IP datagram. This is delivered to the original source of the failed datagram. Table 21.3 compares the **error-reporting messages** of ICMPv4 with ICMPv6. The source-quench message is eliminated in version 6 because the priority and the flow label fields allow the router to control congestion and discard the least important messages. In this version, there is no need to inform the sender to slow down. The packet-too-big message is added because fragmentation is the responsibility of the sender in IPv6. If the sender does not make the right packet size decision, the router has no choice but to drop the packet and send an error message to the sender.

Table 21.3 *Comparison of error-reporting messages in ICMPv4 and ICMPv6*

Type of Message	*Version 4*	*Version 6*
Destination unreachable	Yes	Yes
Source quench	Yes	No
Packet too big	No	Yes
Time exceeded	Yes	Yes
Parameter problem	Yes	Yes
Redirection	Yes	Yes

Destination Unreachable

The concept of the **destination-unreachable message** is exactly the same as described for ICMP version 4.

Packet Too Big

This is a new type of message added to version 6. If a router receives a datagram that is larger than the maximum transmission unit (MTU) size of the network through which the datagram should pass, two things happen. First, the router discards the datagram and then an ICMP error packet—a **packet-too-big message**—is sent to the source.

Time Exceeded

This message is similar to the one in version 4.

Parameter Problem

This message is similar to its version 4 counterpart.

Redirection

The purpose of the **redirection message** is the same as described for version 4.

Query

In addition to error reporting, ICMP can diagnose some network problems. This is accomplished through the **query messages.** Four different groups of messages have been defined: echo request and reply, router solicitation and advertisement, neighbor solicitation and advertisement, and group membership. Table 21.4 shows a comparison between

the query messages in versions 4 and 6. Two sets of query messages are eliminated from ICMPv6: time-stamp request and reply- and address-mask request and reply. The time-stamp request and reply messages are eliminated because they are implemented in other protocols such as TCP and because they were rarely used in the past. The address-mask request and reply messages are eliminated in IPv6 because the subnet section of an address allows the subscriber to use up to $2^{32} - 1$ subnets. Therefore, subnet masking, as defined in IPv4, is not needed here.

Table 21.4 *Comparison of query messages in ICMPv4 and ICMPv6*

Type of Message	*Version 4*	*Version 6*
Echo request and reply	Yes	Yes
Timestamp request and reply	Yes	No
Address-mask request and reply	Yes	No
Router solicitation and advertisement	Yes	Yes
Neighbor solicitation and advertisement	ARP	Yes
Group membership	IGMP	Yes

Echo Request and Reply

The idea and format of the echo request and reply messages are the same as those in version 4.

Router Solicitation and Advertisement

The idea behind the router-solicitation and -advertisement messages is the same as in version 4.

Neighbor Solicitation and Advertisement

As previously mentioned, the network layer in version 4 contains an independent protocol called Address Resolution Protocol (ARP). In version 6, this protocol is eliminated, and its duties are included in ICMPv6. The idea is exactly the same, but the format of the message has changed.

Group Membership

As previously mentioned, the network layer in version 4 contains an independent protocol called IGMP. In version 6, this protocol is eliminated, and its duties are included in ICMPv6. The purpose is exactly the same.

21.5 RECOMMENDED READING

For more details about subjects discussed in this chapter, we recommend the following books and site. The items in brackets [. . .] refer to the reference list at the end of the text.

Books

ARP and RARP are discussed in Chapter 7 of [For06] and Chapters 4 and 5 of [Ste94]. ICMP is discussed in Chapter 9 of [For06] and Chapter 6 of [Ste94]. IGMP is discussed in Chapter 10 of [For06] and Chapter 13 of [Ste94]. BOOTP and DHCP are discussed in Chapter 16 of [For06] and Chapter 16 of [Ste94]. ICMPv6 is discussed in Chapter 27 of [For06].

Site

- www.ietf.org/rfc.html Information about RFCs

RFCs

A discussion of ARP and RARP can be found in following RFCs:

826, 903, 925, 1027, 1293,1329, 1433, 1868, 1931, 2390

A discussion of ICMP can be found in the following RFCs:

777, 792, 1016, 1018, 1256, 1788, 2521

A discussion of IGMP can be found in the following RFCs:

966, 988, 1054, 1112, 1301, 1458, 1469, 1768, 2236, 2357, 2365, 2502, 2588

A discussion of BOOTP and DHCP can be found in the following RFCs:

951, 1048, 1084, 1395, 1497, 1531, 1532, 1533, 1534, 1541, 1542, 2131, 2132

21.6 KEY TERMS

address-mask-reply message
address-mask-request message
address resolution protocol (ARP)
Bootstrap Protocol (BOOTP)
delayed response strategy
destination-unreachable message
dynamic configuration protocol
Dynamic Host Configuration Protocol (DHCP)
dynamic mapping
echo-request and echo-reply messages
error-reporting message
general query message
groupid
group membership
Internet Control Message Protocol (ICMP)
Internet Group Management Protocol (IGMP)
Internetworking Control Message Protocol, version 6 (ICMPv6)
lease
leave report
membership report

multicast address
multicast router
multicasting
neighbor-solicitation and advertisement message
packet-too-big message
parameter-problem message
physical address
proxy ARP
query message
query router
redirection message
relay agent
reverse address resolution protocol (RARP)
router-solicitation and router-advertisement messages
source-quench message
special query message
static mapping
time-exceeded message
timestamp request and timestamp reply messages
traceroute
tunneling

21.7 SUMMARY

- Delivery of a packet to a host or router requires two levels of addresses: logical and physical. A physical address identifies a host or router at the physical level.
- Mapping of a logical address to a physical address can be static or dynamic.
- Static mapping involves a list of logical and physical address correspondences; maintenance of the list requires high overhead.
- The address resolution protocol (ARP) is a dynamic mapping method that finds a physical address, given a logical address.
- In proxy ARP, a router represents a set of hosts. When an ARP request seeks the physical address of any host in this set, the router sends its own physical address. This creates a subnetting effect.
- Reverse Address Resolution Protocol (RARP) is a form of dynamic mapping in which a given physical address is associated with a logical address.
- ICMP sends four pairs of query messages: echo-request and echo-reply, time-stamp request and reply, address-mask-request and -reply, and router solicitation and advertisement.
- The checksum for ICMP is calculated by using both the header and the data fields of the ICMP message.
- Packet InterNet Groper (*ping*) is an application program that uses the services of ICMP to test the reachability of a host.
- Multicasting is the sending of the same message to more than one receiver simultaneously.
- The Internet Group Management Protocol (IGMP) helps multicast routers create and update a list of loyal members related to a router interface.
- The three IGMP message types are the query message, the membership report, and the leave report.

- A delayed response strategy prevents unnecessary traffic on a LAN.
- The IGMP message is encapsulated in an IP datagram.
- Most LANs, including Ethernet, support physical multicast addressing.
- WANs that do not support physical multicast addressing can use a process called tunneling to send multicast packets.
- BOOTP and Dynamic Host Configuration Protocol (DHCP) are client/server applications that deliver vital network information to either diskless computers or computers at first boot.
- A BOOTP request is encapsulated in a UDP user datagram.
- BOOTP, a static configuration protocol, uses a table that maps IP addresses to physical addresses.
- A relay agent is a router that helps send local BOOTP requests to remote servers.
- DHCP is a dynamic configuration protocol with two databases: One is similar to BOOTP, and the other is a pool of IP addresses available for temporary assignment.
- The DHCP server issues a lease for an IP address to a client for a specific time.
- ICMPv6, like version 4, reports errors, handles group memberships, updates specific router and host tables, and checks the viability of a host.

21.8 PRACTICE SET

Review Questions

1. Is the size of the ARP packet fixed? Explain.
2. What is the size of an ARP packet when the protocol is IPv4 and the hardware is Ethernet?
3. What is the size of an Ethernet frame carrying an ARP packet in Question 2?
4. What is the broadcast address for Ethernet?
5. Why is there a restriction on the generation of an ICMPv4 message in response to a failed ICMPv4 error message?
6. What is the purpose of including the IPv4 header and the first 8 bytes of datagram data in the error-reporting ICMPv4 messages?
7. Give an example of a situation in which a host would never receive a redirection message.
8. What is the minimum size of an ICMPv4 packet? What is the maximum size of an ICMPv4 packet?
9. What is the minimum size of an IPv4 packet that carries an ICMPv4 packet? What is the maximum size?
10. How can we determine if an IPv4 packet is carrying an ICMPv4 packet?
11. What is the minimum size of an Ethernet frame that carries an IPv4 packet which in turn carries an ICMPv4 packet? What is the maximum size?
12. Why is there no need for the ICMPv4 message to travel outside its own network?

Exercises

13. A router with IPv4 address 125.45.23.12 and Ethernet physical address 23:45:AB:4F:67:CD has received a packet for a host destination with IP address 125.11.78.10. Show the entries in the ARP request packet sent by the router. Assume no subnetting.
14. Show the entries in the ARP packet sent in response to Exercise 13.
15. Encapsulate the result of Exercise 13 in a data link frame. Fill in all the fields.
16. Encapsulate the result of Exercise 14 in a data link frame. Fill in all the fields.
17. Host A sends a datagram to host B. Host B never receives the datagram, and host A never receives notification of failure. Give two different explanations of what might have happened.
18. Calculate the checksum for the following ICMP packet:
    Type: Echo Request Identifier: 123 Sequence number: 25 Message: Hello
19. A router receives an IPv4 packet with source IP address 130.45.3.3 and destination IP address 201.23.4.6. The router cannot find the destination IP address in its routing table. Which ICMPv4 message should be sent?
20. TCP receives a segment with destination port address 234. TCP checks and cannot find an open port for this destination. Which ICMPv4 message should be sent?
21. A multicast address for a group is 231.24.60.9. What is its 48-bit Ethernet address for a LAN using TCP/IP?
22. If a router has 20 entries in its group table, should it send 20 different queries periodically or just one? Explain your answer.
23. If a host wants to continue membership in five groups, should it send five different membership report messages or just one?
24. A router on an Ethernet network has received a multicast IP packet with groupid 226.17.18.4. When the host checks its multicast group table, it finds this address. Show how the router sends this packet to the recipients by encapsulating the IP packet in an Ethernet frame. Show all the entries of the Ethernet frame. The outgoing IP address of the router is 185.23.5.6, and its outgoing physical address is 4A224512E1E2. Does the router need the services of ARP?
25. A host with IPv4 address 114.45.7.9 receives an IGMP query. When it checks its group table, it finds no entries. What action should the host take? Should it send any messages?
26. A host with IPv4 address 222.5.7.19 receives an IGMP query. When it checks its routing table, it finds two entries in its table: 227.4.3.7 and 229.45.6.23. What action should the host take? Should it send any messages? If so, what type and how many?
27. How many multicast addresses can be supported for the IPv4 protocol in Ethernet? How many multicast addresses can be supported by the IPv4 protocol? What is the size of address space lost when we transform a multicast IPv4 address to an Ethernet multicast address?
28. Change the following IPv4 multicast addresses to Ethernet multicast addresses. How many of them specify the same Ethernet address?

a. 224.18.72.8
b. 235.18.72.8
c. 237.18.6.88
d. 224.88.12.8

Research Activities

29. Use the *ping* program to test your own computer (loopback).
30. Use the *ping* program to test a host inside the United States.
31. Use the *ping* program to test a host outside the United States.
32. Use *traceroute* (or *tracert*) to find the route from your computer to a computer in a college or university.
33. Use *netstat* to find out if your server supports multicast addressing.
34. DHCP uses several messages such as DHCPREQUEST, DHCPDECLINE, DHCPACK, DHCPNACK, and DHCPRELEASE. Find the purpose of these messages.

CHAPTER 22

Network Layer: Delivery, Forwarding, and Routing

This chapter describes the delivery, forwarding, and routing of IP packets to their final destinations. **Delivery** refers to the way a packet is handled by the underlying networks under the control of the network layer. **Forwarding** refers to the way a packet is delivered to the next station. **Routing** refers to the way routing tables are created to help in forwarding.

Routing protocols are used to continuously update the routing tables that are consulted for forwarding and routing. In this chapter, we also briefly discuss common unicast and **multicast routing** protocols.

22.1 DELIVERY

The network layer supervises the handling of the packets by the underlying physical networks.We define this handling as the delivery of a packet.

Direct Versus Indirect Delivery

The delivery of a packet to its final destination is accomplished by using two different methods of delivery, direct and indirect, as shown in Figure 22.1.

Direct Delivery

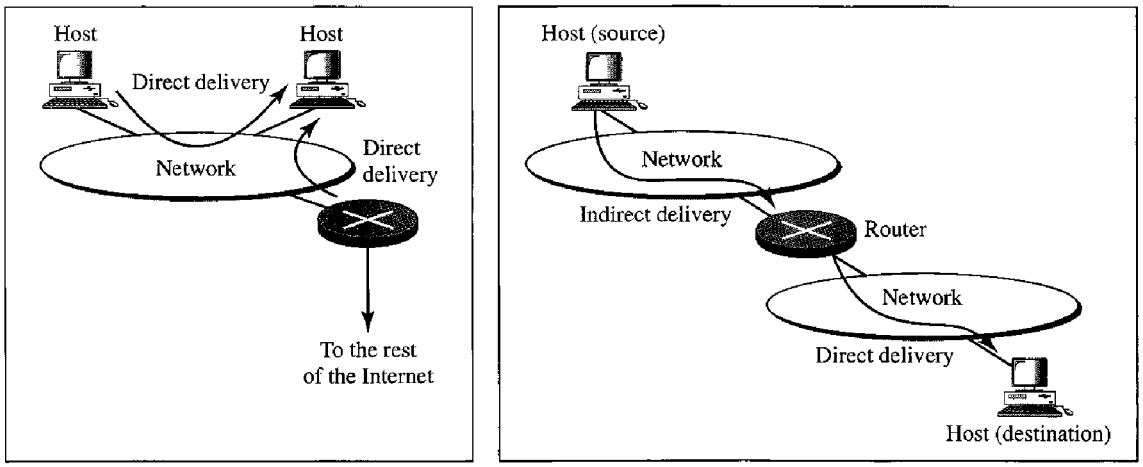
In a **direct delivery,** the final destination of the packet is a host connected to the same physical network as the deliverer. Direct delivery occurs when the source and destination of the packet are located on the same physical network or when the delivery is between the last router and the destination host.

The sender can easily determine if the delivery is direct. It can extract the network address of the destination (using the mask) and compare this address with the addresses of the networks to which it is connected. If a match is found, the delivery is direct.

Indirect Delivery

If the destination host is not on the same network as the deliverer, the packet is delivered indirectly. In an **indirect delivery,** the packet goes from router to router until it reaches the one connected to the same physical network as its final destination. Note

Figure 22.1 *Direct and indirect delivery*



a. Direct delivery **b. Indirect and direct delivery**

that a delivery always involves one direct delivery but zero or more indirect deliveries. Note also that the last delivery is always a direct delivery.

22.2 FORWARDING

Forwarding means to place the packet in its route to its destination. Forwarding requires a host or a router to have a routing table. When a host has a packet to send or when a router has received a packet to be forwarded, it looks at this table to find the route to the final destination. However, this simple solution is impossible today in an internetwork such as the Internet because the number of entries needed in the routing table would make table lookups inefficient.

Forwarding Techniques

Several techniques can make the size of the routing table manageable and also handle issues such as security. We briefly discuss these methods here.

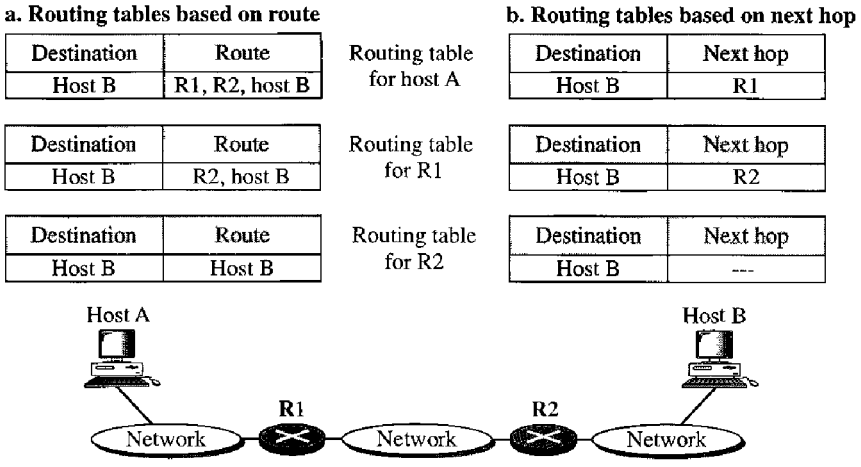
Next-Hop Method Versus Route Method

One technique to reduce the contents of a routing table is called the **next-hop method.** In this technique, the routing table holds only the address of the next hop instead of information about the complete route **(route method).** The entries of a routing table must be consistent with one another. Figure 22.2 shows how routing tables can be simplified by using this technique.

Network-Specific Method Versus Host-Specific Method

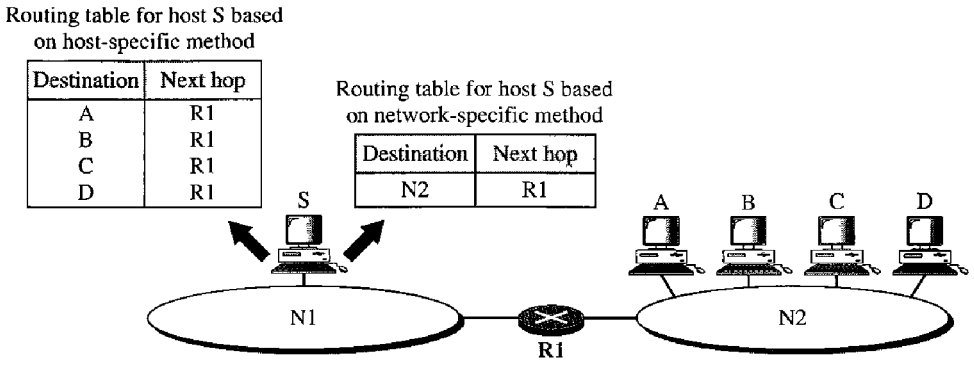
A second technique to reduce the routing table and simplify the searching process is called the **network-specific method.** Here, instead of having an entry for every destination host connected to the same physical network **(host-specific method),** we have

Figure 22.2 *Route method versus next-hop method*



only one entry that defines the address of the destination network itself. In other words, we treat all hosts connected to the same network as one single entity. For example, if 1000 hosts are attached to the same network, only one entry exists in the routing table instead of 1000. Figure 22.3 shows the concept.

Figure 22.3 *Host-specific versus network-specific method*

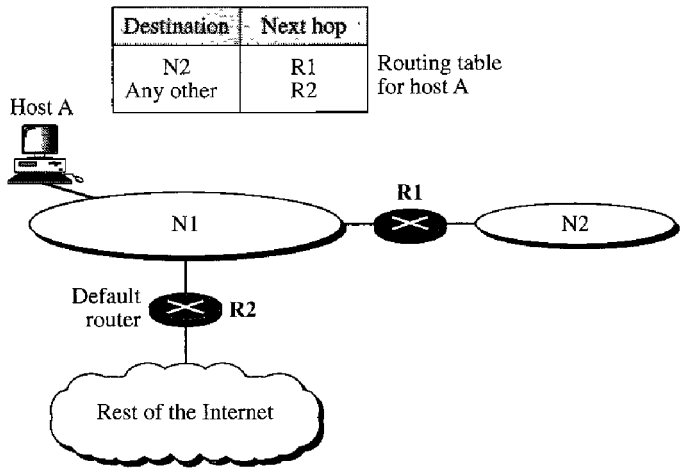


Host-specific routing is used for purposes such as checking the route or providing security measures.

Default Method

Another technique to simplify routing is called the **default method.** In Figure 22.4 host A is connected to a network with two routers. Router R1 routes the packets to hosts connected to network N2. However, for the rest of the Internet, router R2 is used. So instead of listing all networks in the entire Internet, host A can just have one entry called the *default* (normally defined as network address 0.0.0.0).

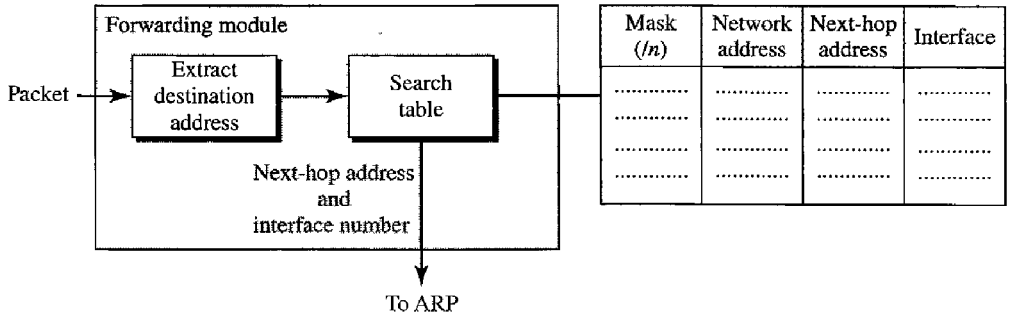
Figure 22.4 *Default method*



Forwarding Process

Let us discuss the forwarding process. We assume that hosts and routers use classless addressing because classful addressing can be treated as a special case of classless addressing. In classless addressing, the routing table needs to have one row of information for each block involved. The table needs to be searched based on the network address (first address in the block). Unfortunately, the destination address in the packet gives no clue about the network address. To solve the problem, we need to include the mask (*/n*) in the table; we need to have an extra column that includes the mask for the corresponding block. Figure 22.5 shows a simple forwarding module for classless addressing.

Figure 22.5 *Simplified forwarding module in classless address*



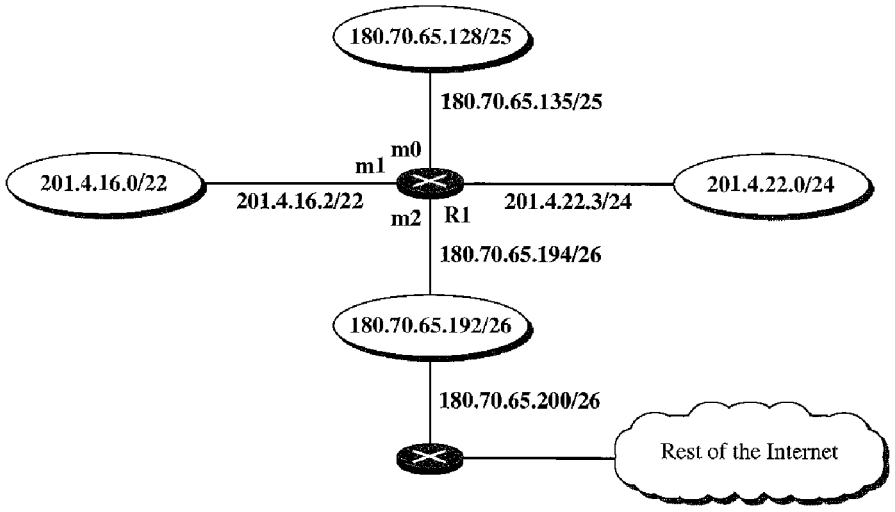
Note that we need at least four columns in our routing table; usually there are more.

In classless addressing, we need at least four columns in a routing table.

Example 22.1

Make a routing table for router R1, using the configuration in Figure 22.6.

Figure 22.6 *Configuration for Example 22.1*



Solution

Table 22.1 shows the corresponding table.

Table 22.1 *Routing table for router R1 in Figure 22.6*

Mask	*Network Address*	*Next Hop*	*Interface*
/26	180.70.65.192	—	m2
/25	180.70.65.128	—	m0
/24	201.4.22.0	—	m3
/22	201.4.16.0	....	m1
Any	Any	180.70.65.200	m2

Example 22.2

Show the forwarding process if a packet arrives at R1 in Figure 22.6 with the destination address 180.70.65.140.

Solution

The router performs the following steps:

1. The first mask (/26) is applied to the destination address. The result is 180.70.65.128, which does not match the corresponding network address.
2. The second mask (/25) is applied to the destination address. The result is 180.70.65.128, which matches the corresponding network address. The **next-hop address** (the destination address of the packet in this case) and the interface number m0 are passed to ARP for further processing.

Example 22.3

Show the forwarding process if a packet arrives at R1 in Figure 22.6 with the destination address 201.4.22.35.

Solution

The router performs the following steps:

1. The first mask (/26) is applied to the destination address. The result is 201.4.22.0, which does not match the corresponding network address (row 1).
2. The second mask (/25) is applied to the destination address. The result is 201.4.22.0, which does not match the corresponding network address (row 2).
3. The third mask (/24) is applied to the destination address. The result is 201.4.22.0, which matches the corresponding network address. The destination address of the packet and the interface number m3 are passed to ARP.

Example 22.4

Show the forwarding process if a packet arrives at R1 in Figure 22.6 with the destination address 18.24.32.78.

Solution

This time all masks are applied, one by one, to the destination address, but no matching network address is found. When it reaches the end of the table, the module gives the next-hop address 180.70.65.200 and interface number m2 to ARP. This is probably an outgoing package that needs to be sent, via the default router, to someplace else in the Internet.

Address Aggregation

When we use classless addressing, it is likely that the number of routing table entries will increase. This is so because the intent of classless addressing is to divide up the whole address space into manageable blocks. The increased size of the table results in an increase in the amount of time needed to search the table. To alleviate the problem, the idea of **address aggregation** was designed. In Figure 22.7 we have two routers.

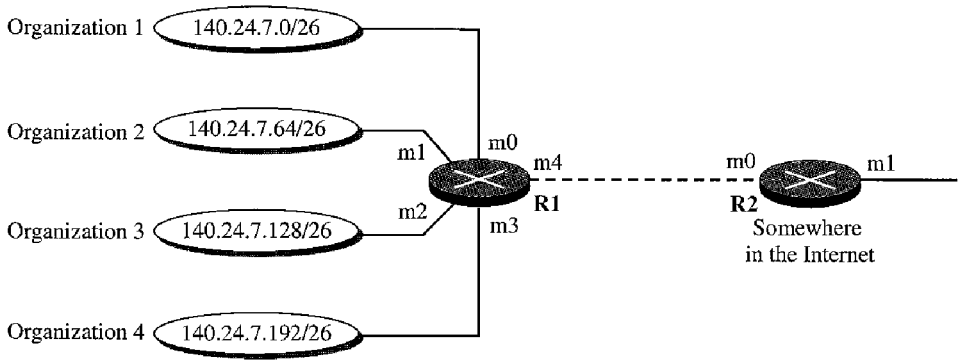
Router R1 is connected to networks of four organizations that each use 64 addresses. Router R2 is somewhere far from R1. Router R1 has a longer routing table because each packet must be correctly routed to the appropriate organization. Router R2, on the other hand, can have a very small routing table. For R2, any packet with destination 140.24.7.0 to 140.24.7.255 is sent out from interface m0 regardless of the organization number. This is called address aggregation because the blocks of addresses for four organizations are aggregated into one larger block. Router R2 would have a longer routing table if each organization had addresses that could not be aggregated into one block.

Note that although the idea of address aggregation is similar to the idea of subnetting, we do not have a common site here; the network for each organization is independent. In addition, we can have several levels of aggregation.

Longest Mask Matching

What happens if one of the organizations in Figure 22.7 is not geographically close to the other three? For example, if organization 4 cannot be connected to router R1 for some reason, can we still use the idea of address aggregation and still assign block 140.24.7.192/26 to organization 4?

Figure 22.7 *Address aggregation*



Mask	Network address	Next-hop address	Interface
/26	140.24.7.0	----------	m0
/26	140.24.7.64	----------	m1
/26	140.24.7.128	----------	m2
/26	140.24.7.192	----------	m3
/0	0.0.0.0	Default	m4

Routing table for R1

Mask	Network address	Next-hop address	Interface
/24	140.24.7.0	----------	m0
/0	0.0.0.0	Default	m1

Routing table for R2

The answer is yes because routing in classless addressing uses another principle, **longest mask matching.** This principle states that the routing table is sorted from the longest mask to the shortest mask. In other words, if there are three masks /27, /26, and /24, the mask /27 must be the first entry and /24 must be last. Let us see if this principle solves the situation in which organization 4 is separated from the other three organizations. Figure 22.8 shows the situation.

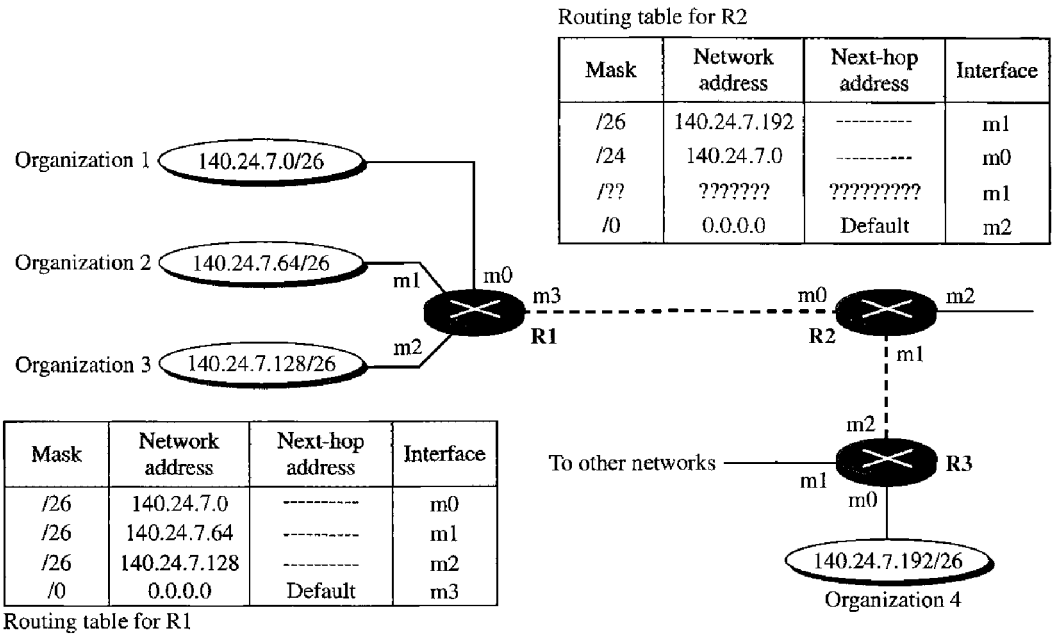
Suppose a packet arrives for organization 4 with destination address 140.24.7.200. The first mask at router R2 is applied, which gives the network address 140.24.7.192. The packet is routed correctly from interface m1 and reaches organization 4. If, however, the routing table was not stored with the longest prefix first, applying the /24 mask would result in the incorrect routing of the packet to router R1.

Hierarchical Routing

To solve the problem of gigantic routing tables, we can create a sense of hierarchy in the routing tables. In Chapter 1, we mentioned that the Internet today has a sense of hierarchy. We said that the Internet is divided into international and national ISPs. National ISPs are divided into regional ISPs, and regional ISPs are divided into local ISPs. If the routing table has a sense of hierarchy like the Internet architecture, the routing table can decrease in size.

Let us take the case of a local ISP. A local ISP can be assigned a single, but large block of addresses with a certain prefix length. The local ISP can divide this block into smaller blocks of different sizes and can assign these to individual users and organizations, both large and small. If the block assigned to the local ISP starts with a.b.c.d/*n*, the ISP can create blocks starting with e.f.g.h/*m*, where *m* may vary for each customer and is greater than *n*.

Figure 22.8 *Longest mask matching*



Routing table for R2

Mask	Network address	Next-hop address	Interface
/26	140.24.7.192	----------	m1
/24	140.24.7.0	----------	m0
/??	???????	?????????	m1
/0	0.0.0.0	Default	m2

Mask	Network address	Next-hop address	Interface
/26	140.24.7.0	----------	m0
/26	140.24.7.64	----------	m1
/26	140.24.7.128	----------	m2
/0	0.0.0.0	Default	m3

Routing table for R1

Mask	Network address	Next-hop address	Interface
/26	140.24.7.192	----------	m0
/??	???????	?????????	m1
/0	0.0.0.0	Default	m2

Routing table for R3

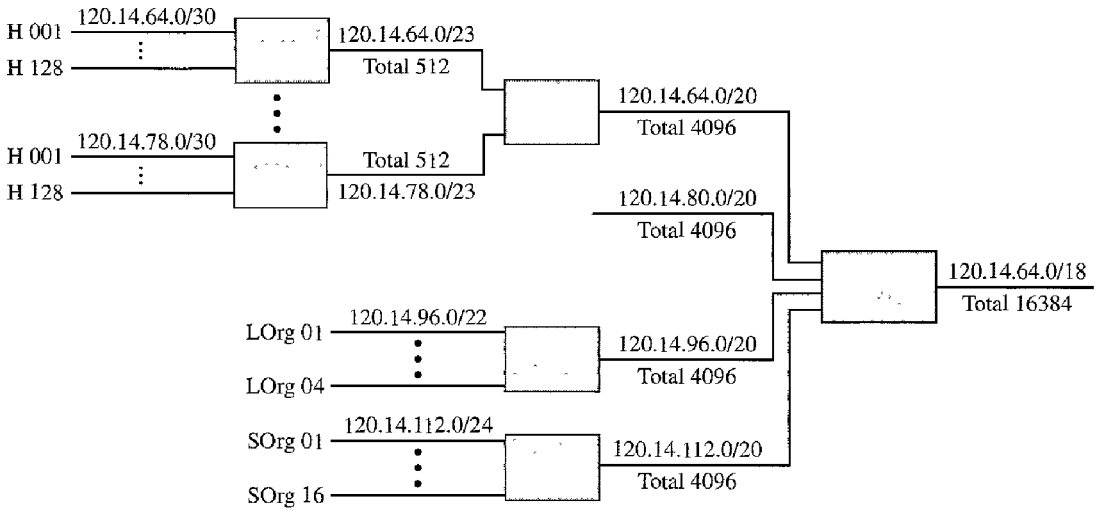
How does this reduce the size of the routing table? The rest of the Internet does not have to be aware of this division. All customers of the local ISP are defined as a.b.c.d/*n* to the rest of the Internet. Every packet destined for one of the addresses in this large block is routed to the local ISP. There is only one entry in every router in the world for all these customers. They all belong to the same group. Of course, inside the local ISP, the router must recognize the subblocks and route the packet to the destined customer. If one of the customers is a large organization, it also can create another level of hierarchy by subnetting and dividing its subblock into smaller subblocks (or sub-subblocks). In classless routing, the levels of hierarchy are unlimited so long as we follow the rules of classless addressing.

Example 22.5

As an example of **hierarchical routing,** let us consider Figure 22.9. A regional ISP is granted 16,384 addresses starting from 120.14.64.0. The regional ISP has decided to divide this block into four subblocks, each with 4096 addresses. Three of these subblocks are assigned to three local ISPs; the second subblock is reserved for future use. Note that the mask for each block is /20 because the original block with mask /18 is divided into 4 blocks.

The first local ISP has divided its assigned subblock into 8 smaller blocks and assigned each to a small ISP. Each small ISP provides services to 128 households (H001 to H128), each using four addresses. Note that the mask for each small ISP is now /23 because the block is further divided into 8 blocks. Each household has a mask of /30, because a household has only four addresses (2^{32-30} is 4).

Figure 22.9 *Hierarchical routing with ISPs*



The second local ISP has divided its block into 4 blocks and has assigned the addresses to four large organizations (LOrg01 to LOrg04). Note that each large organization has 1024 addresses, and the mask is /22.

The third local ISP has divided its block into 16 blocks and assigned each block to a small organization (SOrg01 to SOrg16). Each small organization has 256 addresses, and the mask is /24.

There is a sense of hierarchy in this configuration. All routers in the Internet send a packet with destination address 120.14.64.0 to 120.14.127.255 to the regional ISP.

The regional ISP sends every packet with destination address 120.14.64.0 to 120.14.79.255 to local ISP1. Local ISP1 sends every packet with destination address 120.14.64.0 to 120.14.64.3 to H001.

Geographical Routing

To decrease the size of the routing table even further, we need to extend hierarchical routing to include geographical routing. We must divide the entire address space into a few large blocks. We assign a block to North America, a block to Europe, a block to Asia, a block to Africa, and so on. The routers of ISPs outside Europe will have only one entry for packets to Europe in their routing tables. The routers of ISPs outside North America will have only one entry for packets to North America in their routing tables. And so on.

Routing Table

Let us now discuss routing tables. A host or a router has a routing table with an entry for each destination, or a combination of destinations, to route IP packets. The routing table can be either static or dynamic.

Static Routing Table

A **static routing table** contains information entered manually. The administrator enters the route for each destination into the table. When a table is created, it cannot update

automatically when there is a change in the Internet. The table must be manually altered by the administrator.

A static routing table can be used in a small internet that does not change very often, or in an experimental internet for troubleshooting. It is poor strategy to use a static routing table in a big internet such as the Internet.

Dynamic Routing Table

A **dynamic routing table** is updated periodically by using one of the dynamic routing protocols such as RIP, OSPF, or BGP. Whenever there is a change in the Internet, such as a shutdown of a router or breaking of a link, the dynamic routing protocols update all the tables in the routers (and eventually in the host) automatically.

The routers in a big internet such as the Internet need to be updated dynamically for efficient delivery of the IP packets. We discuss in detail the three dynamic routing protocols later in the chapter.

Format

As mentioned previously, a routing table for classless addressing has a minimum of four columns. However, some of today's routers have even more columns. We should be aware that the number of columns is vendor-dependent, and not all columns can be found in all routers. Figure 22.10 shows some common fields in today's routers.

Figure 22.10 *Common fields in a routing table*

Mask	Network address	Next-hop address	Interface	Flags	Reference count	Use
.....	.....	.....	.....	.....	.....	.....

- **Mask.** This field defines the mask applied for the entry.
- **Network address.** This field defines the network address to which the packet is finally delivered. In the case of host-specific routing, this field defines the address of the destination host.
- **Next-hop address.** This field defines the address of the next-hop router to which the packet is delivered.
- **Interface.** This field shows the name of the interface.
- **Flags.** This field defines up to five flags. Flags are on/off switches that signify either presence or absence. The five flags are U (up), G (gateway), H (host-specific), D (added by redirection), and M (modified by redirection).
  a. **U (up).** The U flag indicates the router is up and running. If this flag is not present, it means that the router is down. The packet cannot be forwarded and is discarded.
  b. **G (gateway).** The G flag means that the destination is in another network. The packet is delivered to the next-hop router for delivery (indirect delivery). When this flag is missing, it means the destination is in this network (direct delivery).

c. **H (host-specific).** The H flag indicates that the entry in the network address field is a host-specific address. When it is missing, it means that the address is only the network address of the destination.

d. **D (added by redirection).** The D flag indicates that routing information for this destination has been added to the host routing table by a redirection message from ICMP. We discussed redirection and the ICMP protocol in Chapter 21.

e. **M (modified by redirection).** The M flag indicates that the routing information for this destination has been modified by a redirection message from ICMP. We discussed redirection and the ICMP protocol in Chapter 21.

- **Reference count.** This field gives the number of users of this route at the moment. For example, if five people at the same time are connecting to the same host from this router, the value of this column is 5.
- **Use.** This field shows the number of packets transmitted through this router for the corresponding destination.

Utilities

There are several utilities that can be used to find the routing information and the contents of a routing table. We discuss *netstat* and *ifconfig*.

Example 22.6

One utility that can be used to find the contents of a routing table for a host or router is ***netstat*** in UNIX or LINUX. The following shows the list of the contents of a default server. We have used two options, r and n. The option r indicates that we are interested in the routing table, and the option n indicates that we are looking for numeric addresses. Note that this is a routing table for a host, not a router. Although we discussed the routing table for a router throughout the chapter, a host also needs a routing table.

```
$ netstat -rn
Kernel IP routing table
```

Destination	Gateway	Mask	Flags	Iface
153.18.16.0	0.0.0.0	255.255.240.0	U	eth0
127.0.0.0	0.0.0.0	255.0.0.0	U	lo
0.0.0.0	153.18.31.254	0.0.0.0	UG	eth0

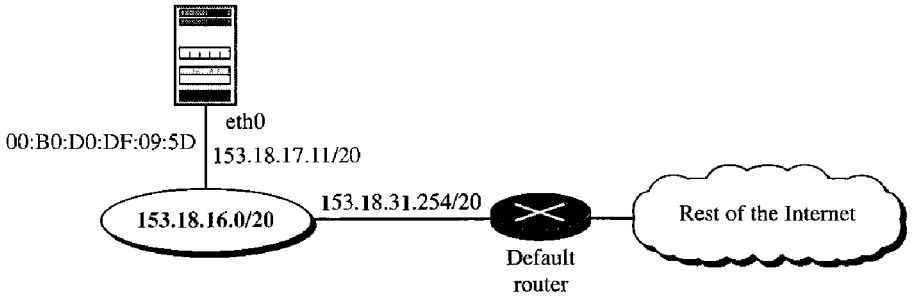
Note also that the order of columns is different from what we showed. The destination column here defines the network address. The term ***gateway*** used by UNIX is synonymous with *router*. This column actually defines the address of the next hop. The value 0.0.0.0 shows that the delivery is direct. The last entry has a flag of G, which means that the destination can be reached through a router (default router). The *Iface* defines the interface. The host has only one real interface, eth0, which means interface 0 connected to an Ethernet network. The second interface, lo, is actually a virtual loopback interface indicating that the host accepts packets with loopback address 127.0.0.0.

More information about the IP address and physical address of the server can be found by using the *ifconfig* command on the given interface (eth0).

```
$ ifconfig eth0
eth0  Link encap:Ethernet  HWaddr 00:B0:D0:DF:09:5D
inet addr:153.18.17.11  Bcast:153.18.31.255  Mask:255.255.240.0
...
```

From the above information, we can deduce the configuration of the server, as shown in Figure 22.11.

Figure 22.11 *Configuration of the server for Example 22.6*



Note that the *ifconfig* command gives us the IP address and the physical (hardware) address of the interface.

22.3 UNICAST ROUTING PROTOCOLS

A routing table can be either static or dynamic. A *static table* is one with manual entries. A *dynamic table,* on the other hand, is one that is updated automatically when there is a change somewhere in the internet. Today, an internet needs dynamic routing tables. The tables need to be updated as soon as there is a change in the internet. For instance, they need to be updated when a router is down, and they need to be updated whenever a better route has been found.

Routing protocols have been created in response to the demand for dynamic routing tables. A routing protocol is a combination of rules and procedures that lets routers in the internet inform each other of changes. It allows routers to share whatever they know about the internet or their neighborhood. The sharing of information allows a router in San Francisco to know about the failure of a network in Texas. The routing protocols also include procedures for combining information received from other routers.

Optimization

A router receives a packet from a network and passes it to another network. A router is usually attached to several networks. When it receives a packet, to which network should it pass the packet? The decision is based on optimization: Which of the available pathways is the optimum pathway? What is the definition of the term *optimum*?

One approach is to assign a cost for passing through a network. We call this cost a **metric.** However, the metric assigned to each network depends on the type of protocol. Some simple protocols, such as the Routing Information Protocol (RIP), treat all networks as equals. The cost of passing through a network is the same; it is one hop count. So if a packet passes through 10 networks to reach the destination, the total cost is 10 hop counts.

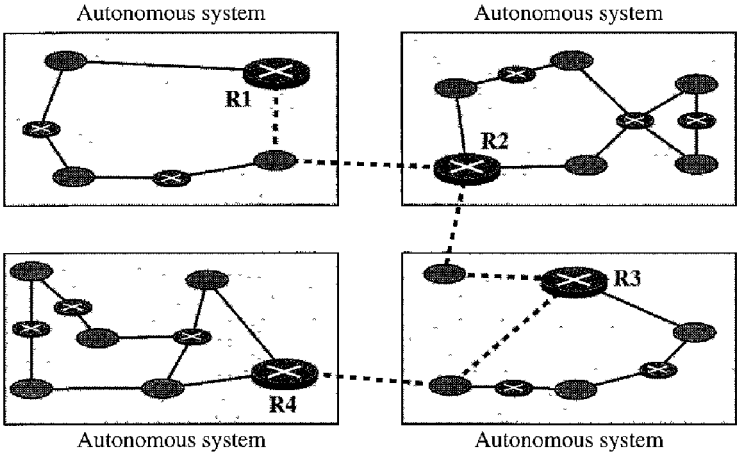
Other protocols, such as Open Shortest Path First (OSPF), allow the administrator to assign a cost for passing through a network based on the type of service required. A route through a network can have different costs (metrics). For example, if maximum throughput is the desired type of service, a satellite link has a lower metric than a fiber-optic line. On the other hand, if minimum delay is the desired type of service, a fiber-optic line has a lower metric than a satellite link. Routers use routing tables to help decide the best route. OSPF protocol allows each router to have several routing tables based on the required type of service.

Other protocols define the metric in a totally different way. In the Border Gateway Protocol (BGP), the criterion is the policy, which can be set by the administrator. The policy defines what paths should be chosen.

Intra- and Interdomain Routing

Today, an internet can be so large that one routing protocol cannot handle the task of updating the routing tables of all routers. For this reason, an internet is divided into autonomous systems. An **autonomous system (AS)** is a group of networks and routers under the authority of a single administration. Routing inside an autonomous system is referred to as **intradomain routing.** Routing between autonomous systems is referred to as **interdomain routing.** Each autonomous system can choose one or more intradomain routing protocols to handle routing inside the autonomous system. However, only one interdomain routing protocol handles routing between autonomous systems (see Figure 22.12).

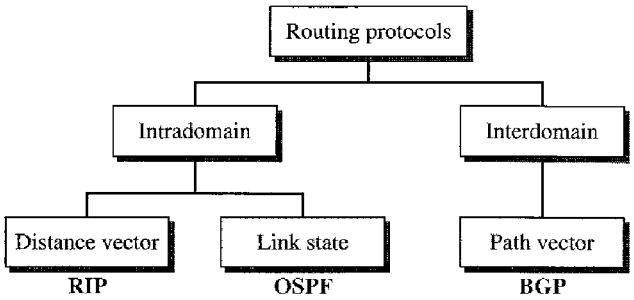
Figure 22.12 *Autonomous systems*



Several intradomain and interdomain routing protocols are in use. In this section, we cover only the most popular ones. We discuss two intradomain routing protocols: distance vector and link state. We also introduce one interdomain routing protocol: path vector (see Figure 22.13).

Routing Information Protocol (RIP) is an implementation of the distance vector protocol. **Open Shortest Path First (OSPF)** is an implementation of the link state protocol. **Border Gateway Protocol (BGP)** is an implementation of the path vector protocol.

Figure 22.13 *Popular routing protocols*



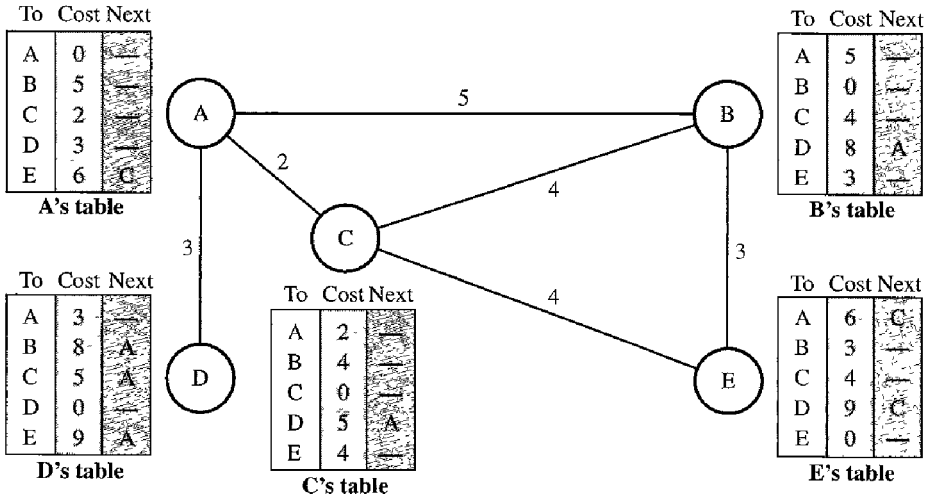
Distance Vector Routing

In **distance vector routing,** the least-cost route between any two nodes is the route with minimum distance. In this protocol, as the name implies, each node maintains a vector (table) of minimum distances to every node. The table at each node also guides the packets to the desired node by showing the next stop in the route (next-hop routing).

We can think of nodes as the cities in an area and the lines as the roads connecting them. A table can show a tourist the minimum distance between cities.

In Figure 22.14, we show a system of five nodes with their corresponding tables.

Figure 22.14 *Distance vector routing tables*



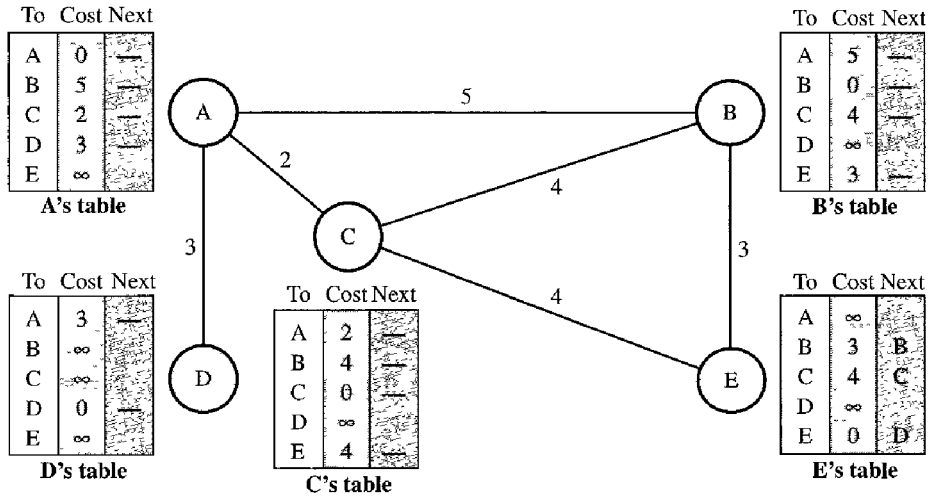
The table for node A shows how we can reach any node from this node. For example, our least cost to reach node E is 6. The route passes through C.

Initialization

The tables in Figure 22.14 are stable; each node knows how to reach any other node and the cost. At the beginning, however, this is not the case. Each node can know only

the distance between itself and its **immediate neighbors,** those directly connected to it. So for the moment, we assume that each node can send a message to the immediate neighbors and find the distance between itself and these neighbors. Figure 22.15 shows the initial tables for each node. The distance for any entry that is not a neighbor is marked as infinite (unreachable).

Figure 22.15 *Initialization of tables in distance vector routing*



Sharing

The whole idea of distance vector routing is the sharing of information between neighbors. Although node A does not know about node E, node C does. So if node C shares its routing table with A, node A can also know how to reach node E. On the other hand, node C does not know how to reach node D, but node A does. If node A shares its routing table with node C, node C also knows how to reach node D. In other words, nodes A and C, as immediate neighbors, can improve their routing tables if they help each other.

There is only one problem. How much of the table must be shared with each neighbor? A node is not aware of a neighbor's table. The best solution for each node is to send its entire table to the neighbor and let the neighbor decide what part to use and what part to discard. However, the third column of a table (next stop) is not useful for the neighbor. When the neighbor receives a table, this column needs to be replaced with the sender's name. If any of the rows can be used, the next node is the sender of the table. A node therefore can send only the first two columns of its table to any neighbor. In other words, sharing here means sharing only the first two columns.

In distance vector routing, each node shares its routing table with its immediate neighbors periodically and when there is a change.

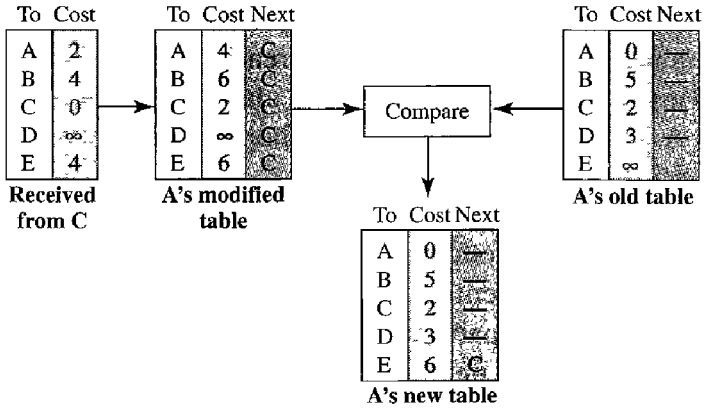
Updating

When a node receives a two-column table from a neighbor, it needs to update its routing table. Updating takes three steps:

1. The receiving node needs to add the cost between itself and the sending node to each value in the second column. The logic is clear. If node C claims that its distance to a destination is x mi, and the distance between A and C is y mi, then the distance between A and that destination, via C, is $x + y$ mi.
2. The receiving node needs to add the name of the sending node to each row as the third column if the receiving node uses information from any row. The sending node is the next node in the route.
3. The receiving node needs to compare each row of its old table with the corresponding row of the modified version of the received table.
   a. If the next-node entry is different, the receiving node chooses the row with the smaller cost. If there is a tie, the old one is kept.
   b. If the next-node entry is the same, the receiving node chooses the new row. For example, suppose node C has previously advertised a route to node X with distance 3. Suppose that now there is no path between C and X; node C now advertises this route with a distance of infinity. Node A must not ignore this value even though its old entry is smaller. The old route does not exist any more. The new route has a distance of infinity.

Figure 22.16 shows how node A updates its routing table after receiving the partial table from node C.

Figure 22.16 *Updating in distance vector routing*



There are several points we need to emphasize here. First, as we know from mathematics, when we add any number to infinity, the result is still infinity. Second, the modified table shows how to reach A from A via C. If A needs to reach itself via C, it needs to go to C and come back, a distance of 4. Third, the only benefit from this updating of node A is the last entry, how to reach E. Previously, node A did not know how to reach E (distance of infinity); now it knows that the cost is 6 via C.

Each node can update its table by using the tables received from other nodes. In a short time, if there is no change in the network itself, such as a failure in a link, each node reaches a stable condition in which the contents of its table remains the same.

When to Share

The question now is, When does a node send its partial routing table (only two columns) to all its immediate neighbors? The table is sent both periodically and when there is a change in the table.

Periodic Update A node sends its routing table, normally every 30 s, in a periodic update. The period depends on the protocol that is using distance vector routing.

Triggered Update A node sends its two-column routing table to its neighbors anytime there is a change in its routing table. This is called a **triggered update.** The change can result from the following.

1. A node receives a table from a neighbor, resulting in changes in its own table after updating.
2. A node detects some failure in the neighboring links which results in a distance change to infinity.

Two-Node Loop Instability

A problem with distance vector routing is instability, which means that a network using this protocol can become unstable. To understand the problem, let us look at the scenario depicted in Figure 22.17.

Figure 22.17 *Two-node instability*

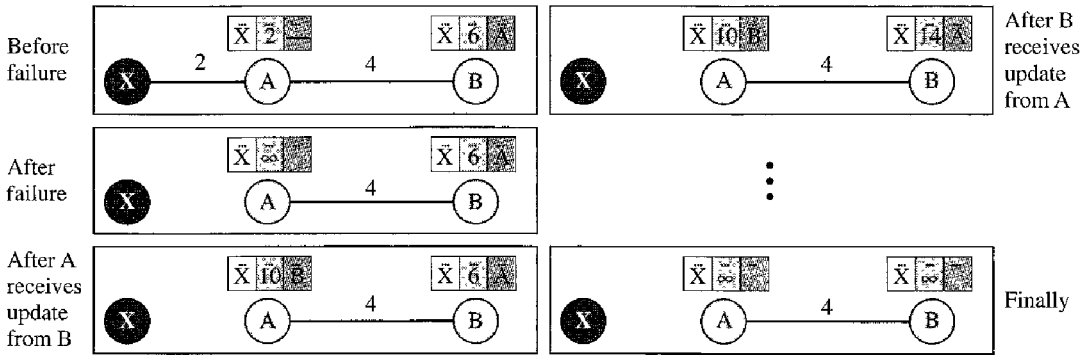


Figure 22.17 shows a system with three nodes. We have shown only the portions of the routing table needed for our discussion. At the beginning, both nodes A and B know how to reach node X. But suddenly, the link between A and X fails. Node A changes its table. If A can send its table to B immediately, everything is fine. However, the system becomes unstable if B sends its routing table to A before receiving A's routing table. Node A receives the update and, assuming that B has found a way to reach X, immediately updates its routing table. Based on the triggered update strategy, A sends its new

update to B. Now B thinks that something has been changed around A and updates its routing table. The cost of reaching X increases gradually until it reaches infinity. At this moment, both A and B know that X cannot be reached. However, during this time the system is not stable. Node A thinks that the route to X is via B; node B thinks that the route to X is via A. If A receives a packet destined for X, it goes to B and then comes back to A. Similarly, if B receives a packet destined for X, it goes to A and comes back to B. Packets bounce between A and B, creating a two-node loop problem. A few solutions have been proposed for instability of this kind.

Defining Infinity The first obvious solution is to redefine infinity to a smaller number, such as 100. For our previous scenario, the system will be stable in less than 20 updates. As a matter of fact, most implementations of the distance vector protocol define the distance between each node to be 1 and define 16 as infinity. However, this means that the distance vector routing cannot be used in large systems. The size of the network, in each direction, can not exceed 15 hops.

Split Horizon Another solution is called **split horizon.** In this strategy, instead of flooding the table through each interface, each node sends only part of its table through each interface. If, according to its table, node B thinks that the optimum route to reach X is via A, it does not need to advertise this piece of information to A; the information has come from A (A already knows). Taking information from node A, modifying it, and sending it back to node A creates the confusion. In our scenario, node B eliminates the last line of its routing table before it sends it to A. In this case, node A keeps the value of infinity as the distance to X. Later when node A sends its routing table to B, node B also corrects its routing table. The system becomes stable after the first update: both node A and B know that X is not reachable.

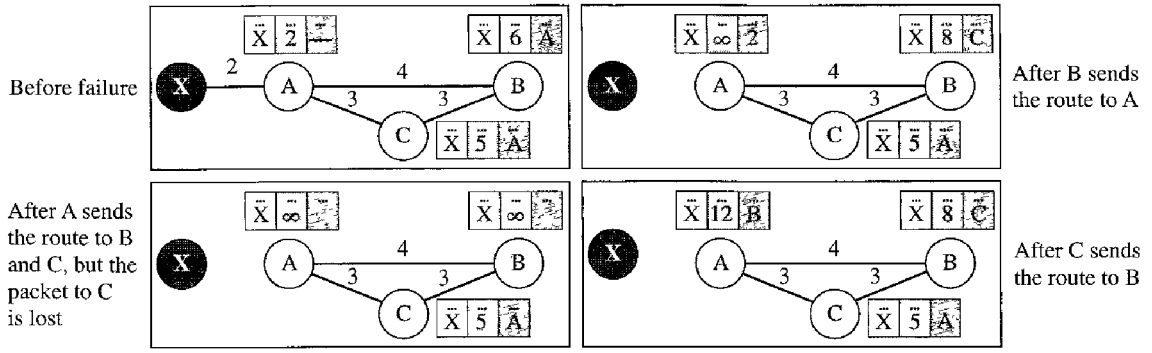
Split Horizon and Poison Reverse Using the split horizon strategy has one drawback. Normally, the distance vector protocol uses a timer, and if there is no news about a route, the node deletes the route from its table. When node B in the previous scenario eliminates the route to X from its advertisement to A, node A cannot guess that this is due to the split horizon strategy (the source of information was A) or because B has not received any news about X recently. The split horizon strategy can be combined with the **poison reverse** strategy. Node B can still advertise the value for X, but if the source of information is A, it can replace the distance with infinity as a warning: "Do not use this value; what I know about this route comes from you."

Three-Node Instability

The two-node instability can be avoided by using the split horizon strategy combined with poison reverse. However, if the instability is between three nodes, stability cannot be guaranteed. Figure 22.18 shows the scenario.

Suppose, after finding that X is not reachable, node A sends a packet to B and C to inform them of the situation. Node B immediately updates its table, but the packet to C is lost in the network and never reaches C. Node C remains in the dark and still thinks that there is a route to X via A with a distance of 5. After a while, node C sends to B its routing table, which includes the route to X. Node B is totally fooled here. It receives information on the route to X from C, and according to the algorithm, it updates its

Figure 22.18 *Three-node instability*



table, showing the route to X via C with a cost of 8. This information has come from C, not from A, so after awhile node B may advertise this route to A. Now A is fooled and updates its table to show that A can reach X via B with a cost of 12. Of course, the loop continues; now A advertises the route to X to C, with increased cost, but not to B. Node C then advertises the route to B with an increased cost. Node B does the same to A. And so on. The loop stops when the cost in each node reaches infinity.

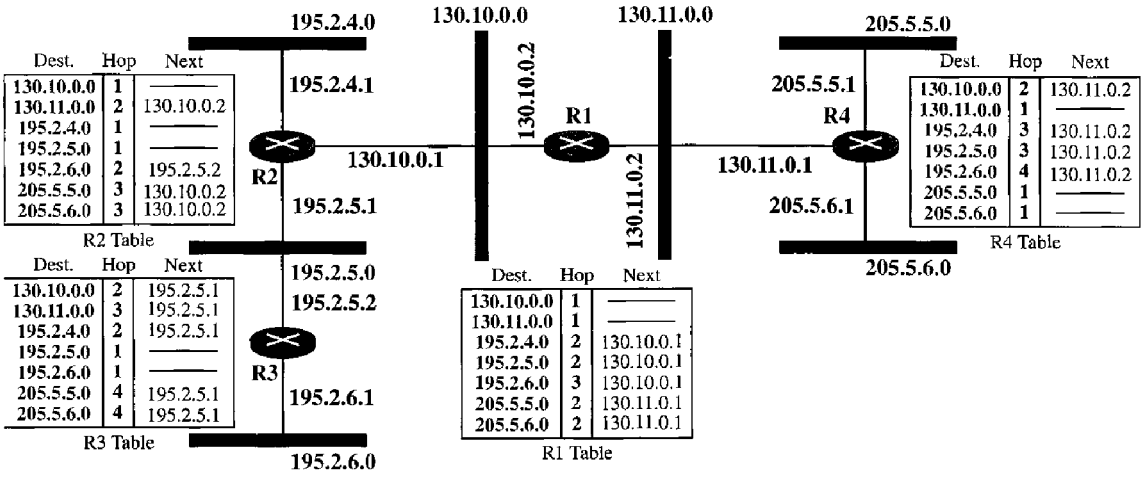
RIP

The **Routing Information Protocol (RIP)** is an intradomain routing protocol used inside an autonomous system. It is a very simple protocol based on distance vector routing. RIP implements distance vector routing directly with some considerations:

1. In an autonomous system, we are dealing with routers and networks (links). The routers have routing tables; networks do not.
2. The destination in a routing table is a network, which means the first column defines a network address.
3. The metric used by RIP is very simple; the distance is defined as the number of links (networks) to reach the destination. For this reason, the metric in RIP is called a **hop count.**
4. Infinity is defined as 16, which means that any route in an autonomous system using RIP cannot have more than 15 hops.
5. The next-node column defines the address of the router to which the packet is to be sent to reach its destination.

Figure 22.19 shows an autonomous system with seven networks and four routers. The table of each router is also shown. Let us look at the routing table for R1. The table has seven entries to show how to reach each network in the autonomous system. Router R1 is directly connected to networks 130.10.0.0 and 130.11.0.0, which means that there are no next-hop entries for these two networks. To send a packet to one of the three networks at the far left, router R1 needs to deliver the packet to R2. The next-node entry for these three networks is the interface of router R2 with IP address 130.10.0.1. To send a packet to the two networks at the far right, router R1 needs to send the packet to the interface of router R4 with IP address 130.11.0.1. The other tables can be explained similarly.

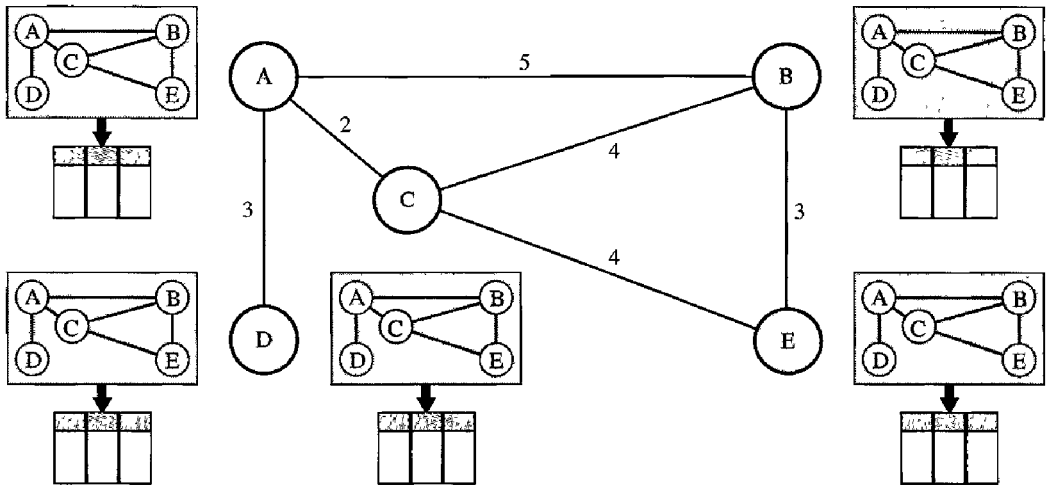
Figure 22.19 *Example of a domain using RIP*



Link State Routing

Link state routing has a different philosophy from that of distance vector routing. In link state routing, if each node in the domain has the entire topology of the domain—the list of nodes and links, how they are connected including the type, cost (metric), and condition of the links (up or down)—the node can use **Dijkstra's algorithm** to build a routing table. Figure 22.20 shows the concept.

Figure 22.20 *Concept of link state routing*

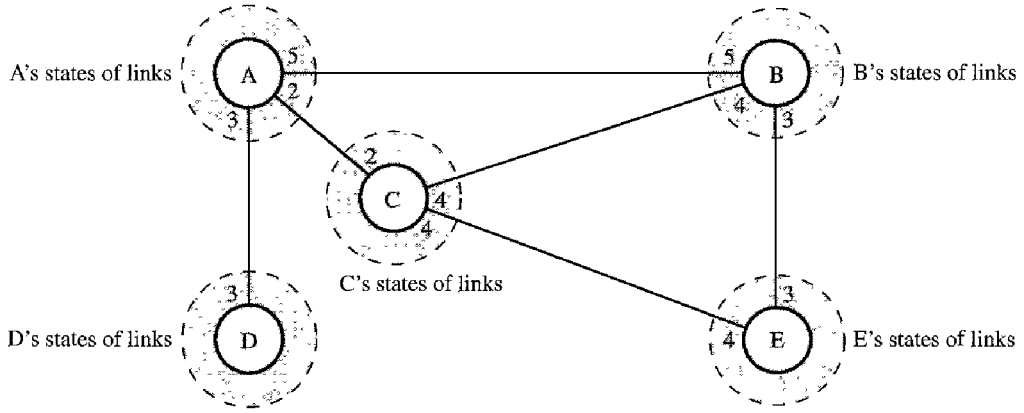


The figure shows a simple domain with five nodes. Each node uses the same topology to create a routing table, but the routing table for each node is unique because the calculations are based on different interpretations of the topology. This is analogous to a city map. While each person may have the same map, each needs to take a different route to reach her specific destination.

The topology must be dynamic, representing the latest state of each node and each link. If there are changes in any point in the network (a link is down, for example), the topology must be updated for each node.

How can a common topology be dynamic and stored in each node? No node can know the topology at the beginning or after a change somewhere in the network. Link state routing is based on the assumption that, although the global knowledge about the topology is not clear, each node has partial knowledge: it knows the state (type, condition, and cost) of its links. In other words, the whole topology can be compiled from the partial knowledge of each node. Figure 22.21 shows the same domain as in Figure 22.20, indicating the part of the knowledge belonging to each node.

Figure 22.21 *Link state knowledge*



Node A knows that it is connected to node B with metric 5, to node C with metric 2, and to node D with metric 3. Node C knows that it is connected to node A with metric 2, to node B with metric 4, and to node E with metric 4. Node D knows that it is connected only to node A with metric 3. And so on. Although there is an overlap in the knowledge, the overlap guarantees the creation of a common topology—a picture of the whole domain for each node.

Building Routing Tables

In **link state routing,** four sets of actions are required to ensure that each node has the routing table showing the least-cost node to every other node.

1. Creation of the states of the links by each node, called the link state packet (LSP).
2. Dissemination of LSPs to every other router, called **flooding,** in an efficient and reliable way.
3. Formation of a shortest path tree for each node.
4. Calculation of a routing table based on the shortest path tree.

Creation of Link State Packet (LSP) A link state packet can carry a large amount of information. For the moment, however, we assume that it carries a minimum amount

of data: the node identity, the list of links, a sequence number, and age. The first two, node identity and the list of links, are needed to make the topology. The third, sequence number, facilitates flooding and distinguishes new LSPs from old ones. The fourth, age, prevents old LSPs from remaining in the domain for a long time. LSPs are generated on two occasions:

1. *When there is a change in the topology of the domain.* Triggering of LSP dissemination is the main way of quickly informing any node in the domain to update its topology.
2. *On a periodic basis.* The period in this case is much longer compared to distance vector routing. As a matter of fact, there is no actual need for this type of LSP dissemination. It is done to ensure that old information is removed from the domain. The timer set for periodic dissemination is normally in the range of 60 min or 2 h based on the implementation. A longer period ensures that flooding does not create too much traffic on the network.

Flooding of LSPs After a node has prepared an LSP, it must be disseminated to all other nodes, not only to its neighbors. The process is called flooding and based on the following:

1. The creating node sends a copy of the LSP out of each interface.
2. A node that receives an LSP compares it with the copy it may already have. If the newly arrived LSP is older than the one it has (found by checking the sequence number), it discards the LSP. If it is newer, the node does the following:
   a. It discards the old LSP and keeps the new one.
   b. It sends a copy of it out of each interface except the one from which the packet arrived. This guarantees that flooding stops somewhere in the domain (where a node has only one interface).

Formation of Shortest Path Tree: Dijkstra Algorithm After receiving all LSPs, each node will have a copy of the whole topology. However, the topology is not sufficient to find the shortest path to every other node; a **shortest path tree** is needed.

A tree is a graph of nodes and links; one node is called the root. All other nodes can be reached from the root through only one single route. A shortest path tree is a tree in which the path between the root and every other node is the shortest. What we need for each node is a shortest path tree with that node as the root.

The **Dijkstra algorithm** creates a shortest path tree from a graph. The algorithm divides the nodes into two sets: tentative and permanent. It finds the neighbors of a current node, makes them tentative, examines them, and if they pass the criteria, makes them permanent. We can informally define the algorithm by using the flowchart in Figure 22.22.

Let us apply the algorithm to node A of our sample graph in Figure 22.23. To find the shortest path in each step, we need the cumulative cost from the root to each node, which is shown next to the node.

The following shows the steps. At the end of each step, we show the permanent (filled circles) and the tentative (open circles) nodes and lists with the cumulative costs.

Figure 22.22 *Dijkstra algorithm*

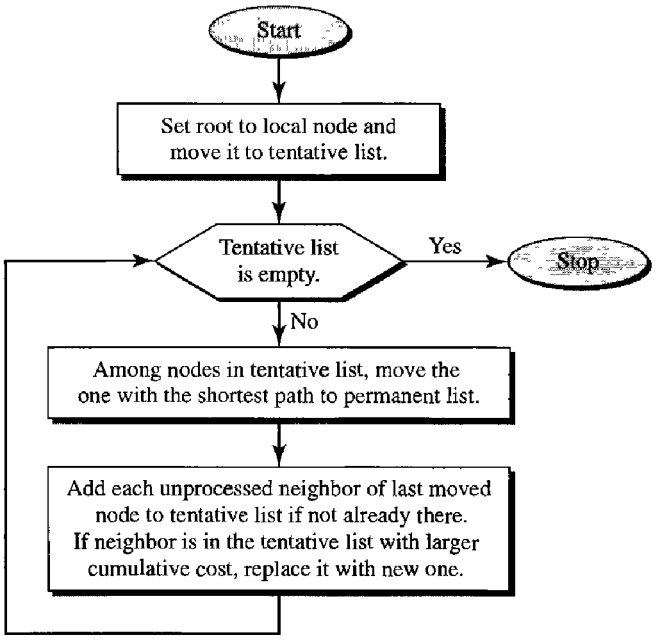
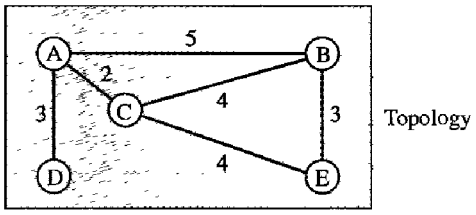
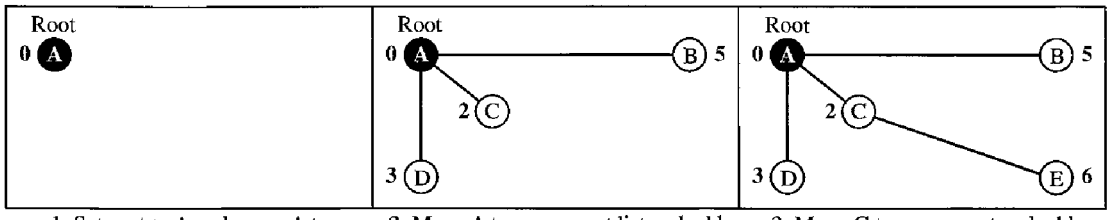


Figure 22.23 *Example of formation of shortest path tree*

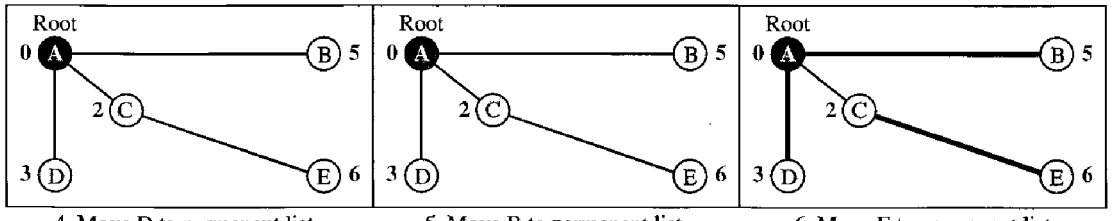




1. Set root to A and move A to tentative list.

2. Move A to permanent list and add B, C, and D to tentative list.

3. Move C to permanent and add E to tentative list.



4. Move D to permanent list.

5. Move B to permanent list.

6. Move E to permanent list (tentative list is empty).

1. We make node A the root of the tree and move it to the tentative list. Our two lists are

   Permanent list: empty Tentative list: A(0)

2. Node A has the shortest cumulative cost from all nodes in the tentative list. We move A to the permanent list and add all neighbors of A to the tentative list. Our new lists are

   Permanent list: A(0) Tentative list: B(5), C(2), D(3)

3. Node C has the shortest cumulative cost from all nodes in the tentative list. We move C to the permanent list. Node C has three neighbors, but node A is already processed, which makes the unprocessed neighbors just B and E. However, B is already in the tentative list with a cumulative cost of 5. Node A could also reach node B through C with a cumulative cost of 6. Since 5 is less than 6, we keep node B with a cumulative cost of 5 in the tentative list and do not replace it. Our new lists are

   Permanent list: A(0), C(2) Tentative list: B(5), D(3), E(6)

4. Node D has the shortest cumulative cost of all the nodes in the tentative list. We move D to the permanent list. Node D has no unprocessed neighbor to be added to the tentative list. Our new lists are

   Permanent list: A(0), C(2), D(3) Tentative list: B(5), E(6)

5. Node B has the shortest cumulative cost of all the nodes in the tentative list. We move B to the permanent list. We need to add all unprocessed neighbors of B to the tentative list (this is just node E). However, E(6) is already in the list with a smaller cumulative cost. The cumulative cost to node E, as the neighbor of B, is 8. We keep node E(6) in the tentative list. Our new lists are

   Permanent list: A(0), B(5), C(2), D(3) Tentative list: E(6)

6. Node E has the shortest cumulative cost from all nodes in the tentative list. We move E to the permanent list. Node E has no neighbor. Now the tentative list is empty. We stop; our shortest path tree is ready. The final lists are

   Permanent list: A(0), B(5), C(2), D(3), E(6) Tentative list: empty

Calculation of Routing Table from Shortest Path Tree Each node uses the shortest path tree protocol to construct its routing table. The routing table shows the cost of reaching each node from the root. Table 22.2 shows the routing table for node A.

Table 22.2 *Routing table for node A*

Node	*Cost*	*Next Router*
A	0	—
B	5	—
C	2	—
D	3	—
E	6	C

Compare Table 22.2 with the one in Figure 22.14. Both distance vector routing and link state routing end up with the same routing table for node A.

OSPF

The Open Shortest Path First or **OSPF protocol** is an intradomain routing protocol based on link state routing. Its domain is also an autonomous system.

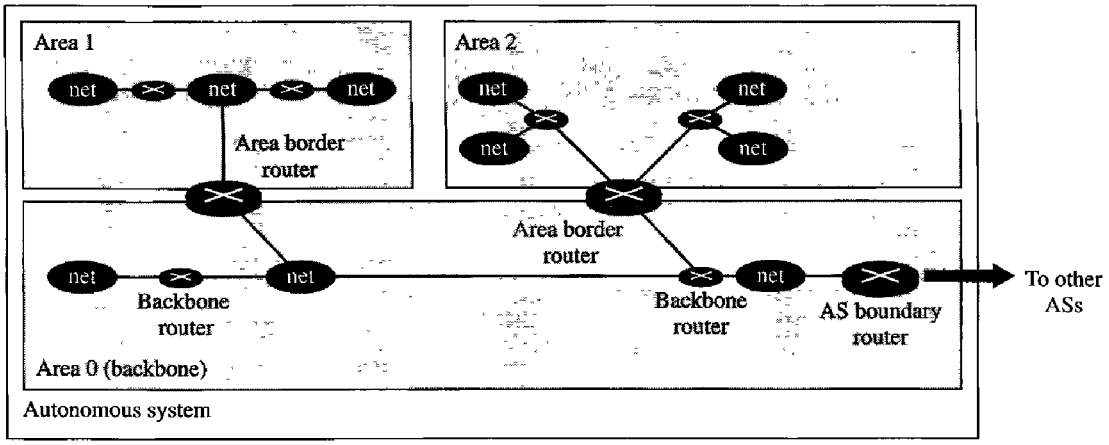
Areas To handle routing efficiently and in a timely manner, OSPF divides an autonomous system into areas. An **area** is a collection of networks, hosts, and routers all contained within an autonomous system. An autonomous system can be divided into many different areas. All networks inside an area must be connected.

Routers inside an area flood the area with routing information. At the border of an area, special routers called **area border routers** summarize the information about the area and send it to other areas. Among the areas inside an autonomous system is a special area called the *backbone;* all the areas inside an autonomous system must be connected to the backbone. In other words, the backbone serves as a primary area and the other areas as secondary areas. This does not mean that the routers within areas cannot be connected to each other, however. The routers inside the backbone are called the **backbone routers.** Note that a backbone router can also be an area border router.

If, because of some problem, the connectivity between a backbone and an area is broken, a **virtual link** between routers must be created by an administrator to allow continuity of the functions of the backbone as the primary area.

Each area has an area identification. The area identification of the backbone is zero. Figure 22.24 shows an autonomous system and its areas.

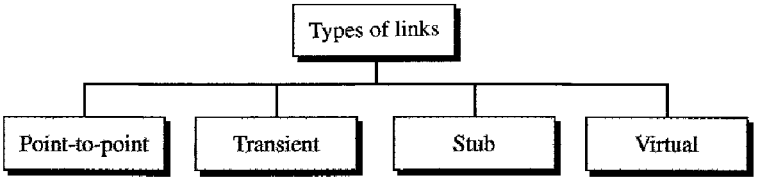
Figure 22.24 *Areas in an autonomous system*



Metric The OSPF protocol allows the administrator to assign a cost, called the **metric,** to each route. The metric can be based on a type of service (minimum delay, maximum throughput, and so on). As a matter of fact, a router can have multiple routing tables, each based on a different type of service.

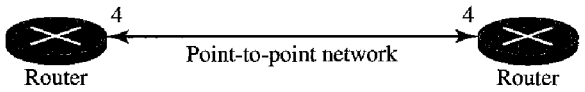
Types of Links In OSPF terminology, a connection is called a *link.* Four types of links have been defined: point-to-point, transient, stub, and virtual (see Figure 22.25).

Figure 22.25 *Types of links*



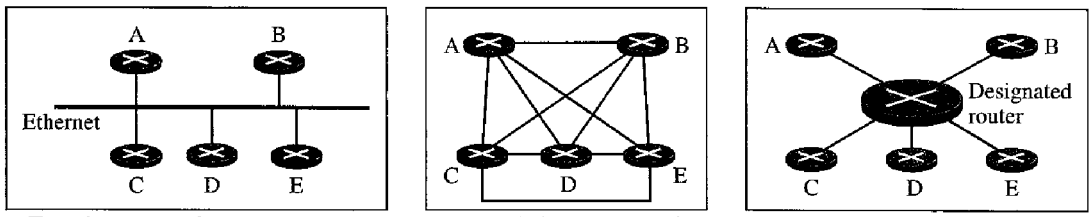
A **point-to-point link** connects two routers without any other host or router in between. In other words, the purpose of the link (network) is just to connect the two routers. An example of this type of link is two routers connected by a telephone line or a T line. There is no need to assign a network address to this type of link. Graphically, the routers are represented by nodes, and the link is represented by a bidirectional edge connecting the nodes. The metrics, which are usually the same, are shown at the two ends, one for each direction. In other words, each router has only one neighbor at the other side of the link (see Figure 22.26).

Figure 22.26 *Point-to-point link*



A **transient link** is a network with several routers attached to it. The data can enter through any of the routers and leave through any router. All LANs and some WANs with two or more routers are of this type. In this case, each router has many neighbors. For example, consider the Ethernet in Figure 22.27a. Router A has routers B, C, D, and E as neighbors. Router B has routers A, C, D, and E as neighbors. If we want to show the neighborhood relationship in this situation, we have the graph shown in Figure 22.27b.

Figure 22.27 *Transient link*



a. Transient network b. Unrealistic representation c. Realistic representation

This is neither efficient nor realistic. It is not efficient because each router needs to advertise the neighborhood to four other routers, for a total of 20 advertisements. It is

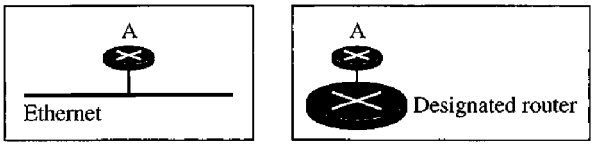
not realistic because there is no single network (link) between each pair of routers; there is only one network that serves as a crossroad between all five routers.

To show that each router is connected to every other router through one single network, the network itself is represented by a node. However, because a network is not a machine, it cannot function as a router. One of the routers in the network takes this responsibility. It is assigned a dual purpose; it is a true router and a designated router. We can use the topology shown in Figure 22.27c to show the connections of a transient network.

Now each router has only one neighbor, the designated router (network). On the other hand, the designated router (the network) has five neighbors. We see that the number of neighbor announcements is reduced from 20 to 10. Still, the link is represented as a bidirectional edge between the nodes. However, while there is a metric from each node to the designated router, there is no metric from the designated router to any other node. The reason is that the designated router represents the network. We can only assign a cost to a packet that is passing through the network. We cannot charge for this twice. When a packet enters a network, we assign a cost; when a packet leaves the network to go to the router, there is no charge.

A **stub link** is a network that is connected to only one router. The data packets enter the network through this single router and leave the network through this same router. This is a special case of the transient network. We can show this situation using the router as a node and using the designated router for the network. However, the link is only one-directional, from the router to the network (see Figure 22.28).

Figure 22.28 *Stub link*



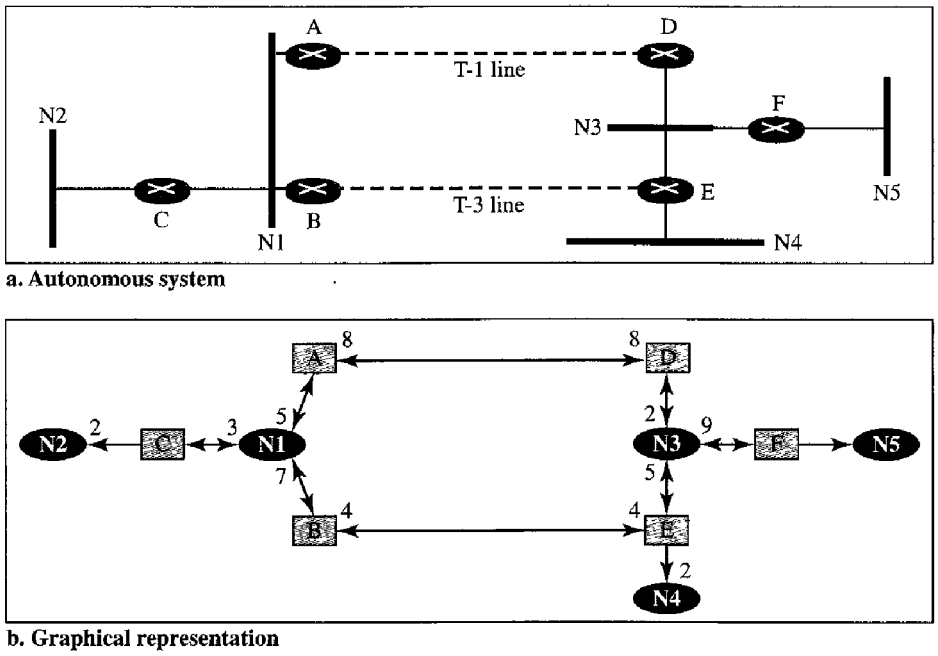
a. Stub network b. Representation

When the link between two routers is broken, the administration may create a **virtual link** between them, using a longer path that probably goes through several routers.

Graphical Representation Let us now examine how an AS can be represented graphically. Figure 22.29 shows a small AS with seven networks and six routers. Two of the networks are point-to-point networks. We use symbols such as N1 and N2 for transient and stub networks. There is no need to assign an identity to a point-to-point network. The figure also shows the graphical representation of the AS as seen by OSPF.

We have used square nodes for the routers and ovals for the networks (represented by designated routers). However, OSPF sees both as nodes. Note that we have three stub networks.

Figure 22.29 *Example of an AS and its graphical representation in OSPF*



Path Vector Routing

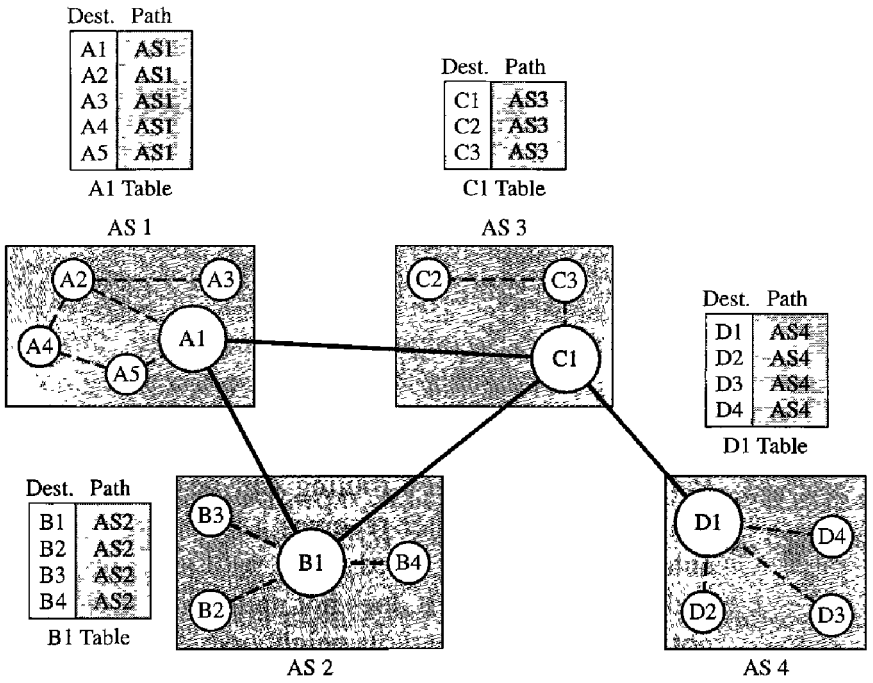
Distance vector and link state routing are both intradomain routing protocols. They can be used inside an autonomous system, but not between autonomous systems. These two protocols are not suitable for interdomain routing mostly because of scalability. Both of these routing protocols become intractable when the domain of operation becomes large. Distance vector routing is subject to instability if there are more than a few hops in the domain of operation. Link state routing needs a huge amount of resources to calculate routing tables. It also creates heavy traffic because of flooding. There is a need for a third routing protocol which we call **path vector routing.**

Path vector routing proved to be useful for interdomain routing. The principle of path vector routing is similar to that of distance vector routing. In path vector routing, we assume that there is one node (there can be more, but one is enough for our conceptual discussion) in each autonomous system that acts on behalf of the entire autonomous system. Let us call it the **speaker node.** The speaker node in an AS creates a routing table and advertises it to speaker nodes in the neighboring ASs. The idea is the same as for distance vector routing except that only speaker nodes in each AS can communicate with each other. However, what is advertised is different. A speaker node advertises the path, not the metric of the nodes, in its autonomous system or other autonomous systems.

Initialization

At the beginning, each speaker node can know only the reachability of nodes inside its autonomous system. Figure 22.30 shows the initial tables for each speaker node in a system made of four ASs.

Figure 22.30 *Initial routing tables in path vector routing*



Node A1 is the speaker node for AS1, B1 for AS2, C1 for AS3, and D1 for AS4. Node A1 creates an initial table that shows A1 to A5 are located in AS1 and can be reached through it. Node B1 advertises that B1 to B4 are located in AS2 and can be reached through B1. And so on.

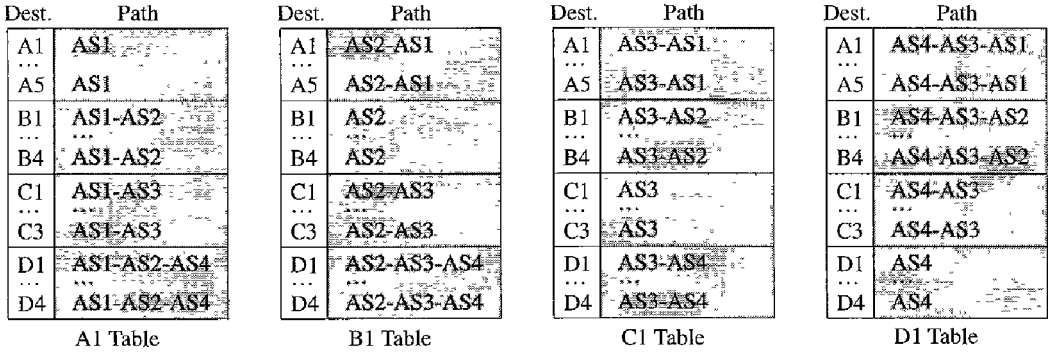
Sharing Just as in distance vector routing, in path vector routing, a speaker in an autonomous system shares its table with immediate neighbors. In Figure 22.30, node A1 shares its table with nodes B1 and C1. Node C1 shares its table with nodes D1, B1, and A1. Node B1 shares its table with C1 and A1. Node D1 shares its table with C1.

Updating When a speaker node receives a two-column table from a neighbor, it updates its own table by adding the nodes that are not in its routing table and adding its own autonomous system and the autonomous system that sent the table. After a while each speaker has a table and knows how to reach each node in other ASs. Figure 22.31 shows the tables for each speaker node after the system is stabilized.

According to the figure, if router A1 receives a packet for nodes A3, it knows that the path is in AS1 (the packet is at home); but if it receives a packet for D1, it knows that the packet should go from AS1, to AS2, and then to AS3. The routing table shows the path completely. On the other hand, if node D1 in AS4 receives a packet for node A2, it knows it should go through AS4, AS3, and AS1.

- **Loop prevention.** The instability of distance vector routing and the creation of loops can be avoided in path vector routing. When a router receives a message, it checks to see if its autonomous system is in the path list to the destination. If it is, looping is involved and the message is ignored.

Figure 22.31 *Stabilized tables for three autonomous systems*



- **Policy routing.** Policy routing can be easily implemented through path vector routing. When a router receives a message, it can check the path. If one of the autonomous systems listed in the path is against its policy, it can ignore that path and that destination. It does not update its routing table with this path, and it does not send this message to its neighbors.
- **Optimum path.** What is the optimum path in path vector routing? We are looking for a path to a destination that is the best for the organization that runs the autonomous system. We definitely cannot include metrics in this route because each autonomous system that is included in the path may use a different criterion for the metric. One system may use, internally, RIP, which defines hop count as the metric; another may use OSPF with minimum delay defined as the metric. The optimum path is the path that fits the organization. In our previous figure, each autonomous system may have more than one path to a destination. For example, a path from AS4 to AS1 can be AS4-AS3-AS2-AS1, or it can be AS4-AS3-AS1. For the tables, we chose the one that had the smaller number of autonomous systems, but this is not always the case. Other criteria, such as security, safety, and reliability, can also be applied.

BGP

Border Gateway Protocol (BGP) is an interdomain routing protocol using path vector routing. It first appeared in 1989 and has gone through four versions.

Types of Autonomous Systems As we said before, the Internet is divided into hierarchical domains called autonomous systems. For example, a large corporation that manages its own network and has full control over it is an autonomous system. A local ISP that provides services to local customers is an autonomous system. We can divide autonomous systems into three categories: stub, multihomed, and transit.

- **Stub AS.** A stub AS has only one connection to another AS. The interdomain data traffic in a stub AS can be either created or terminated in the AS. The hosts in the AS can send data traffic to other ASs. The hosts in the AS can receive data coming from hosts in other ASs. Data traffic, however, cannot pass through a stub AS. A stub AS

is either a source or a sink. A good example of a stub AS is a small corporation or a small local ISP.

- **Multihomed AS.** A multihomed AS has more than one connection to other ASs, but it is still only a source or sink for data traffic. It can receive data traffic from more than one AS. It can send data traffic to more than one AS, but there is no transient traffic. It does not allow data coming from one AS and going to another AS to pass through. A good example of a multihomed AS is a large corporation that is connected to more than one regional or national AS that does not allow transient traffic.
- **Transit AS.** A transit AS is a multihomed AS that also allows transient traffic. Good examples of transit ASs are national and international ISPs (Internet backbones).

Path Attributes In our previous example, we discussed a path for a destination network. The path was presented as a list of autonomous systems, but is, in fact, a list of attributes. Each attribute gives some information about the path. The list of attributes helps the receiving router make a more-informed decision when applying its policy.

Attributes are divided into two broad categories: well known and optional. A **well-known attribute** is one that every BGP router must recognize. An **optional attribute** is one that needs not be recognized by every router.

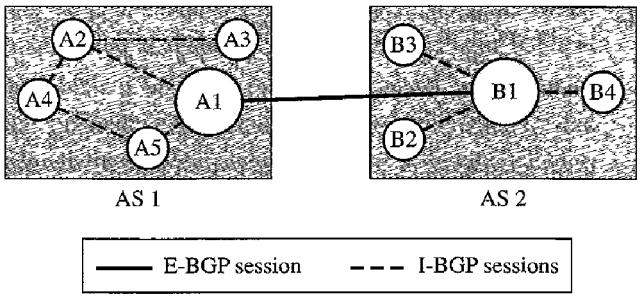
Well-known attributes are themselves divided into two categories: mandatory and discretionary. A *well-known mandatory attribute* is one that must appear in the description of a route. A *well-known discretionary attribute* is one that must be recognized by each router, but is not required to be included in every update message. One well-known mandatory attribute is ORIGIN. This defines the source of the routing information (RIP, OSPF, and so on). Another well-known mandatory attribute is AS_PATH. This defines the list of autonomous systems through which the destination can be reached. Still another well-known mandatory attribute is NEXT-HOP, which defines the next router to which the data packet should be sent.

The optional attributes can also be subdivided into two categories: transitive and nontransitive. An *optional transitive attribute* is one that must be passed to the next router by the router that has not implemented this attribute. An *optional nontransitive attribute* is one that must be discarded if the receiving router has not implemented it.

BGP Sessions The exchange of routing information between two routers using BGP takes place in a session. A session is a connection that is established between two BGP routers only for the sake of exchanging routing information. To create a reliable environment, BGP uses the services of TCP. In other words, a session at the BGP level, as an application program, is a connection at the TCP level. However, there is a subtle difference between a connection in TCP made for BGP and other application programs. When a TCP connection is created for BGP, it can last for a long time, until something unusual happens. For this reason, BGP sessions are sometimes referred to as *semi-permanent connections*.

External and Internal BGP If we want to be precise, BGP can have two types of sessions: external BGP (E-BGP) and internal BGP (I-BGP) sessions. The E-BGP session is used to exchange information between two speaker nodes belonging to two different autonomous systems. The I-BGP session, on the other hand, is used to exchange routing information between two routers inside an autonomous system. Figure 22.32 shows the idea.

Figure 22.32 *Internal and external BGP sessions*



The session established between AS1 and AS2 is an E-BGP session. The two speaker routers exchange information they know about networks in the Internet. However, these two routers need to collect information from other routers in the autonomous systems. This is done using I-BGP sessions.

22.4 MULTICAST ROUTING PROTOCOLS

In this section, we discuss multicasting and multicast routing protocols. We first define the term *multicasting* and compare it to unicasting and broadcasting. We also briefly discuss the applications of multicasting. Finally, we move on to multicast routing and the general ideas and goals related to it. We also discuss some common multicast routing protocols used in the Internet today.

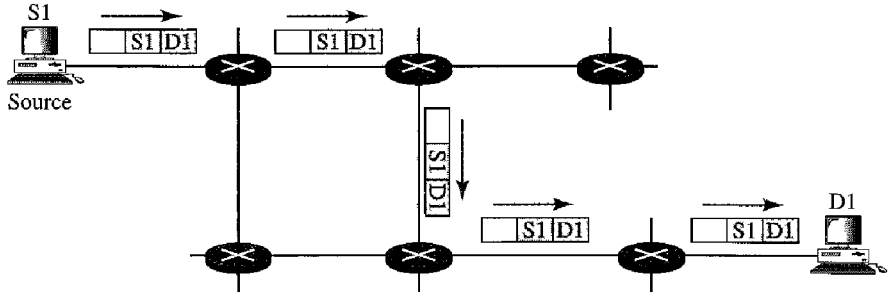
Unicast, Multicast, and Broadcast

A message can be unicast, multicast, or broadcast. Let us clarify these terms as they relate to the Internet.

Unicasting

In unicast communication, there is one source and one destination. The relationship between the source and the destination is one-to-one. In this type of communication, both the source and destination addresses, in the IP datagram, are the unicast addresses assigned to the hosts (or host interfaces, to be more exact). In Figure 22.33, a unicast

Figure 22.33 *Unicasting*



packet starts from the source S1 and passes through routers to reach the destination D1. We have shown the networks as a link between the routers to simplify the figure.

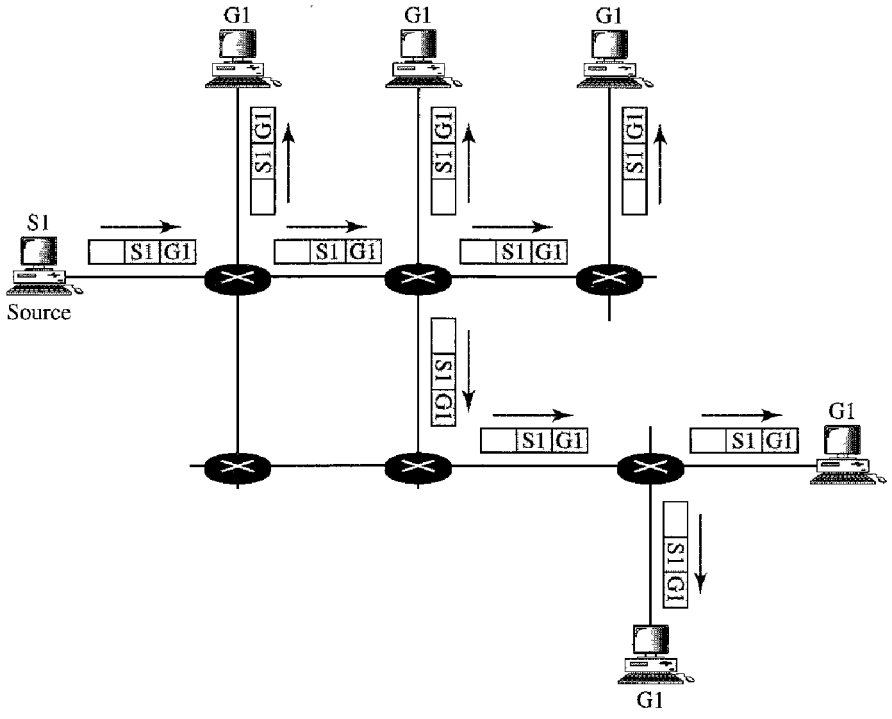
Note that in **unicasting,** when a router receives a packet, it forwards the packet through only one of its interfaces (the one belonging to the optimum path) as defined in the routing table. The router may discard the packet if it cannot find the destination address in its routing table.

In unicasting, the router forwards the received packet through only one of its interfaces.

Multicasting

In multicast communication, there is one source and a group of destinations. The relationship is one-to-many. In this type of communication, the source address is a unicast address, but the destination address is a group address, which defines one or more destinations. The group address identifies the members of the group. Figure 22.34 shows the idea behind **multicasting.**

Figure 22.34 *Multicasting*



A multicast packet starts from the source S1 and goes to all destinations that belong to group G1. In multicasting, when a router receives a packet, it may forward it through several of its interfaces.

In multicasting, the router may forward the received packet through several of its interfaces.

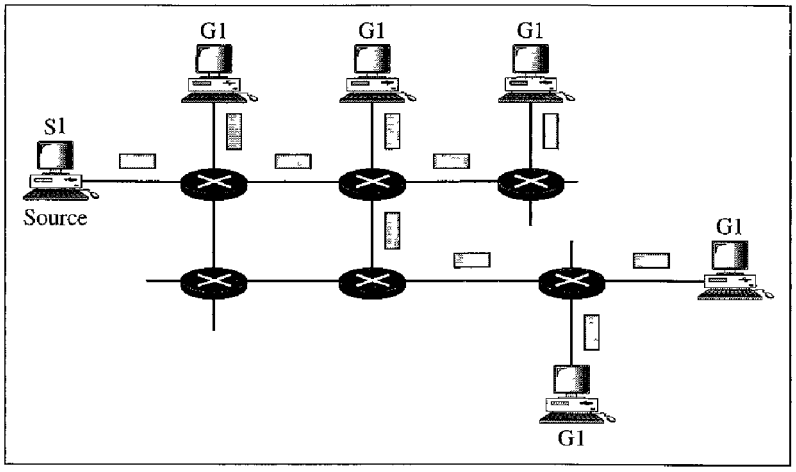
Broadcasting

In broadcast communication, the relationship between the source and the destination is one-to-all. There is only one source, but all the other hosts are the destinations. The Internet does not explicitly support **broadcasting** because of the huge amount of traffic it would create and because of the bandwidth it would need. Imagine the traffic generated in the Internet if one person wanted to send a message to everyone else connected to the Internet.

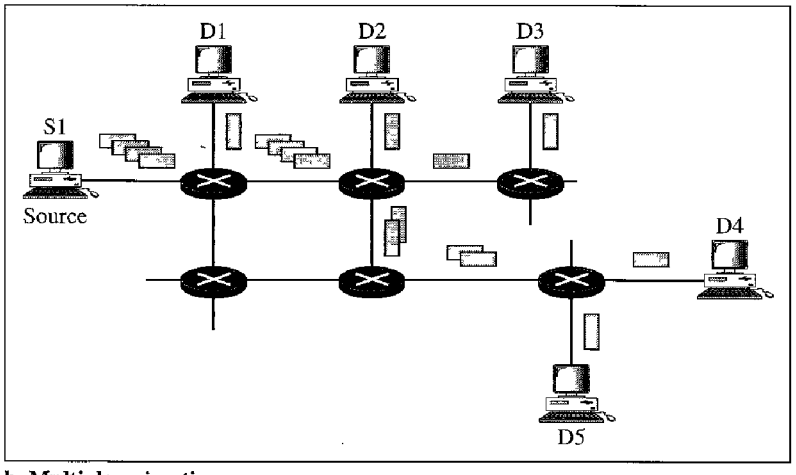
Multicasting Versus Multiple Unicasting

Before we finish this section, we need to distinguish between multicasting and multiple unicasting. Figure 22.35 illustrates both concepts.

Figure 22.35 *Multicasting versus multiple unicasting*



a. Multicasting



b. Multiple unicasting

Multicasting starts with one single packet from the source that is duplicated by the routers. The destination address in each packet is the same for all duplicates. Note that only one single copy of the packet travels between any two routers.

In **multiple unicasting,** several packets start from the source. If there are five destinations, for example, the source sends five packets, each with a different unicast destination address. Note that there may be multiple copies traveling between two routers. For example, when a person sends an e-mail message to a group of people, this is multiple unicasting. The e-mail software creates replicas of the message, each with a different destination address and sends them one by one. This is not multicasting; it is multiple unicasting.

Emulation of Multicasting with Unicasting

You might wonder why we have a separate mechanism for multicasting, when it can be emulated with unicasting. There are two obvious reasons for this.

1. Multicasting is more efficient than multiple unicasting. In Figure 22.35, we can see how multicasting requires less bandwidth than does multiple unicasting. In multiple unicasting, some of the links must handle several copies.
2. In multiple unicasting, the packets are created by the source with a relative delay between packets. If there are 1000 destinations, the delay between the first and the last packet may be unacceptable. In multicasting, there is no delay because only one packet is created by the source.

Emulation of multicasting through multiple unicasting is not efficient and may create long delays, particularly with a large group.

Applications

Multicasting has many applications today such as access to **distributed databases,** information dissemination, teleconferencing, and distance learning.

Access to Distributed Databases

Most of the large databases today are distributed. That is, the information is stored in more than one location, usually at the time of production. The user who needs to access the database does not know the location of the information. A user's request is multicast to all the database locations, and the location that has the information responds.

Information Dissemination

Businesses often need to send information to their customers. If the nature of the information is the same for each customer, it can be multicast. In this way a business can send one message that can reach many customers. For example, a software update can be sent to all purchasers of a particular software package.

Dissemination of News

In a similar manner news can be easily disseminated through multicasting. One single message can be sent to those interested in a particular topic. For example, the statistics of the championship high school basketball tournament can be sent to the sports editors of many newspapers.

Teleconferencing

Teleconferencing involves multicasting. The individuals attending a teleconference all need to receive the same information at the same time. Temporary or permanent groups can be formed for this purpose. For example, an engineering group that holds meetings every Monday morning could have a permanent group while the group that plans the holiday party could form a temporary group.

Distance Learning

One growing area in the use of multicasting is **distance learning.** Lessons taught by one single professor can be received by a specific group of students. This is especially convenient for those students who find it difficult to attend classes on campus.

Multicast Routing

In this section, we first discuss the idea of optimal routing, common in all multicast protocols. We then give an overview of multicast routing protocols.

Optimal Routing: Shortest Path Trees

The process of optimal interdomain routing eventually results in the finding of the *shortest path tree*. The root of the tree is the source, and the leaves are the potential destinations. The path from the root to each destination is the shortest path. However, the number of trees and the formation of the trees in unicast and multicast routing are different. Let us discuss each separately.

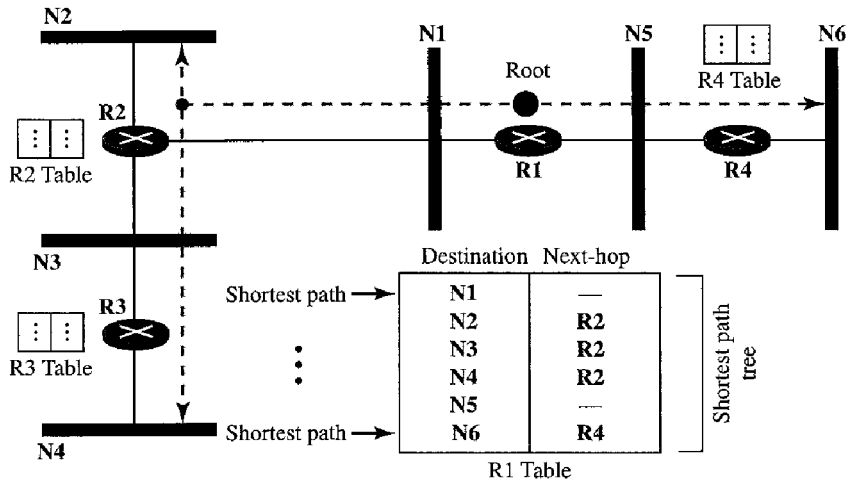
Unicast Routing In unicast routing, when a router receives a packet to forward, it needs to find the shortest path to the destination of the packet. The router consults its routing table for that particular destination. The next-hop entry corresponding to the destination is the start of the shortest path. The router knows the shortest path for each destination, which means that the router has a shortest path tree to optimally reach all destinations. In other words, each line of the routing table is a shortest path; the whole routing table is a shortest path tree. In unicast routing, each router needs only one shortest path tree to forward a packet; however, each router has its own shortest path tree. Figure 22.36 shows the situation.

The figure shows the details of the routing table and the shortest path tree for router R1. Each line in the routing table corresponds to one path from the root to the corresponding network. The whole table represents the shortest path tree.

In unicast routing, each router in the domain has a table that defines a shortest path tree to possible destinations.

Multicast Routing When a router receives a multicast packet, the situation is different from when it receives a unicast packet. A multicast packet may have destinations in more than one network. Forwarding of a single packet to members of a group requires a shortest path tree. If we have n groups, we may need n shortest path trees. We can imagine the complexity of multicast routing. Two approaches have been used to solve the problem: source-based trees and group-shared trees.

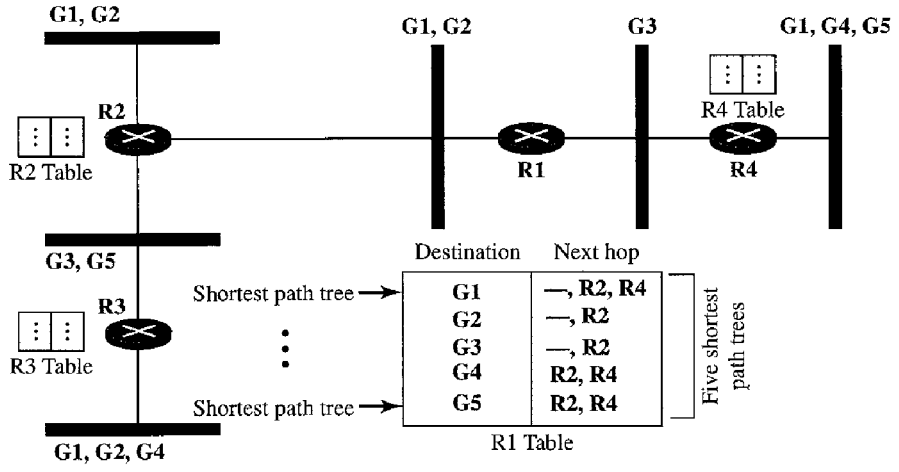
Figure 22.36 *Shortest path tree in unicast routing*



In multicast routing, each involved router needs to construct a shortest path tree for each group.

❑ **Source-Based Tree.** In the source-based tree approach, each router needs to have one shortest path tree for each group. The shortest path tree for a group defines the next hop for each network that has loyal member(s) for that group. In Figure 22.37, we assume that we have only five groups in the domain: G1, G2, G3, G4, and G5. At the moment G1 has loyal members in four networks, G2 in three, G3 in two, G4 in two, and G5 in two. We have shown the names of the groups with loyal members on each network. Figure 22.37 also shows the multicast routing table for router R1. There is one shortest path tree for each group; therefore there are five shortest path trees for five groups. If router R1 receives a packet with destination

Figure 22.37 *Source-based tree approach*

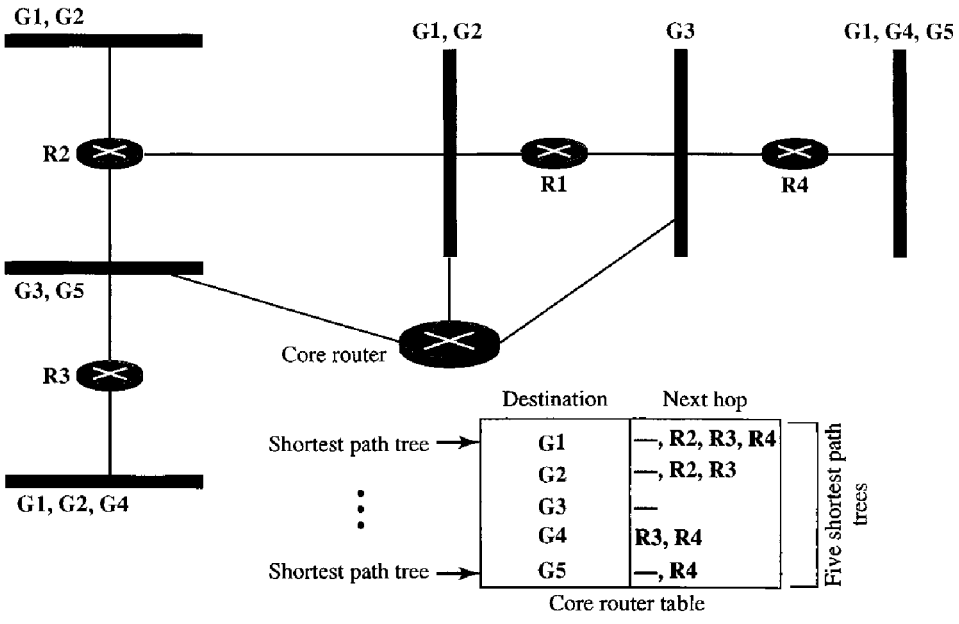


address G1, it needs to send a copy of the packet to the attached network, a copy to router R2, and a copy to router R4 so that all members of G1 can receive a copy. In this approach, if the number of groups is m, each router needs to have m shortest path trees, one for each group. We can imagine the complexity of the routing table if we have hundreds or thousands of groups. However, we will show how different protocols manage to alleviate the situation.

In the source-based tree approach, each router needs to have one shortest path tree for each group.

- **Group-Shared Tree.** In the **group-shared tree** approach, instead of each router having m shortest path trees, only one designated router, called the center core, or **rendezvous router,** takes the responsibility of distributing multicast traffic. The core has m shortest path trees in its routing table. The rest of the routers in the domain have none. If a router receives a multicast packet, it encapsulates the packet in a unicast packet and sends it to the core router. The core router removes the multicast packet from its capsule, and consults its routing table to route the packet. Figure 22.38 shows the idea.

Figure 22.38 *Group-shared tree approach*



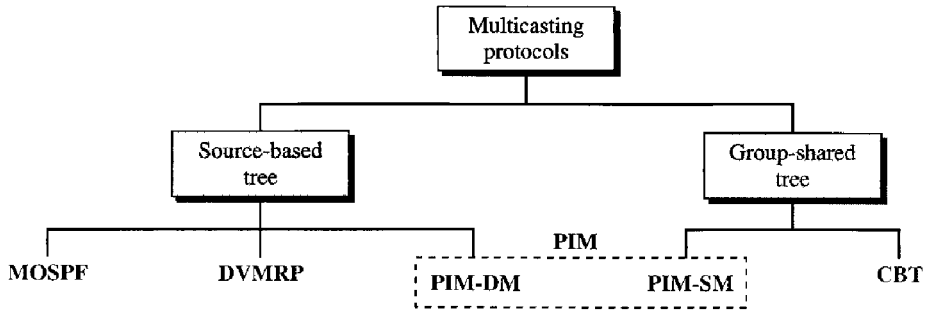
In the group-shared tree approach, only the core router, which has a shortest path tree for each group, is involved in multicasting.

Routing Protocols

During the last few decades, several multicast routing protocols have emerged. Some of these protocols are extensions of unicast routing protocols; others are totally new.

We discuss these protocols in the remainder of this chapter. Figure 22.39 shows the taxonomy of these protocols.

Figure 22.39 *Taxonomy of common multicast protocols*



Multicast Link State Routing: MOSPF

In this section, we briefly discuss multicast link state routing and its implementation in the Internet, MOSPF.

Multicast Link State Routing We discussed unicast link state routing in Section 22.3. We said that each router creates a shortest path tree by using Dijkstra's algorithm. The routing table is a translation of the shortest path tree. Multicast link state routing is a direct extension of unicast routing and uses a source-based tree approach. Although unicast routing is quite involved, the extension to multicast routing is very simple and straightforward.

Multicast link state routing uses the source-based tree approach.

Recall that in unicast routing, each node needs to advertise the state of its links. For multicast routing, a node needs to revise the interpretation of *state*. A node advertises every group which has any loyal member on the link. Here the meaning of state is "what groups are active on this link." The information about the group comes from IGMP (see Chapter 21). Each router running IGMP solicits the hosts on the link to find out the membership status.

When a router receives all these LSPs, it creates n (n is the number of groups) topologies, from which n shortest path trees are made by using Dijkstra's algorithm. So each router has a routing table that represents as many shortest path trees as there are groups.

The only problem with this protocol is the time and space needed to create and save the many shortest path trees. The solution is to create the trees only when needed. When a router receives a packet with a multicast destination address, it runs the Dijkstra algorithm to calculate the shortest path tree for that group. The result can be cached in case there are additional packets for that destination.

MOSPF **Multicast Open Shortest Path First (MOSPF)** protocol is an extension of the OSPF protocol that uses multicast link state routing to create source-based trees.

The protocol requires a new link state update packet to associate the unicast address of a host with the group address or addresses the host is sponsoring. This packet is called the group-membership LSA. In this way, we can include in the tree only the hosts (using their unicast addresses) that belong to a particular group. In other words, we make a tree that contains all the hosts belonging to a group, but we use the unicast address of the host in the calculation. For efficiency, the router calculates the shortest path trees on demand (when it receives the first multicast packet). In addition, the tree can be saved in cache memory for future use by the same source/group pair. MOSPF is a **data-driven** protocol; the first time an MOSPF router sees a datagram with a given source and group address, the router constructs the Dijkstra shortest path tree.

Multicast Distance Vector: DVMRP

In this section, we briefly discuss multicast distance vector routing and its implementation in the Internet, DVMRP.

Multicast Distance Vector Routing Unicast distance vector routing is very simple; extending it to support multicast routing is complicated. Multicast routing does not allow a router to send its routing table to its neighbors. The idea is to create a table from scratch by using the information from the unicast distance vector tables.

Multicast distance vector routing uses source-based trees, but the router never actually makes a routing table. When a router receives a multicast packet, it forwards the packet as though it is consulting a routing table. We can say that the shortest path tree is evanescent. After its use (after a packet is forwarded) the table is destroyed.

To accomplish this, the multicast distance vector algorithm uses a process based on four decision-making strategies. Each strategy is built on its predecessor. We explain them one by one and see how each strategy can improve the shortcomings of the previous one.

- **Flooding.** Flooding is the first strategy that comes to mind. A router receives a packet and, without even looking at the destination group address, sends it out from every interface except the one from which it was received. Flooding accomplishes the first goal of multicasting: every network with active members receives the packet. However, so will networks without active members. This is a broadcast, not a multicast. There is another problem: it creates loops. A packet that has left the router may come back again from another interface or the same interface and be forwarded again. Some flooding protocols keep a copy of the packet for a while and discard any duplicates to avoid loops. The next strategy, reverse path forwarding, corrects this defect.

Flooding broadcasts packets, but creates loops in the systems.

- **Reverse Path Forwarding (RPF).** RPF is a modified flooding strategy. To prevent loops, only one copy is forwarded; the other copies are dropped. In RPF, a router forwards only the copy that has traveled the shortest path from the source to the router. To find this copy, RPF uses the unicast routing table. The router receives a packet and extracts the source address (a unicast address). It consults its unicast routing table as though it wants to send a packet to the source address. The routing table tells the

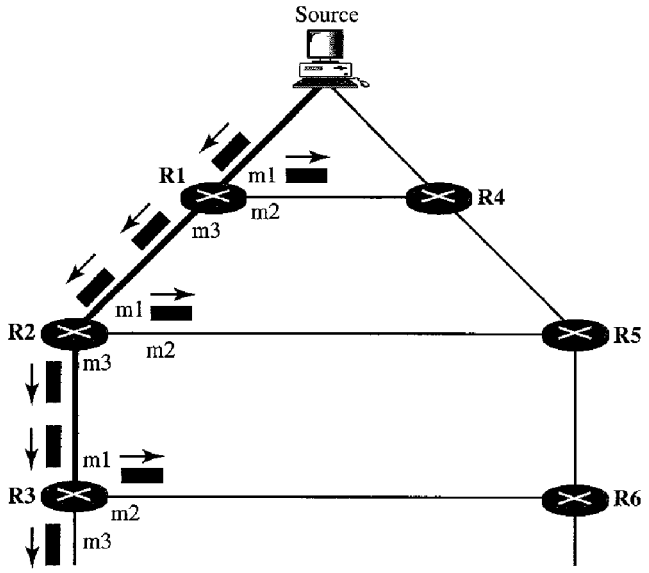
router the next hop. If the multicast packet has just come from the hop defined in the table, the packet has traveled the shortest path from the source to the router because the shortest path is reciprocal in unicast distance vector routing protocols. If the path from A to B is the shortest, then it is also the shortest from B to A. The router forwards the packet if it has traveled from the shortest path; it discards it otherwise.

This strategy prevents loops because there is always one shortest path from the source to the router. If a packet leaves the router and comes back again, it has not traveled the shortest path. To make the point clear, let us look at Figure 22.40.

Figure 22.40 shows part of a domain and a source. The shortest path tree as calculated by routers R1, R2, and R3 is shown by a thick line. When R1 receives a packet from the source through the interface m1, it consults its routing table and finds that the shortest path from R1 to the source is through interface m1. The packet is forwarded. However, if a copy of the packet has arrived through interface m2, it is discarded because m2 does not define the shortest path from R1 to the source. The story is the same with R2 and R3. You may wonder what happens if a copy of a packet that arrives at the m1 interface of R3, travels through R6, R5, R2, and then enters R3 through interface m1. This interface is the correct interface for R3. Is the copy of the packet forwarded? The answer is that this scenario never happens because when the packet goes from R5 to R2, it will be discarded by R2 and never reaches R3. The upstream routers toward the source always discard a packet that has not gone through the shortest path, thus preventing confusion for the downstream routers.

RPF eliminates the loop in the flooding process.

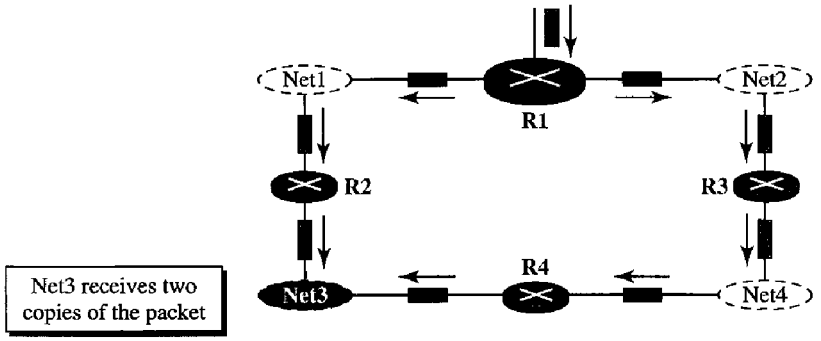
Figure 22.40 *Reverse path forwarding (RPF)*



- **Reverse Path Broadcasting (RPB).** RPF guarantees that each network receives a copy of the multicast packet without formation of loops. However, RPF does not

guarantee that each network receives only one copy; a network may receive two or more copies. The reason is that RPF is not based on the destination address (a group address); forwarding is based on the source address. To visualize the problem, let us look at Figure 22.41.

Figure 22.41 *Problem with RPF*

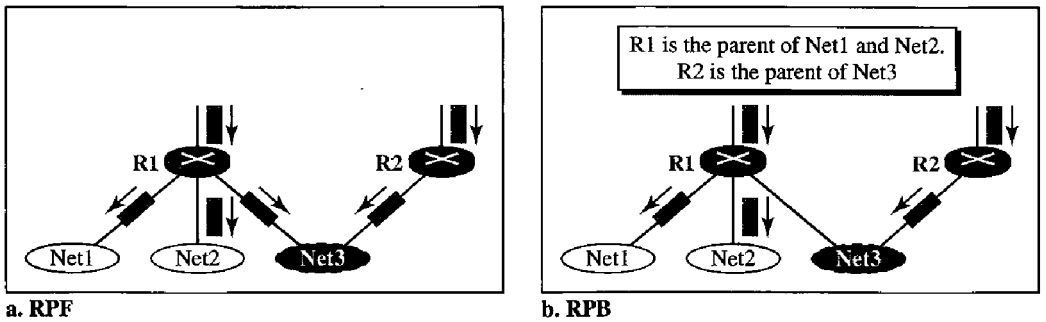


Net3 in this figure receives two copies of the packet even though each router just sends out one copy from each interface. There is duplication because a tree has not been made; instead of a tree we have a graph. Net3 has two parents: routers R2 and R4.

To eliminate duplication, we must define only one parent router for each network. We must have this restriction: A network can receive a multicast packet from a particular source only through a **designated parent router.**

Now the policy is clear. For each source, the router sends the packet only out of those interfaces for which it is the designated parent. This policy is called reverse path broadcasting (RPB). RPB guarantees that the packet reaches every network and that every network receives only one copy. Figure 22.42 shows the difference between RPF and RPB.

Figure 22.42 *RPF Versus RPB*



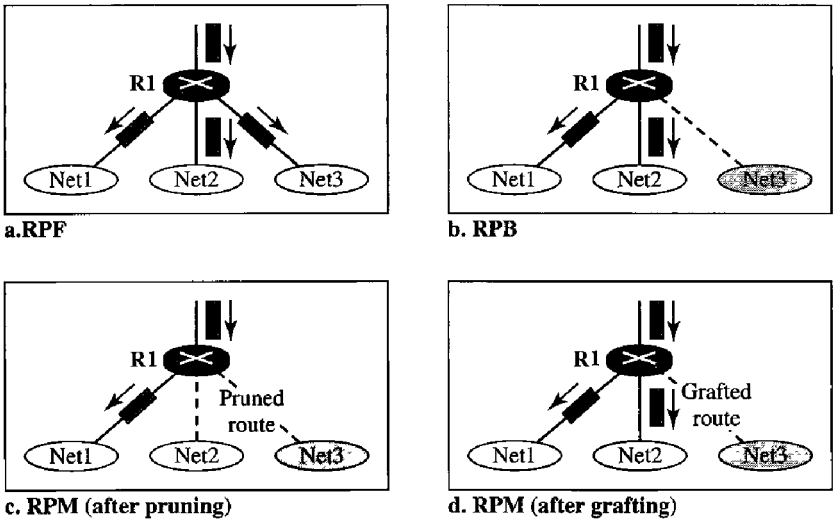
The reader may ask how the designated parent is determined. The designated parent router can be the router with the shortest path to the source. Because routers periodically

send updating packets to each other (in RIP), they can easily determine which router in the neighborhood has the shortest path to the source (when interpreting the source as the destination). If more than one router qualifies, the router with the smallest IP address is selected.

RPB creates a shortest path broadcast tree from the source to each destination. It guarantees that each destination receives one and only one copy of the packet.

- **Reverse Path Multicasting (RPM).** As you may have noticed, RPB does not multicast the packet, it broadcasts it. This is not efficient. To increase efficiency, the multicast packet must reach only those networks that have active members for that particular group. This is called **reverse path multicasting (RPM).** To convert broadcasting to multicasting, the protocol uses two procedures, pruning and grafting. Figure 22.43 shows the idea of pruning and grafting.

Figure 22.43 *RPF, RPB, and RPM*



The designated parent router of each network is responsible for holding the membership information. This is done through the IGMP protocol described in Chapter 21. The process starts when a router connected to a network finds that there is no interest in a multicast packet. The router sends a **prune message** to the upstream router so that it can exclude the corresponding interface. That is, the upstream router can stop sending multicast messages for this group through that interface. Now if this router receives prune messages from all downstream routers, it, in turn, sends a prune message to its upstream router.

What if a leaf router (a router at the bottom of the tree) has sent a prune message but suddenly realizes, through IGMP, that one of its networks is again interested in receiving the multicast packet? It can send a **graft message.** The graft message forces the upstream router to resume sending the multicast messages.

RPM adds pruning and grafting to RPB to create a multicast shortest path tree that supports dynamic membership changes.

DVMRP The **Distance Vector Multicast Routing Protocol (DVMRP)** is an implementation of multicast distance vector routing. It is a source-based routing protocol, based on RIP.

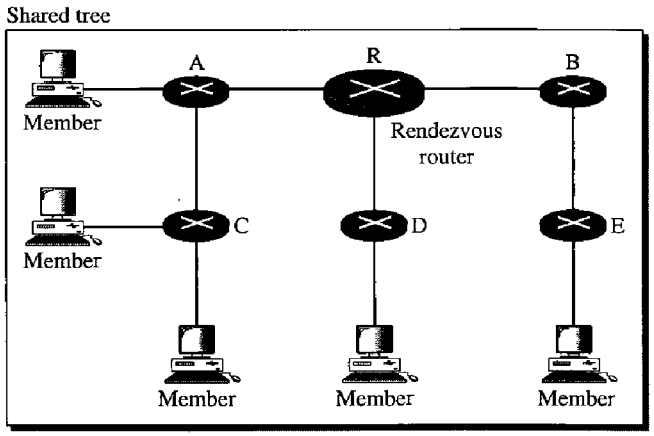
CBT

The **Core-Based Tree (CBT) protocol** is a group-shared protocol that uses a core as the root of the tree. The autonomous system is divided into regions, and a core (center router or rendezvous router) is chosen for each region.

Formation of the Tree After the rendezvous point is selected, every router is informed of the unicast address of the selected router. Each router then sends a unicast join message (similar to a grafting message) to show that it wants to join the group. This message passes through all routers that are located between the sender and the rendezvous router. Each intermediate router extracts the necessary information from the message, such as the unicast address of the sender and the interface through which the packet has arrived, and forwards the message to the next router in the path. When the rendezvous router has received all join messages from every member of the group, the tree is formed. Now every router knows its upstream router (the router that leads to the root) and the downstream router (the router that leads to the leaf).

If a router wants to leave the group, it sends a leave message to its upstream router. The upstream router removes the link to that router from the tree and forwards the message to its upstream router, and so on. Figure 22.44 shows a group-shared tree with its rendezvous router.

Figure 22.44 *Group-shared tree with rendezvous router*

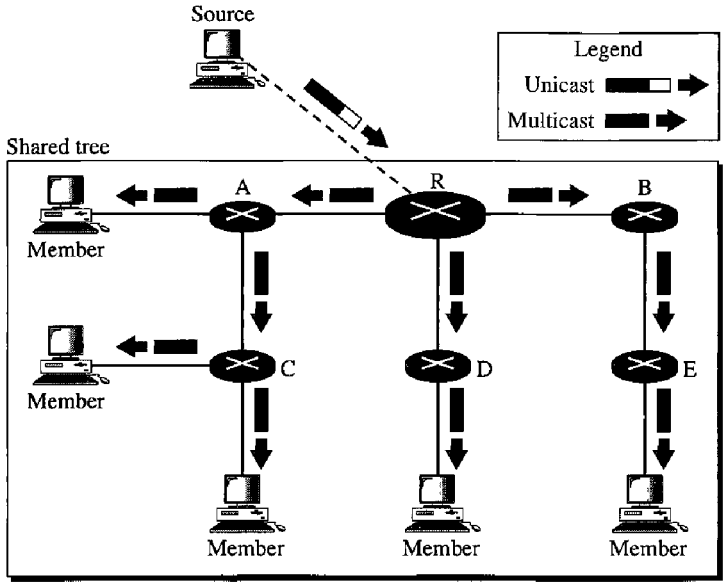


The reader may have noticed two differences between DVMRP and MOSPF, on one hand, and CBT, on the other. First, the tree for the first two is made from the root up; the tree for CBT is formed from the leaves down. Second, in DVMRP, the tree is

first made (broadcasting) and then pruned; in CBT, there is no tree at the beginning; the joining (grafting) gradually makes the tree.

Sending Multicast Packets After formation of the tree, any source (belonging to the group or not) can send a multicast packet to all members of the group. It simply sends the packet to the rendezvous router, using the unicast address of the rendezvous router; the rendezvous router distributes the packet to all members of the group. Figure 22.45 shows how a host can send a multicast packet to all members of the group. Note that the source host can be any of the hosts inside the shared tree or any host outside the shared tree. In Figure 22.45 we show one located outside the shared tree.

Figure 22.45 *Sending a multicast packet to the rendezvous router*



Selecting the Rendezvous Router This approach is simple except for one point. How do we select a rendezvous router to optimize the process and multicasting as well? Several methods have been implemented. However, this topic is beyond the scope of this book, and we leave it to more advanced books.

In summary, the Core-Based Tree (CBT) is a group-shared tree, center-based protocol using one tree per group. One of the routers in the tree is called the core. A packet is sent from the source to members of the group following this procedure:

1. The source, which may or may not be part of the tree, encapsulates the multicast packet inside a unicast packet with the unicast destination address of the core and sends it to the core. This part of delivery is done using a unicast address; the only recipient is the core router.
2. The core decapsulates the unicast packet and forwards it to all interested interfaces.
3. Each router that receives the multicast packet, in turn, forwards it to all interested interfaces.

In CBT, the source sends the multicast packet (encapsulated in a unicast packet) to the core router. The core router decapsulates the packet and forwards it to all interested interfaces.

PIM

Protocol Independent Multicast (PIM) is the name given to two independent multicast routing protocols: **Protocol Independent Multicast, Dense Mode (PIM-DM)** and **Protocol Independent Multicast, Sparse Mode (PIM-SM).** Both protocols are unicast-protocol-dependent, but the similarity ends here. We discuss each separately.

PIM-DM PIM-DM is used when there is a possibility that each router is involved in multicasting **(dense mode).** In this environment, the use of a protocol that broadcasts the packet is justified because almost all routers are involved in the process.

PIM-DM is used in a dense multicast environment, such as a LAN.

PIM-DM is a source-based tree routing protocol that uses RPF and pruning and grafting strategies for multicasting. Its operation is like that of DVMRP; however, unlike DVMRP, it does not depend on a specific unicasting protocol. It assumes that the autonomous system is using a unicast protocol and each router has a table that can find the outgoing interface that has an optimal path to a destination. This unicast protocol can be a distance vector protocol (RIP) or link state protocol (OSPF).

PIM-DM uses RPF and pruning and grafting strategies to handle multicasting. However, it is independent of the underlying unicast protocol.

PIM-SM PIM-SM is used when there is a slight possibility that each router is involved in multicasting (sparse mode). In this environment, the use of a protocol that broadcasts the packet is not justified; a protocol such as CBT that uses a group-shared tree is more appropriate.

PIM-SM is used in a sparse multicast environment such as a WAN.

PIM-SM is a group-shared tree routing protocol that has a rendezvous point (RP) as the source of the tree. Its operation is like CBT; however, it is simpler because it does not require acknowledgment from a join message. In addition, it creates a backup set of RPs for each region to cover RP failures.

One of the characteristics of PIM-SM is that it can switch from a group-shared tree strategy to a source-based tree strategy when necessary. This can happen if there is a dense area of activity far from the RP. That area can be more efficiently handled with a source-based tree strategy instead of a group-shared tree strategy.

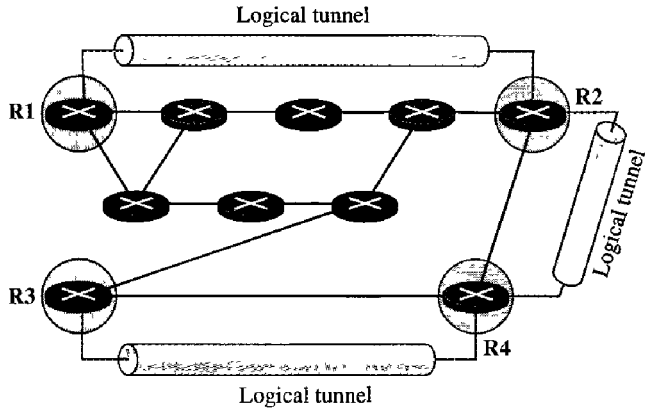
PIM-SM is similar to CBT but uses a simpler procedure.

MBONE

Multimedia and real-time communication have increased the need for multicasting in the Internet. However, only a small fraction of Internet routers are multicast routers. In

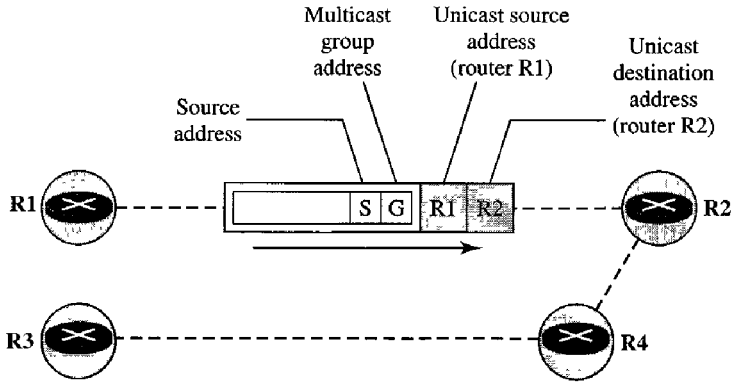
other words, a **multicast router** may not find another multicast router in the neighborhood to forward the multicast packet. Although this problem may be solved in the next few years by adding more and more multicast routers, there is another solution to this problem. The solution is **tunneling.** The multicast routers are seen as a group of routers on top of unicast routers. The multicast routers may not be connected directly, but they are connected logically. Figure 22.46 shows the idea. In Figure 22.46, only the routers enclosed in the shaded circles are capable of multicasting. Without tunneling, these routers are isolated islands. To enable multicasting, we make a **multicast backbone (MBONE)** out of these isolated routers by using the concept of tunneling.

Figure 22.46 *Logical tunneling*



A **logical tunnel** is established by encapsulating the multicast packet inside a unicast packet. The multicast packet becomes the payload (data) of the unicast packet. The intermediate (nonmulticast) routers forward the packet as unicast routers and deliver the packet from one island to another. It's as if the unicast routers do not exist and the two multicast routers are neighbors. Figure 22.47 shows the concept. So far the only protocol that supports MBONE and tunneling is DVMRP.

Figure 22.47 *MBONE*



22.5 RECOMMENDED READING

For more details about subjects discussed in this chapter, we recommend the following books and sites. The items in brackets [. . .] refer to the reference list at the end of the text.

Books

Delivery and forwarding are discussed in Chapter 6 of [For06]. Unicast routing protocols are discussed in Chapter 14 of [For06]. Multicasting and multicast routing are discussed in Chapter 15 of [For06]. For a complete discussion of multicasting see [WZ01]. For routing protocols see [Hui00]. OSPF is discussed in [Moy98].

Sites

- www.ietf.org/rfc.html Information about RFCs

RFCs

A discussion of RIP can be found in the following RFCs:

1131, 1245, 1246, 1247, 1370, 1583, 1584, 1585, 1586, 1587, 1722, 1723, 2082, 2453

A discussion of OSPF can be found in the following RFCs:

1131, 1245, 1246, 1247, 1370, 1583, 1584, 1585, 1586, 1587, 2178, 2328, 2329, 2370

A discussion of BGP can be found in the following RFCs:

1092, 1105, 1163, 1265, 1266, 1267, 1364, 1392, 1403, 1565, 1654, 1655, 1665, 1771, 1772, 1745, 1774, 2283

22.6 KEY TERMS

address aggregation
area
area border routers
area identification
autonomous system (AS)
backbone router
Border Gateway Protocol (BGP)
broadcasting
Core-Based Tree (CBT) protocol
data-driven
default method
delivery
designated parent router
Dijkstra's algorithm
direct delivery
distance learning
Distance Vector Multicast Routing Protocol (DVMRP)
distance vector routing
distributed database
dynamic routing method
dynamic routing table
flooding
forwarding
graft message
group-shared tree

hierarchical routing
hop count
host-specific method
ifconfig
immediate neighbors
indirect delivery
interdomain routing
intradomain routing
least-cost tree
link state routing
logical tunnel
longest mask matching
metric
multicast backbone (MBONE)
Multicast Open Shortest Path First (MOSPF)
multicast router
multicast routing
multicasting
multiple unicasting
netstat
network-specific method
next-hop address
next-hop method
Open Shortest Path First (OSPF)
optional attribute
OSPF protocol
path vector routing
point-to-point link
poison reverse
policy routing
Protocol Independent Multicast (PIM)
Protocol Independent Multicast, Dense Mode (PIM-DM)
Protocol Independent Multicast, Sparse Mode (PIM-SM)
prune message
rendezvous router
rendezvous-point tree
reverse path broadcasting (RPB)
reverse path forwarding (RPF)
reverse path multicasting (RPM)
route method
routing
Routing Information Protocol (RIP)
routing protocols
shortest path tree
slow convergence
source-based tree
speaker node
split horizon
static routing table
stub link
switching fabric
teleconferencing
transient link
triggered update
tunneling
unicasting
update message
virtual link
well-known attribute

22.7 SUMMARY

- The delivery of a packet is called direct if the deliverer (host or router) and the destination are on the same network; the delivery of a packet is called indirect if the deliverer (host or router) and the destination are on different networks.
- In the next-hop method, instead of a complete list of the stops the packet must make, only the address of the next hop is listed in the routing table; in the network-specific method, all hosts on a network share one entry in the routing table.

- In the host-specific method, the full IP address of a host is given in the routing table.
- In the default method, a router is assigned to receive all packets with no match in the routing table.
- The routing table for classless addressing needs at least four columns.
- Address aggregation simplifies the forwarding process in classless addressing.
- Longest mask matching is required in classless addressing.
- Classless addressing requires hierarchical and geographical routing to prevent immense routing tables.
- A static routing table's entries are updated manually by an administrator; a dynamic routing table's entries are updated automatically by a routing protocol.
- A metric is the cost assigned for passage of a packet through a network.
- An autonomous system (AS) is a group of networks and routers under the authority of a single administration.
- RIP is based on distance vector routing, in which each router shares, at regular intervals, its knowledge about the entire AS with its neighbors.
- Two shortcomings associated with the RIP protocol are slow convergence and instability. Procedures to remedy RIP instability include triggered update, split horizons, and poison reverse.
- OSPF divides an AS into areas, defined as collections of networks, hosts, and routers.
- OSPF is based on link state routing, in which each router sends the state of its neighborhood to every other router in the area. A packet is sent only if there is a change in the neighborhood.
- OSPF routing tables are calculated by using Dijkstra's algorithm.
- BGP is an interautonomous system routing protocol used to update routing tables.
- BGP is based on a routing protocol called path vector routing. In this protocol, the ASs through which a packet must pass are explicitly listed.
- In a source-based tree approach to multicast routing, the source/group combination determines the tree; in a group-shared tree approach to multicast routing, the group determines the tree.
- MOSPF is a multicast routing protocol that uses multicast link state routing to create a source-based least-cost tree.
- In reverse path forwarding (RPF), the router forwards only the packets that have traveled the shortest path from the source to the router.
- Reverse path broadcasting (RPB) creates a shortest path broadcast tree from the source to each destination. It guarantees that each destination receives one and only one copy of the packet.
- Reverse path multicasting (RPM) adds pruning and grafting to RPB to create a multicast shortest path tree that supports dynamic membership changes.
- DVMRP is a multicast routing protocol that uses the distance routing protocol to create a source-based tree.
- The Core-Based Tree (CBT) protocol is a multicast routing protocol that uses a router as the root of the tree.

- PIM-DM is a source-based tree routing protocol that uses RPF and pruning and grafting strategies to handle multicasting.
- PIM-SM is a group-shared tree routing protocol that is similar to CBT and uses a rendezvous router as the source of the tree.
- For multicasting between two noncontiguous multicast routers, we make a multicast backbone (MBONE) to enable tunneling.

22.8 PRACTICE SET

Review Questions

1. What is the difference between a direct and an indirect delivery?
2. List three forwarding techniques discussed in this chapter and give a brief description of each.
3. Contrast two different routing tables discussed in this chapter.
4. What is the purpose of RIP?
5. What are the functions of a RIP message?
6. Why is the expiration timer value 6 times that of the periodic timer value?
7. How does the hop count limit alleviate RIP's problems?
8. List RIP shortcomings and their corresponding fixes.
9. What is the basis of classification for the four types of links defined by OSPF?
10. Why do OSPF messages propagate faster than RIP messages?
11. What is the purpose of BGP?
12. Give a brief description of two groups of multicast routing protocols discussed in this chapter.

Exercises

13. Show a routing table for a host that is totally isolated.
14. Show a routing table for a host that is connected to a LAN without being connected to the Internet.
15. Find the topology of the network if Table 22.3 is the routing table for router R1.

Table 22.3 *Routing table for Exercise 15*

Mask	*Network Address*	*Next-Hop Address*	*Interface*
/27	202.14.17.224	—	m1
/18	145.23.192.0	—	m0
Default	Default	130.56.12.4	m2

16. Can router R1 in Figure 22.8 receive a packet with destination address 140.24.7.194? Explain your answer.

17. Can router R1 in Figure 22.8 receive a packet with destination address 140.24.7.42? Explain your answer.
18. Show the routing table for the regional ISP in Figure 22.9.
19. Show the routing table for local ISP 1 in Figure 22.9.
20. Show the routing table for local ISP 2 in Figure 22.9.
21. Show the routing table for local ISP 3 in Figure 22.9.
22. Show the routing table for small ISP 1 in Figure 22.9.
23. Contrast and compare distance vector routing with link state routing.
24. A router has the following RIP routing table:

Net1	4	B
Net2	2	C
Net3	1	F
Net4	5	G

What would be the contents of the table if the router received the following RIP message from router C?

Net1	2
Net2	1
Net3	3
Net4	7

25. How many bytes are empty in a RIP message that advertises *N* networks?
26. A router has the following RIP routing table:

Net1	4	B
Net2	2	C
Net3	1	F
Net4	5	G

Show the response message sent by this router.

27. Show the autonomous system with the following specifications:
    a. There are eight networks (N1 to N8).
    b. There are eight routers (R1 to R8).
    c. N1, N2, N3, N4, N5, and N6 are Ethernet LANs.
    d. N7 and N8 are point-to-point WANs.
    e. R1 connects N1 and N2.
    f. R2 connects N1 and N7.
    g. R3 connects N2 and N8.
    h. R4 connects N7 and N6.
    i. R5 connects N6 and N3.
    j. R6 connects N6 and N4.
    k. R7 connects N6 and N5.
    l. R8 connects N8 and N5.

28. Draw the graphical representation of the autonomous system of Exercise 27 as seen by OSPF.
29. Which of the networks in Exercise 27 is a transient network? Which is a stub network?
30. A router using DVMRP receives a packet with source address 10.14.17.2 from interface 2. If the router forwards the packet, what are the contents of the entry related to this address in the unicast routing table?
31. Does RPF actually create a shortest path tree? Explain.
32. Does RPB actually create a shortest path tree? Explain. What are the leaves of the tree?
33. Does RPM actually create a shortest path tree? Explain. What are the leaves of the tree?

Research Activities

34. If you have access to UNIX (or LINUX), use *netstat* and *ifconfig* to find the routing table for the server to which you are connected.
35. Find out how your ISP uses address aggregation and longest mask match principles.
36. Find out whether your IP address is part of the geographical address allocation.
37. If you are using a router, find the number and names of the columns in the routing table.

PART

5

Transport Layer

Objectives

The transport layer is responsible for process-to-process delivery of the entire message. A process is an application program running on a host. Whereas the network layer oversees source-to-destination delivery of individual packets, it does not recognize any relationship between those packets. It treats each one independently, as though each piece belonged to a separate message, whether or not it does. The transport layer, on the other hand, ensures that the whole message arrives intact and in order, overseeing both error control and flow control at the source-to-destination level.

The transport layer is responsible for the delivery of a message from one process to another.

Computers often run several programs at the same time. For this reason, source-to-destination delivery means delivery not only from one computer to the next but also from a specific process on one computer to a specific process on the other. The transport layer header must therefore include a type of address called a *service-point address* in the OSI model and port number or port addresses in the Internet and TCP/IP protocol suite.

A transport layer protocol can be either connectionless or connection-oriented. A connectionless transport layer treats each segment as an independent packet and delivers it to the transport layer at the destination machine. A connection-oriented transport layer makes a connection with the transport layer at the destination machine first before delivering the packets. After all the data is transferred, the connection is terminated.

In the transport layer, a message is normally divided into transmittable segments. A connectionless protocol, such as UDP, treats each segment separately. A connection-oriented protocol, such as TCP and SCTP, creates a relationship between the segments using sequence numbers.

Like the data link layer, the transport layer may be responsible for flow and error control. However, flow and error control at this layer is performed end to end rather than across a single link. We will see that one of the protocols discussed in this part of