UNIT - III

NETWORK LAYER

Communication at the network layers is host-to-host (Computer-to-Computer).

LOGICAL ADDRESSING

Logical addressing is implemented by network layer.

Logical addressing is a Global Addressing scheme.

Data-link layer handles the addressing problem locally, but if packets passes the network boundary there is a need for logical addressing system to help distinguish source and destination systems.

The network layer adds a header to the packet coming from the upper layer that includes the logical addresses of the sender and receiver.

There are 2 types of addressing mechanisms are present:

1. IPv4 (IP version4)
2. IPv6 (IP version 6)

IPv4 ADDRESSES

An **IPv4** address is a **32-bit** address that **uniquely** and **universally** defines the connection of a device to the Internet.

Unique: Two devices on the Internet can never have the same address at the same time.

Universal: The addressing system must be accepted by any host that wants to be connected to the Internet.

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**).

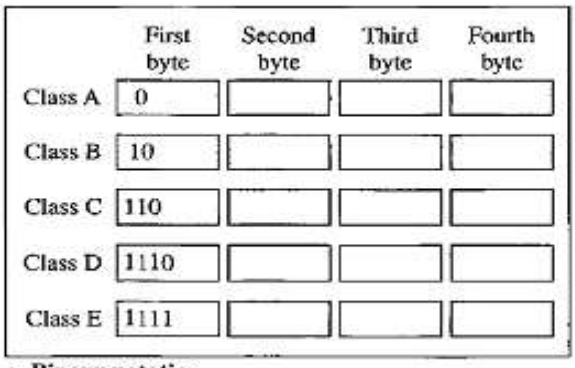
Notations

There are two notations to show an IPv4 address: Binary and Dotted-Decimal Notation.

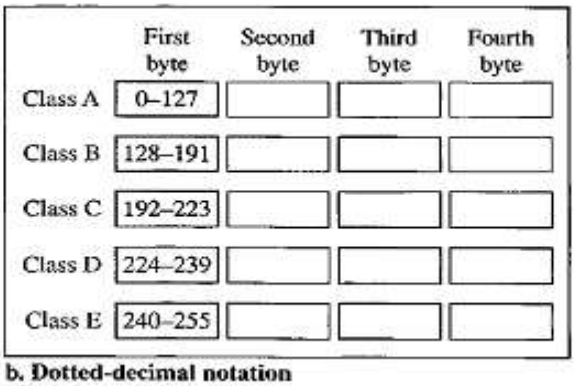
Binary	**Dotted-Decimal**
• IPv4 address is displayed as 32 bits. Each octet is often referred to as a byte. • It is a 4 byte address Ex: 10000000 00001011 00000011 00011111	• Internet addresses are written in decimal form with a dot separating the bytes. • Each number in dotted-decimal notation is a value ranging from 0 to 255. Ex: 128.11.3.31

CLASSFUL ADDRESSING

- Initially IPv4 used the concept of Classful addressing.
- In classful addressing, the address space is divided into five classes: A, B, C, D, and E.
- 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.



a. Binary notation



b. Dotted-decimal notation

Classes and Blocks

- Each class is divided into a fixed number of blocks.
- Size of the each block is also fixed.

Class	No 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

Purpose of classes:

- **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.
- **Class D** addresses were designed for multicasting.
- **Class E** addresses were reserved for future use.

Problem with above classes and block sizes: **A lot of addresses are wasted.**

1. A block in **class A** address is too large for almost any organization. (i.e) most of the addresses in class A were wasted and were not used.
2. A block in **class B** is also very large, probably too large for many of the organizations that received a class B block.
3. A block in **class C** is probably too small for many organizations.

4. **Class D** addresses were designed for multicasting, each address in class D is used to define one group of hosts on the Internet. The Internet authorities wrongly predicted a need for 268,435,456 groups.
5. **Class E** addresses were reserved for future use; only a few addresses were used till now.

Netid and Hostid

- In classful addressing, an IP address in class A, B, or C is divided into Netid and Hostid.
- For Class D,E there were no Netid and Hostid.
- 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 or Default Mask

- A default mask is a 32-bit number made of contiguous 1's followed by contiguous 0's.
- 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 masks for classes A, B, and C are:

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**

Subnetting

- Subnetting is a process of dividing a large block into smaller contiguous groups and assigns each group to smaller networks (subnets) or share a part of the addresses with neighbors.
- Subnetting increases the number of 1's in the mask.

Supernetting

- In supernetting, an organization can combine several blocks to create a larger range of addresses. Supernetting decreases the number of 1's in the mask.

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.

Address Depletion

- The number of available IPv4 addresses is decreasing as the number of internet users are increasing.
- 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**.

CLASSLESS ADDRESSING

Purpose

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

Address Blocks

- In classless addressing, when a small or large entity, needs to be connected to the Internet, it is granted a block of addresses.
- The size of the block (the number of addresses) varies based on the nature and size of the entity.

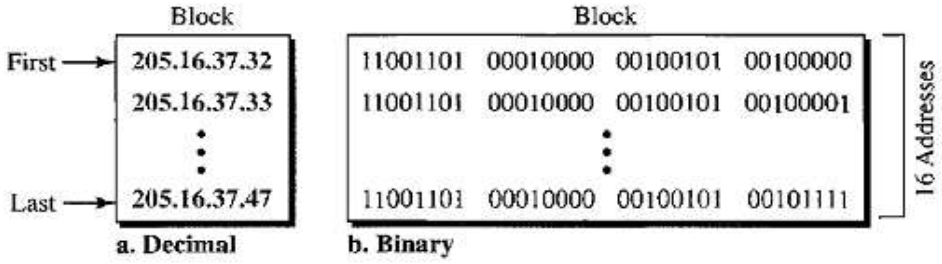
Example:

- For a large organization may be given thousands of addresses
- For a house two addresses are sufficient
- An Internet service provider may be given hundreds of thousands based on the number of customers it may serve.

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.

Consider the below figure for classless addressing that shows a block of addresses, in both binary and dotted-decimal notation, granted to a small business that needs 16 addresses.



It satisfies all 3 restrictions:

- The addresses are contiguous.
- The number of addresses is a power of 2 ($16 = \mathbf{2^4}$).
- 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.

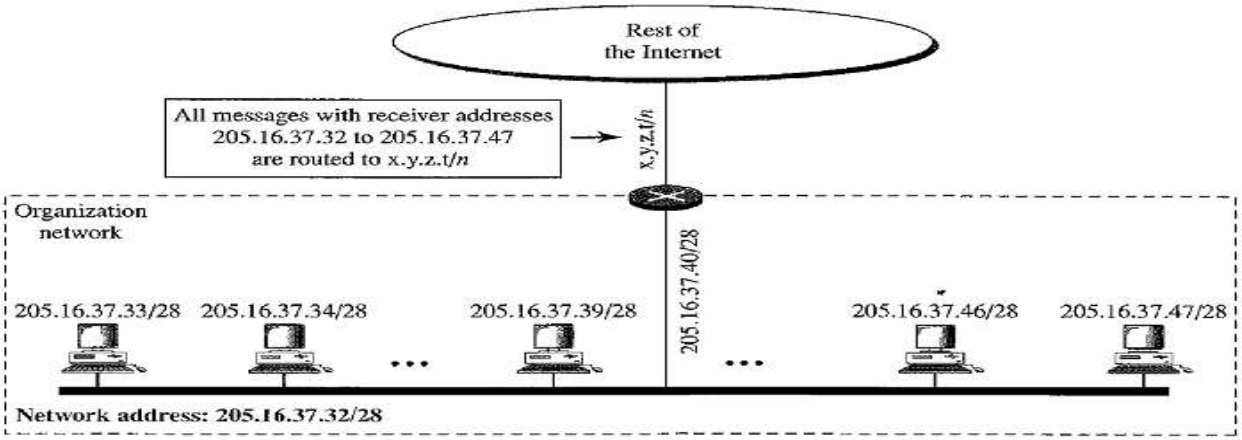
Mask

A mask is a 32-bit number in which the ***n*** leftmost bits are 1's and the **32 - *n*** rightmost bits are 0's, where ***n= 0 to 32.***

In 1Pv4 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. ***/n*** is called as CIDR notation.

- **First Address** in the block can be found by setting the rightmost **32 - *n*** bits to 0's.
- **Last Address** in the block can be found by setting the rightmost **32 - *n*** bits to 1's.
- **Number of Addresses** in the block can be found by using the formula $\boldsymbol{2^{32-n}}$.

- **Network Address** is the first address in the block and defines the organization network. Usually the first address is used by routers to direct the message sent to the organization from the outside.



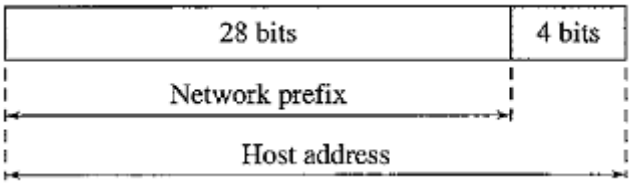
Example: **205.16.37.39/28** or **11001101 00010000 00100101 00100111**

- First address: 11001101 000100000100101 0010000 or 205.16.37.**32**
- Last address: 11001101 00010000 001001010010 1111 or 205.16.37.**47**
- Number of Addresses: $\mathbf{2^{32-28} = 2^4 = 16.}$
- Network Address (First Address) 11001101 000100000100101 0010000 or 205.16.37.32

Netid and Hostid

- The ***n*** leftmost bits of the address **x.y.z.t/n** define the **network address** or **prefix**.
- The **(32 – *n)*** rightmost bits define the particular **suffix** or **host address** (computer or router) connected to the network.

205.16.37.39/28 or **11001101 00010000 00100101 0010 0111**



Network Address Translation (NAT)

- As the number of home users and small business users are increasing day by day it is not possible to give each and every user to one IPv4 address due to shortage of IPv4 addresses.
- In order to overcome this problem the developers designed the concepts of private IP address and Network Address Translation (NAT).
- NAT enables a user to have a large set of addresses internally (private IP addresses) and a small set of addresses externally (public IP addresses).

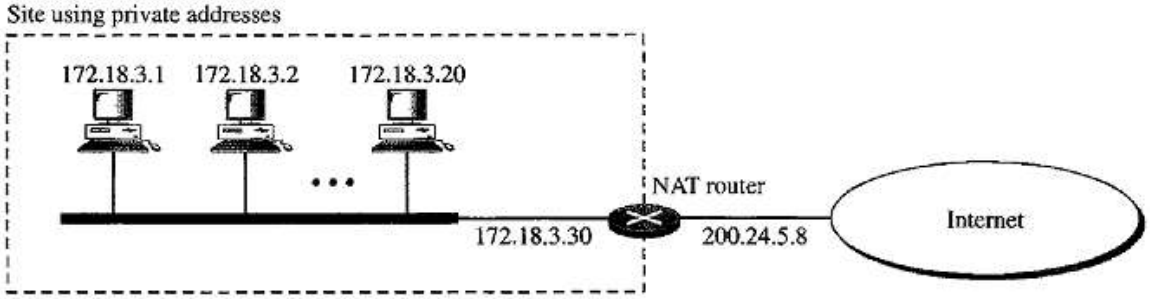
The Internet authorities have reserved three sets of addresses as private addresses:

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.
- 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.

Example:

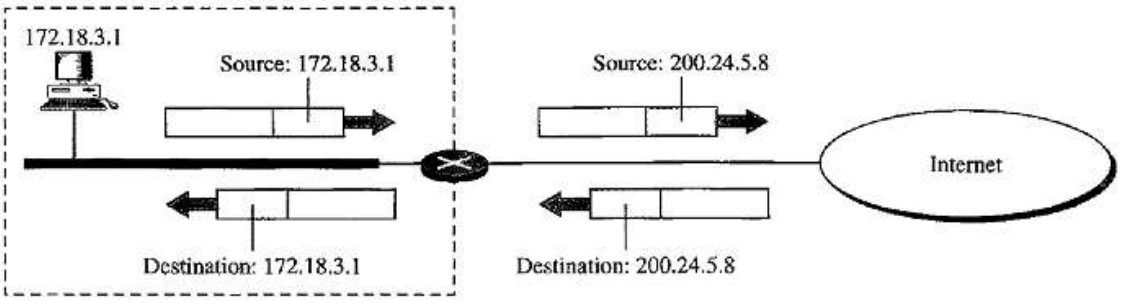
Consider the below figure describes the private network with private addresses:



- The NAT router has one public address **200.24.5.8,** it is a global address.
- The internal devices having addresses from **172.18.3.1** to **172.18.3.30** these are local addresses.

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.**



Translation Table

There are two types of translation tables:

1. Two Column translation table (Using one IP address)
2. Five column translation table (Using IP addresses and Port Numbers)

Two Column Translation Table

- It contains two columns: Private Address and External Address.
- In this strategy, communication must always be initiated by the private network.
- 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.

Translation Table

Private	External
172.18.3.1	25.8.2.10
…	….
….	….

Five Column Translation Table

To allow a many-to-many relationship between private-network hosts and external server programs, we need more information (columns) in the translation table such as:

1. Private Address
2. Private Port
3. External Address
4. External Port
5. Transport Protocol

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.

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.

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: The temporary port numbers must be unique.

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
…	…	…	…	…

IPv6 ADDRESSES (Internetworking Protocol version6)

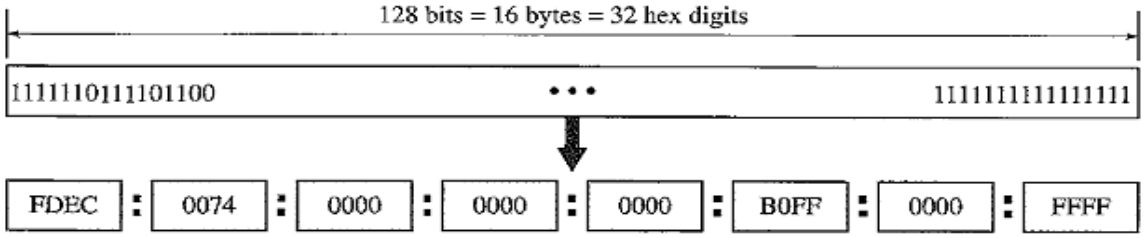
Why IPv6?

- In order to overcome the problems of address depletion.
- It eliminates the concept of NAT and Private Addresses.
- There is no need for classless addressing and DHCP.

Structure

IPv6 specifies hexadecimal colon notation (0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F).

An IPv6 address consists of 16 bytes (octets); it is 128 bits long. 128 bits is divided into eight sections. Each of 4 hex digits separated by a colon.

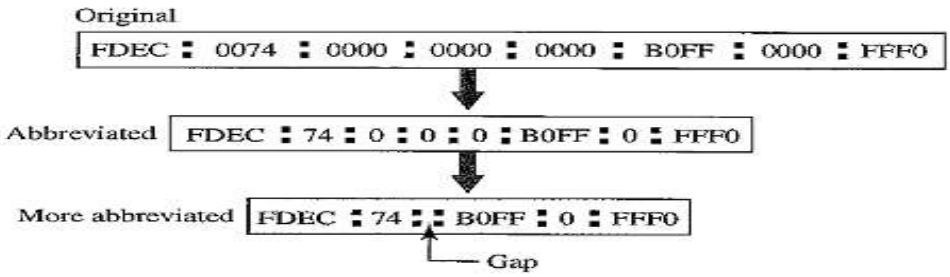


Abbreviation

- Hexa decimal format of IPv6 is very long and many of the digits are zeros.
- We can abbreviate this address as the leading zeros of a section (four digits between two colons) can be omitted.

Note: Only leading zeros are omitted not trailing zeros.

Example:



- In the above example: 0074 written as 74, 0000 written as 0.
- If there are consecutive sections consisting of zeros only. We can remove the zeros altogether and replace them with a double semicolon.

Address Space

- IPv6 has 2^{128} addresses are available. It is a much larger address space than IPv4.
- The designers of IPv6 divided the address into several categories.
- A few leftmost bits called the **type prefix**, in each address define its category. Type Prefix is variable in length.

Type Prefix	Type	Fraction
00000000	Reserved	1/256
0000001	ISO network address	1/128
0000010	IPX Novell network address	1/128
010	Provider-based unicast addresses	1/8
100	Geographic-based unicast addresses	1/8
1111 111010	Link local addresses	1/1024
1111 1110 11	Site local addresses	1/1024
11111111	Multicast addresses	1/256

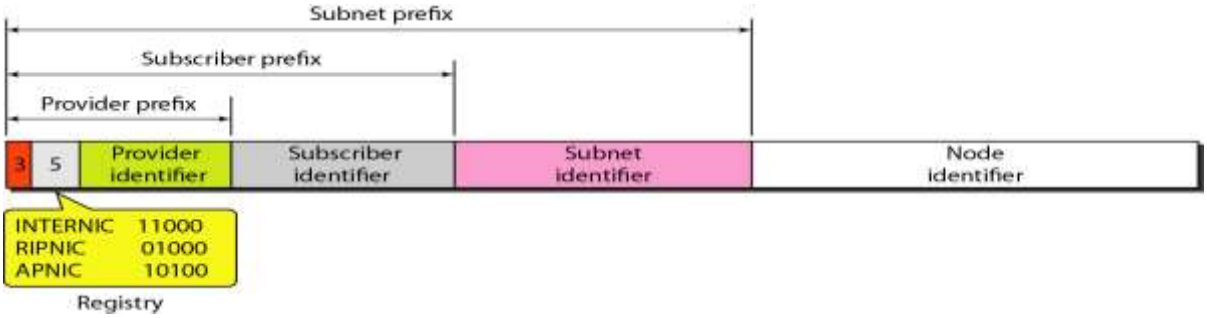
Unicast Addresses

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.

The provider-based address is generally used by a normal host as a unicast address.

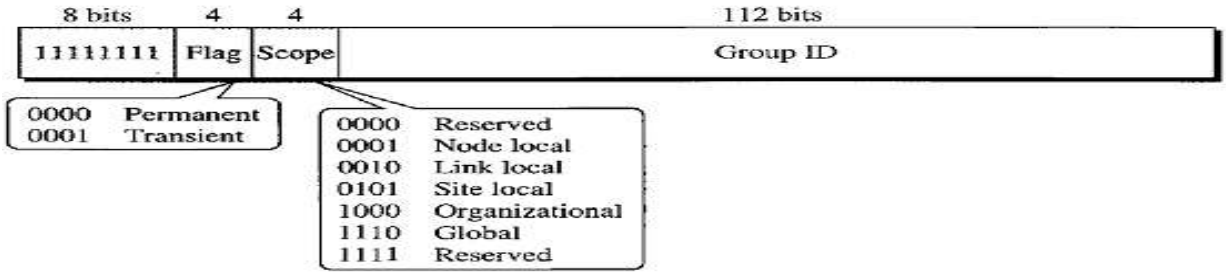


Fields for the provider-based address are as follows:

- **Type identifier (3 bits)** It defines the address as a provider-based address.
- **Registry identifier (5 bits)**
  This 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.
  - APNIC (code 10100) is for Asian and Pacific countries.
- **Provider identifier (16 bits)**
  This variable-length field identifies the provider for Internet access (such as an ISP).
- **Subscriber identifier (24 bits)**
  When an organization subscribes to the Internet through a provider, it is assigned subscriber identification.
- **Subnet identifier (32-bits)**
  The subnet identifier defines a specific subnetwork under the territory of the subscriber.
- **Node identifier (48 bits)**
  It 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.

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.



- 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 is used only temporarily. Systems engaged in a teleconference can use a transient group address.
- The third field defines the scope of the group address.

Reserved Addresses

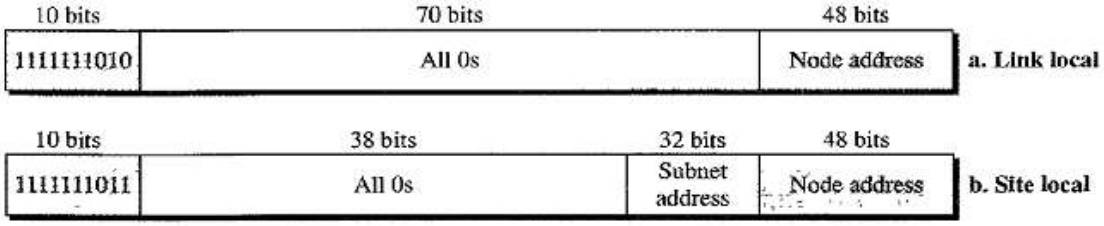
Reserved address are the another category of Address Space. These addresses start with eight 0's (type prefix is 00000000).

Local Addresses

- These addresses are used when an organization wants to use IPv6 protocol without being connected to the global Internet.
- That means they provide addressing for private networks. Nobody outside the organization can send a message to the nodes using these addresses.

There are two types of addresses are defined for this purpose: Link local and Site local.

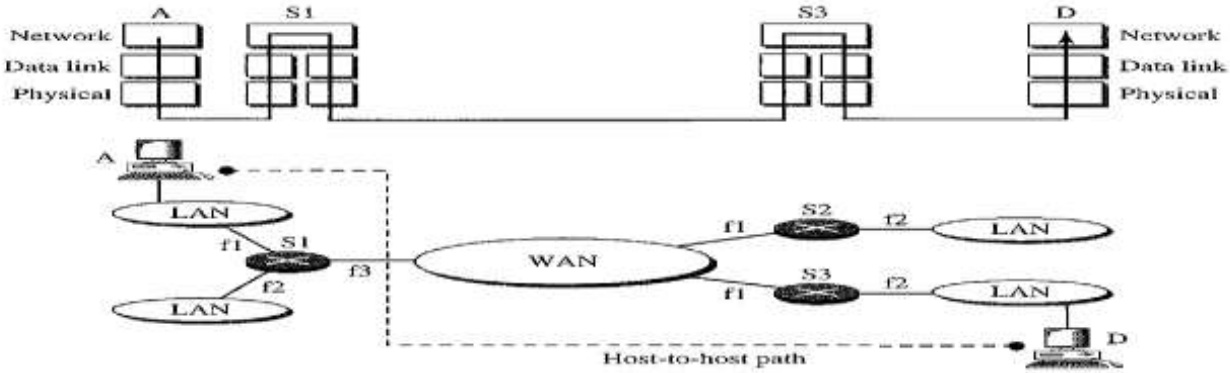
- A link local address is used in an isolated subnet.
- A Site Local address is used in an isolated site with several subnets.



INTERNETWORKING

Problems With Datalink Layer

- The frame in datalink layer does not carry any routing information.
- The frame contains the MAC address of source and the MAC address of destination.
- For a LAN or a WAN, delivery means carrying the frame through one link only.



Need for Network Layer

- To solve the problem of delivery through several links, the network was designed.
- The network layer is responsible for host-to-host delivery and for routing the packets through the routers or switches.

Network Layer at Source

- 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 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 then the packet is fragmented.

Network Layer at Router

- 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.
- After some changes made in the packet Header, the routing information is passed to the Data-link layer again.

Network Layer at Destination

- 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.

Delivery of Packets

The packet delivery will be done in two ways:

1. Connection Oriented Network
2. Connection less Network

Packet delivery in Connection-Oriented 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.
- 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 at the time of connection establishment only.
- 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.

Internet as a Datagram Network (Connectionless Network)

- The Internet at the network layer is a packet-switched network uses datagram approach.
- In connection-less service, the network layer protocol treats each packet independently, with each packet having no relationship to any other packet.
- The packets in a message mayor may not travel the same path to their destination.
- The Internet uses datagram approach to switching the packets in the network layer. It uses the universal addresses defined in the network layer to route packets from the source to the destination.
- **Note**: The reason for using datagram network in internet 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".

IPv4 Delivery Mechanism

- IPv4 delivery mechanism is used in TCP/IP protocols.
- IPv4 is an unreliable and connectionless datagram protocol.
- If reliability is important, IPv4 must be paired with a reliable protocol such as TCP.

Datagram

Packets in the IPv4 layer are called datagrams.

A datagram is a variable-length packet consisting of two parts:

1. Header
2. Data

The header is 20 to 60 bytes in length and contains information essential to routing and delivery. The header is divided into 4 sections:

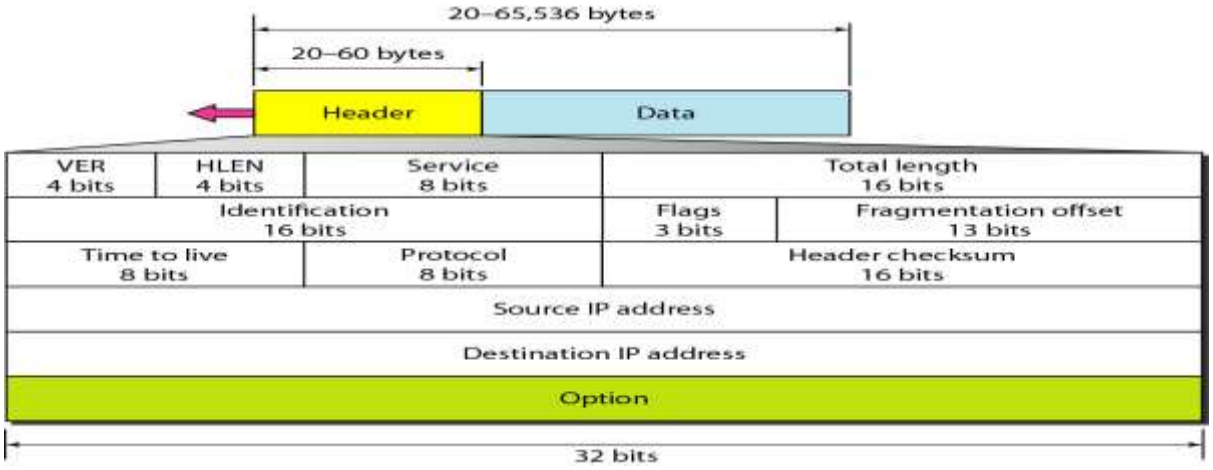
1. Version (VER)
2. Header length (HLEN)
3. Services
4. Total Length

1. Version (VER) – 4 bits

- It defines the version of the IPv4 protocol. Currently 4th version of IPv4 is using.

2. Header length (HLEN) – 4 bits

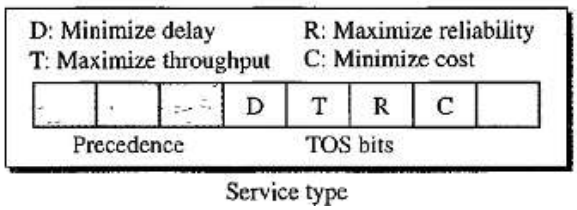
- It defines the total length of the datagram header in 4-byte words.
- The length of the header is variable between 20 and 60 bytes.
- When there are no options, the header length is 20 bytes, and the value of this field is 5 (5x 4 = 20).
- When the option field is at its maximum size, the value of this field is 15 (15 x 4 = 60).



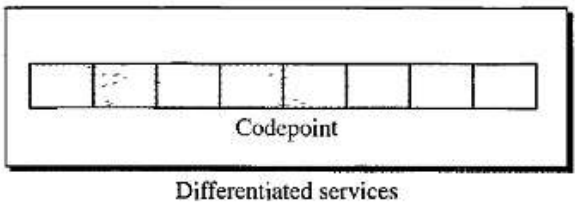
3. Services

IETF has changed the interpretation and name of this 8-bit field.

Previously this field is called **Service Type** but now the name changed to **Differentiated Services**.



Service type



Differentiated services

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.

i. Precedence

- It is a 3-bit subfield ranging from 0 to 7 (000 to 111).
- 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.

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

ii. Type of Service (TOS)

It is a 4-bit subfield with each bit having a special meaning. Out of 4 bits only one bit will have the value of 1.

TOS Bits	Description
0000	Normal (default)
0001	Minimize Cost
0010	Maximize Reliability
0100	Maximize Throughput
1000	Minimize Delay

Differentiated Services

In this interpretation, the first 6 bits make up the code-point subfield, and the last 2 bits are not used. The code-point subfield can be used in two different ways:

a. When the 3 rightmost bits are 0's, the 3 leftmost bits are interpreted the same as the precedence bits in the service type interpretation.

b. When the 3 rightmost bits are not all Os, the 6 bits define 64 services based on the priority assignment by the Internet or local authorities.

Category	**Code-point**	**Assigning Authority**	**No of service types**	**Numbers**
1	XXXXX0	Internet	32	0,2,4,6,8,……60,62
2	XXXX11	Local	16	3,7,11,15,……59,63
3	XXXX01	Temporary or Experiment	16	1,5,9,13,17...……,61

4. Total length

This is a 16-bit field that defines the total length (**header plus data**) of the IPv4 datagram in bytes. Total length of IPv4 is 65,535 (2^{16}-1).

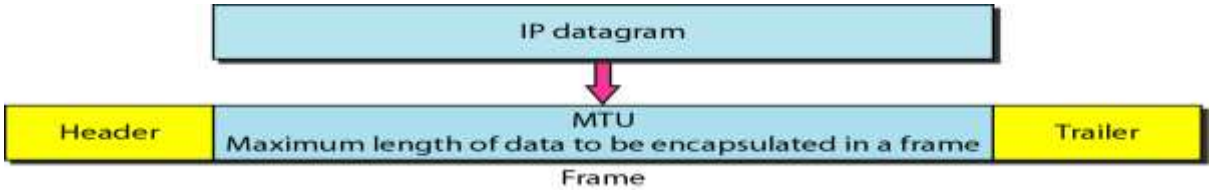
Length of data = Total length - header length

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 sent frame and received frame depends on the protocol used by the physical network through which the frame has been transmitted.
- 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)

- The value of the MTU depends on the physical network protocol.
- When a datagram is encapsulated in a frame, the total size of the datagram must be less than this maximum size.
- The maximum length of IPV4 datagram is 65,535 bytes.



- If the length of the datagram exceeds the MTU then the datagram must be fragmented to make it possible to pass through the networks.
- When a datagram is fragmented, each fragment has its own header with only few fields are changed. Remaining fields are copied by all fragments.

Protocol	MTU in Bytes
Hyper-channel	65,535
Token Ring (16 Mbps)	17,914
Ethernet	1,500

Note: When the datagram is fragmented it is obvious that "total length, flag and flag offset" fields are changed.

- The reassembly of the datagram is done only by the destination host because each fragment becomes an independent datagram and packets received not in the order.
- If all the fragments are arrived at the destination then only the destination host starts reassembly of fragments.

Fields Related to Fragmentation: Identification, Flag, Offset.

Identification (16 bits)

- It 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.
- When a datagram is fragmented, all fragments have the same identification number the same as the original datagram. All fragments having the same identification value must be assembled into one datagram.
- The identification number helps the destination in reassembling the datagram.

Flags (3 bits)

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 its value is 0, the datagram can be fragmented if necessary.

The third bit is called the **More Fragment** bit.

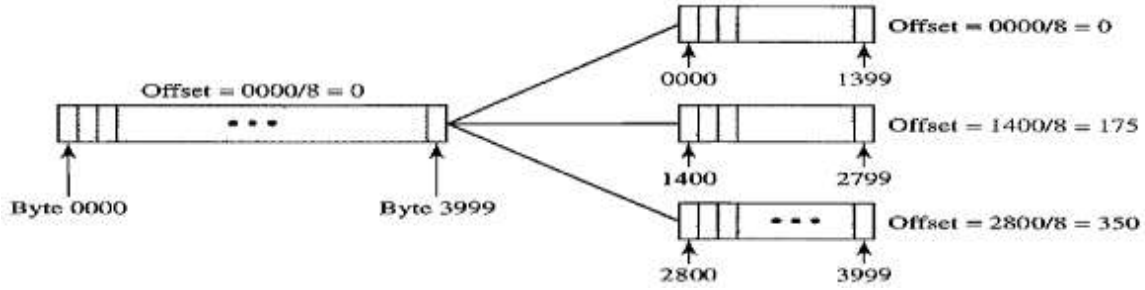
- If its value is 1, it means the datagram is not the last fragment.
- If its value is 0, it means this is the last or only fragment.

Fragmentation offset (13 bits)

It 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.

Example: Consider the below figure 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.

Fragment Number	Range	Offset Value
First	0-1399	0/8=0
Second	1400-2799	1400/8=175
Third	2800-3999	2800/8=350

Time To Live - TTL (8 bits)

A datagram has a limited lifetime in its travel through an internet.

This field can be used in two ways:

i. 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.

ii. This field is used mostly to control the maximum number of hops (routers) visited by the datagram. Each router that processes the datagram decrements this number by 1. The router discards the datagram, if **TTL=0**.
When a source host sends the datagram, it stores a number in TTL field. This value is approximately 2 times the maximum number of routes between any two hosts.

Protocol (8 bits)

- This 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.

Checksum (16 bits)

The checksum in the IPv4 packet covers only the header, not the data. There are two reasons:

i. All higher-level protocols that encapsulate data in the IPv4 datagram have a checksum field that covers the whole packet. The checksum for the IPv4 datagram does not have to check the encapsulated data.

ii. The header of the IPv4 packet changes with each visited router, but the data do not changes. So the checksum includes only the part that has changed.

Options

Options are not required for a datagram. They can be used for network testing and debugging.

Source Address (32 bits) & Destination Address (32 bits)

- These two fields define the IPv4 address of the Source and Destination respectively.
- These fields must remain unchanged during the time the IPv4 datagram travels from the source host to the destination host.

Disadvantages of IPv4

1. Despite all short-term solutions, such as subnetting, classless addressing, and NAT, address depletion is still a long-term problem in the Internet.
2. 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.
3. The Internet must accommodate encryption and authentication of data for some applications. No encryption or authentication is provided by IPv4.

IPV6 DELIVERY MECHANISM

- IPV6 is introduced to overcome the deficiencies of IPv4.
- IPv6 is also called as IPng (Internetworking Protocol next generation).
- In IPv6, the Internet protocol was extensively modified to accommodate the growth of the Internet. Packet format, Length of IP address, ICMP, IGMP, ARP, RARP, RIP routing protocol are also modified in IPv6.

Advantages of IPv6

- **Larger address space** An IPv6 address is 128 bits long whereas IPv4 is 32-bit address.
- **Better header format** IPv6 uses a new header format in which options are separated from the base header. when options are needed it is inserted between the base header and the upper-layer data.
- **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.
- **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.
- **More Security** The encryption and authentication options in IPv6 provide confidentiality and integrity of the packet.

Packet Format

In IPv6 each packet is composed of a mandatory base header followed by the payload.

The **Payload** consists of two parts:

- Optional extension headers
- 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.

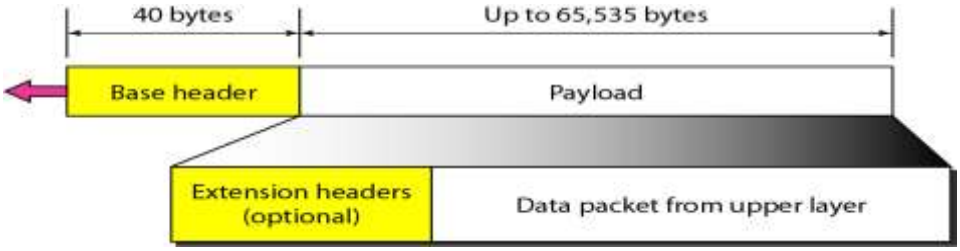
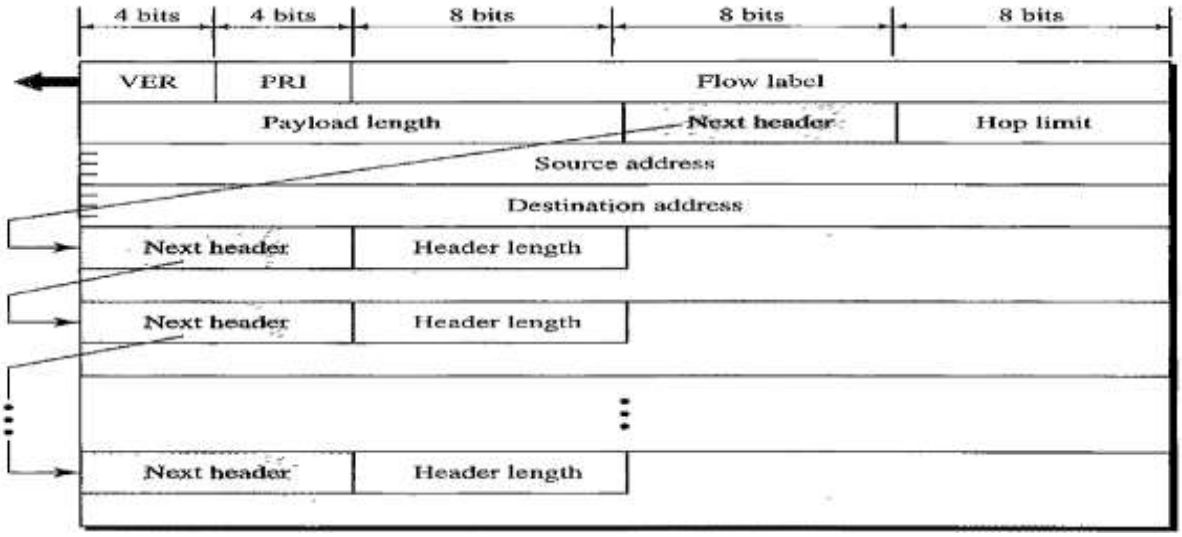


Fig: IPv6 Datagram header and payload

Base Header

Fields in IPv6 datagram are:



- **Version (4-bit)**
  This field defines the version number of the IP. For IPv6, the value is 6.
- **Priority (4-bit)**
  The priority field defines the priority of the packet with respect to traffic congestion.
- **Flow label (24-bit or 3 Byte)**
  Flow label field that is designed to provide special handling for a particular flow of data.
- **Payload length (16 bit or 2 Byte)**
  Payload length field defines the length of the IP datagram excluding the base header.
- **Next header (8-bit)**
  The next header is an 8-bit field defining the header that follows the base header in the datagram. The next header is either optional extension headers used by IP or the header of TCP or UDP encapsulated packet.
  Note: This field in IPv4 is called the *protocol.*
- **Hop limit (8 bit)**
  Hop limit field serves the same purpose as the TTL field in IPv4
- **Source address (128-bit or 16 Byte) and Destination Address (128 bit or 16 Byte)**
  The source address field is a 16-byte (128-bit) Internet address that identifies the original source of the datagram.
  The destination address field is a 16-byte (128-bit) Internet address that usually identifies the final destination of the datagram. If source routing is used, this field contains the address of the next router.

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

Priority

The priority field of the IPv6 packet defines the priority of each packet with respect to other packets from the same source.

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:

i. Congestion-Controlled
ii. Noncongestion-controlled

Congestion-Controlled Traffic

When there is congestion a source adapts itself to slowdown the traffic.

Example: TCP uses sliding window protocol can easily respond to the traffic.

Congestion-controlled data are assigned priorities from 0 to 7

Priority	Meaning	Description
0	No specific traffic	Priority 0 is assigned to a packet when the process does not define a priority.
1	Background data	defines data that are usually delivered in the background. Ex: Delivery of the news.
2	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. Ex: Email
3	Reserved	
4	Attended bulk data traffic	A protocol that transfers data while the user is waiting to receive the data is given a priority of 4 Ex: FTP and HTTP
5	Reserved	
6	Interactive traffic	Protocols that need user interaction are assigned 6. Ex: TELNET
7	Controlled traffic	Routing Protocols are given Highest Priority 7. Ex: OSPF, RIP, SNMP

Noncongestion-Controlled Traffic

- The source does not adapt itself to congestion. It is a type of traffic that expects minimum delay. Priority numbers from 8 to 15 are assigned to Noncongestion-controlled traffic.
- In this traffic Discarding of packets is not desirable and Retransmission in most cases is impossible.

Examples: Real-time audio and video.

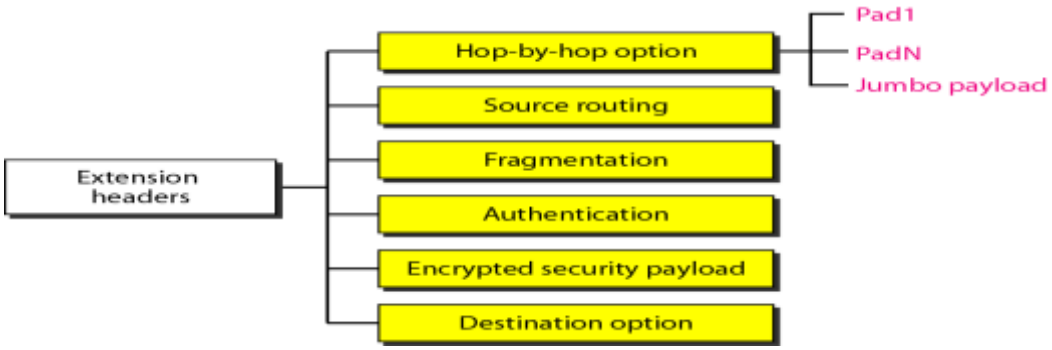
- **Priority 15:** It is given to data containing **Less Redundancy** (low-fidelity audio or video)
- **Priority 8:** It is given to data containing M**ore Redundancy** (high-fidelity audio or video)

Flow Label

- A sequence of packets, sent from a particular source to destination that needs special handling by routers is called a F**low of packets.**
- The combination of the source address and the value of the **Flow Label** uniquely define 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 etc.
- 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 a router receives a packet it consults the flow label table 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.
- This mechanism speed up the processing of a packet by a router.

Extension Headers

To give greater functionality to the IP datagram, the base header can be followed by up to six types of extension headers.



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.

Only three options have been defined: Padl, PadN, and jumbo payload.

- The Padl option is 1 byte long and is designed for 1 byte alignment purposes.
- 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

- 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.

Encrypted Security Payload (ESP)

- ESP is an extension that provides confidentiality and guards against eavesdropping.

Destination Option

- It is used when the source needs to pass information to the destination only.
- Intermediate routers are not permitted access to this information.

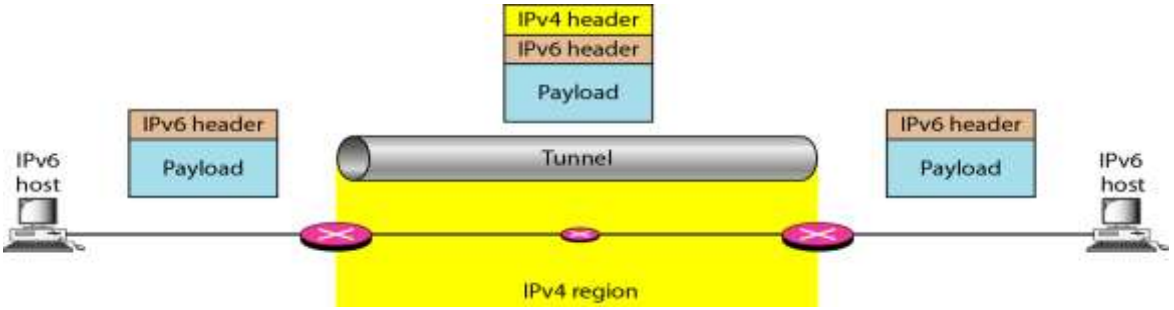
Comparison between IPv4 Options and IPv6 Extension Headers

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 and Encrypted Security Payload extension headers are new in IPv6.

TRANSITION FROM IPv4 TO IPv6

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. The IPv4 packet is carrying an IPv6 packet as data, the protocol value is set to 41.

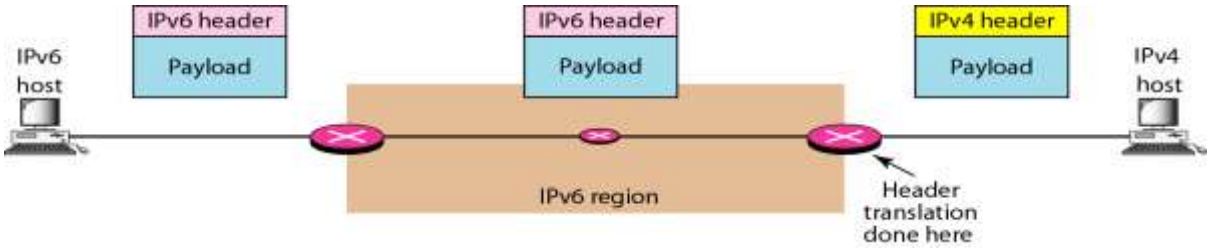


Dual Stack

- A station must run IPv4 and IPv6 simultaneously.
- 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.

Header Translation

- Header translation is necessary when the sender wants to use IPv6, but the receiver does not understand IPv6 , the receiver understand IPv4 only.
- The header of the IPv6 packet is converted to an IPv4 header.
- Header translation uses the mapped address to translate an IPv6 address to an IPv4 address.



Procedure for transforming an IPv6 packet header to an IPv4 packet header:

1. The IPv6 mapped address is changed to 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.

ADDRESS MAPPING

- IP packets use Logical Addresses (Host-to-Host).
- Frame needs Physical Addresses (Node-to-Node).
- IP packets need to be encapsulated in a frame.

The physical address and the logical address are two different identifiers.

- 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.
- 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 Physical address to corresponding Logical Address.

Mapping can be done by using two ways:

1. Static Mapping
2. Dynamic Mapping

Static 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 the IP address of another machine but not its physical address can look it up in the table.

Limitations of Static Mapping

Physical address may change in several ways:

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

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

In order to avoid this overhead and limitation Dynamic Mapping is introduced.

Dynamic Mapping

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

The address mapping protocols are:

1. Address Resolution Protocol (ARP)
2. Reverse Address Resolution Protocol (RARP)
3. Bootstrap protocol (BOOTP)
4. Dynamic Host Configuration Protocol (DHCP)

ARP (Mapping Logical to Physical address)

Need for Physical Address

- 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 in two ways:
  i. If the sender is the host then logical address is obtained from DNS.
  ii. If the sender is router then logical address is obtained from a routing table.
- 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.

In order to know the physical address of the receiver the sender uses ARP protocol.

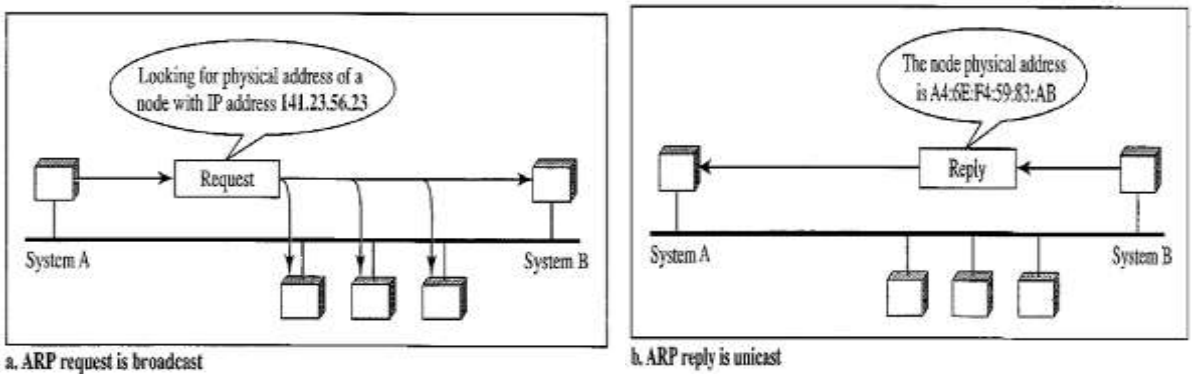
Process of ARP

- The sender knows the IP address of the target. The host or the router sends an ARP query packet.
- IP asks ARP to create an ARP request packet. The packet includes the **Physical address** and **IP addresses** of the **Sender** and the **IP address** of the **Receiver**.
- The target physical address field is filled with all 0's. Because the sender does not know the physical address of the receiver and the query is broadcast over the network.

- 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 reply packet whereas the remaining devices discard the packet.
- The reply packet contains the recipient's IP and physical addresses.
- The packet is unicast directly to the sender by using the physical address received in the query packet.
- The sender receives the reply message. It now knows the physical address of the target machine.
- The IP datagram that carries data for the target machine is now encapsulated in a frame and datagram is unicasted to the destination.

Note

* A system is normally sends several packets to same destination.
* A system that receives an ARP reply stores the mapping in the cache memory and keeps it until 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.



a. ARP request is broadcast b. ARP reply is unicast

The above figure describes:

- System A has a packet that needs to be delivered to another system B with IP address 141.23.56.23.
- System A doesnot know the physical address of 141.23.56.23.
- System A broadcast ARP request packet to ask for physical address of 141.23.56.23
- This packet is received by every system on the physical network, but only system B will respond by sending ARP reply packet that includes its physical address (A4:6E:F4:59:83:AB).
- Now system A can send all the packets it has for the destination (System B) by using the physical address it received.

Packet Format

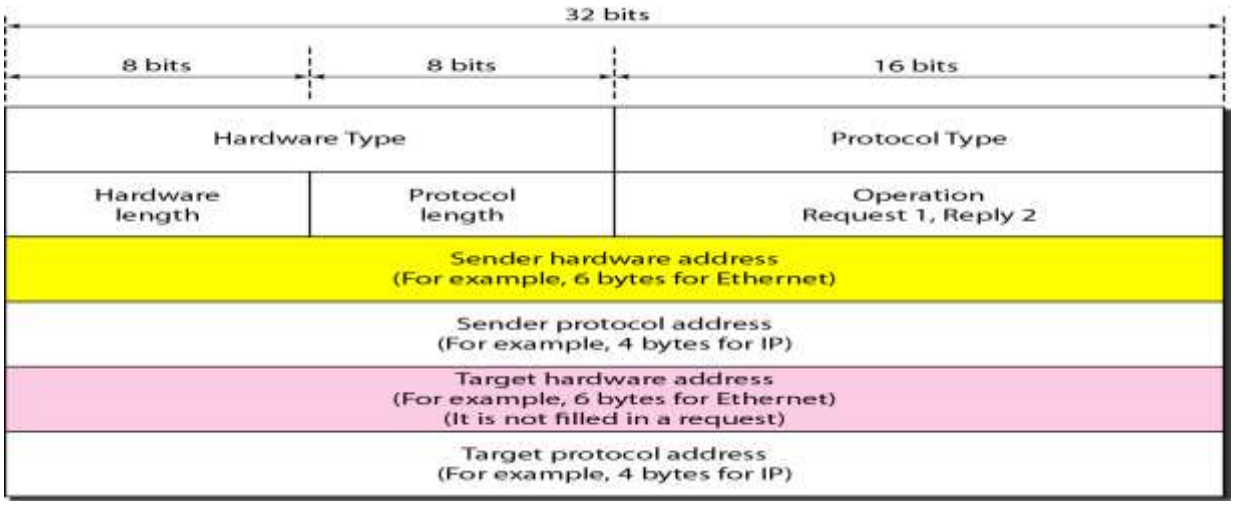
The fields are divided into fixed length fields(5) and Variable length fields(4).

Fixed length fields:

- **Hardware type (16 bit)**
  It defining the type of the network on which ARP is running.
  Each LAN has been assigned an integer based on its type. **Ex:** Ethernet is given type 1.

ARP can be used on any physical network.

- **Protocol type (16-bit)**
  This field defines the protocol. For IPv4 Protocol the value is **0800_{16}.**
- **Hardware length (8-bit)**
  This field defines the length of the physical address in bytes. Ex: for Ethernet value = 6.
- **Protocol length (8-bit)**
  This defines the length of logical address in bytes. For IPv4 protocol the value is 4.
- **Operation (16-bit)**
  This field defines the type of packet. Two packet types are defined: ARP request (1) and ARP reply (2).

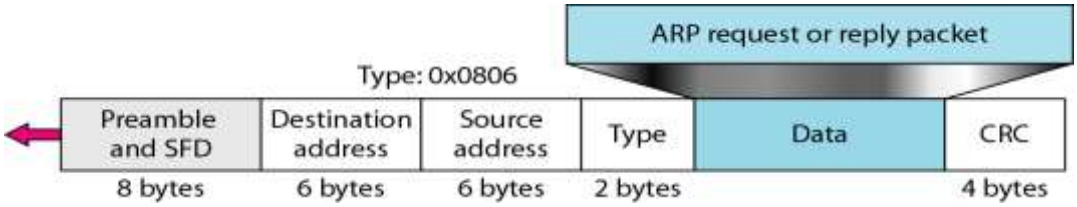


Variable length fields:

- **Sender hardware address**
  This is a variable length field defines the physical address of the sender. For example, for Ethernet this field is 6 bytes long.
- **Sender protocol address**
  It is a variable length field defines the the logical address (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 (receiver).
  For example, for Ethernet this field is 6 bytes long.
  For an ARP request message, this field is all **0's** because the sender does not know the physical address of the target.
- **Target protocol address**
  This is a variable-length field defining the logical address of the target. For the IPv4 protocol, this field is 4 bytes long.

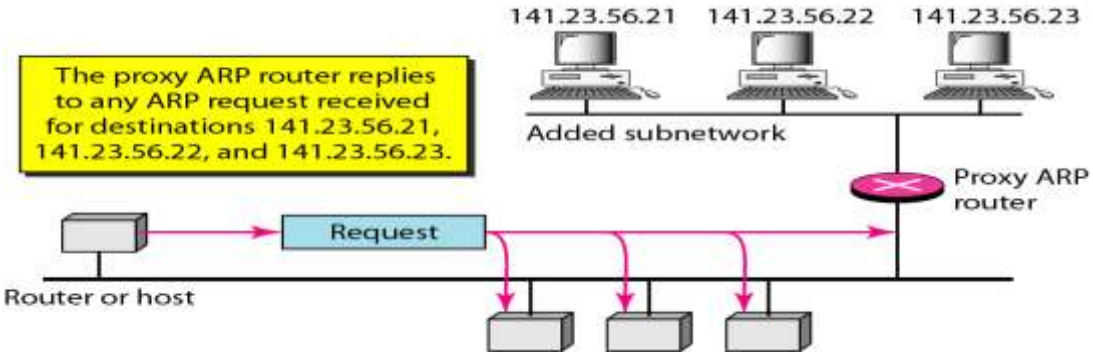
Encapsulation

An ARP packet is encapsulated directly into a data link frame (i.e. Ethernet frame). Note that the type field indicates that the data carried by the frame are an ARP packet.



ProxyARP

- 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.



In the above figure: The ARP installed on the right-hand host will answer only to anARP request with a target IP address of 141.23.56.23.

- 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. 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.
2. 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.

Note:

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.

REVERSE ADDRESS RESOLUTION PROTOCOL (RARP)

- 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.
- IP address of a machine is usually read from its configuration file stored on a disk file.
- The machine can get its physical address by reading its NIC, 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.

Note: The requesting machine must be running a RARP client program and the responding machine must be running a RARP server program.

Problems with RARP

- Broadcasting is done at the data link layer.
- The physical broadcast address, all 1's in the case of Ethernet. It does not pass the boundaries of a network. (i.e.) if an administrator has several networks or several subnets, it needs to assign a RARP server for each network or subnet.

Note: BOOTP and DHCP are implemented to solve the problems with RARP

Bootstrap Protocol (BOOTP)

- 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.
- BOOTP messages are encapsulated in a UDP packet, and the UDP packet itself is encapsulated in an IP packet.

Advantage of BOOTP over RARP

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.

Note:

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 0's as the source address and all 1's as the destination address.

Problem

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.

Solution

- There is a need for intermediary host or router.
- 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 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.

DYNAMIC HOST CONFIGURATION PROTOCOL (DHCP)

DHCP is a static and dynamic configuration protocol whereas BOOTP is a static configuration protocol only.

Why DHCP?

- 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.
- The binding or mapping between the physical address and IP addresses is static and fixed in a table until changed by the administrator. BOOTP is a static configuration protocol.

There are situations where BOOTP fails to handle:

i. What if a host moves from one physical network to another.
ii. What if a host wants a temporary IP address.

DHCP has been devised to provide **Static** and **Dynamic Address Allocation** that can be manual or automatic.

Static Address Allocation

- In this capacity DHCP acts as BOOTP.
- 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.
- 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 then DHCP provides temporary IP addresses for a limited time.

Temporary Addresses

- 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.

Advantages of DHCP over BOOTP

- 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 the administrator needs to manually enter the changes every time there is a change in a physical address or IP address.
- DHCP allows both manual and automatic configurations.
- Static addresses are created manually whereas dynamic addresses are created automatically.

INTERNET CONTROL MESSAGE PROTOCOL (ICMP)

The IP provides unreliable and connectionless datagram delivery. (i.e.) IP does not provide any Error control mechanisms and it does not provide any host management queries.
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:

1. Error-Reporting Messages
2. Query Messages

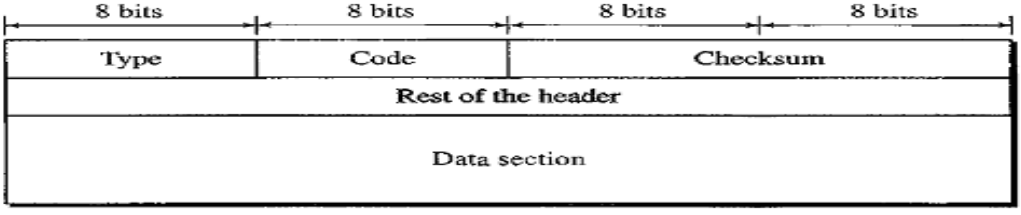
- **Error-reporting messages** report problems that a router or a host (destination) may encounter when it processes an IP packet.
- **Query messages** occur in pairs, help a host or a network manager get specific information from a router or another host.

Example: nodes can discover their neighbors, and 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.

The general format of the header is different for each message type, the first 3 fields Type, Code, Checksum are common to all. These common fields consisting of 4 bytes.

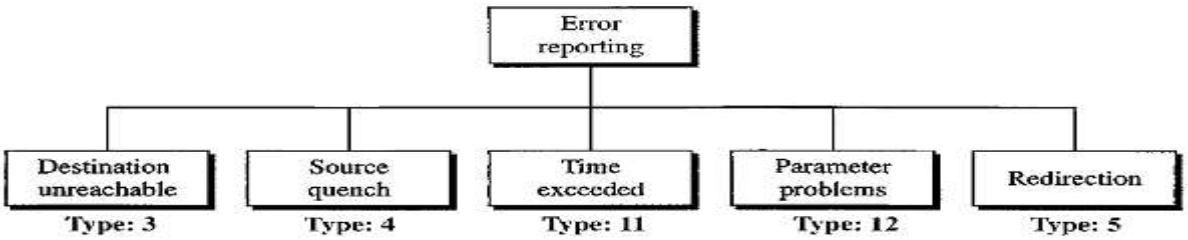


- **ICMP Type** defines the type of the message.
- **Code** field specifies the reason for the particular message type.
- **Rest of the header** is specific for each message type.
- **Checksum** is calculated over the entire message (header and data).
- **Data section** in Error messages carries information for finding the original packet that had the error. **Data Section** in Query messages the data section carries extra information based on the type of the query.

Error Reporting

- The main responsibilities of ICMP are to report errors. ICMP does not correct errors
- 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 original source of the datagram.

ICMP handles 5 types of errors:



Destination Unreachable (Type 3)

- When a router cannot route a datagram or a host cannot deliver a datagram then the datagram is discarded.
- 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 (Type 4)

- IP does not have a flow control mechanism embedded in the protocol.
- The lack of flow control can create major problems such as Congestion in routers or the destination host. When there is a congestion the router or host may discard the packets.
- When a router or host discards a datagram due to congestion, it sends a source-quench message to the sender of the datagram.

Source Quench message has two purposes.

i. It informs the source that the datagram has been discarded.
ii. It warns the source that there is congestion somewhere in the path and that the source should slow down (quench) the sending process.

Redirection (Type 5)

- Routing is dynamic. Routing table will be updated by routers. Host does not involve in the process of updation of routing tables.
- The hosts usually use static routing. Routing table has a limited number of entries.

- Host usually knows the IP address of the default router only. 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.
- To update the routing table of the host, it sends a redirection message to the host.

Time Exceeded (Type 11)

- Each datagram contains a field called Time To Live (TTL).
- When a datagram visits a router, the value of TTL field is decremented by 1.
- When the TTL value reaches 0 the router discards the datagram.
- When the datagram is discarded, a time-exceeded message must be sent by the router to the original source.

Note: 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 (Type 12)

- 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.

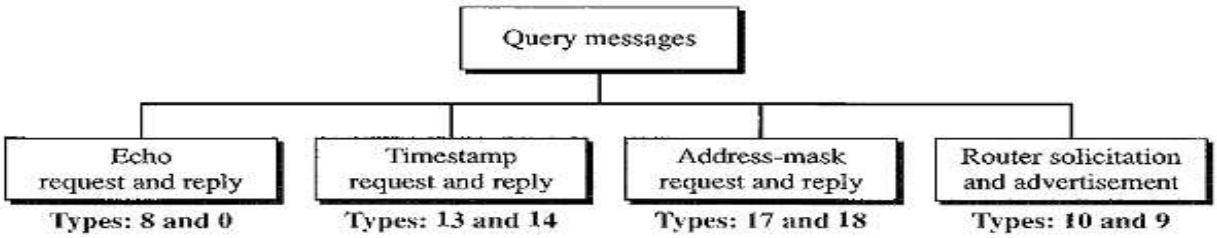
Note: There are some cases where ICMP error messages will not be generated.

i. No ICMP error message will be generated in response to a datagram carrying an ICMP error message.
ii. No ICMP error message will be generated for a fragmented datagram that is not the first fragment.
iii. No ICMP error message will be generated for a datagram having a **Multicast Address**.
iv. 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.

Query Messages

- 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.

Here no bytes of the original IP are included in the message.



Echo Request and Echo Reply

- The echo-request and reply messages can be used to determine if there is communication at the IP level.
- These are used for diagnostic purpose, network managers and users utilize this pair of messages to identify network problems.

- 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.

Example: **ping** command.

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.
- To obtain its mask, a host sends an address-mask-request message to a router on the LAN.
- **Ex:** A host IP address is 159.31.17.24, but it doesn't know its corresponding mask **/24**.
- 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.

Router Solicitation and Advertisement

- The router-solicitation and router-advertisement messages can help whether the router is functioning or not.
- A host can broadcast (or) multicast a router-solicitation message.
- Routers that receive the solicitation message broadcast their routing information using the router-advertisement message.
- In router advertisement message it announces its own presence and all routers on the network which it is aware of.

Note: A router can also periodically send router-advertisement messages even if no host has solicited.

Debugging Tools

There are several tools that can be used in the Internet for debugging.

There are two tools that are used for ICMP debugging: **ping** and **traceroute.**

Ping

- Ping program to find if a host is alive and responding.
- 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.
- *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: ping program to test the server fhda.edu. The result is shown below:

```
$ ping thda.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
--- fhda.edu ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time xxxxx ms
rtt min/avg/max = 1.911/1.955/2.04 ms.
```

- **ttl=62**: means that the packet cannot travel more than 62 hops.
- *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.
- **Note**: In each probe *ping* defines the number of bytes as 64. This is the total number of bytes in the ICMP packet (56 + 8).

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.
- 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

INTERNET GROUP MANAGEMENT PROTOCOL (IGMP)

IGMP is a companion to IP protocol. IGMP is not a multicasting routing protocol. It is a protocol that manages Group Membership.

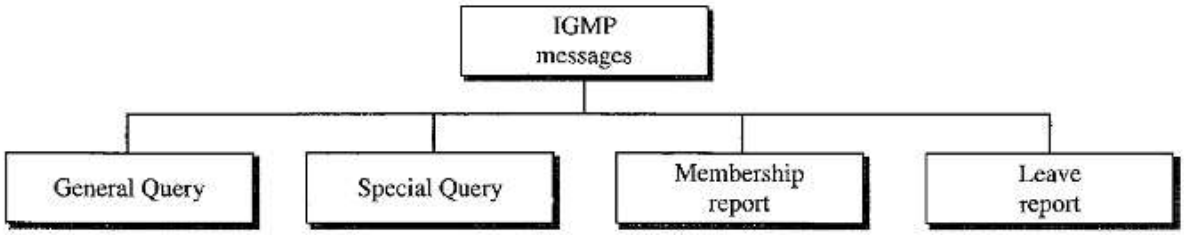
In any network, there are one or more multicast routers that distribute multicast packets to hosts or other routers. The IGMP protocol gives the multicast routers information about the membership status of hosts (routers) connected to the network.

Need for IGMP

- 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.

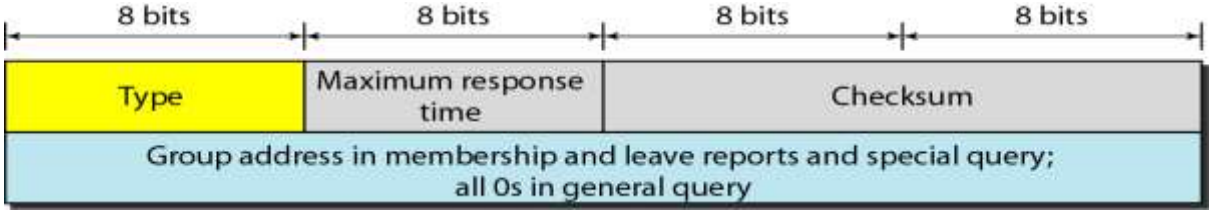
IGMP Messages and Message Format

IGMP has 3 types of messages: Query, Membership report, Leave report.
Query messages can be divided into two types: General and Special



The message format of IGMP contains four fields:

1. Type
2. Maximum Response Time
3. Checksum
4. Group Address



- **Type**
  This 8-bit field defines the type of message.

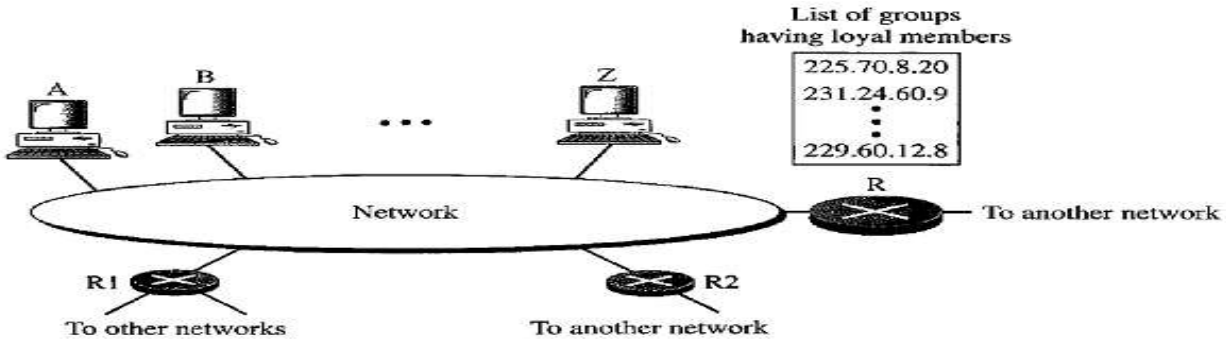
Type	Value
General or Special Query	0x11 or 00010001
Membership Report	0x16 or 00010110
Leave Report	0x17 or 00010111

- **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.
  Example: Value=100 it means 10 sec.
  The value is nonzero in the query message.
  For Membership and Leave report the value is 0.
- **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.
- For each group, there is one router that has the duty of distributing the multicast packets destined for that group.
- If there are three multicast routers connected to a network, their lists of groupid's are mutually exclusive.

Example: Consider the below figure the routers R, R1, R2 are connected to the network but only router R distributes packets with the multicast address of 225.70.8.20.
R1 and R2 are the multicast routers of other groups.



Group Membership or Group Interest

- A host or multicast router can have membership in a group.
- When a host has membership, one of the application program processes of host receives multicast packets from some group.
- When a router has membership, a network connected to one of the other interfaces of the router receives these multicast packets.
- In both of the cases, the host and the router keep a list of Group-id's and relay their interest to the distributing router.
- Membership in a group is also called as an **Interest in the Group.**

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 and they receive responses and update their lists.

There are 4 operations performed in IGMP:

1. Joining a group
2. Leaving a group
3. Monitoring Membership
4. Delayed Response

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 and it already receives multicast packets for this group.

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

Leaving a Group

- When a host sees that no process is interested in a specific group, host sends a leave report.
- 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.
- 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.
- The router sends a special query message and inserts the groupid or multicast address related to the group.
- The router waits for a specified time for any host or router to respond.
- During this time, if no interest (membership report) is received then the router assumes that there are no loyal members in the network and purges the group from its list.

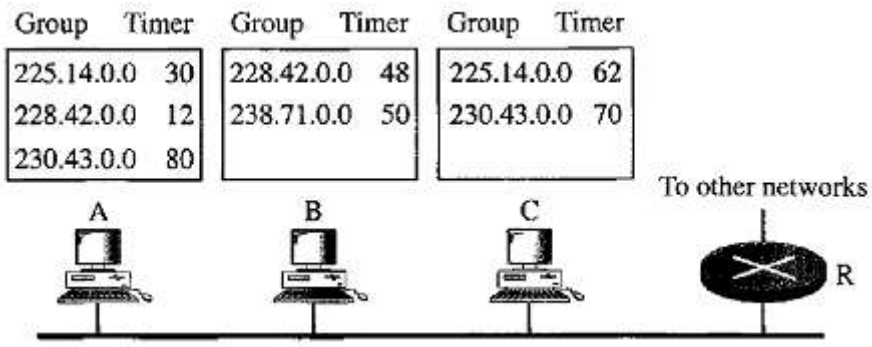
Monitoring Membership

- 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 (Query Router) periodically sends a general query message.
- The group address field of the message is set to 0.0.0.0. (i.e.) the query for membership continuation is sent to all groups in which a host is involved.
- The router expects an answer for each group in its group list.
- The query message has a maximum response time of 10 s.
- When a host or router receives the general query message, it responds with a membership report if it is interested in a group.

Delayed Response

- IGMP uses Delayed response strategy to prevent the traffic.
- If there is a common interest (i.e) two hosts are interested in the same group only one response is sent for that group to prevent unnecessary traffic. This is called a delayed response.
- 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 l0 sec. The expiration time can be in steps of 1sec or less.
- A timer is set for each group in the list.
- 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 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: 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.

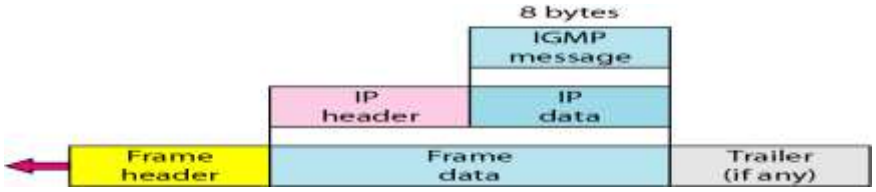


The events occur in this sequence:

- **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.
- **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.
- **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.
- **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.

Encapsulation

The IGMP message is encapsulated in an IP datagram, which is itself encapsulated in a frame. When the message is encapsulated in the IP datagram, the value of TTL must be 1.



Netstat Utility

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

We use **netstat** with three options: **-n, -r**, and **-a**.

-n : gives the numeric versions of IP addresses

-r: gives the routing table

-a: gives all addresses (unicast and multicast).

$ 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	10
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

Gateway defines the router. Iface defines the interface.

Network Layer: Delivery, Forwarding, and Routing

DELIVERY

The network layer supervises the handling of the packets by the underlying physical networks is called Delivery of packet.

Types of Delivery

The delivery of a packet to its final destination is accomplished by using two different methods of delivery.

1. Direct Delivery Method
2. Indirect Delivery Method

Direct Delivery

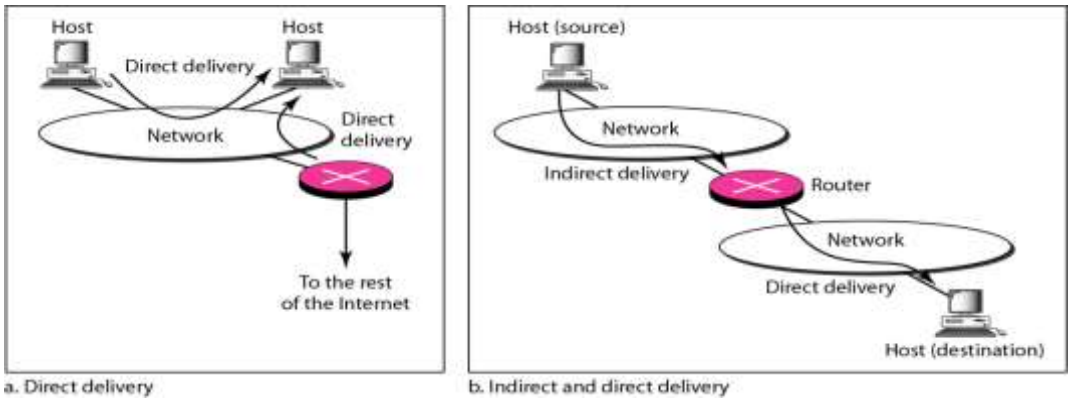
Direct delivery occurs when the source and destination of the packet are located on the same physical network or when the delivery is between the last router and the destination host.

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 router that is connected to the same physical network as its final destination.



a. Direct delivery

b. Indirect and direct delivery

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 Routing table to find the route to the final destination.

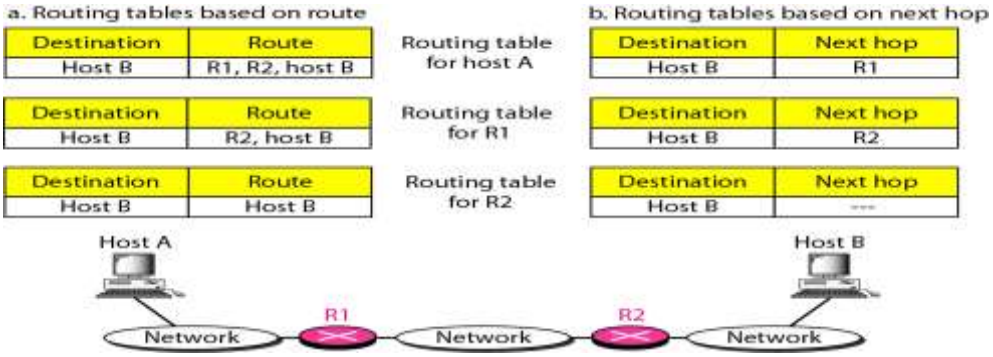
As the number of entries in the routing table increases table lookup process will take a lot of time to find the route.

To manage the size of the routing table several techniques are introduced such as:

1. Next-Hop Method Versus Route Method
2. Network-Specific Method Versus Host-Specific Method
3. Default Method

Next-Hop Method Versus Route Method

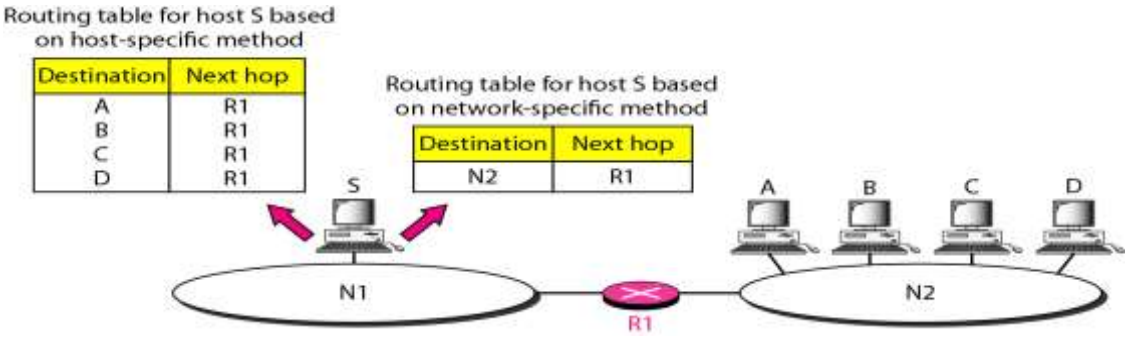
- Routing method the routing table holds information about the complete route.
- Next-Hop method is used to reduce the contents of a routing table. The routing table holds only the address of the next hop. The entries of a routing table must be consistent with one another.



Network-Specific Method Versus Host-Specific Method

- Host-specific method have an entry for every destination host connected to the same physical network.
- Network-specific method used to reduce the routing table and simplifies the searching process by taking only one entry that defines the address of the destination network itself. (i.e.) we treat all hosts connected to the same network as one single entity.

Ex: If 1000 hosts are attached to the same network, only one entry exists in the routing table instead of 1000.

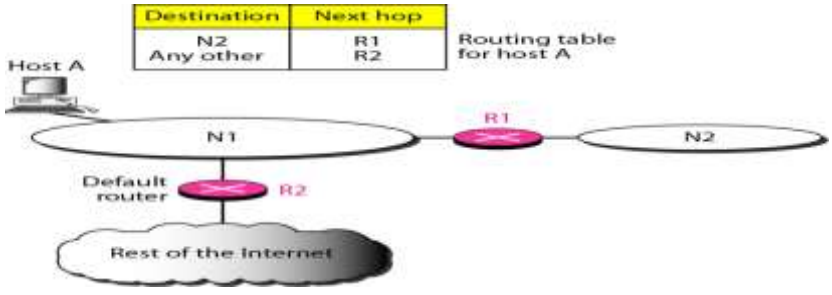


Default Method

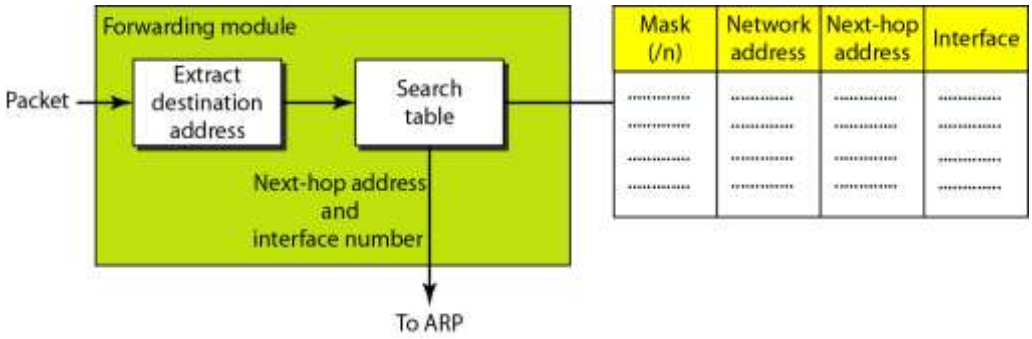
This method is also simplifies the routing table. Consider the below figure:

- Host A is connected to a network N1 with two routers R1 and R2.
- Router Rl routes the packets to hosts connected to network N2.
- For the rest of the Internet, router R2 is used, instead of listing all networks in the entire Internet, host A can just have one entry called the default.

Note: Default is normally defined as network address 0.0.0.0)



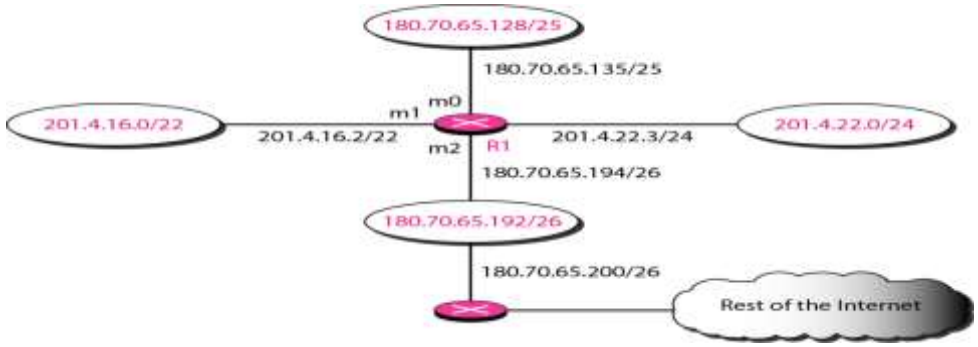
Forwarding Process



Consider the above figure that consists of Forwarding module and Routing table:
The routing table consists of :

- **Network Address** The table needs to be searched based on the first address in the block called Network address. The destination address in the packet gives no clue about the network address. Hence we need to include Subnet Mask in the table.
- **Subnet Mask (/n)** The address used in above figure is Classless address, the routing table have one row of information for each block involved.

In addition to these two columns we need **Next-Hop Address** and **Interface Columns**.
Consider the below figure and the corresponding Routing Table:



Routing table for Router R1 for above figure:

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 1: Show the forwarding process if a packet arrives at R1 in the above figure with the destination address 201.4.22.35.
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 2: Show the forwarding process if a packet arrives at R1 in the above figure with the destination address 18.24.32.78.

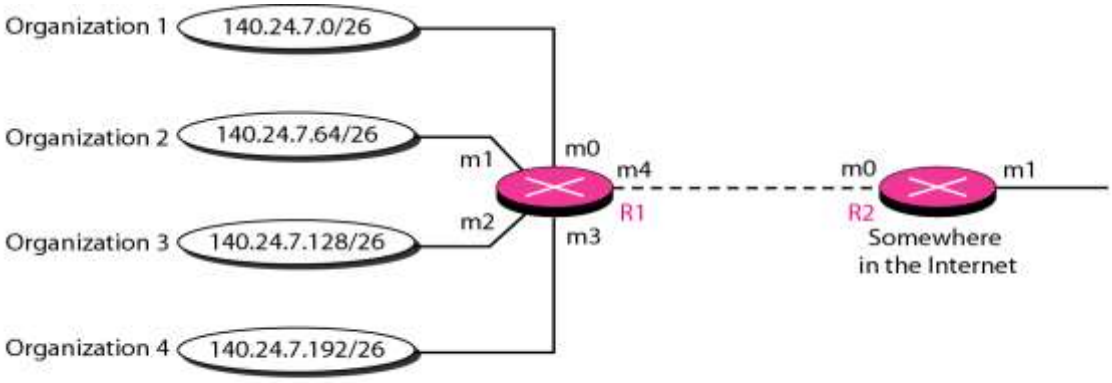
- After applying all the masks to the destination address one by one, there is no matching destination address is found.
- When it reaches the end of the table, the module gives the next-hop address 180.70.65.200 and interface number m2 to ARP.
- This is probably an outgoing package that needs to be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

Address Aggregation

Address Aggregation is the process of aggregating set of smaller blocks into one larger block.

Need for Address aggregation:

- Classless addressing is use to divide the whole address space into manageable blocks, hence when we use classless addressing the number routing table entries will increase.
- The increased size of the table results in an increase in the amount of time needed to search the table. This will affect the efficiency.
- To avoid this problem Address Aggregation is needed.



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

Consider the figure:

- Router Rl is connected to networks of four organizations that each use 64 addresses. Router Rl has a longer routing table because each packet must be correctly routed to the appropriate organization.
- Router R2 is far from Rl and Router R2 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.

Longest Mask Matching

Classless addressing uses the principle of Longest Mask Matching that states that the routing table is sorted from the longest mask to the shortest mask.

Example: If there are three masks **/27**, **/26**, and **/24**, the mask **/27** must be the first entry and **/24** must be last.

Hierarchical Routing

To solve the problem of gigantic routing tables, the hierarchical routing is introduced.

Internet today has a sense of hierarchy.

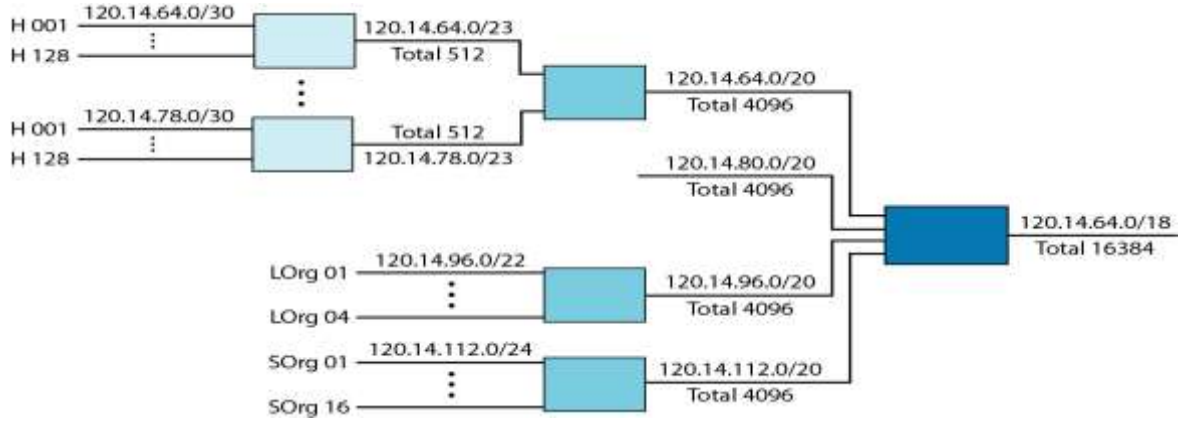
- Internet is divided into international and national ISPs.
- National ISPs are divided into Regional ISPs
- Regional ISPs are divided into Local ISPs.
- A local ISP can be assigned a single, but large block of addresses with a certain prefix length.
- The local ISP can divide this block into smaller blocks of different sizes and can assign these to individual users and organizations.

Process of reducing the size of the routing table by using Hierarchical routing:

- 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.
- Inside the local ISP, the router must recognize the sub-blocks 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 sub-block into smaller sub-blocks.

Note: In Classless addressing the level of hierarchy is unlimited.

Example: Consider the below figure:



- A regional ISP is granted 16,384 addresses starting from 120.14.64.0 with subnet mask /18.
- The regional ISP has divided this block into four sub-blocks. Each sub-block contains 4096 addresses.
- Three subblocks are assigned to 3-local ISPs and one subblock is reserved for future use.

Note: The original block with subnet mask /18 is divided into 4 blocks; hence each sub-block contains the subnet mask /20.

The **First Local ISP** has divided its assigned sub-block 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.
- 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 **4** addresses ($\mathbf{2^{32-30}} = 4$).

The **Second Local ISP** has divided its block into 4 blocks and has assigned the addresses to four large organizations (LOrg01 to LOrg04). 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.

Geographical Routing

- Geographical Routing is an extension of Hierarchical Routing.
- Geographical routing reduces the routing table based on the geographical locations.

Example: It assigns 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.

ROUTING TABLE

A host or a router has a routing table with an entry for each destination or a combination of destinations to route IP packets. A routing table can be of two types:

1. Static Routing
2. Dynamic Routing

Static Routing Table

- A static routing table contains information will be entered manually by the administrator.
- 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.

Note: A static routing table can be used in a small internet that does not change very often.

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 then the dynamic routing protocols update all the tables in the routers automatically.
- The routers in a big internet need to be updated dynamically for efficient delivery of the IP packets.

Format and Fields of Dynamic Routing Table

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

Mask

- This field defines the mask applied for the entry. Example: /26, /18,/14.

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 5 flags. Flags are **On/Off** switches that signify either presence or absence.

i. U (up)
   - 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.

ii. G (gateway)
   - 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).

iii. H (host-specific)
   - 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.

iv. D (added by redirection)
   - D flag indicates that routing information for this destination has been added to the host routing table by a redirection message from ICMP.

v. M (modified by redirection)
   - M flag indicates that the routing information for this destination has been modified by a redirection message from ICMP.

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

To find the routing information and the contents of a routing table there are two utilities are present. They are **netstat** and **ifconfig.**

netstat

- netstat is a utility used in Linux or Unix to find the contents of a routing table for a host or router.
- The host may be a personal computer or server etc.

$ **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

The above command shows the list of the contents of a default server (i.e. host). It used two options r and n.

r - indicates the routing table

n - indicates the numeric address

- The destination column here defines the network address.
- The term *gateway* used by UNIX is synonymous with *router*. This column 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 loop-back** interface indicating that the host accepts packets with loopback address 127.0.0.0.

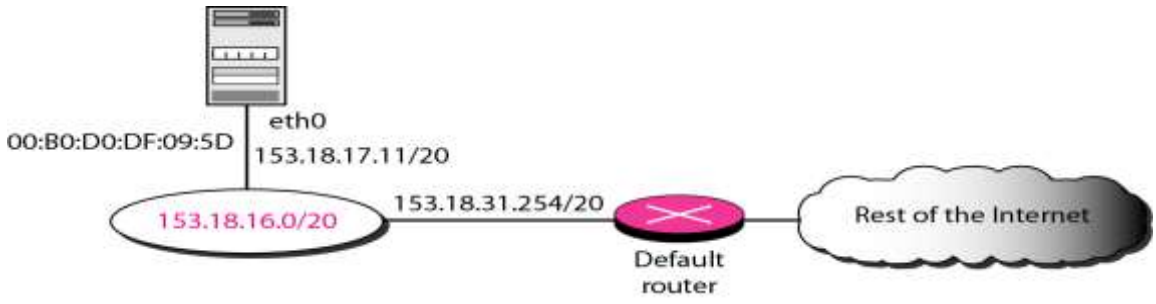
ifconfig

ifconfig command used to give the information about IP address and physical address of the server 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



ROUTING PROTOCOLS

- 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
- Routing Protocols allows routers to share whatever they know about the internet or their neighborhood.
- The sharing of information allows a router in one location (Hyderabad) to know about the failure of a network in other location (Bangalore).
- The routing protocols also include procedures for combining information received from other routers.

Optimization

- A router is usually attached to several networks.
- A router receives a packet from a network and passes it to another network.
- When a router receives a packet it will find an optimum pathway among the available paths to pass the packet to through different networks.

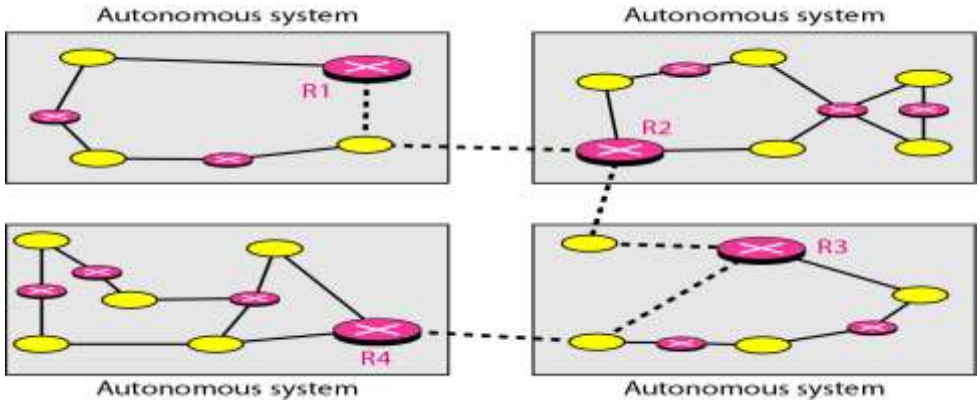
There are several protocols are defined for optimization:

- RIP (Routing Information Protocol)
- OSPF (Open Shortest Path First)
- BGP (Border Gateway Protocol)

Intra and Inter domain Routing

An internet is divided into **Autonomous Systems** because internet is so large that one routing protocol cannot handle the task of updating the routing tables of all routers.

- **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 **Intra-Domain Routing.**
- Routing between autonomous systems is referred to as **Inter-Domain Routing**.

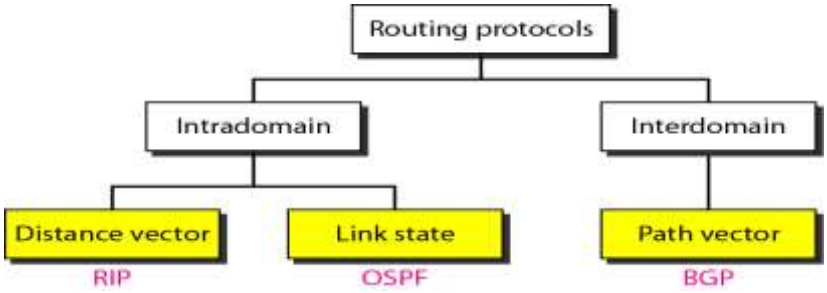


There are two Intra-domain routing protocols:

- Distance Vector Routing (RIP is an implementation of DVR)
- Link State Routing (OSPF is an implementation of LSR)

There is one Inter-Domain routing protocol:

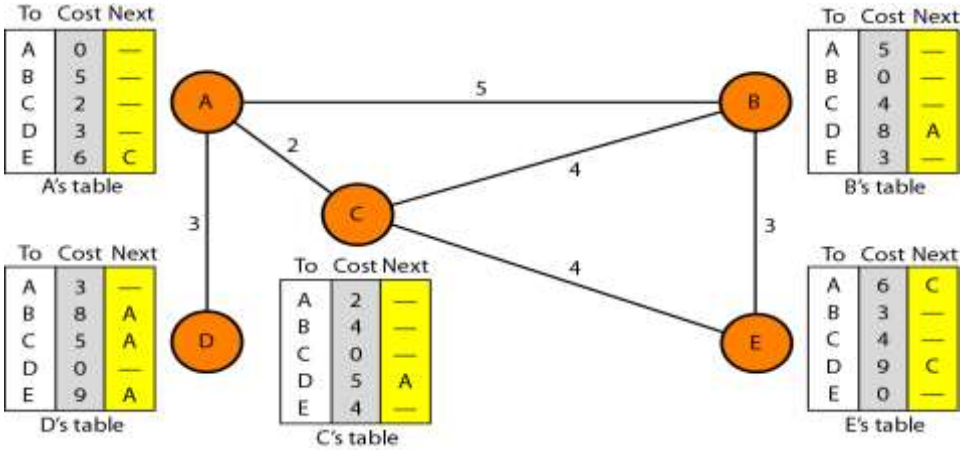
- Path Vector Routing (BGP is an implementation of PVR)



DISTANCE VECTOR ROUTING

In distance vector routing, the least-cost route between any two nodes is the route with minimum distance. The term **Vector** means a **Table**.

- In this protocol each node maintains a table of minimum distances to every node.
- The table at each node also guides the packets to the desired node by showing the next hop in the route.



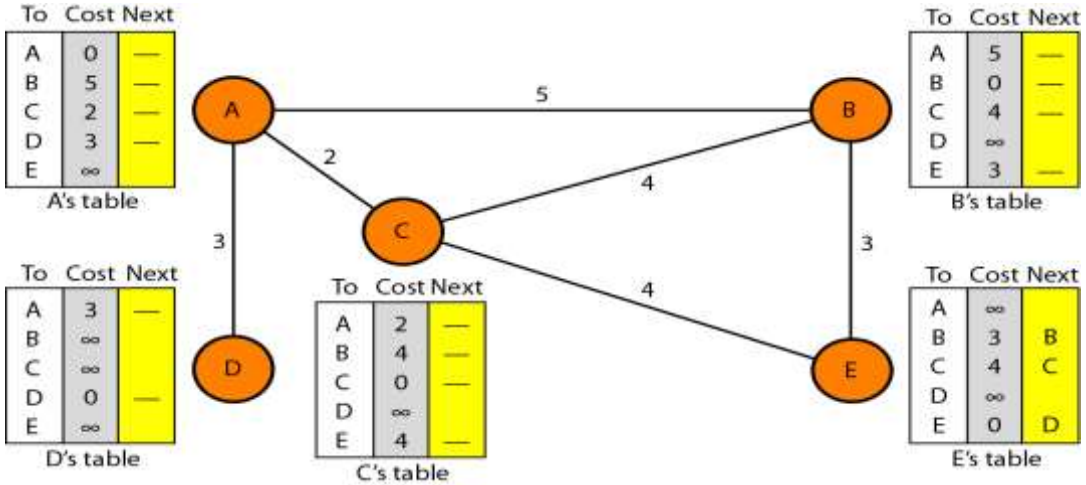
The table for node A shows how we can reach to any node from node A.

Ex: From node A the least cost to reach node E is 6. The route passes through C.

Initialization

- The above table is the final step of Distance vector routing each node knows about each and every node in the network.
- But at the initial stage each node can only know the distance between itself and its immediate neighbors. Neighbors are the nodes which are directly connected to the node.

The below figure shows the initialization of the table:



- Each node will take the immediate neighbor distance into the table.
- The distance for any entry that is not a neighbor is marked as infinite.
- Infinite means Unreachable.

Sharing

The idea behind Distance Vector Routing is the **sharing** of information between **Neighbors**.

By observing the above figure:

1. Node A does not know about node E but node C knows how to reach node E.
   If Node C shares its routing table with node A then node A can also know how to reach E.
2. Node C does not know how to reach Node D but Node A knows how to reach Node D.
   If Node A shares its routing table with node C then node C can also know how to reach node A.
   (i.e.) Immediate Neighbor's nodes A and C can improve their routing tables if they share each other's routing tables.

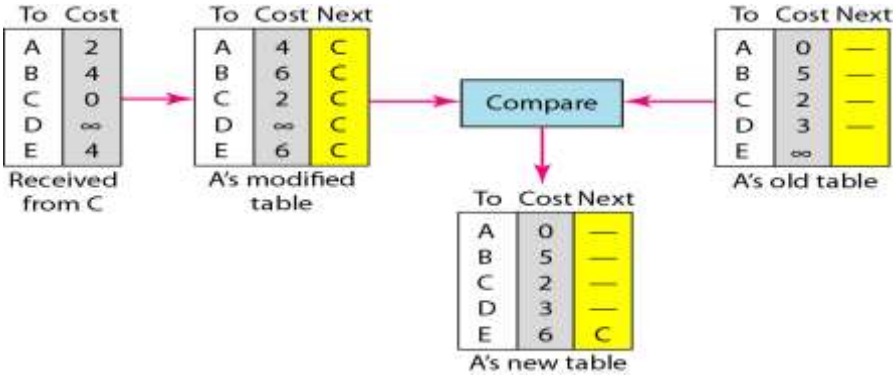
Problem: How many columns of the table must be shared with each neighbor?

- A node is not aware of a neighbor's table.
- A node can send only the first two columns of its table to any neighbor.
- Because the third column of a table next hop is not useful for the neighbor.
- When the neighbor receives a table, third column needs to be replaced with the sender's name.

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.
   Example:
   - If node C claims that its distance to a destination is **x** km
   - The distance between A and C is **y** km then,
   - The distance between A and that destination via C is **(x + y)** km.
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 anymore. The new route has a distance of infinity.



Note:

1. 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, so the distance will be 4 (A→C=2 and C→A=2).
2. The only benefit from this updating of node A is that A now knows how to reach E with cost=6 via C.

Each node can update its table by using the tables received from other nodes. 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 remain the same.

When to Share

When does a node send its partial routing table (only two columns) to all its immediate neighbors?

1. **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.
2. **Triggered Update** A node sends its two-column routing table to its neighbors anytime there is a change in its routing table.

The change can result from the following:

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

<u>Count to Infinity Problems</u>

Two-Node Loop Instability

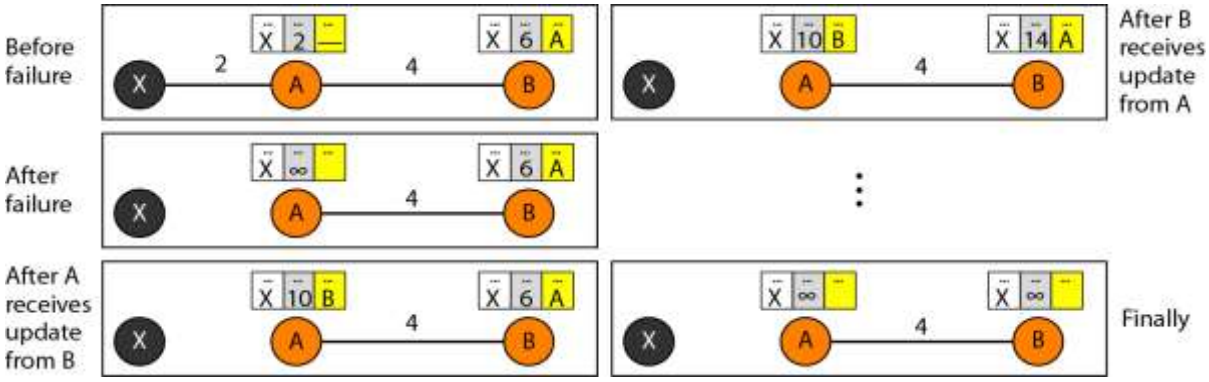
A problem with distance vector routing is instability, which means that a network using this protocol can become unstable.

Consider the above figure:

It shows a system with 3 nodes: X, A and B.

- 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 there will be no problem because B can identify by looking at the table value ∞**.**



Problem is: 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 and 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.

During this counting to infinity 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.
- 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.

Solution for Two-node loop instability or Two Node loop count to Infinity problem:

i. Defining Infinity
ii. Split Horizon
iii. Split Horizon and Poison Reverse.

Defining Infinity

- Here we define Infinity to smaller number.
- Most implementations of the distance vector protocol define the distance between each node to be 1 and define 16 as infinity. This means that the size of the network in each direction cannot exceed 15 hops.
- If we give 16 as infinity, Distance vector routing cannot be used for large systems.

Split Horizon

- In this strategy, instead of flooding the table through each interface, each node sends only part of its table through each interface.
- According to its table, node B thinks that the optimum route to reach X is via A. B 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 this case 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.

Drawback of split horizon:

- 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 case 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.

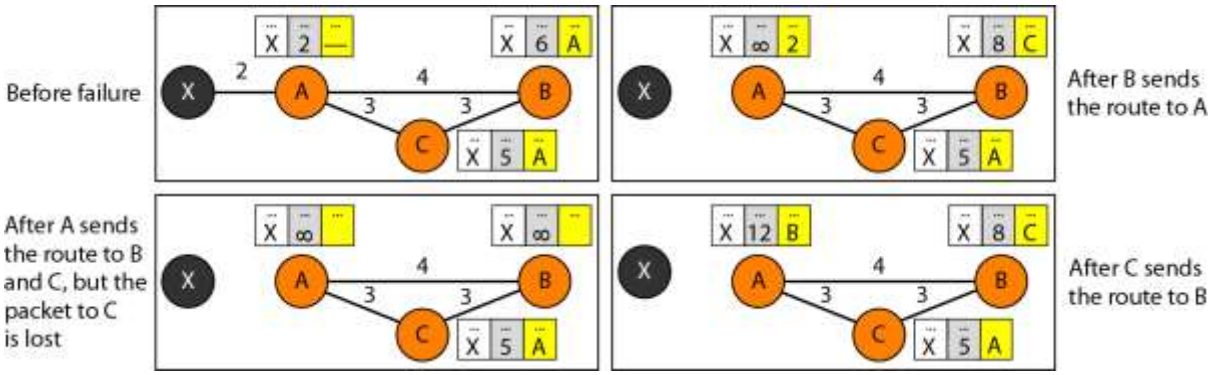
Split Horizon and Poison Reverse

- To overcome the drawback of Split Horizon Strategy we need to combine the Split horizon with 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 or Three node count to infinity problem

Split horizon strategy combined with poison reverse is sufficient for avoiding Two node instability problem but if the instability is between three nodes, stability cannot be guaranteed.

Consider the below figure:



- 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 its routing table to B, which includes the route to X.
- Node B is fooled here. It receives information on the route to X from C and according to the algorithm it updates its table showing the route to X via C with a cost of 8.
- This information has come from C, not from A, so after a while node B may advertise this route to A.

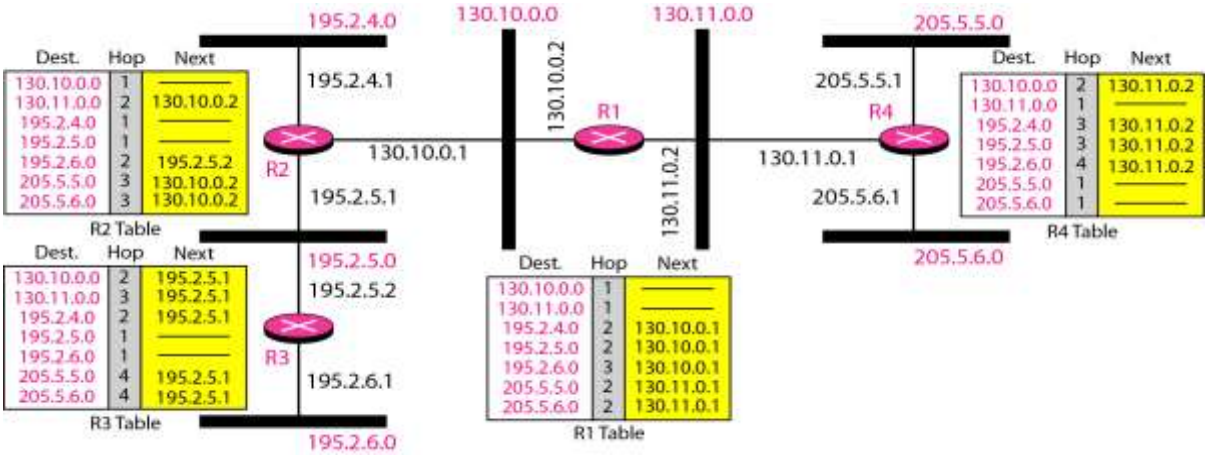
- Now A is fooled and updates its table to show that A can reach X via B with a cost of 12 and the loop continues.
- Now A advertises the route to X to C with increased cost but not to B.
- Node C then advertises the route to B with an increased cost. Node B does the same to A and so on.
- The loop stops when the cost in each node reaches infinity.

RIP (Routing Information Protocol)

The Routing Information Protocol (RIP) is an intra-domain 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 have routing tables.
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 **Hop-Count**. Hop-Count is the distance defined as the number of links (networks) to reach the destination.
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.



The above figure shows an autonomous system with seven networks and four routers. The table of each router has shown.

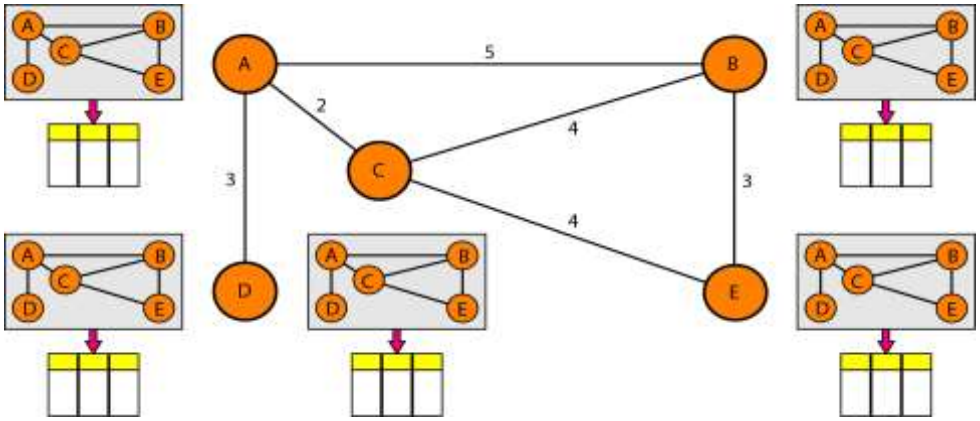
- The R1 routing table has seven entries to show how to reach each network in the autonomous system.
- Router Rl is directly connected to networks 130.10.0.0 and 130.11.0.0, which means that there are no next-hop entries for these two networks.
- To send a packet to one of the three networks at the far left, router Rl 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 Rl needs to send the packet to the interface of router R4 with IP address 130.11.0.1.

Link State Routing (LSR)

In Link State Routing, Each node in the domain has the entire topology of the domain –

- List of nodes and links
- How they are connected including the type
- Cost (metric)
- Condition of the links (up or down)

The node can use Dijkstra's algorithm to build a routing table.



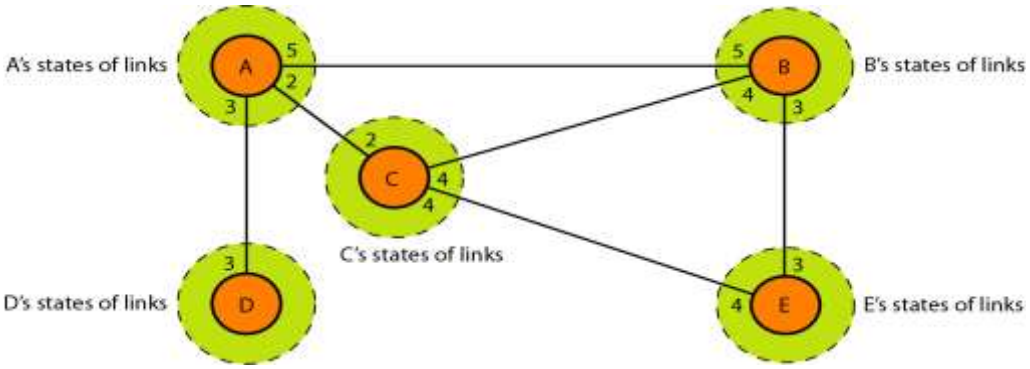
The above figure shows a Simple domain with Five Nodes.

- 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.
- 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 the topology must be updated for each node.

For Example: a link is down then each and every node in the domain should update this change.

How can a common topology be dynamic and stored in each node?

In LSR each node in the domain has the partial knowledge about the state of its links. The state means its type, condition, and cost.



Consider the above figure that shows List of nodes and their partial knowledge.

Node A knows that it is connected

- To Node B with metric 5
- To Node C with metric 2
- 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
- To Node E with metric 4

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 (Distribution) of LSPs to every other router called **Flooding**. The flooding can be done 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 such as the node identity, the list of links, a sequence number, and age etc.

- Node identity and the List of links are needed to make the topology.
- Sequence number facilitates flooding and distinguishes new LSPs from old ones.
- Age prevents old LSP's from remaining LSP's 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:**
   - 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 hours based on the implementation.
   - A longer period ensures that flooding does not create too much traffic on the network.

Note: As a matter of fact, there is no actual need for this type of LSP dissemination.

Flooding of LSP's

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.

Flooding will be done 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.
   Each and every LSP will be given a Sequence number at the time of their creation.
   Comparison of sequence numbers determines which LSP is older and which LSP is latest.
   If the newly arrived LSP is older than the one it already has, then the node discards the LSP.

If LSP arrived is newer, the node does the following:

- It discards the old LSP and keeps the new one.
- 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

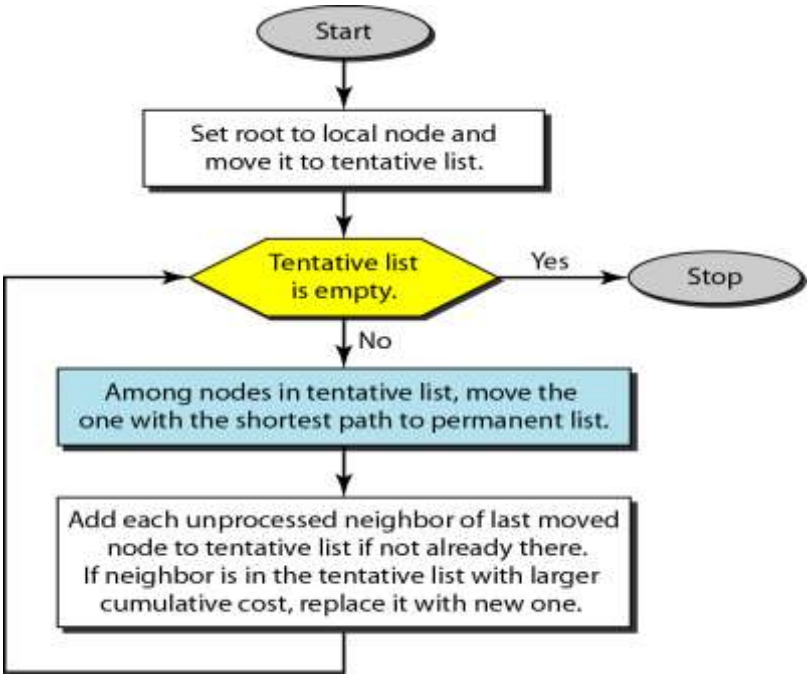
- After receiving all LSPs, each node will have a copy of the whole topology.
- The topology is not sufficient to find the shortest path to every other node; a shortest path tree is needed.
- A tree is a graph of nodes and links, where one node is called Root.
- A shortest path tree is a tree in which the path between the root and every other node is the shortest.
- The Dijkstra's algorithm creates a shortest path tree from a graph.

The algorithm divides the nodes into two sets:

1. Tentative nodes
2. Permanent nodes

Dijkstra's algorithm finds the neighbors of a current node, makes them tentative, examines them, and if they pass the criteria, makes them permanent.

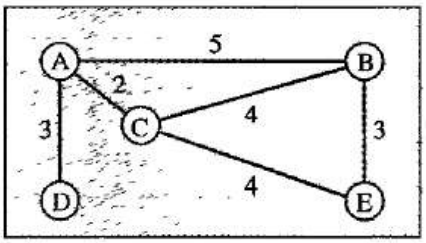
Flow Chart of Dijksta's Algorithm



Example:

Consider the below graph with five nodes: A,B,C,D,E.

- Apply the Dijkstra's algorithm to node A.
- 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.



Topology

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

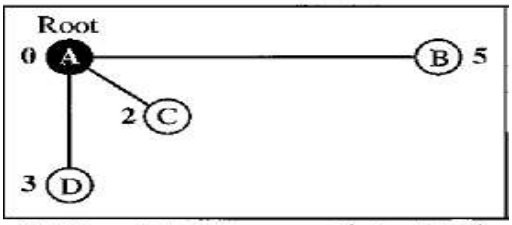
Step 1: We make node A the root of the tree and move it to the tentative list. Our two lists are

Permanent list: **Empty** Tentative list: **A(0)**

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

Step 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)

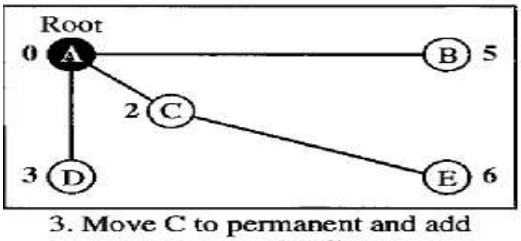


2. Move A to permanent list and add B, C, and D to tentative list.

Step 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).

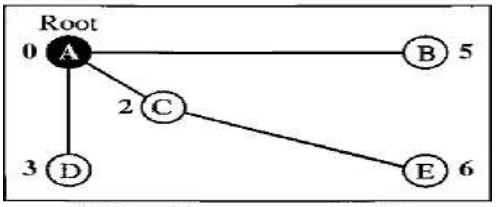


3. Move C to permanent and add E to tentative list.

Step 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).

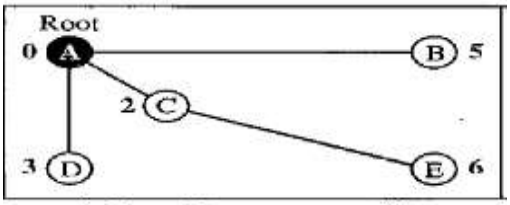


4. Move D to permanent list.

Step 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 (i.e. just node E).
- 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)

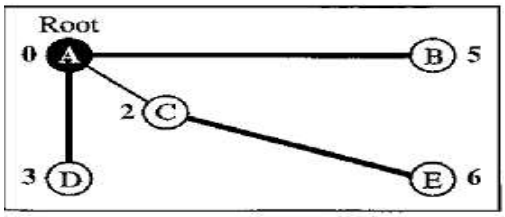


5. Move B to permanent list.

Step 6: Node E has the shortest cumulative cost from all nodes in the tentative list.

- Move E to the permanent list. Node E has no neighbor. Now the tentative list is empty.
- We stop the process here. The shortest path tree is ready for graph ABCDE.

The final lists are: Permanent list: A(0), B(5), C(2), D(3), E(6) Tentative list: Empty



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

The below table shows routing table for Node A.

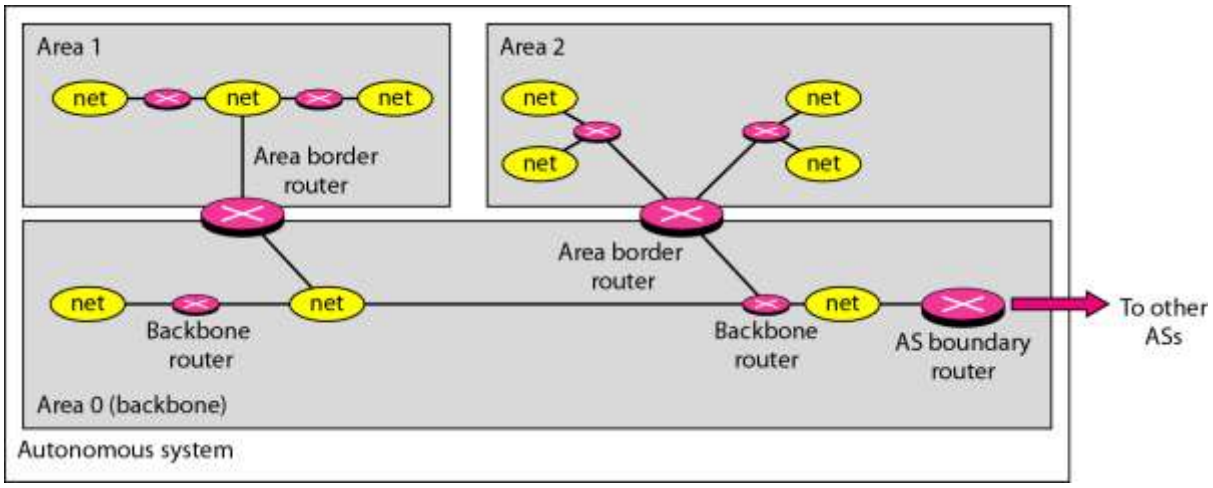
Node	Cost	Next Router
A	0	-
B	5	-
C	2	-
D	3	-
E	6	C

OSPF (Open Shortest Path First Protocol)

OSPF protocol is an intra-domain 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 protocol divides an autonomous system into many different areas.
- An area is a collection of networks, hosts, and routers all contained within an autonomous system.
- Each area has area identification. Ex: Area0, Area1, Area2 so on.
- All networks inside an area must be connected.



Routers inside an area flood the area with routing information.

- **Area Border Routers** are special routers located at the border of an area. These routers summarize the information about the area and send it to other areas.
- **Backbone** is a special area among the areas inside an autonomous system. All the areas inside an autonomous system must be connected to the backbone. Area identification of Backbone is 0.
- **Backbone routers** are a router inside the backbone.
- The backbone area serves as a primary area and the other areas as secondary areas.
- A backbone router can also be an area border router.

Metric

- The OSPF protocol allows the administrator to assign a cost to each route. The cost is called the metric.
- The metric can be based on a type of service such as minimum delay, maximum throughput, and so on.
- A router can have multiple routing tables. Each routing table is based on a different type of service.

Disadvantages of Distance Vector Routing and Link State 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 inter-domain 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.

To overcome these problems Path Vector Routing is introduced.

PATH VECTOR ROUTING (PVR)

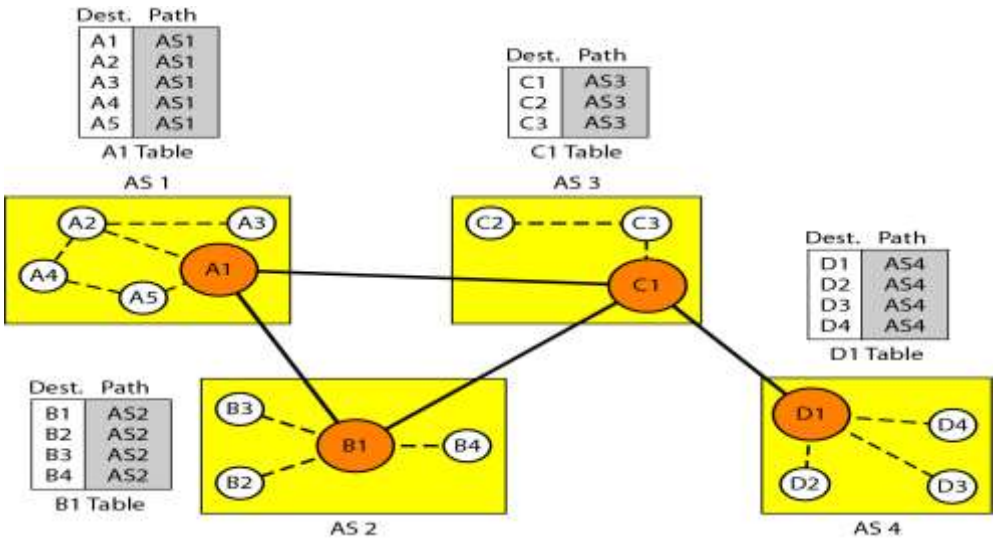
Path vector routing proved to be useful for interdomain routing.

- In path vector routing one node in each autonomous system that acts on behalf of the entire autonomous system. The node is called **Speaker Node**.
- The speaker node in an Autonomous System creates a **Routing Table** and advertises it to speaker nodes in the neighboring Autonomous Systems.
- In PVR only speaker nodes in each Autonomous System can communicate with each other.
- A speaker node advertises the path in its autonomous system or other autonomous systems. (paths such as AS1, AS1-AS2, AS1-AS2-AS4 etc).

Initialization

At the beginning, each speaker node can know only the reachability of nodes inside its autonomous system.

Consider the below figure that shows the initial tables for each speaker node in a system made of four Autonomous Systems.



In the above figure :

- AS1, AS2, AS3, AS4 are the four autonomous systems.
- Node Al, B1, C1, D1 are the Speaker Nodes of Autonomous Systems AS1, AS2, AS3, AS4 respectively.
- The tables of the autonomous systems are created by Speaker Nodes A1, B1, C1 and D1.

- Node Al creates an initial table that shows Al to A5 are located in AS1 and can be reached through it.
- Node Bl advertises that Bl to B4 are located in AS2 and can be reached through Bl and so on.

Sharing

In path vector routing a speaker node in an autonomous system shares its table with immediate neighbors.

- Node Al shares its table with nodes Bl and Cl.
- Node Bl shares its table with A1 and C1.
- Node Cl shares its table with nodes A1, B1, D1.
- Node Dl shares its table with Cl.

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.

The below figure shows the tables for each speaker node after the system is stabilized.

Dest.	Path
A1 … A5	AS1 … AS1
B1 … B4	AS1-AS2 … AS1-AS2
C1 … C3	AS1-AS3 … AS1-AS3
D1 … D4	AS1-AS2-AS4 … AS1-AS2-AS4

A1 Table

Dest.	Path
A1 … A5	AS2-AS1 … AS2-AS1
B1 … B4	AS2 … AS2
C1 … C3	AS2-AS3 … AS2-AS3
D1 … D4	AS2-AS3-AS4 … AS2-AS3-AS4

B1 Table

Dest.	Path
A1 … A5	AS3-AS1 … AS3-AS1
B1 … B4	AS3-AS2 … AS3-AS2
C1 … C3	AS3 … AS3
D1 … D4	AS3-AS4 … AS3-AS4

C1 Table

Dest.	Path
A1 … A5	AS4-AS3-AS1 … AS4-AS3-AS1
B1 … B4	AS4-AS3-AS2 … AS4-AS3-AS2
C1 … C3	AS4-AS3 … AS4-AS3
D1 … D4	AS4 … AS4

D1 Table

By observing the above figure following points to be noted:

- If router Al receives a packet for nodes A3, it knows that the path is in AS1 (i.e.) the packet is at home.
- If Router A1 receives a packet for Dl, it knows that the packet should go from ASl, to AS2 and then AS2 to AS3.
- The routing table shows the path completely.

Advantages of Path Vector Routing

1. Loop Prevention
2. Policy Routing
3. Optimum path

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.

Policy routing

- When a router receives a message, it can check the path. If one of the autonomous systems listed in the path is against router policy, it can ignore that path and that destination.
- Router does not update its routing table with this path, and router does not send this message to its neighbors.

Optimum path

- The optimum path in path vector routing is the path to a destination that is the best for the organization that runs the autonomous system.
- Metrics cannot include 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 RIP that defines hop count as the metric, another system may use OSPF with minimum delay defined as the metric.
- The optimum path is the path that fits the organization.

Example: Consider the above figure after updating the table:

- Each autonomous system may have more than one path to a destination.
- A path from AS4 to AS1 can be AS4-AS3-AS2-AS1, or it can be AS4-AS3-AS1.
- For the table entry we have selected the path that had the smaller number of autonomous systems (i.e. AS4-AS3-AS1).
- But we may take other path when the criteria are related to security, safety and reliability can also be applied.

Border Gateway Protocol (BGP)

Path Vector Routing uses the Border Gateway Protocol as an Inter-Domain routing protocol.

- The first version of BGP developed in 1989. There are 4 versions of BGP are developed.
- The Internet is divided into hierarchical domains called autonomous systems.

Example:

a. A large corporation that manages its own network and has full control over it is an autonomous system.
b. A local ISP that provides services to local customers is an autonomous system.

There are 3 Types of Autonomous Systems are present:

- Stub
- Multihomed
- Transit AS.

Stub Autonomous system (Stub AS)

- ✓ A stub Autonomous System has only one connection to another Autonomous System.
- ✓ The interdomain data traffic in a stub Autonomous System can be either created or terminated in the Autonomous System.
- ✓ The hosts in the Autonomous System can send and receive the data traffic from the hosts in other Autonomous System.
- ✓ Data traffic cannot pass through a stub AS. A Stub AS is either a source or a sink.

Example: A stub Autonomous System is a small corporation or a small local ISP.

Multi-homed Autonomous System (Multi-homed AS)

- A multi-homed Autonomous System has more than one connection to other Autonomous System, but it is still only a source or sink for data traffic.
- It can receive data traffic from more than one Autonomous System.
- It can send data traffic to more than one Autonomous System, but there is no transient traffic.
- It does not allow data coming from one Autonomous System and going to another Autonomous System to pass through.

Example: A multi-homed Autonomous system is a large corporation that is connected to more than one regional or national Autonomous System that does not allow transient traffic.

Transit Autonomous System

- A transit Autonomous System is a multi-homed Autonomous System that also allows transient traffic.

Example: National and International ISPs (Internet backbones).

Path Attributes

- The path is a list of Attributes rather than list of Autonomous Systems.
- 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:

1. **Well-known Attribute:** Every BGP router must recognize this attribute.
2. **Optional Attribute:** Every router need not be recognize this attribute.

Well-known attributes

Well-known attributes are divided into two categories: **Mandatory** and **Discretionary**.

- **Well-known Mandatory Attribute** is one that must appear in the description of a route.
- **Well-known Discretionary Attribute** is one that must be recognized by each router, but is not required to be included in every update message.

The well-known mandatory attributes are: ORIGIN, AS_PATH, NEXT-HOP.

- **ORIGIN:** This defines the source of the routing information (RIP, OSPF, and so on).
- **AS_PATH:** This defines the list of autonomous systems through which the destination can be reached.
- **NEXT-HOP:** This defines the next router to which the data packet should be sent.

Optional Attributes

Optional attributes can also be subdivided into two categories: Transitive and Non-transitive.

- **Optional Transitive attribute** is one that must be passed to the next router by the router that has not implemented optional attribute.
- **Optional Non-transitive attribute** is one that must be discarded if the receiving router has not implemented optional attribute.

BGP Sessions

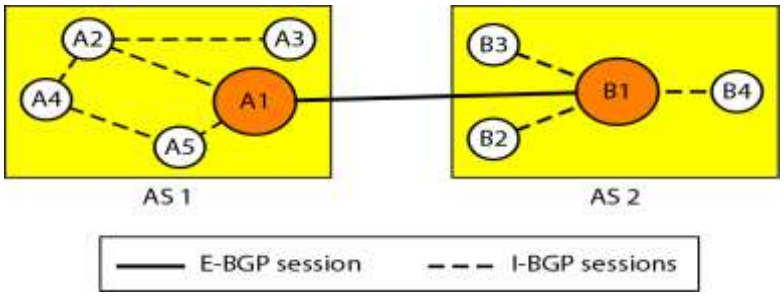
- The exchange of routing information between two routers using BGP takes place in a session.
- A session is a connection that is established between two BGP routers only for the sake of exchanging routing information.
- To create a reliable environment, BGP uses TCP services. (i.e.) a session at the BGP level as an application program, is a connection at the TCP level.
- When a TCP connection is created for BGP, it can last for a long time, until something unusual happens.

BGP sessions are referred to as Semi-permanent connections.

Types of BGP Sessions

BGP can have two types of sessions:

1. **External BGP Session (E-BGP):** It is used to exchange information between two speaker nodes belonging to two different autonomous systems.
2. **Internal BGP sessions (I-BGP):** It is used to exchange routing information between two routers inside an autonomous system.



Consider the above figure:

- 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.
- These two routers need to collect information from other routers in the autonomous systems. This is done using I-BGP sessions.

MULTICAST ROUTING PROTOCOLS

A message can be unicast, multicast, or broadcast.

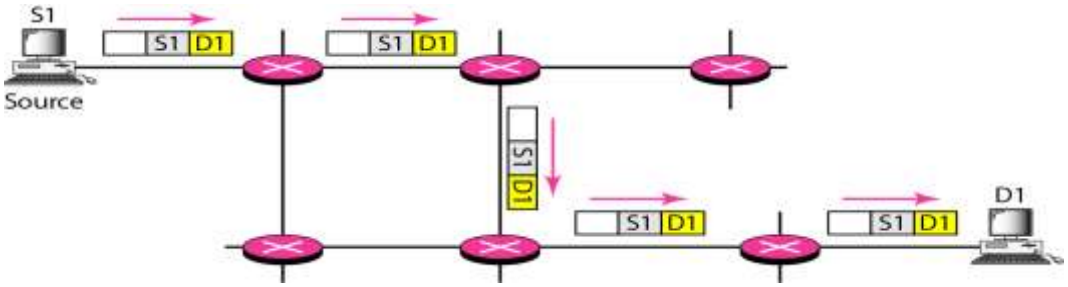
Unicasting

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

Consdier the below figure that shows the networks as a link between the routers.

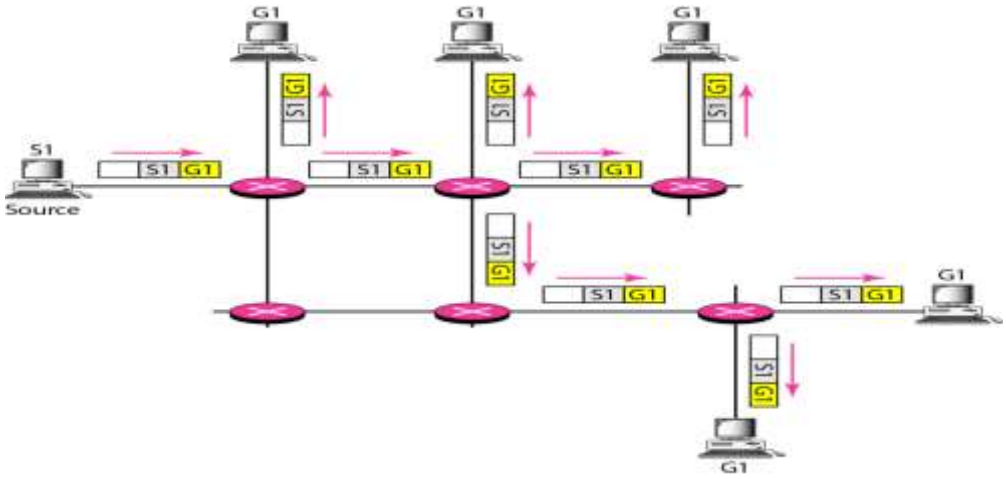
- A unicast packet starts from the source S1 and passes through routers to reach the destination D1.
- In unicasting when a router receives a packet, it forwards the packet through only one of its interfaces as defined in the routing table.

- The router may discard the packet if it cannot find the destination address in its routing table.



Multicasting

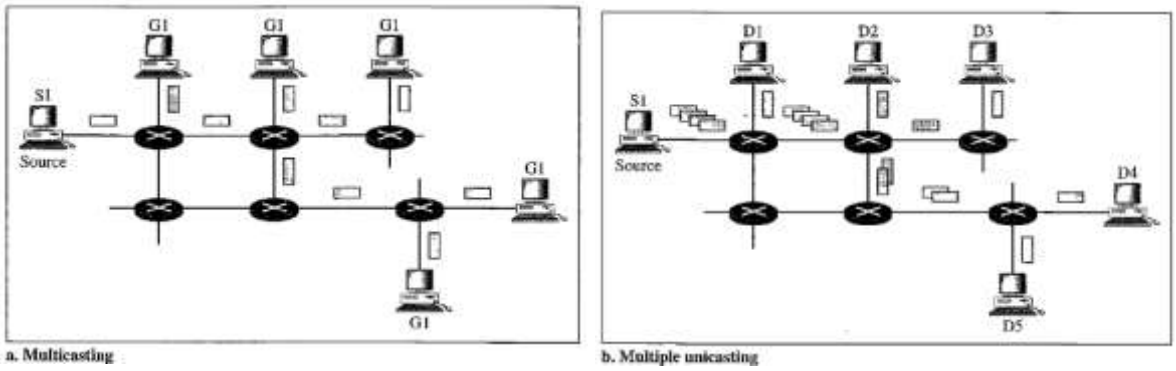
In Multicast communication the source address is a unicast address but the destination address is a group address which defines one or more destinations. The relationship is One-to-Many. The group address identifies the members of the group.



- 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.

Multicasting versus Multiple Unicasting

Multicasting	Multiple Unicasting
1. Multicasting starts with one single packet from the source that is duplicated by the routers.	1. In multiple unicasting several packets start from the source.
2. The destination address in each packet is the same for all duplicates.	2. If there are five destinations the source sends five packets, each with a different unicast destination address.
3. Only one single copy of the packet travels between any two routers.	3. There may be multiple copies traveling between two routers.
4. In multicasting, there is no delay because only one packet is created by the source.	4. In multiple unicasting, the packets are created by the source with a relative delay between packets.



a. Multicasting b. Multiple unicasting

Note: In multiple Unicasting if there are 1000 destinations, the delay between the first and the last packet may be unacceptable.

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.

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 it would create the huge amount of traffic and it needs huge amount of bandwidth.

Applications of Multicasting

Multicasting has many applications:

1. Distributed databases
2. Information dissemination
3. Teleconferencing
4. Distance learning

Distributed Database Access

- Today most of the large databases are distributed. The information is stored in more than one location.
- 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

- Today's Business 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

- News can be easily disseminated through multicasting.
- One single message can be sent to those interested in a particular topic.

Example: The statistics of the championship high school basketball tournament can be sent to the sports editors of many newspapers.

Teleconferencing

- Teleconferencing involves multicasting.
- All the individuals attending a teleconference 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

- Distance learning is today's major application in multicasting.
- 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.

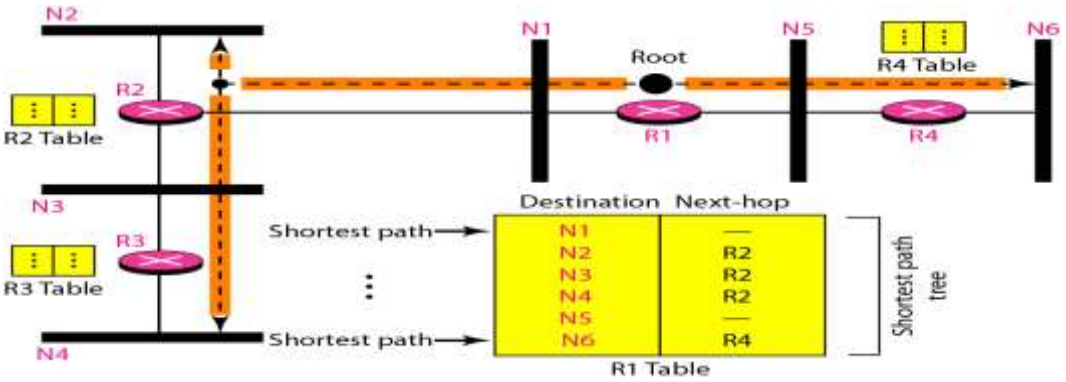
MULTICAST ROUTING

Optimal Routing: Shortest Path Trees

- The process of optimal Inter-Domain Routing 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.
- The number of trees and the formation of the trees in unicast and multicast routing are different.

Unicast Routing

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



- 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.
- Each line of the routing table is a shortest path (i.e.) the whole routing table is a shortest path tree.
- In unicast routing, each router needs only one shortest path tree to forward a packet. Each router has its own shortest path tree.

Multicast Routing

- The multicast routing is more complex than unicast routing.
- 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.

Two approaches have been used to solve the problem of complexity:

- Source-based trees
- Group-shared trees

Source-Based Tree

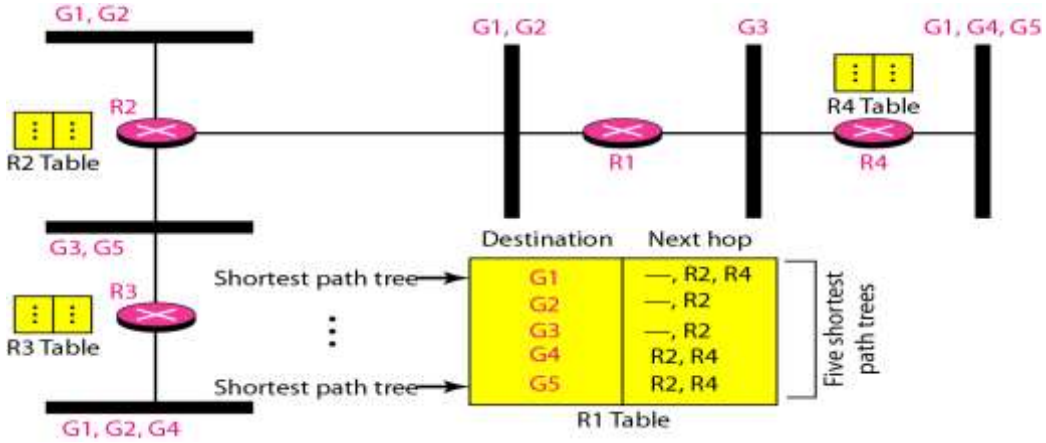
In the source-based tree approach, each router needs to have one shortest path tree for each group.

The shortest path tree for a group defines the next hop for each network that has loyal members for that group.

Consider the below figure with five groups in the domain: G1, G2, G3, G4, and G5 as the names of the Groups with Loyal members on each Network.

The figure shows the routing table for the router R1. At the moment

- G1 has loyal members in four networks
- G2 has loyal members in three networks
- G3 has loyal members in two networks
- G4 has loyal members in two networks
- G5 has loyal members in two networks



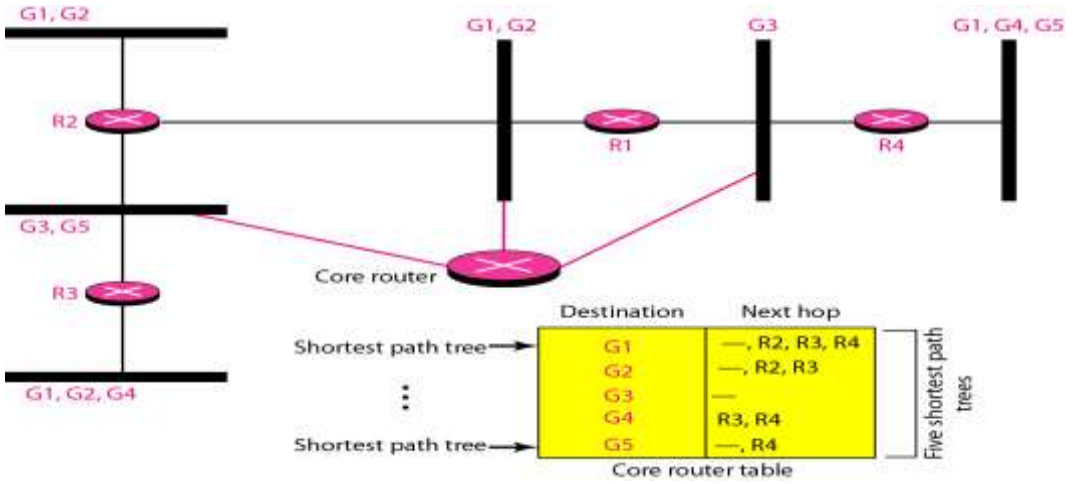
There is one shortest path tree for each group. Therefore there are five shortest path trees for five groups.

- If router R1 receives a packet with destination 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 then each router needs to have m shortest path trees.

The complexity of the routing table increases if we have hundreds or thousands of groups.

Group-Shared Tree

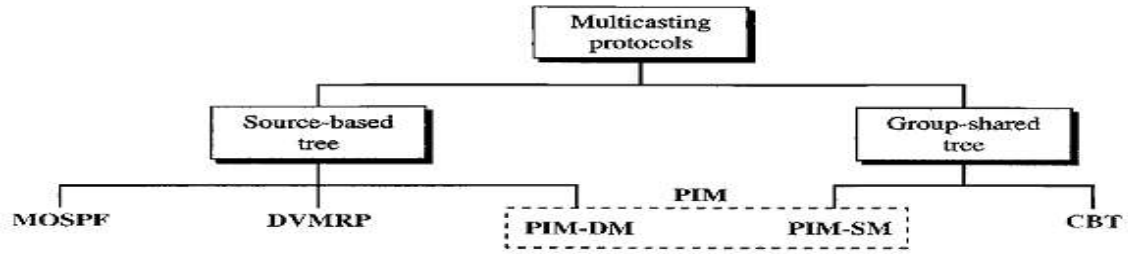
- In the group-shared tree approach, instead of each router having *m* shortest path trees, only one designated router takes the responsibility of distributing multicast traffic, the router is called the Center-Core Router or Rendezvous router.
- The core has *m* shortest path trees in its routing table.
- The rest of the routers in the domain have none.
- If a router receives a multicast packet, it encapsulates the multicast 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.



MULTICAST ROUTING PROTOCOLS

The Multicast Routing protocols are:

1. Multicast Link State Routing
2. Multicast Distance Vector Routing
3. Core Based Tree
4. Protocol Independent Multicast (PIM)
5. MBONE



Multicast Link State Routing

Multicast Open Shortest Path First (MOSPF) is an implementation of multicast link state routing in the internet.

- Multicast link state routing is a direct extension of unicast routing and uses a source-based tree approach.
- In multicast routing a node needs to revise the interpretation of state. State means "what groups are active on this link."
- A node advertises every group which has any loyal member on the link.

- The information about the group comes from IGMP.
- Each router running IGMP solicits the hosts on the link to find out the membership status.

Let us take n number of groups, when a router receives all the Link State Packets (LSP's), it creates n topologies from which n shortest path trees are made by using Dijkstra's algorithm.

Problem:

The only problem with this protocol is the time and space needed to create and save the many shortest path trees.

Solution:

- 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.

Multicast Open Shortest Path First (MOSPF)

- 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.
- Hence the tree 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.
- 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 Routing

Distance Vector Multicast Routing Protocol (**DVMRP**)is the implementation of Multicast Distance Vector routing. It is Source based routing protocol based on RIP.

- Multicast distance vector routing uses source-based trees, but the router never actually makes a routing table.
- When a router receives a multicast packet, the router forwards the packet by consulting its routing table.
- The shortest path tree is temporary. After a packet is forwarded the table is destroyed.

The multicast distance vector algorithm uses a process based on four decision-making strategies. Each strategy is built on its predecessor.

1. Flooding
2. Reverse Path Forwarding (RPF)
3. Reverse Path Broadcasting (RPB)
4. Reverse Path Multicasting (RPM)

Flooding

- Flooding broadcasts packets, but creates loops in the systems.
- A router receives a packet and without even looking at the destination group address, sends the packet out from every interface except the one from which it was received.
- Every network with active members receives the packet and the networks without active members also receive the packets. This is a broadcast not a multicast.

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.

Reverse path forwarding strategy will help to correct this problem.

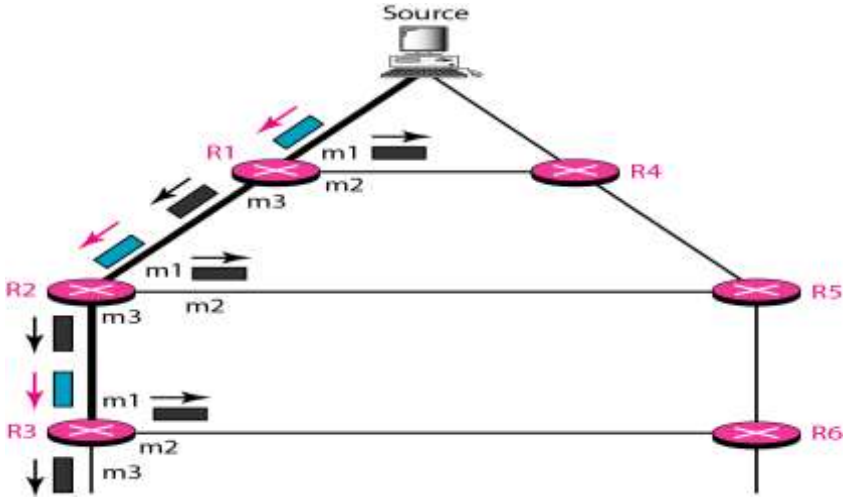
Reverse Path Forwarding (RPF)

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 (i.e. a unicast address).
- Router consults its unicast routing table as though it wants to send a packet to the source address.
- The routing table tells the router about 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.
- The router forwards the packet if it has traveled from the shortest path. Otherwise the packet is discarded.

By using the above strategy we can prevent loops 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.

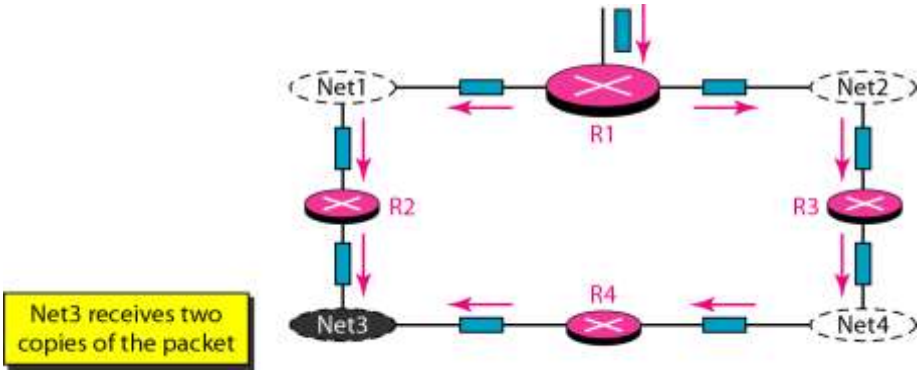
Example: Consider the below figure the shows the RPF process



- 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, router R1 consults its routing table and finds that the shortest path from R1 to the source is through interface m1. The packet is forwarded.
- 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.
- Similar process is applied for Routers R2 and R3 also.

Problem with RPF

- RPF does not guarantee that each network receives only one copy. A network may receive two or more copies.
- The reason is that RPF forwarding is based on source address not based on the destination address (i.e. a group address).



Consider the above figure:

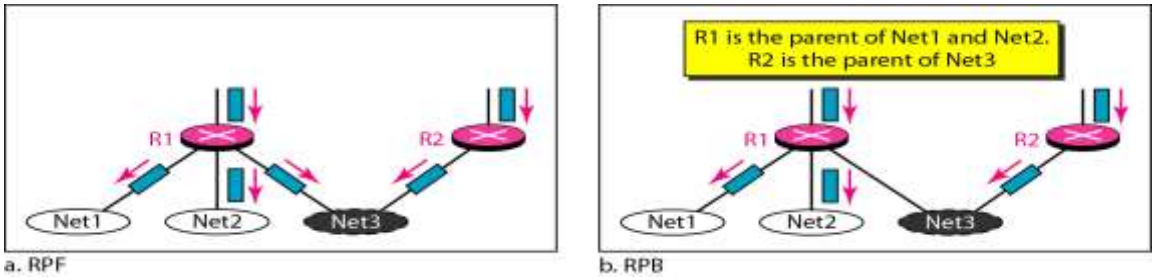
- Net3 receives two copies of the packet even though each router just sends out one copy from each interface.
- There is duplication because a tree has not been made. Instead of a tree we have a graph. Net3 has two parents: routers R2 andR4.

RPB will eliminate the above problem.

Reverse Path Broadcasting (RPB)

- To eliminate duplication in RPF, RPB defines a restriction that only one parent router for each network. The router is called **Designated Parent Router**.
- A network can receive a multicast packet from a particular source only through a designated parent router.
- For each source, the router sends the packet out of those interfaces for which it is the designated parent.

(i.e.) 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.



How the designated parent can be determined?

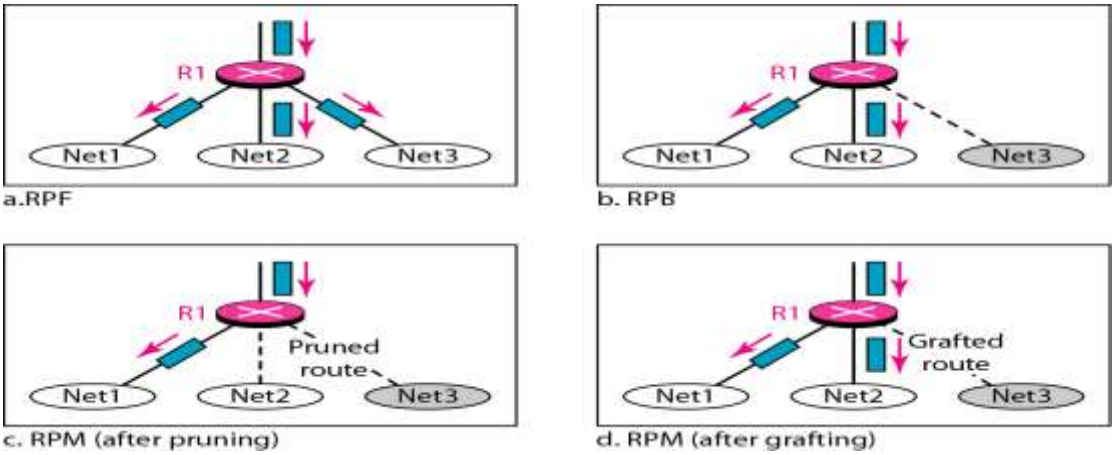
- The designated parent router can be the router with the shortest path to the source.
- Because routers periodically send updating packets to each other they can easily determine which router in the neighborhood has the shortest path to the source.
- If more than one router have same shortest path then the router with the smallest IP address is selected.

Problem with RPB: RPB does not multicast the packet, it broadcasts it. This is not efficient.

Reverse Path Multicasting (RPM)

RPM increases the efficiency by making sure that, the multicast packet must reach only those networks that have active members for that particular group.

To convert broadcasting to multicasting the protocol uses two procedures: Proning & Grafting.



Prune Message

- The designated parent router of each network is responsible for holding the membership information. This is done through the IGMP protocol.
- 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 means, the upstream router can stop sending multicast messages for this group through that interface.

Note: If this router receives prune messages from all downstream routers, then the router sends a prune message to its upstream router.

Graft Message

- A router at the bottom of the tree is called Leaf router.
- If a leaf router has sent a prune message but suddenly realizes through IGMP that, one of its networks is again interested in receiving the multicast packet, then it sends a **Graft message.**
- The Graft message forces the upstream router to resume sending the multicast messages.

Hence RPM adds **Pruning** and **Grafting** to RPB to create a multicast shortest path tree that supports dynamic membership changes.

Core-Based Tree (CBT)

- The Core-Based Tree protocol is a group-shared protocol that uses a core router as the root of the tree.
- The core router is also called as Center router or Rendezvous router.
- The autonomous system is divided into regions. A core is chosen for each region.

Formation of the Tree

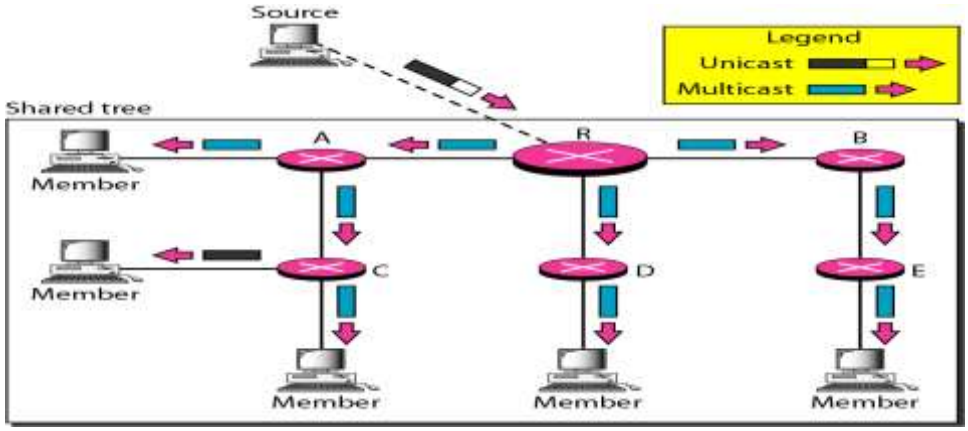
- After the rendezvous point is selected, every router is informed of the unicast address of the selected router.
- Each router then sends a unicast join message to show that it wants to join the group.
- This message passes through all routers that are located between the sender and the rendezvous router.
- 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 and every router knows its upstream router and the downstream router.
- 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.

Note:

1. Upstream Router is the router that leads to the root.
2. Downstream Router is the router that leads to the leaf.

Sending Multicast Packets

Consider the below figure that shows a group-shared tree with its rendezvous router **R** and a host can send a multicast packet to all members of the group.



- After formation of the tree, any source can send a multicast packet to all members of the group.
- The source may belong to the group, may not be belong to the group. (i.e.) the source host can be any of the hosts inside the shared tree or any host outside the shared tree.
- Source host 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.

Difference between DVRMP/MOSPF and CBT

1. The tree for the DVRMP or MOSPF is made from the root up (top to bottom), the tree for CBT is formed from the leaves down (bottom to top).
2. 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.

Protocol Independent Multicast (PIM)

Protocol Independent Multicast is divided into two independent multicast routing protocols:

1. Protocol Independent Multicast-Dense Mode (PIM-DM)
2. Protocol Independent Multicast-Sparse Mode (PIM-SM)

Both protocols are unicast protocol-dependent.

PIM-DM	PIM-SM
1. PIM-DM is used when there is a possibility that each router is involved in multicasting (dense mode).	1. PIM-SM is used when there is a less possibility that each router is involved in multicasting (sparse mode).
2. In this case, the use of a protocol that broadcasts the packet is efficient because almost all routers are involved in the process.	2. In this case, the use of a protocol that broadcasts the packet is not efficient. A protocol such as CBT that uses a group-shared tree is more efficient.
3. PIM-DM is a source-based tree routing protocol that uses RPF and pruning and grafting strategies for multicasting.	3. PIM-SM is a group-shared tree routing protocol that has a Rendezvous Point (RP) as the source of the tree.
4. Its operation is like that of DVMRP. Unlike DVMRP, PIM-DM does not depend on a specific unicasting protocol.	4. Its operation is like CBT but PIM-SM is simpler because it does not require acknowledgment from a join message.
5. PIM-DM 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.	5. PIM-SM creates a backup set of RPs for each region to cover RP failures.
6. This unicast protocol can be a distance vector protocol (RIP) or link state protocol (OSPF).	6. It uses CBT protocol.
7. Example: PIM-DM is used in a dense multicast environment such as a LAN.	7. Example: PIM-SM is used in a sparse multicast environment such as a WAN.

Advantage of PIM-SM

- PIM-SM 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.

Multicast Backbone (MBONE)

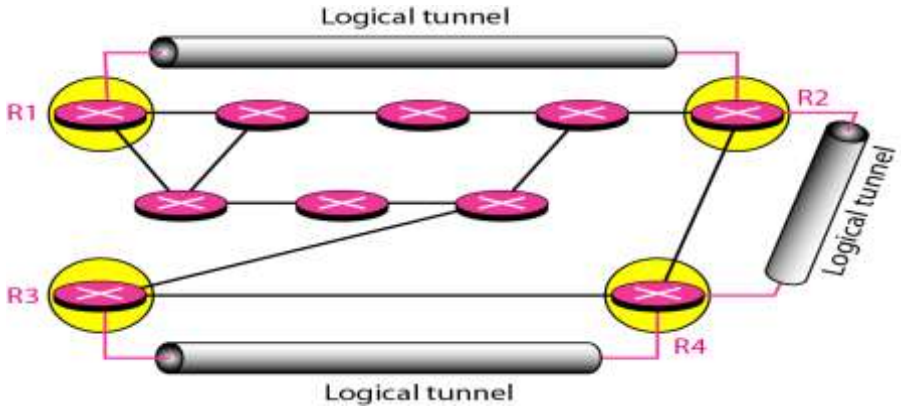
Multimedia and Real-time communication has increased the need for multicasting in the Internet. But only a small fraction of Internet routers are multicast routers.

Problem: A multicast router may not find another multicast router in the neighborhood to forward the multicast packet.

Solution: Tunneling

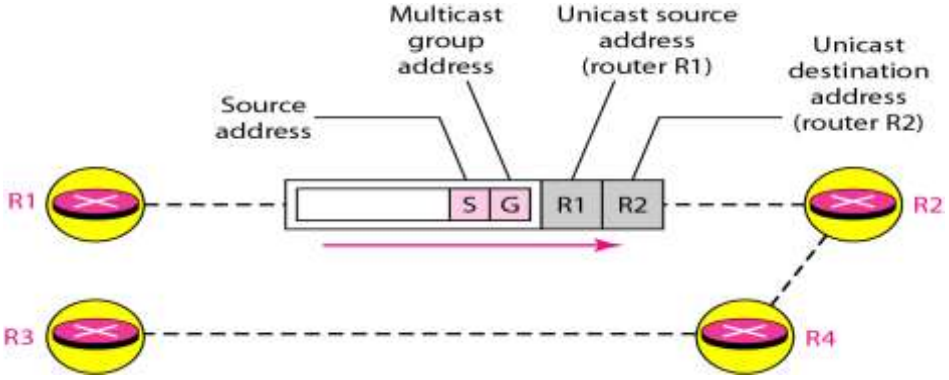
- The multicast routers are seen as a group of routers on top of unicast routers.
- The multicast routers may not be connected directly, but they are connected logically.

Consider the below figure that shows 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.

Consider the below figure that shows the process of MBONE and 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 routers (also called non-multicast routers) forward the packet as unicast routers and deliver the packet from one island to another because the unicast routers do not exist and the two multicast routers are neighbors.

Note: DVMRP is the only protocol that supports MBONE and tunneling.