

3.5

INTERDOMAIN MULTICAST PROTOCOLS

When the members of the groups are spread among different domains (ASs), we need an interdomain multicast routing protocol. One common protocol for interdomain multicast routing is called Multicast Border Gateway Protocol (MBGP), which is the extension of BGP.

BGP provides two paths between ASs: one for unicasting and one for multicasting.

Information about multicasting is exchanged between border routers in different ASs. MBGP is a shared-group multicast routing protocol in which one router in each AS is chosen as the rendezvous point (RP).

The problem with MBGP protocol is that it is difficult to inform an RP about the sources of groups in other ASs. The Multicast Source Discovery Protocol (MSDP) is a new suggested protocol that assigns a source representative router in each AS to inform all RPs about the existence of sources in that AS.

An extension of BGP, MBGP offers the following benefits:

1. A network can support incongruent unicast and multicast topologies.
2. A network can support congruent unicast and multicast topologies that have different policies (BGP filtering configurations).
3. A network can carry routing information for multiple network layer protocol address families (for example, IP Version 4 or VPN Version 4) as specified in RFC 1700, Assigned Numbers.
4. A network that is backward compatible—routers that support the multiprotocol extensions can interoperate with routers that do not support the extensions.
5. All of the routing policy capabilities of BGP can be applied to MBGP.
6. All BGP commands can be used with MBGP.

MBGP is an enhanced BGP that carries routing information for multiple network layer protocols and IP multicast routes. BGP carries two sets of routes, one set for unicast routing and one set for multicast routing. The routes associated with multicast routing are used by the Protocol Independent Multicast (PIM) feature to build data distribution trees.

MBGP is useful when you want a link dedicated to multicast traffic, perhaps to limit which resources are used for which traffic. Perhaps you want all multicast traffic exchanged at one network access point (NAP). MBGP allows you to have a unicast routing topology different from a multicast routing topology. Thus, you have more control over your network and resources.

In BGP, the only way to perform interdomain multicast routing was to use the BGP infrastructure that was in place for unicast routing. If those routers were not multicast-capable, or there were differing policies where you wanted multicast traffic to flow, multicast routing could not be supported without MBGP.

IPv6 ADDRESSING

IPv4 has the small size of the address space.

An IPv6 address is 128 bits or 16 bytes(octets) long, four times the address length in IPv4.

IPv6 address, in hexadecimal format, is very long, many of the digits are zeros.

In this case, the leading zeros of a section can be omitted. Using this form of abbreviation, 0074 can be written as 74, 000F as F, and 0000 as 0.

Address Space

The address space of IPv6 contains 2¹²⁸ addresses. This address space is 2⁹⁶ times the IPv4 address—definitely no address depletion—as shown, the size of the space is

340, 282, 366, 920, 938, 463, 374, 607, 431, 768, 211, 456.

Three Address Types

In IPv6, a destination address can be of three categories: unicast, anycast, and multicast.

Unicast Address

A unicast address defines a single interface (computer or router). The packet sent to a unicast address will be routed to the intended recipient (Receiver).

Anycast Address

An anycast address defines a group of computers that all share a single address. A packet with an anycast address is delivered to only one member of the group, the most reachable one.

An anycast communication is used, for example, when there are several servers that can respond to an inquiry. The request is sent to the one that is most reachable.

Multicast Address

A multicast address defines a group of computers.

Difference between any casting and multicasting.

In any casting, only one copy of the packet is sent to one of the members of the group; in multicasting each member of the group receives a copy.

Address Space Allocation

Like the address space of IPv4, the address space of IPv6 is divided into several blocks of varying size and each block is allocated for a special purpose. Most of the blocks are still unassigned and have been set aside for future use.

IPv6 Packet Format

The IPv6 packet is shown in Figure 3.5.1. Each packet has a base header followed by the payload. The base header occupies 40 bytes, payload is upto 65,535 bytes of information.

The description of fields follows.

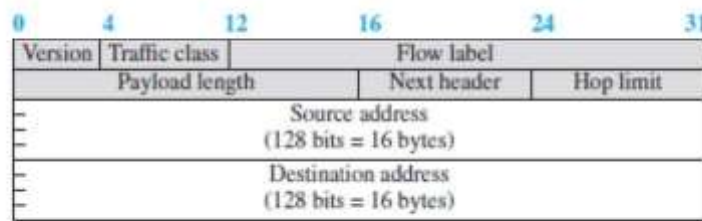
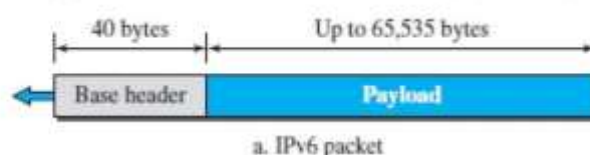


Fig3.5.1: IPv6 datagram.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-674]

Version.The 4-bit version field defines the version number of the IP. For IPv6, the value is 6.

Traffic class.The 8-bit traffic class field is used to distinguish different payloads with different delivery requirements. It replaces the type-of-service field in IPv4.

Flow label. The flow label is a 20-bit field that is designed to provide special handling for a particular flow of data.

Payload length.The 2-byte payload length field defines the length of the IP datagram excluding the header.

Note that IPv4 defines two fields related to the length: header length and total length.

In IPv6, the length of the base header is fixed (40 bytes); only the length of the payload needs to be defined.

Next header. The **next header** is an 8-bit field defining the type of the first extension header (if present) or the type of the data that follows the base header in the datagram.

Hop limit. The 8-bit hop limit field serves the same purpose as the TTL field in IPv4.

Source and destination addresses. The source address field is a 16-byte (128bit) internet address that identifies the original source of the datagram.

The destination address field is a 16-byte (128-bit) Internet address that identifies the destination of the datagram.

Payload. Compared to IPv4, the payload field in IPv6 has a different format and meaning, as shown in Figure .

Extension Header

An IPv6 packet is made of a base header and some extension headers. The length of the base header is fixed at 40 bytes.

To give more functionality to the IP datagram, the base header can be followed by up to six extension headers as shown in figure 3.5.2.

Six types of extension headers have been defined.

These are hop-by-hop option, source routing, fragmentation, authentication, encrypted security payload, and destination option (see Figure below).

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. For example, routers must be informed about certain management, debugging, or control functions.

Destination Option

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

Source Routing

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

Fragmentation

The concept of fragmentation in IPv6 is the same as that in IPv4.

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.

If the source does not use a Path MTU Discovery technique, it fragments the datagram to a size of 1280 bytes or smaller. This is the minimum size of MTU required for each network connected to the Internet.

Authentication

The authentication extension header has a dual purpose:

It validates the message sender and ensures the integrity of data. It is needed so the receiver can be sure that a message is from the genuine sender and not others.

Encrypted Security Payload

The encrypted security payload (ESP) is an extension that provides confidentiality and guards against eavesdropping.

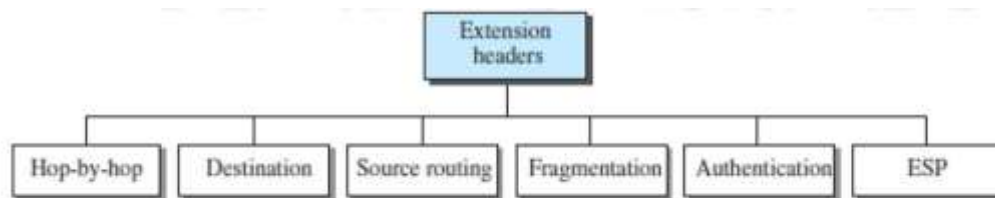


Fig3.5.2: Extension header types.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-677]

UNIT3 ROUTING

3.1

LEAST COST ROUTING

In unicast routing, a packet is routed, hop by hop, from its source to its destination by the help of forwarding tables. The source host needs no forwarding table because it delivers its packet to the default router in its local network.

Routing a packet from its source to its destination means routing the packet from a source router to a destination router.

Least-Cost Routing

When an internet is modeled as a weighted graph, one of the ways to interpret the best route from the source router to the destination router is to find the least cost between the two.

That is, the source router chooses a route to the destination router in such a way that the total cost for the route is the least cost among all possible routes.

In Figure 3.1.1, the best route between A and E is A-B-E, with the cost of 6.

This means that each router needs to find the least-cost route between itself and all the other routers to be able to route a packet towards the destination.

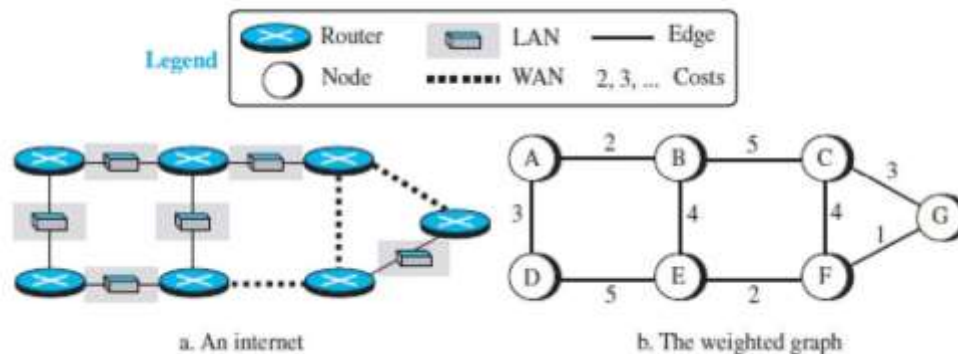


Fig3.1.1:Internet with graph.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-597]

Least-Cost Trees

If there are N routers in an internet, there are $(N - 1)$ least-cost paths from each router to any other router. This means we need $N(N - 1)$ least-cost paths for the whole internet. For example, If we have only 10 routers in an internet, we need 90 least-cost paths. A least-cost tree is a tree with the source router as the root that spans the whole graph (visits all

other nodes) and in which the path between the root and any other node is the shortest.

In this way, we can have only one shortest-path tree for each node; we have N least cost trees for the whole internet. Figure 3.1.2 shows the seven least-cost trees for the internet.

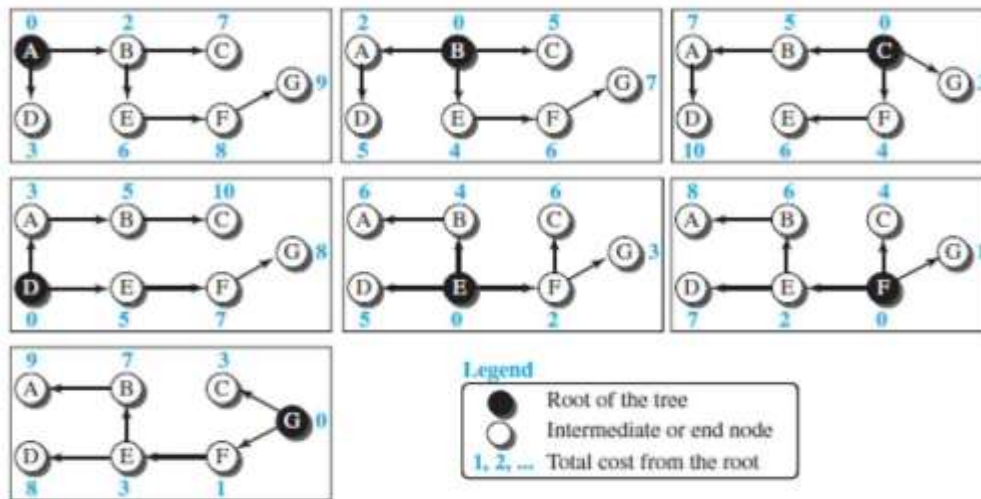


Fig 3.1.2: Seven least-cost trees for the internet.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-597]

1. The least-cost route from X to Y in X's tree is the inverse of the least-cost route from Y to X in Y's tree; the cost in both directions is the same.

For example, in Figure 3.1.2, the route from A to F in A's tree is (A -B-E-F), but the route from F to A in F's tree is (F -E-B-A), which is the inverse of the first route. The cost is 8 in each case.

2. Instead of travelling from X to Z using X's tree, we can travel from X to Y using X's tree and continue from Y to Z using Y's tree.

For example, in Figure 3.1.2 , we can go from A to G in A's tree using the route (A B-E -F - G). We can also go from A to E in A's tree (A -B-E) and then continue in E's tree using the route (E -F-G). The combination of the two routes in the second case is the same route as in the first case. The cost in the first case is 9; the cost in the second case is also 9 (6 3).

ROUTING ALGORITHMS

Distance-Vector Routing

In distance-vector routing, a router continuously tells all of its neighbors what it knows about the whole internet.

Bellman-Ford Equation

In distance-vector routing Bellman-Ford equation is used to find the least cost (shortest distance) between a source node, x , and a destination node, y , through some intermediary nodes (a, b, c, \dots) when the costs between the source and the intermediary nodes and the least costs between the intermediary nodes and the destination are given.

The following shows the general case in which D_{ij} is the shortest distance and c_{ij} is the cost between nodes i and j .

$$D_{xy} = \min \{ (c_{xa} + D_{ay}), (c_{xb} + D_{by}), (c_{xc} + D_{cy}), \dots \}$$

In distance-vector routing, we want to update an existing least cost with a least cost through an intermediary node, such as z , ie, if the intermediate node is shorter. In this case, the equation can be written as:

$$D_{xy} = \min \{ D_{xy}, (c_{xz} + D_{zy}) \}$$

Graphical idea behind Bellman-Ford equation

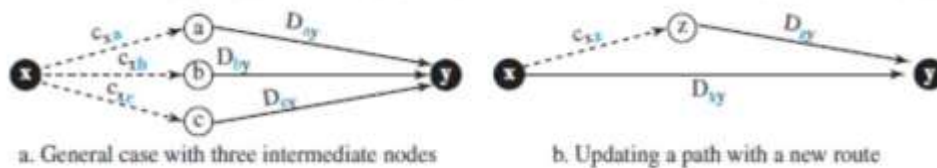


Fig3.1.3: Graphical idea.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-599]

Bellman-Ford equation help us to build a new least-cost path from previously established least-cost paths. In the Figure 3.1.3, we can think of $(a-y)$, $(b-y)$, and $(c-y)$ as previously established least-cost paths and $(x-y)$ as the new least-cost path.

Distance Vectors

The concept of a distance vector is the reason for the name distance-vector routing. A least-cost tree is a combination of least-cost paths from the root of the tree to all destinations.

Figure 3.1.4 shows the tree for node A in the internet and the corresponding distance vector.

A distance vector does not give the path to the destinations as the least-cost tree does; it gives only the least costs to the destinations.

Note that the name of the distance vector defines the root, the indexes define the destinations, and the value of each cell defines the least cost from the root to the destination.

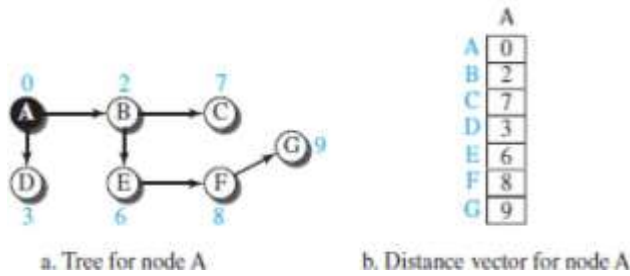


Fig3.1.4: Seven least-cost trees for the internet.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-600]

Each node in an internet, when it starts its function, creates a very basic distance vector with the minimum information the node can obtain from its neighborhood. The node sends some greeting messages out of its interfaces and discovers the identity of the immediate neighbors and the distance between itself and each neighbor.

It then makes a simple distance vector by inserting the discovered distances in the corresponding cells and leaves the value of other cells as infinity.

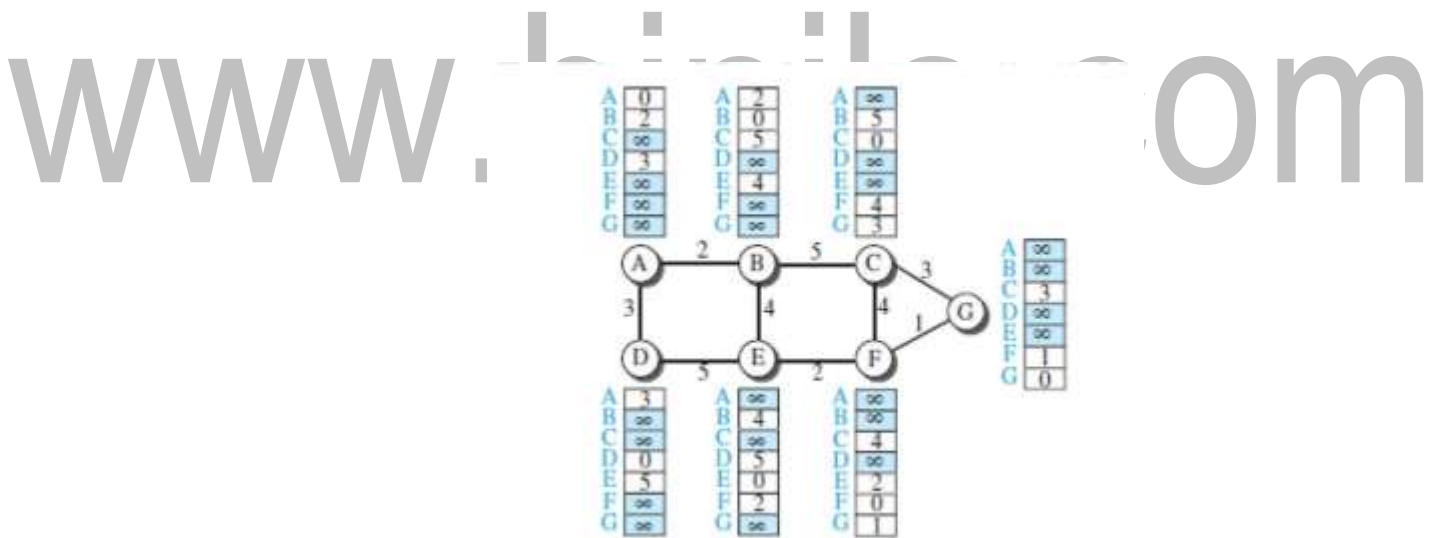


Fig3.1.5: First distance vector for internet.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-600]

Description of above diagram

Consider(For example), Node A in figure 3.1.5, thinks that it is not connected to node G because the corresponding cell shows the least cost of infinity.To improve these vectors, the nodes in the internet need to help each other by exchanging information. After each node has created its vector, it sends a copy of the vector to all its immediate neighbors. After a node receives a distance vector from a neighbor, it updates its distance vector using the Bellman-Ford equation (second case).

The figure 3.1.6, shows two asynchronous events, happening one after another with some time in between. In the first event, node A has sent its vector to node B. Node B updates its vector using the cost $c_{BA}=2$. In the second event, node E has sent its vector to node B. Node B updates its vector using the cost $c_{EA}=4$.

After the first event, node B has one improvement in its vector: its least cost to node D has changed from infinity to 5 (via node A). After the second event, node B has one more improvement in its vector; its least cost to node F has changed from infinity to 6 (via node E). By exchanging the vectors, we can stabilize the system and allows all nodes to find the ultimate least cost between themselves and any other node. After updating a node, it immediately sends its updated vector to all neighbors.

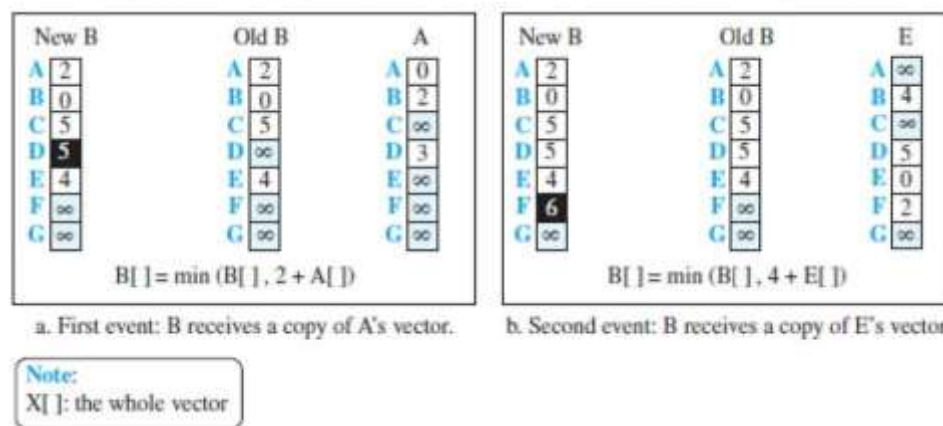


Fig3.1.6:Updating distance vector.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-601]

Count to Infinity

For a routing protocol to work properly, if a link is broken (cost becomes infinity), every other router should be aware of it immediately, but in distance-vector routing, this takes some time.

The problem is called count to infinity.

Two-Node Loop

Example of count to infinity is the two-node loop problem. To understand the problem, consider the Figure 3.1.7. The figure shows a system with three nodes. Initially both nodes A and B know how to reach node X. But suddenly, the link between A and X fails.

Node A changes its table. If A can send its table to B immediately, everything is fine. However, the system becomes unstable if B sends its forwarding table to A before receiving A's forwarding table.

Node A receives the update and, assuming that B has found a way to reach X, immediately updates its forwarding table.

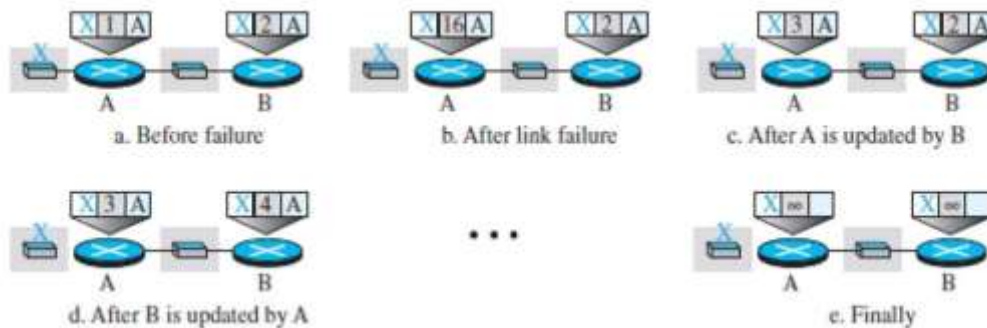


Fig3.1.7: Two node instability.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-603]

Now A sends its new update to B. Now B thinks that something has been changed around A and updates its forwarding 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 time the system is not stable. Node A thinks that the route to X is via B; node B thinks that the route to X is via A. If A receives a packet destined for X, the packet goes to B and then comes back to A. Similarly, if B receives a packet destined for X, it goes to A and comes back to B. Packets bounce between A and B, creating a two-node loop problem.

A few solutions have been proposed for instability of this kind.

Split Horizon

One solution to instability is called split horizon. In this method, instead of flooding the table through each interface, each node sends only part of its table through each interface.

If, according to its table, node B thinks that the optimum route to reach X is via A, it does not need to advertise this piece of information to A; the information has come from A (A already knows). Taking information from node A, modifying it, and sending it back to node A is what creates the confusion. In this method, node B eliminates the last line of its forwarding 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 forwarding table to B, node B also corrects its forwarding table. The system becomes stable after the first update: both node A and node B know that X is not reachable.

3.2

LINK-STATE ROUTING

A routing algorithm that directly creates least-cost trees and forwarding tables is link state (LS) routing. This method uses the term link-state to define the characteristic of a link (an edge) that represents a network in the internet.

Link-State Database (LSDB)

To create a least-cost tree with this method, each node needs to have a complete map of the network, which means it needs to know the state of each link. The collection of states for all links is called the link-state database (LSDB) as in figure 3.2.1.

Example for LSDB

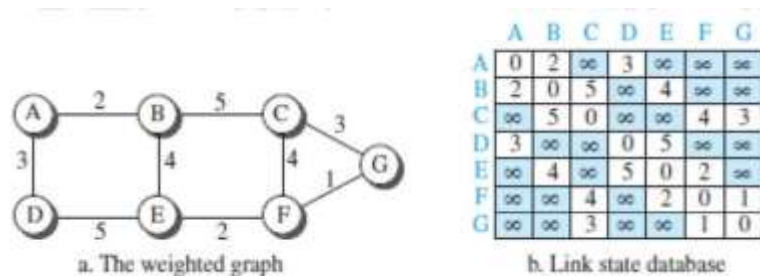


Fig3.2.1: Example Link state data base.

[Source : “Data Communications and Networking” by Behrouz A. Forouzan,Page-604]

This method is called flooding.

Each node can send some greeting messages to all its immediate neighbors (those nodes to which it is connected directly) to collect two pieces of information for each neighboring node: the identity of the node and the cost of the link.

The combination of these two pieces of information is called the LS packet (LSP); the LSP is sent out of each interface, as shown in Figure 3.2.2.

When a node receives an LSP from one of its interfaces, it compares the LSP with the copy it may already have. If the newly arrived LSP is older than the one it has (found by checking the sequence number),it discards the LSP. If it is newer or the first one received, the node discards the old LSP (if there is one) and keeps the received one. It then 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 network (where a node has only one interface).

After receiving all new LSPs, each node creates the comprehensive LSDB as shown in Figure 3.2.2. This LSDB is the same for each node and shows the whole map of the internet. A node can make the whole map if it needs to, using this LSDB.

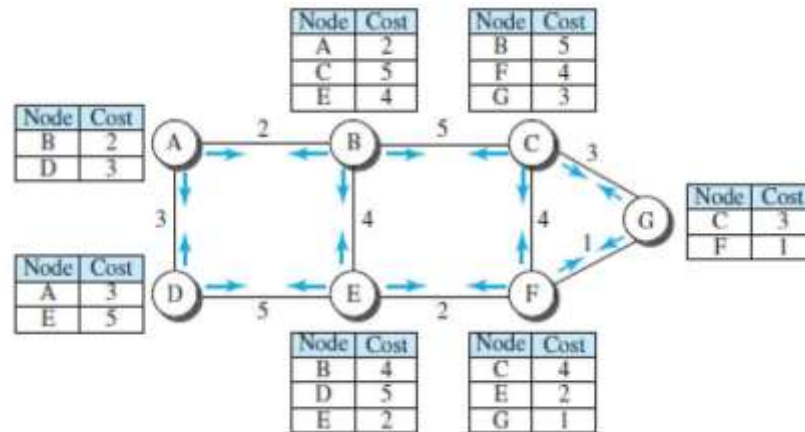


Fig3.2.2:Creation of LSP.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-605]

Note:

In the distance-vector routing algorithm, each router tells its neighbors what it knows about the whole internet; in the link-state routing algorithm, each router tells the whole internet what it knows about its neighbors.

Formation of Least-Cost Trees

To create a least-cost tree for itself, using the shared LSDB, each node needs to run the famous Dijkstra Algorithm. It is illustrated in figure 3.2.3.

This algorithm uses the following steps:

- 1.The node chooses itself as the root of the tree, creating a tree with a single node, and sets the total cost of each node based on the information in the LSDB.
- 2.The node selects one node, among all nodes not in the tree, which is closest to the root, and adds this to the tree. After this node is added to the tree, the cost of all other nodes not in the tree needs to be updated because the paths may have been changed.
- 3.The node repeats step 2 until all nodes are added to the tree.

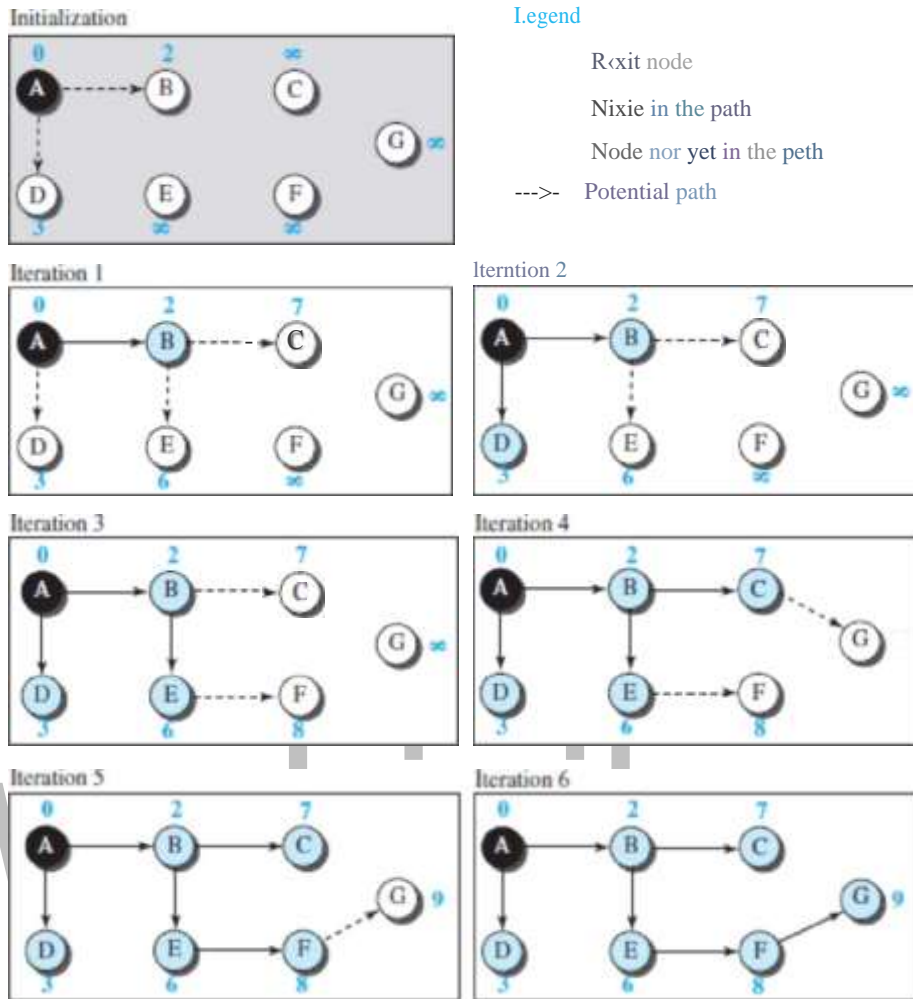


Fig3.2.3: Least Cost Tree.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-607]

3.4 MULTICAST ROUTING

In multicasting, there is one source and a group of destinations. The relationship is one to many. In this type of communication, the source address is a unicast address, but the destination address is a group address, a group of one or more destination networks in which there is at least one member of the group that is interested in receiving the multicast datagram.

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

Multicast Applications

Access to Distributed Databases.

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

Information Dissemination.

Businesses often need to send information to their customers. If the nature of the information is the same for each customer, it can be multicast. In this way a business can send one message that can reach many customers.

Teleconferencing. Teleconferencing involves multicasting. The individuals attending a teleconference all need to receive the same information at the same time.

Distance Learning. One growing area in the use of multicasting is distance learning.

MULTICASTING BASICS

In multicast communication, the sender is only one, but the receiver is many, sometimes thousands or millions spread all over the world. It should be clear that we cannot include the addresses of all recipients in the packet. The destination address of a packet, as described in the Internet Protocol (IP) should be only one. For this reason, we need multicast addresses. A multicast address defines a group of recipients, not a single one.

A multicast address is an identifier for a group. If a new group is formed with some active members, an authority can assign an unused multicast address to this group to uniquely define it.

Multicast Forwarding

Important issue in multicasting is the decision a router needs to make to forward a multicast packet. Forwarding in unicast and multicast communication is different in two aspects:

In unicast communication, the destination address of the packet defines one single destination. The packet needs to be sent only out of one of the interfaces, the interface which is the branch in the shortest-path tree reaching the destination with the minimum cost.

In multicast communication, the destination of the packet defines one group, but that group may have more than one member in the internet. To reach all of the destinations, the router may have to send the packet out of more than one interface. Figure (below) shows the concept.

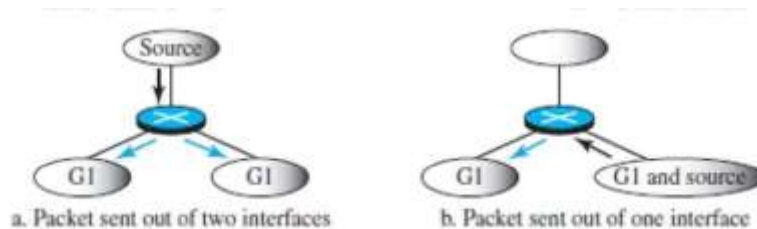


Fig3.4.1: Forwarding depends on destination and source.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-649]

In part a of the figure 3.4.1, the source is in a section of the internet where there is no group member. In part b, the source is in a section where there is a group member. In part a, the router needs to send out the packet from two interfaces; in part b, the router should send the packet only from one interface to avoid sending a second copy of the packet from the interface it has arrived at.

Two Approaches to Multicasting

Source-Based Tree Approach

In the source-based tree approach to multicasting, each router needs to create a separate tree for each source-group combination. If there are m groups and n sources in the internet, a router needs to create $(m \times n)$ routing trees. In each tree, the corresponding source is the root, the members of the group are the leaves, and the router itself is somewhere on the tree.

Group-Shared Tree Approach

In the group-shared tree approach, a router act like a source for each group. The designated router, which is called the core router or the rendezvous point router, acts as the representative for the group. Any source that has a packet to send to a member of that group sends it to the core center (unicast

communication) and the core center is responsible for multicasting. The core center creates one single routing tree with itself as the root and any routers with active members in the group as the leaves. In this approach, there are m core routers (one for each group) and each core router has a routing tree, for the total of m trees. Therefore the number of routing trees is reduced from $(m \times n)$ in the source-based tree approach to m in this approach.

INTRADOMAIN MULTICAST PROTOCOLS

Multicast Distance Vector (DVMRP)

The Distance Vector Multicast Routing Protocol (DVMRP) is the extension of the Routing Information Protocol (RIP) which is used in unicast routing. It uses the source based tree approach to multicasting.

Multicast tree in three steps:

1. The router uses an algorithm called reverse path forwarding (RPF) to simulate creating part of the optimal source-based tree between the source and itself.
2. The router uses an algorithm called reverse path broadcasting (RPB) to create a broadcast (spanning) tree whose root is the router itself and whose leaves are all networks in the internet.
3. The router uses an algorithm called reverse path multicasting (RPM) to create a multicast tree by cutting some branches of the tree that end in networks with no member in the group.

Reverse Path Forwarding (RPF)

The first algorithm, reverse path forwarding (RPF), forces the router to forward a multicast packet from one specific interface: the one which has come through the shortest path from the source to the router. The router does not know the shortest path from the source to itself, but it can find which is the next router in the shortest path from itself to the source (reverse path).

The router simply consults its unicast forwarding table, pretending that it wants to send a packet to the source; the forwarding table gives the next router and the interface the message that the packet should be sent out in this reverse direction.

The router uses this information to accept a multicast packet only if it arrives from this interface. This is needed to prevent looping. In multicasting, a packet may arrive at the same router that has forwarded it. If the router does not drop all arrived packets except the one, multiple copies of the packet will be circulating in the internet.

Reverse Path Broadcasting (RPB)

The RPF algorithm helps a router to forward only one copy received from a source and drop the rest. When we think about broadcasting in the second step, we need to remember that destinations are all the networks (LANs) in the internet. To be efficient, we need to prevent each network from receiving more than one copy of the packet.

If a network is connected to more than one router, it may receive a copy of the packet from each router. RPF cannot help here, because a network does not have the intelligence to apply the RPF algorithm; we need to allow only one of the routers attached to a network to pass the packet to the network.

One way to do so is to designate only one router as the parent of a network related to a specific source. When a router that is not the parent of the attached network receives a multicast packet, it simply drops the packet. There are several ways that the parent of the network related to a network can be selected; one way is to select the router that has the shortest path to the source (using the unicast forwarding table, again in the reverse direction).

Every packet started from the source reaches all LANs in the internet travelling the shortest path.

Figure 3.4.2 shows how RPB can avoid duplicate reception in a network by assigning a designated parent router, R1, for network N. The difference between RPB and RPF is shown in figure 3.4.3.

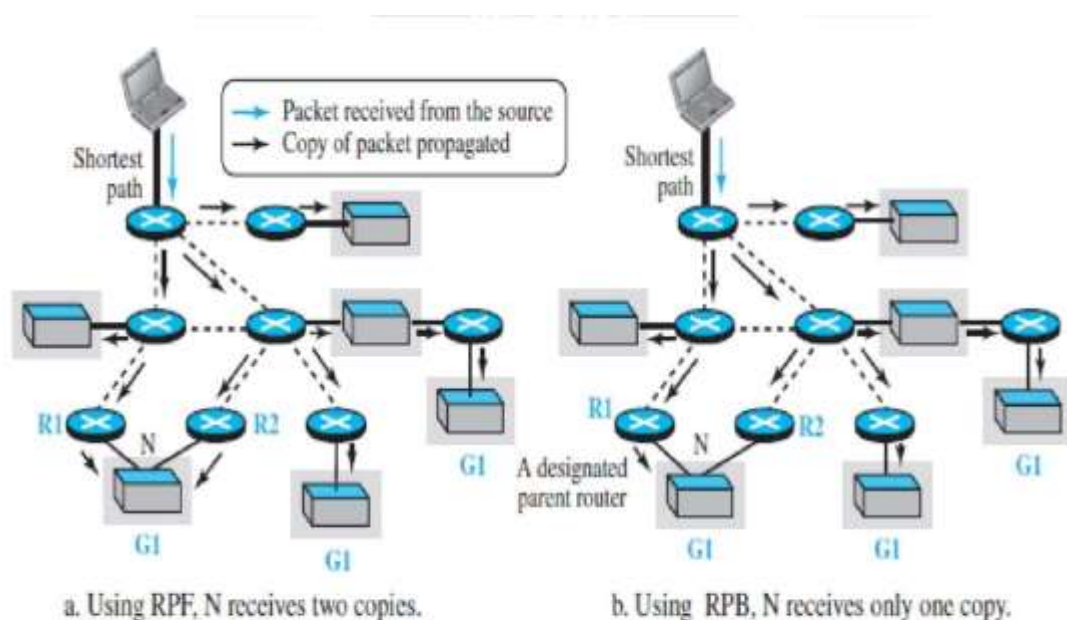


Fig3.4.2: Reverse path broadcasting.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-652]

Reverse Path Multicasting (RPM)

To increase efficiency, the multicast packet must reach only those networks that have active members for that particular group. This is called reverse path multicasting (RPM).

To change the broadcast shortest-path tree to a multicast shortest-path tree, each router needs to prune (make inactive) the interfaces that do not reach a network with active members corresponding to a particular source-group combination.

This step can be done bottom-up, from the leaves to the root. At the leaf level, the routers connected to the network collect the membership information using the IGMP protocol.

The parent router of the network can then disseminate this information upward using the reverse shortest-path tree from the router to the source, the same way as the distance vector messages are passed from one neighbor to another.

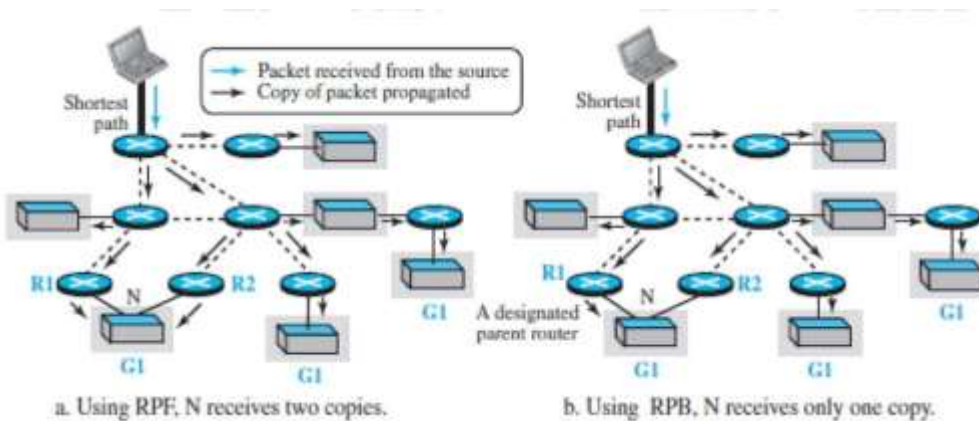


Fig3.4.3:RPM vs RPF.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-653]

Multicast Link State (MOSPF)

Multicast Open Shortest Path First (MOSPF) is the extension of the Open Shortest Path First (OSPF) protocol, which is used in unicast routing as shown in figure 3.4.4. It uses the source based tree approach to multicasting.

In multicasting, each router needs to have a database, as with the case of unicast distance-vector routing, to show which interface has an active member in a particular group.

A router follow these steps to forward a multicast packet received from source S and to be sent to destination G (a group of recipients):

The router uses the Dijkstra algorithm to create a shortest-path tree with S as the root and all destinations in the internet as the leaves. Note that this shortest-path tree is different from the

one the router normally uses for unicast forwarding, in which the root of the tree is the router itself.

Here, the root of the tree is the source of the packet defined in the source address of the packet. The router finds itself in the shortest-path tree created in the first step. In other words, the router creates a shortest-path subtree with itself as the root of the subtree. The shortest-path subtree is actually a broadcast subtree with the router as the root and all networks as the leaves.

The IGMP protocol is used to find the information at the leaf level. The router can now forward the received packet out of only those interfaces that correspond to the branches of the multicast tree.

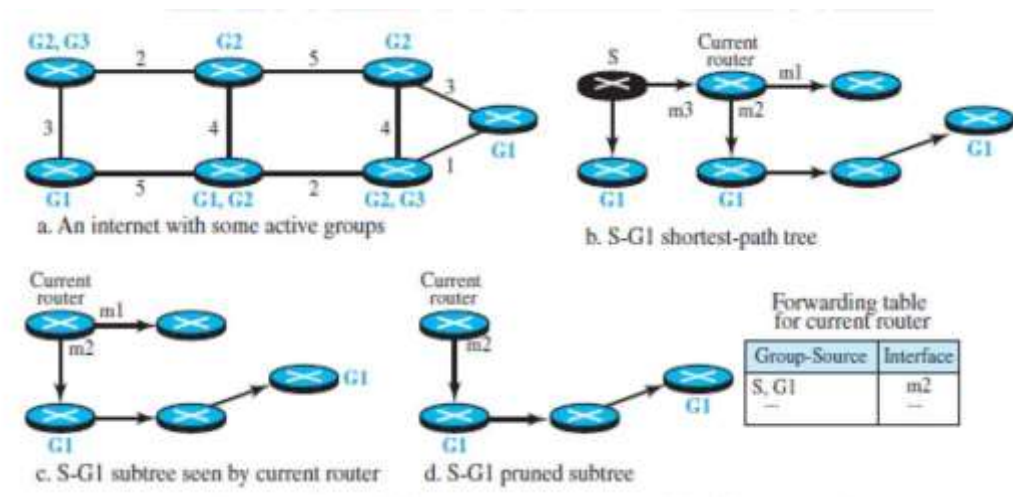


Fig3.4.4: Tree formation in MOSPF.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-654]

Protocol Independent Multicast (PIM)

Protocol Independent Multicast (PIM) is the name given to a common protocol that needs a unicast routing protocol for its operation, but the unicast protocol can be either a distance-vector protocol or a link-state protocol. PIM uses the forwarding table of a unicast routing protocol to find the next router in a path to the destination, but it does not matter how the forwarding table is created.

Feature of PIM: It can work in two different modes: dense and sparse.

The term dense means that the number of active members of a group in the internet is large; the probability that a router has a member in a group is high. For example, in a popular teleconference that has a lot of members.

The term sparse, means that only a few routers in the internet have active members in the group; the probability that a router has a member of the group is low. For example, in a technical teleconference where a number of members are spread somewhere in the internet. When the protocol is working in the dense mode, it is referred to as PIM-DM; when it is working in the sparse mode, it is referred to as PIM-SM.

Protocol Independent Multicast-Dense Mode (PIM-DM)

When the number of routers with attached members is large relative to the number of routers in the internet, PIM works in the dense mode and is called PIM-DM. In this mode, the protocol uses a source-based tree approach as shown in figure 3.4.5.

PIM-DM uses only two strategies described in DVMRP: RPF and RPM.

The two steps used in PIM-DM .

1. A router that has received a multicast packet from the source S destined for the group G first uses the RPF strategy to avoid receiving a duplicate of the packet. It consults the forwarding table of the unicast protocol to find the next router if it wants to send a message to the source S (in the reverse direction). If the packet has not arrived from the next router in the reverse direction, it drops the packet and sends a prune (remove things which are not needed) message in that direction to prevent receiving future packets related to (S, G).

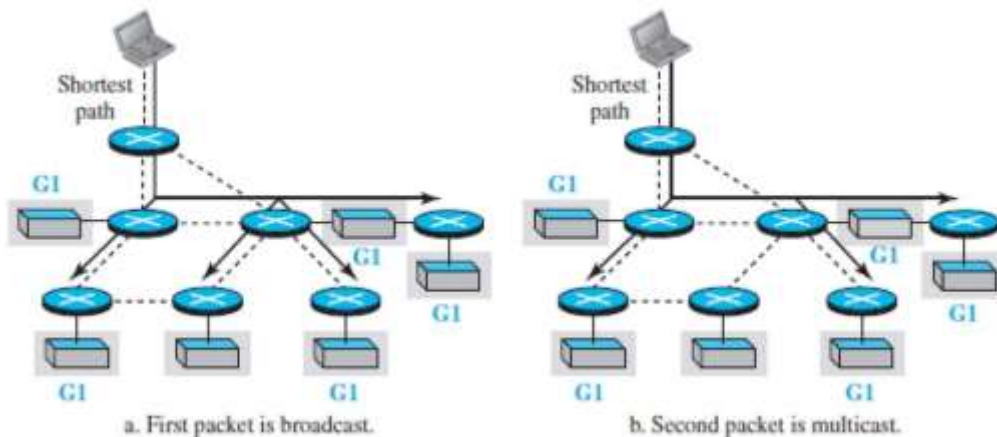


Fig3.4.5: Idea behind PIM- DM.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-656]

2. If the packet in the first step has arrived from the next router in the reverse direction, the receiving router forwards the packet from all its interfaces except the one from which the packet has arrived . Note that this is broadcasting instead of a multicasting if the packet is the first packet from the source S to group G. Each router downstream that receives an unwanted packet sends a prune

message to the router upstream, and eventually the broadcasting is changed to multicasting.

Figure (above) PIM-DM. The first packet is broadcast to all networks, which have or do not have members. After a prune message arrives from a router with no member, the second packet is only multicast.

Protocol Independent Multicast-Sparse Mode (PIM-SM)

When the number of routers with attached members is small relative to the number of routers in the internet, PIM works in the sparse mode and is called PIM-SM as shown in figure 3.4.6.

In this environment, PIM-SM uses a group-shared tree approach to multicasting. The core router in PIM-SM is called the rendezvous point (RP). Multicast communication is achieved in two steps.

Any router that has a multicast packet to send to a group of destinations first encapsulates the multicast packet in a unicast packet (tunneling) and sends it to the RP. The RP then decapsulates the unicast packet and sends the multicast packet to its destination.

PIM-SM uses a complex algorithm to select one router among all routers in the internet as the RP for a specific group. This means that if we have m active groups, we need m RPs, although a router may serve more than one group.

After the RP for each group is selected, each router creates a database and stores the group identifier and the IP address of the RP for tunneling multicast packets to it.

PIM-SM uses a spanning multicast tree rooted at the RP with leaves pointing to designated routers connected to each network with an active member. A very interesting point in PIM-SM is the formation of the multicast tree for a group.

To create a multicast tree rooted at the RP, PIM-SM uses join and prune messages.

Figure (below) shows the operation of join and prune messages in PIM-SM. First, three networks join group G1 and form a multicast tree. Later, one of the networks leaves the group and the tree is pruned. The join message is used to add possible new branches to the tree; the prune message is used to cut branches that are not needed.

When a designated router finds out that a network has a new member in the corresponding group (via IGMP), it sends a join message in a unicast packet destined for the RP. The packet travels through the unicast shortest-path tree to reach the RP. Any router in the path receives and forwards the packet, but at the same time, the router adds two pieces of information to its multicast forwarding table.

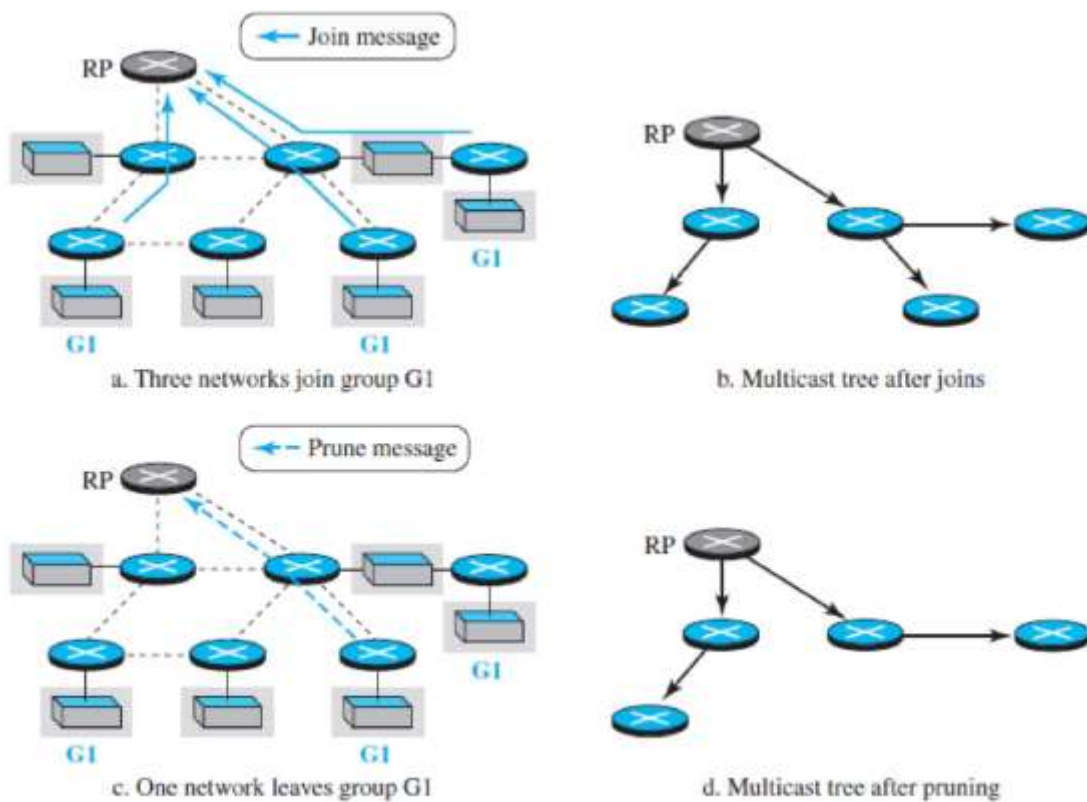


Fig3.4.6: Idea behind PIM-SM.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-657]

The number of the interface through which the join message was sent to the RP is marked (if not already marked) as the only interface through which the multicast packet destined for the same group should be received. In this way, the first join message sent by a designated router creates a path from the RP to one of the networks with group members. To avoid sending multicast packets to networks with no members, PIM-SM uses the prune message.

3.6

.6 TRANSITION FROM IPv4 TO IPv6

Because of the huge number of systems in the Internet, the transition from IPv4 to IPv6 cannot happen suddenly. It will take a considerable amount of time before every system in the Internet can move from IPv4 to IPv6. The transition must be smooth to prevent any problems between IPv4 and IPv6 systems.

Strategies

Three strategies are used for transition: dual stack, tunneling, and header translation.

One or all of these three strategies can be implemented during the transition period.

Dual Stack

It is recommended that all hosts, before migrating completely to version 6, have a dual stack of protocols during the transition. In other words, a station must run IPv4 and IPv6 simultaneously until all the Internet uses IPv6.

Figure 3.6.1 shows the layout of a dual-stack configuration.

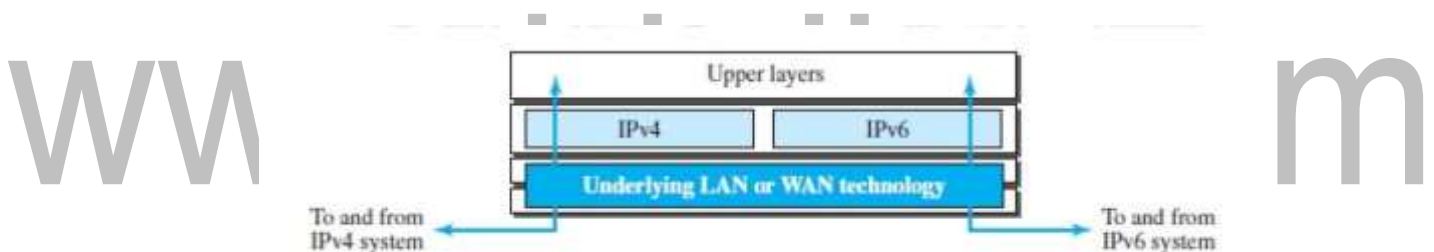


Fig3.6.1: Dual stack.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-683]

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

Tunneling

Tunneling is a strategy used when two computers using IPv6 want to communicate with each other and the packet must pass through a region that uses IPv4.

To pass through this region, the packet must have an IPv4 address. So the IPv6 packet is encapsulated in an IPv4 packet when it enters the region, and it leaves its capsule when it exits the region.

It seems as if the IPv6 packet enters a tunnel at one end and emerges at the other end. To make it clear that the IPv4 packet is carrying an IPv6 packet as data, the protocol value is set to 41. Tunneling is shown in Figure 3.6.2.

Header Translation

Header translation is necessary when the majority of the Internet has moved to IPv6 but some systems still use IPv4.

The sender wants to use IPv6, but the receiver does not understand IPv6. Tunneling does not work in this situation because the packet must be in the IPv4 format to be understood by the receiver. In this case, the header format must be totally changed through header translation. The header of the IPv6 packet is converted to an IPv4 header.

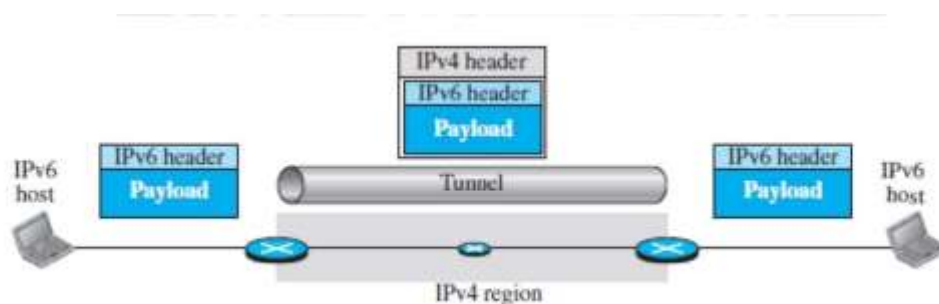


Fig3.6.2: Tunneling strategy.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-684]

Use of IP Addresses

During the transition a host may need to use two addresses, IPv4 and IPv6. When the transition is complete, IPv4 addresses should disappear. The DNS servers need to be ready to map a host name to either address type during the transition, but the IPv4 directory will disappear after all hosts in the world have migrated to IPv6.

3.3 UNICAST ROUTING PROTOCOLS

Hierarchical Routing in Internet

The Internet today is made of a huge number of networks and routers that connect them. Routing in the Internet cannot be done using a single protocol for two reasons:

A scalability problem and an administrative issue.

Scalability problem means that the size of the forwarding tables becomes huge, searching for a destination in a forwarding table becomes time-consuming, and updating creates a huge amount of traffic.

Hierarchical routing means considering each ISP as an autonomous system (AS). Each AS can run a routing protocol that meets its needs, but the global Internet runs a global protocol to join and connect all ASs together. The routing protocol run in each AS is referred to as intra-AS routing protocol, intradomain routing protocol, or interior gateway protocol (IGP); The global routing protocol is referred to as inter-AS routing protocol, interdomain routing protocol, or exterior gateway protocol (EGP).

Routing Information Protocol (RIP)

The Routing Information Protocol (RIP) is one of the most widely used intradomain routing protocols based on the distance-vector routing algorithm.

Hop Count

A router in this protocol implements the distance-vector routing algorithm. First, since a router in an AS needs to know how to forward a packet to different networks (subnets) in an AS, RIP routers advertise the cost of reaching different networks instead of reaching other nodes in a theoretical graph.

The cost is defined between a router and the network in which the destination host is located.

Second, to make the implementation of the cost simpler (independent from performance factors of the routers and links, such as delay, bandwidth, and so on), the cost is defined as the number of hops, which means the number of networks (subnets) a packet needs to travel through from the source router to the final destination host.

Note that the network in which the source host is connected is not counted in this calculation because the source host does not use a forwarding table; the packet is delivered to the default router.

Figure 3.3.1 shows the concept of hop count advertised by three routers from a source host to a destination host. In RIP, the maximum cost of a path can be 15, which means 16 is considered as infinity (no connection). For this reason, RIP can be used only in autonomous systems in which the diameter of the AS is not more than 15 hops.

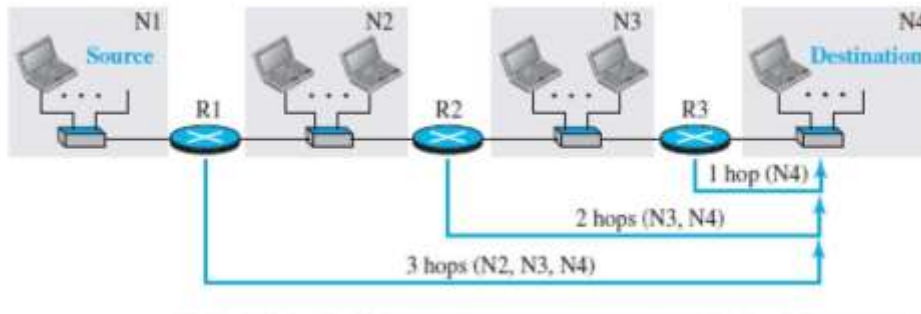


Fig3.3.1: Hop counts in RIP.

[Source : “Data Communications and Networking” by Behrouz A. Forouzan,Page-613]

Forwarding Table

A forwarding table in RIP is a three-column table in which the first column is the address of the destination network, the second column is the address of the next router to which the packet should be forwarded, and the third column is the cost (the number of hops) to reach the destination network. Figure 3.3.2 shows the three forwarding tables for the routers in Figure (above).

Note that the first and the third columns together convey the same information as does a distance vector, but the cost shows the number of hops to the destination networks.

Forwarding table for R1			Forwarding table for R2			Forwarding table for R3		
Destination network	Next router	Cost in hops	Destination network	Next router	Cost in hops	Destination network	Next router	Cost in hops
N1	—	1	N1	R1	2	N1	R2	3
N2	—	1	N2	—	1	N2	R2	2
N3	R2	2	N3	—	1	N3	—	1
N4	R2	3	N4	R3	2	N4	—	1

Fig3.3.2: Forwarding tables in RIP.

[Source : “Data Communications and Networking” by Behrouz A. Forouzan,Page-614]

For example, R1 defines that the next router for the path to N4 is R2; R2 defines that the next router to N4 is R3; R3 defines that there is no next router for this path. The tree is then R1 -R2- R3-N4.

The third column is not needed for forwarding the packet, but it is needed for updating the forwarding table when there is a change in the route.

RIP Implementation

RIP is implemented as a process that uses the service of UDP on the port number 520. RIP is a routing protocol to help IP route its datagrams through the AS, the RIP messages are encapsulated inside UDP user datagrams, which in turn are encapsulated inside IP datagrams.

That is, RIP runs at the application layer, but creates forwarding tables for IP at the network layer.

RIP Messages

Two RIP processes, a client and a server, need to exchange messages. RIP-2 defines the format of the message, as shown in figure 3.3.3. The message Entry, can be repeated as needed in a message. Each entry carries the information related to one line in the forwarding table of the router that sends the message.

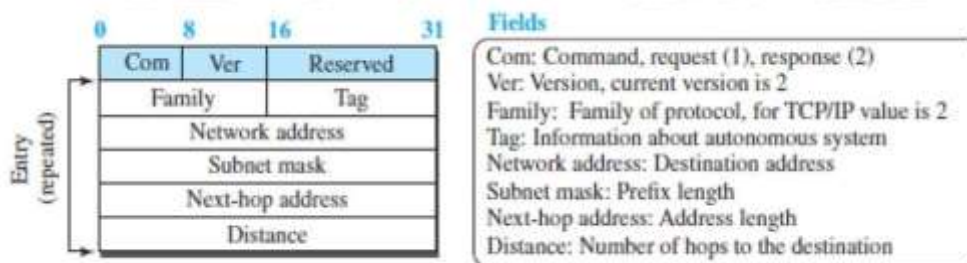


Fig3.3.3: RIP message format.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan, Page-615]

RIP has two types of messages:

Request and response. A request message is sent by a router that has just come up or by a router that has some time-out entries.

A request message can ask about specific entries or all entries.

A response (or update) message can be either solicited or unsolicited. A solicited response message is sent only in answer to a request message. It contains information about the destination specified in the corresponding request message.

RIP Algorithm

RIP implements the same algorithm as the distance-vector routing algorithm.

- ☐ Instead of sending only distance vectors, a router needs to send the whole contents of its forwarding table in a response message.
- ☐ The receiver adds one hop to each cost and changes the next router field to the address of the sending router.
- ☐ The received router selects the old routes as the new ones except in the following three cases:
 1. If the received route does not exist in the old forwarding table, it should be added to the route.
 2. If the cost of the received route is lower than the cost of the old one, the received route should be selected as the new one.
 3. If the cost of the received route is higher than the cost of the old one, but the value of the next router is the same in both routes, the received route should be selected as the new one.

Timers in RIP

RIP uses three timers to support its operation.

The periodic timer controls the advertising of regular update messages. Each router has one periodic timer that is randomly set to a number between 25 and 35 seconds (to prevent all routers sending their messages at the same time and creating excess traffic). The timer counts down; when zero is reached, the update message is sent, and the timer is randomly set once again.

The expiration timer governs the validity of a route. When a router receives update information for a route, the expiration timer is set to 180 seconds for that particular route. Every time a new update for the route is received, the timer is reset.

If there is a problem on an internet and no update is received within the allotted 180 seconds, the route is considered expired and the hop count of the route is set to 16, which means the destination is unreachable. Every route has its own expiration timer. The garbage collection timer is used to purge a route from the forwarding table.

The garbage collection timer is used to purge a route from the forwarding table. When the information about a route becomes invalid, the router does not immediately purge that route from its table. Instead, it continues to advertise the route with a metric value of 16. At the same time, a garbage collection timer is set to 120 seconds for that route. When the count reaches zero, the route is purged from the table.

Open Shortest Path First (OSPF)

Open Shortest Path First (OSPF) is an intradomain routing protocol like RIP. It is based on the link-state routing protocol.

Metric

In OSPF, like RIP, the cost of reaching a destination from the host is calculated from the source router to the destination network.

However, each link (network) can be assigned a weight based on the throughput, round-trip time, reliability, and so on

Figure 3.3.4 shows the idea of the cost from a router to the destination host network.

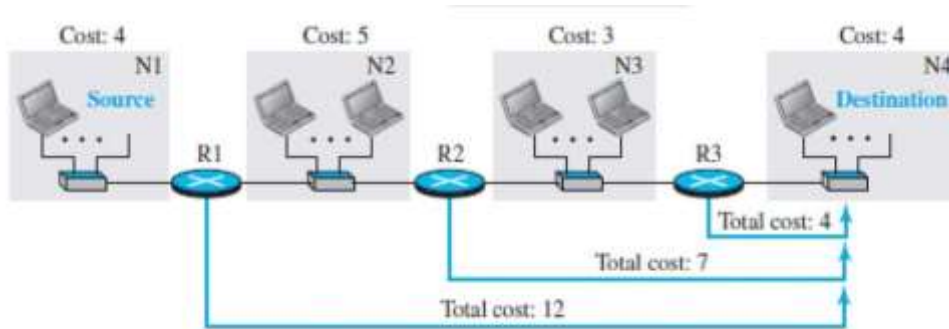


Fig3.3.4:Metric in OSPF.

[Source : “Data Communications and Networking” by Behrouz A. Forouzan,Page-618]

Forwarding Tables

Each OSPF router can create a forwarding table as in figure 3.3.5, after finding the shortest-path tree between itself and the destination using Dijkstra’s algorithm.

Forwarding table for R1			Forwarding table for R2			Forwarding table for R3		
Destination network	Next router	Cost	Destination network	Next router	Cost	Destination network	Next router	Cost
N1	—	4	N1	R1	9	N1	R2	12
N2	—	5	N2	—	5	N2	R2	8
N3	R2	8	N3	—	3	N3	—	3
N4	R2	12	N4	R3	7	N4	—	4

Fig3.3.5:Forwarding table in OSPF.

[Source : “Data Communications and Networking” by Behrouz A. Forouzan,Page-619]

Areas

OSPF was designed to handle routing in a small or large autonomous system.

The formation of shortest-path trees in OSPF requires that all routers flood the whole AS with their LSPs to create the global LSDB. This may not create a problem in a small AS, but create traffic in

large AS. To prevent this, the AS needs to be divided into small sections called areas as shown in figure 3.3.6 .

Each area acts as a small independent domain for flooding .Each router in an area needs to know the information about the link states not only in its area but also in other areas.

For this reason, one of the areas in the AS is designated as the backbone area, responsible for gluing the areas together.

The routers in the backbone area are responsible for passing the information collected by each area to all other areas. In this way, a router in an area can receive all LSPs generated in other areas. For the purpose of communication, each area has an area identification.

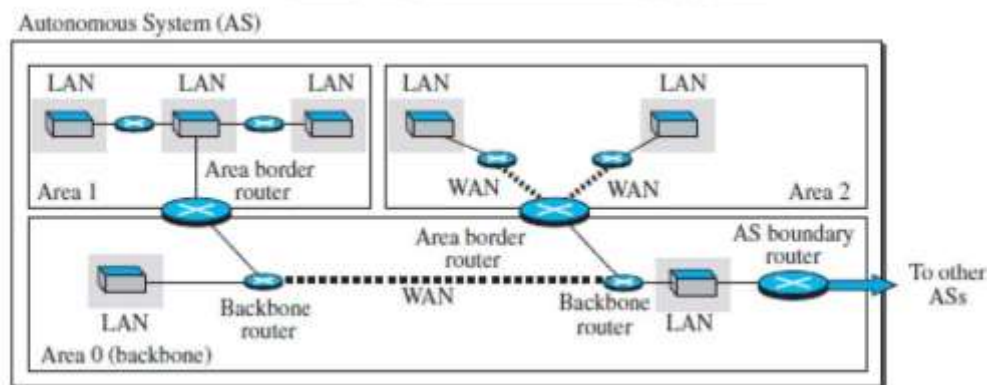


Fig3.3.6:Areas in AS.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-619]

OSPF Implementation

OSPF is implemented as a program in the network layer, using the service of the IP for propagation. An IP datagram that carries a message from OSPF sets the value of the protocol field to 89. This means that, the OSPF messages are encapsulated inside datagrams.

OSPF has two versions: version 1 and version 2.

OSPF Messages

OSPF is a very complex protocol; it has five different types of messages as shown in figure 3.3.7 .

.The hello message (type 1) is used by a router to introduce itself to the neighbors.

The database description message (type 2) is sent in response to the hello message to allow a newly joined router to acquire the full LSDB.The link state request message (type 3) is sent by a router that needs information about a specific LS.The link-state update message (type 4) is the main OSPF message used for building the LSDB. This message, has five different versions (router link, network

link, summary link to network, summary link to AS border router, and external link). The link-state acknowledgment message (type 5) is used to create reliability in OSPF; each router that receives a link-state update message needs to acknowledge it.

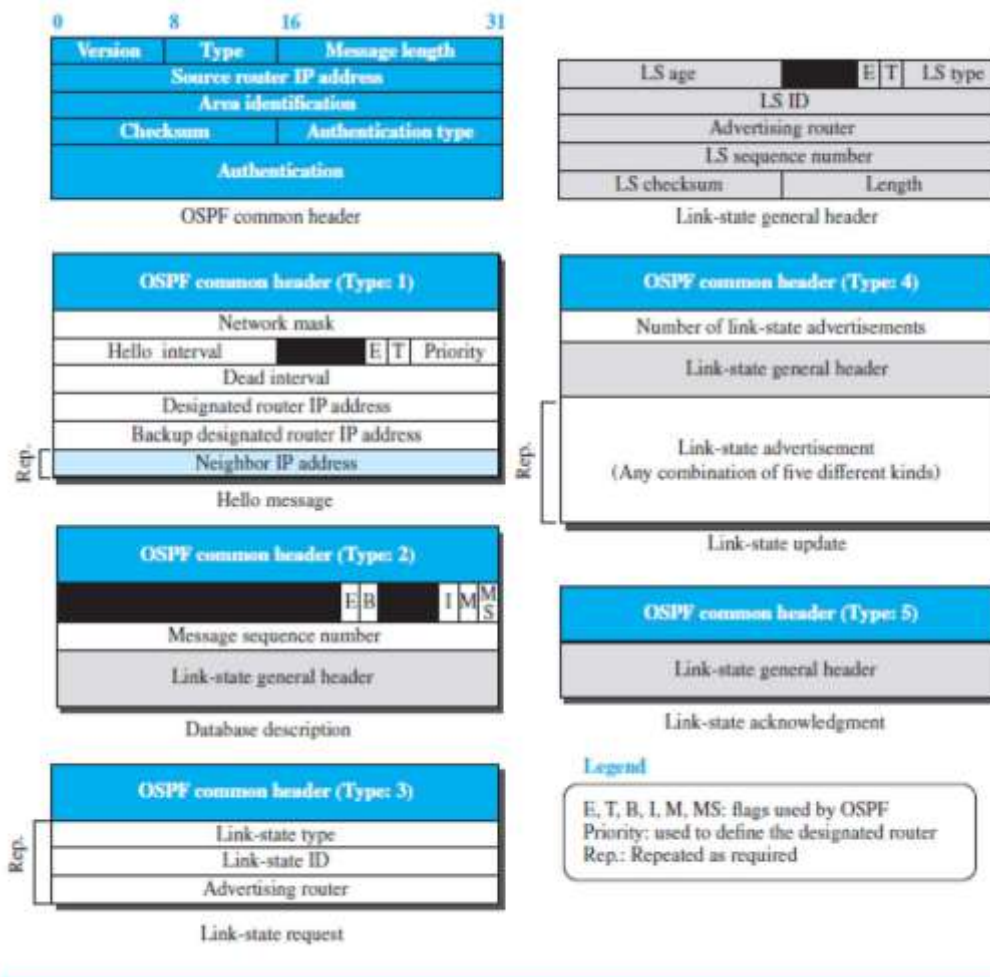


Fig3.3.7: OSPF message format.

[Source : “Data Communications and Networking” by Behrouz A. Forouzan, Page-622]

OSPF Algorithm

OSPF implements the link-state routing algorithm .

After each router has created the shortest-path tree, the algorithm needs to use it to create the corresponding routing algorithm. The algorithm needs to be augmented to handle sending and receiving all five types of messages.

Border Gateway Protocol Version 4 (BGP4)

The Border Gateway Protocol version 4 (BGP4) is the only inter domain routing protocol used in the Internet today.

Consider an example of an internet with four autonomous systems. AS2, AS3, and AS4 are stub autonomous systems; AS1 is a transient one as shown in figure 3.3.8 .

. Here, data exchange between AS2, AS3, and AS4 should pass through AS1.

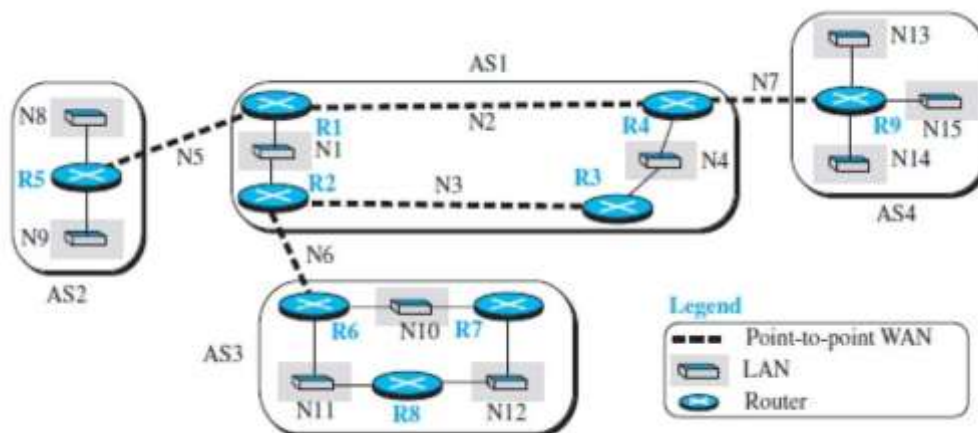


Fig3.3.8:Sample internet with four AS.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-623]

Each router in each AS knows how to reach a network that is in its own AS, but it does not know how to reach a network in another AS.

To enable each router to route a packet to any network in the internet, we first install a variation of BGP4, called external BGP (eBGP), on each border router (the one at the edge of each AS which is connected to a router at another AS). We then install the second variation of BGP, called internal BGP (iBGP), on all routers.

The border routers will be running three routing protocols (intradomain, eBGP, and iBGP), but other routers are running two protocols (intradomain and iBGP).

Operation of External BGP (eBGP)

BGP is a point-to-point protocol. When the software is installed on two routers, they try to create a TCP connection using the well-known port 179.

The two routers that run the BGP processes are called BGP peers or BGP speakers. The eBGP variation of BGP allows two physically connected border routers in two different ASs to form pairs of eBGP speakers and exchange messages.

The routers that we use has three pairs: R1-R5, R2-R6, and R4-R9. The connection between these pairs is established over three physical WANs (N5,N6, and N7). There is a need for a logical TCP connection to be created over the physical connection to make the exchange of information possible. Each logical connection in BGP is referred to as a session. This means that we need three sessions, as shown in Figure 3.3.9.

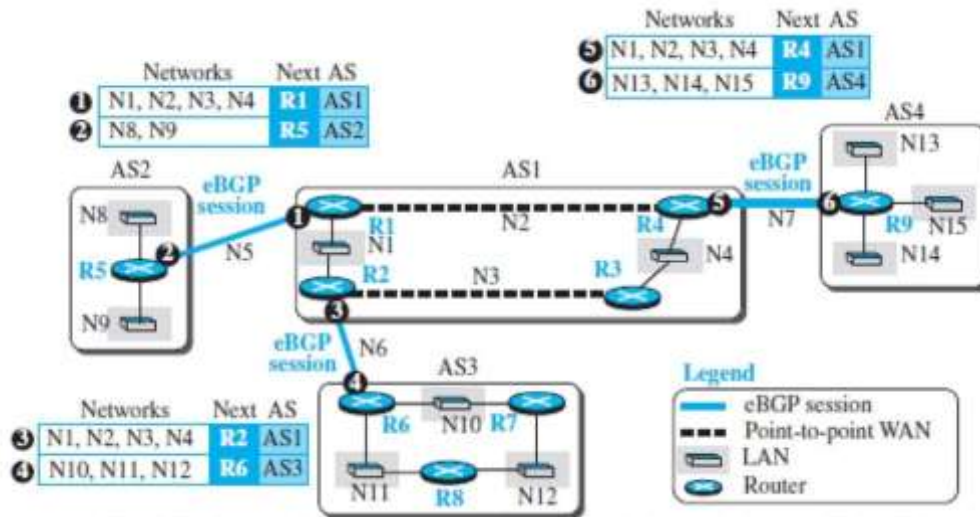


Fig3.3.9: EBGP operation.

[Source : "Data Communications and Networking" by Behrouz A. Forouzan,Page-624]

The circled number defines the sending router in each case.

For example, message number 1 is sent by router R1 and tells router R5 that N1, N2, N3, and N4 can be reached through router R1 (R1 gets this information from the corresponding intradomain forwarding table).

Router R5 can now add these pieces of information at the end of its forwarding table. When R5 receives any packet destined for these four networks, it can use its forwarding table and find that the next router is R1.

Messages

BGP has four types of messages for communication between the BGP speakers across the ASs and inside an AS:

Four messages are open, update, keep alive, and notification .

All BGP packets share the same common header.

Open Message. To create a neighborhood relationship, a router running BGP opens a TCP connection with a neighbor and sends an open message.

Update Message.

The update message is used by a router to withdraw destinations that have been advertised previously, to announce a route to a new destination, or both.

Note that BGP can withdraw several destinations that were advertised before, but it can only advertise one new destination in a single update message.

Keep alive Message. The BGP peers that are running exchange keep alive messages regularly (before their hold time expires) to tell each other that they are alive.

Notification. A notification message is sent by a router whenever an error condition is detected or a router wants to close the session.

Performance

BGP performance can be compared with RIP. BGP speakers exchange a lot of messages to create forwarding tables, but BGP is free from loops and count-to-infinity.

www.binils.com