# Introduction to IP Multicast Routing

**by Chuck Semeria and Tom Maufer**

## Abstract

The first part of this paper describes the benefits of multicasting, the Multicast Backbone (MBONE), Class D addressing, and the operation of the Internet Group Management Protocol (IGMP). The second section explores a number of different algorithms that may potentially be employed by multicast routing protocols:

- Flooding
- Spanning Trees
- Reverse Path Broadcasting (RPB)
- Truncated Reverse Path Broadcasting (TRPB)
- Reverse Path Multicasting (RPM)
- Core-Based Trees

The third part contains the main body of the paper. It describes how the previous algorithms are implemented in multicast routing protocols available today.

- Distance Vector Multicast Routing Protocol (DVMRP)
- Multicast OSPF (MOSPF)
- Protocol-Independent Multicast (PIM)

## Introduction

There are three fundamental types of IPv4 addresses: unicast, broadcast, and multicast. A unicast address is designed to transmit a packet to a single destination. A broadcast address is used to send a datagram to an entire subnetwork. A multicast address is designed to enable the delivery of datagrams to a set of hosts that have been configured as members of a multicast group in various scattered subnetworks.

Multicasting is not connection oriented. A multicast datagram is delivered to destination group members with the same “best-effort” reliability as a standard unicast IP datagram. This means that a multicast datagram is not guaranteed to reach all members of the group, or arrive in the same order relative to the transmission of other packets.

The only difference between a multicast IP packet and a unicast IP packet is the presence of a “group address” in the Destination Address field of the IP header. Instead of a Class A, B, or C IP address, multicasting employs a Class D destination address format (224.0.0.0- 239.255.255.255).

### Multicast Groups

Individual hosts are free to join or leave a multicast group at any time. There are no restrictions on the physical location or the number of members in a multicast group. A host may be a member of more than one multicast group at any given time and does not have to belong to a group to send messages to members of a group.

### Group Membership Protocol

A group membership protocol is employed by routers to learn about the presence of group members on their directly attached subnetworks. When a host joins a multicast group, it transmits a group membership protocol message for the group(s) that it wishes to receive, and sets its IP process and network interface card to receive frames addressed to the multicast group. This receiver-initiated join process has excellent scaling properties since, as the multicast group increases in size, it becomes ever more likely that a new group member will be able to locate a nearby branch of the multicast distribution tree.

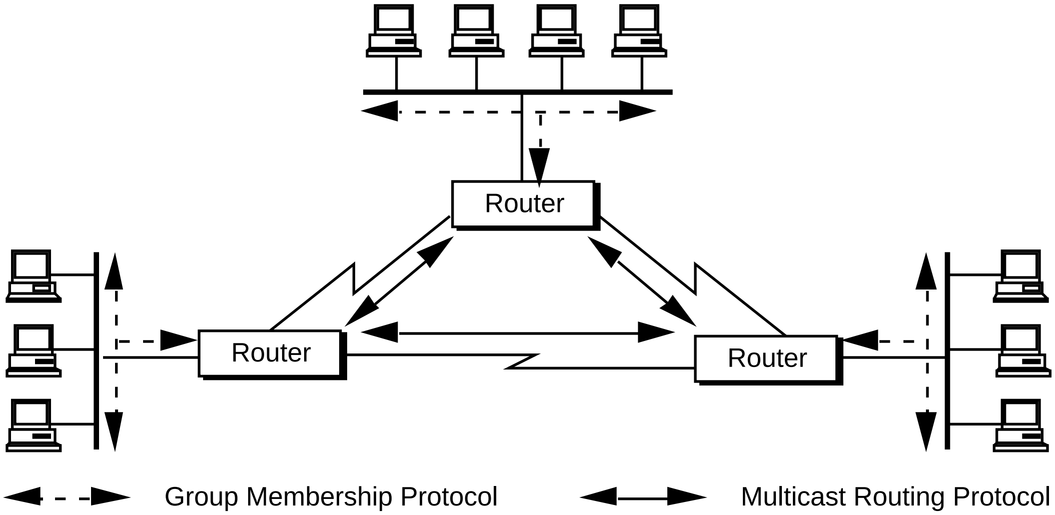

**Figure 1: Multicast IP Delivery Service**

### Multicast Routing Protocol

Multicast routers execute a multicast routing protocol to define delivery paths that enable the forwarding of multicast datagrams across an internetwork. The Distance Vector Multicast Routing Protocol (DVMRP) is a distance-vector routing protocol, and Multicast OSPF (MOSPF) is an extension to the OSPF link-state unicast routing protocol.

## Multicast Support for Emerging Internet Applications

Today, the majority of Internet applications rely on point-to-point transmission. The utilization of point-to-multipoint transmission has traditionally been limited to local area network applications. Over the past few years the Internet has seen a rise in the number of new applications that rely on multicast transmission. Multicast IP conserves bandwidth by forcing the network to do packet replication only when necessary, and offers an attractive alternative to unicast transmission for the delivery of network ticker tapes, live stock quotes, multiparty video-conferencing, and shared whiteboard applications (among others). It is important to note that the applications for IP Multicast are not solely limited to the Internet. Multicast IP can also play an important role in large distributed commercial networks.

### Reducing Network Load

Assume that a stock ticker application is required to transmit packets to 100 stations within an organization’s network. Unicast transmission to the group of stations will require the periodic transmission of 100 packets where many packets may be required to traverse the same link(s). Multicast transmission is the ideal solution for this type of application since it requires only a single packet transmission by the source which is then replicated at forks in the multicast delivery tree.

Broadcast transmission is not an effective solution for this type of application since it affects the CPU performance of each and every end station that sees the packet and it wastes bandwidth.

### Resource Discovery

Some applications implement multicast group addresses instead of broadcasts to transmit packets to group members residing on the same network. However, there is no reason to limit the extent of a multicast transmission to a single LAN. The time-to-live (TTL) field in the IP header can be used to limit the range (or “scope”) of a multicast transmission.

### Support for Datacasting Applications

Since 1992 the Internet Engineering Task Force (IETF) has conducted a series of “audiocast” experiments in which live audio and video were multicast from the IETF meeting site to destinations around the world. Datacasting takes compressed audio and video signals from the source station and transmits them as a sequence of UDP packets to a group address. Multicasting might eliminate an organization’s need to maintain parallel networks for voice, video, and data.

## The Internet’s Multicast Backbone (MBONE)

The Internet Multicast Backbone (MBONE) is an interconnected set of subnetworks and routers that support the delivery of IP multicast traffic. The goal of the MBONE is to construct a semipermanent IP multicast testbed to enable the deployment of multicast applications without waiting for the ubiquitous deployment of multicast-capable routers in the Internet.

The MBONE has grown from 40 subnets in four different countries in 1992 to more than 2800 subnets in over 25 countries by April 1996. With new multicast applications and multicast-based services appearing, it appears likely that the use of multicast technology in the Internet will continue to grow at an ever-increasing rate.

The MBONE is a virtual network that is layered on top of sections of the physical Internet. It is composed of islands of multicast routing capability connected to other islands by virtual point-to-point links called “tunnels.” The tunnels allow multicast traffic to pass through the non-multicast-capable parts of the Internet. IP multicast packets are encapsulated as IP-over-IP (i.e., the protocol number is set to 4), so they look like normal unicast packets to intervening routers. The encapsulation is added on entry to a tunnel and stripped off on exit from a tunnel. This set of multicast routers, their directly connected subnetworks, and the interconnecting tunnels define the MBONE.

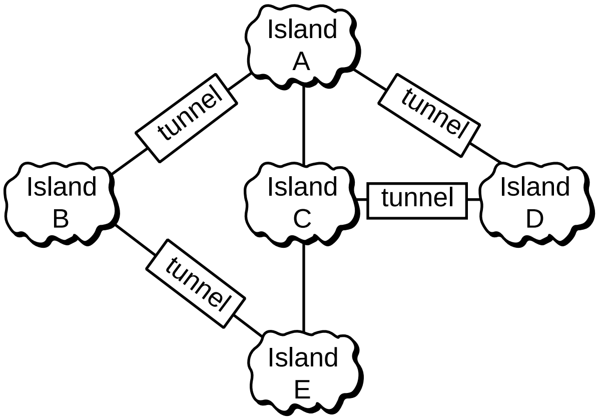

**Figure 2: Internet Multicast Backbone (MBONE)**

Since the MBONE and the Internet have different topologies, multicast routers execute a separate routing protocol to decide how to forward multicast packets. The majority of the MBONE routers currently use the Distance Vector Multicast Routing Protocol (DVMRP), although some portions of the MBONE execute either Multicast OSPF (MOSPF) or the Protocol-Independent Multicast (PIM) routing protocols. The operation of each of these protocols is discussed later in this paper.

The MBONE carries audio and video multicasts of IETF meetings, NASA space shuttle missions, U.S. House and Senate sessions, and live satellite weather photos. The Session Directory (SD) tool provides users with a listing of the active multicast sessions on the MBONE and allows them to define or join a conference.

# Multicast Addressing

A multicast address is assigned to a set of receivers defining a multicast group. Senders use the multicast address as the destination IP address of a packet that is to be transmitted to all group members.

## Class D Addresses

An IP multicast group is identified by a Class D address. Class D addresses have their high-order four bits set to “1110” followed by a 28-bit multicast group ID. Expressed in standard “dotted-decimal” notation, multicast group addresses range from 224.0.0.0 to 239.255.255.255. Figure 3 shows the format of a 32- bit Class D address.

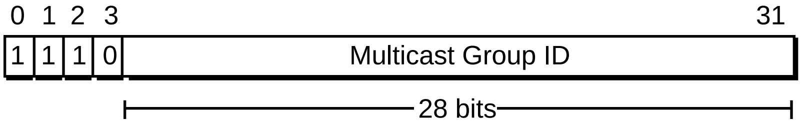

**Figure 3: Class D Multicast Address Format**

The Internet Assigned Numbers Authority (IANA) maintains a list of registered IP multicast groups. The base address 224.0.0.0 is reserved and cannot be assigned to any group. The block of multicast addresses ranging from 224.0.0.1 to 224.0.0.255 is reserved for the use of routing protocols and other low-level topology discovery or maintenance protocols. Multicast routers should not forward a multicast datagram with a destination address in this range, regardless of its TTL.

The remaining groups ranging from 224.0.1.0 to 239.255.255.255 are assigned to various multicast applications or remain unassigned. From this range, 239.0.0.0 to 239.255.255.255 are to be reserved for site-local “administratively scoped” applications, not Internet-wide applications. Some of the well-known groups include: “all systems on this subnet” (224.0.0.1), “all routers on this subnet”(224.0.0.2), “all DVMRP routers” (224.0.0.4), “all OSPF routers” (224.0.0.5), “IETF-1-Audio” (224.0.1.11), “IETF-1-Video” (224.0.1.12), “AUDIONEWS” (224.0.1.7), and “MUSIC- SERVICE” (224.0.1.16).

## Mapping a Class D Address to an Ethernet Address

The IANA has been allocated a reserved portion of the IEEE-802 MAC-layer multicast address space. All of the addresses in IANA’s reserved block begin with 01-00-5E (hex). A simple procedure was developed to map Class D addresses to this reserved address block. This allows IP multicasting to take advantage of the hardware-level multicasting supported by network interface cards.

For example, the mapping between a Class D IP address and an Ethernet multicast address is obtained by placing the low-order 23 bits of the Class D address into the low-order 23 bits of IANA’s reserved address block.

Figure 4 illustrates how the multicast group address 224.10.8.5 (E0-0A-08-05) is mapped into an Ethernet (IEEE-802) multicast address.

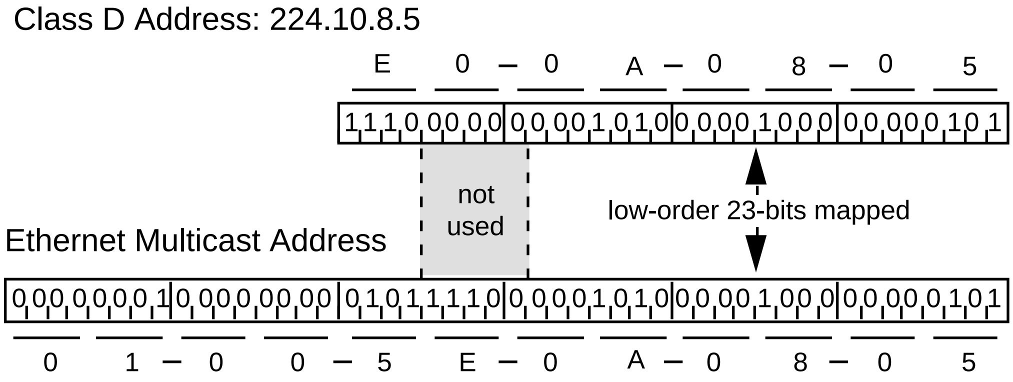

**Figure 4: Mapping Between Class D and IEEE-802 Multicast Addresses**

The mapping in Figure 4 places the low-order 23 bits of the IP multicast group ID into the low order 23 bits of the Ethernet address. You should note that the mapping may place up to 32 different IP groups into the same Ethernet address because the upper five bits of the IP multicast group ID are ignored. For example, the multicast addresses 224.138.8.5 (E0-8A-08-05) and 225.10.8.5 (E1-0A-08-05) would also be mapped to the same Ethernet address (01-00-5E-0A-08-05) used in this example.

### Transmission and Delivery of Multicast Datagrams

When the sender and receivers are members of the same (LAN) subnetwork, the transmission and reception of multicast frames are relatively simple processes. The source station simply addresses the IP packet to the multicast group, the network interface card maps the Class D address to the corresponding IEEE-802 multicast address, and the frame is sent. Receivers that wish to capture the frame notify their IP layer that they want to receive datagrams addressed to the group.

Things become much more complicated when the sender is attached to one subnetwork and receivers reside on different subnetworks. In this case, the routers are required to implement a multicast routing protocol that permits the construction of multicast delivery trees and supports multicast data packet forwarding. In addition, each router needs to implement a group membership protocol that allows it to learn about the existence of group members on its directly attached subnetworks.

## Internet Group Management Protocol (IGMP)

The Internet Group Management Protocol (IGMP) runs between hosts and their immediately neighboring multicast routers. The mechanisms of the protocol allow a host to inform its local router that it wishes to receive transmissions addressed to a specific multicast group. Also, routers periodically query the LAN to determine if known group members are still active. If there is more than one router on the LAN performing IP multicasting, one of the routers is elected “querier” and assumes the responsibility of querying the LAN for group members.

Based on the group membership information learned from the IGMP, a router is able to determine which (if any) multicast traffic needs to be forwarded to each of its “leaf” subnetworks. Multicast routers use this information, in conjunction with a multicast routing protocol, to support IP multicasting across the Internet.

## IGMP Version 1

IGMP Version 1 was specified in RFC-1112. According to the specification, multicast routers periodically transmit Host Membership Query messages to determine which host groups have members on their directly attached networks. Query messages are addressed to the all-hosts group (224.0.0.1) and have an IP TTL = 1. This means that Query messages sourced by a router are transmitted onto the directly attached subnetwork but are not forwarded by any other multicast router.

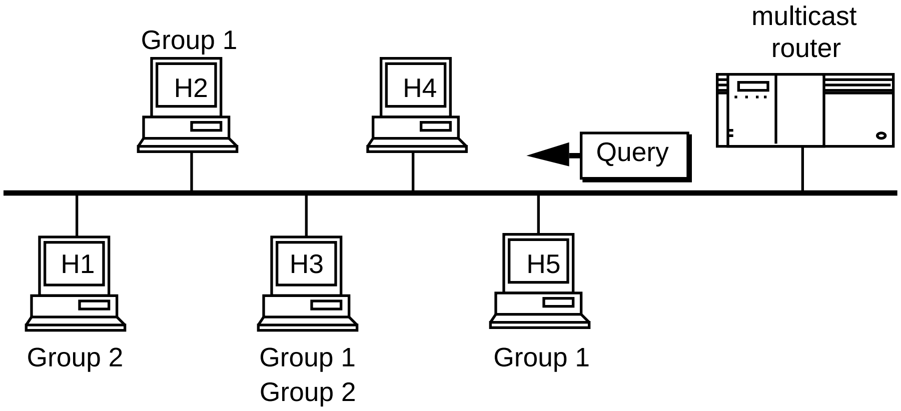

**Figure 5: Internet Group Management Protocol—Query Message**

When a host receives a Query message, it responds with a Host Membership Report for each host group to which it belongs. In order to avoid a flurry of Reports, each host starts a randomly chosen Report delay timer for each of its group memberships. If, during the delay period, another Report is heard for the same group, the local host resets its timer to a new random value. Otherwise, the host transmits a Report to the reported group address, causing all other members of the group to reset their Report message timers. This procedure guarantees that Reports are spread out over a period of time and that Report traffic is minimized for each group with at least one member on the subnetwork.

It should be noted that multicast routers do not need to be directly addressed since their interfaces are configured to receive all multicast IP traffic. Also, a router does not need to maintain a detailed list of which hosts belong to each multicast group; the router only needs to know that at least one group member is present on a network interface.

Multicast routers periodically transmit Queries to update their knowledge of the group members present on each network interface.

If the router does not receive a Report for a particular group after a number of Queries, the router assumes that group members are no longer present on the interface and the group is removed from the list of group memberships for the directly attached subnetwork.

When a host first joins a group, it immediately transmits a Report for the group rather than waiting for a router Query. This guarantees that the host will receive traffic addressed to the group if it is the first member of that group on the subnetwork.

## IGMP Version 2

IGMP Version 2 was distributed as part of the IP Multicasting (Version 3.3 through Version 3.8) code package. Initially, there was no detailed specification for IGMP Version 2 other than the source code. However, the complete specification has recently been published in <draft-ietf-idmr-igmp-v2-01.txt>, which will replace the informal specification contained in Appendix I of RFC-1112. IGMP Version 2 enhances and extends IGMP Version 1 while also providing backward compatibility with Version 1 hosts.

IGMP Version 2 defines a procedure for the election of the multicast querier for each LAN. In IGMP Version 2, the router with the lowest IP address on the LAN is elected the multicast querier. In IGMP Version 1, the querier election was determined by the multicast routing protocol. This could lead to potential problems because each multicast routing protocol used different methods for determining the multicast querier.

IGMP Version 2 defines a new type of Query message—the Group-Specific Query message. Group-Specific Query messages allow a router to transmit a Query to a specific multicast group rather than all groups residing on a directly attached subnetwork.

Finally, IGMP Version 2 defines a Leave Group message to lower IGMP's "leave latency." When the last host to respond to a Query with a Report wishes to leave that specific group, the host transmits a Leave Group message to the all-routers group (224.0.0.2) with the group field set to the group to be left. In response to a Leave Group message, the router begins the transmission of Group-Specific Query messages on the interface that received the Leave Group message. If there are no Reports in response to the Group-Specific Query messages, the group is removed from the list of group memberships for the directly attached subnetwork.

## IGMP Version 3

IGMP Version 3 is a preliminary draft specification published in <draft-cain-igmp-00.txt>. IGMP Version 3 introduces support for Group-Source Report messages so that a host can elect to receive traffic from specific sources of a multicast group. An Inclusion Group-Source Report message allows a host to specify the IP addresses of the specific sources it wants to receive. An Exclusion Group-Source Report message allows a host to explicitly identify the sources that it does not want to receive. With IGMP Version 1 and Version 2, if a host wants to receive any sources from a group, the traffic from all sources for the group has to be forwarded onto the subnetwork.

IGMP Version 3 will help conserve bandwidth by allowing a host to select the specific sources from which it wants to receive traffic. Also, multicast routing protocols will be able to make use of this information to conserve bandwidth when constructing the branches of their multicast delivery trees.

Finally, support for Leave Group messages first introduced in IGMP Version 2 has been enhanced to support Group-Source Leave messages. This feature allows a host to leave an entire group or to specify the specific IP address(es) of the (source, group) pair(s) that it wishes to leave.

## Multicast Forwarding Algorithms

IGMP provides the final step in a multicast packet delivery service since it is only concerned with the forwarding of multicast traffic from the local router to group members on directly attached subnetworks. IGMP is not concerned with the delivery of multicast packets between neighboring routers or across an internetwork.
To provide an Internet-wide delivery service, it is necessary to define multicast routing protocols. A multicast routing protocol is responsible for the construction of multicast packet delivery trees and performing multicast packet forwarding. This section explores a number of different algorithms that may potentially be employed by multicast routing protocols:

- Flooding
- Spanning Trees
- Reverse Path Broadcasting (RPB)
- Truncated Reverse Path Broadcasting (TRPB)
- Reverse Path Multicasting (RPM)
- Core-Based Trees

Later sections will describe how these algorithms are implemented in the most prevalent multicast routing protocols in the Internet today.

- Distance Vector Multicast Routing Protocol (DVMRP)
- Multicast OSPF (MOSPF)
- Protocol-Independent Multicast (PIM)

### Flooding

The simplest technique for delivering multicast datagrams to all routers in an internetwork is to implement a flooding algorithm. The flooding procedure begins when a router receives a packet that is addressed to a multicast group. The router employs a protocol mechanism to determine whether this is the first time that it has seen this particular packet or whether it has seen the packet before. If it is the first reception of the packet, the packet is forwarded on all interfaces except the one on which it arrived, guaranteeing that the multicast packet reaches all routers in the internetwork. If the router has seen the packet before, it is simply discarded.

A flooding algorithm is very simple to implement since a router does not have to maintain a routing table and only needs to keep track of the most recently seen packets. However, flooding does not scale for Internet-wide applications since it generates a large number of duplicate packets and uses all available paths across the internetwork instead of just a limited number. Also, the flooding algorithm makes inefficient use of router memory resources since each router is required to maintain a distinct table entry for each recently seen packet.

## Spanning Tree

A more effective solution than flooding would be to select a subset of the Internet topology that forms a spanning tree. The spanning tree defines a tree structure where only one active path connects any two routers on the Internet. Figure 6 shows an internetwork and a spanning tree rooted at router RR.

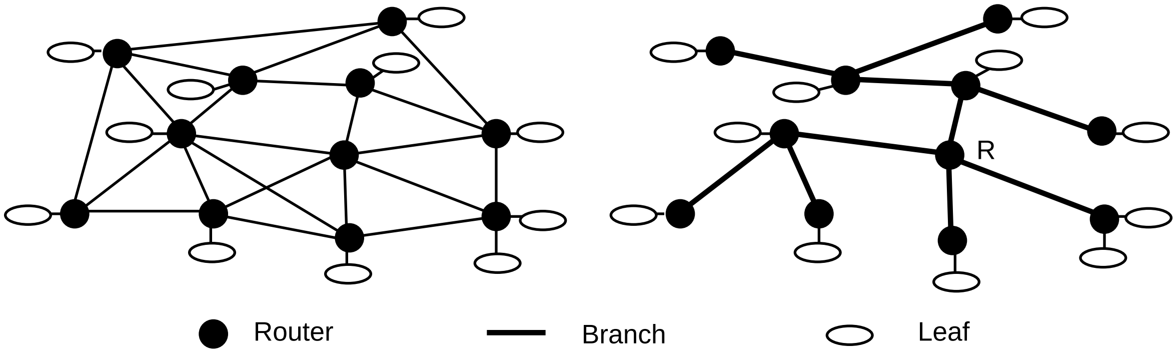

**Figure 6: Spanning Tree**

Once the spanning tree has been built, a multicast router simply forwards each multicast packet to all interfaces that are part of the spanning tree except the one on which the packet originally arrived. Forwarding along the branches of a spanning tree guarantees that the multicast packet will not loop and that it will eventually reach all routers in the internetwork.

A spanning tree solution is powerful and is relatively easy to implement since there is a great deal of experience with spanning tree protocols in the Internet community. However, a spanning tree solution can centralize traffic on a small number of links, and may not provide the most efficient path between the source subnetwork and group members.

## Reverse Path Broadcasting (RPB)

An even more efficient solution than building a single spanning tree for the entire Internet would be to build a group-specific spanning tree for each potential source subnetwork. These spanning trees would result in source-rooted delivery trees emanating from the subnetwork directly connected to the source station. Since there are many potential sources for a group, a different spanning tree is constructed for each active (source, group) pair.

### Operation

The fundamental algorithm to construct these source-rooted trees is referred to as Reverse Path Broadcasting (RPB). The RPB algorithm is actually quite simple. For each (source, group) pair, if a packet arrives on a link that the local router considers to be the shortest path back to the source of the packet, then the router forwards the packet on all interfaces except the incoming interface. Otherwise, the packet is discarded. The interface over which the router expects to receive multicast packets from a particular source is referred to as the “parent” link. The outbound links over which the router forwards the multicast packet are called the “child” links.

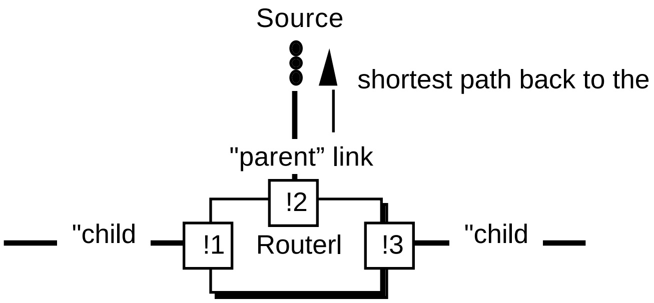

**Figure 7: Reverse Path Broadcasting—Forwarding Algorithm**

This basic algorithm can be enhanced to reduce unnecessary packet duplication if the router making the forwarding decision can determine whether a neighboring router on a potential child link considers the local router to be on its shortest path back to the source [i.e., on the remote router’s parent link for the (source, group) pair]. If this is the case, the packet is forwarded to the neighbor. Otherwise, the datagram is not forwarded on the potential child link because the neighboring router will be required to discard the packet since it arrives on a non-parent link for the (source, group) pair.

The information needed to make this “downstream” decision is relatively easy to derive from a link-state routing protocol since each router maintains a topological database for the entire routing domain. If a distance-vector routing protocol is employed, a neighbor can either advertise its previous hop for the (source, group) pair as part of its routing update messages or “poison reverse” the route. Either of these techniques allows an upstream router to determine if a downstream neighboring router considers it to be on the downstream router’s shortest path back to the source.

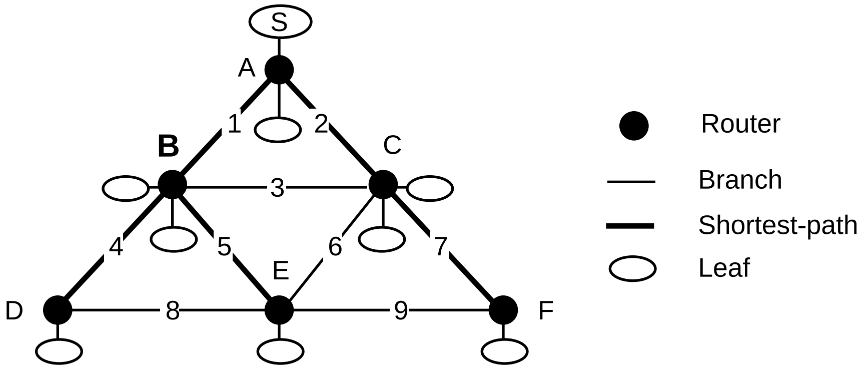

**Figure 8: Example of Reverse Path Broadcasting**

Figure 8 shows an example of  the basic operation of the enhanced RPB algorithm. Note that the source station (S) is attached to a leaf subnetwork directly connected to Router A. For this example, we will look at the RPB algorithm from Router B’s perspective. Router B receives the multicast packet from Router A on link 1. Since Router B considers link 1 to be the parent link for the (source, group) pair, it forwards the packet on link 4, link 5, and the local leaf subnetworks if they have group members. Router B does not forward the packet on link 3 because it knows from routing protocol exchanges that Router C considers link 2 as its parent link for the (source, group) pair. If Router B were to forward the packet on link 3 it would be discarded by Router C since it would arrive on a non-parent link for the (source, group) pair.

### Benefits and Limitations

The key benefit to reverse path broadcasting is that it is reasonably efficient and easy to implement. It does not require that the router know about the entire spanning tree nor does it require a special mechanism to stop the forwarding process, as flooding does. In addition, it guarantees efficient delivery since multicast packets always follow the “shortest” path from the source station to the destination group. Finally, the packets are distributed over multiple links, resulting in better network utilization since a different tree is computed for each (source, group) pair.

One of the major limitations of the RPB algorithm is that it does not take into account multicast group membership when building the distribution tree for a (source, group) pair. As a result, datagrams may be unnecessarily forwarded to subnetworks that have no members in the destination group.

## Truncated Reverse Path Broadcasting (TRPB)

Truncated Reverse Path Broadcasting (TRPB) was developed to overcome the limitations of Reverse Path Broadcasting. With the help of IGMP, multicast routers determine the group memberships on each leaf subnetwork and avoid forwarding datagrams onto a leaf subnetwork if it does not have a member of the destination group present. The spanning delivery tree is “truncated” by the router if a leaf subnetwork does not have group members.

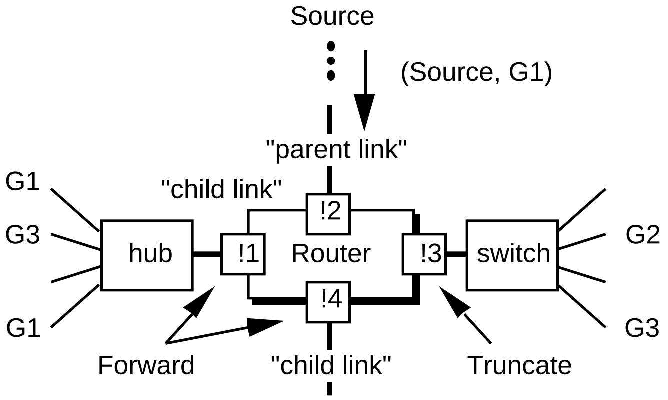

**Figure 9: Truncated Reverse Path Broadcasting (TRPB)**

Figure 9 illustrates the operation of the TRPB algorithm. In this example the router receives a multicast packet on its parent link for the (Source, G1) pair. The router forwards the datagram on !1 since the interface has at least one member of G1. The router does not forward the datagram to !3 since this interface has no members in the destination group. The datagram is forwarded on !4 if and only if a downstream router considers the interface as part of its parent link for the (Source, G1) pair.

TRPB removes some limitations of RPB, but it solves only part of the problem. It eliminates unnecessary traffic on leaf subnetworks but it does not consider group memberships when building the branches of the distribution tree.

## Reverse Path Multicasting (RPM)

Reverse Path Multicasting (RPM) is an enhancement to Reverse Path Broadcasting and Truncated Reverse Path Broadcasting. RPM creates a delivery tree that spans only:

-Subnetworks with group members, and
-Routers and subnetworks along the shortest path to subnetworks with group members

RPM allows the source-rooted spanning tree to be pruned so that datagrams are only forwarded along branches that lead to members of the destination group.

### Operation

When a multicast router receives a packet for a (source, group) pair, the first packet is forwarded following the TRPB algorithm to all routers in the internetwork. Routers that are at the edge of the network and have no further downstream routers in the TRPB tree are called leaf routers. The TRPB algorithm guarantees that each leaf router receives the first multicast packet. If there is a group member on one of its leaf subnetworks, a leaf router forwards the packet based on its IGMP information. If none of the subnetworks connected to the leaf router have group members, the leaf router may transmit a “prune” message on its parent link informing the upstream router that it should not forward packets for the particular (source, group) pair on the child interface receiving the prune message. Prune messages are sent only one hop back toward the source.

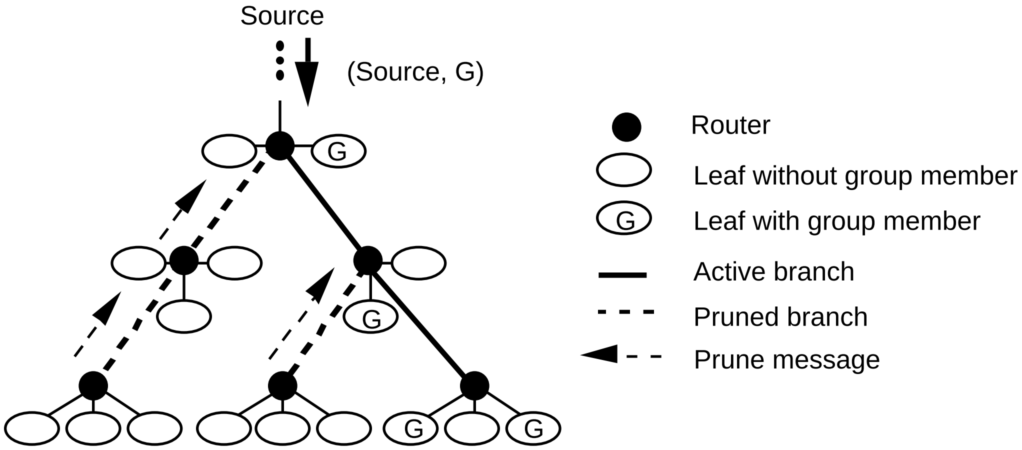

**Figure 10: Reverse Path Multicasting (RPM)**

An upstream router receiving a prune message is required to record the prune information in memory. If the upstream router has no local recipient and receives prune messages on each of the child interfaces that it used to forward the first packet, the upstream router does not need to receive additional packets for the (source, group) pair.

This means that the upstream router can generate a prune message of its own, one hop back toward the source. This succession of prune messages creates a multicast forwarding tree that contains only “live” branches (i.e., branches that lead to active group members).

Since both the group membership and network topology are dynamically changing, the pruned state of the multicast forwarding tree must be refreshed at regular intervals. Periodically, the prune information is removed from the memory of all routers and the next packet for the (source, group) pair is forwarded to all leaf routers. This results in a new burst of prune messages, allowing the multicast forwarding tree to adapt to the ever-changing multicast delivery requirements of the internetwork.

### Limitations

Despite the improvements offered by the RPM algorithm, there are still several scaling issues that need to be addressed when attempting to develop an Internet-wide delivery service. The first limitation is that multicast packets must be periodically forwarded to every router in the internetwork. The second drawback is that each router is required to maintain state information for all groups and each source. The significance of these shortcomings is amplified as the number of sources and groups in the multicast internetwork expands.

## Core-Based Trees (CBT)

The latest addition to the existing set of multicast forwarding algorithms is Core Based Trees (CBT). Unlike existing algorithms which build a source-rooted, shortest-path tree for each (source, group) pair, CBT constructs a single delivery tree that is shared by all members of a group. The CBT algorithm is quite similar to the spanning tree algorithm except it allows a different core-based tree for each group. Multicast traffic for each group is sent and received over the same delivery tree, regardless of the source.

### Operation

A core-based tree may involve a single router or set of routers, which acts as the core of a multicast delivery tree. Figure 11 illustrates how multicast traffic is forwarded across a CBT "backbone" to all members of the group. Note that the CBT backbone may contain both core and non-core routers.

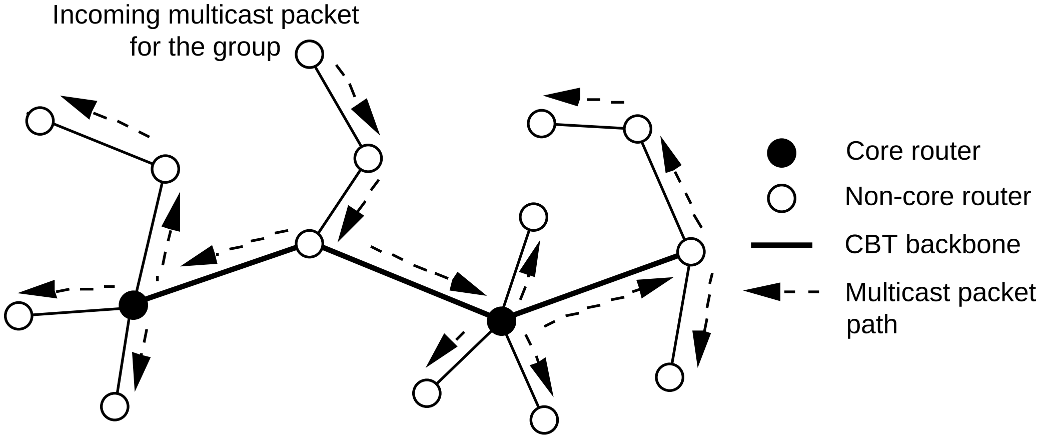

**Figure 11: Multi-Core CBT Multicast Delivery Tree**

Each station that wishes to receive traffic that has been addressed to a multicast group is required to send a "join" message toward the "core tree" of the particular multicast group. A potential group member only needs to know the address of one of the group's core routers in order to transmit a unicast join request. The join request is processed by all intermediate routers that identify the interface on which the join was received as belonging to the group's delivery tree. The intermediate routers continue to forward the join message toward the core and marking local interfaces until the request reaches a core router.

Similar to other multicast forwarding algorithms, CBT does not require that the source of a multicast packet be a member of the destination group. Packets sourced by a non-group member are simply unicast toward the core until they reach the first router that is a member of the group's delivery tree. When the unicast packet reaches a member of the delivery tree, the packet is multicast to all outgoing interfaces that are part of the tree except the incoming link. This guarantees that the multicast packet is forwarded to all routers on the delivery tree.

### Benefits

In terms of scalability, CBT has several advantages over the Reverse Path Multicasting (RPM) algorithm. CBT makes efficient use of router resources since it only requires a router to maintain state information for each group, not for each (source, group) pair. Also, CBT conserves network bandwidth since it does not require that multicast frames be periodically forwarded to all multicast routers in the internetwork.

### Limitations

Despite these benefits, there are still several limitations to the CBT approach. CBT may result in traffic concentration and bottlenecks near core routers since traffic from all sources traverses the same set of links as it approaches the core. In addition, a single shared delivery tree may create suboptimal routes resulting in increased delay—a critical issue for some multimedia applications. Finally, new algorithms still need to be developed to support core management which encompasses all aspects of core router selection and (potentially) dynamic placement strategies.

# Distance Vector Multicast Routing Protocol (DVMRP)

The Distance Vector Multicast Routing Protocol (DVMRP) is a distance-vector routing protocol designed to support the forwarding of multicast datagrams through an internetwork. DVMRP constructs source-rooted multicast delivery trees using variants of the Reverse Path Broadcasting (RPB) algorithm. Some version of DVMRP is currently deployed in the majority of MBONE routers.

DVMRP was first defined in RFC-1075. The original specification was derived from the Routing Information Protocol (RIP) and employed the Truncated Reverse Path Broadcasting (TRPB) algorithm. The major difference between RIP and DVMRP is that RIP is concerned with calculating the next hop to a destination, while DVMRP is concerned with computing the previous hop back to a source. It is important to note that the latest mrouted version 3.8 and vendor implementations have extended DVMRP to employ the Reverse Path Multicasting (RPM) algorithm. This means that the latest implementations of DVMRP are quite different from the original RFC specification in many regards.

## Physical and Tunnel Interfaces

The ports of a DVMRP router may be either a physical interface to a directly attached subnetwork or a tunnel interface to another multicast island. All interfaces are configured with a metric that specifies the cost for the given port and a TTL threshold that limits the scope of a multicast transmission. In addition, each tunnel interface must be explicitly configured with two additional parameters—the IP address of the local router’s interface and the IP address of the remote router’s interface.

**Table 1 TTL Scope Control Values**

| Initial TTL | Scope |
|---|---|
| 0 | Restricted to the same host |
| 1 | Restricted to the same subnetwork |
| 32 | Restricted to the same site |
| 64 | Restricted to the same region |
| 128 | Restricted to the same continent |
| 255 | Unrestricted in scope |

A multicast router will only forward a multicast datagram across an interface if the TTL field in the IP header is greater than the TTL threshold assigned to the interface. Table 1 lists the conventional TTL values that are used to restrict the scope of an IP multicast.

For example, a multicast datagram with a TTL of less than 32 is restricted to the same site and should not be forwarded across an interface to other sites in the same region.

## Basic Operation

DVMRP implements the Reverse Path Multicasting (RPM) algorithm. According to RPM, the first datagram for any (source, group) pair is forwarded across the entire internetwork, providing that the packet's TTL and router interface thresholds permit. The initial datagram is delivered to all leaf routers, which transmit prune messages back toward the source if there are no group members on their directly attached leaf subnetworks. The prune messages result in the removal of branches from the tree that do not lead to group members, thus creating a source-specific shortest path tree with all leaves having group members. After a period of time, the pruned branches grow back and the next datagram for the (source, group) pair is forwarded across the entire internetwork, resulting in a new set of prune messages.

DVMRP implements a mechanism to quickly "graft" back a previously pruned branch of a group's delivery tree. If a router that previously sent a prune message for a (source, group) pair discovers new group members on a leaf network, it sends a graft message to the group's previous-hop router. When an upstream router receives a graft message, it cancels out the previously received prune message. Graft messages may cascade back toward the sourc,e allowing previously pruned branches to be restored as part of the multicast delivery tree.

## DVMRP Router Functions

When there is more than one DVMRP router on a subnetwork, the Dominant Router is responsible for the periodic transmission of IGMP Host Membership Query messages. Upon initialization, a DVMRP router considers itself to be the Dominant Router for the subnetwork until it receives a Host Membership Query message from a neighbor router with a lower IP address. Figure 12 illustrates how the router with the lowest IP address functions as the Designated Router for the subnetwork .

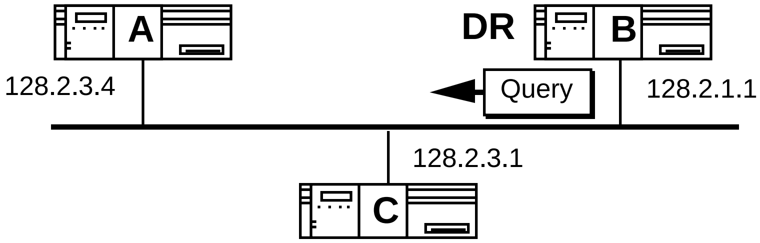

**Figure 12: DVMRP Designated Router**

In order to avoid duplicate multicast datagrams when there is more than one DVMRP router on a subnetwork, one router is elected the Dominant Router for the particular source subnetwork (see Figure 12). In Figure 13, Router C is downstream and may potentially receive datagrams from the source subnetwork from Router A or Router B. If Router A's metric to the source subnetwork is less than Router B's metric, Router A is dominant to Router B for this source. This means that Router A forwards traffic from the source subnetwork and Router B discards traffic from the source subnetwork.

However, if Router A’s metric is equal to Router B’s metric, the router with the lowest IP address on its downstream interface (child link) becomes the Dominant Router.

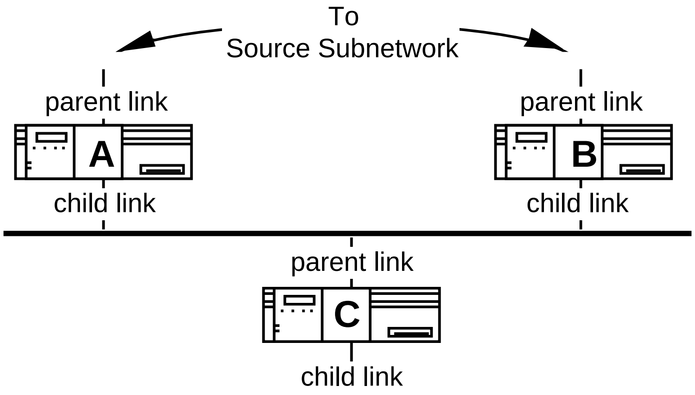

**Figure 13: DVMRP Dominant Router**

## DVMRP Routing Table

Since the DVMRP was developed to route multicast and not unicast traffic, a router may be required to run multiple routing processes—one for the delivery of unicast traffic and another for the delivery of multicast traffic. The DVMRP process periodically exchanges routing table update messages with multicast-capable neighbors. These updates are independent of those generated by any Interior Gateway Protocol that provides support for unicast routing.

DVMRP relies on the receipt of “poison reverse” updates for leaf router detection. This technique requires that a downstream neighbor advertise “infinity” for a source subnetwork to the previous hop router on its shortest-path back to that source subnetwork. If an upstream router does not receive a “poison reverse” update for the source subnetwork on a downstream interface, the upstream router assumes that the downstream subnetwork is a leaf and removes the downstream port from its list of forwarding ports.

A sample routing table for a DVMRP router is shown in Figure 14. Unlike the typical table created by a unicast routing protocol such as RIP, the DVMRP routing table contains Source Subnets and From-Gateways instead of Destinations and Next-Hop Gateways. The routing table represents the shortest path source-rooted spanning tree to every participating subnetwork in the internetwork—the Reverse Path Broadcasting (RPB) tree. The DVMRP routing table does not consider group membership or received prune messages.

| Source Subnet | Subnet Mask | From Gateway | Metric | Status | TTL | InPort | OutPorts |
|---|---|---|---|---|---|---|---|
| 128.1.0.0 | 255.255.0.0 | 128.7.5.2 | 3 | Up | 200 | 1 | 2,3 |
| 128.2.0.0 | 255.255.0.0 | 128.7.5.2 | 5 | Up | 150 | 2 | 1 |
| 128.3.0.0 | 255.255.0.0 | 128.6.3.1 | 2 | Up | 150 | 2 | 1,3 |
| 128.4.0.0 | 255.255.0.0 | 128.6.3.1 | 4 | Up | 200 | 1 | 2 |

**Figure 14: DVMRP Routing Table**

The key elements in DVMRP routing table include the following items:

Source Subnet
A subnetwork containing a host sourcing multicast datagrams.

Subnet Mask
The subnet mask assigned to the Source Subnet. Note that the DVMRP provides the subnet mask for each source subnetwork.

From-Gateway
The previous hop router leading back to the Source Subnet.

TTL
The time-to-live is used for table management and indicates the number of seconds before an entry is removed from the routing table.

## DVMRP Forwarding Table

Since the DVMRP routing table is not aware of group membership, the DVMRP process builds a forwarding table based on a combination of the information contained in the multicast routing table, known groups, and received prune messages. The forwarding table represents the local router’s understanding of the shortest path source-rooted delivery tree for each (source, group) pair—the Reverse Path Multicasting (RPM) tree.

| Source Subnet | Multicast Group | TTL | InPort | OutPorts |
|---|---|---|---|---|
| 128.1.0.0 | 224.1.1.1 | 200 | 1 Pr | 2p 3p |
| | 224.2.2.2 | 100 | 1 | 2p 3 |
| | 224.3.3.3 | 250 | 1 | 2 |
| 128.2.0.0 | 224.1.1.1 | 150 | 2 | 2p 3 |

**Figure 15: DVMRP Forwarding Table**

The forwarding table for a typical DVMRP router is shown in Figure 15. The elements in this display include the following items:

Source Subnet
The subnetwork containing a host sourcing multicast datagrams addressed to the specified groups.

Multicast Group
The Class D IP address to which multicast datagrams are addressed. Note that a given Source Subnet may contain sources for many different Multicast Groups.

InPort
The parent port for the (source, group) pair. A “Pr” in this column indicates that a prune message has been sent to the upstream router.

OutPorts
The child ports over which multicast datagrams for the (source, group) pair are forwarded. A lower-case “p” in this column indicates that the router has received a prune message from a downstream router.

## Hierarchical DVMRP

The rapid growth of the MBONE is beginning to place increasing demands on its routers. The current version of the DVMRP treats the MBONE as a single, “flat” routing domain where each router is required to maintain detailed routing information to every subnetwork on the MBONE. As the number of subnetworks continues to increase, the size of the routing tables and of the periodic update messages will continue to grow. If nothing is done about these issues, the processing and memory capabilities of the MBONE routers will eventually be depleted and routing on the MBONE will fail.

### Benefits of Hierarchical Multicast Routing

To overcome these potential threats, a hierarchical version of the DVMRP is under development. In hierarchical routing, the MBONE is divided into a number of individual routing domains. Each routing domain executes its own instance of a multicast routing protocol. Another protocol, or another instance of the same protocol, is used for routing between the individual domains. Hierarchical routing reduces the demand for router resources because each router only needs to know the explicit details about routing packets to destinations within its own domain, but knows nothing about the detailed topological structure of any of the other domains. The protocol running between the individual domains maintains information about the interconnection of the domains, but not about the internal topology of each domain.

In addition to reducing the amount of routing information, there are several other benefits gained from the development of a hierarchical version of the DVMRP:

- Different multicast routing protocols may be deployed in each region of the MBONE. This permits the testing and deployment of new protocols on a domain-by-domain basis.

- The effects of an individual link or router failures are limited to only those routers operating within a single domain. Likewise, the effects of any change to the topological interconnection of regions is limited to only inter-domain routers. These enhancements are especially important when deploying a distance-vector routing protocol that can result in relatively long convergence times.

- The count-to-infinity problem associated with distance-vector routing protocols places limitations on the maximum diameter of the MBONE topology. Hierarchical routing limits these diameter constraints to a single domain, not to the entire MBONE.

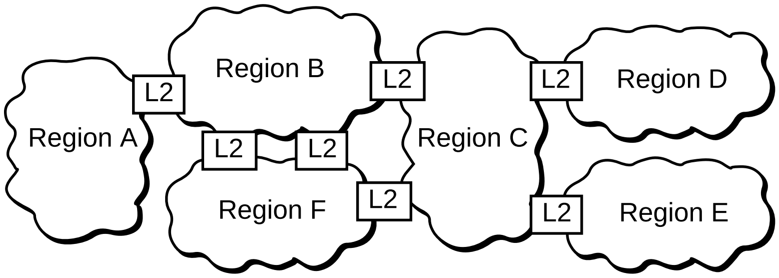

**Figure 16: Hierarchical DVMRP**

### Hierarchical Architecture

Hierarchical DVMRP proposes the creation of non-intersecting regions where each region has a unique Region-Id. The routers internal to a region execute any multicast routing protocols such as DVMRP, MOSPF, PIM, or CBT as a "Level 1" (L1) protocol. Each region is required to have at least one "boundary router" that is responsible for providing inter-regional connectivity. The boundary routers execute DVMRP as a "Level 2" (L2) protocol to forward traffic between regions (Figure 16).

The L2 routers exchange routing information in the form of Region-Ids instead of the individual subnetwork addresses contained within each region. With DVMRP as the L2 protocol, inter-regional multicast delivery tree is constructed based on the (region_ID, group) pair rather than the standard (source, group) pair.

When a multicast packet originates within a region, it is forwarded according to the L1 protocol to all subnetworks containing group members. In addition, the datagram is forwarded to each of the boundary routers (L2) configured for the source region. The L2 routers tag the packet with the Region-Id and place it in an encapsulation header for delivery to other regions. When the packet arrives at a remote region, the encapsulation header is removed before delivery to group members by the L1 routers.

## Multicast Extensions to OSPF (MOSPF)

Version 2 of the Open Shortest Path First (OSPF) routing protocol is defined in RFC-1583. It is an Interior Gateway Protocol (IGP) specifically designed to distribute unicast topology information among routers belonging to a single Autonomous System. OSPF is based on link-state algorithms that permit rapid route calculation with a minimum of routing protocol traffic. In addition to efficient route calculation, OSPF is an open standard that supports hierarchical routing, load balancing, and the import of external routing information.

The Multicast extensions to OSPF (MOSPF) are defined in RFC-1584. MOSPF routers maintain a current image of the network topology through the unicast OSPF link-state routing protocol. MOSPF enhances the OSPF protocol by providing the ability to route multicast IP traffic. The multicast extensions to OSPF are built on top of OSPF Version 2 so that a multicast routing capability can be easily introduced into an OSPF Version 2 routing domain. The enhancements that have been added are backward compatible so that routers running MOSPF will interoperate with non-multicast OSPF routers when forwarding unicast IP data traffic.

MOSPF, unlike DVMRP, does not provide support for tunnels.

## Intra-Area Routing with MOSPF

Intra-Area Routing describes the basic routing algorithm employed by MOSPF. This elementary algorithm runs inside a single OSPF area and supports multicast forwarding when the source and all destination group members reside in the same OSPF area, or when the entire Autonomous System is a single OSPF area. The following discussion assumes that the reader is familiar with the basic operation of the OSPF routing protocol.

### Local Group Database

Similar to the DVMRP, MOSPF routers use the Internet Group Management Protocol (IGMP) to monitor multicast group membership on directly attached subnetworks. MOSPF routers are required to implement a "local group database" that  maintains a list of directly attached group members and determines the local router's responsibility for delivering multicast datagrams to these group members.

On any given subnetwork, the transmission of IGMP Host Membership Queries is performed solely by the Designated Router (DR). Also, the responsibility of listening to IGMP Host Membership Reports is performed only by the Designated Router (DR) and the Backup Designated Router (BDR). This means that in a mixed environment containing both MOSPF and OSPF routers, an MOSPF router must be elected the DR for the subnetwork if IGMP Queries are to be generated. This can be achieved by simply assigning all non-MOSPF routers a RouterPriority of 0 to prevent them from becoming the DR or BDR, thus allowing an MOSPF router to become the DR for the subnetwork.

The DR is responsible for communicating group membership information to all other routers in the OSPF area by flooding Group-Membership LSAs. The DR originates a separate Group-Membership LSA for each multicast group having one or more entries in the DR's local group database. Similar to Router-LSAs and Network-LSAs, Group Membership-LSAs are flooded throughout a single area only. This ensures that all remotely originated multicast datagrams are forwarded to the specified subnetwork for distribution to local group members.

### Datagram's Shortest Path Tree

The datagram's shortest path tree describes the path taken by a multicast datagram as it travels through the internetwork from the source subnetwork to each of the individual group members. The shortest path tree for each (source, group) pair is built "on demand" when a router receives the first multicast datagram for a particular (source, group) pair.

When the initial datagram arrives, the source subnetwork is located in the MOSPF link state database. The MOSPF link state database is simply the standard OSPF link state database with the addition of Group-Membership LSAs. Based on the Router-LSAs and Network-LSAs in the MOSPF link state database, a source-rooted shortest-path tree is constructed using Dijkstra's algorithm. After the tree is built, Group-Membership LSAs are used to prune those branches that do not lead to subnetworks containing individual group members. The result of the Dijkstra calculation is a pruned shortest-path tree rooted at the datagram's source.

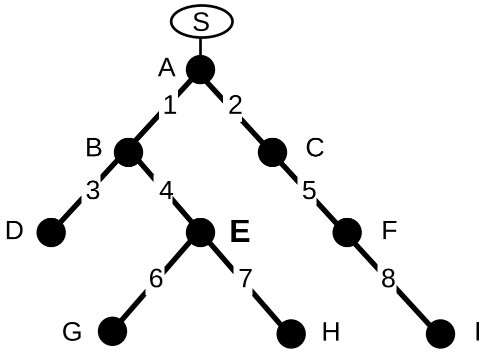

**Figure 17: Shortest Path Tree for (S, G)**

To forward a multicast datagram to downstream members of the group, each router must determine its position in the datagram's shortest path delivery tree. Assume that Figure 17 illustrates the shortest path tree for a particular (source, group) pair. Router E's upstream node is Router B and there are two downstream interfaces: one connecting to Subnetwork 6 and another connecting to Subnetwork 7.

Note the following properties of the basic MOSPF routing algorithm:

- For a given multicast datagram, all routers within an OSPF area calculate the same source-rooted shortest path delivery tree. Tie-breakers have been defined to guarantee that if several equal- cost paths exist, all routers agree on a single path through the area. Unlike unicast OSPF, MOSPF does not support the concept of equal-cost multipath routing.

- Synchronized link state databases containing Group-Membership LSAs allow an MOSPF router to effectively perform the Reverse Path Multicasting (RPM) computation "in memory". Unlike DVMRP, this means that the first datagram of a group transmission does not have to be forwarded to all routers in the area.

- The "on demand" construction of the shortest-path delivery tree has the benefit of spreading calculations over time, resulting in a lesser impact for participating routers.

### Forwarding Cache

Each MOSPF router makes its forwarding decision based on the contents of its forwarding cache. The forwarding cache is built from the source-rooted shortest-path tree for each (source, group) pair and the router's local group database. After the router discovers its position in the shortest path tree, a forwarding cache entry is created containing the (source, group) pair, the upstream node, and the downstream interfaces. At this point, the Dijkstra shortest path tree is discarded, releasing all resources associated with the creation of the tree. From this point on, the forwarding cache entry is used to forward all subsequent datagrams for the (source, group) pair.

| Destination | Source | Upstream | Downstream | TTL |
|---|---|---|---|---|
| 224.1.1.1 | 128.1.0.2 | !1 | !2 !3 | 5 |
| 224.1.1.1 | 128.4.1.2 | !1 | !2 !3 | 2 |
| 224.1.1.1 | 128.5.2.2 | !1 | !2 !3 | 3 |
| 224.2.2.2 | 128.2.0.3 | !2 | !1 | 7 |

**Figure 18: MOSPF Forwarding Cache**

Figure 18 displays the forwarding cache for an example MOSPF router. The elements in the display include the following items:

Destination The destination group address to which matching datagrams are forwarded.

Source The datagram’s source subnetwork. Each Destination/Source pair identifies a separate forwarding cache entry.

Upstream The interface from which a matching datagram must be received.

Downstream The interfaces over which a matching datagram should be forwarded to reach Destination group members.

TTL The minimum number of hops a datagram will travel to reach the multicast group members. This allows the router to discard datagrams that do not have a chance of reaching a destination group member.

The information in the forwarding cache is not aged or periodically refreshed. It is maintained as long as there are system resources available (i.e., memory) or until the next topology change. In general, the contents of the forwarding cache will change when:

- The topology of the OSPF internetwork changes, forcing all of the datagram shortest-path trees to be recalculated.

- There is a change in the Group-Membership LSAs indicating that the distribution of individual group members has changed.

## Mixing MOSPF and OSPF Routers

MOSPF routers can be combined with non-multicast OSPF routers. This permits the gradual deployment of MOSPF and allows experimentation with multicast routing on a limited scale. When MOSPF and non-multicast OSPF routers are mixed within an

Autonomous System, all routers will interoperate in the forwarding of unicast datagrams.

It is important to note that an MOSPF router is required to eliminate all non-multicast OSPF routers when it builds its source-rooted shortest-path delivery tree. An MOSPF router can easily determine the multicast capability of any other router based on the setting of the multicast bit (MC-bit) in the Options field of each router’s link state advertisements. The omission of non-multicast routers can create a number of potential problems when forwarding multicast traffic:

- Multicast datagrams may be forwarded along suboptimal routes since the shortest path between two points may require traversal of a non-multicast OSPF router.

- Even though there is unicast connectivity to a destination, there may not be multicast connectivity. For example, the network may partition with respect to multicast connectivity since the only path between two points requires traversal of a non-multicast OSPF router.

- The forwarding of multicast and unicast datagrams between two points may follow entirely different paths through the internetwork. This may make some routing problems a bit more difficult to debug.

- The Designated Router for a multi-access network must be an MOSPF router. If a non-multicast OSPF router is elected the DR, the subnetwork will not be selected to forward multicast datagrams since a non-multicast DR cannot generate Group-Membership LSAs for its subnetwork.

## Inter-Area Routing with MOSPF

Inter-area routing involves the case where a datagram’s source and some of its destination group members reside in different OSPF areas. It should be noted that the forwarding of multicast datagrams continues to be determined by the contents of the forwarding cache which is still built from the local group database and the datagram shortest-path trees. The major differences are related to the way that group membership information is propagated and the way that the inter-area shortest-path tree is constructed.

### Inter-Area Multicast Forwarders

In MOSPF, a subset of an area’s Area Border Routers (ABRs) function as “inter-area multicast forwarders.” An inter-area multicast forwarder is responsible for the forwarding of group membership information and multicast datagrams between areas. Configuration parameters determine whether or not a particular ABR also functions as an inter-area multicast forwarder.

Inter-area multicast forwarders summarize their attached areas' group membership information to the backbone by originating new Group-Membership LSAs into the backbone area (Figure 19). It is important to note that the summarization of group membership in MOSPF is asymmetric. This means that group membership information from non-backbone areas is flooded into the backbone. However, the backbone does not readvertise either backbone group membership information or group membership information learned from other non-backbone areas into any non-backbone areas.

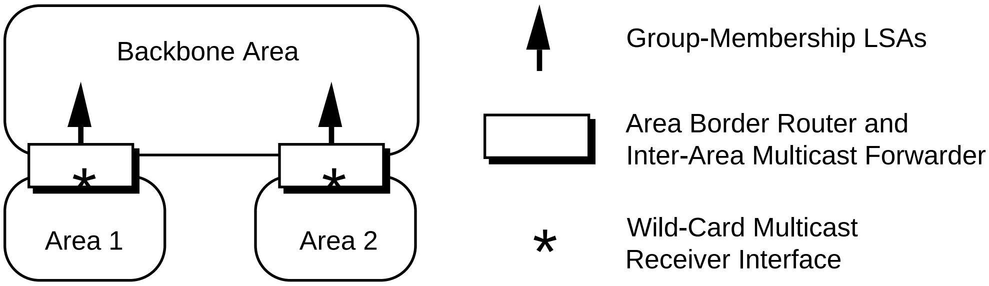

**Figure 19: Inter-Area Routing Architecture**

To permit the forwarding of multicast traffic between areas, MOSPF introduces the concept of a "wild-card multicast receiver." A wild-card multicast receiver is a router that receives all multicast traffic generated in an area, regardless of the multicast group membership. In non-backbone areas, all inter-area multicast forwarders operate as wild-card multicast receivers. This guarantees that all multicast traffic originating in a non-backbone area is delivered to its inter-area multicast forwarder, and then if necessary into the backbone area. Since the backbone has group membership knowledge for all areas, the datagram can then be forwarded to group members residing in the backbone and other, non-backbone areas. The backbone area does not require wild-card multicast receivers because the routers in the backbone area have complete knowledge of group membership information for the entire OSPF system.

### Inter-Area Datagram Shortest-Path Tree

In the case of inter-area multicast routing, it is often impossible to build a complete datagram shortest-path delivery tree. Incomplete trees are created because detailed topological and group membership information for each OSPF area is not distributed between OSPF areas. To overcome these limitations, topological estimates are made through the use of wild-card receivers and OSPF Summary-Links LSAs.

There are two cases that need to be considered when constructing an inter-area shortest-path delivery tree. The first involves the condition when the source subnetwork is located in the same area as the router performing the calculation. The second situation occurs when the source subnetwork is located in a different area than the router performing the calculation.

If the source of a multicast datagram resides in the same area as the router performing the calculation, the pruning process must be careful to ensure that branches leading to other areas are not removed from the tree (Figure 20). Only those branches having no group members nor wild-card multicast receivers are pruned. Branches containing wild-card multicast receivers must be retained, since the local routers do not know if there are group members residing in other areas.

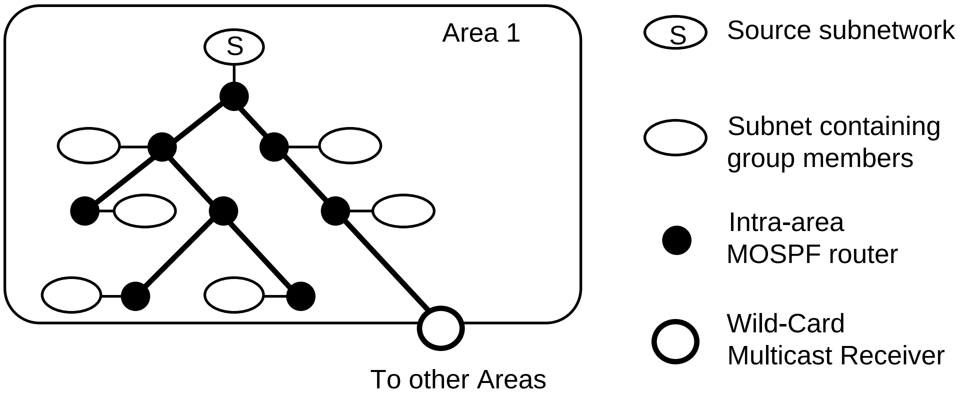

**Figure 20: Datagram Shortest-Path Tree—Source in Same Area**

If the source of a multicast datagram resides in a different area than the router performing the calculation, the details describing the local topology surrounding the source station are not known. However, this information can be estimated using information provided by Summary-Links LSAs for the source subnetwork. In this case, the base of the tree begins with branches directly connecting the source subnetwork to each of the local area's inter-area multicast forwarders (Figure 21). The inter-area multicast forwarders must be included in the tree, since any multicast datagrams originating outside the local area will enter the area via an inter-area multicast forwarder.

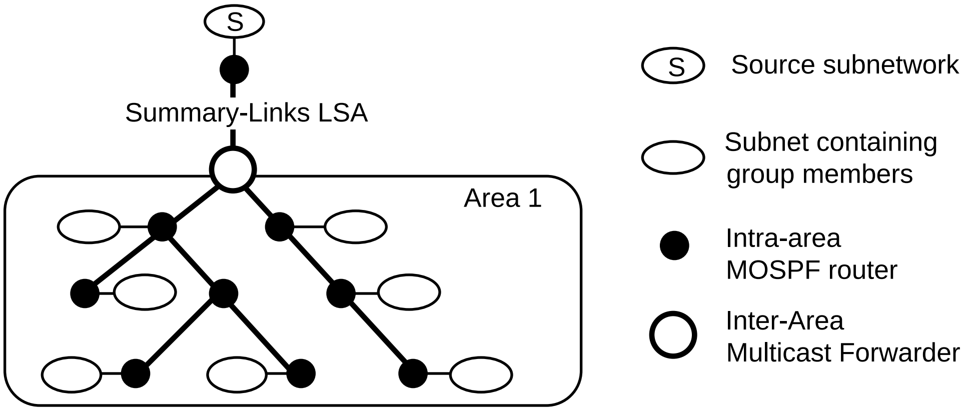

**Figure 21: Datagram Shortest-Path Tree—Source in a Remote Area**

Since each inter-area multicast forwarder is also an ABR, it must maintain a separate link-state database for each attached area. This means that each inter-area multicast forwarder is required to calculate a separate forwarding tree for each of its attached areas. After the individual trees are calculated, they are merged into a single forwarding cache entry for the (source, group) pair and then the individual trees are discarded.

## Inter-Autonomous System Multicasting with MOSPF

Inter-Autonomous System Multicasting involves the situation where a datagram's source and at least some of its destination group members reside in different Autonomous Systems. It should be emphasized that in OSPF terminology "inter-AS" communication also refers to connectivity between an OSPF domain and another routing domain that could be within the same Autonomous System.

To facilitate inter-AS multicast routing, selected Autonomous System Boundary Routers (ASBRs) are configured as "inter-AS multicast forwarders." MOSPF makes the assumption that each inter-AS multicast forwarder executes an inter-AS multicast routing protocol (such as DVMRP), which forwards multicast datagrams in a reverse path forwarding (RPF) manner. Each inter-AS multicast forwarder functions as a wild-card multicast receiver in each of its attached areas. This guarantees that each inter-AS multicast forwarder remains on all pruned shortest-path trees and receives all multicast datagrams, regardless of the multicast group membership.

Three cases need to be considered when describing the construction of an inter-AS shortest-path delivery tree. The first occurs when the source subnetwork is located in the same area as the router performing the calculation. For the second case, the source subnetwork resides in a different area than the router performing the calculation. The final case occurs when the source subnetwork is located in a different AS (or in another routing domain within the same AS) than the router performing the calculation.

The first two cases are similar to the inter-area examples described in the previous section. The only enhancement is that inter-AS multicast forwarders must also be included on the pruned shortest path delivery tree. Branches containing inter-AS multicast forwarders must be retained since the local routers do not know if there are group members residing in other Autonomous Systems. When a multicast datagram arrives at an inter-AS multicast forwarder, it is the responsibility of the ASBR to determine whether the datagram should be forwarded outside of the local Autonomous System. Figure 22 illustrates a sample inter-AS shortest path delivery tree when the source subnetwork resides in the same area as the router performing the calculation.

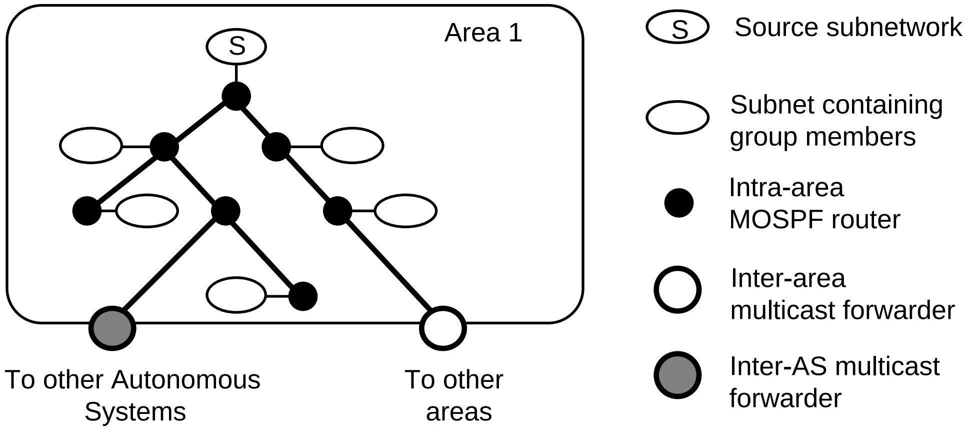
Area 1
S
To other Autonomous Systems
To other areas
S Source subnetwork
Subnet containing group members
Intra-area MOSPF router
Inter-area multicast forwarder
Inter-AS multicast forwarder

**Figure 22: Inter-AS Datagram Shortest-Path Tree—Source in Same Area**

If the source of a multicast datagram resides in a different Autonomous System than the router performing the calculation, the details describing the local topology surrounding the source station are not known. However, this information can be estimated using the multicast-capable AS-External Links describing the source subnetwork. In this case, the base of the tree begins with branches directly connecting the source subnetwork to each of the local area's inter-AS multicast forwarders.

Figure 23 shows a sample inter-AS shortest-path delivery tree when the inter-AS multicast forwarder resides in the same area as the router performing the calculation. If the inter-AS multicast forwarder is located in a different area than the router performing the calculation, the topology surrounding the source is approximated by combining the Summary-ASBR Link with the multicast-capable AS-External Link.

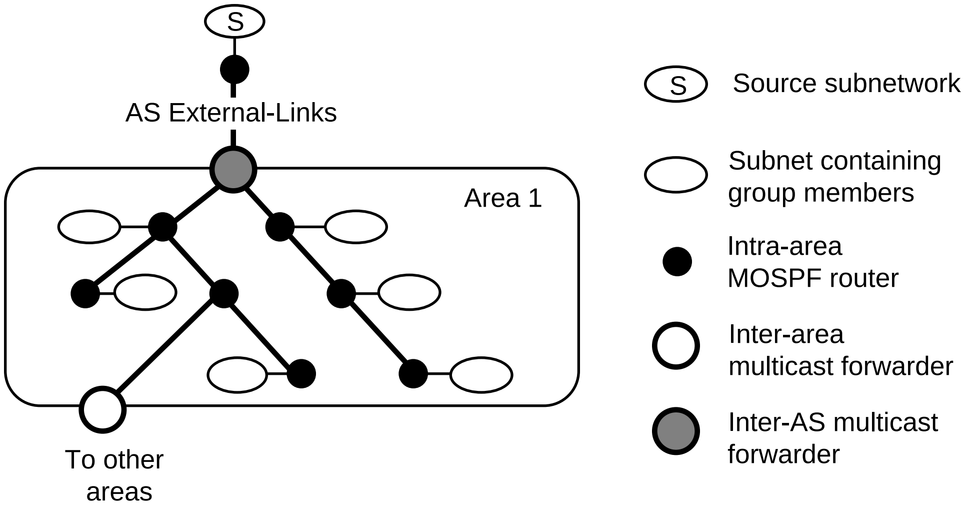
S
AS External-Links
Area 1
To other areas
S
Source subnetwork
Subnet containing group members
Intra-area MOSPF router
Inter-area multicast forwarder
Inter-AS multicast forwarder

**Figure 23: Inter-AS Datagram Shortest-Path Tree—Source in Different AS**

As a final point, it is important to note that AS External Links are not imported into Stub areas. If the source is located outside of the stub area, the topology surrounding the source is estimated by the Default Summary Links originated by the stub area's intra-area multicast forwarder rather than the AS-External Links.

## Protocol-Independent Multicast (PIM)

The Protocol-Independent Multicast (PIM) routing protocol is currently under development by the Inter-Domain Multicast Routing (IDMR) working group of the IETF. The objective of the IDMR working group is to develop a standard multicast routing protocol that can provide scalable inter-domain multicast routing across the Internet.

PIM receives its name because it is not dependent on the mechanisms provided by any particular unicast routing protocol. However, any implementation supporting PIM requires the presence of a unicast routing protocol to provide routing table information and to adapt to topology changes.

PIM makes a clear distinction between a multicast routing protocol that is designed for dense environments and one that is designed for sparse environments. Dense-mode refers to a protocol that is designed to operate in an environment where group members are relatively densely packed and bandwidth is plentiful. Sparse-mode refers to a protocol that is optimized for environments where group members are distributed across many regions of the Internet and bandwidth is not necessarily widely available. It is important to note that sparse-mode does not imply that the group has few members, just that they are widely dispersed across the Internet.

The designers of PIM argue that DVMRP and MOSPF were developed for environments where group members are densely distributed. They emphasize that when group members and senders are sparsely distributed across a wide area, DVMRP and MOSPF do not provide the most efficient multicast delivery service. DVMRP periodically sends multicast packets over many links that do not lead to group members, while MOSPF can send group membership information over links that do not lead to senders or receivers.

### PIM Dense Mode (PIM-DM)

While the PIM architecture was driven by the need to provide scalable sparse-mode delivery trees, it also defines a new dense-mode protocol instead of relying on existing dense-mode protocols such as DVMRP and MOSPF. It is envisioned that PIM-DM will be deployed in resource-rich environments, such as a campus LAN where group membership is relatively dense and bandwidth is likely to be readily available.

PIM Dense Mode (PIM-DM) is similar to DVMRP in that it employs the Reverse Path Multicasting (RPM) algorithm. However, there are several important differences between PIM-DM and DVMRP:

- PIM-DM relies on the presence of an existing unicast routing protocol to adapt to topology changes, but it is independent of the mechanisms of the specific unicast routing protocol. In contrast, DVMRP contains an integrated routing protocol that makes use of its own RIP-like exchanges to compute the required unicast routing information. MOSPF uses the information contained in the OSPF link-state database, but MOSPF is specific to only the OSPF unicast routing protocol.

- Unlike DVMRP, which calculates a set of child interfaces for each (source, group) pair, PIM-DM simply forwards multicast traffic on all downstream interfaces until explicit prune messages are received. PIM-DM is willing to accept packet duplication to eliminate routing protocol dependencies and to avoid the overhead involved in building the parent/child database.

For those cases where group members suddenly appear on a pruned branch of the distribution tree, PIM-DM, like DVMRP, employs graft messages to add the previously pruned branch to the delivery tree. Finally, PIM-DM control message processing and data packet forwarding are integrated with PIM-Sparse Mode operation so that a single router can run different modes for different groups.

## PIM Sparse Mode (PIM-SM)

PIM Sparse Mode (PIM-SM) is being developed to provide a multicast routing protocol that provides efficient communication between members of sparsely distributed groups - the type of groups that are most common in wide-area internetworks. Its designers believe that a situation in which several hosts wish to participate in a multicast conference do not justify flooding the entire internetwork with periodic multicast traffic. They fear that existing multicast routing protocols will experience scaling problems if several thousand small conferences are in progress, creating large amounts of aggregate traffic that would potentially saturate most wide-area Internet connections. To eliminate these potential scaling issues, PIM-SM is designed to limit multicast traffic so that only those routers interested in receiving traffic for a particular group “see” it.

PIM-SM differs from existing dense-mode multicast algorithms in two essential ways:

- Routers with directly attached or downstream members are required to join a sparse-mode distribution tree by transmitting explicit join messages. If a router does not become part of the predefined distribution tree, it will not receive multicast traffic addressed to the group. In contrast, dense-mode multicast routing protocols assume downstream group membership and continue to forward multicast traffic on downstream links until explicit prune messages are received. The default forwarding action of the other dense-mode multicast routing protocols is to forward traffic, while the default action of a sparse-mode multicast routing protocol is to block traffic unless it is explicitly requested.

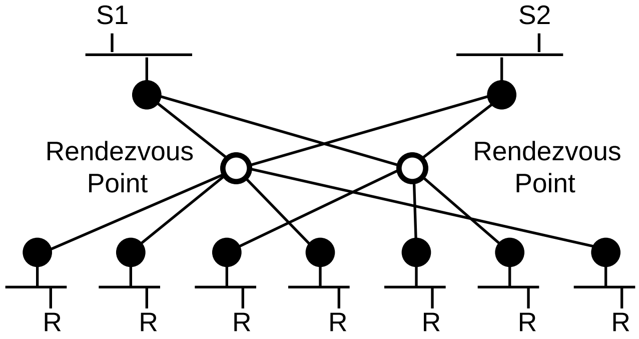

**Figure 24: Rendezvous Points**

- PIM-SM is similar to the Core-Based Tree (CBT) approach in that it employs the concept of a rendezvous point (RP) where receivers “meet” sources. The initiator of each multicast group selects a primary RP and a small ordered set of alternative RPs, known as the RP-list. For each multicast group, there is only a single active RP. Each receiver wishing to join a multicast group contacts its directly attached router, which in turn joins the multicast distribution tree by sending an explicit join message to the group's primary RP. A source uses the RP to announce its presence and to find a path to members that have joined the group. This model requires sparse-mode routers to maintain some state (i.e., the RP-list) prior to the arrival of data packets. In contrast, dense-mode multicast routing protocols are data driven, since they do not define any state for a multicast group until the first data packet arrives.

### Directly Attached Host Joins a Group

When there is more than one PIM router connected to a multi-access LAN, the router with the highest IP address is selected to function as the Designated Router (DR) for the LAN. The DR is responsible for the transmission of IGMP Host Query messages, for sending Join/Prune messages toward the RP, and for maintaining the status of the active RP for local senders to multicast groups (Figure 25).

To facilitate the differentiation between DM and SM groups, a part of the Class D multicast address space is being reserved for use by SM groups. When the DR receives an IGMP Report message for a new group, the DR determines if the group is RP-based by examining the group address. If the address indicates a SM group, the DR performs a lookup in the associated group's RP-list to determine the primary RP for the group. The draft specification describes a procedure for the selection of the primary RP and the use of alternate RPs if the primary RP becomes unreachable.

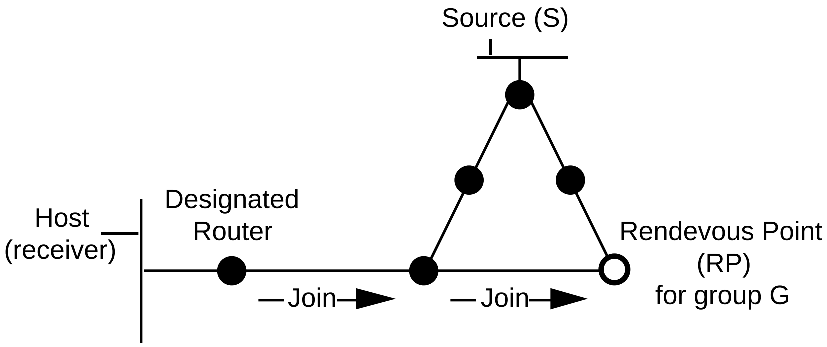

**Figure 25: Host Joins a Multicast Group**

After performing the lookup, the DR creates a multicast forwarding cache for the (*, group) pair and transmits a unicast PIM-Join message to the primary RP. The (*, group) notation indicates an (any source, group) pair. The intermediate routers forward the unicast PIM-Join message and create a forwarding cache entry for the (*, group) pair. Intermediate routers create the forwarding cache entry so that they will know how to forward traffic addressed to the (*, group) pair downstream to the DR originating the PIM-Join message.

### Directly Attached Source Sends to a Group

When a host first transmits a multicast packet to a group, its DR must forward the datagram to the primary RP for subsequent distribution across the group's delivery tree. The DR encapsulates the multicast packet in a PIM-SM-Register packet and unicasts it to the primary RP for the group. The PIM-SM-Register packet informs the RP of a new source, which causes the active RP to transmit PIM-Join messages back to the source station's DR. The routers lying between the source's DR and the RP maintain state from received PIM-Join messages so that they will know how to forward subsequent unencapsulated multicast packets from the source subnetwork to the RP.

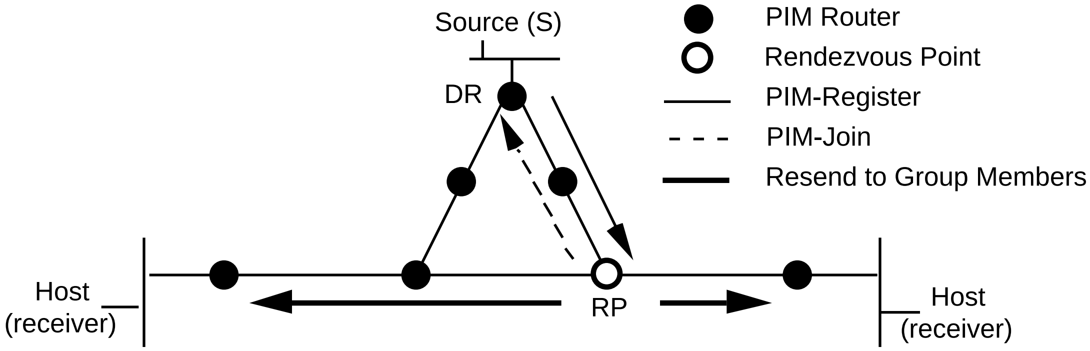

**Figure 26: Source Sends to a Multicast Group**

The source’s DR ceases to encapsulate data packets in PIM-SM- Registers when it receives Join/Prune messages from the RP. At this point, data traffic is forwarded by the DR in its native multicast format to the RP. When the RP receives multicast packets from the source station, it resends the datagrams on the RP-shared tree to all downstream group members.

### RP-Shared Tree or Shortest Path Tree (SPT)

The RP-shared tree provides connectivity for group members but does not optimize the delivery path through the internetwork. PIM-SM allows receivers to either continue to receive multicast traffic over the RP-shared tree or over a source-rooted shortest-path tree that a receiver subsequently creates. The shortest-path tree allows a group member to reduce the delay between itself and a particular source.

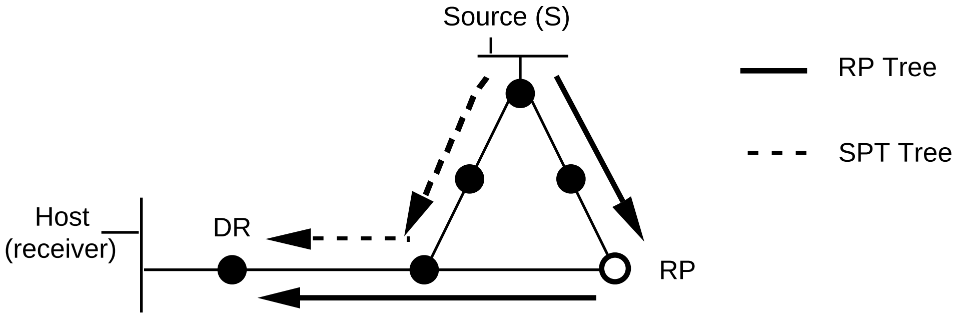

**Figure 27: RP-Shared Tree (RP Tree) and Shortest-Path Tree (SPT)**

A PIM router with local receivers has the option of switching to the source’s shortest-path tree as soon as it starts receiving data packets from the source station. The changeover may be triggered if the data rate from the source station exceeds a predefined threshold. The local receiver’s DR does this by sending a Join message toward the active source. At the same time, protocol mechanisms guarantee that a Prune message for the same source is transmitted to the active RP. Alternatively, the DR may be configured to continue using the RP-based tree and never switch over to the source’s shortest-path tree.

### Unresolved Issues

It is important to note that PIM is an Internet draft. This means that it is still early in its development cycle and clearly a work in progress. There are several important issues that require further research, engineering, and/or experimentation:

- PIM-SM still requires routers to maintain a significant amount of state information to describe sources and groups.

- Some multicast routers will be required to have both PIM interfaces and non-PIM interfaces. The interaction and sharing of multicast routing information between PIM and other multicast routing protocols is still in the early stages of definition.

- The future deployment of PIM-SM will probably require more coordination between Internet service providers to support an Internet-wide delivery service.

- Finally, PIM-SM is considerably more complex than DVMRP or the MOSPF extensions.

## References

### Requests for Comment (RFCs)

1075 "Distance Vector Multicast Routing Protocol," D. Waitzman, C. Partridge, and S. Deering, November 1988.

1112 "Host Extensions for IP Multicasting," Steve Deering, August 1989.

1583 "OSPF Version 2," John Moy, March 1994.

1584 "Multicast Extensions to OSPF," John Moy, March 1994.

1585 "MOSPF: Analysis and Experience," John Moy, March 1994.

1800 "Internet Official Protocol Standards," Jon Postel, Editor, July 1995.

1812 "Requirements for IP Version 4 Routers," Fred Baker, Editor, June 1995.

## Internet Drafts

"Core Based Trees (CBT) Multicast: Architectural Overview," <draft-ietf-idmr-cbt-arch-02.txt>, A. J. Ballardie, June 20, 1995.

"Core Based Trees (CBT) Multicast: Protocol Specification," <draft-ietf-idmr-cbt-spec-03.txt>, A. J. Ballardie, November 21, 1995.

"Hierarchical Distance Vector Multicast Routing for the MBONE," Ajit Thyagarajan and Steve Deering, July 1995.

"Internet Group Management Protocol, Version 2," <draft-ietf- idmr-igmp-v2-01.txt>, William Fenner, Expires April 1996.

"Internet Group Management Protocol, Version 3," <draft-cain- igmp-00.txt>, Brad Cain, Ajit Thyagarajan, and Steve Deering, Expires March 8, 1996

"Protocol-Independent Multicast (PIM), Dense-Mode Protocol Specification," <draft-ietf-idmr-PIM-DM-spec-01.ps>, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, P. Sharma, and A. Helmy, January 17, 1996.

"Protocol-Independent Multicast (PIM): Motivation and Architecture," <draft-ietf-idmr-pim-arch-01.ps>, S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, and L. Wei, January 11, 1995.

"Protocol-Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," <draft-ietf-idmr-PIM-SM-spec-02.ps>, S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, P. Sharma, and A Helmy, September 7, 1995.

## Textbooks

Comer, Douglas E. *Internetworking with TCP/IP: Volume 1—Principles, Protocols, and Architecture,* Second Edition. Englewood Cliffs, NJ: Prentice Hall, 1991.

Huitema, Christian. *Routing in the Internet*. Englewood Cliffs, NJ: Prentice Hall, 1995.

Stevens, W. Richard. *TCP/IP Illustrated: Volume 1 The Protocols*. Reading, MA: Addison Wesley, 1994.

Wright, Gary, and W. Richard Stevens. *TCP/IP Illustrated: Volume 2—The Implementation*. Reading, MA: Addison Wesley, 1995

## Other

Deering, Steven E. "Multicast Routing in a Datagram Internetwork," Ph.D. Thesis, Stanford University, December 1991.

Ballardie, Anthony J. "A New Approach to Multicast Communication in a Datagram Internetwork," Ph.D. Thesis, University of London, May 1995.