IP MULTICAST EXPLAINED

June 2004

Jon Hardwick
Data Connection Ltd.
Jon.Hardwick@dataconnection.com

Data Connection Limited
100 Church Street
Enfield, UK
Tel: +44 20 8366 1177
http://www.dataconnection.com/

TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW

1.1 Introduction

In contrast to the one-to-one model of IP unicast, in which data packets are sent from a single source to a single recipient, IP multicast provides a method of efficient many-to-many communication. This concept is becoming increasingly important, both in the Internet and in private networks, for providing services such as multimedia content delivery.

This is especially true for services that require material to be propagated simultaneously to a subset of subscribers. For example, multicast may be used in the provision of e-learning courses, in the distribution of information to stock-market feeds or news feeds, or in broadcasting webinars, radio and television.

There is a large and ever-expanding range of protocols used to provide and support various aspects of multicast, including address allocation, group membership, and multicast routing and signaling. It can be hard to keep track of and understand the protocols in all these areas.

This white paper gives an overview of the most important network-layer protocols used in multicast. We aim to provide a guide for developers or network managers familiar with unicast routing but new to multicast.

We begin by giving an introduction to the key concepts of multicast, starting from the ground up. We then go into more depth about the main protocols in each area, focusing particularly on multicast routing. Where there are multiple protocols offering similar functionality, we compare and contrast the different approaches.

1.2 What is Multicast?

Normal IP packets are sent from a single source to a single recipient. Along the way these packets are forwarded by a number of routers between the source and recipient, according to forwarding table information that has been built up by configuration and routing protocol activity. This form of IP packet delivery is known as *unicast*.

However, some scenarios (for example, audio/video streaming broadcasts) need individual IP packets to be delivered to multiple destinations. Sending multiple unicast packets to achieve this is unacceptable because

- it would require the source to hold a complete list of recipients
- multiple identical copies of the same data would flow over the same links, increasing bandwidth requirements and costs.[1]

Instead, data to multiple destinations can be delivered using *multicast*.

[1] This is particularly inefficient on transport networks (such as Ethernet) that support one-to-many transmission.

Multicast allows the source to send a single copy of data, using a single address for the entire group of recipients. Routers between the source and recipients use the group address to route the data. The routers forward duplicate data packets wherever the path to recipients diverges.

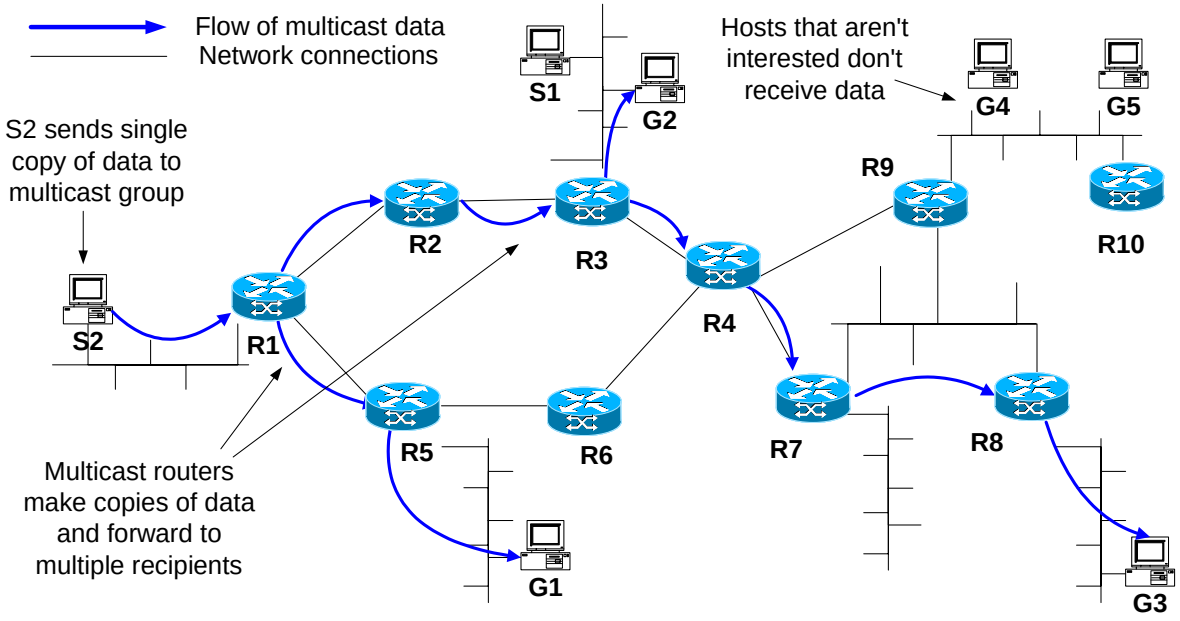
A *multicast group* identifies a set of recipients that are interested in a particular data stream, and is represented by an IP address from a well-defined range, as discussed in Chapter 2, **Multicast Addressing**. Data addressed to this IP address is forwarded to all members of the multicast group.

A source host sends data to a multicast group by simply setting the destination IP address of the datagram to be the multicast group address. Sources do not need to register in any way before they can begin sending data to a group, and do not need to be members of the group themselves.

In the following diagram, S2 sends a single copy of its multicast data addressed to the multicast group. The group consists of hosts G1, G2 and G3. The data is duplicated at routers R1 and R3 to ensure that it reaches all the hosts that are interested in this multicast data. G4 and G5 do not belong to the multicast group, and hence do not receive copies of the data.

In this and future diagrams

- **S** indicates a host sending multicast data
- **G** indicates a host which may or may not be a member of the multicast group
- **R** indicates a multicast-capable router.



Information about which parts of the network contain members of a particular multicast group is distributed as follows.

- Hosts who wish to receive data from the multicast group join the group by sending a message to a multicast router on a local interface, using a multicast group membership discovery protocol such as IGMP or MLD. These protocols are discussed in Chapter 3, **Multicast Group Membership Discovery Protocols**.

- Multicast routers communicate among themselves, using a multicast routing protocol such as PIM-SM, as discussed in Chapter 4, **Multicast Routing Protocols**. These protocols ensure that

    - multicast traffic reaches all of the recipients that have joined the multicast group

    - multicast traffic does not reach networks that do not have any such recipients (unless the network is a transit network on the way to other recipients)

    - the number of identical copies of the same data flowing over the same link is minimized.

To satisfy these requirements, multicast routing protocols calculate a *multicast distribution tree* of recipients.

1.3 History of Multicast

The need for IP multicast has been around for many years. In the early 1990s, some multicast-enabled routers did exist, but most routers on the IP network did not have multicast capabilities. In order to overcome this, multicast data packets were encapsulated within unicast packets and sent down 'tunnels' (predetermined routes though non-multicast routers) between the multicast-enabled routers. This network of tunnels was known as the Multicast Backbone (MBONE). On arriving at a multicast router, the unicast packet was decapsulated into a multicast packet again.

In the late 1990s, ISPs began to replace tunnels with *native multicast routing*, in which intermediate routers, although not multicast-enabled, were able to forward raw multicast packets without encapsulating them into unicast packets first.

To date, multicast has not been deployed in networks nearly as widely as unicast. One way of measuring the extent to which a networking technology is deployed is to count the number of Autonomous Systems (ASs) in the Internet that have deployed it. Whereas there are over 17,000 ASs in the Internet that support unicast routing today, there are only 485 that support multicast routing.

Support for multicast routing is, however, being slowly rolled out by service providers. The number of ASs supporting multicast routing has gradually increased over the last three years (in January 2004, the figure was 320). This reflects growing customer demand for multicast services, as well as the availability of next-generation technology making multicast a viable business case for service providers.

1.4 Overview of Multicast

Section 1.2, **What is Multicast?**, gave a basic introduction to the concept of IP multicast. In this section we look in a little more detail at the way that IP multicast is carried out, and introduce the main protocols used today. This leads into Chapters 2 to 7, where this material is explained and discussed in depth.

Multicast Addressing

Before a host can send data to a multicast group, or join the list of recipients for a group, it must know the address of the multicast group. Multicast IP address allocation is discussed in detail in Chapter 2, **Multicast Addressing**, but in outline works as follows.

- The Internet Assigned Numbers Authority (IANA) assigns multicast group addresses for well-known protocols and services.
- Other addresses are delegated for allocation by network administrators.
- Certain address ranges are assigned for use within local or administratively scoped boundaries.

In general, identification of the multicast group address is an application-level issue.

When joining a multicast group, hosts can opt to receive data sent to the group from any source, or to receive data sent to the group from one specific source only.

- To receive data from any source, hosts need to specify only the IP address of the multicast group. This is known as Any Source Multicast (ASM).
- To receive data from one particular source only, hosts must specify both the IP address of the multicast group and the IP address of the source. This is known as Source-Specific Multicast (SSM).

The advantages and disadvantages of ASM and SSM are discussed in Section 2.3, **ASM versus SSM**.

Multicast Group Membership Discovery Protocols

Once a host has identified a multicast group (and optionally a specific source) in which it is interested, the next stage is for the host to register that interest with a local multicast router. This router needs to know the multicast group memberships of all hosts directly connected to it.

Host memberships are communicated using a Multicast Group Membership Discovery (MGMD) protocol, which runs between a router and directly connected hosts. The MGMDs used in the Internet are IGMP for IPv4 and MLD for IPv6. Chapter 3, **Multicast Group Membership Discovery Protocols**, gives more details of MGMD protocols.

Multicast Routing Protocols

When a multicast router knows the group memberships of its directly connected hosts, it exchanges information with other routers enabling it to join or leave trees of multicast group recipients. These exchanges use a *multicast routing protocol*.

The three multicast routing protocols used to any significant extent today are Protocol Independent Multicast Sparse Mode (PIM-SM), Protocol Independent Multicast Dense Mode (PIM-DM) and Distance Vector Multicast Routing Protocol (DVMRP), with PIM-SM being particularly widespread. DVMRP was widely used in the MBONE in the past, but is now recommended for use only for compatibility with existing deployments. Bi-directional PIM (BIDIR-PIM) is still somewhat experimental and is not yet widely deployed. Two further protocols exist, Multicast OSPF (MOSPF) and Core-Based Trees Multicast Routing (CBT), but these have never been seriously deployed.

Once a tree has been constructed, multicast data is forwarded down the tree of recipients, with duplicate copies of the packets generated where the tree branches. The tree of multicast group recipients may be constructed in one of two ways, depending on the multicast routing protocol in use.

- In *opt-in* protocols, routers indicate which multicast groups they want to receive data for, in advance of that data flowing.

- In *opt-out* or *broadcast-and-prune* protocols, every router is initially assumed to want to receive multicast data, and so each tree initially spans every link in the network. Routers prune themselves from a given tree.

For further details, see Section 4.1.1, **Opt-in and Opt-out Protocols**.

The location of the root of any given multicast tree also depends on the multicast routing protocol in use. Multicast protocols use one or both of the following methods.

- *Source-based tree* protocols build a separate tree for each source sending data to a multicast group. Each tree is rooted at a router adjacent to the source, and sources send data directly to the root of the tree.

- *Shared tree* protocols build a single tree used for all sources sending to a multicast group. The tree is rooted at some selected node (in PIM, this router is called the Rendezvous Point, or RP). The protocols then use a protocol-specific mechanism to transport the multicast datagrams from the source to the root of the tree.

For further explanation, see Section 4.1.2, **Source-Based and Shared Tree Protocols**.

Another important difference between multicast routing protocols is the mechanism they provide for routers to locate the upstream multicast router (towards the root of the multicast tree). Routers need to communicate with the next upstream router when joining or leaving a multicast tree. This is discussed in Section 4.1.3, **Determining the Upstream Router**.

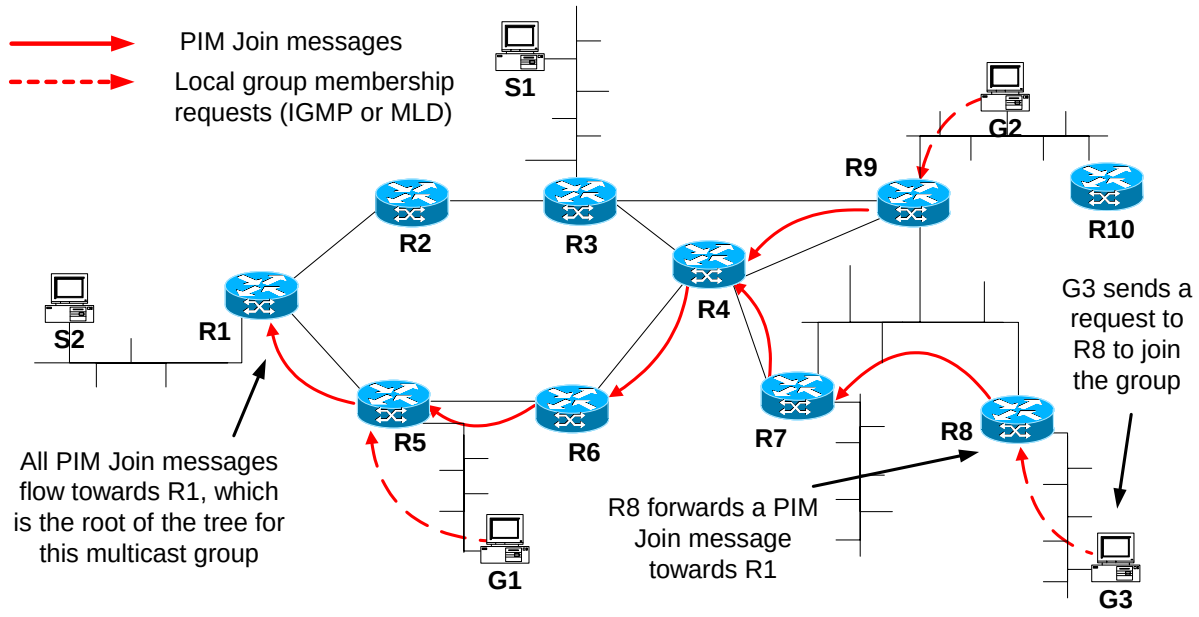
An important aspect of any multicast routing protocol is the interactions with the data plane, which by comparison with unicast routing protocols can be very complicated. In all multicast routing protocols, information gathered in the data plane, such as measurements of bandwidth used by each tree, must be passed to the control plane, because it influences protocol operation. This level of interaction with the data plane is not required in unicast routing. This is discussed briefly in Section 4.1.4, **Data Plane Interactions**, and in greater detail in Chapter 7, **Multicast Data Plane Operation**.

A detailed description of each of the most important multicast routing protocols (PIM-SM, PIM-DM, BIDIR-PIM, DVMRP and MOSPF) is given in Chapter 4, **Multicast Routing Protocols**. As an introduction, a brief overview of each is given here.

PIM-SM

PIM-SM is an opt-in multicast routing protocol, and can use either source-based trees or shared trees. It is the most widely used multicast routing protocol.

The diagram below shows how a multicast tree is constructed in PIM-SM. Hosts wishing to join the group, such as G1, G2 and G3, send group membership requests (using IGMP or MLD) to local routers. The routers then send PIM Join requests towards the root of the tree, which is R1.

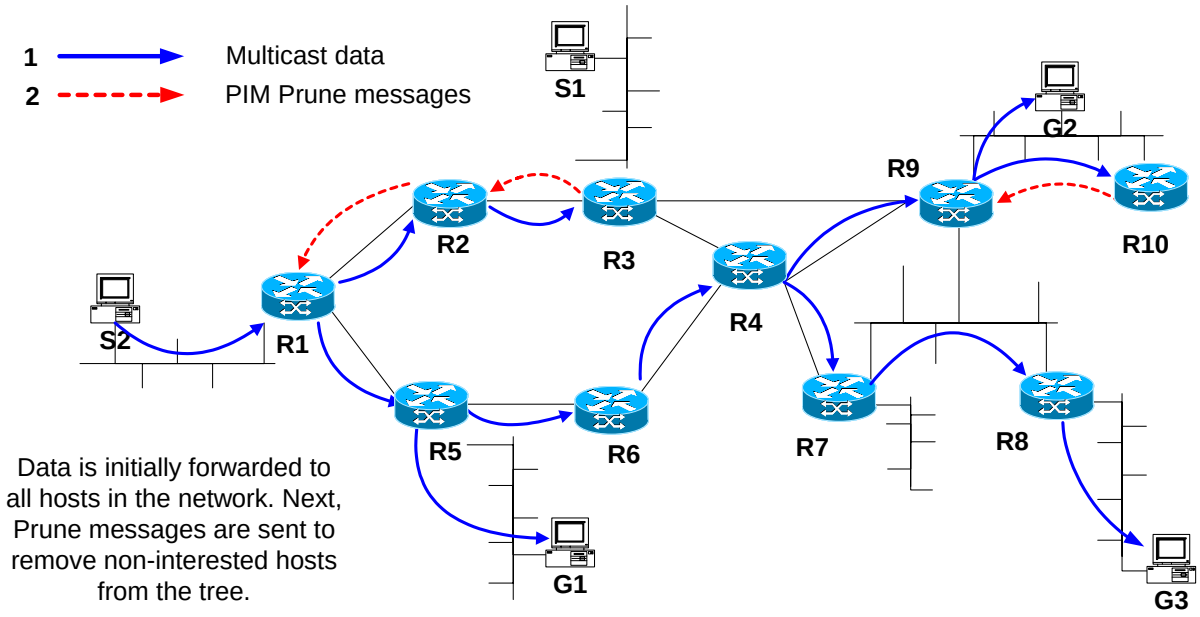


When data is sent to the multicast group, it will flow from R1 to G1, G2 and G3, as in the diagram shown earlier in Section 1.2, **What is Multicast?**.

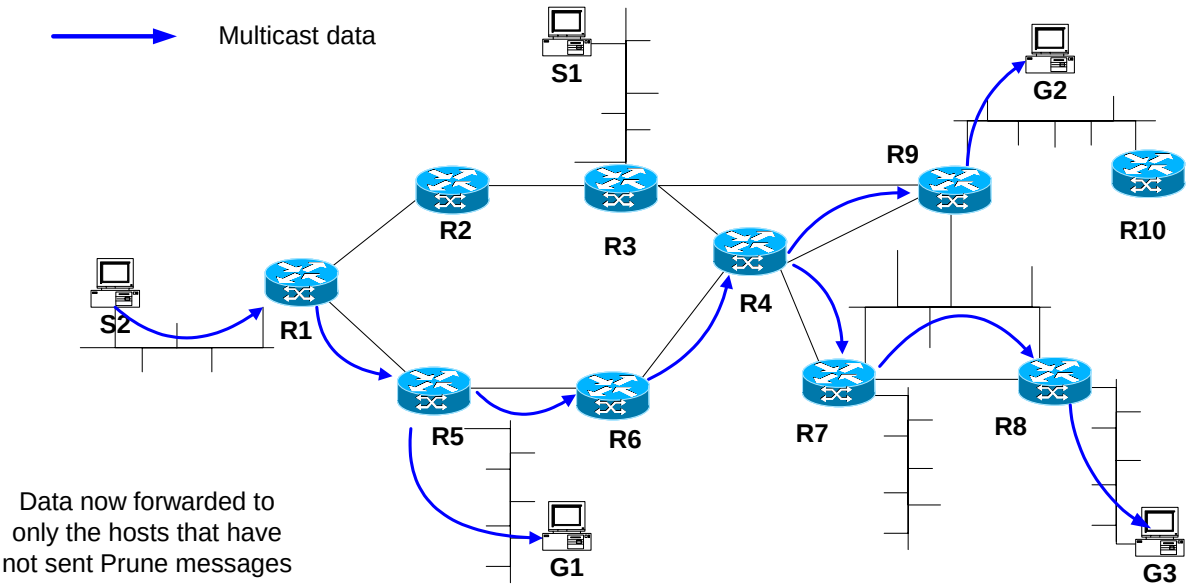
Various methods are used by PIM-SM to identify the Rendezvous Point (root of the tree) in shared tree networks, including the PIM Bootstrap Router (BSR) mechanism, Auto-RP, Embedded RP and Anycast RP. These are described in Section 4.2.4, **RP Discovery**.

PIM-DM

PIM-DM is an opt-out multicast routing protocol that uses source-based trees. The following diagram shows how a multicast distribution tree is constructed in PIM-DM. Multicast data is initially sent to all hosts in the network. Routers that do not have any interested hosts, such as R3 and R10, then send Prune messages to remove themselves from the tree.



The next time data is sent to the multicast group, it does not flow to the routers that have sent Prune messages; hence R3 and R10 do not receive multicast data.



BIDIR-PIM

BIDIR-PIM is based on PIM-SM, but allows data to flow in both directions along branches of the multicast tree. This makes the interactions with the data plane very much simpler. BIDIR-PIM does not support source-based trees, so SSM cannot be used with this protocol.

Typically, either PIM-SM or PIM-DM will be used throughout a multicast domain. However, they may be used together, or in conjunction with BIDIR-PIM, within a single domain. More information on mixed-mode PIM configurations is given in Section 4.2.5, **Mixed-mode PIM Configurations**.

DVMRP

DVMRP is an opt-out protocol that uses its own distance vector algorithm to compute routes in the network. It was widely used in the MBONE, but is now recommended for use only for compatibility with existing deployments.

MOSPF

MOSPF is an extension of the OSPF unicast routing protocol that acts as a combined unicast and multicast routing protocol. It differs significantly from every other multicast routing protocol in that MOSPF neighbors do not exchange tree-building messages, such as the Join messages used by PIM-SM or the Prune messages used by PIM-DM. Instead, the routers flood information about the location of all receivers throughout the network, and then each MOSPF router calculates each distribution tree on its own. MOSPF has never been seriously deployed.

For more information on the advantages, disadvantages and current deployment of each of these multicast routing protocols, see Chapter 4, **Multicast Routing Protocols**.

Interdomain Multicast Routing

The protocols described above provide *intradomain multicast routing*, which is routing within a single autonomous multicast routing domain. In some cases, multicasting across different networks requires the use of additional protocols. This and other aspects of interdomain multicasting are discussed in Chapter 5, **Interdomain Multicast Routing**.

Multicast Signaling

Multi-protocol label switching (MPLS) is the protocol of choice for extending IP unicast packet forwarding to improve efficiency and to provide traffic engineering and guaranteed quality of service.

There have been various proposals for performing IP multicast packet forwarding using MPLS, making use of point-to-multipoint Label Switched Paths. An overview of this topic is given in RFC 3353, but none of the detailed approaches has yet been standardized.

Chapter 6, **Multicast Signaling**, gives an overview of the main Internet-Drafts proposing MPLS extensions for multicast signaling.

1.5 Document Roadmap

This section summarizes the content of the rest of this paper.

- Chapter 2, **Multicast Addressing**, describes various aspects of multicast IP addressing, including how addresses are allocated, and the properties of Any Source Multicast (ASM) versus Source Specific Multicast (SSM).

- Chapter 3, **Multicast Group Membership Discovery Protocols**, describes the protocols used to manage membership of multicast groups. These include IGMP, MLD, IGMP Proxying and IGMP Snooping, and Multicast over ATM.

- Chapter 4, **Multicast Routing Protocols**, describes the major multicast routing protocols used today, including PIM-SM, PIM-DM, BIDIR-PIM, DVMRP and MOSPF. This chapter also includes a discussion of the advantages, disadvantages and extent of the current deployment of each major protocol.

- Chapter 5, **Interdomain Multicast Routing**, describes methods used for multicast routing between different multicast domains. These include MSDP, Multicast Border Routers, and BGMP.

- Chapter 6, **Multicast Signaling**, describes how the principles of MPLS are applied to multicast IP packet forwarding situations.

- Chapter 7, **Multicast Data Plane Operation**, describes how the data plane forwards multicast data, and the information that the data plane must make available to the control plane. These interactions are crucial to the design of a multicast router.

- Chapter 8, **Summary**, provides a summary of the conclusions drawn by this paper on the relative importance of the multicast protocols currently available.

- Chapter 9, **About Data Connection**, contains details of the author of this paper, of Data Connection and of Data Connection's range of portable software (including IP multicast).

- Chapter 10, **Glossary**, contains a glossary of some of the important terms used in this paper.

- Chapter 11, **References**, provides details of references made in this paper.

2. MULTICAST ADDRESSING

Multicast groups are allocated an IP address within a well-defined range (224.0.0.0/4 for IPv4 and FF::/8 for IPv6).

This chapter discusses various aspects of multicast addressing, namely

- the division and allocation of the multicast address space
- multicast addresses with local, administrative and global scope
- the difference between Any Source Multicast (ASM) and Source Specific Multicast (SSM).

2.1 Multicast Address Allocation

The IP addresses of multicast groups are allocated very differently from unicast IP address.

- Some addresses are allocated globally by IANA for well-known services, in the same way as TCP port numbers are allocated for unicast. In general, these are only allocated to services that are used for network control. For example, the address 224.0.0.4 is used by the DVMRP routing protocol. See http://www.iana.org/assignments/multicast-addresses for a full list.
- Other addresses are delegated for allocation by network administrators. For IPv4, the GLOP scheme (defined in RFC 3180) allocates the range 233.0.0.0/8 to the owners of Autonomous Systems (ASs). The Unicast-Prefix-based addressing scheme (defined in RFC 3306) provides a similar solution for IPv6 addresses. Rather than being based on an AS number, with this system every single subnet in the Internet gets a 32-bit multicast address range.
- Other addresses are only intended for local use, for example the administratively scoped region 239.0.0.0/8 (see below), and so are assigned independently in each individual network.

Source Specific Multicast (SSM), which is described in Section 2.3, **ASM versus SSM**, below, has a significant effect on the address allocation issue. When SSM is used, multicast group addresses are unique only for a particular source, and no longer have to be globally unique. To avoid clashes, the address range 232.0.0.0/8 has been allocated for use only with SSM.

2.2 Multicast Address Scoping

Not all multicast addresses have a global scope. In particular, the following address ranges have limited scope.

- 224.0.0.0/24 is the link-local scope region. Traffic sent to these addresses is only transmitted over a single link. This is used for control traffic, for example that from multicast routing protocols.

- 239.0.0.0/8 is the *administratively scoped* region. The key properties of administratively scoped multicast are as follows.

    - Packets addressed to administratively scoped multicast addresses do not cross configured administrative boundaries.

    - Administratively scoped multicast addresses are locally assigned, and hence are not required to be unique across administrative boundaries.

An older form of multicast scoping was *TTL scoping*, which used TTL limits to restrict the flow of multicast traffic. This form of scoping proved hard to configure and manage. Nowadays it is recommended to use administrative scoping instead.

2.3 ASM versus SSM

The basic difference between Any Source Multicast (ASM) and Source Specific Multicast (SSM) is as described in Section 1.4, **Overview of Multicast**. With SSM a host identifies a multicast data stream with a source and group address pair (S,G), rather than by group address alone (*,G).

SSM has the following advantages over ASM.

- *Simpler address allocation*. Since with SSM the multicast group address is local to the source, no global allocation mechanism or protocol is required.

- *Improved security and access control*. SSM is less susceptible to denial-of-service attacks, where unauthorized senders send to a multicast group.

- *Simplified interdomain multicast routing*. Receiving SSM traffic across domain boundaries is easier than with ASM. For example, there is no need to coordinate the root of a global shared tree, or use some other mechanism for source discovery.

However, SSM has the following drawbacks.

- *Knowledge of source address*. A recipient needs to know the source address of the SSM traffic. In scenarios where there is only one source, this may be easily distributed along with the group address. However, in situations where there are multiple, varying sources, the problem becomes more serious and some other mechanism is necessary.

- *IGMPv3*. As described in Chapter 3, **Multicast Group Membership Discovery Protocols**, in order for hosts to indicate their interest in particular sources, hosts and routers require support for IGMPv3 (or MLDv2), as previous versions of these protocols do not support this capability. IGMPv3 is becoming more widespread, but is still not ubiquitous.

- *Routing protocol support*. SSM requires a multicast routing protocol that supports source-based trees (see Section 4.1.2.1, **Source-Based Tree Protocols**) and source filtering. While the most important protocols support these, BIDIR-PIM does not support source-based trees, and MOSPF does not support source filtering.

In the future, SSM is likely to become very important, particularly for applications where there is one source, or a small number of sources, which are unchanging. Such applications may include audio/video broadcasting. However, despite the advantages of SSM, there are likely to remain applications where the sources are many and varying, such as video-conferencing, for which SSM is inappropriate.

3. MULTICAST GROUP MEMBERSHIP DISCOVERY PROTOCOLS

If a host wants to join a particular multicast group, it must inform the routers on its LAN of the address of the multicast group it wishes to join. It is then the responsibility of a router on the LAN to join the multicast group, as described in Chapter 4, **Multicast Routing Protocols**. Data from the multicast group is sent to this router.

This chapter describes the protocols that hosts use to communicate their multicast group memberships to neighboring routers. These are the Multicast Group Membership Discovery protocols (MGMDs). The protocols discussed in this section include IGMP (used for IPv4 multicast groups), MLD (used for IPv6), and ATM multicast.

3.1 Internet Group Management Protocol (IGMP)

Internet Group Management Protocol (IGMP) is the MGMD protocol used for IPv4 multicast groups. There have been three versions of IGMP. Versions 1 and 2 are very widely deployed, version 3 less so.

The following table shows the level of support for IGMP in Microsoft Windows hosts.

IGMP Version	Microsoft Windows Version
IGMPv1	Windows 95, Windows NT 4.0 (SP3 and earlier)
IGMPv2	Windows 98, Windows ME, Windows NT 4.0 (SP4 and later), Windows 2000
IGMPv3	Windows XP, Windows Server 2003

Each implementation of IGMP is required to interoperate with all previous implementations.

3.1.1 Basic IGMP operation

The following describes the basic operation of IGMP, common to all versions.

Note that a multicast router acts as both an IGMP host and an IGMP router in this and following descriptions, and as a result can respond to its own IGMP messages.

- If a host wishes to join a new multicast group, it sends an unsolicited IGMP Report message for that group.
- A local router picks up the IGMP Report message and uses a multicast routing protocol to join the multicast group.
- Periodically, a special router called the Querier (see Section 3.1.2, **Sending Group Membership Queries**) broadcasts IGMP Query messages onto the LAN to check which groups the local hosts are subscribed to.

- Hosts respond to the Query messages by sending IGMP Report messages indicating their group memberships.

- All routers on the LAN receive the Report messages and note the memberships of hosts on the LAN. If a router does not receive a Report message for a particular group for a period of time, the router assumes there are no more members of the group on the LAN, and removes itself from the multicast group (but see Section 3.1.4, **Improving Group Membership Latency**).

Note that all IGMP messages are raw IP datagrams, and are sent to multicast group addresses, with a TTL of 1. Since raw IP does not provide reliable transport, some messages are sent multiple times to aid reliability.

3.1.2 Sending Group Membership Queries

Only one router sends IGMP Query messages onto a particular LAN. This router is called the *Querier*. IGMPv1 depended on the multicast routing protocol to decide which router was the Querier. IGMPv2 introduced a Querier election process, which works as follows.

By default, a router takes the role of Querier. If a Querier receives an IGMP Query message from a router on the same interface and with a lower IP address, it stops being the Querier. If a router has stopped being the Querier, but does not receive an IGMP Query message within a configured interval, it becomes the Querier again.

3.1.3 Responding to Group Membership Queries

Ordinary LAN routers typically forward multicast traffic onto all other LAN segments. Therefore, the Querier does not need to know exactly which hosts on the LAN require data for a particular multicast group. It only needs to know that one host requires the multicast data.

To avoid a 'storm' of responses to an IGMP Query message, each host that receives this message starts a randomized timer for each group that it is a member of. When this timer pops, the host sends an IGMP Report message, which is addressed to that group. Any other hosts that are members of the group also receive the message, at which point they cancel their timer for the group.

This mechanism ensures that at most one IGMP Report message is sent for each multicast group in response to a single Query.

3.1.4 Improving Group Membership Latency

IGMPv2 introduced a Leave Group message, which is sent by a host when it leaves a multicast group for which it was the last host to send an IGMP Report message. Receipt of this message causes the Querier possibly to reduce the remaining lifetime of its state for the group, and to send a group-specific IGMP Query message to the multicast group.

Note that the Leave Group message in not used with IGMPv3, as its source address filtering mechanism (see below) provides the same functionality.

3.1.5 Source Address Filtering

IGMPv3 introduced an IGMP Version 3 Report message, to allow a host to include or exclude a list of source addresses for each multicast group that the host is a member of. Routers merge the source address requirements of different hosts for each group.

This feature is required to support SSM.

3.2 Multicast Listener Discovery (MLD)

Multicast Listener Discovery (MLD) is the MGMD protocol used for IPv6 multicast groups. MLD messages are a subset of the Internet Control Message Protocol for IPv6 (ICMPv6) messages.

There are two versions of MLD. MLDv1 provides equivalent functionality to IGMPv2, and MLDv2 provides equivalent functionality to IGMPv3. The message flows and mechanisms of MLD are identical to those of IGMP.

3.3 IGMP/MLD Proxying

In certain network topologies, a multicast router does not actually need to run a multicast routing protocol in order to deliver multicast traffic. Instead it can perform IGMP/MLD proxying.

One of the router's interfaces is configured to be its 'upstream' interface. This is the interface to the core network. All of its other interfaces are 'downstream' interfaces. The router performs the router and host parts of IGMP/MLD on its downstream interfaces as normal, but only performs the host part of IGMP/MLD on the upstream interface.

The router maintains a database of multicast group memberships on each of its downstream interfaces. This has two purposes.

- The router uses a merged version of these memberships when acting as a host on the upstream interface. It relays IGMP messages from its downstream interfaces directly to its upstream interface, so that if one of its downstream interfaces requests a multicast group membership, it passes the message on to the upstream interface to deal with.

- The router uses the database to forward multicast packets. When it receives a multicast packet from its upstream interface, it forwards the packet out of all the downstream interfaces that its database tells it have members of that multicast group.

There is a small additional requirement on multicast routing protocols in a network that uses IGMP/MLD proxying. This is that multicast routers upstream of the IGMP/MLD proxy router(s) must treat multicast traffic sent from within the IGMP/MLD proxying tree as if it came from a directly connected source.

The main advantages of IGMP/MLD proxying are that such routers are cheaper, easier to administer, and are independent of the multicast routing protocol used in the core network. The disadvantage is a restriction on the network topology. Only a simple tree topology is supported, in which a router has a single connection towards the core network and connections to one or more disjoint edge networks.

3.4 IGMP/MLD Snooping

Ordinary LAN switches typically forward multicast traffic onto all other LAN segments, to ensure that all receivers see it. However, if the receivers are sparsely distributed, this is a waste of network resources.

IGMP/MLD snooping allows LAN switches to forward multicast traffic more intelligently (albeit by violating the separation of functionality between the OSI communication layers). By processing IGMP/MLD Report messages, the switch can determine the locations of interested receivers, and suppress its forwarding of multicast traffic onto LAN segments where there are no interested receivers.

3.5 Multicast over ATM

Some ATM networks are capable of supporting layer 3 (IP) multicast using point-to-multipoint virtual circuits (P2MP VCs). These are set up using the Multicast Address Resolution Server (MARS) protocol.

The MARS protocol works as follows.

- Following layer 3 requests for local interfaces to join or leave multicast groups, an ATM endpoint registers its group memberships with the MARS for the ATM network.
- When it has a multicast packet to send, and no existing state for that destination address, an ATM endpoint queries the MARS to determine what VC it needs to create.

There are two VCs models, which can be chosen administratively on a per-group basis.

- Full mesh – each sender creates its own P2MP VC to the receivers.
- Multicast Server (MCS) – each sender creates its own point-to-point VC to the MCS, and the MCS creates a P2MP VC to the receivers. The MCS can be administratively chosen on a per-group basis.

To avoid the overhead of duplicating multicast group membership discovery at layer 2 and layer 3, MARS routers are recommended to suppress IGMP Query messages on interfaces to networks containing a MARS, and instead query the MARS.

4. MULTICAST ROUTING PROTOCOLS

Once a multicast router knows the group memberships of its directly connected hosts, it exchanges information with other routers to join the tree of multicast group recipients. Data sent to the multicast group is forwarded to all branches on the tree. This section describes the protocols used between routers to build a spanning tree of recipients for multicast data.

The three multicast routing protocols used to any significant extent today are Protocol Independent Multicast Sparse Mode (PIM-SM), Protocol Independent Multicast Dense Mode (PIM-DM) and Distance Vector Multicast Routing Protocol (DVMRP), with PIM-SM being particularly widespread. DVMRP was widely used in the MBONE in the past, but is now recommended for use only for compatibility with existing deployments. Bi-directional PIM (BIDIR-PIM) is still somewhat experimental and is not yet widely deployed. Multicast OSPF (MOSPF) and Core-Based Trees Multicast Routing (CBT) have never been seriously deployed.[2]

We first describe some general properties of multicast routing protocols. The remainder of this section gives a brief overview of each of the most important multicast routing protocols: PIM-SM, PIM-DM, BIDIR-PIM, DVMRP and MOSPF.

4.1 Properties of Multicast Routing Protocols

Four of the most important features of multicast routing protocols are the following.

- Whether they use opt-in or opt-out routing protocols.
- Whether they use *source-based* or *shared* trees.
- The methods they use to find the upstream router.
- The interactions they require with the data plane.

We will consider each of these features in turn. The table in Section 4.1.5, **Summary of Multicast Routing Protocols**, summarizes the properties of each multicast routing protocol with reference to the features above.

4.1.1 Opt-in and Opt-out Protocols

The basic difference between opt-in and opt-out protocols was described in Section 1.4, **Overview of Multicast**. A more detailed explanation is given here.

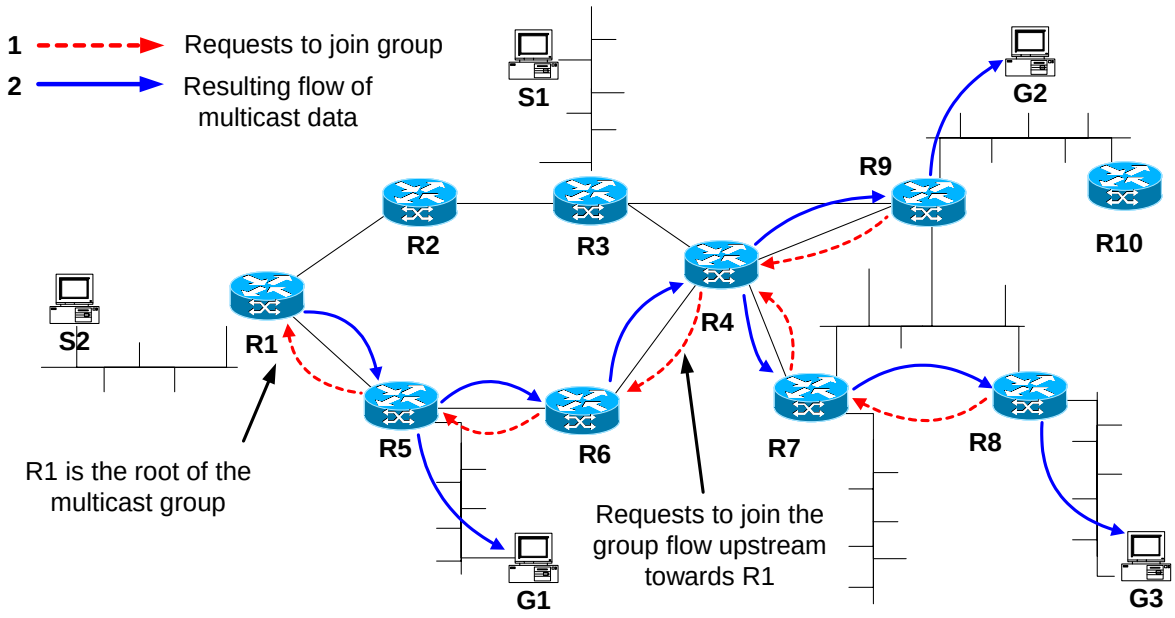
[2] Core Based Trees (CBT) is a shared tree multicast routing protocol which has never been deployed. It builds bi-directional trees, rooted at a 'core router', in a similar way to BIDIR-PIM. CBT version 2 is described in RFC 2189, dating from 1997.

4.1.1.1 Opt-in Protocols

Opt-in or sparse protocols are designed on the assumption that the receivers for any particular multicast group will be sparsely distributed throughout the network. In other words, they assume that most subnets in the network will not want any given multicast packet. PIM-SM, BIDIR-PIM, MOSPF and CBT are opt-in protocols.

Routers must indicate explicitly which multicast groups they want to receive data for, in advance of that data flowing, by sending a *Join* message to the upstream router. By default, they are not connected to multicast trees.

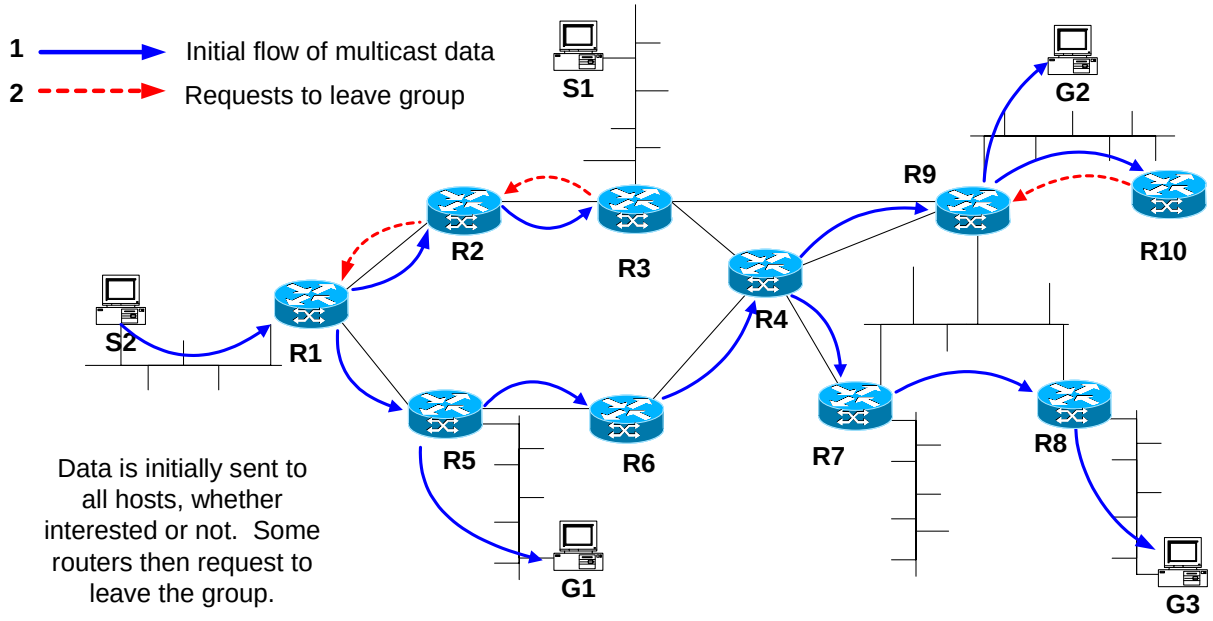
The diagram below shows routers R5, R8 and R9, which have interested hosts, sending requests to join the multicast group rooted at R1, and the resulting flow of multicast data. Data is only sent to routers that have opted to join the group.



4.1.1.2 Opt-out protocols

In *opt-out* or *broadcast-and-prune* or dense protocols, it is initially assumed that every router on the network wishes to receive multicast data, and data is sent to all routers. Routers wishing to remove themselves from the multicast tree must then send a *Prune* message to the upstream router. Messages are sent when they receive a multicast datagram sent to a group or from a source that they are not interested in. PIM-DM and DVMRP are opt-out protocols.

The following diagram shows how data is initially sent to all routers. R3 and R10, which do not have any interested connected hosts, then send requests to leave the multicast group. The next time data is sent down the multicast tree, it is not forwarded to these routers.



Opt-out protocols do not scale well in domains where most receivers do not wish to receive data, such as the Internet, so they are mostly used for individual small domains.

4.1.2 Source-Based and Shared Tree Protocols

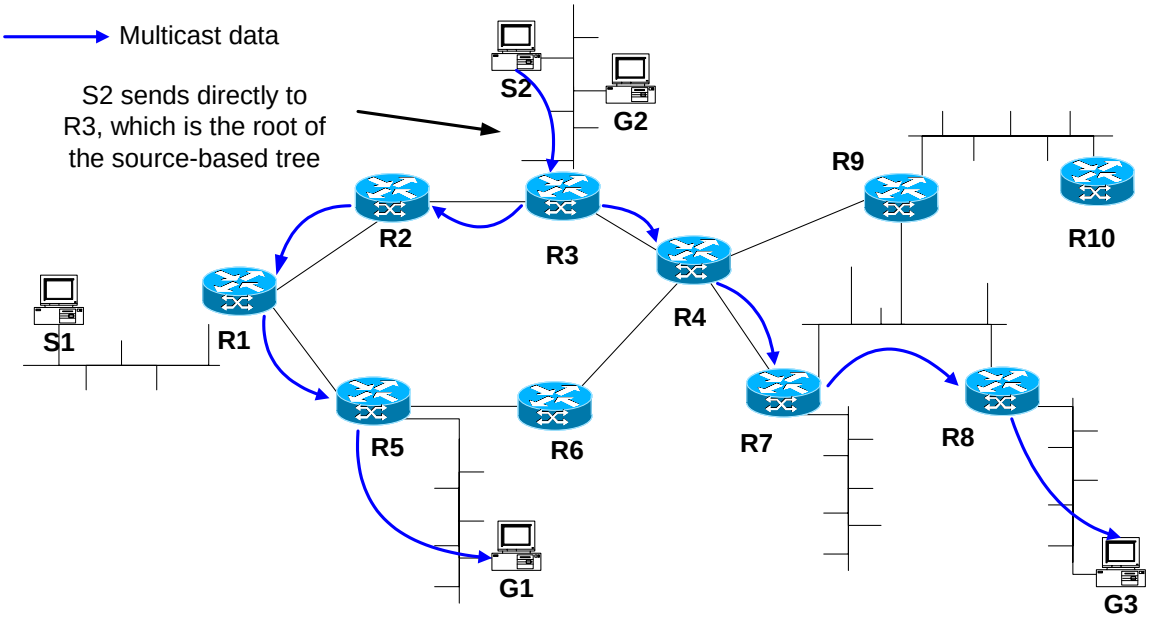
The location of the root of any given multicast tree depends on the multicast routing protocol in use. Multicast routing protocols may use source-based or shared tree methods to ensure that multicast data reaches the root of the tree in order to be forwarded downstream to all of the recipients.

4.1.2.1 Source-Based Tree Protocols

Source-based tree protocols build a separate tree for each source that sends data to a multicast group. Each tree is rooted at a router adjacent to the source. PIM-DM, DVMRP and MOSPF are source-based tree protocols. In addition, PIM-SM can run in a mode where it acts as a source-based tree protocol.

Routers wishing to join the multicast group must specify both the source and the group of the multicast data they would like to receive, by sending an (S,G) message to the next upstream router.

The following diagram shows a source-based tree rooted at R3. S2 sends data directly to the root of the tree. If S1 wished to send data to the same multicast group, it would need to use a new source-based tree based at R1.



The advantages of source-based tree protocols are that multicast data paths are always efficient, and they benefit from a simpler configuration than is necessary with shared tree protocols. However, source-based tree protocols suffer from scalability problems when there are large numbers of varying sources.

Source Specific Multicast (SSM), described in Section 2.3, **ASM versus SSM**, requires the use of source-based trees. Protocols that cannot support source-based trees, such as BIDIR-PIM, are therefore unable to use SSM.

4.1.2.2 Shared Tree Protocols

Shared tree protocols build a single tree that is used for all sources for a multicast group. The tree is rooted at some selected node (in PIM, this router is called the Rendezvous Point, or RP). The protocols then use a protocol-specific mechanism to transport the multicast datagrams from the source to the root of the tree. Typically, they are encapsulated in a unicast datagram and sent from a router adjacent to the source to the router at the root of the tree.

BIDIR-PIM and CBT are shared tree protocols. In addition, PIM-SM can run in a mode where it acts as a shared tree protocol. Note that only opt-in multicast protocols can use shared trees.

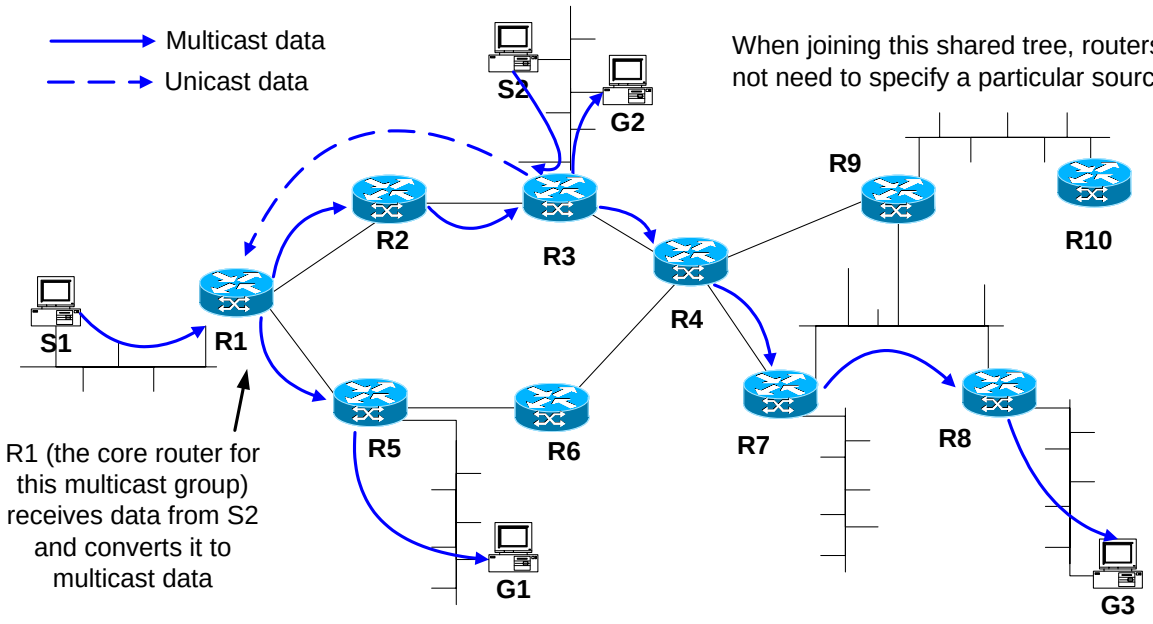
When a LAN router wishes to join a multicast group, it does not specify the source of the group it would like to join, but sends a (*, G) message to the next upstream router.

Shared tree protocols involve two additional factors.

- Multicast data must reach the root of the shared multicast distribution tree before it can be forwarded on. This is achieved in one of two different ways.
  - With *unidirectional* shared trees, each data packet is encapsulated by a source router, sent to the root of the tree (using unicast delivery) and decapsulated.
  - With *bi-directional* shared trees, the data flows natively back up the shared tree towards the root.

  PIM-SM uses the former approach and CBT and BIDIR-PIM use the latter, although CBT also performs unicast encapsulation in certain circumstances.
- The root of each shared multicast tree must be selected in some manner, such as pre-configuration or election.

The following diagram illustrates the flow of data in a unidirectional shared tree system, where dashed arrows indicate encapsulated unicast data flowing from S2 to the root of the tree, which is R1.



Shared tree protocols are better than source-based tree protocols when there are many potential sources, but may involve inefficient data paths, as multicast packets must always be sent to the root before being forwarded onward.

In a shared tree, the root of the tree (RP) must be selected in some manner, such as pre-configuration or election. The methods PIM protocols use to select the RP are described in Section 4.2.4, **RP Discovery**.

4.1.3 Determining the Upstream Router

All multicast routing protocols (with the exception of MOSPF) need to be able to perform a *reverse path forwarding* (RPF) lookup on the unicast address of either a data source or the root of a shared tree, in order to determine the next upstream interface for the multicast group.

The router uses the upstream interface as the outgoing interface for control packets (such as Join and Prune messages), and as the incoming interface for multicast data. Multicast data received on other interfaces is usually dropped or ignored, to reduce forwarding of duplicate packets and avoid forwarding loops.

Some protocols use their own mechanism to exchange the routing information necessary to perform RPF lookups. MOSPF uses its own link state mechanism, and DVMRP uses a distance vector mechanism. Other protocols, namely PIM and CBT, rely on a Multicast Routing Information Base (MRIB) populated by an external source.

The MRIB is similar to a unicast forwarding table, and may indeed be the same table used for unicast forwarding. However, in some cases it is desirable for the MRIB and the unicast forwarding table to differ, for example, when some routers do not support multicast. The PIM protocol is independent of the particular unicast routing protocol used to populate the MRIB, hence its name, Protocol Independent Multicast.

The MRIB can be configured with static routes, and can be populated by routing protocols such as M-ISIS (Multi Topology Routing in IS-IS) and MBGP (Multiprotocol Extensions for BGP), which distribute tagged routing information, in this case the information being tagged as being for use by multicast routing protocols.

It is generally accepted that the principle of separating the construction of the MRIB from the exchange of multicast routing information is a good one. Therefore PIM-DM is generally preferred over DVMRP, except where required for backwards compatibility.

4.1.4 Data Plane Interactions

In contrast to unicast routing protocols, multicast routing protocols involve a tangled set of interactions between the control plane and the data plane functionality of the router. A summary of this is given below, and a more detailed discussion is presented in Chapter 7, **Multicast Data Plane Operation**.

Unicast routing protocols use control messages to generate a forwarding table that the data plane can then use to forward unicast packets. Multicast routing protocols also generate a set of multicast forwarding table information that describes how to forward multicast packets; however, the interactions with the data plane do not stop there.

- In some protocols, a situation may arise where multiple upstream routers are forwarding the same traffic onto a LAN. The data plane must detect this, and inform the control plane, to allow the control plane to elect a single forwarder and avoid the duplication of data traffic.

- In some shared tree protocols, the data plane must be programmed with enough information to allow it to perform the encapsulation/decapsulation described in Section 4.1.2.2, **Shared Tree Protocols**.

- In some protocols, the state associated to multicast groups and sources is *soft state*, which is kept alive by the arrival of multicast data. This means that the data plane must inform the control plane of the arrival of multicast packets so that it can reset its state expiry timers.

- Some protocols, such as PIM-SM, allow transfer from a shared tree to a source-based tree. To enable this, the data plane must inform the control plane when traffic arrives from a new source on the shared tree, so that the control plane knows when to switch to a source-based tree.

- In opt-out protocols, the arrival of a multicast packet may trigger a control plane Prune message to prevent the arrival of future packets for that group on that tree.

It should be noted that while some multicast routing protocols require these interactions, others do not. In fact BIDIR-PIM and MOSPF do not require any interactions with the data plane at all, other than to program it with forwarding table information.

4.1.5 Summary of Multicast Routing Protocols

The following table summarizes the characteristics of each of the main multicast routing protocols, including the version of IP with which they are compatible.

Protocol	**Opt-in / Opt-out**	**Supports SSM**	**Tree Type**	**Upstream Router Info Via**	**IP Version**
PIM-SM	Opt-in	Yes	Shared or source-based	MRIB	IPv4 & IPv6
PIM-DM	Opt-out	Yes	Source-based	MRIB	IPv4 & IPv6
BIDIR-PIM	Opt-in	No	Shared	MRIB	IPv4 & IPv6
DVMRP	Opt-out	Yes	Source-based	Distance vector mechanism	IPv4
MOSPF	Opt-in	No	Source-based	Link state mechanism	IPv4

The following sections describe each of the protocols listed above in more detail.

4.2 Protocol Independent Multicast (PIM)

Protocol Independent Multicast (PIM) is a collection of multicast routing protocols that share

- a common control message format
- the property of depending on unicast routing information, as described in Section 4.1.3, **Determining the Upstream Router**.

PIM control messages are sent as raw IP datagrams (protocol number 103), either multicast to the link-local ALL-PIM-ROUTERS multicast group, or unicast to a specific destination.

There are two main PIM protocols, as follows.

- PIM Sparse Mode (PIM-SM) is an opt-in protocol that uses both shared and source-based trees. PIM-SM is the multicast routing protocol most widely used today.
- PIM Dense Mode (PIM-DM) is an opt-out protocol that uses source-based trees only. PIM-DM is mostly used for individual small domains.

In addition there is a third PIM protocol, Bi-directional PIM (BIDIR-PIM), which is based on PIM Sparse Mode but has significant differences. BIDIR-PIM is not yet particularly well deployed compared to PIM-SM or PIM-DM, though it is supported in the latest Cisco and Juniper routers.

Sections 4.2.1 to 4.2.3 describe these PIM protocols, including a discussion of the advantages and disadvantages of each protocol.

One of the important requirements of PIM-SM and BIDIR-PIM is the ability to discover the address of the root of the group's shared multicast distribution tree, known as its Rendezvous Point (RP). The various methods of RP discovery are described in Section 4.2.4, **RP Discovery**.

Typically, either PIM-SM or PIM-DM will be used throughout a multicast domain. However, they may also be used together within a single domain, using Sparse Mode for some groups and Dense Mode for others. This is known as Sparse-Dense Mode. Similarly, BIDIR-PIM may be used on its own, or it may be used in conjunction with one or both of PIM-SM and PIM-DM. Section 4.2.5, **Mixed-mode PIM Configurations**, describes how these mixed-mode configurations work.

4.2.1 PIM Sparse Mode

There have been many implementations of PIM Sparse Mode (PIM-SM) and it is widely used today. For a full discussion of the advantages and disadvantages of PIM-SM, see Section 4.2.1.5, **PIM-SM Summary**.

PIM-SM is an opt-in multicast routing protocol. To receive multicast data, routers must explicitly inform their upstream neighbors of their interest in particular groups and sources. This is done using PIM Join and Prune messages to join and leave a multicast distribution tree.

PIM-SM by default uses shared trees, with the trees rooted at a router called the Rendezvous Point (RP) for a group. Data is sent from a source to the RP via encapsulation in PIM control messages sent by unicast. PIM-SM also supports source-based trees, which may be used in the following circumstances.

- To avoid having to encapsulate data sent to an RP, the RP may join a source-based tree.
- To optimize the data path, a last-hop router may choose to switch from the shared tree to a source-based tree.
- For source-specific multicast (SSM), a last-hop router will join a source-based tree from the outset.

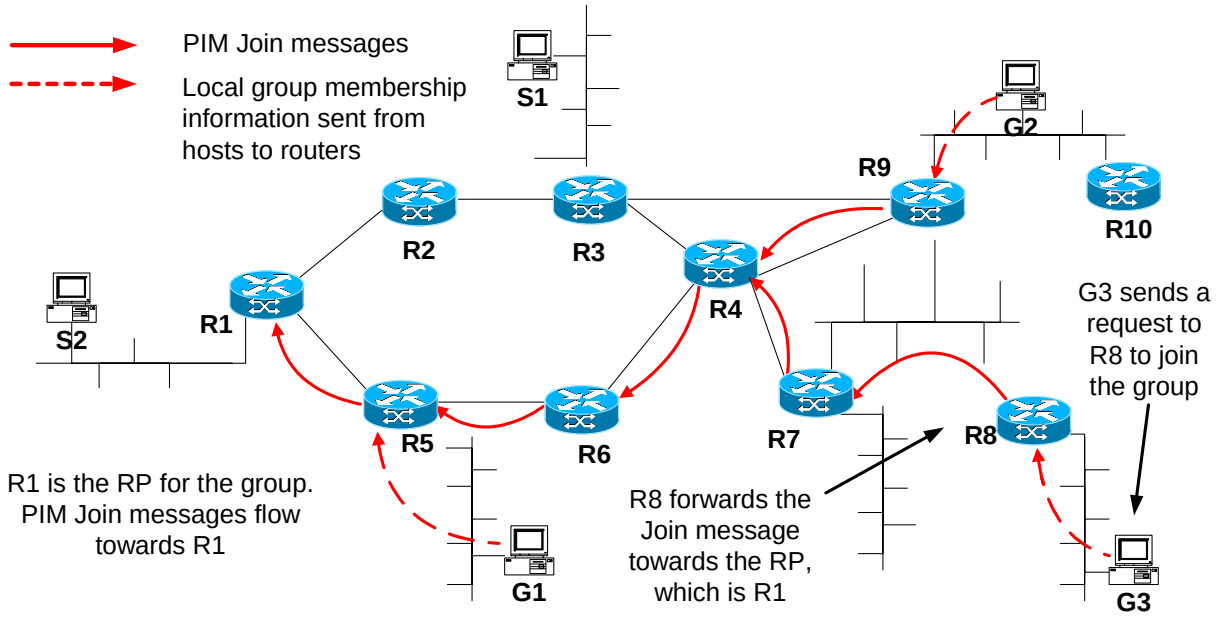
PIM-SM is a soft-state protocol. That is, all state is timed-out a while after receiving the control message that instantiated it. All PIM Join messages are periodically re-transmitted to keep the state alive.

The following sections describe significant aspects of PIM-SM in more detail.

4.2.1.1 Basic shared tree forwarding procedure

In an opt-in multicast routing protocol like PIM-SM, a multicast routing tree is constructed before any data is sent. Hosts indicate their interest in receiving data for a particular multicast group G using a mechanism such as IGMP or MLD. One of the routers on the host's LAN is elected the Designated Router (DR) for the LAN. The DR is responsible for joining the multicast group and forwarding multicast traffic to the LAN on behalf of its local receivers.

The following diagram shows how a multicast routing tree is constructed. G3 sends an expression of interest in a particular multicast group to its DR (R8), which then sends a PIM (*,G) Join message towards the RP for that group. G2 and G1 also send messages to their DRs in the same way. The RP in the diagram below is R1.

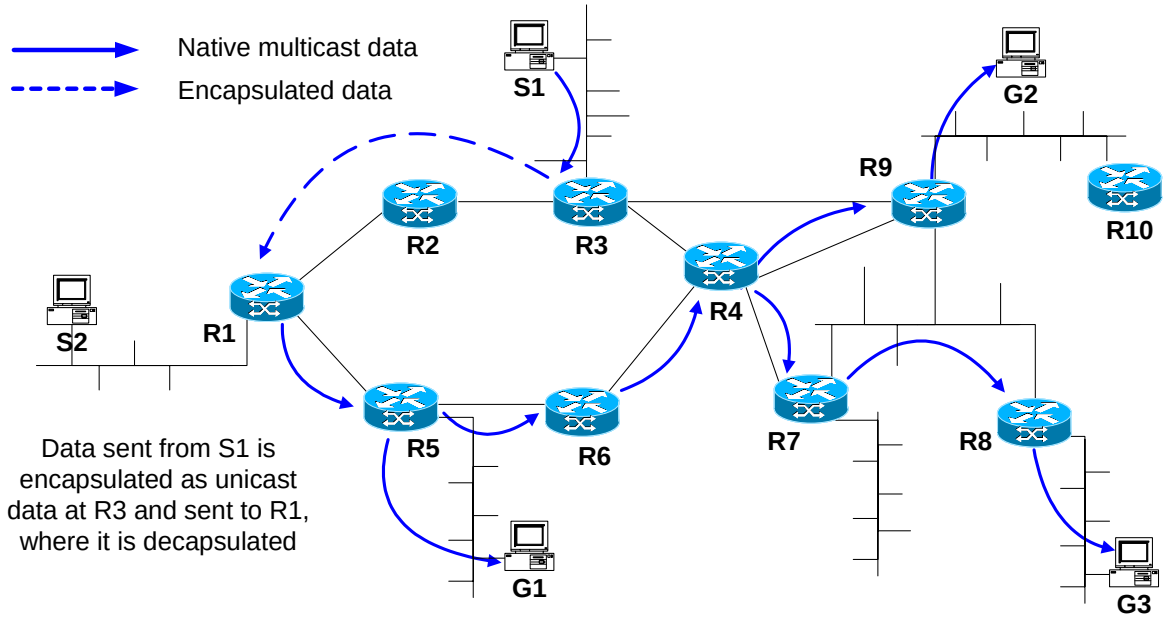


Routers forward the Join message by looking for the closest interface upstream towards the RP. As the Join message is forwarded hop-by-hop towards the root of the tree, it causes (*,G) forwarding state to be instantiated at each router it passes through.

Eventually, the Join message reaches either the RP or another router that already has (*, G) forwarding state. As more hosts join the group G, their Join messages converge on the RP, forming a shared distribution tree for G called the Rendezvous Point Tree (RPT).

The following diagram illustrates the flow of multicast data through the shared RPT constructed in the previous diagram. When the source S1 sends data to the group G, the DR on the source's LAN (R3 in the diagram below) receives the data packets, encapsulates each one in a PIM Register message and unicasts them to the RP (R1 in the diagram below).

When the RP receives the encapsulated data packets, it decapsulates them and forwards them out down each branch of the RPT. Each router on the RPT receives the data from its upstream neighbor and forwards it on, replicating the data packets as necessary. In this way the data is sent to all interested receivers, following the opposite path to the PIM Join messages.



When a host no longer wishes to receive data for a multicast group, it signals this with IGMP, MLD or whatever other mechanism is in use. If the host's DR then has no dependent receivers, it will remove itself from the RPT by sending a PIM (*,G) Prune message towards the RP.

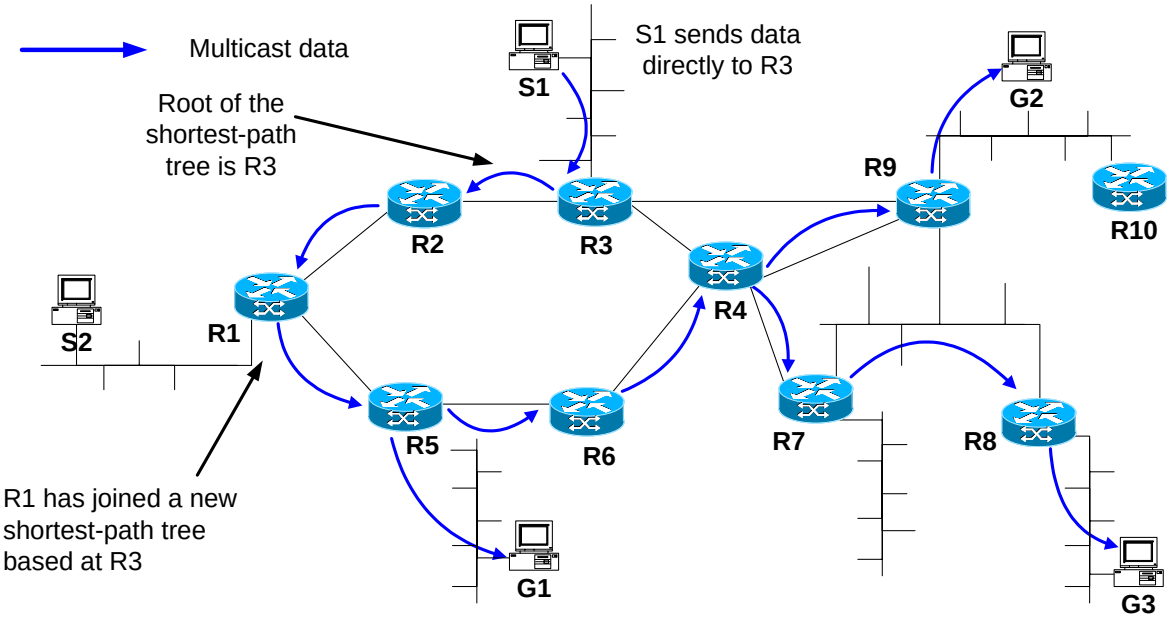
4.2.1.2 RP switch to source-based tree

Encapsulation and decapsulation of the data packets sent from the source to the RP may be expensive in terms of router CPU time. Therefore, the RP is permitted to join a source-based tree, which does not require encapsulation. Instead, the root of the tree is a router adjacent to the source.

This type of multicast tree is constructed in much the same way as above, but when sending Join messages, routers specify both the source and the group for the multicast data they would like to receive, sending a (S,G) message to the next upstream router.

This Join message is forwarded hop-by-hop, as with the (*,G) Join messages, causing (S,G) forwarding state to be instantiated at each router it passes through. Eventually, the Join message reaches either a router on the same LAN as S or another router that already has (S,G) state, forming a branch of a source-based distribution tree called the Shortest-Path Tree (SPT). The (S,G) data then gets forwarded down this SPT towards the RP.

The following diagram illustrates the flow of multicast data when a SPT is constructed for S1. The RP (R1) has joined an SPT based at R3, which is on the same LAN as S1.



Once the source-based tree has been joined and the data is flowing natively down the SPT to the RP, the RP forwards data on down the RPT. It also begins discarding any received encapsulated Register messages, since these are no longer needed after switching away from a shared tree. To indicate that it is no longer interested in the Register messages, the RP sends a PIM Register-Stop message to the DR. While the source is still active, the DR periodically sends Null-Register messages to the RP ('Null' indicating that they contain no encapsulated data) to indicate that the source remains active. The RP responds to each by sending another Register-Stop message.

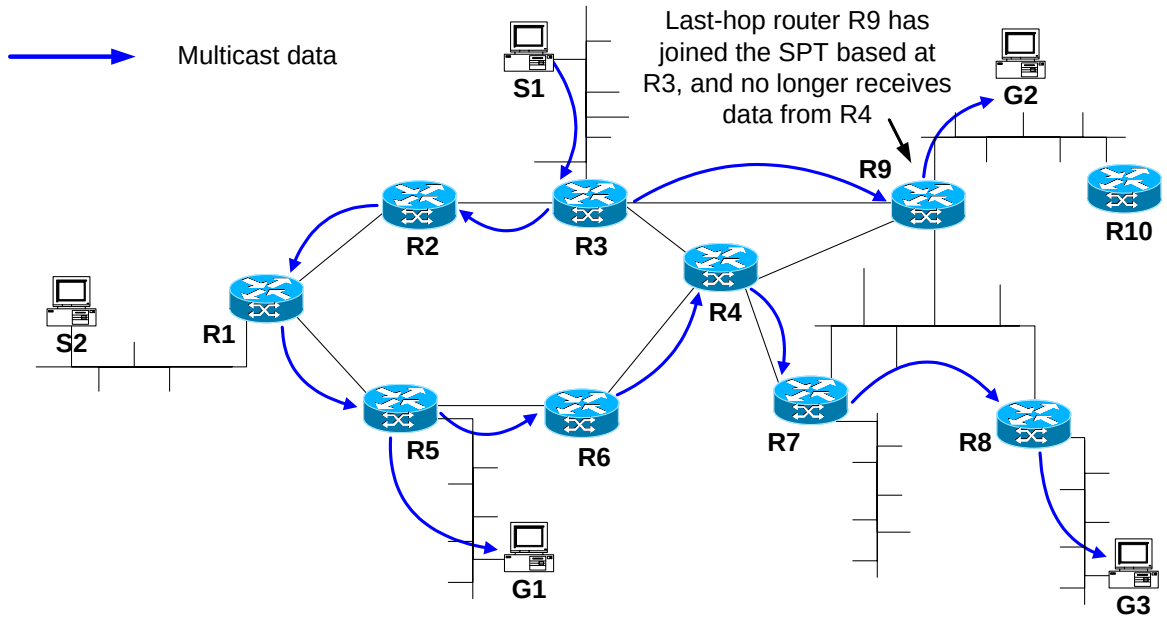
If another source wished to send data to the same multicast group, it would be necessary to construct a new SPT. Source-based tree protocols generate a separate SPT for each source in the network.

4.2.1.3 Last-hop switch to source-based tree

When data is flowing from a source to the RP (either Register-encapsulated or natively over the SPT) and then down the RPT to receivers, it is unlikely to be following the most efficient route. Therefore, last-hop routers are permitted to switch from the RPT to the SPT. If and when they do so is implementation-dependent. The most common behavior is either to switch to the SPT immediately (as soon as the first packet arrives over the RPT) or never to switch at all. However, in principle an implementation could switch to the SPT when the data flow rate for an (S,G) reaches a certain threshold.

To switch to the SPT, the last-hop router sends an (S,G) Join message towards the source S, as when the RP switches to the SPT (see above). Eventually the SPT branch is set up and data starts flowing down it. At this point, the last-hop router (or a router upstream of it) starts to receive two copies of the (S,G) data, one from the RPT and one from the SPT. It therefore needs to prune off the RPT, but for this source only. To do this, it sends a special Prune message, called an (S,G,rpt) Prune, towards the RP. This overrides the (*,G) Join message and prevents (S,G) data being forwarded down the RPT for G.

The following diagram illustrates the flow of data when last-hop router R9 joins the SPT for S1.



4.2.1.4 SSM forwarding procedure

When hosts join a multicast group and specify a specific source, PIM-SM does not use the RP mechanism described above. Instead, the DR simply sends an (S,G) Join message towards the source, joining the SPT immediately.

4.2.1.5 PIM-SM Summary

The advantages of PIM-SM are as follows.

- Like all PIM protocols, it is protocol-independent. In other words, its operation as a multicast routing protocol is independent of the particular type of unicast routing protocol operating alongside it in the network.
- It scales well across large networks.
- Sparse mode means that information only needs to be held at those routers in the network that are part of a distribution tree.
- It supports the use of both SSM and ASM.
- It can support either shared trees (with the advantage of not needing to record per-source state) or source-based trees (in which the data path is more efficient).
- It can use MSDP, SSM or Embedded RP for an inter-domain solution (see Chapter 5, **Interdomain Multicast Routing**).

The disadvantages of PIM-SM are as follows.

- In shared trees, register-encapsulation and decapsulation between the source and RP can be inefficient.
- Group-to-RP mapping has yet to be standardized.
- Many interactions with the data plane are required, which can affect the overall efficiency of routers.

Version 1 of PIM-SM was created in 1995, but was never standardized by the IETF. It is now considered obsolete, though it is still supported by Cisco and Juniper routers. Version 2 of PIM-SM was standardized in RFC 2117 (in 1997) and updated by RFC 2362 (in 1998). Version 2 is significantly different from and incompatible with version 1. However, there were a number of problems with RFC 2362, and a new specification of PIM-SM version 2 is currently being produced by the IETF.

4.2.2 PIM Dense Mode

PIM Dense Mode (PIM-DM) is less common than PIM-SM, and is mostly used for individual small domains, as it does not scale well across larger domains. For a discussion of the advantages and disadvantages of PIM-DM, see Section 4.2.2.2, **PIM-DM Summary**.

PIM-DM is a multicast routing protocol designed with the opposite assumption to PIM-SM. It assumes that the receivers for any multicast group are distributed *densely* (rather than sparsely) throughout the network, and therefore that most (or at least many) subnets in the network will want any given multicast packet. In other words, it is an opt-out protocol.

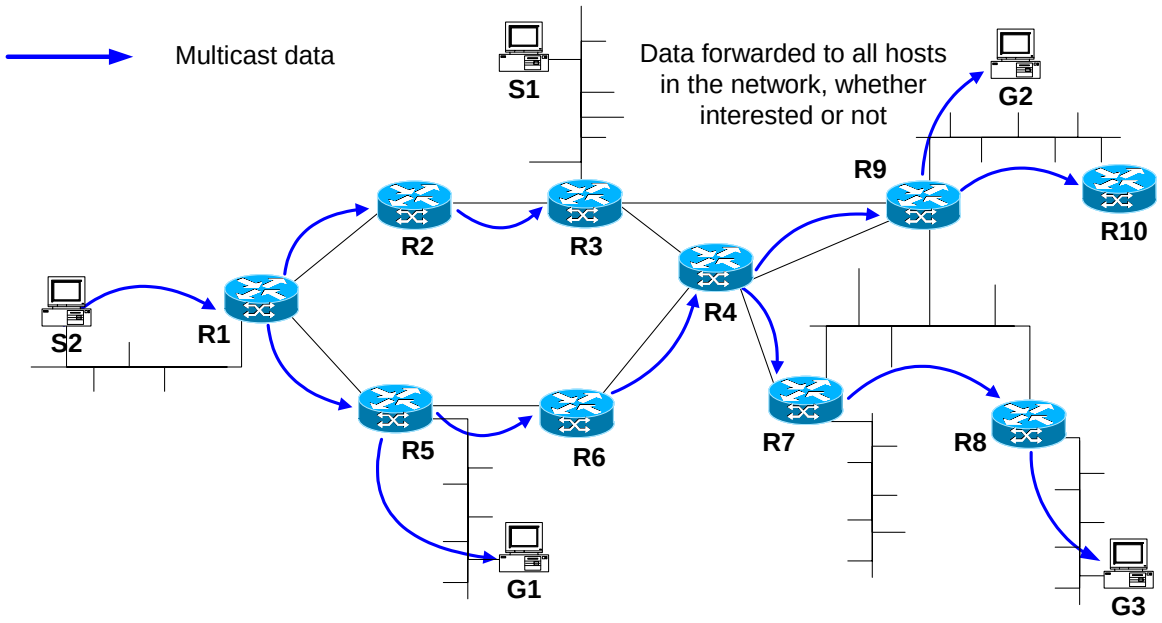
PIM-DM uses source-based trees to distribute data, as described in Section 4.1.2.1, **Source-Based Tree Protocols** (shared trees are never used with opt-out protocols). However, rather than having a mechanism in which routers explicitly join the tree, it is assumed to begin with that every link is a branch of the tree. Links on which the data is not required are removed from the tree using PIM Prune messages.

4.2.2.1 Forwarding operation

In a dense multicast routing protocol, when a source first starts sending data to the multicast group, each router on the source's LAN receives the data and forwards it to all its PIM neighbors and to all links with directly attached receivers for the data.

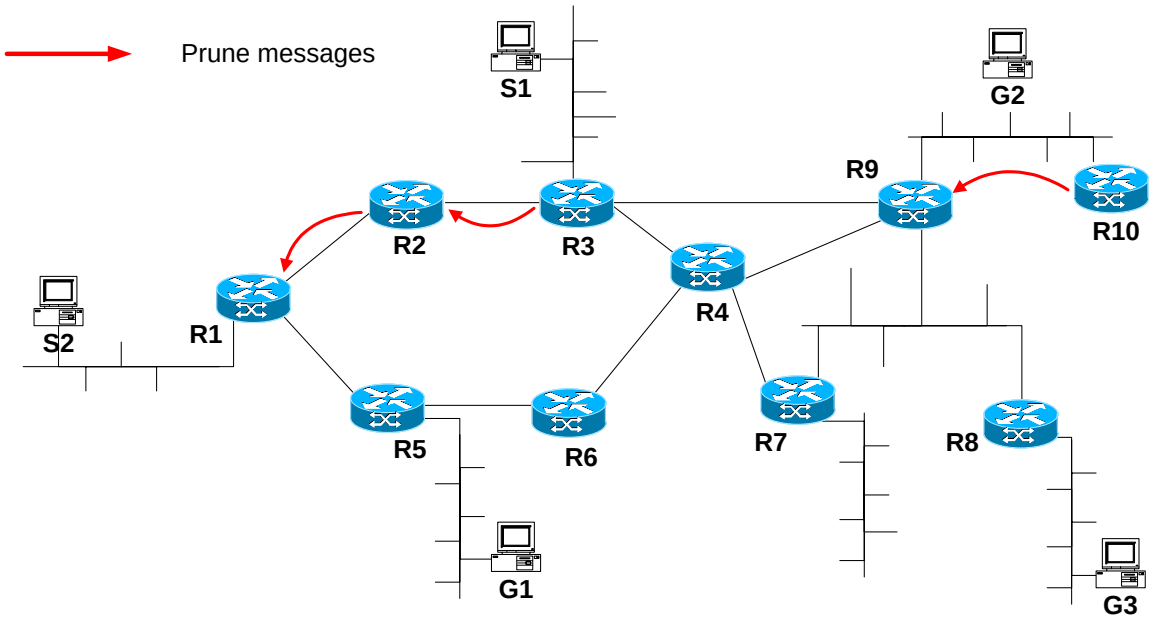
Each router that receives a forwarded packet also forwards it in the same way, but only after checking that the packet arrived on the interface closest to the source. If not, the packet is dropped. This mechanism prevents forwarding loops from occurring.

In this way, the data is flooded to all parts of the network. The following diagram illustrates the flooding of data from source S2 throughout a PIM-DM domain, where only hosts G1 and G2 are interested in receiving the multicast data stream.

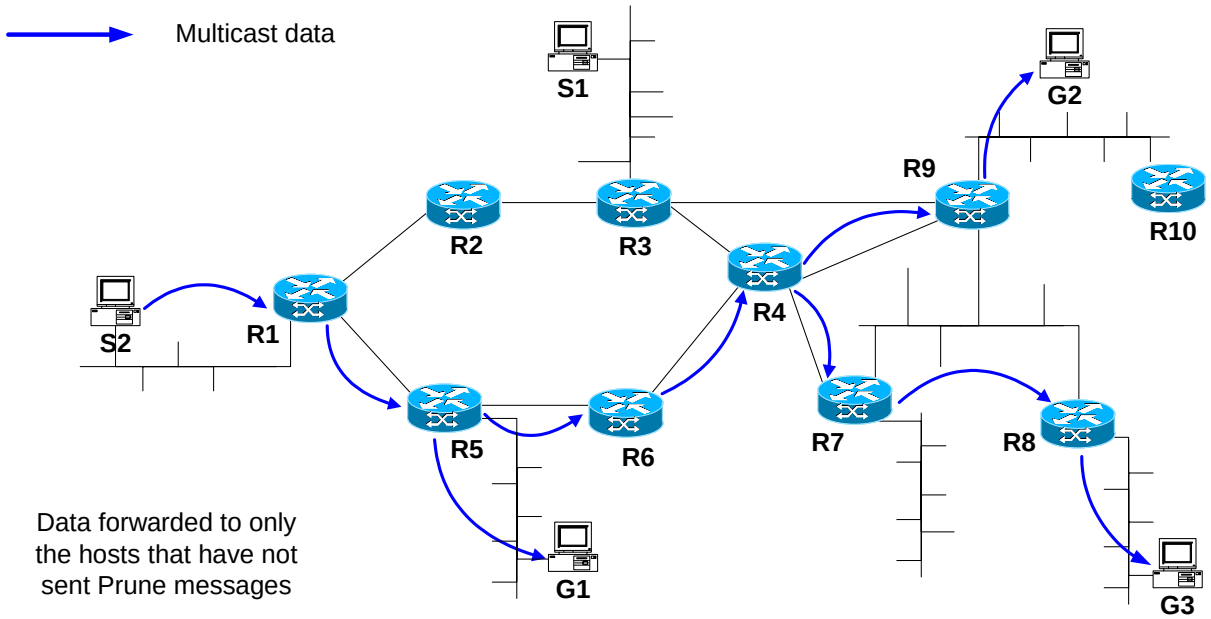


There may be routers that have no need of the data, either for other PIM neighbors or for directly connected receivers (for example, R3 in the previous diagram has no need of the data, since it has no neighbors and no connected receivers). These routers respond to receipt of the data by sending a PIM (S,G) Prune message upstream, which instantiate (S,G) Prune state in the upstream router, causing it to stop forwarding the data to its downstream neighbor.

This may mean that the upstream router (R2 in the previous diagram) also no longer needs the data, triggering it to send a Prune message to its upstream neighbor in the same way. This 'broadcast and prune' behavior means that eventually the data is only sent to those parts of the network that require it. The following diagram shows the flow of Prune messages that would occur in the situation shown in the previous diagram.



The following diagram shows the resulting data flow after all the Prune messages shown above have been processed.



Eventually, the Prune state at each router will time out, and data will begin to flow back into the parts of the network that had previously been pruned. This will trigger further Prune messages from receivers, and the Prune state will be instantiated once more.

If a new receiver wishes to join the multicast group but is located in a part of the network that is currently pruned from the tree, its local router sends a PIM (S,G) Graft message upstream, instructing the upstream router to rejoin the multicast tree. Graft messages are acknowledged with PIM Graft-Ack messages. This is the only time explicit acknowledgement is performed in PIM.

4.2.2.2 PIM-DM Summary

The advantages of PIM-DM are as follows.

- It is an efficient protocol when the receivers genuinely are densely distributed throughout the network.
- Like all PIM protocols, it is protocol-independent.
- It supports both SSM and ASM.
- It does not use RPs, which makes it simpler than PIM-SM to implement and deploy.

The disadvantages of PIM-DM are as follows.

- All routers need to store per-source state for every source in the domain.
- It does not scale well in domains where most receivers do not wish to receive data, such as the Internet, so it is mostly used for individual small domains.

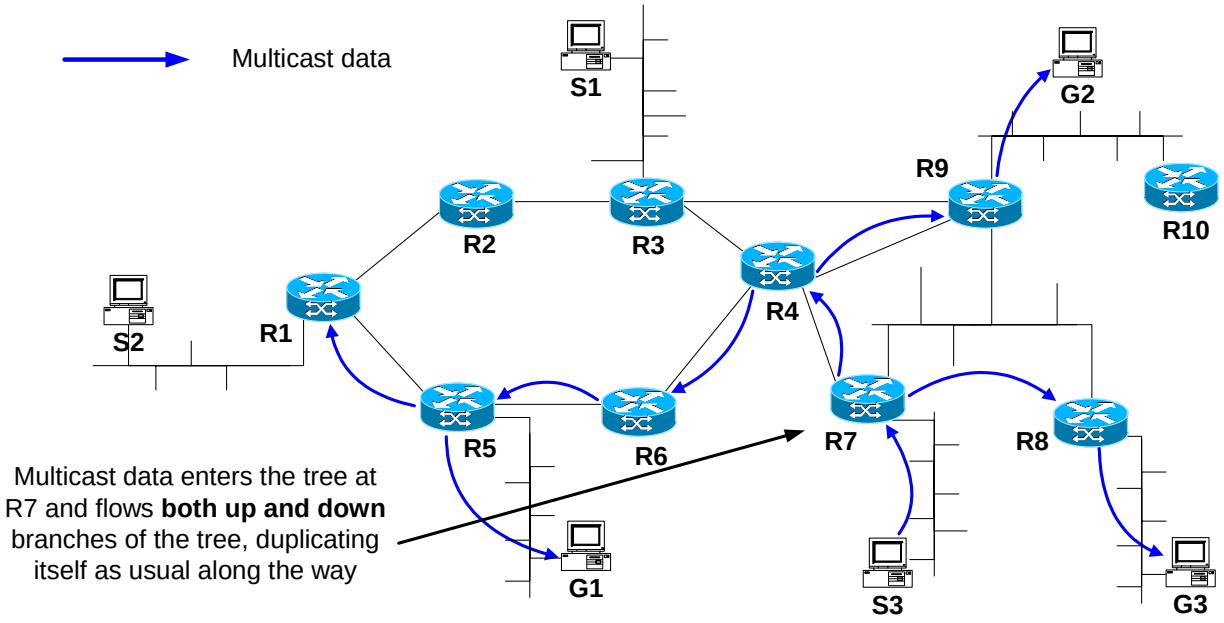
The development of PIM-DM has paralleled that of PIM-SM. Version 1 was created in 1995, but was never standardized. It is now considered obsolete, though it is still supported by Cisco and Juniper routers. Version 2 of PIM-DM is currently being standardized by the IETF. As with PIM-SM, version 2 of PIM-DM is significantly different from and incompatible with version 1.

4.2.3 Bi-directional PIM

Bi-directional PIM (BIDIR-PIM) is a third PIM protocol based on PIM-SM. BIDIR-PIM is not yet particularly well deployed compared to PIM-SM, PIM-DM or DVMRP, though it is supported in the latest Cisco and Juniper routers. For a discussion of the advantages and disadvantages of BIDIR-PIM, see Section 4.2.3.1, **BIDIR-PIM Summary**.

The main way BIDIR-PIM differs from PIM-SM is the method used to send data from a source to the RP. In PIM-SM data is sent using either Register-encapsulation between the source and the RP of a shared tree, or directly via a source-based tree. However, in BIDIR-PIM data can flow from the source to the RP along a branch of the shared tree, which is *bi-directional*. This means that data can flow in either direction along any given branch.

The following diagram illustrates the flow of data from source S3 to a group G though a BIDIR-PIM domain, where R1 is the RP for G, and G1, G2 and G3 are subscribed to receive multicast data from G. Data enters the tree at R7, which is on a branch of the multicast tree. It flows both upstream towards the RP, and downstream towards receivers, duplicating itself where required. Note that this is carried out using (*,G) ASM forwarding state.



BIDIR-PIM's other major differences from PIM-SM are as follows.

- There are no source-based trees, and in fact no (S,G) state at all. There is therefore no option for routers to switch from the shared tree onto a source-based tree, and SSM is not supported.
- To avoid forwarding loops, for each RP one router on each LAN is elected the Designated Forwarder (DF), which is a different method to PIM-SM. The DF is elected at RP discovery time using a new PIM DF-Election message.
- As there is no Register-encapsulation, the PIM Register and Register-Stop messages are never used.
- The forwarding rules are very much simpler than in PIM-SM, and there are no data-driven events in the control plane at all.

4.2.3.1 BIDIR-PIM Summary

The advantages of BIDIR-PIM are as follows.

- The lack of (S,G) state means that it scales very well when there are many sources for each group.
- In contrast to PIM-SM, no Register-encapsulation is required.
- It is a simple protocol to implement, with very simple interactions with the data plane.
- Like all PIM protocols, it is protocol-independent.

The disadvantages of BIDIR-PIM are as follows.

- The lack of source-based trees means that traffic is forced to remain on the possibly inefficient shared tree.
- It only supports ASM, though in practice BIDIR-PIM would always be used in combination with PIM-SM, with some groups (including SSM) supported on SM and some on BIDIR.
- It is not yet clear how inter-domain BIDIR-PIM is to be implemented, since it is not possible to use MSDP, SSM or Embedded RP with BIDIR-PIM.
- Group-to-RP mapping has yet to be standardized.

There have been two proposed specifications for Bi-directional PIM. The first was described in **draft-farinacci-BIDIR-PIM**, which dates from 1999. The protocol described here is a replacement that is simpler than the first version, and has some improvements. It is described in **draft-ietf-pim-bidir**.

4.2.4 RP Discovery

A PIM-SM or BIDIR-PIM router needs to be able to discover the address of the RP (the Rendezvous Point, which is the single router at the base of the shared tree) for any given multicast group. RP discovery protocols must be able to address the following issues.

- *Load balancing / scalability*. A router encounters a certain amount of load for each group for which it is the RP, in particular for Register decapsulation in PIM-SM. Therefore, it should ideally be possible to distribute this load evenly across several RPs.
- *Fault-tolerance*. An RP is a single point of failure for a multicast network, so ideally there should be a mechanism to cope with the failure of an RP.
- *Ease of configuration*. Every router in a PIM-SM or BIDIR-PIM domain must have the same set of group-to-RP mappings. Ideally, this should be achievable without requiring too much work from the network operators.

Various methods of RP discovery are described in the following sections. These include the following.

- Static configuration
- Bootstrap Router (BSR)
- Auto-RP
- Embedded RP
- Anycast RP

4.2.4.1 Static configuration

The simplest method from a technical point of view is to configure the group-to-RP mappings on each router statically. The advantages of this scheme are its technical simplicity and ability to support any desired group-to-RP mapping scheme. The disadvantages are the potentially large amount of configuration effort required in a large network, and the lack of any support for fault tolerance.

4.2.4.2 Bootstrap Router (BSR) Mechanism

The PIM Bootstrap Router (BSR) mechanism, which is specified in **draft-ietf-PIM-SM-bsr**, works as follows.

- Certain routers in the domain are configured as Candidate BSRs. One of these routers is elected as the BSR using PIM Bootstrap messages flooded throughout the domain.
- Certain routers in the domain are configured as Candidate RPs. Each Candidate RP sends PIM Candidate-RP-Advertisement messages to the elected BSR.
- The BSR chooses a subset of Candidate RPs and floods this information as an RP-Set to all other routers within the domain using PIM Bootstrap messages.
- Individual routers algorithmically determine the RP for any given group from the information in the received RP-Set.

The BSR mechanism provides fault tolerance and load balancing, and requires little configuration except at the Candidate BSRs and Candidate RPs. Its disadvantage is its technical complexity.

4.2.4.3 Auto-RP

Auto-RP is a non-standardized protocol designed by Cisco that is similar to BSR, with the same advantages and disadvantages. It works as follows.

- Certain routers in the domain are configured as Candidate RPs. Each Candidate RP sends RP Announcement messages over UDP to the multicast group address CISCO-RP-ANNOUNCE.

- Certain routers in the domain are configured as RP-Mapping Agents. These routers listen to the CISCO-RP-ANNOUNCE multicast group and receive the RP Announcement messages. They select a subset of Candidate RPs and send RP Discovery messages for these RPs over UDP to the multicast group address CISCO-RP-DISCOVERY.

- All routers listen to the CISCO-RP-DISCOVERY group and receive the RP Discovery messages. This information enables them to determine the RP for any given group.

In order to avoid the chicken-and-egg situation of multicast to the CISCO-RP-ANNOUNCE and CISCO-RP-DISCOVERY groups, either these groups are configured to be in dense mode, or all routers must have a static RP assignment for these groups.

4.2.4.4 Embedded RP

Embedded RP is a method of embedding the RP address of a group within the group address itself. It is available only on IPv6, and uses a subset of the IPv6 address space. The encoding of the IPv6 addresses is specified in **draft-ietf-mboned-embeddedrp**.

The advantage of using Embedded RP is that no configuration is required. The main disadvantage is that the RP address becomes globally visible, making the RP an easier target for denial-of-service attacks. Furthermore, as Embedded RP is only used for a subset of the IPv6 multicast address space, another RP discovery mechanism is required for all other addresses.

Embedded RP is also incompatible with BIDIR-PIM, as BIDIR-PIM relies on being able to do DF election at RP discovery time, before data starts to arrive. This is not possible with Embedded RP.

4.2.4.5 Anycast RP

Anycast RP (defined in RFC 3446) is not a method of RP discovery, but is a means of bypassing the restriction of having a single RP per multicast group, giving the scalability and fault-tolerance advantages of the more complex RP discovery systems.

Multiple RPs are configured with the same unicast IP address, and this shared address is advertised as the RP address for the group. Sources and receivers use the nearest RP, as determined by the IGP. In order that all the RPs know about all the sources for their groups, MSDP (see Section 5.1, **Multicast Source Discovery Protocol**) must be run between the RPs, though an extension to the PIM-SM protocol to remove the need for MSDP with Anycast RP has been proposed (see **draft-ietf-pim-anycast-rp**).

The advantages of Anycast RP are two-fold.

- *Load balancing / scalability*. The burden of Register decapsulation is distributed across several RPs. This has the advantage over other methods of RP load balancing that the work is split for each group, removing the need to predict traffic patterns across the multicast group address space.

- *Fault-tolerance*. If one RP fails, its sources and receivers are distributed to other RPs by the unicast routing infrastructure.

As Anycast RP solves two of the problems Auto-RP and BSR are designed to avoid, it is usually preferable to use Anycast RP in conjunction with a static RP configuration, rather than use Auto-RP or BSR, as this avoids the complexities of the latter two systems while retaining their benefits.

4.2.5 Mixed-mode PIM Configurations

Typically, a single one of PIM-SM, PIM-DM or BIDIR-PIM would be used throughout a multicast domain. However it is possible to use a combination of the three by distributing multicast groups between the different protocols. Each group must operate in either sparse, dense or bi-directional mode; it is not possible to use a single group in more than one mode at once. Because of this division, the protocols co-exist largely independently of one another.

The one way in which these protocols interact is that the same Hello protocol is used by each, and is only run once on each LAN. The information learned from the Hello message exchange must be shared among the three routing protocols.

The method of distributing groups between the three protocols is outside the scope of the PIM protocols and is a matter of local configuration. Note that it is important that every router in the domain has the same assignment of groups to protocols. The following techniques are used.

- The Bootstrap Router protocol has been extended to add a “Bi-directional” bit for each group range. This method may be used to assign groups between sparse and bi-directional modes if using BSR.

- Routers may be configured to use dense mode if the RP discovery mechanism (whatever that may be) fails to find an available RP for a group, and to use sparse or bi-directional mode otherwise.

- Router may be manually configured with group ranges for sparse, dense and bi-directional modes.

4.3 Distance Vector Multicast Routing Protocol (DVMRP)

Distance Vector Multicast Routing Protocol (DVMRP) is generally regarded as being a suitable protocol for small networks, particularly those running only multicast traffic (and hence lacking in a suitable unicast routing protocol). However, it is not considered to scale suitably for larger networks or more general applications.

The first version of DVMRP was created in 1988 and is documented in RFC 1075. It was (and still is) widely used in the MBONE. There have since been two more versions, the second of which is currently being standardized by the IETF. It is documented in **draft-ietf-idmr-dvmrp-v3**, which is shortly to be published as a Proposed Standard RFC. It is worth noting that due to the limited applicability of DVMRP there have been recent calls to class the draft as "Historic" rather than "Proposed Standard".

DVMRP is a dense multicast routing protocol that uses its own distance vector algorithm to compute routes in the network. It is similar to PIM-DM but differs in the following significant ways.

- It uses its own distance vector algorithm for computing routes through the network, rather than depending on an external MRIB, as discussed in Section 4.1.3, **Determining the Upstream Router**.
- The process of route exchange automatically resolves forwarding conflicts.
- Its routing algorithm has native support for tunnels between DVMRP routers through unicast-only networks. These tunnels typically use IP-in-IP or Generic Routing Encapsulation.
- Control messages are sent as raw IP datagrams with protocol number 2 (the same as IGMP).

The advantages and disadvantages of DVMRP are similar to those of PIM-DM, except that DVMRP is not independent of unicast routing protocols in the same way as PIM.

4.4 Multicast Extensions to OSPF (MOSPF)

Multicast Extensions to OSPF (MOSPF) is an extension of the unicast routing protocol OSPF that turns it into a combined unicast and multicast routing protocol. MOSPF is documented in RFC 1584, which dates from 1994. There has been only one version of the protocol, and it has never been seriously deployed.

MOSPF differs significantly from every other multicast routing protocol in that there are no protocol messages exchanged between MOSPF neighbors for building multicast distribution trees. All the information necessary to compute the trees for every source and group is broadcast to every router in the domain using the same mechanisms as for distributing OSPF link-state information, but including information about multicast memberships.

MOSPF builds multicast routing trees as follows.

- MOSPF uses the same link-state mechanism as OSPF for computing and distributing unicast routes through the domain, except that in MOSPF some links are tagged as being multicast-capable and some are not. This allows incongruent network topologies for unicast and multicast data.

- The same mechanism is also used to distribute local membership information to all routers in the domain. This state is (*,G) only – there is no support for SSM.

- On receipt of an (S,G) multicast data packet, a router uses the route and local membership information it has learned to compute the whole source-based tree for that (S,G). If the router finds that it is on the tree, it forwards the data packet to just those links that require it.

Because every router knows the same information, they all compute the same tree, and so no Assert, Designated Router or Designated Forwarder mechanisms are necessary.

An advantage of MOSPF is that it is not necessary to define RPs. MOSPF may also have a particular appeal in networks that already use OSPF as their unicast routing protocol, although it has the same inherent scalability limitations as OSPF. Another disadvantage of MOSPF is that it does not support SSM.

MOSPF is documented in RFC 1584, which dates from 1994.

5. INTERDOMAIN MULTICAST ROUTING

The multicast routing protocols described in the previous chapter apply to multicast routing within a single domain. This chapter describes various methods of *interdomain* multicast routing.

We first describe the Multicast Source Discovery Protocol (MSDP), which is used to exchange source information between PIM-SM RPs in different domains. In current networks, multicast is mainly deployed with domains running PIM-SM, connected using MSDP. This is not a long-term scalable solution, as the amount of traffic flowing over MSDP scales linearly with the number of sources in the whole Internet. We discuss potential future alternatives to MSDP.

Next, we discuss the rules for the operation of multicast border routers, which are routers that belong to more than one multicast domain. Finally, we describe the Border Gateway Multicast Protocol (BGMP), which is intended to provide true interdomain multicast routing between domains running different multicast routing protocols.

5.1 Multicast Source Discovery Protocol (MSDP)

PIM-SM (described in Section 4.2.1, **PIM Sparse Mode**) enables multicast packets from sources in a domain to reach receivers in the same domain, using an RP inside that domain.

The Multicast Source Discovery Protocol (MSDP) is a companion protocol that allows PIM-SM RPs in different domains to exchange information about the multicast sources in their domains. This information is essentially a list of source/group pairs for the active multicast sources in their domain, which allows other PIM-SM RPs to discover and join to sources in the domain. MSDP is described in RFC 3618 and in **draft-ietf-mboned-msdp-deploy**.

MSDP can also be used to exchange multicast routing information between Anycast RPs in the same domain (see Section 4.2.4.5, **Anycast RP**).

MSDP works as follows.

- The RPs in each domain run MSDP. It is also possible for non-RPs to run MSDP; they can relay MSDP information to other MSDP nodes, but do not originate or act on it.
- Each MSDP node is explicitly configured with a set of MSDP peers. MSDP peers set up TCP connections between themselves.
- When an RP receives a PIM Register message from a new source within its domain, it creates an MSDP Source-Active (SA) message, containing the address of the source, the destination group and the originating MSDP node. It periodically resends this SA MESSAGE for as long as the source is active.

- MSDP messages, containing one or more SA messages, are flooded throughout the MSDP network. To reduce unnecessary forwarding of MSDP information, this flooding mechanism usually relies on the BGP topology and on BGP-distributed information about the route to the originating MSDP node.

- When an RP receives a new SA message, and one or more receivers in its domain want to receive the multicast traffic for that source/group pair, it creates a shortest-path tree to the source by sending a PIM Join message to the source, exactly as if it had received a PIM Register from the source.

MSDP nodes are permitted to filter the SA messages that they originate, and for administratively scoped groups (as described in Section 2.2, **Multicast Address Scoping**), the SA messages that they forward.

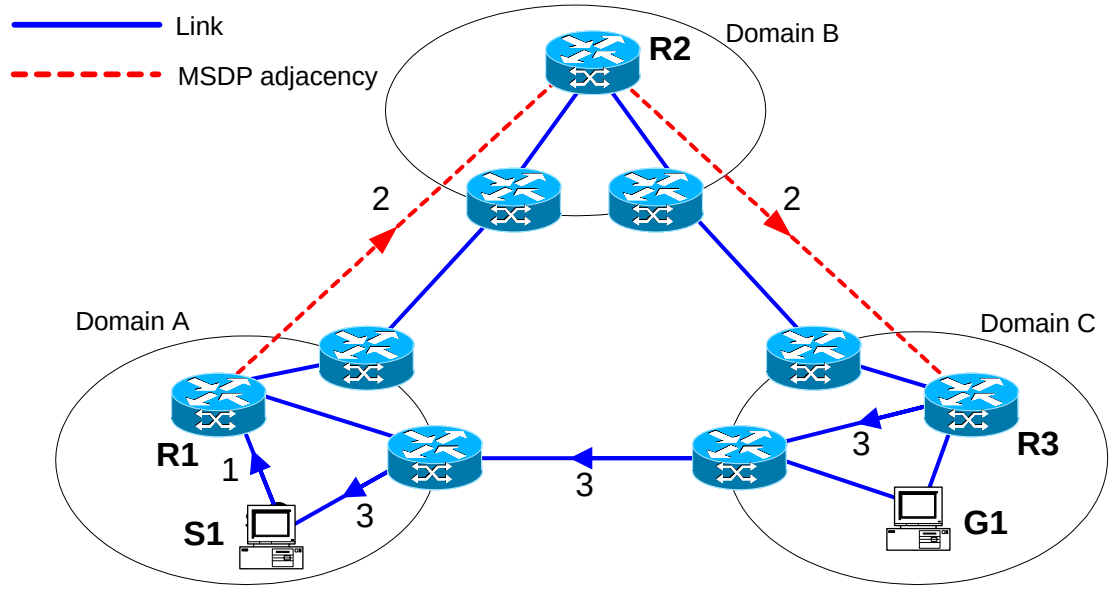
Using MSDP with protocols other than PIM-SM is permitted, but there are currently no specifications that do this.

In the following example, G1 has already sent a PIM Join message to R3, instructing it to join the multicast group.

1. In domain A, source S1 begins to send multicast traffic to R1.

2. R1 creates an MSDP SA message, and floods it throughout the MSDP network, sending it to R2 and R3.

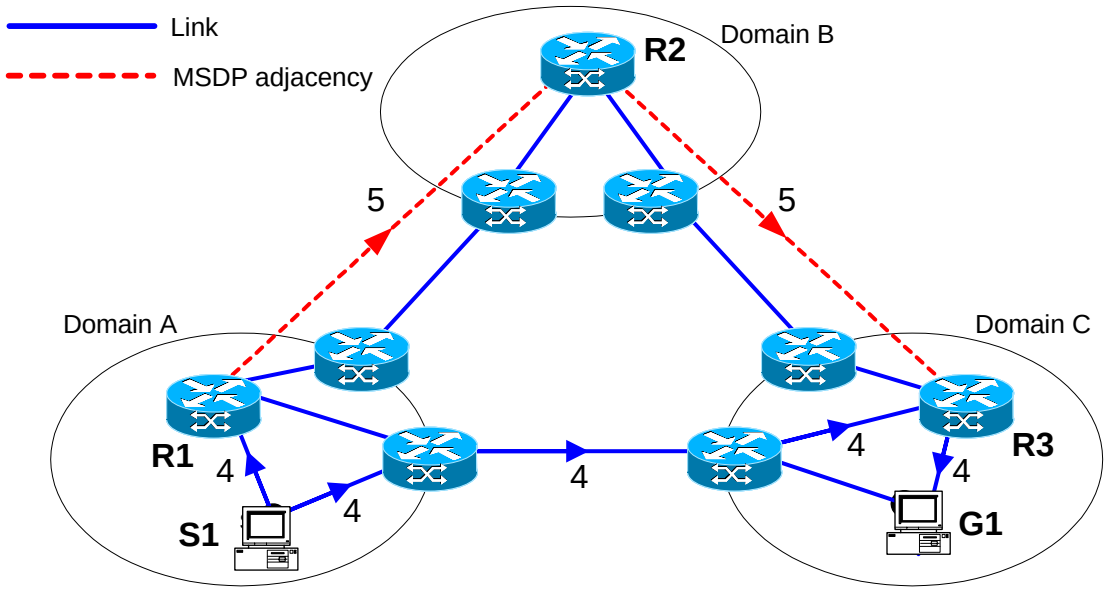
3. R3 knows that it has a receiver G1 that is interested in this multicast group, so it sends a PIM Join to S. R2 does not have any interested receivers, so it does not send a PIM Join message to S.



R3 has now indicated that it wishes to receive multicast data, and R2 has indicated it does not.

4. S1 sends multicast traffic to R1 and R3.[3]

5. R1 periodically continues to flood the MSDP SA MESSAGE throughout the MSDP network.



5.1.1 Alternatives to MSDP

MSDP is problematic as the amount of traffic flowing over MSDP scales linearly with the number of sources in the whole Internet, meaning that it is not a long-term scalable solution. There are two alternatives to the MSDP protocol.

- SSM (described in Section 2.3, **ASM versus SSM**): MSDP is not needed for SSM, as no source discovery is necessary.

- Embedded RP (described in Section 4.2.4.4, **Embedded RP**): MSDP is not needed with Embedded RP, as for an Embedded RP group there is only a single RP in the whole Internet (rather than an RP per domain with normal groups). Embedded RP is only supported for IPv6.

[3] In this example, there are no receivers in Domain A, so R1 would actually have sent a PIM Register-Stop to S1, causing it to stop sending the multicast traffic to R1. S1 would still send periodic PIM Null-Register messages to R1, causing it to continue flooding the MSDP SA message. Meanwhile, PIM-SM could choose to optimize the multicast data flow between S1 and G1 by switching G1 onto the SPT.

In practice, both SSM and Embedded RP are relatively new. There are applications for which SSM is not ideally suited, and disadvantages to Embedded RP, notably that it is only supported for IPv6.

Interdomain PIM-SM will require MSDP for the foreseeable future, but SSM and Embedded RP are likely to become more important with time.

5.2 Multicast Border Routers

RFC 2715 specifies the behavior required at multicast border routers to allow domains running different multicast routing protocols to interoperate. In order for these rules to apply, there must either be a domain hierarchy (a tree with one domain as root), or an interdomain routing protocol (such as BGMP, which is described in Section 5.3, **Border Gateway Multicast Protocol**) must be used.

The important features of the rules for multicast border routers may be summarized as follows.

- The router consists of two or more multicast routing components, each owning a subset of the router's interfaces over which it runs some multicast routing protocol.
- No interface runs more than one multicast routing protocol. Hence each multicast-enabled interface is owned by exactly one multicast routing component.
- The components communicate to tell each other which multicast data is required in each domain.
- All the components share a common forwarding cache.
- The component that owns an interface chooses whether to accept or reject packets received on that interface.
- Once an incoming packet has been accepted, the component that owns an interface chooses whether or not the packet should be forwarded out of that interface.

5.3 Border Gateway Multicast Protocol (BGMP)

The goal of the Border Gateway Multicast Protocol (BGMP) is to allow full interdomain multicast routing in a network with domains running different multicast routing protocols. BGMP is not currently deployable, and is not likely to be in the near future. Eventually, however, BGMP or a similar protocol may be the answer for interdomain multicast.

BGMP supports both Source-Specific Multicast (SSM) and Any Source Multicast (ASM) models, and allows receivers to build bi-directional shared trees that span domains. Within each domain, the multicast routing protocol routes the multicast traffic. For a given tree of domains, a single domain is designated (by configuration) the 'root domain' for the tree. This is analogous to an RP being the root of a tree of routers.

The ASM model requires that each multicast group is associated with a single root domain. This association is encoded in IPv6 multicast group addresses using Unicast-Prefix-based addressing. Alternatively, another mechanism, such as the Multicast Address-Set Claim (MASC) protocol, can be used to distribute the association.

BGMP is still evolving, but its key features are as follows.

- BGMP runs on the border routers between multicast-capable domains.
- BGMP creates TCP connections between peers.
- BGMP routers send Join and Prune messages to peers in other domains.
- A BGMP router uses these Join and Prune messages to build up state for forwarding packets between domains.
- If a BGMP router does not have forwarding state for a multicast data packet, it forwards it towards the root domain.

BGMP is described in **draft-ietf-bgmp-spec**, which is soon to be published as an Informational RFC. The status of “Informational” was chosen to underline the fact that the protocol is not currently deployable, and that its future as an IETF standard is uncertain.

6. MULTICAST SIGNALING

Signaling protocols are used to enhance the efficiency of unicast IP packet forwarding, and to provide traffic engineering and guaranteed quality of service.

Multi Protocol Label Switching (MPLS) is a collection of such signaling protocols. The protocols specify point-to-point Label Switched Paths (P2P LSPs) which traverse routers along a given data path. IP traffic is then tunneled through these LSPs, providing two key advantages.

- When a data packet enters the LSP, it is assigned a label. Transit routers along that LSP then forward the data based on a quick look-up on that label. This is more efficient than, say, an IP address best-match.

- When the RSVP-TE protocol is used to set up the LSP, the path of the LSP is chosen to provide a specific quality of service. Each router along that path can set aside resources for use solely by that LSP.

LSP tunneling is also an efficient and secure way to implement standards-based Virtual Private Networks (VPNs), because the forwarding of data packets is based solely on the MPLS label, so inspection of the packet contents (including the destination address) is not required. More information on these issues is given in Data Connection's White Paper, *VPN Technologies – A Comparison*, which is available at http://www.dataconnection.com/products/whitepapers.htm.

The principles described above can also be applied to multicast IP packet forwarding situations by creating point-to-multipoint (P2MP) LSPs. There are a number of protocol drafts that target this area. This section gives an overview of these approaches, which include

- RSVP-TE P2MP LSPs

- PIM P2MP LSPs

- tunneling IP multicast traffic through an MPLS network

- tunneling VPN IP multicast traffic through a provider network.

6.1 RSVP-TE P2MP LSPs

RSVP-TE is the only standardized MPLS signaling protocol which provides traffic engineering, and hence is the obvious choice to adapt to multicast signaling for applications (such as streaming video broadcast) which have strict requirements on reliability. The requirements for RSVP-TE extensions for P2MP are set out in **draft-ietf-mpls-p2mp-requirement**.

All of the suggested solutions require the ingress node to calculate the entire P2MP LSP, so that it must somehow determine the set of receivers. This is in contrast to plain IP multicast, where the source/RP knows only the next hop. This requirement affects the routing protocol. See Chapter 4, **Multicast Routing Protocols**, for more details.

However, this requirement allows far more flexibility in choosing the tree. Rather than simply using a shortest-path tree, the path could be optimized to, for example, have the least overall cost (a Steiner tree).

6.1.1 LSP Association and Secondary Explicit Route Object (SERO) Approach

draft-yasukawa-mpls-rsvp-p2mp defines a method of grouping together a set of P2P LSPs to make a P2MP LSP such that there is no duplication of resources either in the control plane or the data plane.

This is achieved by the ingress (source) node first discovering the set of egress (receiver) nodes and calculating the tree route for the P2MP. It then sends a single Path message with a single Explicit Route Object (ERO) corresponding to one of the egresses, and multiple Secondary Explicit Route Objects (SEROs) corresponding to the other egresses. The first hop in each SERO is a branch node. Each branch node takes its corresponding SERO, and sets up a new P2P LSP (in a different session) with an ERO taken from that SERO. The ASSOCIATION object is used to associate all these P2P LSPs together as a P2MP LSP.

- The key advantage of this approach is that it does not duplicate state in the control plane. The P2P LSPs making up the P2MP LSP do not overlap.

- The main disadvantage is that it does not support all of the normal RSVP-TE extensions (e.g. Fast Reroute and other protection/restoration mechanisms), although it does support make-before-break of the entire tree.

- A complete set of SEROs for a P2MP LSP can also be quite large, making the RSVP Path message harder to handle.

One of the approaches defined by **draft-choi-mpls-grouplabel-requirement** is similar to this approach.

6.1.2 P2P LSP Merging Approach

draft-raggarwa-mpls-p2mp-te also defines a method of grouping together P2P LSPs, but it does this such that each P2P LSP extends right from the ingress of the P2MP LSP, rather than from a branch node. Upstream of a branch node, there will be multiple P2P LSPs sharing a common path.

However, the set of P2P LSPs is grouped using a modified RSVP Session Object, which has a P2MP LSP ID field. The P2MP LSP ID allows nodes on a common path to assign the same upstream label to each P2P LSP that is part of the same P2MP LSP.

While there are multiple overlapping P2P LSPs in the control plane, there is a single P2MP LSP in the data plane.

- The key advantage of this approach is that because each P2P LSPs extends the full length of the P2MP LSP in the data plane, existing fault protection/recovery mechanisms such as Fast Reroute, error reporting, etc. all just work.

- The disadvantage is that nodes on a common path have to hold duplicate control plane state, wasting resources. It is suggested that this can be reduced by using hierarchical LSPs to tunnel otherwise-duplicated state directly to branch nodes, but this makes tree modifications potentially more complex.

6.1.3 RSVP Broadcast Approach

The second of the two approaches defined by **draft-choi-mpls-grouplabel-requirement** uses an RSVP broadcast-type mechanism. To date, this approach has generated less interest in the MPLS community.

The ingress node sends a Path message to all other RSVP nodes in the network, and any node that wants to receive multicast traffic sends a Resv message back. This draft defines RSVP Join and Prune messages to allow egress nodes to join or leave a P2MP LSP after its initial creation.

This approach allows the ingress node to determine the set of receivers without having to rely on an external mechanism. It is very different to the way that RSVP currently works.

6.2 PIM P2MP LSPs

draft-farinacci-mpls-multicast defines extensions to PIM to allow MPLS labels to be distributed in a downstream unsolicited fashion along with the multicast routes.

Key attributes of these extensions are as follows.

- The P2MP LSP is restricted to following the same tree as the IP multicast tree.

- LSPs are set up even if there are no senders.

- The ingress node does not know the set of receivers, only the next-hop(s).

6.3 Tunneling IP Multicast Traffic Through an MPLS Network

There are times when it is undesirable to run a multicast routing protocol on every node within a network, for example, it may place too much demand on the stored state.
draft-raggarwa-PIM-SM-mpls-te describes a mechanism to tunnel PIM-routed IP multicast traffic through an MPLS network.

Key attributes of this mechanism are as follows.

- Within the MPLS network, PIM and BGP run on the Provider Edge routers only.
- The PIM routers form adjacencies to each other using the Directed Hello extensions to PIM specified by **draft-raggarwa-PIM-SM-remote-nbr**. These extensions allow PIM routers to unicast PIM messages to specific non-adjacent (in IP terms) PIM neighbors.
- Downstream Provider Edge (PE) routers send PIM Join/Prune messages to the upstream PE router (the BGP next-hop towards the source/RP).
- The upstream PE router is responsible for creation/modification/destruction of the P2MP LSP to the downstream PE routers.
- The upstream PE router sends a newly-defined PIM Join Acknowledge message to the downstream PEs, containing the ID of the P2MP LSP that it will send the multicast traffic down.

The upstream PE router could create a P2MP LSP for each source/group pair, but to improve scalability, it is permitted to multiplex multicast traffic down the same P2MP LSP, even if this results in downstream PE routers receiving multicast traffic when they have no interested receivers.

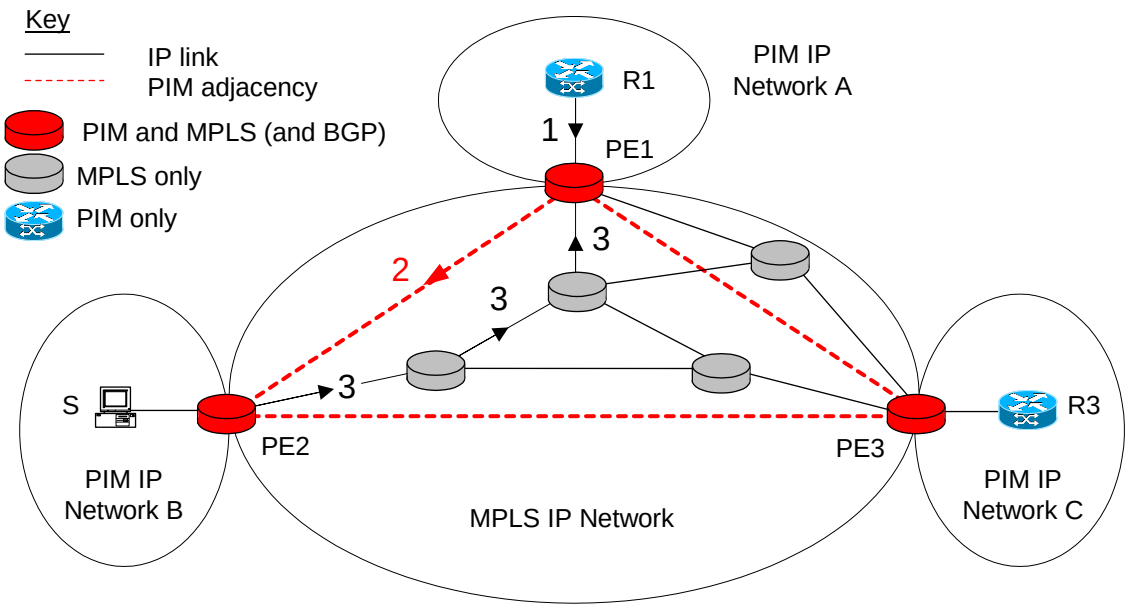
For simplicity, the following example involves SSM in PIM-SM networks, but the same PIM/MPLS interactions occur when using ASM (whether the networks represent separate domains or subdivisions of the same PIM domain).

The following diagram shows routers R1 and R3 joining the multicast network with a source at B. In this and in future diagrams, PE indicates a PE router.

1. R1 sends a PIM Join message to PE1.

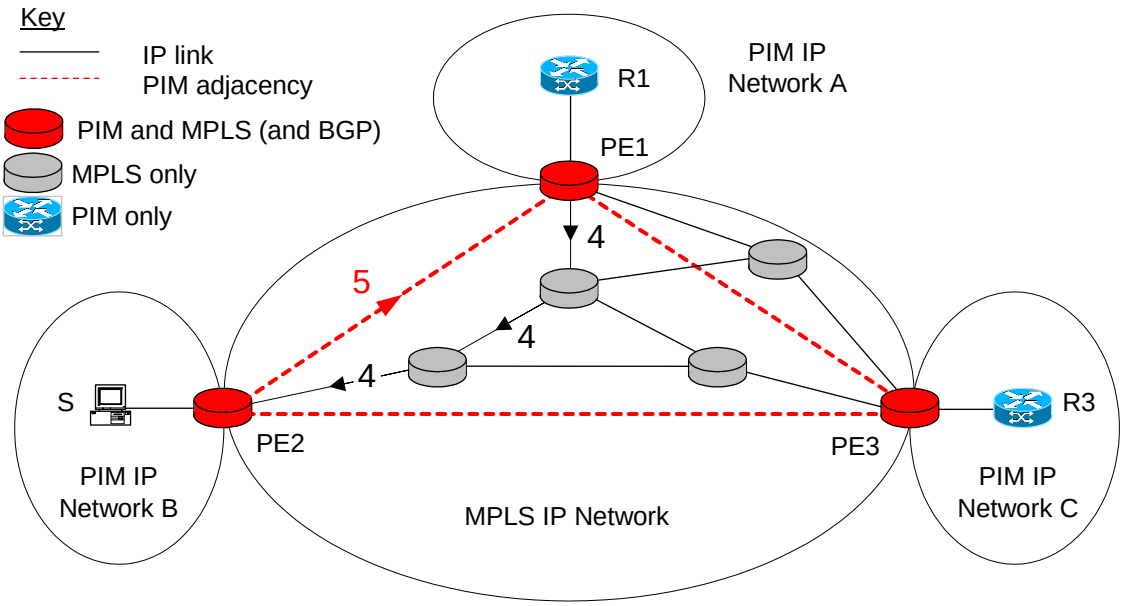
2. PE1 creates a remote PIM adjacency to PE2 (unless such an adjacency already exists), and unicasts a PIM Join message to it.

3. PE2 starts to set up a P2MP LSP, with PE1 as the only egress node. Note that this can follow a different path to the PIM messages between PE2 and PE1.

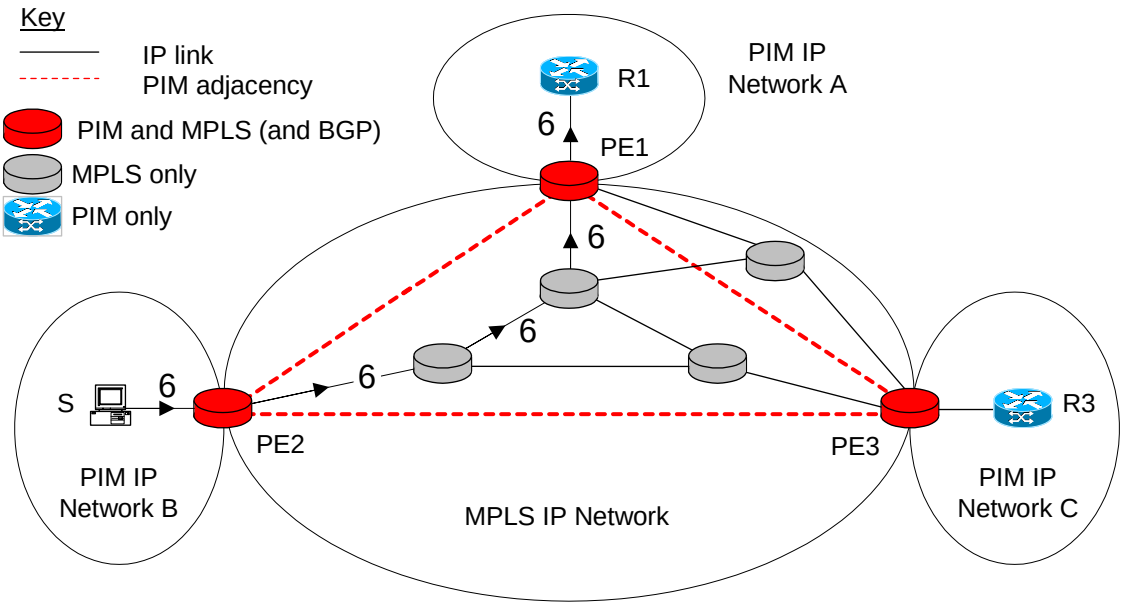


4. MPLS completes setting up the P2MP LSP.

5. PE2 unicasts a PIM Join Acknowledge message to PE1, including the P2MP LSP ID.



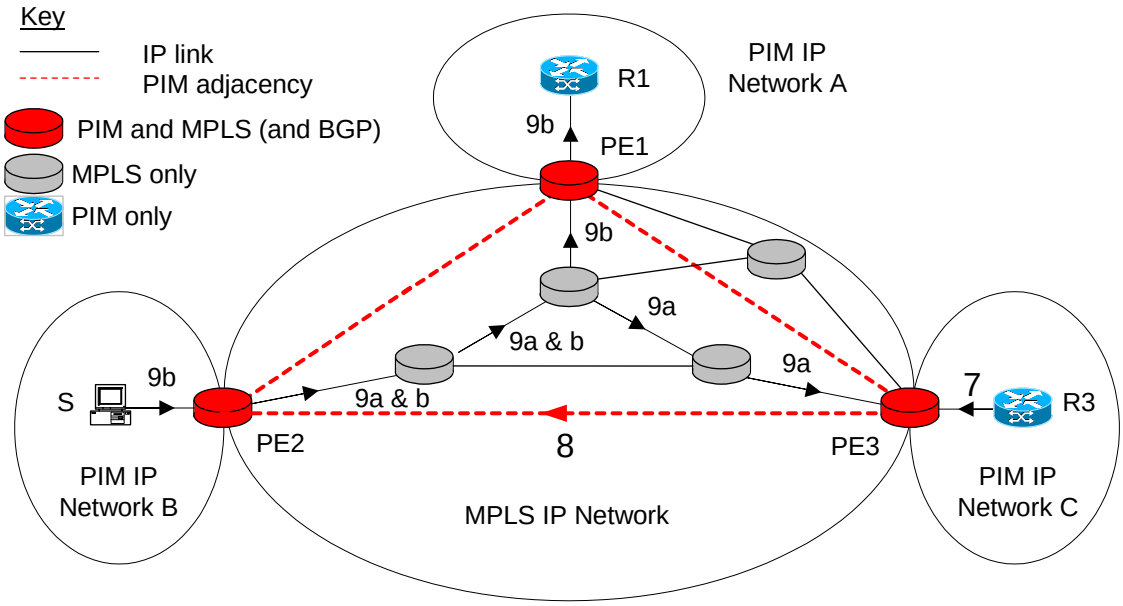
6. S begins to send. The traffic flows from S to PE2 as normal, then down the MPLS tunnel to PE1, then as normal to R1.



7. R3 sends a PIM Join message to PE3.

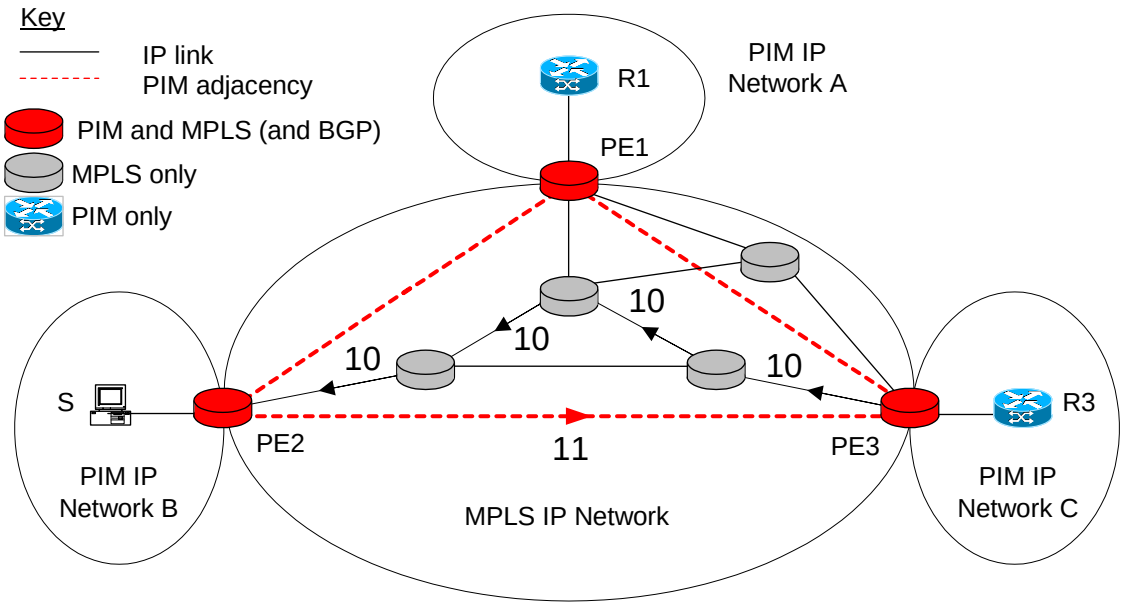
8. PE3 creates a remote PIM adjacency to PE2 (unless such an adjacency already exists), and unicasts a PIM Join message to it.

9. (a) PE2 starts to modify the P2MP LSP to include PE3 as another egress node.
   (b) Multicast traffic flowing via the LSP to PE1 is unaffected.

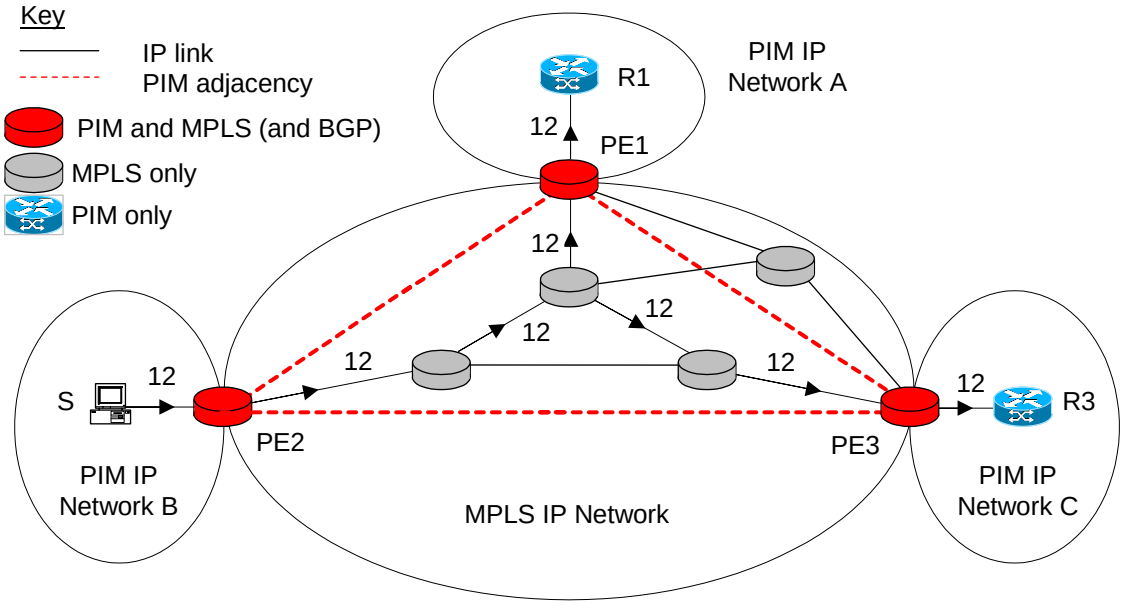


10. MPLS completes modifying the P2MP LSP.

11. PE2 unicasts a PIM Join Acknowledge message to PE3, including the P2MP LSP ID.



12. Multicast traffic now flows to both R1 and R3, and is duplicated at the branch node within the MPLS network.



6.4 Tunneling VPN IP Multicast Traffic Through a Provider Network

draft-rosen-vpn-mcast extends RFC 2547 Virtual Private Network (VPN) functionality to support multicast data transport.

The key features of this approach are as follows.

- A *multicast domain* is a set of VPN Routing and Forwarding Tables (VRFs) associated with interfaces that can send multicast traffic to each other. A VRF can belong to more than one multicast domain.

- Each multicast domain is associated with an otherwise unused multicast group address from the address space of the service provider network.

- A service provider network instance of PIM-SM runs on each PE router. These PE routers join the service provider multicast groups corresponding to the multicast domains to which they are connected.

  For scalability reasons, this simple mechanism does not consider whether there are actually any receivers in a particular PE router's part of a multicast domain, and which customer-space multicast groups that they are members of – if a PE router is connected to a multicast domain, it will receive all multicast traffic for that multicast domain.

- Within the service provider network, there is now a multicast tree for each multicast domain (it is recommended that this be a bi-directional tree, for scalability). This is referred to as the *multicast tunnel* for the domain.

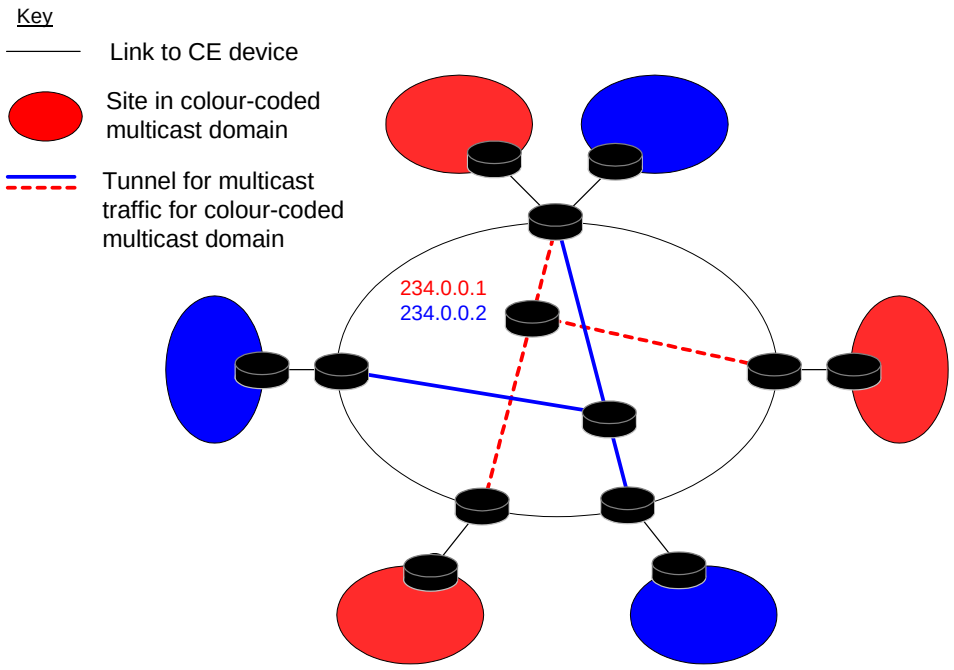
- Customer multicast data can now be transmitted across the service provider network, but it must be encapsulated within a packet whose outer destination address is the per-domain multicast group address (since the inner packet will have a customer site specific multicast group address). This encapsulation could use MPLS tunnels or Generic Routing Encapsulation (GRE).

- Each PE router runs a per-VRF instance of PIM-SM that communicates with the CE device. This PIM-SM instance treats the multicast tunnel as a multidrop interface that reaches all of the other PE routers that are associated with this domain.

  If a single VRF is in several multicast domains, then PIM-SM requires additional information to make the choice of which multicast tunnel to use for which customer-space multicast groups. This information is provided by an out-of-band mechanism.

- Note that from the point of view of the CE devices, there is no change to PIM-SM or packet forwarding behavior.

The following diagram illustrates non-overlapping trees connecting multicast VPN sites.



7. MULTICAST DATA PLANE OPERATION

This chapter describes how the data plane forwards multicast data. It also discusses the information that the data plane must make available to the control plane. The sum total of these interactions between the data plane and the control plane is vital to the overall architecture of a multicast router.

7.1 Host Functionality

7.1.1 Sending Data

From the point of view of an application, sending multicast data is easy. The application just specifies a multicast IP address as the destination IP address for the packet. For Ethernet, the sockets stack then sets the destination MAC address to a multicast MAC address that is derived from the IP multicast address.

7.1.2 Receiving Data

A multicast datagram arrives on an interface when a multicast router sends it to that interface, with (for Ethernet) a multicast MAC address derived from the IP multicast address. A host that is interested in that multicast group will thus receive the packet.

In order to know whether it is interested in a particular multicast group, the host's sockets stack will have collated the requirements of all of the applications on the host. When a multicast packet arrives, the sockets stack will distribute copies of the packet to all, some, or none of the applications, as follows.

- Prior to IGMPv3, applications could only register interest in the whole multicast group. Datagrams would typically be delivered to all applications that had opened a multicast socket for that group on that interface.

- With IGMPv3 source filtering, a packet from a particular source may only be delivered to those applications that have requested that the source be included (or not excluded).

7.2 Router Functionality

7.2.1 Forwarding Data

On receipt of a multicast datagram, a multicast router must first determine whether it should forward the packet at all. In general, a router only expects to receive multicast datagrams over the interface that leads upstream towards the source of the data or the root of the multicast tree. A datagram that arrives over a different interface is normally dropped as it is typically

- a packet on a multidrop interface where another router has been designated responsible for forwarding the packet onwards
- a late-arriving packet left over from before the routing topology changed
- a packet arriving via an old route during the process of changing the distribution tree of a particular multicast group (for example, when PIM-SM shifts from a shared tree to a source-based tree for a particular group).

Note that this is not the case for bi-directional shared trees; packets arriving on any branch of such a tree are forwarded out over all other branches.

If a datagram is accepted for forwarding, the data plane forwards it out over a list of interfaces, making copies of the packet as necessary.

7.2.2 Unicast Encapsulation and Decapsulation

In some shared tree routing protocols, multicast data is initially sent over unicast between the router adjacent to the source node and the router that forms the root of the shared multicast forwarding tree (as described in Section 4.1.4, **Data Plane Interactions**). In terms of data flow this means that

- the router adjacent to the source node receives the multicast datagram, encapsulates it and sends it to the root using normal IP unicast
- the router at the root of the shared tree decapsulates the incoming unicast packet and forwards it as a multicast packet.

Therefore, the data plane must either be able to perform this encapsulation and decapsulation itself, or be able to send packets for encapsulation to, and receive decapsulated packets from, the control plane.

7.2.3 Packet Arrival Information

As described in Section 4.1.4, **Data Plane Interactions**, there are several situations in which the arrival of a multicast packet may trigger communication between the data plane and the control plane.

- In some protocols, such as PIM-DM or MOSPF, it is difficult or impossible to set up the correct forwarding state for every source and group in advance. Therefore, the data plane must inform the control plane when a packet from a new source or for a new group arrives so that the control plane can set up the state correctly.

- In protocols such as PIM-SM, when a directly connected source starts sending data to a multicast group, the data plane must inform the control plane so that it can start encapsulating the data and sending it to the root of the shared tree.

- The control plane typically has soft state associated to a particular multicast group and source for that group. This state is kept alive for as long as multicast data is being received from that source for that group. Consequently, the data plane must inform the control plane of timing information about the arrival of multicast packets.

- In certain network configurations, it is possible that multiple upstream routers may forward multicast traffic from the same source onto the same LAN segment. Multicast protocols such as PIM-SM and PIM-DM do not attempt to prevent this occurring. Rather, they rely on the data plane to detect the arrival of multicast packets from different upstream routers. This triggers the control plane to elect a single forwarder for these packets.

- In protocols such as PIM-SM, it is possible to initiate a switch from a shared tree to a source-specific tree. To enable the control plane to decide when to do this, the data plane needs to inform the control plane of the bandwidth consumed by data being forwarded from individual sources over the shared tree.

- In protocols such as PIM-SM, when the control plane has initiated a switch to a source-specific tree, it needs to know when the source-specific tree has been established so that the source can be pruned from the shared tree. In general, multicast routing protocols do not have any control plane mechanism to signal that a tree towards a source has been established. Therefore, the data plane must inform the control plane when data first arrives over the source-specific tree.

8. SUMMARY

IP multicast is an important and growing area, involving a wide and sometimes competing range of network protocols. There are more and more areas in which people need to distribute information to a subset of subscribers, for example, e-learning, webinars, news feeds, stock market feeds, and radio and television broadcasts.

Hosts use the IGMP and MLD protocols to communicate their multicast group membership requirements to adjacent routers. Different versions of these group membership protocols provide different functionality; significantly, IGMPv3 and MLDv2 support Source-Specific Multicast. MLD is required if using IPv6.

A range of protocols is used between routers within a domain to build spanning trees of multicast group members, enabling efficient delivery of multicast traffic. These multicast routing protocols have different properties, making them suitable for use in differing environments.

PIM-SM is at present the most widely used multicast routing protocol. BIDIR-PIM is not widely used at present, but is becoming increasingly important. PIM-DM may also be used in small networks. DVMRP was popular in the past, but is now mostly required only to interoperate with existing DVMRP deployments.

MSDP is the protocol most widely used today for interdomain multicast routing, between PIM-SM domains. In the future, a true interdomain multicast routing protocol such as BGMP may provide a more scalable solution.

There are various proposals in circulation for extending MPLS to perform multicast packet forwarding, using point-to-multipoint LSPs. None of these have yet been standardized, but as the use of both MPLS and multicast grows, this is likely to become an important area for future network designers.

Although multicast is not currently widely deployed within the Internet, it has been slowly gaining influence over the last few years. The emerging next generation technologies described in this document, together with an ever-widening area of applications and business cases, guarantee that multicast will soon be a highly sought-after area of technology. Hence, multicast is a strategic area for investment for many vendors and service providers.

9. ABOUT DATA CONNECTION

Data Connection Limited is the leading independent developer and supplier of (G)MPLS, OSPF(-TE), ISIS(-TE), BGP, RIP, VPN, PIM, IGMP, MLD, ATM, MGCP, Megaco, SCTP, SIP, VoIP Conferencing, Messaging, Directory and SNA portable products. Customers include Alcatel, Cabletron, Cisco, Fujitsu, Hewlett-Packard, Hitachi, IBM Corp., Microsoft, Nortel, SGI and Sun.

Data Connection is headquartered in London UK, with US offices in Reston, VA and Alameda, CA. It was founded in 1981 and is privately held. During each of the past 21 years its profits have exceeded 20% of revenue. Last year, sales exceeded $36 million, of which 95% were outside the UK, mostly in the US. Even through the current severe downturn, Data Connection's financial position remains secure, as does its employee base: our 200 software engineers have an average length of service of 8 years, with turnover of <3% annually.

Our routing protocols are designed from the ground up to address next generation networking issues such as massive Internet scalability, optical routing at multiple layers, virtual routing, MPLS and TE/CSPF, and VPNs.

Data Connection's Multicast IP Routing products, DC-PIM and DC-IGMP, are designed to support the functionality required by existing multicast devices, as well as to be extensible as new features and protocols are developed.

Our products share a common architecture to facilitate integration of the family members. The architecture allows components to be combined in a variety of ways to create routers with specific functionality, and allows seamless integration with Data Connection's suite of unicast IP Routing products. These IP Routing products include implementations of ISIS, OSPF, RIP and BGP, which can be used to provide the multicast routing protocols (e.g. PIM) with the routing data to build multicast distribution trees.

All of the Data Connection protocol implementations are built with scalability, distribution across multiple processors and fault tolerance architected in from the beginning. We have developed extremely consistent development processes that result in on-time delivery of highly robust and efficient software. This is backed up by an exceptionally responsive and expert support service, staffed by engineers with direct experience in developing the protocol solutions.

Jon Hardwick is the senior architect and project manager for Data Connection's multicast routing implementations. He plays a leading role in product architecture and standards-based development in Data Connection's Networking Protocols Group. He has six years' experience in the field of communications protocols, having worked on IP Multicast, BGP, MPLS, ATM, SIP, Megaco, MGCP, SNA and APPN.

10. GLOSSARY

This section provides a brief review of some of the terminology used in the document.

Anycast RP	Anycast RP is a mechanism of having more than one RP for a PIM-SM multicast group by configuring multiple RP routers to have the same IP.
ASM	Any Source Multicast (ASM) is a model of multicast data transmission where receivers request all data sent to a multicast group, regardless of its source address.
Auto-RP	Auto-RP is a mechanism that was used by Cisco routers in PIM-SM v1 to determine the mapping from multicast groups to Rendezvous Points. A central location would multicast out group-to-RP mappings; the chicken-and-egg problem was avoided by using PIM-DM multicast to do this.
BGMP	Border Gateway Multicast Protocol (BGMP) is a multicast routing protocol that builds shared multicast distribution trees, but where these trees are made up of entire domains rather than individual routers.
BIDIR-PIM	Bi-directional PIM (BIDIR-PIM) is a shared tree multicast routing protocol, based on PIM-SM but using bi-directional trees.
BSR	A Bootstrap Router (BSR) is the Rendezvous Point for a PIM-SM or BIDIR-PIM multicast tree, as (automatically) chosen by the BSR election procedure.
CBT	The Core Based Tree (CBT) routing protocol is a shared tree multicast routing protocol.
DF	The Designated Forwarder (DF) in BIDIR-PIM is the router on a LAN that acts as the upstream neighbor for a particular RP. One DF is elected per LAN, per RP.
DR	The Designated Router (DR) in PIM-SM is the router responsible for Register-encapsulating data from directly attached sources, and for forwarding on behalf of directly connected receivers. The Designated Router (DR) in CBT is like the DF in BIDIR-PIM, except that there is a single one per LAN.
DVMRP	Distance Vector Multicast Routing Protocol (DVMRP) is a source-based tree multicast routing protocol which uses distance vector mechanisms to determine the next upstream router on the way to the source. DVMRP is used as the multicast routing protocol in the MBONE.

IGMP	The Internet Group Membership Protocol (IGMP) is the protocol used between hosts and routers on an interface to indicate membership of IPv4 multicast groups.
MBGP	Multiprotocol Extensions for BGP (MBGP) is an extension to the BGP unicast routing protocol that allows different types of addresses (known as *address families*) to be distributed in parallel—for example, IPv4 addresses, IPv6 addresses or RFC 2547 VPN-IPv4 addresses. This allows information about the topology of multicast-capable routers to be exchanged separately from the topology of normal IPv4 unicast routers.
MBONE	The Multicast Backbone (MBONE) is an experimental network for multicast data that is layered on top of the Internet, allowing multicast data to be sent by unicast.
MLD	The Multicast Listener Discovery (MLD) protocol is the protocol used between hosts and routers on an interface to indicate membership of IPv6 multicast groups.
MOSPF	Multicast Extensions to OSPF (MOSPF) is a multicast variant of the Open Shortest Path First (OSPF) routing protocol. This protocol uses source-based multicast trees, and distributes link state information to allow routers to determine the next upstream router on the way to the source.
MRIB	The Multicast Routing Information Base (MRIB) is a unicast forwarding table, used by protocols such as PIM-SM to determine the upstream router for a particular multicast group or multicast source/group combination.
MSDP	The Multicast Source Discovery Protocol (MSDP) describes a mechanism to connect multiple PIM-SM domains together (each with its own domain-specific Rendezvous Point).
PIM	Protocol Independent Multicast (PIM) is a set of protocols used between multicast routers to distribute information about multicast group membership. The protocol is “protocol independent” because it relies on underlying unicast forwarding table information to find the root of the multicast distribution tree, regardless of which unicast routing protocol has been used to generated the information.
PIM-DM	PIM Dense Mode (PIM-DM) is a source-based tree multicast routing protocol appropriate for use when the set of members of any particular multicast group are densely distributed through the network.

PIM-SM	PIM Sparse Mode (PIM-SM) is a shared tree multicast routing protocol appropriate for use when the set of members of any particular multicast group are sparsely distributed across the network.
PIM Sparse-Dense Mode	Mode of operation for a router capable of both PIM-SM and PIM-DM operation, where PIM-SM is used for some multicast groups and PIM-DM is used for others.
PIM-SSM	PIM Source-Specific Multicast (PIM-SSM) is a subset of the PIM-SM protocol that is intended for use in pure source-specific multicast scenarios.
RP	A Rendezvous Point (RP) is the router that forms the root of a shared multicast distribution tree in PIM-SM.
SSM	Source-Specific Multicast (SSM) is a model of multicast data transmission where the channel is identified by the <source address, multicast group address> pair rather than just the multicast group address.
TIB	The Tree Information Base is the collection of state at a multicast router that describes the multicast distribution trees for all of the multicast groups.

11. REFERENCES

The following documents provide more information on the topics covered by this white paper. All RFCs and current Internet-Drafts may be downloaded from the IETF web site at http://www.ietf.org/.

Note that all Internet-Drafts are work in progress and may be subject to change or may be withdrawn without notice.

11.1 Multicast Addressing

RFC 2373	IP Version 6 Addressing Architecture
RFC 3180	GLOP Addressing in 233/8
RFC 3306	Unicast-Prefix-based IPv6 Multicast Addresses
RFC 2974	Session Announcement Protocol (SAP)
RFC 2327	Session Description Protocol (SDP)
RFC 2909	The Multicast Address-Set Claim (MASC) Protocol
RFC 2730	Multicast Address Dynamic Client Allocation Protocol (MADCAP)
RFC 2365	Administratively Scoped IP Multicast
RFC 3569	An Overview of Source-Specific Multicast (SSM)
draft-ietf-ssm-arch	Source-Specific Multicast for IP

11.2 Multicast Group Membership Discovery Protocols

RFC 1054	Host Extensions For IP Multicasting (IGMPv1)
RFC 2236	Internet Group Management Protocol, Version 2
RFC 3376	Internet Group Management Protocol, Version 3
RFC 2710	Multicast Listener Discovery (MLD) for IPv6
draft-vida-mld-v2	Multicast Listener Discovery version 2 (MLDv2) for IPv6
draft-ietf-magma-igmp-proxy	IGMP/MLD-based Multicast Forwarding (“IGMP/MLD Proxying”)
draft-ietf-magma-snoop	Considerations for IGMP and MLD Snooping Switches
RFC 2022	Support for Multicast over UNI 3.0/3.1 based ATM Networks

11.3 Multicast Routing Protocols

draft-ietf-PIM-SM-v2-new	Protocol Independent Multicast – Sparse Mode (PIM-SM) Protocol Specification (Revised)
draft-ietf-PIM-DM-new-v2	Protocol Independent Multicast – Dense Mode (PIM-DM) Protocol Specification (Revised)
draft-ietf-pim-bidir	Bi-directional Protocol Independent Multicast
draft-ietf-PIM-SM-bsr	Bootstrap Router (BSR) Mechanism for PIM Sparse Mode
RFC 3446	Anycast-RP mechanism using PIM and MSDP
draft-ietf-mboned-embeddedrp	Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address
draft-ietf-pim-anycast-rp	Anycast-RP using PIM
RFC 1075	Distance Vector Multicast Routing Protocol (DVMRP)
draft-ietf-idmr-dvmrp-v3	Distance Vector Multicast Routing Protocol (DVMRP)
RFC 1584	Multicast Extensions to OSPF (MOSPF)
RFC 2189	Core Based Trees (CBT version 2) Multicast Routing

11.4 Interdomain Multicast Routing

RFC 3618	Multicast Source Discovery Protocol (MSDP)
draft-ietf-mboned-msdp-deploy	MSDF Deployment Scenarios
RFC 2715	Interoperability Rules for Multicast Routing Protocols
draft-ietf-bgmp-spec	Border Gateway Multicast Protocol (BGMP)

11.5 Multicast Signaling

draft-ietf-mpls-p2mp-requirement	Requirements for Point to Multipoint Extension to RSVP-TE
draft-yasukawa-mpls-rsvp-multicast	Extended RSVP-TE for Multicast LSP Tunnels
RFC 3353	Overview of IP Multicast in an MPLS Environment
draft-choi-mpls-grouplabel-requirement	Requirements for multicast service using a group label over MPLS
draft-farinacci-mpls-multicast	Using PIM to Distribute MPLS Labels for Multicast Routes
draft-raggarwa-mpls-p2mp-te	Establishing Point to Multipoint MPLS TE LSPs

draft-raggarwa-PIM-SM-mpls-te	IP Multicast With PIM-SM Over a MPLS Traffic Engineered Core
draft-raggarwa-PIM-SM-remote-nbr	PIM-SM Extensions for Supporting Remote Neighbors
draft-rosen-vpn-mcast	Multicast in MPLS/BGP VPNs
RFC 2547	BGP/MPLS VPNs