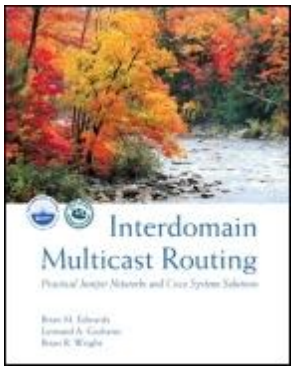


- Table of Contents
- Index

Interdomain Multicast Routing: Practical Juniper Networks and Cisco Systems Solutions

By Brian M. Edwards, Leonard A. Giuliano, Brian R. Wright

Publisher : Addison Wesley
Pub Date : April 24, 2002
ISBN : 0-201-74612-3
Pages : 384

Increasing numbers of ISPs have begun implementing multicast infrastructure. Soon the Internet will provide multicast connectivity between any two points on the Internet the way it provides for unicast traffic today. Long-evolving protocols are reaching maturity, and enterprise networks and ISPs around the world are ramping up their multicast infrastructure. Now, more than ever, network engineers must be ready to deal with new applications that capitalize on the simultaneous, efficient delivery of data and imagery to multiple recipients.

Interdomain Multicast Routing is the key to unlocking the complexities of this growing technology. Starting with a summary of the technology and its relevant protocols, this book shows readers the big picture before revealing a detailed analysis of important protocols and the way they work with one another. Throughout, the authors focus on both Cisco Systems and Juniper Networks technology--the two leading vendors of routers and routing technology. Real-life examples are used to clearly illustrate key concepts. Specific topics covered in Interdomain Multicast Routing include:

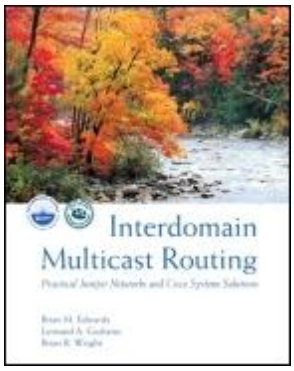
- Background and in-depth analyses of multicast routing using PIM-SM and MSDP
- Comparison of Any-Source and Source-Specific multicast delivery models
- Explanation of how MBGP and M-ISIS can be used side by side to build a dedicated multicast environment
- A detailed breakdown of the differences between IGMP versions 1, 2, and 3
- A step-by-step guide to understanding the MSDP RPF-peer selection rules
-

Lists of packet formats for IGMP, PIM, and MSDP

- A complete glossary that clarifies important terms and acronyms and provides their definitions

Practical and thorough in coverage, Interdomain Multicast Routing is an important addition to any network engineer's bookshelf.

NEXT



- Table of Contents
- Index

Interdomain Multicast Routing: Practical Juniper Networks and Cisco Systems Solutions

By Brian M. Edwards, Leonard A. Giuliano, Brian R. Wright

Publisher : Addison Wesley
Pub Date : April 24, 2002
ISBN : 0-201-74612-3
Pages : 384

Copyright

Library of Congress Cataloging-in-Publication Data

Edwards, Brian M.

Interdomain mulitcast routing : practical Juniper Networks and Cisco Systems solutions / Brian M. Edwards, Leonard A. Giuliano, Brian R. Wright.

p. cm.

Includes bibliographical references and index.

ISBN 0-201-74612-3

1. Routers (Computer networks) I. Giuliano, Leonard A. II. Wright, Brian R. III. Title.

TK5105.543 .E38 2002

004.6—dc21

2002018254

Foreword

It is with great pleasure that I introduce Brian Edwards, Leonard Giuliano, and Brian Wright's book Interdomain Multicast Routing. I expect the publication of this text will improve the networking community's understanding of the promise of multicast.

In thinking about multicast, I'm drawn to two topics that the authors discuss in the introductory chapter: the question of multicast's killer application and the complexity of multicast (reverse path forwarding, packet replication, the many routing protocols in the control plane, the unique requirements of multicast with respect to scaling and network management, etc.).

One reaction to the killer application question is the observation that the ability to deliver the same content to many users in an efficient way is an important capability of a multi-service network. In other words, using a specific example, IP needs scalable multicast in order for it to subsume the delivery of services such as broadcast television. The point isn't that the broadcast television infrastructure needs to be replaced with an IP network—that's just an example—but instead the point is that scalable multicast allows a whole new set of applications and services to leverage IP networks. To make the same statement negatively, if IP networks don't support scalable multicast then a set of applications and services will be precluded from being able to leverage IP networks. The Internet community has an honorable trait of solving hard problems, even if some questions remain unanswered. The work done with multicast over the last fifteen years (!) is one manifestation of that trait.

The more interesting reaction to these topics, in my opinion, is a reaction to them as an intertwined pair. Some could argue that in the absence of a killer application, tackling a problem as complex as multicast isn't well advised. I see this argument as very shortsighted exactly because of multicast's complexity. Multicast presents an opportunity for learning about the science of networking far more than if we conservatively stay in our comfort zone. A recent example of this phenomenon in networking is MPLS. In the case of MPLS, the killer application started as fast packet forwarding but then morphed to IP VPNs, then traffic engineering, then "layer 2" VPNs and, most recently, the generalization of the MPLS signaling suite to non-packet-switching technologies (optical switching, TDM switching, and others). Work on MPLS continued, in spite of some killer applications withering away and/or becoming less trendy than newer ones, and that work taught the networking community extremely valuable lessons such as the distinction between the control plane and data plane and the advantage of having a suite of signaling protocols that can be leveraged for many applications. In the case of multicast, I believe our community of protocol designers, system vendors and, most importantly, network operators has already benefited—we have had to think very creatively about the interaction between unicast and multicast routing, the impact of various multicast routing approaches to dynamics in the control and data planes, how to design, implement and deploy multicast in highly scalable ways, how to engineer and operate services that require more than simple point-to-point connections, and so forth. Independent of multicast itself, we gain a richer intuition about networking in general and become better at designing all kinds of protocols, implementing all kinds of networking systems and deploying all kinds of network infrastructures. The fact that networking is still young enough for us to learn such lessons is one of the reasons why it is such an exciting industry!

So I invite you to read the following text with a curious and open mind. You will certainly walk away from the experience with a greater understanding of networking in general, which will better arm you for the future. And perhaps in being particularly curious and particularly open minded, you might come up with the World Wide Web for multicast!

—John W. Stewart, III
JUNOS Product Line Manager, Juniper Networks
San Francisco
January 2002

Preface

Interdomain Multicast Routing is a book on the timely technology of multicasting and is written, mainly, for network engineers responsible for configuring and maintaining that capability within their networks. It is a practical reference guide that includes Cisco Systems and Juniper Networks technology. The authors' goals are to explain the rationale and benefits of multicast routing on the Internet, to include the two leading vendors of routers and routing technology and note how they differ when applying interdomain multicast routing (IMR), and to explain the underpinnings of interdomain multicasting in simple, clear language. For a preview of the topics within this book, the following chapter listings detail the topic matter.

Chapter 1, "Interdomain Multicast Fundamentals," begins with a definition of multicast transmisson of data in contrast to other means of data delivery, within and outside the Internet, and then provides an introductory explanation of some of the issues affecting successful routing of multicast traffic on the Internet. Those seeking to understand the enormous potential for multicast may wish to tune in directly to this section.

Chapter 2, "IMR Overview," is a general description of how to generate and receive multicast traffic, including a description of methods for routers to detect sources and receivers of multicast traffic. The discussion then proceeds from multicast single-domain routing using PIM-SM (Protocol Independent Multicast-Sparse Mode) to interdomain multicast routing using MSDP (Multicast Source Discovery Protocol).

Chapter 3, "Multicast Routing Protocols," examines the two primary types of multicast routing protocols, describing the main features and examples of each.

Chapter 4, "Protocol Independent Multicast-Sparse Mode (PIM-SM)," lays out PIM-SM, the predominant multicast routing protocol for interdomain routing. Since PIM-SM is commonly used in the initial sequence of activities that gets multicast up and running within a single domain, the procedure dominates the scope of this chapter. PIM messages for both version 1 and version 2 of the protocol are covered, as is the use of anycast rendezvous point (RP) to improve load balancing and redundancy. Ample diagrams and corresponding examples describe distribution tree construction and teardown for various topologies, and the chapter ends with a discussion of multicast scoping.

Chapter 5, "Multicast Source Discovery Protocol (MSDP)," demonstrates how to use MSDP to connect multiple PIM-SM domains and subdomains. MSDP is an any source multicast (ASM) mechanism for giving Internet multicast routing its "interdomain" reach. This chapter contains a number of illustrations of the rules that determine the reverse path forwarding (RPF) peer, a critical component in MSDP. Recognizing the paucity of clear information about MSDP peer-RPF rules, which are quite complex, the authors have provided detailed rule descriptions, as well as diagrams and realistic examples. The intent is for Chapter 5 to become the most definitive guide available on the subject (MSDP peer-RPF rules). The chapter concludes with sections on mesh groups, susceptibility to operational problems, and a discussion of the prospects for the widely used MSDP vis a vis the upcoming version of Border Gateway Multicast Protocol (BGMP).

Chapter 6, "Source-Specific Multicast (SSM)," is a critical component of the book. SSM, a recent addition to the ever-changing multicast routing landscape, holds the greatest amount of promise for deployment, considering that many believe the most dominant commercial use of Internet multicast will likely conform to a one-to-many model. This chapter explains the rationale for development of this SSM service model versus ASM and how SSM can serve as a basis for learning the more complex world of ASM.

Chapter 7, "Multiprotocol Extensions for BGP (MBGP)," and Chapter 8, "Multitopology Routing in Intermediate System to Intermediate System (M-ISIS)," focus on how to create two separate virtual topologies, one for unicast and one for multicast. MBGP and M-ISIS can be used side-by-side to build a dedicated multicast RPF table, just as BGP and ISIS have traditionally coexisted in unicast intra-AS and inter-AS environments.

The remaining chapters of Interdomain Multicast Routing cover critical hands-on, real-world examples and tools. Chapter 9, "Configuring and Verifying Multicast Routing on Juniper Networks Routers," and Chapter 10, "Configuring and Verifying Multicast Routing on Cisco Systems Routers," provide practical methods and guidelines for actually configuring and verifying multicast routing on Juniper Networks and Cisco Systems routers.

Acknowledgments

The authors have been blessed with many excellent reviewers, ensuring our approach and description were accurate and unbiased. Among the many, those who spent considerable time reading our manuscripts are Ravi Prakash, Matthew Naugle, Dave Thaler, John N. Stewart, Bill Fenner, Brian Haberman, Matthew Davy, Greg Shepherd, Marshall Eubanks, Jill Gemmill, Jennifer Joy, Liming Wei, Naiming Shen, John Brassil, Walter Weiss, Tom Pusateri, Paras Trivedi, Hannes Gredler, Amir Tabdili, William Lemons, Supratik Bhattacharyya, Aviva Garrett, Patrick Ames, Hallie Giuliano, and Margaret Searing.

Further, the authors would like to thank Karen Gettman and Emily Frey of Addison-Wesley for helping to guide us through the intricacies of turning manuscripts into printed works. We would also like to pause and acknowledge the support and assistance of Juniper Networks for allowing us to work on this project and to occasionally use corporate resources.

Finally, the authors, individually, would like to thank the following friends, family, peers, and coworkers for their fortitude and inspiration:

Brian Edwards wants to thank the following people because each has provided a tremendous amount of support and guidance throughout his life and professional career: Christine Hatchett, Mabry and Linda Edwards, John Madding, Sr., Ronald Smallwood, and Chip Leonard.

Leonard A. Giuliano first thanks his loving wife, Hallie, along with his parents and sisters for their endless support and encouragement. Over the years, he has had the pleasure to work with and learn a great deal from the following individuals: Amir Tabdili, Gary Barnhart, Rob Rockell, James Milne, Dale Morey, James Zahniser, Timothy Flynn, Tom Pusateri, Greg Shepherd, Mujahid Khan, Jeff Loughridge, Paras Trivedi, Peter Lothberg, Supratik Bhattacharyya, and Christophe Diot. Finally, he would like to thank the following individuals for challenging and encouraging him to learn more: Patricia Kendall, Ralph Lane, Bill Lemons, Daemon Morrell, Stephen Miller, Marty Schulman, and Kaydon Stanzione.

Brian Wright especially salutes his coauthors for their practical contributions to the development of IMR, as well as their idea that a hands-on book on the subject would be worthwhile. Thanks to the aforementioned first-rate talent at Juniper Networks and Addison-Wesley. And he would like to mention, in particular, the following individuals among many who, through the seasons or from time to time, inspired or helped him to develop personally and/or professionally: Truman and Phyllis Wright, Rose Wright, Trese Hercher, James Cline, James Castner, Charles Nelson, Al Suggs, Kathy Kennelly, Serita Lockhart, Sam Mills, Cathy Keller, Mary Jo David, Pat Markey, Jordan Mergist, Glen Gibbons, Charlie Christal, Teena Thompson, Michael Boughner, Nora Kryza, Paul Swantek, Stephen Brancaleone, Kimberly Hall, and Brenda Ackerman.

— Brian Edwards, Leonard Giuliano, Brian Wright
January 2002

Chapter 1. Interdomain Multicast Fundamentals

This chapter introduces and describes the fundamental concepts of multicast. Subsequent chapters build upon these concepts, illustrating how they are specifically used in the protocols and technologies that enable the operation of interdomain multicast. This chapter also defines terms and conventions that will be used throughout the book.

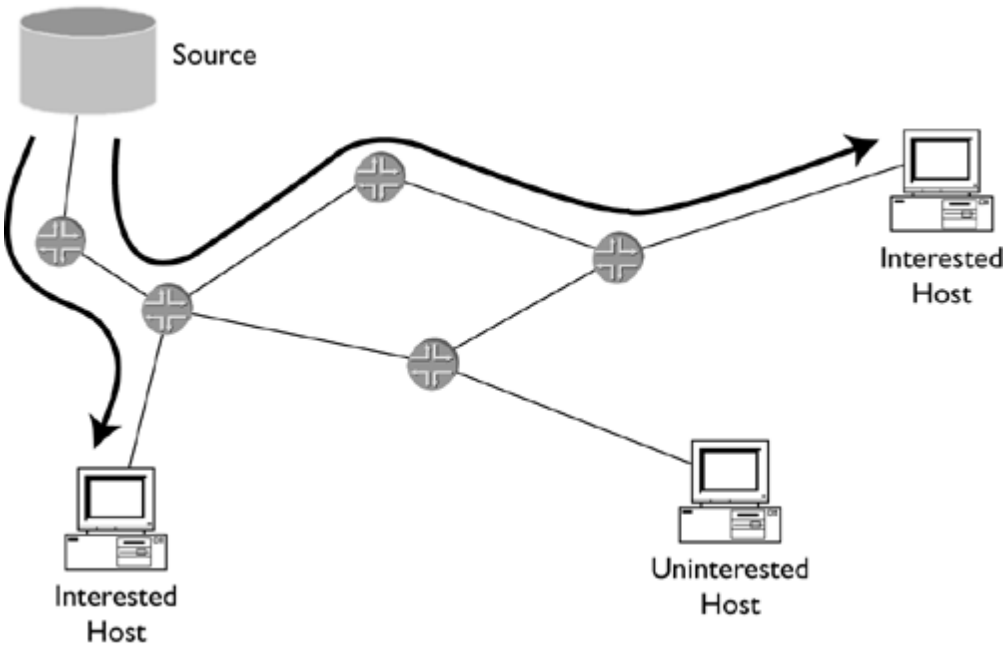
1.1 What Is Multicast?

The three main methods of data delivery are unicast, broadcast, and multicast. These methods are summarized as follows:

- Unicast: Data is delivered to one specific recipient, providing one-to-one delivery.
- Broadcast: Data is delivered to all hosts, providing one-to-all delivery.
- Multicast: Data is delivered to all hosts that have expressed interest. This method provides one-to-many delivery.

The Internet was built primarily on the unicast model for data delivery (see Figure 1-1). However, unicast does not efficiently support certain types of traffic.

Figure 1-1. Unicast delivery



Multicast, originally defined in RFC 1112 by Steve Deering, provides an efficient method for delivering traffic that can be characterized as "one-to-many" or "many-to-many."

Radio and television are examples of traffic that fit the one-to-many model. With unicast, a radio station would have to set up a separate session with each interested listener. A duplicate stream of packets would be contained in each session. The processing load and the amount of bandwidth consumed by the transmitting server increase linearly as more people tune in to the station. This might work fine with a handful of listeners; however, with hundreds or thousands of listeners, this method would be extremely inefficient. With unicast, the source bears the burden of duplication.

Using broadcast (see Figure 1-2), the radio station would transmit only a single stream of packets, whether destined for one listener or for one million listeners. The network would replicate this stream and deliver it to every listener. Unfortunately, people who had not even tuned in to the station would be delivered this traffic. This method becomes very inefficient when many uninterested listeners exist. Links that connect to uninterested end hosts must carry unwanted traffic, wasting valuable network resources. With broadcast, the network carries the burden of delivering the traffic to every end host.

Figure 1-2. Broadcast delivery

1.2 Internetworking Basics

To facilitate the reader's understanding, this section covers some of the notation and conventions used in the book and thus indicates the level of the typical reader's internetworking knowledge anticipated by the authors.

Throughout the book we use the slash notation for bit mask when describing IP address ranges. The slash notation indicates how many bits of the address remain constant throughout the range of addresses. For example, 10.0.0.0/8 indicates a range of IP addresses all with the first 8 bits equal to 10. The range is from address 10.0.0.0 to 10.255.255.255.

We also make reference to classful networks. The class A, B, and C networks constitute all unicast IP addresses as follows:

- Class A networks: Describe the range of networks from 1.0.0.0/8 through 126.0.0.0/8.
- Class B networks: Describe the range of networks from 128.0.0.0/16 through 191.255.0.0/16.
- Class C networks: Describe the range of networks from 192.0.0.0/24 through 223.255.255.0/24.

Originally, networks were assigned to organizations along classful boundaries. That meant class A networks were assigned in /8 blocks, class B in /16 blocks, and class C in /24 blocks. Classful allocation was inefficient because organizations that required slightly more than 254 addresses could be assigned an entire class B. Classless interdomain routing (CIDR) enabled the assignment and routing of addresses outside of classful boundaries. An organization that needed enough addresses for 500 hosts could be assigned one /23, instead of an entire class B network.

All multicast addresses fall in the class D range of the IPv4 address space. The class D range is 224.0.0.0 through 239.255.255.255. Multicast addresses do not have a mask length associated with them for forwarding purposes. Each address is treated independently so the mask used for forwarding is always assumed to be /32. We use shorter mask lengths on multicast addresses in some parts of the book for reasons other than forwarding. These masks generally are used to describe ranges of multicast addresses. For example, the address range reserved for Source-Specific Multicast (SSM) is 232.0.0.0/8.

We refer throughout the book to unicast and multicast routing protocols. Unicast routing protocols are used by routers to exchange routing information and build routing tables. Unicast IP routing protocols are further categorized into interior gateway protocols (IGPs) and exterior gateway protocols (EGPs).

IGPs provide routing within an administrative domain known as an autonomous system (AS). EGPs provide routing between ASs. Routing Information Protocol (RIP), Open Shortest Path First (OSPF), and Intermediate System to Intermediate System (IS-IS) are examples of IGPs, while Border Gateway Protocol (BGP) is an example of an EGP. Multicast routing protocols are used by routers to set up multicast forwarding state and to exchange this information with other multicast routers. Examples of multicast IP routing protocols are Distance Vector Multicast Routing Protocol (DVMRP), Protocol Independent Multicast–Dense Mode (PIM-DM), and Protocol Independent Multicast–Sparse Mode (PIM-SM).

The terms control packets and data packets are used to differentiate the types of packets being routed through the network. Control packets include any packets sent for the purpose of exchanging information between routers about how to deliver data packets through the network. Control packets are typically protocol traffic that network devices use to communicate with one another to make such things as routing possible.

Data packets use the network to communicate data between hosts; they do not influence the way the network forwards traffic. Letters delivered via postal mail are analogous to data packets. Information exchanged between post offices to describe what ZIP codes mean is analogous to control packets. In the IP world, all packets sent for an FTP

1.3 Multicast Basics

A multicast address is also called a multicast group address. A group member is a host that expresses interest in receiving packets sent to a specific group address. A group member is also sometimes called a receiver or listener. A multicast source is a host that sends packets with the destination IP address set to a multicast group. A multicast source does not have to be a member of the group; sourcing and listening are mutually exclusive.

Because there can be multiple receivers, the path that multicast packets take may have several branches. A multicast data path is known as a distribution tree. Data flow through the multicast distribution trees is sometimes referenced in terms of upstream and downstream. Downstream is in the direction toward the receivers. Upstream is in the direction toward the source. A downstream interface is also known as an outgoing or outbound interface; likewise, an upstream interface is also known as an incoming or inbound interface.

Routers keep track of the incoming and outgoing interfaces for each group, which is known as multicast forwarding state. The incoming interface for a group is sometimes referred to as the IIF. The outgoing interface list for a group is sometimes referred to as the OIL or olist. The OIL can contain 0 to N interfaces, where N is the total number of logical interfaces on the router.

Multicast forwarding state in a router is typically kept in terms of "(S,G)" and "(*,G)" state, which usually are pronounced "ess comma gee" and "star comma gee," respectively. In (S,G), the "S" refers to the unicast IP address of the source. The IP header of the multicast data packet contains S as the packet's source address. The "G" represents the specific multicast group IP address of concern. The IP header of the multicast data packet contains G as the packet's destination address. So for a host whose IP address is 10.1.1.1 acting as a source for the multicast group 224.1.1.1, (S,G) state would read (10.1.1.1,224.1.1.1).

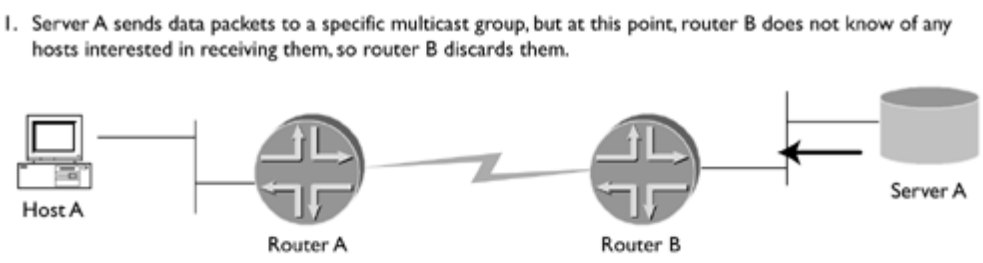
In (*,G) notation, the asterisk (*) is a wild card used to denote the state that applies to any source sending to group G. A multicast group can have more than one source. If two hosts are both acting as sources for the group 224.1.1.2, (*,224.1.1.2) could be used to represent the state a router could contain to forward traffic from both sources to the group. The significance of (S,G) and (*,G) state will become more apparent when we discuss shortest path and shared trees in Chapters 2 and 3.

1.3.1 Reverse Path Forwarding

Multicast routing involves a significant paradigm change from standard unicast routing. In general, routers make unicast routing decisions based on the destination address of the packet. When a unicast packet arrives, the router looks up the destination address of the packet in its routing table. The routing table tells the router out from which interface to forward packets for each destination network. Unicast packets are then routed from source to destination.

In multicast, routers set up forwarding state in the opposite direction of unicast, from receiver to the root of the distribution tree. Routers perform a reverse path forwarding (RPF) check to determine the interface that is topologically closest to the root of the tree (see Figure 1-4). RPF is a central concept in multicast routing. In an RPF check, the router looks in a routing table to determine its RPF interface, which is the interface topologically closest to the root. The RPF interface is the incoming interface for the group.

Figure 1-4. Reverse path forwarding (RPF)



1.4 Interdomain Multicast Routing

For years multicast has enjoyed niche success in many financial and enterprise networks. Financial institutions have applications, such as stock tickers, that require sharing the same data across the network. Using unicast for these applications is inefficient and not cost effective. Likewise, some enterprise networks serve companies with applications ideally suited to multicast delivery—for example, a central headquarters that must feed hundreds of branch sites with price lists and product information. Transferring these identical files to all sites individually with unicast simply is not efficient.

In the past, enterprise networks have frequently looked much different than the networks managed by Internet service providers (ISPs). This difference existed because these networks had to meet a set of radically different requirements. Enterprise networks connect the offices of a single company, which often involves transporting primarily a single type of data (for example, file transfer). Transporting only a single type of data enables the network to be built in a way that optimizes delivery of that type of traffic. Also, few, if any, of the routers in an enterprise network connect to routers controlled by another entity.

ISP networks couldn't be more different. ISPs can have up to thousands of different customers, each a separate administrative entity. The data can include an unclassifiable mix of voice, video, e-mail, Web, and so on. Providing ubiquitous support for these various traffic types across the interdomain world of the Internet has always set ISPs apart from enterprises in the way they are designed and operated.

Unicast and multicast routing on enterprise and financial networks has often involved deploying protocols and architectures that best meet the needs of the companies they connect. These protocols and architectures often do not address the scalability and interdomain requirements of ISPs. However, recent trends have shown that the networking needs of enterprises have evolved to more closely resemble those of ISPs. Accordingly, many enterprise networks today are beginning to use the same principles and philosophies found in the engineering of ISPs' networks, albeit on a smaller scale.

The focus of this book is to describe the technologies and challenges faced by ISPs when deploying and operating multicast across the Internet. The first reason for this focus is neglect. Most networking books concentrate on enterprise networks rather than the unique demands of service provider networks. Second, ISP networks generally possess the superset of requirements that are found on other types of networks. For example, financial networks typically need to support many-to-many applications. Other enterprise networks may need to support only one-to-many applications. Because ISPs may be delivering service to both types of networks, they must be equipped to handle both types of applications. Additionally, ISP networks have scalability demands that are rarely found on any other types of networks.

While ISPs continue to have unique requirements for scalability and interdomain stability, most of the same multicast technologies found in ISP networks can be applied for use on other networks. By adopting these ISP philosophies, financial and enterprise networks are capable of ubiquitously supporting all types of multicast traffic. This flexibility enables a network to be prepared if traffic types change in the future.

The scope of this book is confined to the protocols and technologies currently used in the production networks of service providers. In order to provide a pragmatic examination of the challenges faced by ISPs today, little to no mention is made of protocols that have not been implemented by routing vendors or deployed by service providers at the time of writing. Accordingly, IPv6 is outside the scope of this book.

1.5 Where Is Multicast?

The Multicast Backbone, or MBone, refers to the networks on the Internet that are enabled for multicast. The original MBone was built in the early 1990s as a network of multicast-enabled routers that were connected by tunnels. These routers were frequently UNIX servers running multicast routing software developed before router vendors had stable implementations of multicast software.

Tunnels allowed these early multicast-enabled "islands" to appear to be virtually connected to one another. Multicast packets were encapsulated within unicast packets and sent in the tunnel. Routers that were not multicast-enabled simply saw the unicast IP packet and routed it toward the tunnel destination. When the unicast packet reached the tunnel destination, the router decapsulated the unicast header to find the multicast packet within. If that packet had to be forwarded to another tunneled router, it was once again encapsulated and sent out another tunnel.

As router vendors implemented more stable multicast routing code, ISPs began to replace tunnels with native multicast routing in the late 1990s. Native multicast routing means routers forward raw multicast packets without encapsulating the multicast data within unicast packets. Most of the world's largest ISPs are multicast-enabled in at least some portion of their production networks today.

Multicast Internet Exchanges (MIXs) were built to connect multicast-enabled ISPs. MIXs are usually found in network access points (NAPs) where ISPs publicly peer with one another. A MIX enables ISPs to exchange multicast traffic on separate equipment from what is used for unicast peering. SprintNAP, in Pennsauken, New Jersey, and the NASA Ames Research Center Federal Internet Exchange (FIX-West), in Mountain View, California, contain two of the most popular MIXs used for public multicast peering.

Most people think of the old tunneled network of UNIX boxes when they hear the word "MBone," but it technically refers to any network that is multicast-enabled. Unanimous agreement has not been reached on a catchy word or phrase to colloquially refer to the native multicast-enabled portion of the Internet.

1.6 Multicast on the LAN

Throughout this book we focus primarily on the protocols that enable multicast packets to be forwarded within and between different domains. However, to provide a complete picture, we should examine what occurs on the link, or local area network (LAN), on which group members reside.

1.6.1 IGMP

When a host wants to become a multicast receiver, it must inform the routers on its LAN. The Internet Group Management Protocol (IGMP) is used to communicate group membership information between hosts and routers on a LAN.

To join a multicast group that is not already being forwarded on its LAN, a host sends an IGMP Report to a well-known multicast group. All IGMP-enabled routers on that LAN are listening to this group. Upon hearing a host's IGMP Report for a multicast group, G, one of the routers on the LAN uses a multicast routing protocol to join that group. In the case of PIM-SM, this router sends a (*,G) Join toward the RP for the specified group.

IGMP versions 1 and 2 allow a host to specify only the group address that it is interested in receiving. IGMP version 3 allows a host to express interest in only specified sources of a group, triggering an (S,G) Join by a PIM-SM router on the LAN. This is a key component of Source-Specific Multicast, which we examine in section 1.7.

A host must support IGMP in order to receive multicast packets. The version of IGMP supported is a function of the host's operating system. For example, unless otherwise modified, PCs running Windows 95 support IGMPv1. Likewise, PCs running Windows 98 or 2000 support IGMPv2, while IGMPv3 is available in Windows XP.

1.6.2 IGMP Proxying

When a host reports interest in a multicast group from a source outside its LAN, it is the responsibility of a router on the LAN to join that group using a multicast routing protocol like PIM-SM. However, some routers do not support any multicast routing protocols. Low-end routers and legacy equipment such as dialup remote access servers (RAS) are examples of routing devices that sometimes do not support any multicast routing protocols.

Nearly all routing devices support IGMP. A common technique used in routers that do not support any multicast routing protocols is IGMP proxying. A router that hears an IGMP Report from a host simply relays that IGMP message to an upstream router that does support a multicast routing protocol. IGMP messages simply "hop over" a local router and reach a router that is capable of joining the group via a protocol like PIM-SM. IGMP proxying lowers the bar that low-end routing devices need to meet in order to deliver multicast.

1.6.3 Layer 3 to Layer 2 Mapping

The layers of the OSI reference model that we are most concerned with in this book are the data link, or layer 2, and the network, or layer 3. Here we focus on Ethernet, by far the most common layer 2 LAN technology. All layer 3 packets, in this case IP, are encapsulated with an Ethernet header and trailer and transmitted onto a LAN as an Ethernet frame.

All devices on the Ethernet have a unique 48-bit Media Access Control (MAC) address. To speak to one another, devices on the LAN keep a table that maps unicast IP addresses to MAC addresses. When packets are encapsulated in frames, the destination MAC address in the frame header is set to the MAC address corresponding to the IP address in the header of the IP packet.

IP multicast packets are destined to class D group addresses, which do not correspond with a single end host. Likewise, the MAC address used for multicast packets cannot be the address of a single station on the LAN. A special range of MAC addresses must be used for multicast.

The high order four bits of the first octets of class D addresses are always the same. Thus 28 bits may be varied in a

1.7 ASM versus SSM

The original vision for multicast in RFC 1112 supported both one-to-many and many-to-many communication models and has come to be known as Any-Source Multicast (ASM). Radio and television, as we have already discussed, are obvious examples of the one-to-many model. Applications such as online gaming and videoconferencing, in which some or all of the participants become sources, are examples of the many-to-many model. To support the many-to-many model, the network is responsible for source discovery. When a host expresses interest in a group, the network must determine all of the sources of that group and deliver them to the receiving host.

The mechanisms that provide this control plane of source discovery contribute the majority of the complexity surrounding interdomain multicast. However, applications that are believed to possess the greatest potential for commercial viability on the Internet use the one-to-many model. Since the bulk of the complexity is providing the least important functionality, the "ratio of annoyance" is disproportionately high in ASM.

It recently has been suggested that by abandoning the many-to-many model, multicast could deliver more "bang for the buck" on the Internet. By focusing on the one-to-many model, the most appealing of multicast applications could be supported while vastly reducing the amount of complexity required. Source-Specific Multicast (SSM) is a service model that supports multicast delivery from only one specified source to its receivers.

By sacrificing functionality that many may consider less important on the Internet, the network no longer needs to provide the control plane for source discovery. This control plane is now the responsibility of receivers. Typically, the application layer (via a mouse click, for example) informs the receiver who the source is. When the receiver informs its directly connected router that it is interested in joining a group, it specifies the source as well as the group. This last-hop router is then able to join the SPT directly, instead of having to join the RPT.

SSM eliminates the need for RPTs, RPs, and Multicast Source Discovery Protocol (MSDP), radically simplifying the mechanisms needed to deliver multicast. Best of all, this service model is realized through a subset of functionality already present in existing protocols. Very little needs to be added.

It is important to note that ASM and SSM are service models, not protocols. Different protocols are implemented and configured to deliver the service model. For example, SSM is a service model that is realized through a subset of functionality of PIM-SM and IGMPv3. The first five chapters of this book examine interdomain multicast generally from an ASM point of view because ASM is much more interesting from a protocol perspective. With a clear understanding of ASM, the operation and benefits of SSM become apparent. Chapter 6 describes SSM in detail.

1.8 Addressing Issues

The addresses available for multicast usage range from 224.0.0.0 to 239.255.255.255. This plentiful, but finite, range is controlled by the Internet Assigned Numbers Authority (IANA). Certain subranges within the class D range of addresses are reserved for specific uses:

- 224.0.0.0/24: The link-local multicast range
- 224.2.0.0/16: The Session Announcement Protocol (SAP)/Session Description Protocol (SDP) range
- 232.0.0.0/8: The SSM range
- 233.0.0.0/8: The AS-encoded, statically assigned GLOP range (RFC 3180)
- 239.0.0.0/8: The administratively scoped multicast range (RFC 2365)

For a complete list of IANA assigned multicast addresses, refer to the http://www.iana.org/assignments/multicast-addresses Web site.

If class D addresses had been assigned in the same manner unicast addresses were allocated, this address space would have been exhausted long ago. In general, IANA allocates static multicast addresses only used for protocol control. Examples of this type of address include

- 224.0.0.1: All systems on this subnet
- 224.0.0.2: All routers on this subnet
- 224.0.0.5: OSPF routers
- 224.0.0.6: OSPF designated routers (DRs)
- 224.0.0.12: DHCP server/relay agent

To protect against address exhaustion, a simple dynamic address allocation mechanism is used in the SAP/SDP block. Applications such as Session Directory Tool (SDR) that use this mechanism randomly select an unused address in this range. This dynamic allocation mechanism for global multicast addresses is somewhat analogous functionally to DHCP, which dynamically assigns unicast addresses on a LAN.

Unfortunately, some applications require the use of static multicast addresses. GLOP, described in RFC 3180, provides static multicast ranges for organizations that already have reserved an AS number. In GLOP, an AS number is used to derive a /24 block within the 233/8 range. The static multicast range is created in the following form:

```
233.[first byte of AS number].[second byte of AS number].0/24
```

For example, AS 12345 is automatically allocated 233.48.57.0/24. Here is an easy way to compute this:

1.9 Applications

The most widely used application on the old MBone was SDR. By launching SDR, a host listens to the well-known SAP group, 224.2.127.254. Any source host that wants to advertise a session (usually audio and/or video) describes its session in SDP messages. These SDP messages contain the address of the source, type of session, contact information for the source, and so on and are transmitted on the SAP multicast group.

Thus every host running SDR learns about every session on the Internet by receiving these SDP messages on the SAP group. By clicking one of the sessions listed in the SDR window, applications such as VIC (video conferencing tool) or VAT (visual audio tool) are launched to display the video or audio. An interesting feature of many of the applications launched by SDR is that by joining a session, you also become a source for that session. Because every receiver is also a source, each participant can see the others, which makes these applications ideal for collaboration and videoconferencing (and unscalable for sessions with lots of participants!).

Most agree that SDR is a "neat little toy" but not really a commercially viable application. Most SDR sessions are "cube-cams" or video camera shots of ISP parking lots. Because SDR acts as a global directory service for all multicast content on the Internet, it is not expected to scale to support large numbers of sessions.

Windows Media Player (WMP) is currently a popular application for accessing multicast audio and video content. WMP has excellent scaling potential for the Internet because, unlike many SDR-launched applications, receivers do not become sources to the group they join. Also, WMP has the capability to attempt to join a multicast session first, failing over to a unicast session if unsuccessful, which is ideal for content providers seeking the efficiency of multicast and the availability of unicast. Cisco System's IP/TV is another promising application for delivering multicast multimedia content. IP/TV supports multicast content only.

Juniper Networks and Cisco System routers can be configured to listen to the SAP group and keep a cache of SDR sessions. Joining the SAP group is useful in troubleshooting. It is a quick and easy way to determine whether the router has multicast connectivity with the rest of the Internet.

1.10 Multicast Performance in Routers

When deploying multicast, it is important to consider whether the routers in a network are well suited to support multicast. Just as some cars provide speed at a cost of safety, some routers provide unicast performance at a cost of multicast. As high-end routers are built to scale to terabits and beyond, router designers sometimes compromise multicast performance to optimize unicast forwarding. The two most important considerations when evaluating a router for multicast are state and forwarding performance.

A router must keep forwarding state for every multicast group that flows through it. Pragmatically, this means (S,G) and (*,G) state for PIM-SM. It is important to know how many state entries a router can support without running out of memory. MSDP-speaking routers typically keep a cache of Source-Active messages. Likewise, knowing the maximum number of Source-Active entries a router can hold in memory is crucial.

The obvious next question is "how many entries should a router support?" Like many questions in life, there is no good answer. Past traffic trends for multicast are not necessarily a reliable forecast for the future. Traffic trends for the Internet in general are rarely linear. Growth graphs of Internet traffic frequently resemble step functions, where stable, flat lines suddenly yield to drastic upward surges that level off and repeat the cycle.

The best policy is to select a router that can hold far more state than even the most optimistic projections require and monitor memory consumption. When state in a router begins to approach maximum supportable levels, take appropriate action (upgrade software or hardware, redesign, apply rate limits or filters, update your resume, and so on). With the exception of the Ramen worm attacks (see Chapter 5), state has not been much of a problem yet. Of course, as with mutual funds, past performance does not ensure future success.

Forwarding performance is characterized by throughput and fanout. Throughput describes the maximum amount of multicast traffic a router can forward (in packets per second or bits per second). Fanout describes the maximum number of outgoing interface for which a router can replicate traffic for a single group. As port densities in routers increase, maximum supported fanout becomes a critical factor. Also, it should be understood how increasing fanout levels affects throughput. As is the case with state, it is important to be aware of the performance limits, even if the exact amount of multicast traffic on the network is not known.

Forwarding performance is primarily a function of hardware. The switching architecture a router uses to forward packets is usually the most important factor in determining the forwarding performance of a hardware platform. Shared memory switching architectures typically provide the best forwarding performance for multicast. A shared memory router stores all packets in a single shared bank of memory.

Juniper Networks' M-series routers employ a shared memory architecture that is very efficient for multicast. In this implementation, multicast packets are written into memory once and read out of the same memory location for each outgoing interface. Because multicast packets are not written across multiple memory locations, high throughput levels can be realized regardless of fanout.

Some routers are based on a crossbar switching architecture. The "crossbar" is a grid connecting all ports on the router. Each port shows up on both the X and Y axes of the grid, where the X axis is the inbound port and the Y axis is the outbound port. With the crossbar architecture, packets wait at the inbound port until a clear path is on the crossbar grid to the outbound port. Inbound traffic that is destined for multiple egress ports must be replicated multiple times and placed in multiple memory locations. Because of this, routers with crossbar architectures usually exhibit multicast forwarding limitations.

Router designers sometimes work around this inherent challenge by creating a separate virtual output queue dedicated to multicast and giving the queue higher priority than the unicast queues. Unfortunately, this technique can cause multicast traffic to suffer head-of-line blocking, which occurs when packets at the head of the queue are unable to be serviced, preventing the rest of the packets in the queue from being serviced as well. Such a design assumes multicasts are a small percentage of total traffic because a router incorporating this design would be inefficient under a high multicast load.

1.10.1 RP Load

1.11 Disclaimers and Fine Print

Throughout this book, reference is made to RFCs (Request for Comments) and Internet Engineering Task Force (IETF) Internet-Drafts. Internet-Drafts are submitted to the IETF as working documents for its working groups. If a working group decides to advance an Internet-Draft for standardization, it is submitted to the Internet Engineering Steering Group (IESG) to become an RFC. RFCs are the closest things to the official laws of the Internet. For a good description of Internet-Drafts and the various types of RFCs, visit http://www.ietf.org/ID.html.

It is not uncommon for protocol-defining Internet-Drafts never to reach RFC status. Likewise, vendors do not always implement protocols exactly as they are defined in the specification. Internet-Drafts that are not modified after six months are considered expired and are deleted from the IETF Web site. All RFCs and current Internet-Drafts can be found at the IETF's Web site. A good way to find an expired Internet-Draft is by searching for it by name at http://www.google.com. A search there will usually find it on a Web site that mirrors the IETF Internet-Drafts directory without deleting old drafts. Unless otherwise stated, all Internet-Drafts and RFCs mentioned in this book are current at the time of writing. These documents are constantly revised and tend to become obsolete very quickly.

Similarly, the implementations of Juniper Networks and Cisco System routers, the routers most commonly found in ISP networks, are described throughout this book. The descriptions and configurations are meant to assist engineers in understanding the predominant implementations found in production networks and provide a starting point for configuration. They are not the official recommendations of these vendors. It is also important to note that these vendors are constantly updating and supplementing their implementations. For officially supported configurations, it is best to contact these vendors directly.

1.12 Why Multicast?

In less than a decade, the Internet has gone from a little known research tool to a dominant influence in the lives of people around the globe. It has created an age in which information can be disseminated freely and equally to everyone. The Internet has changed the way people communicate, interact, work, shop, and even think. It has forced us to reconsider many of our ideas and laws that had been taken for granted for decades.

Any person on earth with a thought to share can do so with a simple Web page, viewable to anyone with a connection onto the network. When considering the revolutionary impact their achievements have had on the way people interact, it is not ludicrous to mention names like Cerf, Berners-Lee, and Andreessen in the same breath as Gutenberg and Bell.

Nearly every aspect of communication in our lives is tied in one way or another to the Internet. Noticeably absent, however, in the amalgamation of content that is delivered prominently across the Internet is video. Video is an ideal fit for the Internet. While text and pictures do well to convey ideas, video provides the most natural, comfortable, and convenient method of human communication.

Even the least dynamic examples of video reveal infinitely more than the audio-only versions. For example, accounts of the 1960 Nixon-Kennedy debates varied widely between those who had watched on TV and those who had listened on the radio. So why then is video restricted primarily to the occasional brief clip accessible on the corner of a Web page and not a dominant provider of content for the Internet?

The answer is simple: The unicast delivery paradigm predominant in today's Internet does not scale to support the widespread use of video. Earlier attempts, such as the webcasts of the Starr Hearings and the Victoria's Secret fashion show, have failed to demonstrate otherwise.

The easiest target for video's lack of pervasiveness on the Internet has always been the limited bandwidth of the "last mile." It has often been argued that potential viewers simply do not have pipes large enough to view the content. However, with the proliferation of technologies like digital subscriber line (DSL) and cable modems, widespread residential access to video of reasonably adequate quality exists. Furthermore, for years, the number of people employed in offices with broadband Internet connectivity has been substantial. Finally, with nearly every college dorm room in the United States (and increasingly throughout the world) equipped with an Ethernet connection, client-side capacity is quickly becoming a nonissue.

The server side, on the other hand, has principally relied on unicast to deliver this content. The cost required to build an infrastructure of servers and networks capable of reaching millions of viewers is simply too great, if even possible. Compare that to the cost of delivery with multicast, where a content provider with only a server powerful enough and bandwidth sufficient to support a single stream is potentially able to reach every single user on the Internet.

Interestingly, while it has always been viewed as a bandwidth saver, the previously mentioned efficiency underscores multicast's capability as a bandwidth multiplier. With a multicast-enabled Internet, every home can be its own radio or television station with the ability to reach an arbitrarily large audience. If Napster created interesting debates on copyright laws, imagine the day when everyone on earth will be able to watch a cable television channel multicast from your very own PC.

It is worth noting that multicast need not be used solely for video. Multicast provides efficient delivery for any content that uses one-to-many or many-to-many transmission. File transfer, network management, online gaming, and stock tickers are some examples of applications ideally suited to multicast. However, multimedia, and more specifically video, is widely agreed to be the most interesting and compelling application for this delivery mechanism.

The brief history of the Internet suggests the inevitability that it someday will be a prevalent vehicle for television and radio, as all data networks converge onto a single common IP infrastructure. Accepting this, multicast provides the only scalable way to realize this vision. With such great potential for providing new services, it is logical to wonder why multicast has not been deployed ubiquitously across the Internet. In fact, to this point, the deployment has actually been somewhat slow.

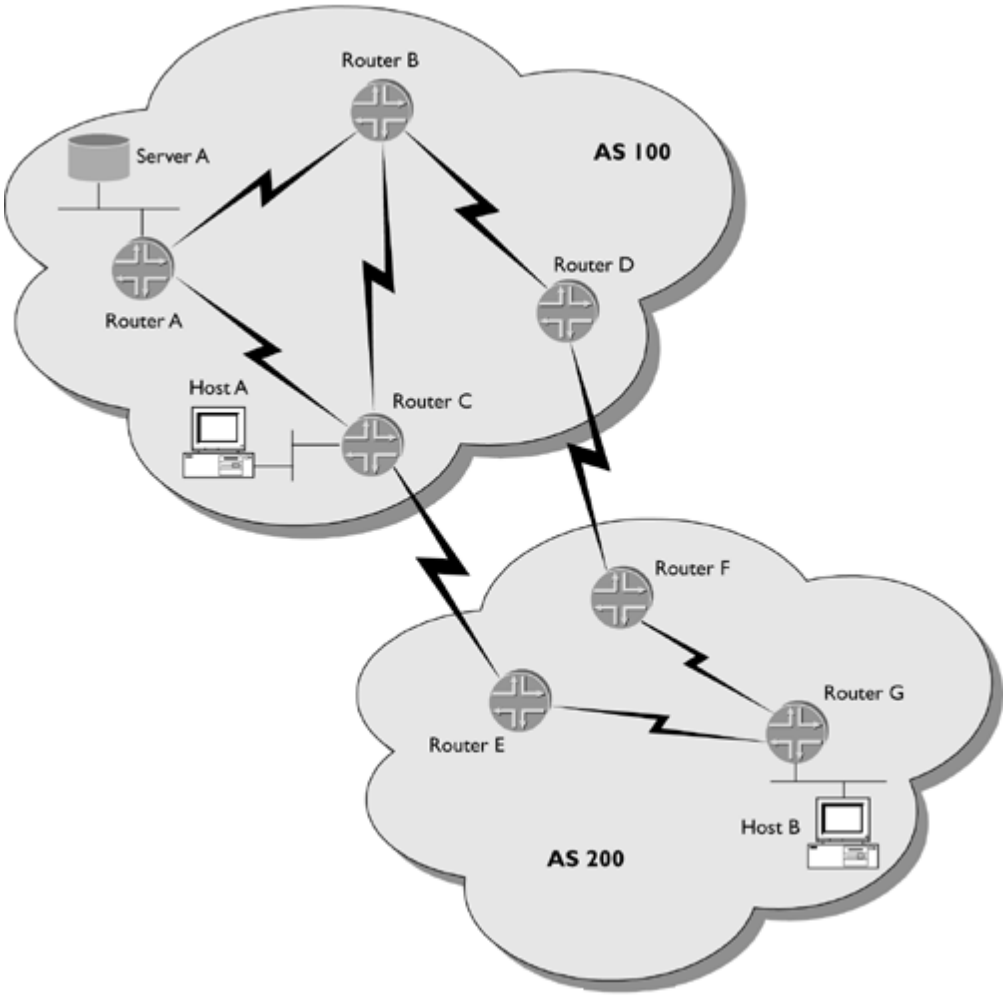
Chapter 2. IMR Overview

Several protocols are required to enable IP multicast over multiple domains. These protocols, as well as their documentation, have been developed independently, and each has its own set of specific terms. The multiplicity of protocol development makes it challenging to create a workable, integrated multicast implementation.

The point of this chapter is to resolve the difficulty of implementation using these several protocols and to provide a working-level illustration of an interdomain multicast routing (IMR) system, end-to-end across multiple domains.

Figure 2-1 shows a simple network and serves as a reference for subsequent discussion in this chapter. The figure shows two interconnected autonomous systems (ASs) with fully functional unicast routing.

Figure 2-1. Base internetwork



Each AS is controlled by an independent organization and has its own IGP. Based on information gained from the independent IGPs, router C and router D, in Figure 2-1, run an External Border Gateway Protocol (EBGP) session to exchange routing information with router E and router F, respectively.

In the remainder of this chapter, we show the step-by-step process of how to enable these two autonomous systems to route multicast traffic between them. Each step concentrates on a portion of Figure 2-1 and describes the mechanisms that must be put in place for operational multicast routing. Specifically, the focus is on enabling host B to receive multicast traffic from server A.

2.1 Receiving Multicast Traffic: IGMP from the Perspective of the Host

A host must run IGMP in order to receive multicast packets. Currently, three versions of IGMP exist:

- Version 1: Described in RFC 1112
- Version 2: Described in RFC 2236
- Version 3: Described in draft-ietf-idmr-igmp-v3-09.txt

Hosts use IGMP to express their interest in specific multicast groups. A properly behaving host performs two tasks to join a multicast group:

- Begins listening on the layer 2 address that maps to the IP multicast group address
- Reports interest in joining a group by sending a Host Membership Report message, which triggers one of the routers on the LAN to join the group using a multicast routing protocol such as PIM-SM

To refresh state, a router on the host's LAN periodically sends IGMP Host Membership Query messages. Hosts send Report messages in response to these Query messages for each group in which they are interested.

Note

Multiple hosts on the same subnet may be interested in the same group, but it is necessary for only one host to respond to the IGMP Query in order for the router to forward traffic destined to the group onto the subnet.

To avoid a condition in which all hosts bombard the local network with redundant information, two strategies are used:

- When a Query message is received, a host waits a random amount of time to respond for each group in which it is interested.
- The host, in its IGMP version 1 or 2 Report message responses, sets the destination address to the group address being reported. In IGMP version 3, the destination address of Report messages is 224.0.0.22. If a host hears a report from another host on the subnet, it suppresses the sending of its own report for that group.

The primary difference between IGMPv1 and IGMPv2 is the way they handle hosts leaving a group. In IGMPv1, when a host is no longer interested in listening to a group, it simply stops sending reports for that group. After some amount of time has passed without hearing any reports from any hosts, the router assumes all hosts on the LAN have left the group and stops forwarding traffic for that group onto the LAN. IGMPv2 introduced the concept of explicit leave with the addition of a Leave-Group message. This message enables hosts to report they are no longer interested in a group. The router responds to this message with a group-specific Query message to determine whether any other hosts are still interested in the group. If no other hosts respond with interest in the group, the router stops forwarding traffic onto the LAN immediately. This mechanism dramatically reduces leave latency.

IGMPv3 adds support for exclude and include modes. Exclude mode enables a host to request multicast packets for

2.2 Detecting Multicast Receivers: IGMP from the Perspective of the Router

IGMP Query messages are sent to the ALL-SYSTEMS multicast group (224.0.0.1) with an IP time-to-live (TTL) value set to 1. If more than one router exists on a subnet, the router whose Query messages contain the lowest-numbered IP source address is elected as the active querier for the subnet. Other routers on the subnet suppress sending of IGMP queries but still listen to and cache the information included in all IGMP messages.

On each IGMP-enabled interface, a router keeps an IGMP group cache, which is a simple table that keeps track of the following information:

- A list of all groups that have interested hosts
- The IP address of the host that last reported interest for each group
- The timeout value for each entry in the table

The timeout is the amount of time remaining for each group before the router determines no more members of a group exist on the interface. This timer is reset each time the router receives a Membership Report for that group on that interface. The value to which the timer is reset is called the group membership interval (GMI) and is calculated as follows:

GMI = (robustness variable x query interval) + query response interval

The query interval is the interval between general queries sent by the querier. Its default value is 125 seconds. By varying the query interval, an administrator may control the number of IGMP messages on the network; larger values cause IGMP queries to be sent less frequently.

The query response interval is the amount of time hosts have to respond to a IGMP Query. A 10-second default value is encoded in the Query message, which provides a time limit for the host's initial response. The host then randomly selects a response time between zero and this maximum, as described in the previous section.

The robustness variable defines the number of queries that can be sent without receiving a response before the cache entry times out. Increasing the robustness variable from its default of 2 safeguards against packet loss but increases the amount of time before the router detects that additional interested hosts truly do not exist.

2.3 Generating Multicast Traffic

Generating multicast traffic requires no additional protocols. Server A simply starts sending traffic to a multicast group address. It is worth noting that the source can be transmitting multicast data into the network while there are no interested receivers. Nothing tells the source to "start sending." This provides inefficiency on the source's LAN because the source can be blasting traffic that no one is interested in receiving.

There have been some preliminary discussions about adding mechanisms that would enable the network to inform a source when receivers exist. With such a mechanism, the source could wait until at least one listener is on the network before transmitting. At the time of writing, no such mechanism has been implemented or deployed.

2.4 Detecting Multicast Sources

Routers recognize multicast packets sent by directly connected sources by examining the source and destination addresses of the packet as well as the TTL. The addresses are examined for the following conditions:

- Destination address is in the class D range and is not of link-local scope.
- Source address is part of a directly connected subnet.
- TTL is greater than 1.

Under these conditions, the router knows it is the first-hop router, or designated router (DR), for the multicast source and acts appropriately as described in the following discussion on PIM-SM operation.

2.5 Routing Multicast Traffic within a Domain Using PIM-SM

The first step is to get multicast routing up and running within a single domain. PIM–SM is the most commonly used multicast routing protocol for this task. We describe a domain in the context of PIM-SM. A PIM-SM domain is group of PIM-SM speakers interconnected with physical links and/or tunnels that agree on the same RP-to-group mapping matrix for all or a subset of the 224/4 address range.

Figure 2-1 shows two separate domains. PIM-SM domains are commonly mapped to BGP ASs, which are collections of routers controlled by the same administrative entity. However, PIM-SM domains and BGP ASs are mutually exclusive, so this need not be the case.

Multicast routing centers on the building of distribution trees. Unlike unicast routing, each router may have multiple interfaces out of which it forwards packets on behalf of a particular multicast group. Packets do not traverse the network in a straight-line path; instead, a multifingered distribution tree is rooted at one router with various branches heading toward each interested receiver.

At the core of the PIM-SM domain is a router that serves a very special role, known as the rendezvous point (RP). The RP serves as the "well-known meeting place" for multicast sources and interested listeners. PIM-SM supports the following three major phases to deliver multicast packets from a source to a receiver:

- Build the RPT that delivers packets from the RP to interested listeners
- Build the distribution tree that delivers packets from the source to the RP
- Build the SPT that delivers packets directly from the source to the interested listeners

These phases can occur for each source-receiver pair, and the distribution trees for different sources and groups may have reached different phases at any given time, depending on the existence of a source and interested listeners for that group.

The order of operation is not strict. Multicast sources can be created before receivers are created, and PIM-SM still enables delivery of multicast packets to any newly appearing receivers. Likewise, for a given multicast group, PIM-SM handles the condition in which a listener emerges after the SPT has already been built from the source to other receivers.

At this point, it is best to walk through the simplest example to explain the operation of the protocol, using a single source and a single receiver.

2.5.1 Phase 1: Building the RPT That Delivers Packets from the RP to Interested Listeners

The traffic flow of the RPT starts at the RP and flows to all hosts interested in receiving on the multicast group. The RPT is constructed in the opposite direction of traffic flow, starting at the receivers and building hop-by-hop toward the RP.

When a PIM-SM router receives an IGMP Host Membership Report from a host on a directly attached subnet, the router is responsible for initiating the creation of a branch of the RPT for the specific group of interest. This last-hop router is known as the receiver's DR.

The DR sends a (*,G) PIM Join[1] message to its RPF neighbor for the RP's IP address. An RPF neighbor is defined as the next-hop address in the RPF table for a specific IP address. The RPF table is the routing table used by multicast routing protocols when performing RPF checks. In PIM-SM, it is possible for the RPF table to be the same routing table used for unicast packet forwarding.

2.6 Routing Multicast Traffic across Multiple Domains with MSDP

The PIM-SM protocol in itself does not have a mechanism to enable multicast packets from a source in one PIM-SM domain to reach a receiver in another domain. The PIM-SM DR for the source sends its Register messages to the RP in its own domain, while the PIM-SM DR for the receiver sends Join messages toward the RP in its own domain.

The delivery of multicast packets from the source to the RP in the source's domain is disconnected from the RPT in the receiver's domain. Thus the following conditions are required to transit multicast traffic across multiple PIM-SM domains:

- The RP in domains that have receivers must have knowledge of the IP address of active sources.
- All routers along the path from the source to the receivers must have a route to the source's IP address in their RPF table.[2]

[2] It is possible for a receiver's DR to always remain on the RPT and never join the SPT. In this case, routers on the path between the receiver and the RP (in the receiver's domain) need to have only a route to their RP in their RPF table. This situation is not very common, though, because most DRs join the SPT immediately.

The first requirement is accomplished by using MSDP. MSDP provides a way to connect multiple PIM-SM domains so that RPs can exchange information on the active sources of which they are aware. Each domain relies on its own RP instead of having to share an RP with another domain.

MSDP sessions use TCP for reliable transport and can be multihop. MSDP sessions are formed between the RPs of various domains. MSDP-speaking RPs send MSDP Source-Active (SA) messages to notify the RPs in other domains of active sources. An RP constructs an SA message each time it receives a PIM Register message from a DR advertising a new source. SA messages include the multicast data packet encapsulated in the Register message in current implementations of MSDP.

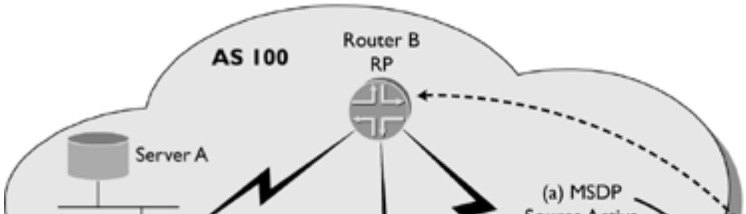
When an RP receives an SA message for a group for which interested receivers exist, the RP delivers the encapsulated data down the RPT to all the receivers in its domain. When the receiver's DR receives the multicast packets down the RPT, it joins the SPT directly to the source.

The second requirement usually is not a concern because most networks have any-to-any connectivity for unicast traffic, even for addresses in other ASs. Keep in mind the multicast RPF table need not be the same routing table used for unicast routing. In this case, the dedicated multicast RPF table must have routes for all potential multicast sources. MBGP is used to populate such an RPF table and is discussed in the next section.

2.6.1 MSDP in the Example Network

To show the functionality of MSDP in the example network, we continue with reference to Figure 2-5 . The data packets are being delivered from server A to host A via the SPT. Now host B in AS 200 reports its interest in the 230.1.1.1 group by sending an IGMP Report message to router G.

Figure 2-5. MSDP



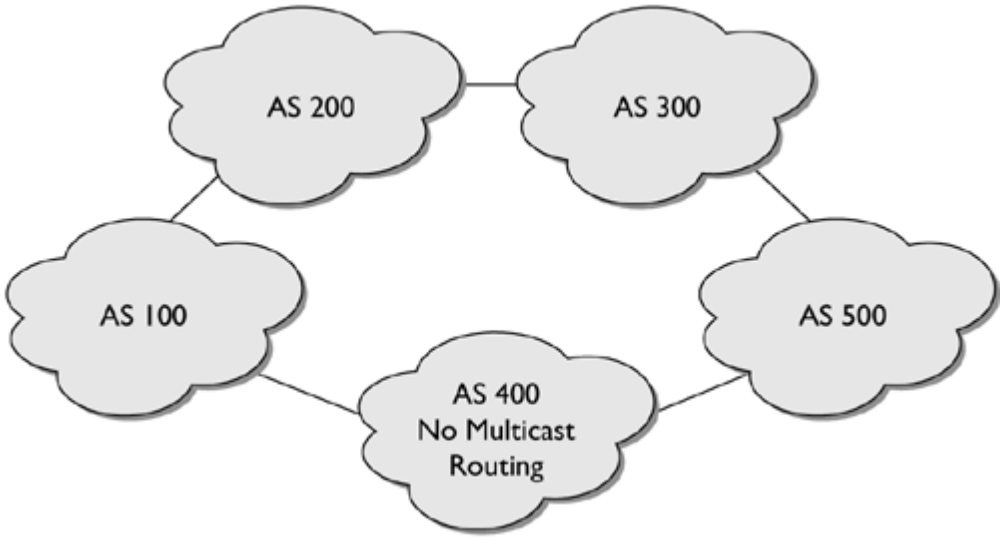
2.7 Populating a Routing Table Dedicated to RPF Checks with MBGP

The previous sections describe how the RPF mechanism uses information learned from a unicast routing table to determine the path of a multicast distribution tree. In PIM-SM it is possible for the RPF table to be populated from the same routing table used for unicast forwarding.

By taking this approach, unicast and multicast traffic follow the same path but in opposing directions. For example, a multicast packet traveling from server A to host B would traverse all the same routers and links, but in the exact opposite order, as a unicast packet traveling from host B to server A.

Some situations make such congruent routing of unicast and multicast traffic less than optimal. Figure 2-6 illustrates when it is beneficial for multicast and unicast traffic to travel separate paths.

Figure 2-6. Unicast and multicast paths



Based on fewest AS hops, the optimal path for unicast traffic traveling from AS 100 to AS 500 is through AS 400. However, AS 400 does not support multicast routing. If the same routing table used to forward unicast traffic is used for the RPF table in all routers and multicast traffic must flow from AS 500 to AS 100, AS 100 is compelled to use a suboptimal path for its unicast traffic destined for AS 500. Unicast traffic from AS 100 destined for AS 500 would be forced to traverse the path across AS 200 and AS 300.

To circumvent this limitation, a table other than the one used for unicast forwarding can be used for multicast RPF. The question is how to populate such a table: How are unicast routes introduced into a separate RPF table, with next-hop information different from the table used for unicast forwarding?

One solution is to configure static routes specifically for the RPF table. Note that static routing for multicast RPF faces the same scalability limitations as static routing for unicast forwarding. That is, static routes lack dynamic failover and can be administratively burdensome because changes to topology are not automatically updated.

In real networks, it is desirable to dynamically update the entries in the RPF table. The RPF table consists of unicast routes so there is no need to invent a new routing protocol. Instead, the need is to somehow differentiate between route-control information intended for unicast forwarding and the multicast RPF table. Theoretically, this differentiation could be implemented by modifying any of the existing unicast routing protocols. However, the structure of some unicast routing protocols makes them inherently more extensible, such that adding support for multicast requires relatively few modifications to the protocol. BGP is one of the best candidates for adding such functionality.

BGP is a dynamic routing protocol that can differentiate between multiple types of routing information. This capability is designated Multiprotocol Extensions for BGP (MBGP) and is defined in RFC 2858. MBGP works identically to BGP in all respects; it simply adds functionality to BGP, such as the capability for BGP updates to tag routing

Chapter 3. Multicast Routing Protocols

This chapter provides an overview of multicast routing protocols. For a protocol to be deemed a multicast routing protocol, it must at minimum provide the functionality of setting up multicast forwarding state and exchange information about this forwarding state with other multicast routers. By this definition PIM-DM, PIM-SM, and DVMRP are all multicast routing protocols; IGMP, MBGP, and MSDP are not.

Multicast routing protocols can generally be classified into two categories: dense and sparse. By understanding the advantages and disadvantages of each, it is clear to see why PIM-SM has become the protocol of choice for IMR. We describe the characteristics of dense and sparse protocols and briefly discuss examples of each. PIM-SM is examined in detail in Chapter 4.

3.1 Dense Protocols

Dense protocols assume a dense distribution of receivers exists throughout the domain, which means each subnet likely has at least one interested receiver for every active group. This assumption may be valid on enterprise networks where only a few groups are active and most of the subnets contain interested listeners.

On networks where few or no prunes occur, dense protocols are actually more efficient than sparse protocols. However, on the Internet, where prunes are more prevalent, dense protocols are not well suited to interdomain deployment.

Dense protocols follow a flood-and-prune model. To inform the routers of multicast sources, this traffic is initially broadcast throughout the domain. Upon first receiving traffic to a dense group on its interface closest to the source, a router forwards this traffic out all of its interfaces except the interface on which it received the data. Thus the IIF initially is the RPF interface toward the source, and the OIL contains all other interfaces.

If traffic is received on the interface that is not the RPF interface toward the source, the traffic is discarded, and a Prune message is sent upstream. If a router has no interested receivers for the data (that is, its OIL becomes empty), it sends a Prune upstream. Periodic reflooding is used to refresh state.

The primary benefit of dense protocols is simplicity. The flood-and-prune mechanism enables these protocols to easily build a multicast distribution tree rooted at the source. A source-based tree guarantees the shortest and most efficient path from source to receiver. The obvious limitation is scalability; any mechanism that relies on flooding across the entire network does not scale particularly well on the Internet.

3.1.1 DVMRP

DVMRP is the multicast routing protocol first used to support the MBone. It performs the standard flood-and-prune behavior common to dense protocols. It also implements a separate routing protocol used to build the routing tables on which RPF checks are performed.

As its name suggests, DVMRP has a distance-vector routing protocol very similar to RIP. It has the same limitations found in other distance-vector protocols, which include slow convergence and limited metric (that is, hop count). Although most DVMRP deployments have been replaced by PIM-SM, DVMRP can still be found on networks with legacy equipment, such as dialup RASs. Most RASs made today support either PIM-SM or IGMP proxying.

3.1.2 PIM-DM

PIM-DM implements the same flood-and-prune mechanism mentioned previously and is quite similar to DVMRP. The primary difference between DVMRP and PIM-DM is that the latter aptly named protocol introduces the concept of protocol independence. PIM, in both dense and sparse modes of operation, can use the routing table populated by any underlying unicast routing protocol to perform RPF checks.

This ability to use any underlying unicast protocol was seen by ISPs to be a significant enhancement because they did not want to manage a separate routing protocol just for RPF (ironically, MBGP and M-ISIS were later used to do just that). PIM-DM can consult the unicast routing table, populated by OSPF, IS-IS, BGP, and so on, or it can be configured to use a multicast RPF table populated by MBGP or M-ISIS when performing RPF checks.

3.2 Sparse Protocols

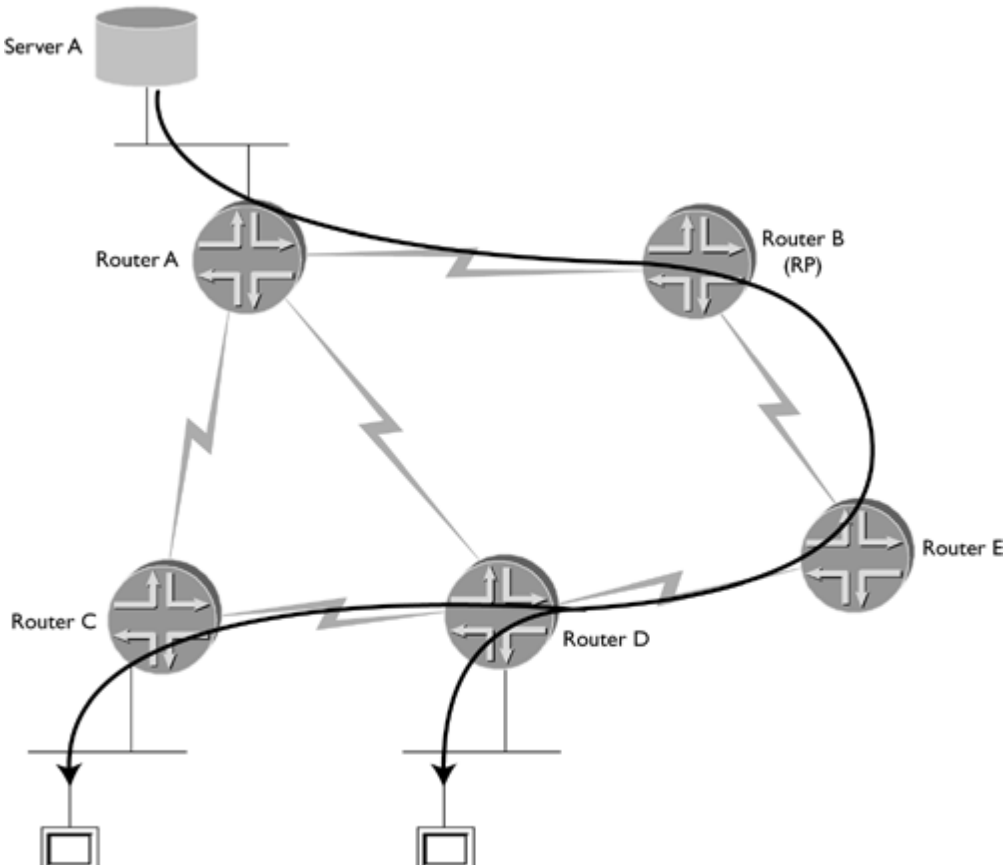
Sparse protocols make the implicit assumption that a sparse distribution of subnets with at least one interested receiver for each active group exists, which is much more consistent with what is found on the Internet. The primary difference between sparse and dense protocols is in the way the protocols handle source discovery. Where dense protocols use flooding of the actual data to inform the routers in the domain of active sources, sparse protocols designate a core node to keep track of all of the active sources in a domain. The mechanisms involved in source discovery in sparse protocols, while much more complex than those in dense, provide the scalability needed to support multicast across the Internet.

Sparse protocols follow an explicit join model. In this model, multicast data is forwarded only to routers that explicitly request it. In sparse protocols, the root of the distribution tree is at a core node. This core node, or rendezvous point (RP) as it is known in PIM-SM, can receive traffic from the source via the SPT. When a host wants to join a group, its directly connected router joins the distribution tree toward the RP. So traffic is received by the RP along the SPT and forwarded to interested receivers across the domain via the shared tree, or rendezvous point (RPT).

The benefit of the RPT is that it reduces the amount of state required in the non-RP routers and does not require flooding across the network to inform routers of active sources. Instead, the RP is the only router that needs to be aware of all of the active sources for a domain. All other routers simply need to know who the RP is and how to reach it.

One disadvantage of the RPT is that it introduces the potential for suboptimal routing (see Figure 3-1). Multicast data must first flow to the RP and then to the receivers, even if the receivers are much closer to the source. The potentially inefficient path of an RPT illustrates how the SPT always provides the most efficient path (see Figure 3-2). To eliminate this inefficiency, PIM-SM enables the receiver's DR to join toward the source along the SPT if traffic reaches a certain threshold. Juniper Networks and Cisco System routers implement a threshold of 0 by default (this threshold is configurable on Cisco routers), which means once the DR receives the first multicast packet and learns the source, it sends an (S,G) Join toward the source. When it starts receiving data from the SPT, it then sends a Prune for traffic received via the RPT toward the RP. Accordingly, PIM-SM exhibits the best of both worlds by enjoying the benefits of both the RPT and the SPT without suffering their limitations.

Figure 3-1. The RP tree with suboptimal routing



3.3 Sparse-Dense Mode

Sparse-dense mode is a mode of PIM implemented by both Juniper Networks and Cisco Systems that supports both types of operation concurrently. Sparse-dense mode enables the interface to operate in either sparse mode or dense mode on a per-group basis. Groups specified as dense groups are not mapped to an RP, and data packets destined for those groups are forwarded based on the rules of PIM-DM. Sparse groups are mapped to an RP and are forwarded based on the rules of PIM-SM.

Initially, sparse-dense mode was ideal for indecisive enterprise network designers who were not sure if they should deploy PIM-SM or PIM-DM. Today it is mainly used in networks that implement auto-RP for PIM-SM.

Chapter 4. Protocol Independent Multicast-Sparse Mode (PIM-SM)

This chapter provides a detailed description of PIM-SM, the predominant multicast routing protocol for interdomain routing.

4.1 Specifications

The version numbers (version 1 and version 2) mostly pertain to packet format, and each version is described in several iterations of specifications. For example, version 2, sparse mode, was described in RFC 2117, which then was made obsolete by RFC 2362. The most recent specification of the PIM-SM protocol is draft-ietf-pim-sm-v2-new-04.txt. No current RFC describes PIM version 1; the Internet-Drafts that originally defined version 1 have expired.

4.2 PIM Versions

The implementation of PIM by Juniper Networks and Cisco Systems enables configuration of a distinct PIM version on each interface of a router. This setting identifies the format of PIM messages sent out the interface.

4.2.1 Version 1

PIM version 1 messages are sent as an IGMP message (IP protocol number 2) with IGMP version set to 1 and IGMP type set to 4 (4 = router PIM messages).

The type of PIM message is distinguished by the IGMP Code field. The following are all the PIM version 1 message types:

0: Router-Query

1: Register (used in PIM-SM only)

2: Register-Stop (used in PIM-SM only)

3: Join/Prune

4: RP-Reachability (not used)

5: Assert

6: Graft (used in PIM-DM only)

7: Graft-Ack (used in PIM-DM only)

4.2.2 Version 2

PIM version 2 messages use IP protocol number 103. The first four bits of a version 2 message represent the version number (2). The next four bits represent the type of message. The following is a list of all PIM version 2 message types:

0: Hello

1: Register (used in PIM-SM only)

2: Register-Stop (used in PIM-SM only)

3: JoinJoin/Prune

4.3 Group-to-RP Mapping

For PIM-SM to work properly, all routers in a domain must know and agree on the active RP for each multicast group. In fact, the definition of PIM-SM domain is just that: a group of PIM-SM speakers interconnected with physical links and/or tunnels that agree on the same RP-to-group mapping matrix for all or a subset of the 224/4 address range. There are three ways to map the RP:

- Static group-to-RP mapping
- Cisco Systems auto-RP (dynamic)
- PIM bootstrap router (BSR) (dynamic)

Typically, only one of these methods is used for setting the RP. The following sections describe each of these mechanisms.

4.3.1 Static Group-to-RP Mapping

Static RP is by far the least elaborate method. Every router in the PIM domain must be manually configured with the address of the RP for each multicast group. The major advantage to using this method is simplicity. The drawback is that it requires configuration on every router in the domain each time the address of the RP changes. Also, failover to a backup RP requires additional configuration in the event the primary RP is unreachable. Anycast RP alleviates this limitation.

4.3.2 Dynamic Group-to-RP Mapping: Cisco Systems Auto-RP

Auto-RP, originally a Cisco Systems proprietary mechanism for dynamic group-to-RP mapping, is fully supported by Juniper Networks routers as well. Auto-RP relies on dense mode of operation to forward control messages to two well-known group addresses (224.0.1.39 and 224.0.1.40). Because of this reliance, all routers in an auto-RP-enabled PIM-SM domain should be configured in sparse-dense mode. With auto-RP, RPs in a domain announce themselves as such with these groups. All other routers in the domain join one or both of these dense groups and learn dynamically the address of the RPs.

Each router in a domain using auto-RP fits into one of the following roles:

- Candidate RP
- Mapping agent
- Discovery-only

Every 60 seconds, a candidate RP sends an RP-Announcement message detailing the group ranges for which it intends to serve as RP. This message is sent to 224.0.1.39 (CISCO-RP-ANNOUNCE).

The routers configured as mapping agents join the 224.0.1.39 group and listen for RP-Announcement messages. Each mapping agent uses the following criteria to determine which RPs to announce as the active RP for each group:

- When multiple RPs announce the same group prefix and mask, accept the announcement only from the RP with the highest IP address

4.4 Anycast RP

In PIM-SM, only one RP can be active for any single multicast group. This limitation provides a great challenge when trying to deliver load balancing and redundancy. Anycast RP is a clever mechanism that circumvents this limitation. Anycast means that multiple hosts, or in this case routers, share the same unicast IP address. This address is then advertised by a routing protocol, such as OSPF, (M-)ISIS, or (M)BGP. Packets destined for the anycast address are then delivered to the closest host with this address. If that host becomes unreachable, packets are delivered to the next closest host with the anycast address.

With anycast RP, multiple routers are configured with the same IP address, typically on their loopback interface. This shared address is used in the RP-to-group mapping, which allows multicast groups to have multiple active RPs in a PIM-SM domain. PIM-SM control messages are sent toward the shared address, and they will reach an RP with the best routing metric from the originator of the message. Register messages and (*,G) Joins are sent to the topologically closest RP.

Thus anycast RP essentially forms multiple PIM-SM subdomains within the domain, with each subdomain consisting of one of the RPs and all of the PIM-SM routers with the best routing metric for the shared address pointing toward that RP. Because the domain is broken into subdomains, it is necessary to run MSDP between the RPs to exchange information about active sources between subdomains.

The anycast RP address is typically configured as a secondary address on the loopback interface. Care should be taken to ensure that routing protocols such as OSPF, IS-IS, or BGP do not select the anycast address as the router ID. Duplicate router IDs in these protocols can cause disastrous results. For this reason, it is wise to configure a unique unicast address as the primary loopback address. This unique address is used as the router ID for routing protocols as well as the peering address for MSDP sessions.

Anycast RP enables RP tasks for a PIM-SM domain to be shared across multiple routers by localizing their responsibility to their respective subdomains. This localization provides very intelligent load balancing from a routing perspective. Anycast RP also provides redundancy around a failed RP that is as fast as the convergence of the routing protocol carrying the anycast address. If one of the anycast RPs becomes unavailable, all PIM-SM control messages that were originally destined for the failed RP are delivered to the RP with the next best routing metric. Forthcoming PIM Register and (*,G) messages will be sent to the next closest RP, and RPTs will be rooted at the next closest RP.

Anycast RP is mutually exclusive with the group-to-RP mapping mechanism, so it can be used in conjunction with static RP, auto-RP, or BSR. While auto-RP and BSR have their own methods of delivering load balancing and redundancy, most ISPs have found anycast RP provides these benefits in a much simpler and more intuitive way.

Unless the IP address of the RP changes frequently, BSR and auto-RP provide little benefit over a statically defined anycast RP. Furthermore, these dynamic mapping mechanisms introduce a great deal of complexity in a realm already replete with confusion. For example, auto-RP requires a sparse mode protocol to use a dense mode control plane. When troubleshooting, both of these topologies must be examined.

While simplicity is always a desired goal in network design, it is even more valuable when building and operating multicast networks. For this reason, it is highly recommended that anycast RP with static group-to-RP mapping be used when deploying interdomain multicast. Interestingly, most ISP engineers strongly prefer this method, while protocol designers usually insist on BSR. These differing preferences are probably due to the same reason, reflecting the opposing biases these groups frequently hold. Static anycast RP is unsophisticatedly simple; BSR is elegantly complex.

4.5 PIM Register Message Processing

When a PIM-SM DR receives a multicast packet sourced by a directly connected host, the DR encapsulates the packet in a Register message and sends it as a unicast packet to the RP for the group. The Register message conveys the source address, S, and group address, G. Upon receiving the Register message, the response of the RP is based on two factors:

- Whether it has an RPT set up for the group (that is, does it know of any receivers interested in the group?)
- Whether it is receiving data natively for this (S,G) pair down a distribution tree

The RP ignores the Register message and immediately sends a Register-Stop message to the DR if either of the following conditions is met:

- No RPT is set up.
- An RPT exists, but the RP is already receiving data natively from the source.

If these two conditions are not met, the RPT is set up, and the RP is not receiving packets natively yet. According to the PIM-SM specifications, the RP can decapsulate the register packets and forward them natively down the RPT. Or, optionally, the RP can join the SPT and receive packets natively from the source. For low data rate sources, not joining the SPT and decapsulating register packets may be desirable because it reduces the amount of state created. However, this strategy can lead to high join latency because the RP must wait for the DR to send register packets. Additionally, decapsulation can be a resource-intensive process for a router. Accordingly, Juniper Networks and Cisco System RPs implement this option and always join the SPT.

In this case, the RP joins the SPT by sending a (S,G) Join to its RPF neighbor for the source and waits until it receives packets natively before it sends the Register-Stop message. Meanwhile, it extracts the data packet from every PIM Register message it receives for the (S,G) pair and forwards it down the RPT natively.

Upon receiving the Register-Stop message, the DR stops sending Register messages and starts a Register-Stop timer for the (S,G) pair. The DR periodically sends a Null-Register to the RP. The Null-Register is a Register message with no encapsulated data and with the Null-Register bit set. The Null-Register message is used to probe the RP to determine whether the DR needs to start sending normal Register messages to the RP for the (S,G) pair.

The RP handles receipt of a Null-Register the same way as a normal Register message. It decides whether to send a Register-Stop back to the DR based on the rules described previously. If the RP sends a Register-Stop, the Register-Stop timer on the DR is reset before it expires, and the DR does not start encapsulating data again.

If the Register-Stop is not sent and the DR's Register-Stop timer expires, the DR starts sending normal Register messages with encapsulated data to the RP until it receives another Register-Stop. The purpose of the Null-Register is to avoid having the DR encapsulate data that is not needed by the RP.

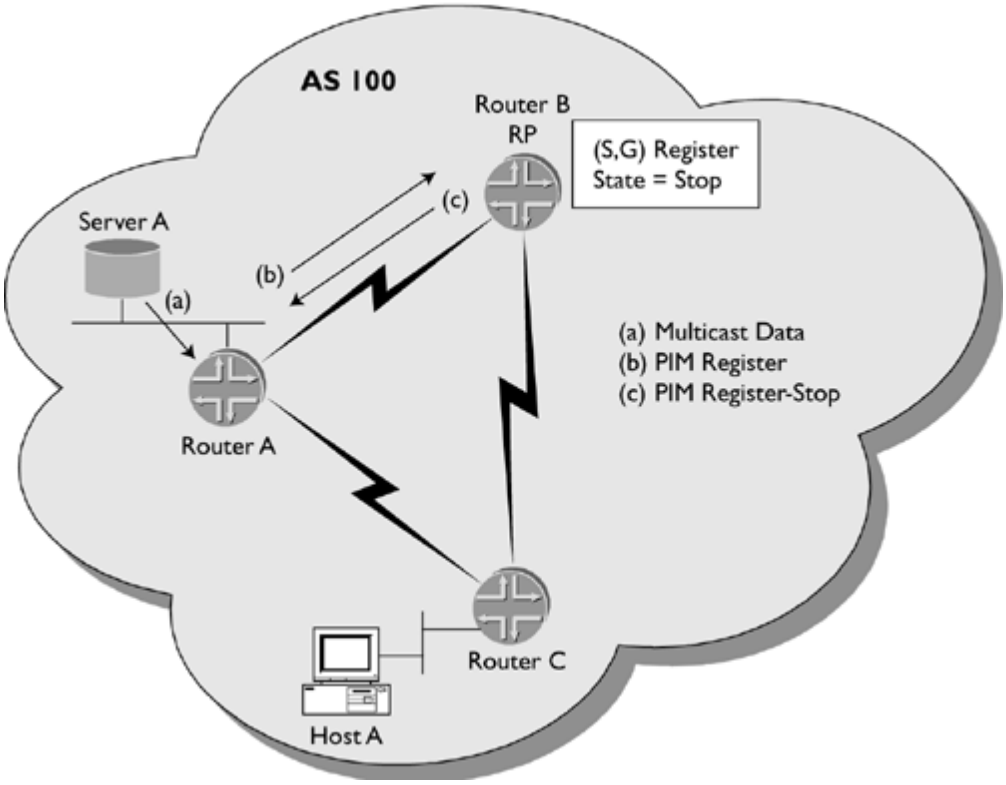
4.6 Distribution Tree Construction and Teardown

The model presented in phases 1–3 in Chapter 2 looked at the creation of PIM distribution trees in its simplest form. This section, although it does not cover all possibilities, describes how the protocol reacts in other common scenarios. PIM-SM is a very complex protocol, and covering every possible scenario could fill an entire (very boring) book. The important thing to notice in the four scenarios presented here is that PIM-SM reacts in the same way. This section should reinforce the reader's trust that PIM-SM does behave in the expected manner in the various common situations. If it seems repetitive at times, then that goal is accomplished.

4.6.1 Scenario 1: Source Comes Online First, Then a Receiver Joins

This scenario, shown in Figure 4-1 , essentially flips the order of phases 1 and 2 from Chapter 2 (see sections 2.5.1 and 2.5.2). When the source begins to send traffic to a group, G, the DR for its subnet encapsulates the multicast packets in Register messages and sends them to the RP. No receivers in the domain have joined the group, so the RP's OIL for (*,G) is empty. The data packets encapsulated in the Register messages are discarded by the RP, and the RP sends a Register-Stop to the DR.

Figure 4-1. Source comes online first.



After receiving a Register-Stop message from the RP, the DR stops sending Register messages for the group and initializes its Register-Stop timer. The DR periodically sends a Null-Register message to the RP. If still no receivers have expressed interest in the group, the RP responds with a Register-Stop. The DR reinitializes its Register-Stop timer.

In Figure 4-2 , we see that when a receiver wants to join the group, its DR sends a (*,G) Join toward the RP to build an RPT. Upon receiving a (*,G) Join, the RP adds the interface on which it is received to the OIL for (*,G). The RP then joins the SPT by sending an (S,G) Join toward the source. The RP is now receiving traffic natively via the SPT and distributing it down the RPT.

Figure 4-2. Receiver joins existing source. (The dashed line represents data flow after steps d, e, and f are completed.)

4.7 Designated Routers and Hello Messages

PIM Hello messages are sent periodically on each interface that has PIM enabled. The primary purpose of Hello messages is to announce each router's existence on the subnet as a PIM router, so all routers can decide on a single DR for the subnet.

PIM Hello messages are sent on both multiaccess and point-to-point interfaces. Hello messages are sent to the multicast address 224.0.0.13 (ALL-PIM-ROUTERS group) with a TTL of 1. When a router first boots or is first configured for PIM, it sends out the initial Hello message and then sets its Hello timer to 30 seconds (the default).

Each time the Hello timer counts down to 0, Hello messages are sent out, and the timer is reset. If a router does not hear from a neighbor for a period of 3.5 times the Hello timer (105 seconds is the default), the neighbor is dropped (possibly causing the election of a new DR).

Note

This hold-time value is actually carried in the Hello message, so routers on the same subnet can have different hold timers and not experience problems with incorrectly dropping neighbors.

The Hello messages contain the configured DR priority of the router sending the message. The router with the highest DR priority is elected DR for the subnet. If any of the routers does not support the DR priority option, the DR is the router with the highest IP address.

Note

Each router on the subnet elects the DR, and the election results are never communicated to the other routers on the subnet. This is not a problem as long as each router has the same information and uses the same algorithm to determine the DR.

If the IP address of an interface is changed, the router first sends a Hello message from the old address with a hold-time of 0, which forces the other routers on the subnet to immediately purge the old address from their neighbor tables. Then a Hello message with the new address and standard hold-time is sent. Each time the neighbor table is changed, each router runs through the DR election algorithm again.

If a router loses its DR status, it no longer sends Register messages for new sources to the RP. It also stops sending (null) Registers for current sources on the subnet.

4.8 PIM Assert Messages

PIM Assert messages are needed for multiaccess networks that serve as a transit for multicast traffic. Ordinarily, multiaccess networks (in the form of LANs such as Ethernet) serve as end-points of distribution trees, housing multiple hosts serving as either receivers or sources; typically only one or two routers provide access to the rest of the Internet for the hosts.

It is not uncommon, though, for a multiaccess network to connect multiple routers and no hosts. These transit LANs introduce a number of complications for setting up multicast distribution trees as compared to point-to-point links.

The PIM assert mechanism accomplishes the following tasks:

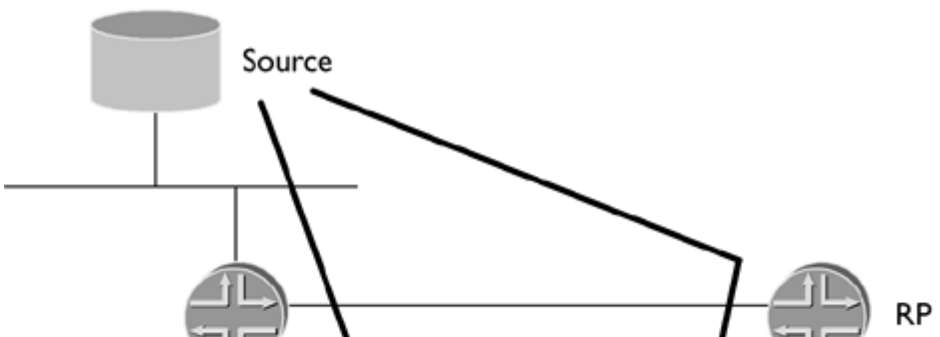
1. Recognizes when multiple routers are forwarding duplicate multicast data packets to a transit LAN
2. Holds an election to choose a single forwarder for the LAN
3. Advertises the winner of the election to all routers on the LAN
4. Overrides the RPF rules; PIM Join/Prune messages for the group are sent to the assert election winner instead of the RPF neighbor

If the assert mechanism were not available, problems could arise that involve data packets being forwarded to the LAN by two different routers. The following three situations lead to duplicate packets being forwarded to a transit LAN:

1. One router may send a (*,G) Join to its RPF neighbor for the RP, while another router on the same LAN sends an (S,G) Join to its RPF neighbor for the source. If the RPF neighbor for the RP and the source are different routers, redundant traffic will be delivered to the LAN.
2. Two routers on a LAN can have different RPF neighbors for the RP (which should be the case only when the routing policy is designed poorly). If both of these routers send a (*,G) Join to their respective RPF neighbor for the RP, the same multicast traffic is delivered to the LAN twice.
3. Problem 2 can also occur if two routers on the LAN have different RPF neighbors for the source and send (S,G) Joins. This situation can be common on MIXs.

Figure 4-4 illustrates a network topology that could easily lead to the occurrence of problem 1 in the preceding list.

Figure 4-4. Situation that requires the PIM assert mechanism



4.9 Multicast Scoping

Multicast scoping enables a network operator to configure interfaces not to receive or transmit packets for specific multicast groups. These routers are boundaries for the groups specified in the configuration. RFC 2365, "Administratively Scoped IP Multicast," is the specification for this functionality.

When a router is configured to scope group G on an interface, the router does not forward packets destined for G out the interface nor does it accept packets destined for G received on the interface. This is to say that multicast scoping is bidirectional. The router does not accept any Join messages received on the boundary interface for group G. If group G is a dense group, the router prunes the boundary interface from the OIL for group G.

Multicast scoping enables selected groups to remain within the domain without fear of the data being leaked outside the domain. Receivers within the domain will not receive data from external sources.

A router can be configured for multicast scoping on any range of group addresses on any of its interfaces. All routers on the boundary of a domain with one or more connections to routers in other PIM domains should share the same scoping configuration on their boundary interfaces. If auto-RP is enabled in a PIM-SM domain, multicast scoping should be used to block all packets destined for 224.0.1.39 and 224.0.1.40 from entering or leaving the domain. This prevents the accidental leaking of control packets to other domains.

Prior to the availability of multicast scoping, the only way to achieve the same effect was to use the TTL field in the IP header. This sloppy solution was hard to maintain because the network diameter is different from each router's perspective. The intended purpose of TTL is to eventually discard packets caught in a routing loop. Using TTL for any other reason is not good practice because it may not always produce desired results and can waste bandwidth.

RFC 2365 defines addresses within the 239/8 address range as administratively scoped. Packets destined for these addresses should not be forwarded beyond an administratively defined boundary, which is somewhat analogous to a private unicast address space, such as 10/8. Further subranges within 239/8 are defined with additional scoping classification. The 224.0.0/24 address range has link-local scope. Packets destined for these addresses should never be forwarded outside a LAN by a router.

Chapter 5. Multicast Source Discovery Protocol (MSDP)

This chapter describes MSDP in depth. MSDP establishes a mechanism to connect xmultiple PIM-SM domains. With MSDP, each PIM-SM domain has its own RP and does not rely on the RP of another organization's network.

5.1 Introduction

In PIM-SM, the RP is configured to serve a range of multicast groups. The RP is responsible for knowing all of the active sources of all multicast groups in this range. There can be only one active RP for a given group. This requirement of PIM-SM presents interesting challenges when trying to support redundancy, load balancing, and interdomain connectivity. MSDP was developed to address these challenges.

Before MSDP, one technique for achieving IMR was to connect each ISP's RP on a multiaccess interface at a multicast peering exchange. This multiaccess interface was configured for PIM-DM so each RP could flood its source information to all other RPs on this interconnecting LAN.

The limitations of this hybrid PIM-SM/PIM-DM approach are obvious. First, an RP is forced to sit at the edge of the domain. Ideally, the RP is placed in a well-connected part of the core of a network to minimize suboptimal routing on the shared tree. Second, only one RP can be in each domain, and it must be located at a single interconnect point for all multicast domains in the Internet. This single interconnect limits the redundancy and scalability of each domain individually and collectively. Imagine what would happen if this LAN failed!

Another approach that might have been considered would create a centralized RP shared by all ISPs. Aside from the scalability issues of this idea, ISPs would not like relying on a third party for RP service. The concept of owning and managing their own RPs is important to network operators.

MSDP introduced the ability for RPs to connect to one another and to exchange information about the active sources in their respective PIM-SM domains. With this capability, each domain can have one or more RPs, enabling support for redundancy, load balancing, and interdomain connectivity.

At the time of writing, MSDP is defined in an IETF Internet-Draft (draft-ietf-msdp-spec-13.txt). The evolution of this protocol has been interesting, to say the least. The implementations of Juniper Networks and Cisco Systems are based on version 2 of the original draft. Certain items in later versions of the draft have been added to these implementations along with undocumented optimizations based on deployment experience. Other implementations have been reported to be based on later versions of the draft. By supporting various components of various versions of the specification, no implementation operates exactly the same. Despite all this, these implementations generally interoperate with one another.

The discussion in this chapter is based on the current specification but points out any major differences between the current specification and the predominant implementations. At the end of this chapter, we discuss some of the limitations of MSDP and what the future could hold for this protocol.

5.1.1 MSDP Operation

MSDP-speaking routers form peer relationships, similar to BGP peers, over a TCP connection. Two MSDP peers can be in the same PIM-SM domain or in two separate domains. Within a domain, MSDP enables creation of multiple RPs, facilitating redundancy and load balancing. Anycast RP is the primary example of intradomain MSDP. Between different domains, MSDP enables RPs to exchange source information from their respective domains, allowing interdomain source discovery to occur.

An RP that wants to participate in IMR must speak MSDP. However, an MSDP speaker does not necessarily have to be an RP. Non-RP routers can be configured for MSDP, which may be useful in a domain that does not support multicast sources but does support multicast transit. A non-RP MSDP speaker does not originate any source information but provides transit for source information from other domains.

When an MSDP-speaking RP receives a PIM Register message, it generates an MSDP Source-Active message for the source-group pair and forwards the message to all of its configured MSDP peers. The SA message contains the source address, the group address, and the address of the RP. Additionally, the encapsulated data in the Register message is copied by the RP into the MSDP SA.

Subsequent Register messages for the same source-group pair do not cause the creation of other SA messages

5.2 MSDP Peering Sessions

The MSDP peer with the higher IP address listens for new connections on the well-known MSDP TCP port 639. The MSDP peer with the lower address repeatedly attempts to initiate a TCP session with its peer on port 639. This method prevents TCP session set-up collisions, which occur when both sides initiate a connection at approximately the same time and one session has to be dropped (as is the case with the setup of BGP peering sessions). The drawback is a possible longer set-up time for the passive (higher IP address) side of the connection.

The MSDP peer state machine has the following five possible states:

- DISABLED: MSDP peer is not configured.
- INACTIVE: MSDP peer is configured but not listening or connecting.
- CONNECT: Active peer attempts to initiate TCP session.
- LISTEN: Passive peer is configured and listening on TCP port 639.
- ESTABLISHED: TCP session is established.

The normal, successful state transition for the passive peer is as follows:

1. DISABLED
2. INACTIVE
3. LISTEN
4. ESTABLISHED

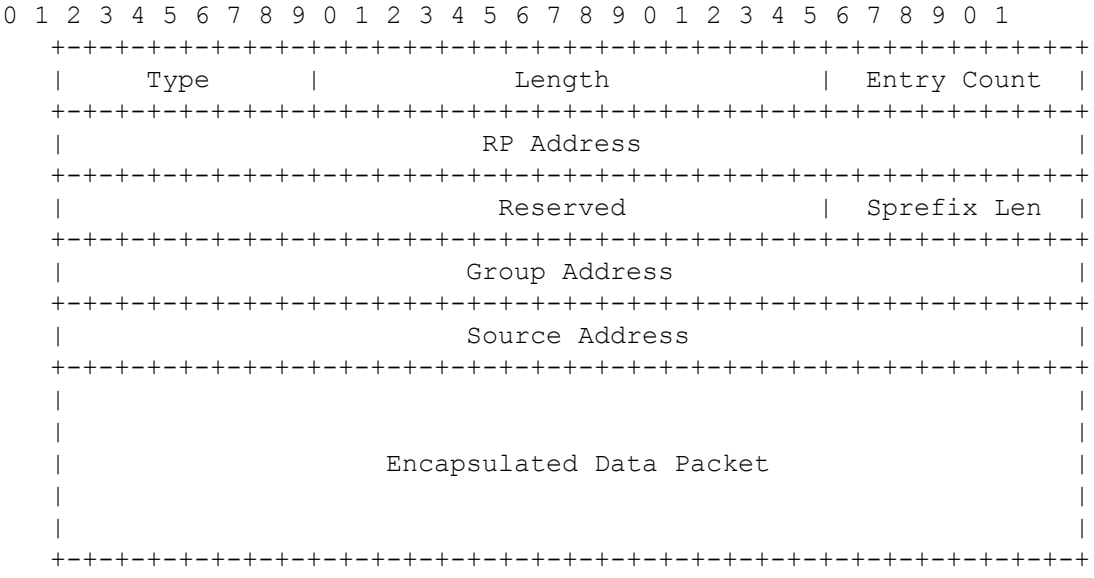
For the active peer (lower IP address), the normal, successful state transition is:

1. DISABLED
2. INACTIVE
3. CONNECT
4. ESTABLISHED

The active peer swaps between the INACTIVE and CONNECT states until the passive peer accepts the connection. Each time the active peer reverts from CONNECT to INACTIVE state, the active peer waits a default of 30 seconds before trying to connect again.

5.3 The MSDP SA Message

Like IS-IS, MSDP messages use structures known as type-length values (TLVs). The packet format of the MSDP SA message is as follows:



The first byte is the message Type code. The Length field contains the length of the MSDP SA message in octets. The length includes everything from the Type field to the end of the encapsulated data packet. Entry Count is the number of source-group pairs listed in the message. Each source-group pair is encoded with its own Reserved and Sprefix Len (source address prefix length) fields. Each source-group pair adds 12 octets to the length of the SA message.

The RP Address field indicates the address of the router that created the SA. When originating an SA message, the address selected for this field might not be the address used for the purposes of the PIM-SM RP. For example, in the case of anycast RP, the address placed in the RP Address field of the SA message should be the unique local address of the RP, not the anycast address. In some implementations, a router discards a received SA message that contains an RP address that matches one of its own addresses. To better illustrate, imagine the following scenario.

Routers A and B are anycast RPs. When a PIM-SM DR sends Register messages to router A, an SA message is created by router A and forwarded to its MSDP peers, including router B. If router A places the anycast address in the RP Address field of the SA, router B sees its own address in this received message. Believing this might be a looped or spoofed SA message, router B may discard this message.[1] If, on the other hand, router A places its own unique loopback address in the RP Address field of the SA, router B will clearly see that this message is originated by another router and will accept the message.

[1] Discarding an SA message for this reason is not the behavior of all implementations.

The Reserved field is all zeros and the Sprefix Len field is always 32 (0x20). The Group Address field encodes the group address. The Source Address field encodes the source address.

5.4 Determining the RPF Peer

The nature of MSDP is to flood SA messages to all peers except the peer from which the SA was received. Because of this behavior, it is possible for an MSDP-speaking router to receive SA messages containing duplicate information from one or more of its peers, which is normal operation of the protocol. However, if the router were to accept and flood all of these duplicate messages, it would cause unneeded traffic on the network. The problem would grow exponentially with complex meshes of MSDP peers. To avoid this problem, MSDP uses peer-RPF flooding to choose a peer from which to accept an SA message containing certain information.

The originating RP contained in the SA message is used to determine the RPF peer. All SA messages with the same originating RP have the same RPF peer. An SA message is accepted and forwarded to other peers only if it was received from the RPF peer; otherwise, it is ignored and silently discarded.

The rules for determining the RPF peer of a particular SA message have changed considerably throughout the revisions of the MSDP specification. As stated earlier, at the time of writing, the MSDP implementations of Juniper Networks and Cisco Systems follow version 2 of the MSDP specification (draft-ietf-msdp-spec-02.txt). This draft has now expired and is no longer available on the IETF's Web site. One easy way to find it is by searching by name for the draft at http://www.google.com.

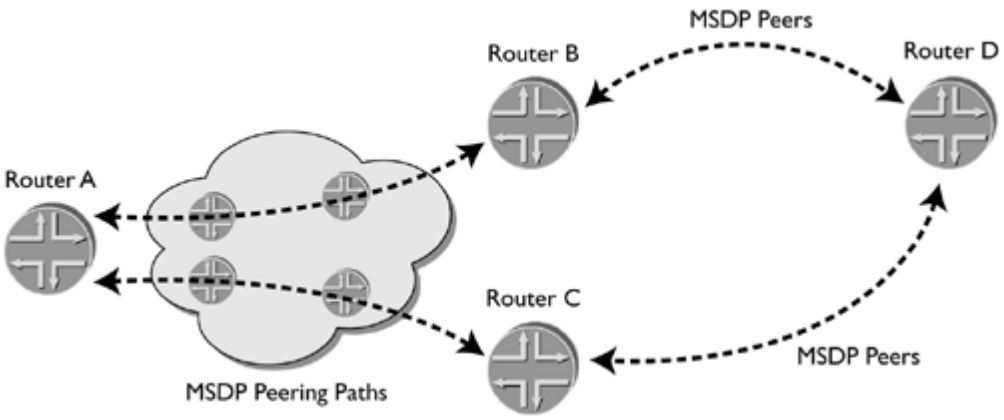
The core aspects of the RPF-peer rules have remained the same throughout the various revisions of the specification. When an SA is received, the rules are evaluated in order against all MSDP peers. The only information that is considered from the SA message is the originating RP address. The source and group addresses are not involved in determining the RPF peer. The peer that matches the earliest rule is declared the RPF peer for the originating RP. SA messages are accepted and forwarded only when received from the RPF peer. The specifics of the rules have changed quite a bit, but all of these core guidelines have stayed intact.

These changes cause implementations based on varying versions of the draft to act differently. It is important to know which version of the draft your vendor supports. This information can be found in the vendor's technical publications.

5.4.1 The Current Versions RPF-Peer Rules

Figure 5-1 represents a basic internetwork containing MSDP speakers. Figure 5-1 is used to discuss the RPF-peer rules. With each rule, we provide a subsequent figure, based on Figure 5-1, to illustrate the specific situation.

Figure 5-1. Network for explaining RPF-peer rules



All rules are examined from the perspective of router D. Router D has only two peers, which simplifies the explanation of the rules. The rules are applied to all MSDP peers that are in ESTABLISHED state at the time that the SA is received. Figure 5-1 is a generic diagram that does not include information about BGP peering sessions or other details on unicast routing that affect MSDP RPF-peer selection. These details are filled in as we discuss the rules.

The MSDP peering paths denoted in the figure are simply a chain of routers that provide MSDP connectivity from router A to router B and from router A to router C. In this discussion, we are interested only in router D's RPF-peer decision, but keep in mind that each router must make its own independent decision.

5.5 Mesh Groups

An MSDP mesh group can be configured for a group of MSDP peers that are completely meshed; that is, each router in the group has an MSDP peering session with every other router in the group. The idea behind mesh groups is borrowed from IS-IS. MSDP mesh groups are able to reduce SA flooding by identifying a group of MSDP peers that are fully meshed. Using the knowledge that certain peers are fully meshed, an MSDP speaker can modify the way it behaves upon receipt of an SA message.

If an SA message is accepted from a nonmesh group peer (per the RPF-peer rules), the message is sent to all mesh group peers. If an SA message is received from a mesh group peer, the message is sent to all nonmesh group peers. If a message is received from a mesh group peer, it is not forwarded back to any other peer in the mesh group. If these fully connected peers were not configured in a mesh group, the copies of the same SA message would be flooded between these peers. Each peer might receive the same SA message from every other peer if a mesh group weren't configured.

While mesh groups were originally created to reduce SA flooding, mesh groups are used today primarily because of a side effect of RPF-peer behavior. SA messages received from mesh group peers are always accepted and are not subject to RPF-peer rules. This relaxation of the RPF rules may be desirable within a domain, for example, where SA messages exchanged between anycast RPs should always be accepted.

Because MSDP RPF-peer rules are so complicated, poorly understood, and difficult to troubleshoot, many ISPs configure anycast RPs in a mesh group to circumvent the RPF rules among these peers. All other peers not in the mesh group are subject to the RPF rules. This method creates what has been referred to as internal MSDP (IMSDP) peers and external MSDP (EMSDP) peers. The relationship between IMSDP and EMSDP resembles the relationship between IBGP and EBGP.

Mesh groups are somewhat of a necessary evil. A true protocol hack, mesh groups are often used to eliminate the need for complex MSDP/MBGP session interdependencies. However, this mechanism circumvents the entire goal of peer-RPF flooding. Furthermore, mesh groups are not supported in MSDP traceroute, which is briefly described in Chapter 12.

A router can be a member of multiple mesh groups, but this is strongly discouraged in an attempt to avoid SA looping.

5.6 MSDP Policy

MSDP policy can be enforced using SA message filters. SA filtering can typically be performed on source address, group address, and MSDP peer address. Care should be taken before applying SA filters in transit domains because if the MSDP speakers in the domain where the filtering occurs are the RPF peer for other domains, it can cause loss of connectivity. For example, imagine the following scenario:

- A peer in domain A originates an SA for a local source and sends it to an MSDP peer in domain B.
- The peer in domain B is the RPF peer of domain C for sources in domain A.
- Domain B does not forward this SA to its peer in domain C.

In this scenario, domain C is blackholed from sources in domain A. For this reason, it is much better practice to influence the path of interdomain multicast traffic by using an MBGP policy to change the RPF information.

MSDP policy is most useful in preventing the leaking of SA messages that should not leave a local domain. These include SA messages containing the following:

- Sources in private address space (for example, 10/8)
- Groups that are reserved for protocol use (for example, auto-RP groups, 224.0.1.39 and 224.0.1.40)
- Administratively scoped groups (239/8)
- SSM groups (232/8)

It is good practice to apply SA filters to all MSDP sessions with peers outside a domain to prevent SA messages containing these sources or groups from leaking into or out of the domain.

5.7 SA Storms, Ramen, and MSDP Rate Limiting

Far more destructive than the delicious snack (or meal) for college students that is its namesake, the Ramen worm is a self-propagating program that caused major problems for multicast-enabled networks in early 2001. Worms are similar to viruses in that they are annoying or harmful and self-replicating, but they do not attach themselves to other files or programs as viruses do. The intent of Ramen was more to annoy than to harm.

Once Ramen infects a PC, it scans a range of addresses to find other vulnerable hosts to which it can attach itself. Because of sloppy coding in Ramen, multicast addresses can be scanned as well. When Ramen scans a multicast-enabled network, a Register message and an SA message are generated for every multicast address that is scanned. Ramen can scan through a /16 range of addresses in about 15 minutes, causing 65,000+ SAs to be generated. This number is compounded if multiple hosts are infected. This flood of SAs can crash routers that are not able to process all the SA messages.

No satisfactory method exists to proactively avoid such storms. Some networks have applied rate limits to MSDP traffic. However, rate limiting of control packets always provides vulnerability to denial-of-service attacks because the rate limiters cannot tell the difference between good traffic and bad traffic. For instance, imagine a rate limit is applied that allows only 200Kbps of packets to or from TCP port 639 to enter or leave an MSDP-speaking router. If a malicious attacker flooded that router with a high rate of traffic destined for TCP port 639, the allowable limit of MSDP traffic would be filled with useless data. The "good" MSDP packets from valid peers would be dropped as well, causing MSDP sessions to drop. Multicast in this domain would fail to operate properly.

IMR is still in its adolescence. When interdomain unicast routing was at the same point in its development, similar growing pains were experienced. MSDP SA storms are roughly analogous to the leaking of bad BGP routes, which is no longer the crippling common occurrence it once was. Over time, vendors augment their implementations, and ISPs develop best practices that reduce the likelihood of network disasters.

Accordingly, the safest current defense against SA storms is vigilance. By monitoring the size of an SA cache and being prepared to take action, such as adding temporary filters when levels become extraordinarily high, networks can become hardened to attack without adding new vulnerabilities.

5.8 Outlook for MSDP

MSDP has been affectionately referred to as a cocktail napkin protocol. The protocol was created as a temporary solution for multicast routing between PIM-SM domains prior to the advent of SSM. MSDP's ability to scale can best be described as somewhere between "good enough for now" and "a disaster waiting to happen."

MSDP is not needed in SSM. In fact, MSDP is prohibited from advertising source information for SSM groups. The reduced dependence upon MSDP is actually one of the principle benefits of SSM. The arrival of SSM has delayed the necessity for the Border Gateway Multicast Protocol (BGMP). BGMP is the IMR protocol expected to meet the long-term scalability needs of the Internet. BGMP is still in the early development stage, and a full discussion of it is beyond the scope of this book. It is discussed briefly in Chapter 13.

The current stance of the IETF is to use BGMP as the IMR protocol for IPv6. In the meantime, the combination of MSDP and SSM is expected to provide adequate scalability for IMR in IPv4. After all, despite its inherent weaknesses, MSDP has been successfully deployed in many production networks.

Chapter 6. Source-Specific Multicast (SSM)

To this point in the book, we have generally examined multicast routing from an ASM perspective. With a clear understanding of ASM, the operation and benefits of Source-Specific Multicast (SSM) become very apparent. This chapter describes SSM and how multicast protocols are modified in order to support this service model.

6.1 Introduction

The original vision for multicast in RFC 1112 supported both one-to-many and many-to-many communication models and has come to be known as Any-Source Multicast (ASM). To support these models, an ASM network must determine all of the sources of a group and deliver all of them to interested listeners. In ASM, this function of source discovery rests squarely in the hands of the network.

We have already seen in Chapter 3 that dense protocols provide source discovery by flooding the actual data to all of the routers in a domain. While it is probably the simplest way to inform all routers of multicast sources, flooding presents significant scalability issues and inefficiently uses network resources. Sparse protocols achieve the same functionality with mechanisms that are much more scalable and efficient but present a substantial amount of added complexity. In PIM-SM, we saw how only one router in the domain (the RP) is responsible for knowing all the multicast sources, and the distribution tree is rooted around that router.

Thus it can be said that the primary shortcomings of dense protocols are inefficiency and lack of scalability, while the primary shortcoming of sparse protocols is complexity. In both cases, the mechanisms that cause these shortcomings are trying to accomplish the same goal: source discovery.

The primary beneficiary of a network-provided source discovery control plane is the many-to-many model, where sources for any given group come and go. However, applications now believed to possess the greatest potential for commercial viability across the Internet generally use the one-to-many model. Thus the primary deficiencies of the ASM do provide certain functionality; however, this functionality is now considered less important for Internet applications.

By ignoring the many-to-many model and focusing on the one-to-many model, the vast majority of "interesting" applications can be supported by mechanisms that are much simpler than those found in ASM. SSM, while supporting a subset of ASM functionality, enables this vision of desired functionality through simplicity. Moreover, SSM provides a number of added benefits as a side effect of having to support only the one-to-many model.

SSM, which is currently defined in an Internet-Draft within the IETF's SSM Working Group (draft-ietf-ssm-arch-00.txt), is a service model that supports one-to-many multicast delivery through the use of shortest path trees (SPTs). While it is theoretically possible to support this service model with any protocol that meets its requirements, SSM is generally supported through a subset of functionality in PIM-SM and IGMPv3. We focus on how these protocols specifically support SSM.

Perhaps to make up for an inherent lack of complexity, a new set of terminology is introduced in SSM that describes the same terms that we have used in ASM. When describing SSM, it is preferred to use the words subscribe and unsubscribe instead of the ASM terms join and leave. (This usage of join and leave should not be confused with the various protocol message names, such as PIM Join messages and IGMP Leave-Group messages). The idea behind subscribe and unsubscribe is to differentiate SSM from ASM, even though the operations are identical.

The following table compares ASM and SSM terminology:

Term	Any-Source Multicast (ASM)	Source-Specific Multicast (SSM)
Address identifier	G	S,G
Address designation	group	channel
Receiver operations	join, leave	subscribe, unsubscribe
Group range	224/4 excluding 232/8	224/4[a]

[a] SSM is permitted in all of 224/4 but guaranteed only in 232/8.

6.1.1 Overview of SSM Operation

In SSM, source discovery is provided by some sort of out-of-band means from the perspective of the network; that

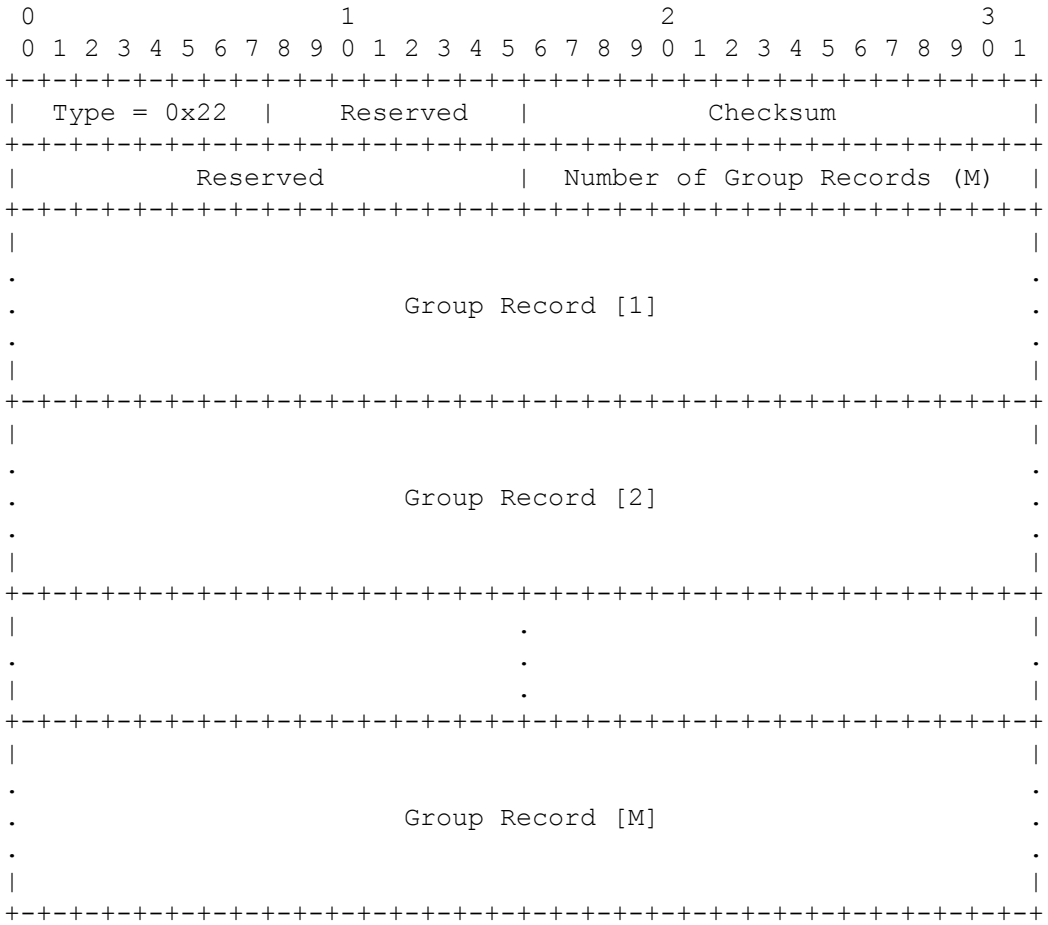
6.2 IGMPv3 in SSM

As we have discussed throughout, routers use IGMP to discover directly connected group members. This section details the new features in version 3 of IGMP, which add the ability for a host to subscribe to or unsubscribe from an SSM channel. It is important to note that IGMPv3 is not used solely for SSM. IGMPv3 introduces two new source-filtering modes, exclude mode and include mode, and only one of them provides SSM functionality.

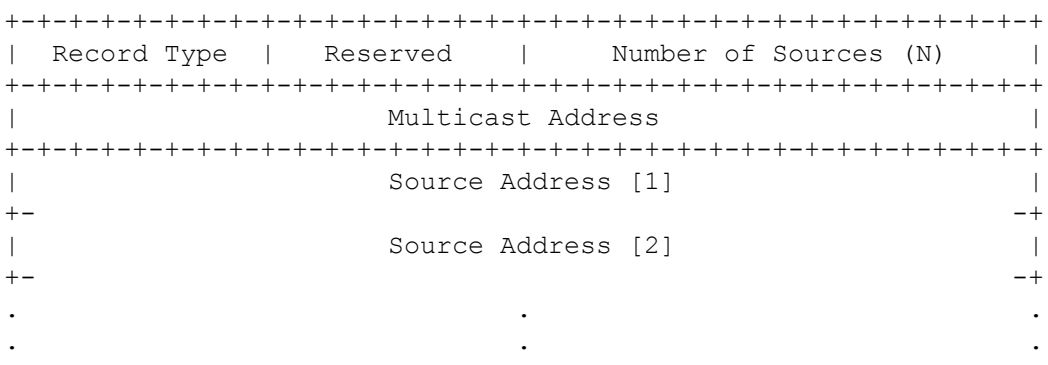
As stated in Chapter 2, exclude mode enables a host to request traffic for a group from all sources except those specified. Include mode enables a host to request traffic for a group from only specified sources. Include mode with a single specified source is used to support SSM.

To enable SSM functionality, IGMPv3 modifies the format of various messages to enable a host to specify the source address of interest in addition to the group address. IGMPv3 must be running on the host and on the host's directly connected router for SSM functionality to work.

The format of IGMPv3 Membership Report messages enables the host to specify both the source address and group address, fully describing the SSM channel. IGMPv3 reports have the following format:



Where each Group record has the following format:



6.3 PIM-SM in SSM

Because it was already capable of building SPTs, PIM-SM required very little to be added to support SSM. The additions to PIM-SM for SSM primarily involved defining behavior in the SSM address range.

When a host subscribes to an SSM channel through IGMPv3, the directly connected PIM-SM router (the receiver's DR) initiates the creation of the SPT by sending an (S,G) Join message to its RPF neighbor for the source. The SPT is built hop by hop until it either reaches a router already on the SPT or a router connected to the source itself. Once the SPT is built, data packets for the SSM channel are delivered to the subscribing host.

Shared tree behavior is prohibited for groups in the SSM range. Accordingly, SSM routers must never send (*,G) Joins for groups in the 232/8 range. If an SSM router receives a (*,G) Join for a group in 232/8 (presumably from an ill-behaving router), it ignores the message. Likewise, it ignores nonsource-specific IGMP reports for groups in this range.

The DR for a source must not send Register messages for groups in 232/8. An RP ignores all Register messages and never creates, sends, or accepts an MSDP SA message for a group in this range.

For the most part, this behavior was possible on Juniper Networks and Cisco Systems routers prior to the existence of SSM by configuring filters and policy. Recently, these implementations have simply added commands that enforce SSM behavior in the SSM range in a much simpler and friendlier way. Thus it can be said that these implementations have always supported SSM for all router roles except the receiver's DR, even before SSM came into being.

Likewise, any domain that supports ASM can support SSM with the addition of this minor configuration. Thus the investment of building an RP-based ASM infrastructure does not go to waste, as both of these models can be supported side-by-side.

Finally, the potential of SSM-only deployments is very attractive for networks with engineers who are unfamiliar with multicast. By simply turning on PIM-SM on all router interfaces and configuring the commands that ensure SSM behavior in 232/8, a network can support SSM. The effort required to design, deploy, and operate such a network is minimal. An SSM-only network is also an ideal stepping-stone for deploying an ASM network. In this case, SSM provides the "training wheels" for engineers to become familiar with multicast, which prepares them to handle the far more complex ASM world.

Chapter 7. Multiprotocol Extensions for BGP (MBGP)

This chapter describes in detail how routers use MBGP to transfer route information used for the reverse path forwarding (RPF) checks of the PIM and MSDP protocols. RPF checks can be performed on the same routing table used by the router to forward unicast traffic…However, splitting these two functions over two separate tables provides the flexibility to enforce different policies for unicast and multicast traffic and to create incongruent topologies for each.

7.1 Overview

MBGP is used to populate a separate routing table dedicated for the RPF mechanisms used when forwarding multicast packets, forwarding certain PIM-SM messages, and deciding whether to accept MSDP Source-Active (SA) messages.

The reason for having a dedicated multicast RPF table is to achieve an incongruent next-hop selection for unicast versus multicast RPF routes; that is to say, the next hop used to route unicast packets to prefix P can be different from the next hop used for multicast RPF checks.

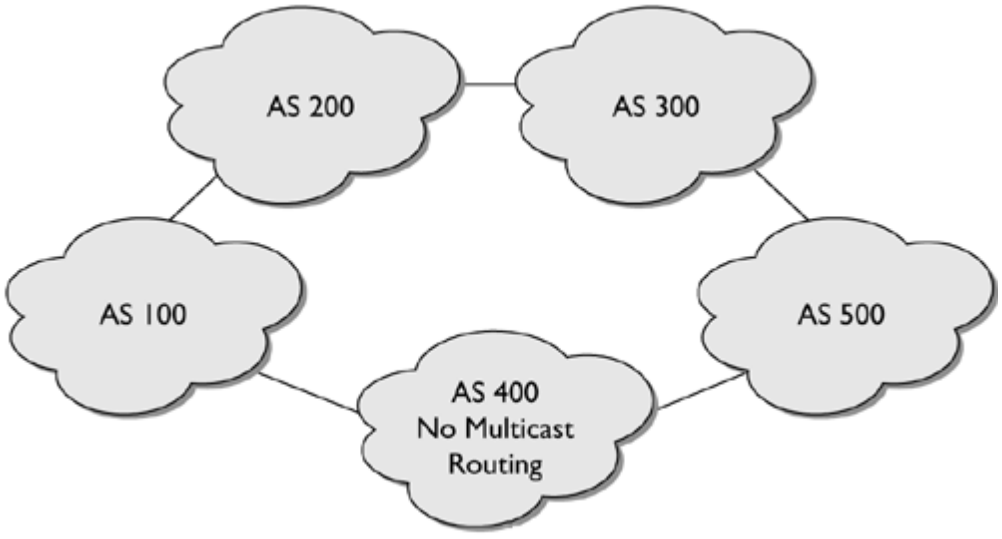
This incongruence is advantageous in two situations:

- To avoid routers not capable of multicast routing as multicast RPF next hops but still use them as unicast forwarding next hops
- To use different links and routers for multicast and unicast traffic—for example, where ISPs publicly peer at a Multicast Internet Exchange (MIX) to exchange multicast traffic only

It is completely possible to use PIM-SM and MSDP for multicast routing without MBGP, as long as the next hops for any prefix P can be the same for both unicast forwarding and multicast RPF. However, it is wise to use MBGP from the start, even if it appears that incongruent paths are not a necessity. Topology changes, additional MSDP peers, and so on can potentially lead to such a need in the future. Using MBGP does not force you to have incongruent paths. And if the need arises, it is much easier to modify the policy of an existing MBGP setup than it is to convert from standard BGP to MBGP throughout your network and with all your external peers.

Figure 7-1 shows an example of a common topology where MBGP is quite useful. In this figure, we see that without MBGP, routers in AS 100 using standard BGP for RPF checks would typically select a path through AS 400 to get to sources in AS 500. Because AS 400 is not multicast-enabled, multicast sources in AS 500 would be blackholed for receivers in AS 100. With MBGP, RPF could select the path through AS 200 and AS 300, while unicast routing could continue to use AS 400 to reach AS 500 destinations.

Figure 7-1. Topology with incongruent unicast and multicast paths



Furthermore, while it is possible to control the unicast and multicast topologies within a domain, the world outside your own autonomous system is a much trickier place. Many of the current deployments on the Internet have used multicast RPF topologies significantly different from those used for unicast routing, especially when connecting to other domains. In practice, this means multicast multihoming without MBGP can easily lead to multicast black holes, where RPF selects a path that is not multicast-enabled.

7.2 BGP and Related Terminology

This chapter assumes the reader has working knowledge of BGP and specifically the use of BGP for interdomain routing of IP unicast packets. This section contains a brief overview of BGP to enable discussion of MGBP using consistent terminology.

The Internet is made up of thousands of heterogeneous internetworks, each maintained by a separate organization. An internetwork maintained by a single operating group is termed an autonomous system (AS). Each AS has a primary interior gateway protocol (IGP) that handles the routing of IP unicast packets within the AS. OSPF (Open Shortest Path First), IS-IS (Intermediate System to Intermediate System), and RIP (Routing Information Protocol) are examples of IGPs.

In order to exchange routes with other ASs, an exterior gateway protocol (EGP) is required. BGP4 is the EGP used in the Internet for routing IPv4 unicast packets across multiple autonomous systems and is defined in RFC 1771.

Three main functions separate the various IGPs and BGP:

- BGP can handle a much larger number of routes.
- BGP has a much more versatile array of attributes that can be used for enforcing policies.
- BGP is focused on routing packets by means of AS hops, not router hops.

Note

In this book, BGP and BGP4 refer to the same thing, with the "4" simply indicating the current version of Border Gateway Protocol and the specific version of the protocol we refer to.

IGPs are very trusting of the information received from their neighbors. They assume neighboring routers are part of their own domain and therefore share the same routing policy. BGP assumes the opposite is true because the information it receives from external peers is from a different organization. BGP tries to ensure the credibility and stability of that information.

Connections between BGP speakers of different ASs are referred to as external links. BGP connections between BGP speakers within the same AS are referred to as internal links. Similarly, a peer in a different AS is referred to as an external peer, while a peer in the same AS may be described as an internal peer.

7.3 BGP Internals—Foundation for Understanding MBGP

A router configured to send and receive route information via BGP is known as a BGP speaker. To exchange route information, a BGP speaker forms adjacencies with peer routers. BGP peers establish a TCP session using port number 179.

Both peers attempt to open a connection to port 179, but only one connection is kept up for the peer adjacency. The peer initiating the connection uses a random port number on its side of the BGP session. All BGP packets have the following format:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| IP Header | TCP Header  | BGP Header  | BGP Message |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The BGP header contains three fields:

- Marker: Can be used for authentication and synchronization of the BGP session
- Length: Indicates the total length of the message, including the BGP header, in octets
- Type: Indicates the type of message to follow

BGP has four types of messages:

- Open
- Update
- Notification
- Keepalive

An explanation of how these messages are used requires an understanding of the BGP adjacency finite state machine. The clearest way to explain this state machine is to walk through the flow when everything works properly—that is, the connection comes up and routes are exchanged. The following example describes what happens when two routers, router A and router B, are configured as BGP peers for the first time.

1. Router A CONNECT; router B IDLE

   Router A is configured, identifying router B as a potential BGP peer. This configuration shifts router A into the CONNECT state, and it tries to initiate a TCP connection with router B on port 179.

2. Router A ACTIVE; router B IDLE

   This connection fails because router B has yet to be configured. This failure shifts router A into ACTIVE state. In ACTIVE state, router A listens on port 179 for a connection initiated by router B. Every few seconds, router A again attempts to initiate a TCP session.

7.4 Extending BGP: MBGP

MBGP is not a separate protocol but an extension of BGP, so the specifics of MBGP peering are similar to conventional BGP peering. RFC 2858 defines the multiprotocol extensions for BGP4, and the extensions are implemented as optional path attributes. A standard BGP Update message may contain multiple destination prefixes that share the same path attributes such as AS_PATH, NEXT_HOP, MED, and so on.

As previously stated, a BGP Update message contains a single instance of each path attribute, plus a list of prefixes that share those particular attribute values. This strategy is unlike that of most IGPs, whose updates contain a list of prefixes, each listed with its own attributes.

BGP's method of exchanging updates leads to the efficient use of bandwidth for a protocol with so many attributes, especially considering that many of the attributes are optional and do not pertain to every prefix.

MBGP adds two new path attributes called MP_REACH_NLRI and MP_ UNREACH_NLRI. MP_REACH_NLRI is used instead of the standard BGP NLRI for prefixes from protocols other than IPv4 or for IPv4 prefixes intended for a routing table other than the unicast forwarding table. The MP_UNREACH_NLRI attribute is used in place of the Withdrawn Routes field of the standard BGP UPDATE message to indicate that the specified prefixes are unreachable.

MBGP can be used to carry forwarding information for any protocol that has a prefix-mask hierarchical address space. Possible protocols include IPX (Novell's Internetwork Packet Exchange) and IPv6. The most popular implementation of MBGP currently is for multicast routing. This application is so popular that it is common to hear MBGP translated to "Multicast Border Gateway Protocol."

7.5 MBGP Internals

An MBGP speaker must have an IPv4 address in order to establish sessions to its peers, even if it is only exchanging routing information for protocols other than IPv4. BGP4 has three attributes that are IPv4 specific:

- NEXT_HOP
- AGGREGATOR
- NLRI

MBGP does not specify a way to use other protocols' addresses in the AGGREGATOR attribute. It is possible to aggregate prefixes of other network layer pro tocols, but the router performing the route aggregation is denoted by its IPv4 address in the AGGREGATOR attribute for the path.

MBGP does add the ability to associate other network layer protocols' prefix addresses with next-hop information specific to that protocol. None of the other BGP path attributes is specific to IPv4, so the path attributes are used "as is" for non-IPv4 MBGP reachability information.

MBGP uses the numbers assigned to address families in RFC 1700. The assigned numbers are listed in Table 7-1

Table 7-1. MBGP Address Family Numbers

Number	Description
0	Reserved
1	IPv4
2	IPv6
3	NSAP
4	HDLC (8-bit multidrop)
5	BBN 1822
6	802 (includes all 802 media plus Ethernet "canonical format")
7	E.163
8	E.164 (SMDS, Frame Relay, ATM)
9	F.69 (Telex)
10	X.121 (X.25, Frame Relay)
11	IPX
12	AppleTalk
13	DECnet IV
14	Banyan Vines
65535	Reserved

These numbers are referred to as address family identifiers (AFIs). MBGP also uses subsequent address family identifiers (SAFIs) to provide additional information about the type of the NLRI included in the MBGP Update message. The SAFIs for IPv4 are as follows:

- 1: NLRI used for unicast forwarding
-

7.6 Using MGBP for Multicast Routing

While MBGP is not used to feed the (S,G) and (*,G) tables used for forwarding multicast packets out the correct interfaces, MBGP can indirectly influence the flow of multicast traffic. PIM Join/Prune messages are sent to a router's RPF neighbor. Upon receiving a Join, a router adds the interface on which the Join was received to the outgoing interface list for the multicast group. Thus, by populating the routing table used to determine the RPF neighbor, MBGP can influence which direction the multicast packets take through the network.

The primary application for MBGP is to create different topologies for unicast and multicast traffic. A benefit exists, however, of having MBGP running in the case of congruent topologies. Although both unicast and multicast packets traverse the same links, disparate policies can be applied to unicast and multicast BGP routes.

7.6.1 Manipulation of Path Attributes

Path attributes can be manipulated separately for both unicast and multicast paths. In particular, this section illustrates how routers use MBGP to transfer unicast route information employed specifically for the RPF checks of the PIM-SM and MSDP protocols.

MBGP can be used to achieve incongruent routing within a domain. Doing so typically requires manual manipulation of the NEXT_HOP attribute across IBGP sessions, which can be an administrative burden and can reduce redundancy. Recall the primary purpose of BGP is to provide policy-based routing between two ASs. BGP relies on the underlying IGP to make routing decisions within a particular AS. For this reason, it is much easier to manipulate the IGP to support disparate routing topologies for unicast and multicast within an AS. Chapter 8 describes how M-ISIS can be used for this purpose.

Using MBGP to achieve incongruent routing across ASs is much cleaner, and there are plenty of options (for example, MED, LOCAL_PREF, and AS_PATH prepending). In this example, the LOCAL_PREF attribute is manipulated to accomplish the goal.

We start out with congruent routing. The local router has two EBGP peers. They are 10.0.0.1 in AS 100 and 10.0.1.1 in AS 200. Both of these peers are advertising a SAFI = 1 and 2 route to 192.168.1.0/24. The route learned from 10.0.0.1 is chosen as the active route for both unicast forwarding and multicast RFP because its router ID is lower (indicated with the asterisk).

```
kalamata> show route 192.168.1.0/24

Unicast Routing Table
192.168.1.0/24    *[BGP/170] , localpref 100, from 10.0.0.1
                    AS path: 100 300 I > via so-6/1/0.0
                   [BGP/170] , localpref 100, from 10.0.1.1
                    AS path: 200 300 I > via so-6/2/0.0

Multicast RPF Table
192.168.1.0/24    *[BGP/170] , localpref 100, from 10.0.0.1
                    AS path: 100 300 I > via so-6/1/0.0
                   [BGP/170] , localpref 100, from 10.0.1.1
                    AS path: 200 300 I > via so-6/2/0.0
```

To achieve incongruent routing, a policy is applied to the multicast RPF route learned from 10.0.1.1. This policy sets the LOCAL_PREF for that route to 110. This change causes the preferred route for multicast RPF to be different from the route preferred for unicast forwarding. Notice that the route through 10.0.1.1 is now the active route for multicast RPF.

```
kalamata> show route 192.168.1.0/24

Unicast Routing Table
192.168.1.0/24    *[BGP/170] , localpref 100, from 10.0.0.1
                    AS path: 100 300 I > via so-6/1/0.0
```

Chapter 8. Multitopology Routing in Intermediate System to Intermediate System (M-ISIS)

This chapter describes Multitopology Routing in IS-IS (M-ISIS), which extends the capabilities of the IS-IS routing protocol. These extensions have enabled M-ISIS to evolve into a general-purpose tool for providing multiple-topology support for technologies such as in-band management, multicast, and IPv6. As we did in Chapter 7 with MBGP, we focus on how M-ISIS can be used to create two separate virtual topologies, one for unicast and another for multicast.

In interdomain unicast routing, recursive routing is used to select the best path to a destination. When a BGP-learned route is selected as the best path to a destination, an IGP such as OSPF or IS-IS determines the path to the BGP next hop of the selected BGP route. Thus BGP generally makes routing decisions between different autonomous systems, while an IGP makes routing decisions within an autonomous system.

M-ISIS and MBGP can be used side-by-side to build a dedicated multicast RPF table, much as IS-IS and BGP have traditionally coexisted. M-ISIS provides the ability to create incongruent topologies within the AS, whereas MBGP provides this ability between ASs.

At the time of writing, Juniper Networks routers support M-ISIS, while Cisco Systems routers do not.

8.1 Overview of IS-IS

IS-IS is the most common IGP found in the networks of the world's largest ISPs, for reasons that are mainly due to circumstance. In the early 1990s, the first large ISPs such as UUNet, MCI, and Sprint were beginning to build IP backbones and needed to select an IGP. Because a link-state routing protocol was desired, IS-IS, developed by ISO (International Organization for Standardization), and OSPF, developed by IETF, were the main candidates.

At that time, Cisco Systems had just implemented a link-state routing protocol for Internetwork Packet Exchange (IPX) called NetWare Link Services Protocol (NLSP). Because NLSP is very similar to IS-IS, Cisco Systems software developers rewrote the IS-IS code at the same time. The newer, more stable implementation of IS-IS was selected by these ISPs, where it continues to run today.

Over time, some have suggested that IS-IS is more stable for large carrier networks than its rival, OSPF. However, the operation of both is very similar, and there is nothing inherently better about one than the other. Further, with years of evolution and deployment experience, software implementations of both protocols by router vendors such as Juniper Networks and Cisco Systems have matured to be equally stable. In spite of this, the IS-IS versus OSPF debate continues to provide intense discussion among the fervent partisans of each.

8.1.1 IS-IS Background

The Intermediate System to Intermediate System (IS-IS) routing protocol is specified in ISO 10589 (republished as RFC 1142). The ISO standard only describes how IS-IS can be used to route ISO's Connectionless Network Protocol (CLNP) packets. RFC 1195 integrated IS-IS as a routing protocol capable of carrying IPv4 prefixes. In ISO terminology, an intermediate system is a router. A host in ISO terms is known as an end system. The following lists some of the specification documents for IS-IS:

- ISO/IEC 10589, "IS-IS Intra-Domain Routing Information Exchange Protocol"
- RFC 1195, "Use of OSI IS-IS for Routing in TCP/IP and Dual Environments"
- RFC 2763, "Dynamic Hostname Exchange Mechanism for IS-IS"
- RFC 2966, "Domain-wide Prefix Distribution with Two-Level IS-IS"
- RFC 2973, "IS-IS Mesh Groups"
- draft-ietf-isis-traffic-04.txt, "IS-IS Extensions for Traffic Engineering"
- draft-ietf-isis-wg-multi-topology-02.txt, "M-ISIS: Multi Topology Routing in IS-IS"

Unlike most other unicast IP routing protocols, IS-IS is rarely documented outside of the standards themselves. Thus, in this chapter, we provide some detail of how IS-IS works at a core level to ensure that subsequent sections in this chapter on multitopology extensions make sense.

As stated earlier, IS-IS, like OSPF, is a link-state routing protocol. Because OSPF tends to be better understood throughout the networking community, we compare IS-IS to the operation of OSPF throughout this chapter. With an understanding of OSPF, learning other link-state protocols such as IS-IS becomes straightforward. IS-IS and OSPF strive for the same goal, simply approaching it from different angles.

8.2 Specifics of IS-IS

In this section, we discuss several important IS-IS topics, including the use of packets, establishing adjacencies on point-to-point links, determination of designated routers on multiaccess networks, and exchanging link-state information with neighbors.

8.2.1 IS-IS Packets

Every IS-IS packet (or PDU in OSI-speak) begins with a list of mandatory fields specific to that packet type. The fixed PDU fields are followed by various TLVs. Some TLVs apply to certain PDUs and not to others. IS-IS uses the following packets to exchange protocol information:

- Level 1 IS-IS Hello PDU (IIH): Used on LANs to discover the identity of neighboring level 1 IS-IS systems, elect a designated intermediate system, and keep up the adjacencies.
- Level 2 IS-IS Hello PDU (IIH): Used on LANs to discover the identity of neighboring level 2 IS-IS systems, elect a designated intermediate system, and keep up the adjacencies.
- Point-to-Point Hello (IIH) PDU: Used on point-to-point links to discover the identity of the neighboring IS-IS system, determine whether the neighbor is a level 1 or level 2 router, and keep up the adjacency. This message is the only one that is not level dependent. The same packet format is used for both levels.
- Level 1 link-state PDU (LSP): Contains information about the state of adjacencies to neighboring level 1 IS-IS systems. LSPs are flooded periodically throughout an area.
- Level 2 link-state PDU (LSP): Contains information about the state of adjacencies to neighboring level 2 IS-IS systems. LSPs are flooded periodically.
- Level 1 Complete Sequence Number PDU (CSNP): Used to synchronize level 1 link-state databases when adjacency first comes up and periodically thereafter.
- Level 2 Complete Sequence Number PDU (CSNP): Used to synchronize level 2 link-state databases when adjacency first comes up and periodically thereafter.
- Level 1 Partial Sequence Number PDU (PSNP): Used to request one or more level 1 LSPs that were detected to be missing from a level 1 CSNP. The local router sends a level 1 PSNP to the neighbor that transmitted the incomplete level 1 CSNP. That router, in turn, forwards the missing level 1 LSPs to the requesting router.
- Level 2 Partial Sequence Number PDU (PSNP): Used to request one or more level 2 LSPs that were detected to be missing from a level 2 CSNP. The local router sends a level 2 PSNP to the neighbor that transmitted the incomplete level 2 CSNP. That router, in turn, forwards the missing level 2 LSPs to the requesting router.

8.2.2 IS-IS Neighbor State Machine on Point- to-Point Links

To establish adjacencies on point-to-point links, each side declares the other side to be reachable if a Hello packet is

8.3 Overview of M-ISIS

The multitopology extensions to IS-IS provide four main features:

- A way to tag Hello packets as belonging to certain topologies
- A means of tagging LSP information as being specific to a topology
- Separate SPF calculations for each topology
- Backward compatibility with legacy IS-IS implementations

IS-IS is already "multiprotocol" because it can carry routing information for both ISO and IPv4 network addresses. M-ISIS gives IS-IS the added ability of being able to view the underlying topology in a different way for each independent IP topology. Each of these multitopologies (MTs) views the cost of each link throughout the network independently from other MTs.

Similar to the multiprotocol extensions to BGP, M-ISIS can be used for purposes other than populating a dedicated multicast RPF table. For example, each topology can have overlapping IP prefixes, so it might be possible to employ M-ISIS in some VPN schemes.

M-ISIS is backward compatible with standard IS-IS implementations. The protocol overcomes the following two challenges in order to achieve backward compatibility:

- Establishing IS-IS adjacencies
- Advertising prefixes within each MT

Level boundaries are consistent across all MTs, which enables only one adjacency to be required for each level the router is exchanging with each of its peers. For example, two level 1 neighbors running M-ISIS establish only a single level 1 adjacency; they need not have a separate adjacency for each MT in which they both participate.

MT 0 is a special MT. It is equivalent to the standard IS-IS topology. LSPs tagged with MT 0 are placed in the same link-state database as untagged LSPs. Tagging an LSP with MT 0 is optional if it is the only MT on the interface. Untagged routes are considered to be in MT 0.

MT 3 is reserved for multicast RPF topology. LSPs tagged with MT 3 are placed in the link-state database dedicated to the multicast RPF topology. The router runs the SPF algorithm separately on each MT link-state database to determine the best paths for each prefix in that MT.

The best routes, determined by running the SPF algorithm on the MT 3 link-state database, are placed in the dedicated multicast RPF routing table. From this routing table, they are used for the RPF checks of such protocols as PIM-SM and MSDP. Figure 8-2 presents a conceptual view of the information flow within a router running MBGP and M-ISIS.

Figure 8-2. Information flow for a router running MGBP and M-ISIS

8.4 Specifics of M-ISIS

M-ISIS adds functionality to various parts of the base IS-IS protocol. Affected parts include the forming of adjacencies and advertising of prefixes. New IS-IS TLVs are specified to enable the new functionality.

8.4.1 Forming Adjacencies

On point-to-point links, an M-ISIS adjacency is associated with a set of MTs. In their Hello packets, both routers advertise the MTs they have configured for the interface. The set of MTs associated with the adjacency consists of those present in both routers' Hellos. If the two routers do not share any MTs, the adjacency does not need to be formed. Absence of M-ISIS TLVs in the Hello is interpreted as MT 0.

Broadcast media present more complications for backward-compatible M-ISIS adjacencies. All M-ISIS routers on a LAN announce that they are MT-capable in their Hellos. To maintain backward compatibility, the DIS function is not MT-enabled; that is to say, the pseudonode LSP created by the DIS does not contain any MT information. On a LAN with mixed M-ISIS and legacy IS-IS speakers, the DIS can be any of the routers. Even if the DIS is MT-capable, it does not include MT information in the pseudonode LSP. Because of this operation, it is possible to make a graceful transition from a legacy IS-IS network to an M-ISIS network.

8.4.2 M-ISIS TLVs

The M-ISIS draft defines three new TLVs:

- TLV 222: Multitopology Intermediate Systems TLV
- TLV 229: Multitopology TLV
- TLV 235: Multitopology Reachable IPv4 Prefixes TLV

TLV 222 leverages the format and functionality of TLV 22 (Extended IS Reachability). TLV 222 is found only in LSPs. The format of the Value field for the multicast topology is as follows:

```
                     1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0|0|0|      MT ID = 3         |                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
|                                                               |
|    0 - 251 octets of various structures used in TLV 22        |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

TLV 229 lists the MT ID numbers in which the local router is participating. This TLV is applicable to all Hello PDUs and LSP fragment 0. For a router that supports unicast and multicast topologies, the format of the Value field is as follows:

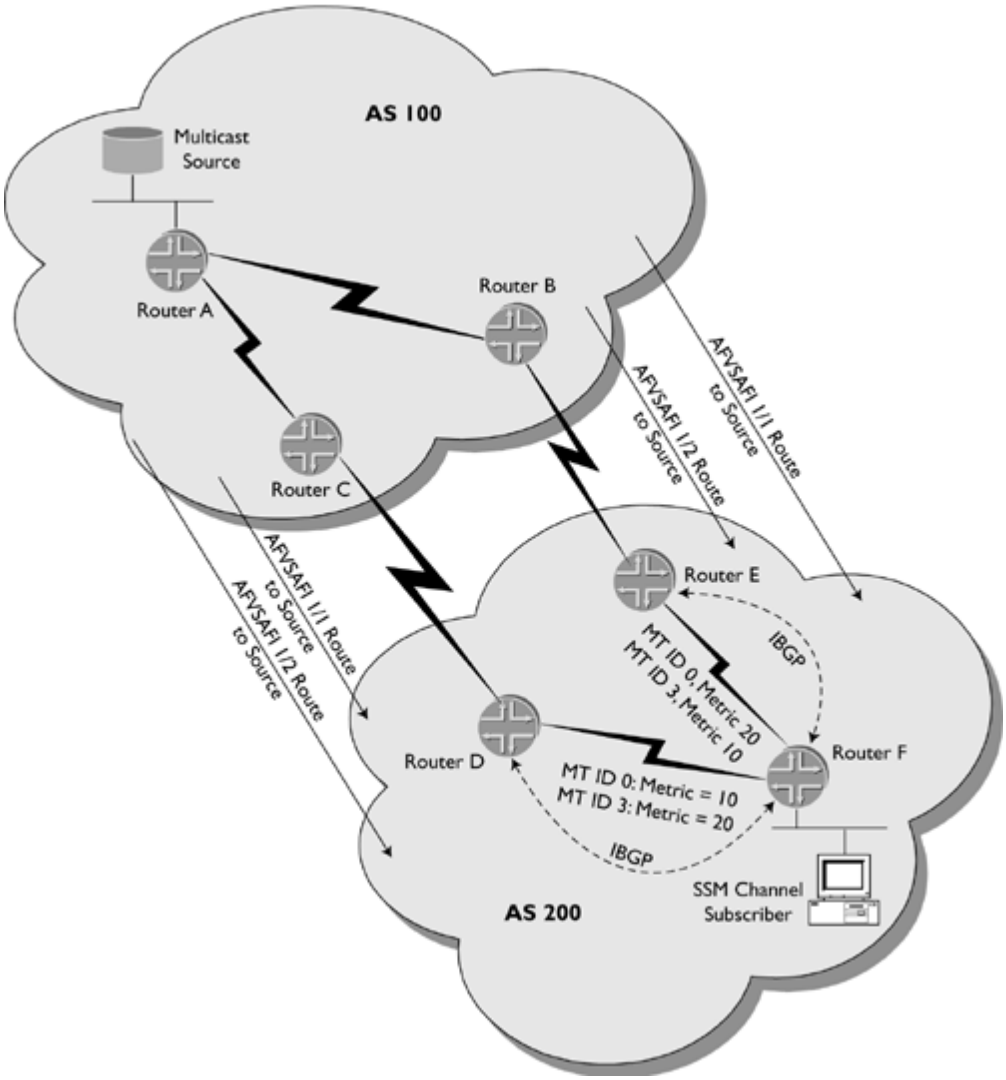
```
                     1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|0|0|0|      MT ID = 0         |0|A|0|0|      MT ID = 3         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Value field is divided into 2-byte sections for each supported MT. The most significant 4 bits are control bits.

8.5 Examples of Using M-ISIS

Within the realm of multicast routing, there are two main uses of M-ISIS. The first is to set up a separate RPF topology for sources inside an autonomous system. The second is to set up a multicast-specific topology for resolving MBGP next hops for routes to sources outside the local AS; these routes are learned from IBGP peers that are not directly connected. Figure 8-3 shows a generic example of the this second usage.

Figure 8-3. M-ISIS used to resolve MBGP next hops (Routers D and E use a next-hop self-export policy for their IBGP peers.)



In Figure 8-3, the routers in AS 200 learn the route to the multicast source in AS 100 over two External BGP (EBGP) connections: one between routers C and D and one between routers B and E. Both of these connections are MBGP-enabled. These two routes have identical path attributes except for the BGP next hop.

Router F receives the two BGP routes to the source with AFI/SAFI equal to 1/1 (unicast routes) and two BGP routes to the source with AFI/SAFI equal to 1/2 (multicast RPF routes). Router F places the unicast routes in its BGP Loc-RIB for unicast routes, and it places the multicast RPF routes in its BGP Loc-RIB for multicast RPF routes.

Lookups for BGP-learned routes are done recursively using an IGP. When a router runs the BGP path selection algorithm to select the best BGP route, it uses the IGP to determine the best path to the BGP next hop of the BGP route.

In this example, router F learns its routes to the BGP next hops (via D and E) in IS-IS. M-ISIS is used here to provide different metrics over the same links for each topology. In the unicast routing table, the IS-IS route to router D has a metric of 10, and the route to router E has a metric of 20. In the multicast RPF table, the IS-IS route to router D has a metric of 20, and the route to router E has a metric of 10.

Chapter 9. Configuring and Verifying Multicast Routing on Juniper Networks Routers

Juniper Networks produces various router platforms targeted for deployment in the core and at the edge of ISP networks. The JUNOS operating system runs Juniper Networks routers. The commands described in this chapter give a general understanding of the minimum configuration needed to enable multicast in JUNOS software. This chapter does not describe every possible multicast command. The full technical documentation for configuring JUNOS software is available at http://www.juniper.net/techpubs/software.html.

All of Juniper Networks platforms keep the route control function and the packet-forwarding function on completely separate hardware modules. The Routing Engine (RE) handles the route control function. The Packet Forwarding Engine (PFE) handles the packet-forwarding function. The RE is a single module. Some platforms can house two Routing Engines, but one is always in backup mode. The PFE is composed of multiple hardware modules: mainly, the Physical Interface Cards (PICs), the Flexible PIC Concentrators (FPCs), and the packet-switching board (the abbreviation for the latter depends on the platform).

The RE has an external management Ethernet interface named fxp0. This interface can be used for remote access and monitoring, but it is not used as a transit interface. The RE has an internal Ethernet interface named fxp1 that it uses to communicate with all the other modules in the chassis. The RE uses the Trivial Network Protocol (TNP) for internal communication, so there is no need to configure an IP address on fxp1. The M160 platform has two internal Ethernet interfaces, the second one being fxp2. All of the external interfaces on the router other than the RE's fxp0 interface are termed PFE interfaces. The PFE interfaces can be used for transit IP and Multiprotocol Label Switching (MPLS) traffic as well as IP and ISO Connectionless Network Service (CLNS—used for IS-IS) traffic destined for the RE.

9.1 Configuring IGMP and PIM

The following sections describe how to configure and manage JUNOS software to support multicast routing within a domain.

9.1.1 Enabling Interfaces for IGMP and PIM

Without any configuration, none of the router's interfaces is enabled for PIM or IGMP. Enabling PIM on the router automatically enables IGMP on all LAN interfaces. Use the following configuration to enable IGMP and PIM-SM with the version 2 packet format on all nonmanagement interfaces:

```
protocols {
    igmp {
        interface fxp0.0 {
            disable;
        }
    }
    pim {
        interface all {
            mode sparse;
            version 2;
        }
        interface fxp0.0 {
            disable;
        }
    }
}
user@m20-a> show pim interfaces
Name          Stat Mode     V State    Priority    DR address    Neighbors
fe-3/3/0.3    Up   Sparse   2 DR       1           10.0.3.1              0
lo0.0         Up   Sparse   2 DR       1           10.0.5.3              0
t3-1/0/0.0    Up   Sparse   2 P2P                                        0

user@m20-a> show igmp interface
Interface         State         Querier       Timeout  Version  Groups
fxp0.0            Disabled                          0        2       0
fxp1.0            Disabled                          0        2       0
t3-1/0/0.0        Disabled                          0        2       0
fe-3/3/0.3        Up            10.0.3.1          None        2       0

Configured Parameters:
IGMP Query Interval (1/10 secs): 1250
IGMP Query Response Interval (1/10 secs): 100
IGP Last Member Query Interval (1/10 secs): 10
IGMP Robustness Count: 2

Derived Parameters:
IGMP Membership Timeout (usecs): 260000000
IGMP Other Querier Present Timeout (usecs): 255000000
```

Notice that enabling PIM on the t3-1/0/0.0 interface does not automatically enable IGMP because this interface is a point-to-point interface and is most likely not connected to any end hosts that could become group members. If a point-to-point interface needs to speak IGMP, it can be explicitly enabled.

Even if a router is attached to a LAN that has only IGMP-speaking hosts and no other PIM-speaking routers, the interface connected to that LAN must still run PIM for the router to function properly. If PIM is disabled on an interface, IGMP shows the UP state, but group membership reports will not be processed correctly. Use the show igmp group command to show the groups joined by directly connected hosts.

```
user@m20-a> show igmp group
Interface         Group          Source          Last Reported   Timeout
fe-3/3/0.3        224.0.0.2      0.0.0.0         10.0.3.2            193
```

9.2 Configuring MSDP

The following shows the configuration of an MSDP speaker with a unique IP address 10.0.0.1 (preferably on lo0.0) peering with two other MSDP-speaking routers:

```
protocols {
    msdp {
        local-address 10.0.0.1;
        peer 10.0.0.3;
        peer 10.0.0.4;
    }
}
```

Use the following command to display the MSDP Source-Active cache:

```
user@m20-a> show msdp source-active
Group address   Source address  Peer address    Originator      Flags
224.0.1.11      10.222.100.32   10.9.201.205    10.222.1.5      Accept
                                10.9.201.254    10.222.1.5      Reject
                                10.9.202.2      10.222.1.5      Reject
                                10.9.202.253    10.222.1.5      Reject
```

From this output, it can be inferred that 10.9.201.205 is the RPF peer for originator 10.222.1.5 because the Flags column shows that it is the only peer from which the router accepted the SA. Another way to show the RPF peer for a particular originator address is as follows:

```
user@m20-a> test msdp rpf-peer 10.222.1.5
MSDP peer is 10.9.201.205 for Originator 10.222.1.5/32
```

To configure peers in an MSDP mesh group, use the following configuration:

```
protocols {
    msdp {
        local-address 10.0.0.1;
        group g {
            mode mesh-group;
            peer 10.0.0.3;
            peer 10.0.0.4;
    }
}
```

To configure a default peer, use the following configuration:

```
protocols {
    msdp {
        local-address 10.0.0.1;
        peer 10.0.0.3 {
            default-peer;
        }
    }
}
```

The following is an example of an SA filter in which SA messages for the 229.9.9.9 group are discarded:

```
protocols {
    msdp {
        peer 10.0.0.3 {
            import msdp-p;
```

9.3 Configuring a Dedicated RPF Table

In JUNOS software, the primary IPv4 unicast routing table is inet.0. This table imports routes from the unicast routing protocols running on the router. By default, PIM and MSDP use inet.0 for their RPF checks. JUNOS software has another routing table, inet.2, that is reserved for use as a dedicated multicast RPF table. The following demonstrates how to configure PIM and MSDP to use inet.2 as the dedicated multicast RPF table:

```
routing-options {
    rib-groups {
        mcast-rpf-rib {
            import-rib inet.2;
        }
    }
}
protocols {
    msdp {
        rib-group inet mcast-rpf-rib;
    }
    pim {
        rib-group inet mcast-rpf-rib;
    }
}
```

A routing information base (RIB) is a routing table. A RIB group associates zero or more import RIBs with zero or one export RIBs. This association does not have any meaning until you apply the RIB group to a routing protocol. The RIB group is interpreted differently depending on the protocol. PIM and MSDP use the first import RIB listed in the configuration as their respective RPF tables. Any other RIBs listed in the RIB group are ignored by PIM and MSDP.

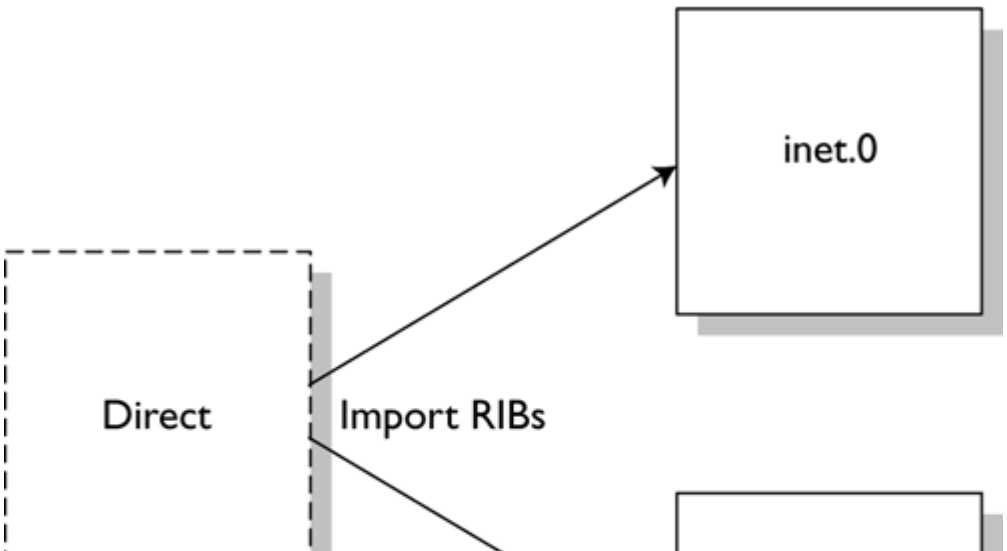
The previously shown configuration tells PIM and MSDP to use inet.2 for their RPF checks. It does not place any routes in inet.2. To verify that the PIM is now using inet.2 as its RPF table, use the following command:

```
user@m20-a> show multicast rpf
Multicast RPF table: INET.2

Source prefix      Protocol   RPF interface    RPF neighbor
```

At this point, PIM is using inet.2 as its RPF table, but there are no routes in inet.2. The first step in building an RPF table is to get the directly connected routes into inet.2, as illustrated in Figure 9-1. Use the following configuration to do so:

Figure 9-1. Use of RIB groups for interface routes



Chapter 10. Configuring and Verifying Multicast Routing on Cisco Systems Routers

Cisco Systems produces a wide variety of router platforms. The IOS (Internetwork Operating System) runs on most Cisco router platforms. The commands described in this chapter provide a general understanding of the minimum configuration needed to enable multicast in IOS. This chapter does not describe every possible multicast command. The full technical documentation for configuring IOS software is available at http://www.cisco.com/univercd/cc/td/doc/product/software/index.htm.

Chapter 10 assumes the reader has experience configuring IOS software and understands such terms as global configuration mode and interface configuration mode. Experience configuring multicast routing features in the IOS software is not assumed.

Note

The organization of this chapter parallels Chapter 9, which describes JUNOS software.

10.1 Configuring PIM and IGMP

In this section, we provide practical configuration guidelines for PIM and IGMP on IOS platforms.

10.1.1 Enabling Interfaces for IGMP and PIM

The first step when configuring IOS software for multicast is to enable multicast routing on the entire router with the following command in global configuration mode:

```
ip multicast-routing
```

Enabling PIM on an interface automatically enables IGMP for that interface. Use the following command in interface configuration mode to enable PIM-SM and IGMP:

```
ip pim sparse-mode
```

Use the following command to show which interfaces have PIM enabled:

```
Router# show ip pim interface

Address         Interface    Mode    Neighbor    Query       DR
                                     Count       Interval
10.92.37.6      Ethernet0    Sparse  2           30          10.92.37.33
10.92.36.129    Ethernet1    Sparse  2           30          10.92.36.131
10.1.37.2       Tunnel0      Sparse  1           30          0.0.0.0
```

Use the following command to show which interfaces have IGMP enabled:

```
Router# show ip igmp interface

Ethernet0 is up, line protocol is up
  Internet address is 10.92.37.6, subnet mask is 255.255.255.0
  IGMP is enabled on interface
  IGMP query interval is 60 seconds
  Inbound IGMP access group is not set
  Multicast routing is enabled on interface
  Multicast TTL threshold is 0
  Multicast designated router (DR) is 10.92.37.33
  No multicast groups joined
Ethernet1 is up, line protocol is up
  Internet address is 10.92.36.129, subnet mask is 255.255.255.0
  IGMP is enabled on interface
  IGMP query interval is 60 seconds
  Inbound IGMP access group is not set
  Multicast routing is enabled on interface
  Multicast TTL threshold is 0
  Multicast designated router (DR) is 10.92.36.131
  No multicast groups joined
Tunnel0 is up, line protocol is up
  Internet address is 10.1.37.2, subnet mask is 255.255.0.0
  IGMP is enabled on interface
  IGMP query interval is 60 seconds
  Inbound IGMP access group is not set
  Multicast routing is enabled on interface
  Multicast TTL threshold is 0
  No multicast groups joined
```

Use the following command to show the groups joined by directly connected hosts:

10.2 Configuring MSDP

To configure an MSDP peer, use the following command in global configuration mode:

```
ip msdp peer 192.168.1.2 connect-source loopback 0
```

By default, the IOS software does not cache Source-Active state. To enable Source-Active caching, use the following command in global configuration mode:

```
ip msdp cache-sa-state
```

Use the following command to display the MSDP Source-Active cache:

```
Router# show ip msdp sa-cache

MSDP Source-Active Cache - 5 entries
(10.39.41.33, 238.105.148.0), RP 10.39.3.111, MBGP/AS 65000, 2d10h/00:05:33
(10.240.112.8, 224.2.0.1), RP 10.9.200.65, MBGP/AS 65001, 00:03:21/00:02:38
(10.69.10.13, 227.37.32.1), RP 10.39.3.92, MBGP/AS 65002, 05:22:20/00:03:32
(10.67.66.18, 234.0.0.1), RP 10.39.3.111, MBGP/AS 65002, 2d10h/00:05:35
(10.67.66.148, 234.0.0.1), RP 10.39.3.111, MBGP/AS 65002, 2d10h/00:05:35
```

The AS listed is the AS in which the originating RP resides according to the MBGP table.

To configure peers in an MSDP mesh group, use the following commands in global configuration mode:

```
ip msdp mesh-group mesh-group-01 192.168.1.2
ip msdp mesh-group mesh-group-01 192.168.1.3
```

The mesh-group-01 parameter is a user-assigned mesh group name. This string is not carried in MSDP messages, so it is significant only to the local router. A router can participate in multiple mesh groups, and the name is used to distinguish among them.

To configure a default peer, use the following command in global configuration mode:

```
ip msdp default-peer 192.168.1.2
```

To filter SA messages for the 229.9.9.9 group received from a particular peer (192.168.1.2) by defining an MSDP import policy, use the following commands in global configuration mode:

```
ip msdp sa-filter in 192.168.1.2 route-map msdp-import
route-map msdp-import permit 10
  match ip address 1
access-list 1 deny 229.9.9.9 0.0.0.0
access-list 1 permit any
```

Care should be taken when using SA filters in transit domains. Preventing SAs from being flooded to other domains can lead to multicast "blackholes."

10.3 Configuring a Dedicated RPF Table

In IOS, the following four routing tables are used for RPF:

- Unicast routing table (which includes tables for each unicast routing protocol)
- MBGP routing table
- DVMRP routing table
- Static mroute table

When performing an RPF check, the software searches each table to find the longest-match prefix. If it finds a matching prefix in more than one of the tables, it uses the protocol preference (known as administrative distance in IOS software) to determine the route that is used for RPF.

It is important to note that the RPF route-selection process is different from that of unicast routing. With unicast routing, the longest-match prefix is chosen even if a matching prefix with a shorter mask exists from a more preferred protocol. In the case of RPF, the same selection criteria are used on each of the four routing tables, and then the route from the most preferred protocol is selected regardless of mask length. For example, the route 10.0.0.0/14, learned via MBGP, will be selected over the route 10.0.0.0/16, learned via BGP, when performing an RPF check. Unlike the unicast routing table, the longest-match prefix rule does not take effect because administrative distance is considered prior to prefix length.

Table 10-1 lists the default protocol preferences (lower preference values are more preferred).

Table 10-1. Routing Preferences

Protocol	Preference
Directly connected	0
Static mroute	0
DVMRP	0
Static route	1
External MBGP	20
External BGP	20
OSPF	110
IS-IS	115
RIP	120
Local MBGP	200
Internal MBGP	200
Local BGP	200
Internal BGP	200

You can override the default preference for each protocol. For example, to change the preference value for Internal MBGP to 20, use the following commands in global configuration mode:

```
router bgp 65005
    distance bgp 20 20 200
```

Chapter 11. Case Study: Service Provider Native Deployment

In this chapter, we combine all of the concepts we have discussed in previous chapters and describe a working model of IMR. Throughout this chapter, we illustrate the exact blueprint that is typically used by the world's largest ISPs when deploying native multicast.

Juniper Networks and Cisco Systems router configurations for all router roles in this example network are provided. Additionally, configurations are provided for the routers that a customer can use to connect to the example network.

While our example network supports both ASM and SSM, most of the "heavy lifting" of this design involves ASM support. In fact, the configuration needed to enable SSM in this network involves the addition of only one or two commands in each router, which illustrates how current ASM networks require so little to add support for SSM. At the end of this chapter, configurations for an SSM-only domain are provided for comparison.

11.1 Network Architecture

In the previous chapters, we described most of the design options that are possible. In this case study, the focus is on what is recommended. This practical architecture reflects all of the best current practices of deployment on the Internet.

11.1.1 PIM-SM

PIM-SM is the multicast routing protocol used in this network. The PIM-SM domain contains five statically mapped anycast RPs. As is the case in most native ISP deployments, static RP mapping is selected over BSR and auto-RP because it provides maximum simplicity. Anycast delivers RP load balancing and redundancy.

The number of RPs deployed in an ISP network typically ranges from four to eight. Fewer than four RPs in a domain may not supply enough redundancy or load balancing. More than eight RPs adds extra administrative burden with little benefit.

11.1.1.1 RP Placement

To reduce the potential for suboptimal routing on the RPT, the routers selected as RPs should be well connected and in the core of the network. Because the RPT is usually short-lived, it is not essential to have centrally located RPs, but it makes more sense. In our example network, one core router from each of the five largest hub sites is chosen as RP. The names and unique IP addresses of the loopbacks of these routers are as follows:

- NY-RP: 10.1.1.1/32
- Atlanta-RP: 10.1.1.2/32
- Chicago-RP: 10.1.1.3/32
- Denver-RP: 10.1.1.4/32
- LA-RP: 10.1.1.5/32
- Anycast RP address: 10.1.1.100

PIM-SM is configured on all of the nonmanagement interfaces of all routers in the network. PIM borders are configured on all customer-facing links to prevent certain multicast traffic from leaking onto the service provider network. This type of traffic includes protocol and administratively scoped addresses.

Customers of this ISP have the choice of using the provider's RP or their own RP. Likewise, BGP customers may elect to run MBGP with the provider.

11.1.2 IGP

In our example network, configuration for IS-IS is provided with the equivalent OSPF configuration shown in italics. All routers in the domain are assumed to be level 2-only routers in the same area with congruent unicast and multicast topologies. Likewise, a single backbone area is the only OSPF area in the network. The IGP carries routing information for only the loopback and network-facing (that is, noncustomer-facing) interfaces of the ISP's routers. The anycast RP address is also carried by the IGP.

11.2 ISP Router Configurations

This section shows relevant configurations for Juniper Networks and Cisco Systems routers acting as both RPs and non-RPs. In our example network, NY-RP and LA-nonRP are Juniper Networks routers, while LA-RP and NY-nonRP are Cisco Systems routers. On the Juniper Networks non-RP routers, interfaces so-0/0/0 and so-0/0/1 are backbone links, while interface t3-1/0/0 connects to a customer. On the Cisco non-RP routers, interfaces POS0/0/0 and POS0/0/1 are backbone links, while Serial1/0/0 connects to a customer.

To reduce repetition, MBGP, IS-IS, and OSPF configurations are shown only for the RP routers. Configuration for the non-RP routers would look the same.

11.2.1 ISP RP Configuration: Juniper Networks

This configuration describes a Juniper Networks router acting as an RP in a typical service provider's network.

```
system {
    host-name NY-RP;
}
interfaces {
    so-0/0/0 {
        unit 0 {
            description "Backbone Link";
            family inet;
            family iso;
      }
  }
  lo0 {
      unit 0 {
          family inet {
              address 10.1.1.1/32 {     /*  Unique IP Address  */
                  primary;
              }
              address 10.1.1.100/32;    /*  Anycast RP Address  */
          }
          family iso {
              address 49.0001.0100.0100.1001.00; /* ISO Address */
          }
      }
  }
}
protocols {
    sap;                          /* Listen to SDR announcements */
    bgp {
        family inet {
            unicast;                      /* SAFI=1  */
            multicast;            /* SAFI=2  */
        }
        export static-connected; /* Redistribute static and connected */
        group IBGP-Peers {       /* routes into BGP */
            type internal;
            local-address 10.1.1.1;
            neighbor 10.1.1.x;
        }
        group BGP-Customers {
            type external;
            neighbor 10.2.2.x {
                peer-as 65001;
            }
        }
    }
    isis {
        multicast-topology;               /* M-ISIS */
        level 1 disable;
```

11.3 Customer Router Configurations

This section shows relevant router configurations for customers of our example ISP. Configurations for a customer who does not operate his own RP and uses the ISP's RP are shown first. Configuration for a customer RP follows.

In the customer RP scenario shown in sections 11.3.3 and 11.3.4, a single statically mapped RP is used. Configuration for a non-RP router in a customer domain with an RP is not shown. In that case, the configuration looks identical to that found in sections 11.3.1 and 11.3.2 with the exception of the IP address in the RP mapping command.

These configurations assume that the customer uses the standard unicast routing table for RPF. To populate a separate routing table for RPF, the same configuration shown for the ISP's RPs can be used.

11.3.1 Customer Without RP Configuration: Juniper Networks

This configuration describes a Juniper Networks router acting as a non-RP in a typical customer network.

```
interfaces {
    t3-1/0/0 {
        unit 0 {
            description "To ISP";
        }
    }
}
protocols {
   sap;          /* Listen to SDR announcements */
   pim {
        rp {
           bootstrap-import block-bsr; /* Prevents BSR messages from */
           bootstrap-export block-bsr; /* entering or leaving router */
           static {
               address 10.1.1.100;     /* RP address of provider */
           }
        }
        interface all {      /* Enable PIM-SM on all interfaces */
            mode sparse;
        }
        interface fxp0.0 {   /* ... except management interface */
            disable;
        }
   }
}
routing-options {
    multicast {
        scope SGI-Dogfight {
            prefix 224.0.1.2/32;
            interface all;
        }
        scope RWHOD {
            prefix 224.0.1.3/32;
            interface all;
        }
        scope SVRLOC {
            prefix 224.0.1.22/32;
            interface all;
        }
        scope MICROSOFT-DS {
            prefix 224.0.1.24/32;
            interface all;
        }
        scope SVRLOC-DA {
            prefix 224.0.1.35/32;
            interface all;
```

11.4 SSM-Only Domain

Creating an SSM-only domain is a much simpler task because an RP-based infrastructure is not necessary. This option is likely to be attractive for network operators with limited multicast experience. Deploying an SSM-only domain requires nothing more than enabling PIM-SM on interfaces, defining the SSM address range, and creating boundaries between other domains.

Once again, these configurations assume the customer uses the standard unicast routing table for RPF. To populate a separate routing table for RPF, the same configuration shown for the ISP's RPs can be used.

11.4.1 SSM-Only Configuration: Juniper Networks

This configuration describes a Juniper Networks router in a typical customer network that supports SSM-only.

```
interfaces {
    t3-1/0/0 {
        unit 0 {
            description "To ISP";
        }
    }
}
protocols {
    pim {
        rp {
           bootstrap-import block-bsr;  /* Prevents BSR messages from */
           bootstrap-export block-bsr;  /* entering or leaving router */
        }
        interface all {        /* Enable PIM-SM on all interfaces */
            mode sparse;
        }
        interface fxp0.0 {     /* ... except management interface */
            disable;
        }
    }
}
routing-options {
    multicast {
        scope SGI-Dogfight {
            prefix 224.0.1.2/32;
            interface all;
        }
        scope RWHOD {
            prefix 224.0.1.3/32;
            interface all;
        }
        scope SVRLOC {
            prefix 224.0.1.22/32;
            interface all;
        }
        scope MICROSOFT-DS {
            prefix 224.0.1.24/32;
            interface all;
        }
        scope SVRLOC-DA {
            prefix 224.0.1.35/32;
            interface all;
        }
        scope AutoRP-Announce {
            prefix 224.0.1.39/32;
            interface all;
        }
        scope AutoRP-Discovery {
            prefix 224.0.1.40/32;
            interface all;
```

Chapter 12. Management Tools for Multicast Networks

This chapter provides an overview of the various tools available to monitor and troubleshoot multicast-enabled internets. Some of these tools run on the router itself, and some run on a separate host. Similar to unicast monitoring and troubleshooting tools, the multicast counterparts help to identify loss of connectivity and provide insight into the cause of that loss of connectivity.

12.1 SNMP MIBs

The Simple Network Management Protocol (SNMP) provides a means to help manage networks. SNMP is essentially the only network management protocol in use today.

SNMP uses Management Information Bases (MIBs) to establish a consistent language between all SNMP-speaking devices. A MIB is written in human-readable MIB language and is saved as a text file. Network management administrators must compile the MIBs that interest them into the proprietary binary format of their network management software. Once this is done, users can browse the MIBs using the network management software.

Public MIBs exist as RFCs. A MIB RFC provides a brief description of the intended use of the MIB and includes the contents of the MIB. If you need to compile a MIB in your network management software, you can copy and paste the RFC text, being careful to delete each page's header and footer.

You can usually find versions of public MIBs on Web sites in a ready-to-compile form. One such site is http://www.aciri.org/fenner/mibs/mib-index.html. Router vendors support a subset of the public MIBs available, depending on the protocols that run on their routers. Check your router vendor's technical documentation to determine the MIBs it supports. Juniper Networks lists the public MIBs it supports in its Installation and System Management guide.

In addition to public MIBs, many vendors provide proprietary MIBs for information specific to their products. Juniper Networks' proprietary MIBs are available at http://www.juniper.net/techpubs/mibs.html.

An SNMP daemon runs on each router, and SNMP software runs on one or more network management systems (NMS). SNMP MIBs contain two types of entities, namely traps and objects. SNMP traps are pushed from the router to the NMS, without the NMS requesting the information. SNMP traps are sent when a specific event occurs (for example, a link or adjacency in the networks goes down).

MIB objects can be polled from the NMS. They can either be polled manually—a practice known as MIB browsing—or they can be polled periodically and used for graphs of historical data. You can find free NMS software on the Web. The authors' favorite is Active SNMP, which is available at http://www.cscare.com/ActiveSNMP/.

When an NMS sends an SNMP request to a router, it identifies the MIB object in which it is interested by the object identifier (OID). OIDs are hierarchical, with each level separated by a dot in the standard notation. Each named level of hierarchy is assigned a number. The following lists where various MIBs are rooted:

- Public MIBs: .iso.org.dod.internet.mgmt.mib-2 = .1.3.6.1.2.1
- Proprietary MIBs: .iso.org.dod.internet.private.enterprises = .1.3.6.1.4.1
- Experimental MIBs: .iso.org.dod.internet.experimental = .1.3.6.1.3

To know where a MIB fits into the hierarchy, read the MODULE-IDENTITY section of the MIB. For example the MODULE-IDENTITY of the ipMRouteStdMIB MIB is the following:

```
ipMRouteStdMIB MODULE-IDENTITY
    LAST-UPDATED "200009220000Z" -- September 22, 2000
    ORGANIZATION "IETF IDMR Working Group"
    CONTACT-INFO
            " Dave Thaler
              Microsoft Corporation
              One Microsoft Way
              Redmond, WA  98052-6399
              US
```

12.2 The mtrace Facility

Currently mtrace is specified in an IETF draft titled "A 'traceroute' facility for IP Multicast." On a UNIX host, you can use man mtrace to learn about the specific mtrace application installed on the system. The mtrace utility is intended to be used for assessing IP multicast connectivity problems. It provides a method for troubleshooting multicast problems similar to what standard IP traceroute does for unicast connectivity problems.

The mtrace utility uses the IGMP protocol number. A host initiates an mtrace query specifying a source hostname or IP address. The mtrace query is passed along hop-by-hop using each router's RPF route for the source address. Along the way, information is collected about hop addresses, packet counts, and routing error conditions.

The final hop (either a router directly connected to the source or a router that has no RPF route for the source) returns an mtrace response to the host that initiated the mtrace query. The standard mtrace query is sent to the ALL-ROUTERS link-local multicast group and has a TTL of 1.

Optionally, a receiver address and group address can be specified in the mtrace query. If no receiver address is specified, the address of the host that generated the query is used. If a receiver address is specified, mtrace finds the router connected directly to that receiver. It is able to locate the last-hop router by sending the mtrace query to the group address of interest.

This mechanism for finding the last-hop router requires that the intended receiver has joined the ASM group or subscribed to two SSM channels. These channels consist of one with the local host being the source of the SSM channel and the second with the specified source as the source of the SSM channel. Alternatively, the last-hop router can be specified in the mtrace command line.

If there is no response from the initial mtrace query, the application automatically switches to hop-by-hop mode. In hop-by-hop mode, tracing queries are started with a maximum hop count of 1 and are incremented until the last-hop router is reached or there is no response.

It is also possible to specify the number of hops an mtrace query can travel. When the maximum number of hops is specified, the mtrace query travels along the RPF patch from the receiver to the source incrementing the hop count at each router. Once the maximum number of hops has been reached, the router returns the mtrace response as if it were directly connected to the source.

Specifying the maximum hop count is useful because it enables a partial trace if the mtrace query is blackholed because a router along the path does not support mtrace or because an RPF route exists but some other problem (such as a firewall filter) is causing an outage.

Each router inserts the following information into the mtrace query:

- IP address of the hop
- TTL required to forward
- Flags to indicate routing errors
- Total number of packets on the incoming interface
- Total number of packets on the outgoing interfaces
-

12.3 The MSDP Traceroute Facility

The MSDP traceroute utility traces the control path for MSDP Source-Active messages from any MSDP-speaking router to the originating RP for the message. This is accomplished by each router along the path forwarding the MSDP traceroute packet to its RPF peer for the originating RP address.

The MSDP traceroute utility is currently described in an Internet-Draft (draft-ietf-msdp-traceroute-06.txt).

Chapter 13. Other Related Topics

This chapter provides a brief overview of various protocols that are not core to current deployments of IMR but may be of interest in future deployments or as potential Trivial Pursuit questions. We provide a brief overview of where the protocol fits in the grand scheme, how it operates, and where to find the specs.

13.1 Border Gateway Multicast Protocol (BGMP)

BGMP is the proverbial promised land for IMR. It is specified in an Internet-Draft titled "Border Gateway Multicast Protocol (BGMP): Protocol Specification" (draft-ietf-bgmp-spec.txt). BGMP requires that each multicast group be associated with a single root domain. BGMP-speaking routers require a mechanism that maps a group address to a next hop toward that group's root domain. Multicast Address Set Claim Protocol (MASC) is one mechanism that can be used to create such a mapping. MASC dynamically distributes information about the associations of group addresses to root domains. This information is stored in the G-RIB table. Using the information in the G-RIB, BGMP builds shared trees for active groups and then enables each domain to build source-based trees.

BGMP uses TCP (port 264) as its transport protocol. Like BGP, BGMP is an incremental protocol, meaning that routing updates are only sent once and are explicitly withdrawn (periodic refresh of state is not required).

One roadblock that has slowed the deployment of BGMP is the complexities involved with the G-RIB and a dynamic protocol (such as MASC) used to fill it with information. The root domain is encoded in IPv6 multicast groups, so BGMP deployment in an all IPv6 Internet will be a much easier task. A mechanism similar to the one described for IPv6 could be devised for IPv4 multicast addresses. An Internet-Draft, "Unicast-Prefix-Based IPv6 Multicast Addresses" (draft-ietf-ipngwg-uni-based-mcast-03.txt), describes the format of IPv6 multicast addresses with encoded root domains. The format is as follows:

```
|   8    |  4 |  4 |   8    |    8   |       64       |    32    |
+--------+----+----+--------+--------+----------------+----------+
|11111111|00PT|Scop|00000000|  Plen  | Network Prefix | Group ID |
+--------+----+----+--------+--------+----------------+----------+
```

P = 0 indicates a multicast address that is not assigned based on the network prefix. P = 1 indicates a multicast address that is assigned based on the network prefix. The setting of the T bit is defined in RFC 2373. When P = 1, the Plen field indicates the actual length of the network prefix portion of the address, and the Network Prefix field identifies the unicast subnet that owns the multicast group.

13.2 Multicast Address Set Claim Protocol (MASC)

MASC, which is specified in RFC 2909, is used to declare a group prefix as being owned by a domain. The multicast groups that are associated with a domain are injected into MBGP with the AFI/SAFI set to 1/4 and are used to populate a G-RIB that can be used by BGMP to construct interdomain shared trees.

13.3 Bi-Directional PIM (Bi-Dir PIM)

Bi-Directional PIM (Bi-Dir PIM) is another mode of operation for PIM (in contrast to sparse mode and dense mode). When a group is forwarded based on Bi-Dir rules, data packets are routed along a bidirectional shared tree to the RP for the group. Bi-Dir PIM is designed for multicast applications with many sources, where all sources and receivers are in the same PIM domain. It is not intended to be used for IMR.

Bi-Dir PIM does not keep (S,G) Join state, which reduces the overall amount of state that is kept on routers throughout the domain. Sources join the shared tree (even if they do not want to receive traffic for the group) and send traffic upstream. Bi-Dir PIM is specified in an Internet-Draft, "Bi-directional Protocol Independent Multicast (BIDIR-PIM)" (draft-ietf-pim-bidir-03.txt).

13.4 Multicast Data Packets and Real-Time Transport Protocol (RTP)

Real-Time Transport Protocol (RTP), which is defined in RFC 1889, provides host-to-host transport over IP networks that is suitable for real-time applications such as video and audio streaming. The underlying IP forwarding mechanism can be either unicast or multicast. One of the common uses of IP multicast is to transport live and scheduled multimedia traffic, where RTP is often used as the transport layer protocol for multicast data packets. RTP is commonly run on top of UDP, with both protocols sharing part of the transport layer responsibilities. Using RTP over UDP instead of just UDP for transporting real-time data has several advantages, including the following:

- Payload type identification
- Sequence numbering
- Time stamping
- Delivery monitoring

RTP in itself does not provide any guarantee of timely delivery. It relies on network layer mechanisms to provide differentiated quality of service.

Appendix A. IGMP Packet Formats

This appendix includes the packet formats from the specifications of the three versions of IGMP.

A.1 IGMP Version 3 Packet Formats

IGMPv3 is the result of the collective work of Brad Cain, Steve Deering, Bill Fenner, Isidor Kouvelas, and Ajit Thyagarajan. The packet formats are described in section 4 of the current specification. Here is section 4 from the specification in its entirety (it has been reformatted for consistency).

4 Message Formats

IGMP messages are encapsulated in IPv4 datagrams, with an IP protocol number of 2. Every IGMP message described in this document is sent with an IP Time-to-Live of 1 and carries an IP Router Alert option [RFC-2113] in its IP header.

There are two IGMP message types of concern to the IGMPv3 protocol described in this document:

- 0x11: Membership Query
- 0x22: Version 3 Membership Report

An implementation of IGMPv3 MUST also support the following three message types, for interoperation with previous versions of IGMP:

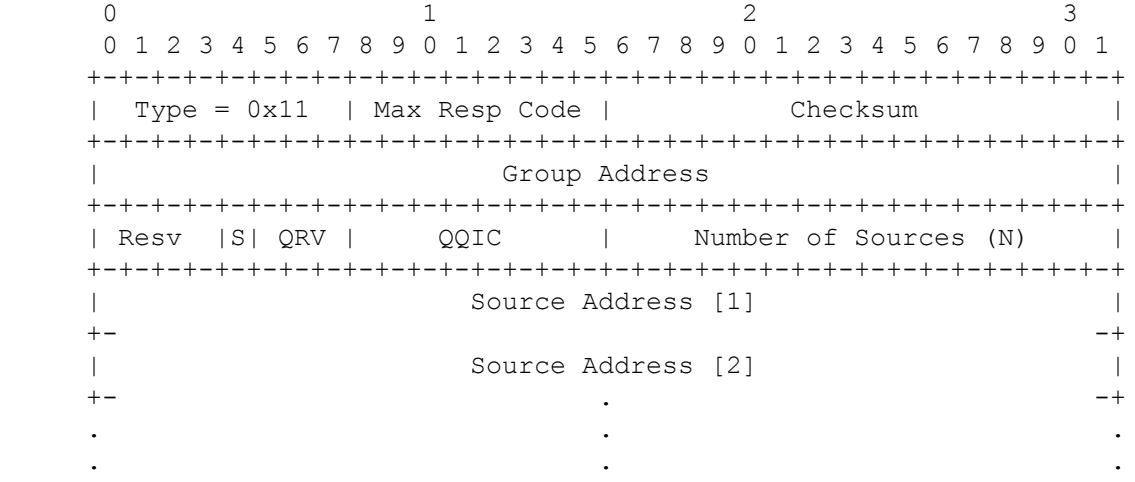
- 0x12: Version 1 Membership Report [RFC-1112]
- 0x16: Version 2 Membership Report [RFC-2236]
- 0x17: Version 2 Leave Group [RFC-2236]

Unrecognized message types MUST be silently ignored. Other message types may be used by newer versions or extensions of IGMP, by multicast routing protocols, or for other uses.

In this document, unless otherwise qualified, the capitalized words "Query" and "Report" refer to IGMP Membership Queries and IGMP Version 3 Membership Reports, respectively.

4.1 Membership Query Message

Membership Queries are sent by IP multicast routers to query the multicast reception state of neighboring interfaces. Queries have the following format:



A.2 IGMP Version 2 Packet Formats

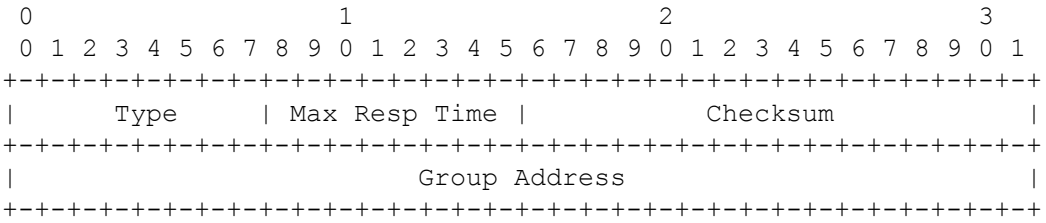
IGMPv2 is the result of work by Bill Fenner. The packet formats are described in section 2 of RFC 2236. Here is section 2 from the specification in its entirety (it has been reformatted for consistency).

2 Introduction

The Internet Group Management Protocol (IGMP) is used by IP hosts to report their multicast group memberships to any immediately neighboring multicast routers. This memo describes only the use of IGMP between hosts and routers to determine group membership.

Routers that are members of multicast groups are expected to behave as hosts as well as routers and may even respond to their own queries. IGMP may also be used between routers, but such use is not specified here.

Like ICMP, IGMP is a integral part of IP. It is required to be implemented by all hosts wishing to receive IP multicasts. IGMP messages are encapsulated in IP datagrams, with an IP protocol number of 2. All IGMP messages described in this document are sent with IP TTL 1 and contain the IP Router Alert option (RFC 2113) in their IP header. All IGMP messages of concern to hosts have the following format:



2.1 Type

There are three types of IGMP messages of concern to the host-router interaction:

- 0x11: Membership Query

  There are two subtypes of Membership Query messages:
  - General Query: Used to learn which groups have members on an attached network

  - Group-Specific Query: Used to learn if a particular group has any members on an attached network

  - These two messages are differentiated by the Group Address. Membership Query messages are referred to simply as "Query" messages.
- 0x16: Version 2 Membership Report
- 0x17: Leave Group

  There is an additional type of message, for backwards-compatibility with IGMPv1:
- 0x12: Version 1 Membership Report

This document refers to Membership Reports simply as "Reports." When no version is specified, the statement applies equally to both versions.

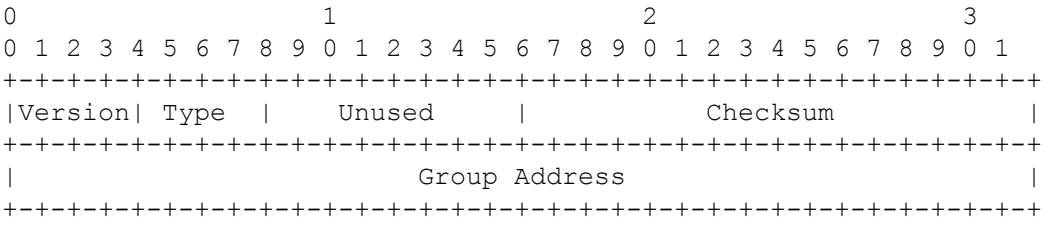
A.3 IGMP Version 1 Packet Formats

IGMPv1 is the result of work completed by Steve Deering. The packet formats are described in Appendix I of RFC 1112. Here is the part of Appendix I from the specification that describes packet formats (it has been reformatted for consistency).

Appendix I. Internet Group Management Protocol (IGMP)

The Internet Group Management Protocol (IGMP) is used by IP hosts to report their host group memberships to any immediately-neighboring multicast routers. IGMP is an asymmetric protocol and is specified here from the point of view of a host rather than a multicast router. (IGMP may also be used, symmetrically or asymmetrically, between multicast routers. Such use is not specified here.)

Like ICMP, IGMP is a integral part of IP. It is required to be implemented by all hosts conforming to level 2 of the IP multicasting specification. IGMP messages are encapsulated in IP datagrams, with an IP protocol number of 2. All IGMP messages of concern to hosts have the following format:



- Version

  This memo specifies version 1 of IGMP. Version 0 is specified in RFC-988 and is now obsolete.

- Type

  There are two types of IGMP message of concern to hosts:

  - Host Membership Query

  - Host Membership Report

- Unused

  Unused field, zeroed when sent, ignored when received.

- Checksum

  The checksum is the 16-bit one's complement of the one's complement sum of the 8-octet IGMP message. For computing the checksum, the checksum field is zeroed.

- Group Address

  In a Host Membership Query message, the Group Address field is zeroed when sent, ignored when received. In a Host Membership Report message, the Group Address field holds the IP host group address of the group being reported.

Appendix B. PIM Packet Formats

This appendix includes the packet formats from the specifications of the two versions of PIM.

B.1 PIM Version 2 Packet Formats

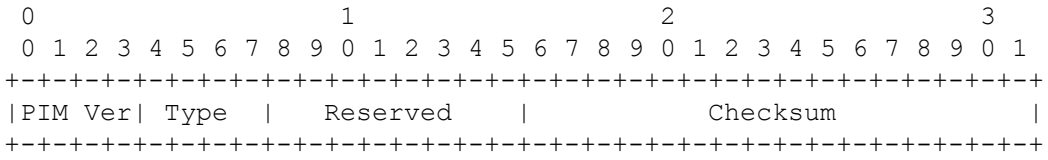
PIMv2 is the result of the collective work of Bill Fenner, Mark Handley, Hugh Holbrook, and Isidor Kouvelas. The packet formats are described in section 4.9 of the current specification, which is included here in its entirety (it has been reformatted for consistency).

4.9 PIM Packet Formats

This section describes the details of the packet formats for PIM control messages.

All PIM control messages have IP protocol number 103. PIM messages are either unicast (e.g., Registers and Register-Stop) or multicast with TTL 1 to the ALL-PIM-ROUTERS group (e.g., Join/Prune, Asserts, etc.). The source address used for unicast messages is a domainwide reachable address; the source address used for multicast messages is the link-local address of the interface on which the message is being sent.

The IPv4 ALL-PIM-ROUTERS group is 224.0.0.13. The IPv6 ALL-PIM- ROUTERS group is ff02::d.



- PIM Ver

  PIM Version number is 2.

- Type

  Types for specific PIM messages. PIM Types are listed in the table that follows:

Message Type	Destination
0 = Hello	Multicast to ALL-PIM-ROUTERS
1 = Register	Unicast to RP
2 = Register-Stop	Unicast to source of Register packet
3 = Join/Prune	Multicast to ALL-PIM-ROUTERS
4 = Bootstrap	Multicast to ALL-PIM-ROUTERS
5 = Assert	Multicast to ALL-PIM-ROUTERS
6 = Graft (used in PIM-DM only)	Multicast to ALL-PIM-ROUTERS
7 = Graft-Ack (used in PIM-DM only)	Unicast to source of Graft packet
8 = Candidate-RP-Advertisement	Unicast to Domain's BSR

- Reserved

  Set to zero on transmission. Ignored upon receipt.

- Checksum

  The checksum is a standard IP checksum, i.e., the 16-bit one's complement of the one's complement sum of

B.2 PIM Version 1 Packet Formats

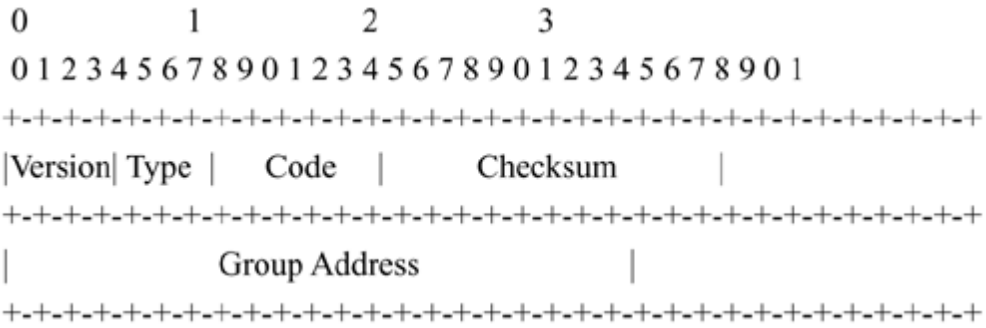
PIMv1 is the result of the collective work of Steve Deering, Deborah Estrin, Dino Farinacci, and Van Jacobsen. The packet formats are described in section 4 in the specification, which is included here in its entirety (it has been reformatted for consistency).

4 Packet Types

RFC-1112 specifies two types of IGMP packets for hosts and routers to convey multicast group membership and reachability information. An IGMP-Host-Query packet is transmitted periodically by routers to ask hosts to report which multicast groups they are members of. An IGMP-Host-Report packet is transmitted by hosts in response to received queries advertising group membership.

This document introduces new types of IGMP packets that are used by PIM routers. The packet format is shown in Figure 1

Figure 1. (Figure 8 in orig.) IGMP packet format



- Version: This memo specifies version 1 of IGMP. Version 0 is specified in RFC-988 and is now obsolete.
- Type: There are five types of IGMP messages:

  - 1 = Host Membership Query

  - 2 = Host Membership Report

  - 3 = Router DVMRP Messages

  - 4 = Router PIM Messages

  - 5 = Trace Messages
- Code: Codes for specific message types. Used only by DVMRP and PIM. PIM codes are:

  - 0 = Router-Query

  - 1 = Register

  - 2 = Register-Stop

  - 3 = Join/Prune

  - 4 = RP-Reachability

Appendix C. MSDP Packet Formats

This appendix includes the packet formats from the latest specification of versions of MSDP.

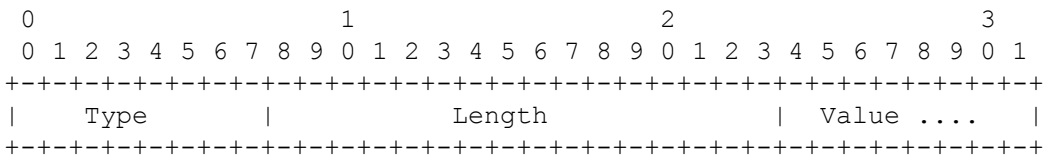
C.1 MSDP Packet Formats

MSDP is the result of the collective work of many individuals. David Meyer and Bill Fenner serve as editors of the specification. The packet formats are described in section 16 of the current specification, which is included here in its entirety (it has been reformatted for consistency).

16 Packet Formats

MSDP messages will be encoded in TLV format. If an implementation receives a TLV that has length that is longer than expected, the TLV SHOULD be accepted. Any additional data SHOULD be ignored.

16.1 MSDP TLV format



- Type (8 bits)

  Describes the format of the Value field.

- Length (16 bits)

  Length of Type, Length, and Value fields in octets. The minimum length required is 4 octets, except for Keepalive messages. The maximum TLV length is 1400.

- Value (variable length)

  Format is based on the Type value. See below. The length of the Value field is Length field minus 3. All reserved fields in the Value field MUST be transmitted as zeros and ignored on receipt.

16.2 Defined TLVs

The following TLV Types are defined:

Code	Type
1	IPv4 Source-Active
2	IPv4 Source-Active Request
3	IPv4 Source-Active Response
4	KeepAlive
5	Notification

Each TLV is described below.

In addition, the following TLV Types are assigned but not described in this memo:

Code	Type
6	MSDP traceroute in progress
7	MSDP traceroute reply

Glossary

This glossary lists key terms, abbreviations, and acronyms with their definitions and indicates whether the term is applicable to a particular protocol, routing environment, or "context."

Term	Context	Description
(*,G) route entry	PIM-SM	Group members join the RP tree for a particular group. This tree is represented by (*,G) multicast route entries along the shared tree branches between the RP and the group members.
ABR	OSPF	area border router—Routers within a nonbackbone area that connect to area 0.
AFI	MBGP	address family identifier—A number referencing network protocols.
AFI	IS-IS	authority and format indicator—A byte in NSAP format used to describe the organization that assigned the address and the meaning of the fields that follow.
anycast	Packet delivery	A method of delivering packets to exactly one member of the anycast group. The specific host to which the packet is delivered cannot be determined by the sender.
anycast RP	PIM-SM	A method in which multiple routers are configured with the same IP address, typically on their loopback interface. This shared address is used in the RP-to-group mapping, which enables multicast groups to have multiple active RPs in a PIM-SM domain for the same group range.
AS	BGP	autonomous system—A collection of routers, typically operated by a single administrative function, coordinated to implement the same routing policy.
ASM	IP multicast	Any-Source Multicast—One-to-many and many-to-many communications model outlined in RFC 1112. ASM is the original vision of multicast.
ASP	Networking	application service provider—An organization that provides content-hosting services.
ASSERT message	PIM	Provides a mechanism for avoiding a condition where multiple routers exist for a LAN and more than one router forwards the same multicast data packets to the LAN.
ATM	Link layer	Asynchronous Transfer Mode—A circuit-switched, link-layer protocol.
auto-RP	PIM-SM	A method for dynamically learning of the RP. Originally proprietary to Cisco Systems, it is now fully supported by Juniper Networks.
BGMP	Multicast routing protocols	Border Gateway Multicast Protocol—Interdomain multicast routing protocol that is expected to be implemented in

Bibliography

Albanna, Z., K. Almeroth, D. Meyer, M. Schipper. "*IANA Guidelines for IPv4 Multicast Address Assignments,*" RFC 3171, August 2001.

Bates, T., Y. Rekhter, R. Chandra, D. Katz. "*Multiprotocol Extensions for BGP-4,*" RFC 2858, June 2000.

Bhattacharyya, Supratik, Christophe Diot, Leonard Giuliano, Rob Rockell, John Meylor, David Meyer, Greg Shepherd, Brian Haberman. "*An Overview of Source-Specific Multicast (SSM) Deployment,*" Work in progress.

Cain, Brad, Steve Deering, Bill Fenner, Isidor Kouvelas, and Ajit Thyagarajan. "*Internet Group Management Protocol, Version 3,*" work in progress.

Callon, R. "*Use of OSI IS-IS for Routing in TCP/IP and Dual Environments,*" RFC 1195, December 1990.

Cisco Product Documentation. http://www.cisco.com/univercd/home/home.htm.

Deering, S. "*Host Extensions for IP Multicasting,*" RFC 1112, August 1989.

Deering, S., Deborah Estrin, Dino Farinacci, Van Jacobson, Ahmed Helmy, David Meyer, Liming Wei. "*Protocol Independent Multicast Version 2 Dense Mode Specification,*" work in progress.

Doyle, Jeff. Routing TCP/IP, Volume I: A Detailed Examination of Interior Routing Protocols, Cisco Press, 1998.

Estrin, D., D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei. "*Protocol Independent Multicast Sparse Mode (PIM-SM): Protocol Specification,*" RFC 2362, June 1998.

Fenner, Bill, and Dave Thaler. "*Multicast Source Discovery protocol MIB,*" work in progress.

Fenner, Bill, Mark Handley, Hugh Holbrook, and Isidor Kouvelas. "*Protocol Independent Multicast—Sparse Mode (PIM-SM): Protocol Specification (Revised),*" work in progress.

Fenner, Bill, Mark Handley, Roger Kermode, and David Thaler. "*Bootstrap Router (BSR) Mechanism for PIM Sparse Mode,*" work in progress.

Fenner, W. "*Internet Group Management Protocol, Version 2,*" RFC 2236, November 1997.

Giuliano, Leonard. "*Deploying Native Multicast across the Internet,*" whitepaper.

Halabi, Bassam. Internet Routing Architectures: The Definitive Resource for Internetworking Design Alternatives and Solutions. Cisco Press, 1997.

Handley, M., and V. Jacobson. "*SDP: Session Description Protocol,*" RFC 2327, April 1998.

Handley, M., C. Perkins, and E. Whelan. "*Session Announcement Protocol,*" RFC 2974, October 2000.

Handley, Mark, Van Jacobson, and Colin Perkins. "*SDP: Session Description Protocol,*" work in progress.

Holbrook, H. and B. Cain. "*Source-Specific Multicast for IP,*" work in progress.

Holbrook, H. and B. Cain. "*Using IGMPv3 For Source-Specific Multicast,*" work in progress.

JUNOS Internet Software Documentation. http://www.juniper.net/techpubs/software.html.

Katz, Dave. "*OSPF and IS-IS: A Comparative Anatomy,*" In Proceedings of NANOG, June 2000.

About the Authors

Brian M. Edwards

Leonard A. Giuliano

Brian R. Wright

Brian M. Edwards

Brian is the customer support engineer for premium accounts in Juniper Networks Technical Assistance Center (JTAC). On a daily basis, he troubleshoots problems affecting the largest ISP networks in the world. He is the designated subject matter expert for multicast routing at Juniper Networks and has completed the highest levels of Cisco Systems and Juniper Networks certification programs (CCIE #6187 and JNCIE #9). He earned a B.S. in computer engineering from the University of Florida in 1997.

Leonard A. Giuliano

Leonard is a systems engineer for Juniper Networks, supporting large ISPs in the architecture, design, and operation of backbone networks. He specializes in IP multicast, IP core routing, and traffic engineering. Leonard previously worked as a multicast architect for SprintLink, the world's first native multicast-enabled Internet backbone. He has coauthored many published documents on multicast networking including the IETF's SSM Framework specification. He is also a member of the IETF's MSDP Protocol Design Team and is a Juniper Networks Certified Internet Specialist (JNCIS). He earned his B.S.E. in electrical engineering from Duke University in 1997.

Brian R. Wright

Brian is a technical documentation specialist currently developing the system documentation set for the MasterCard Debit System (MDS) application. He wrote the operations guide for the Tandem computer-based point-of-sale (POS) system of Bank One and the system document for the Exxon retail store POS system. He also helped compose the documentation set for MPACT EDI systems messaging and EDI translation software, and was senior writer/editor for EDS corporate communications, automotive product engineer for American Motors, senior design engineer for gas turbine engine accessories at Williams International, and project engineer at the Wayne State University Biomechanics Department. He has worked as a freelance journalist and writer and is a member of the Society for Technical Communication. Brian earned his BSME at Wayne State University in 1975.

[M] [P]

[M] [P]

PIM-SM