Internet Engineering Task Force (IETF)
Request for Comments: 6513
Category: Standards Track
ISSN: 2070-1721

E. Rosen, Ed.
Cisco Systems, Inc.
R. Aggarwal, Ed.
Juniper Networks

February 2012

Multicast in MPLS/BGP IP VPNs

Abstract

In order for IP multicast traffic within a BGP/MPLS IP VPN (Virtual Private Network) to travel from one VPN site to another, special protocols and procedures must be implemented by the VPN Service Provider. These protocols and procedures are specified in this document.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6513.

Table of Contents

1. Introduction

[RFC4364] specifies the set of procedures that a Service Provider (SP) must implement in order to provide a particular kind of VPN service ("BGP/MPLS IP VPN") for its customers. The service described therein allows IP unicast packets to travel from one customer site to another, but it does not provide a way for IP multicast traffic to travel from one customer site to another.

This document extends the service defined in [RFC4364] so that it also includes the capability of handling IP multicast traffic. This requires a number of different protocols to work together. The document provides a framework describing how the various protocols fit together, and it also provides a detailed specification of some of the protocols. The detailed specification of some of the other protocols is found in preexisting documents or in companion documents.

A BGP/MPLS IP VPN service that supports multicast is known as a "Multicast VPN" or "MVPN".

Both this document and its companion document [MVPN-BGP] discuss the use of various BGP messages and procedures to provide MVPN support. While every effort has been made to ensure that the two documents are consistent with each other, it is possible that discrepancies have crept in. In the event of any conflict or other discrepancy with respect to the use of BGP in support of MVPN service, [MVPN-BGP] is to be considered to be the authoritative document.

Throughout this document, we will use the term "VPN-IP route" to mean a route that is either in the VPN-IPv4 address family [RFC4364] or in the VPN-IPv6 address family [RFC4659].

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Overview

2.1. Optimality vs. Scalability

In a "BGP/MPLS IP VPN" [RFC4364], unicast routing of VPN packets is achieved without the need to keep any per-VPN state in the core of the SP's network (the "P routers"). Routing information from a particular VPN is maintained only by the Provider Edge routers (the "PE routers", or "PEs") that attach directly to sites of that VPN. Customer data travels through the P routers in tunnels from one PE to another (usually MPLS Label Switched Paths, LSPs), so to support the

VPN service the P routers only need to have routes to the PE routers. The PE-to-PE routing is optimal, but the amount of associated state in the P routers depends only on the number of PEs, not on the number of VPNs.

However, in order to provide optimal multicast routing for a particular multicast flow, the P routers through which that flow travels have to hold state that is specific to that flow. A multicast flow is identified by the (source, group) tuple where the source is the IP address of the sender and the group is the IP multicast group address of the destination. Scalability would be poor if the amount of state in the P routers were proportional to the number of multicast flows in the VPNs. Therefore, when supporting multicast service for a BGP/MPLS IP VPN, the optimality of the multicast routing must be traded off against the scalability of the P routers. We explain this below in more detail.

If a particular VPN is transmitting "native" multicast traffic over the backbone, we refer to it as an "MVPN". By "native" multicast traffic, we mean packets that a Customer Edge router (a "CE router" or "CE") sends to a PE, such that the IP destination address of the packets is a multicast group address, the packets are multicast control packets addressed to the PE router itself, or the packets are IP multicast data packets encapsulated in MPLS.

We say that the backbone multicast routing for a particular multicast group in a particular VPN is "optimal" if and only if all of the following conditions hold:

- When a PE router receives a multicast data packet of that group from a CE router, it transmits the packet in such a way that the packet is received by every other PE router that is on the path to a receiver of that group;

- The packet is not received by any other PEs;

- While in the backbone, no more than one copy of the packet ever traverses any link.

- While in the backbone, if bandwidth usage is to be optimized, the packet traverses minimum cost trees rather than shortest path trees.

Optimal routing for a particular multicast group requires that the backbone maintain one or more source trees that are specific to that flow. Each such tree requires that state be maintained in all the P routers that are in the tree.

Potentially, this would require an unbounded amount of state in the P routers, since the SP has no control of the number of multicast groups in the VPNs that it supports. The SP also doesn't have any control over the number of transmitters in each group, nor over the distribution of the receivers.

The procedures defined in this document allow an SP to provide multicast VPN service, without requiring the amount of state maintained by the P routers to be proportional to the number of multicast data flows in the VPNs. The amount of state is traded off against the optimality of the multicast routing. Enough flexibility is provided so that a given SP can make his own trade-offs between scalability and optimality. An SP can even allow some multicast groups in some VPNs to receive optimal routing, while others do not. Of course, the cost of this flexibility is an increase in the number of options provided by the protocols.

The basic technique for providing scalability is to aggregate a number of customer multicast flows onto a single multicast distribution tree through the P routers. A number of aggregation methods are supported.

The procedures defined in this document also accommodate the SP that does not want to build multicast distribution trees in his backbone at all; the ingress PE can replicate each multicast data packet and then unicast each replica through a tunnel to each egress PE that needs to receive the data.

2.1.1. Multicast Distribution Trees

This document supports the use of a single multicast distribution tree in the backbone to carry all the multicast traffic from a specified set of one or more MVPNs. Such a tree is referred to as an "Inclusive Tree". An Inclusive Tree that carries the traffic of more than one MVPN is an "Aggregate Inclusive Tree". An Inclusive Tree contains, as its members, all the PEs that attach to any of the MVPNs using the tree.

With this option, even if each tree supports only one MVPN, the upper bound on the amount of state maintained by the P routers is proportional to the number of VPNs supported rather than to the number of multicast flows in those VPNs. If the trees are unidirectional, it would be more accurate to say that the state is proportional to the product of the number of VPNs and the average number of PEs per VPN. The amount of state maintained by the P routers can be further reduced by aggregating more MVPNs onto a single tree. If each such tree supports a set of MVPNs, (call it an "MVPN aggregation set"), the state maintained by the P routers is

proportional to the product of the number of MVPN aggregation sets and the average number of PEs per MVPN. Thus, the state does not grow linearly with the number of MVPNs.

However, as data from many multicast groups is aggregated together onto a single Inclusive Tree, it is likely that some PEs will receive multicast data for which they have no need, i.e., some degree of optimality has been sacrificed.

This document also provides procedures that enable a single multicast distribution tree in the backbone to be used to carry traffic belonging only to a specified set of one or more multicast groups, from one or more MVPNs. Such a tree is referred to as a "Selective Tree" and more specifically as an "Aggregate Selective Tree" when the multicast groups belong to different MVPNs. By default, traffic from most multicast groups could be carried by an Inclusive Tree, while traffic from, e.g., high bandwidth groups could be carried in one of the Selective Trees. When setting up the Selective Trees, one should include only those PEs that need to receive multicast data from one or more of the groups assigned to the tree. This provides more optimal routing than can be obtained by using only Inclusive Trees, though it requires additional state in the P routers.

2.1.2. Ingress Replication through Unicast Tunnels

This document also provides procedures for carrying MVPN data traffic through unicast tunnels from the ingress PE to each of the egress PEs. The ingress PE replicates the multicast data packet received from a CE and sends it to each of the egress PEs using the unicast tunnels. This requires no multicast routing state in the P routers at all, but it puts the entire replication load on the ingress PE router and makes no attempt to optimize the multicast routing.

2.2. Multicast Routing Adjacencies

In BGP/MPLS IP VPNs [RFC4364], each CE (Customer Edge) router is a unicast routing adjacency of a PE router, but CE routers at different sites do not become unicast routing adjacencies of each other. This important characteristic is retained for multicast routing -- a CE router becomes a multicast routing adjacency of a PE router, but CE routers at different sites do not become multicast routing adjacencies of each other.

We will use the term "C-tree" to refer to a multicast distribution tree whose nodes include CE routers. (See Section 3.1 for further explication of this terminology.)

The multicast routing protocol on the PE-CE link is presumed to be PIM (Protocol Independent Multicast) [PIM-SM]. Both the ASM (Any-Source Multicast) and the SSM (Source-Specific Multicast) service models are supported. Thus, both shared C-trees and source-specific C-trees are supported. Shared C-trees may be unidirectional or bidirectional; in the latter case, the multicast routing protocol is presumed to be the BIDIR-PIM [BIDIR-PIM] "variant" of PIM-SM. A CE router exchanges "ordinary" PIM control messages with the PE router to which it is attached.

Support for PIM-DM (Dense Mode) is outside the scope of this document.

The PEs attaching to a particular MVPN then have to exchange the multicast routing information with each other. Two basic methods for doing this are defined: (1) PE-PE PIM and (2) BGP. In the former case, the PEs need to be multicast routing adjacencies of each other. In the latter case, they do not. For example, each PE may be a BGP adjacency of a route reflector (RR) and not of any other PEs.

In order to support the "Carrier’s Carrier" model of [RFC4364], mLDP (Label Distribution Protocol Extensions for Multipoint Label Switched Paths) [MLDP] may also be supported on the PE-CE interface. The use of mLDP on the PE-CE interface is described in [MVPN-BGP]. The use of BGP on the PE-CE interface is not within the scope of this document.

2.3. MVPN Definition

An MVPN is defined by two sets of sites: the Sender Sites set and the Receiver Sites set, with the following properties:

- Hosts within the Sender Sites set could originate multicast traffic for receivers in the Receiver Sites set.

- Receivers not in the Receiver Sites set should not be able to receive this traffic.

- Hosts within the Receiver Sites set could receive multicast traffic originated by any host in the Sender Sites set.

- Hosts within the Receiver Sites set should not be able to receive multicast traffic originated by any host that is not in the Sender Sites set.

A site could be both in the Sender Sites set and Receiver Sites set, which implies that hosts within such a site could both originate and receive multicast traffic. An extreme case is when the Sender Sites set is the same as the Receiver Sites set, in which case all sites could originate and receive multicast traffic from each other.

Sites within a given MVPN may be either within the same organization or in different organizations, which implies that an MVPN can be either an Intranet or an Extranet.

A given site may be in more than one MVPN, which implies that MVPNs may overlap.

Not all sites of a given MVPN have to be connected to the same service provider, which implies that an MVPN can span multiple service providers.

Another way to look at MVPN is to say that an MVPN is defined by a set of administrative policies. Such policies determine both the Sender Sites set and Receiver Sites set. Such policies are established by MVPN customers, but implemented/realized by MVPN Service Providers using the existing BGP/MPLS VPN mechanisms, such as Route Targets (RTs), with extensions, as necessary.

2.4. Auto-Discovery

In order for the PE routers attaching to a given MVPN to exchange MVPN control information with each other, each one needs to discover all the other PEs that attach to the same MVPN. (Strictly speaking, a PE in the Receiver Sites set need only discover the other PEs in the Sender Sites set, and a PE in the Sender Sites set need only discover the other PEs in the Receiver Sites set.) This is referred to as "MVPN Auto-Discovery".

This document discusses two ways of providing MVPN auto-discovery:

- BGP can be used for discovering and maintaining MVPN membership. The PE routers advertise their MVPN membership to other PE routers using BGP. A PE is considered to be a "member" of a particular MVPN if it contains a VRF (Virtual Routing and Forwarding table, see [RFC4364]) that is configured to contain the multicast routing information of that MVPN. This auto-discovery option does not make any assumptions about the methods used for transmitting MVPN multicast data packets through the backbone.

- If it is known that the PE-PE multicast control packets (i.e., PIM packets) of a particular MVPN are to be transmitted through a non-aggregated Inclusive Tree supporting the ASM service model (e.g., through a tree that is created by non-SSM PIM-SM or by BIDIR-PIM), and if the PEs attaching to that MVPN are configured with the group address corresponding to that tree, then the PEs can auto-discover each other simply by joining the tree and then multicasting PIM Hellos over the tree.

2.5. PE-PE Multicast Routing Information

The BGP/MPLS IP VPN [RFC4364] specification requires a PE to maintain, at most, one BGP peering with every other PE in the network. This peering is used to exchange VPN routing information. The use of route reflectors further reduces the number of BGP adjacencies maintained by a PE to exchange VPN routing information with other PEs. This document describes various options for exchanging MVPN control information between PE routers based on the use of PIM or BGP. These options have different overheads with respect to the number of routing adjacencies that a PE router needs to maintain to exchange MVPN control information with other PE routers. Some of these options allow the retention of the unicast BGP/MPLS VPN model letting a PE maintain, at most, one BGP routing adjacency with other PE routers to exchange MVPN control information. BGP also provides reliable transport and uses incremental updates. Another option is the use of the currently existing "soft state" PIM standard [PIM-SM] that uses periodic complete updates.

2.6. PE-PE Multicast Data Transmission

Like [RFC4364], this document decouples the procedures for exchanging routing information from the procedures for transmitting data traffic. Hence, a variety of transport technologies may be used in the backbone. For Inclusive Trees, these transport technologies include unicast PE-PE tunnels, using encapsulation in MPLS, IP, or GRE (Generic Routing Encapsulation), multicast distribution trees created by PIM (either unidirectional in the SSM or ASM service models or bidirectional) using IP/GRE encapsulation, point-to-multipoint LSPs created by RSVP - Traffic Engineering (RSVP-TE) or mLDP, and multipoint-to-multipoint LSPs created by mLDP.

In order to aggregate traffic from multiple MVPNs onto a single multicast distribution tree, it is necessary to have a mechanism to enable the egresses of the tree to demultiplex the multicast traffic received over the tree and to associate each received packet with a particular MVPN. This document specifies a mechanism whereby upstream label assignment [MPLS-UPSTREAM-LABEL] is used by the root of the tree to assign a label to each flow. This label is used by

the receivers to perform the demultiplexing. This document also describes procedures based on BGP that are used by the root of an Aggregate Tree to advertise the Inclusive and/or Selective binding and the demultiplexing information to the leaves of the tree.

This document also describes the data plane encapsulations for supporting the various SP multicast transport options.

The specification for aggregating traffic of multiple MVPNs onto a single multipoint-to-multipoint LSP or onto a single bidirectional multicast distribution tree is outside the scope of this document.

The specifications for using, as Selective Trees, multicast distribution trees that support the ASM service model are outside the scope of this document. The specification for using multipoint-to-multipoint LSPs as Selective Trees is outside the scope of this document.

This document assumes that when SP multicast trees are used, traffic for a particular multicast group is transmitted by a particular PE on only one SP multicast tree. The use of multiple SP multicast trees for transmitting traffic belonging to a particular multicast group is outside the scope of this document.

2.7. Inter-AS MVPNs

[RFC4364] describes different options for supporting BGP/MPLS IP unicast VPNs whose provider backbones contain more than one Autonomous System (AS). These are known as "inter-AS VPNs". In an inter-AS VPN, the ASes may belong to the same provider or to different providers. This document describes how inter-AS MVPNs can be supported for each of the unicast BGP/MPLS VPN inter-AS options. This document also specifies a model where inter-AS MVPN service can be offered without requiring a single SP multicast tree to span multiple ASes. In this model, an inter-AS multicast tree consists of a number of "segments", one per AS, that are stitched together at AS boundary points. These are known as "segmented inter-AS trees". Each segment of a segmented inter-AS tree may use a different multicast transport technology.

It is also possible to support inter-AS MVPNs with non-segmented source trees that extend across AS boundaries.

2.8. Optionally Eliminating Shared Tree State

This document also discusses some options and protocol extensions that can be used to eliminate the need for the PE routers to distribute to each other the (*,G) and (*,G,rpt) states that occur when the VPNs are creating unidirectional C-trees to support the ASM service model.

3. Concepts and Framework

3.1. PE-CE Multicast Routing

Support of multicast in BGP/MPLS IP VPNs is modeled closely after the support of unicast in BGP/MPLS IP VPNs. That is, a multicast routing protocol will be run on the PE-CE interfaces, such that PE and CE are multicast routing adjacencies on that interface. CEs at different sites do not become multicast routing adjacencies of each other.

If a PE attaches to n VPNs for which multicast support is provided (i.e., to n "MVPNs"), the PE will run n independent instances of a multicast routing protocol. We will refer to these multicast routing instances as "VPN-specific multicast routing instances", or more briefly as "multicast C-instances". The notion of a "VRF" (VPN Routing and Forwarding Table), defined in [RFC4364], is extended to include multicast routing entries as well as unicast routing entries. Each multicast routing entry is thus associated with a particular VRF.

Whether a particular VRF belongs to an MVPN or not is determined by configuration.

In this document, we do not attempt to provide support for every possible multicast routing protocol that could possibly run on the PE-CE link. Rather, we consider multicast C-instances only for the following multicast routing protocols:

- PIM Sparse Mode (PIM-SM), supporting the ASM service model
- PIM Sparse Mode, supporting the SSM service model
- PIM Bidirectional Mode (BIDIR-PIM), which uses bidirectional C-trees to support the ASM service model.

In order to support the "Carrier’s Carrier" model of [RFC4364], mLDP may also be supported on the PE-CE interface. The use of mLDP on the PE-CE interface is described in [MVPN-BGP].

The use of BGP on the PE-CE interface is not within the scope of this document.

As the only multicast C-instances discussed by this document are PIM-based C-instances, we will generally use the term "PIM C-instances" to refer to the multicast C-instances.

A PE router may also be running a "provider-wide" instance of PIM, (a "PIM P-instance"), in which it has a PIM adjacency with, e.g., each of its IGP neighbors (i.e., with P routers), but NOT with any CE routers, and not with other PE routers (unless another PE router happens to be an IGP adjacency). In this case, P routers would also run the P-instance of PIM but NOT a C-instance. If there is a PIM P-instance, it may or may not have a role to play in the support of VPN multicast; this is discussed in later sections. However, in no case will the PIM P-instance contain VPN-specific multicast routing information.

In order to help clarify when we are speaking of the PIM P-instance and when we are speaking of a PIM C-instance, we will also apply the prefixes "P-" and "C-", respectively, to control messages, addresses, etc. Thus, a P-Join would be a PIM Join that is processed by the PIM P-instance, and a C-Join would be a PIM Join that is processed by a C-instance. A P-group address would be a group address in the SP's address space, and a C-group address would be a group address in a VPN's address space. A C-tree is a multicast distribution tree constructed and maintained by the PIM C-instances. A C-flow is a stream of multicast packets with a common C-source address and a common C-group address. We will use the notation "(C-S,C-G)" to identify specific C-flows. If a particular C-tree is a shared tree (whether unidirectional or bidirectional) rather than a source-specific tree, we will sometimes speak of the entire set of flows traveling that tree, identifying the set as "(C-*,C-G)".

3.2. P-Multicast Service Interfaces (PMSIs)

A PE must have the ability to forward multicast data packets received from a CE to one or more of the other PEs in the same MVPN for delivery to one or more other CEs.

We define the notion of a "P-Multicast Service Interface" (PMSI). If a particular MVPN is supported by a particular set of PE routers, then there will be one or more PMSIs connecting those PE routers and/or subsets thereof. A PMSI is a conceptual "overlay" on the P-network with the following property: a PE in a given MVPN can give a packet to the PMSI, and the packet will be delivered to some or all of the other PEs in the MVPN, such that any PE receiving the packet will be able to determine the MVPN to which the packet belongs.

As we discuss below, a PMSI may be instantiated by a number of different transport mechanisms, depending on the particular requirements of the MVPN and of the SP. We will refer to these transport mechanisms as "P-tunnels".

For each MVPN, there are one or more PMSIs that are used for transmitting the MVPN's multicast data from one PE to others. We will use the term "PMSI" such that a single PMSI belongs to a single MVPN. However, the transport mechanism that is used to instantiate a PMSI may allow a single P-tunnel to carry the data of multiple PMSIs.

In this document, we make a clear distinction between the multicast service (the PMSI) and its instantiation. This allows us to separate the discussion of different services from the discussion of different instantiations of each service. The term "P-tunnel" is used to refer to the transport mechanism that instantiates a service.

PMSIs are used to carry C-multicast data traffic. The C-multicast data traffic travels along a C-tree, but in the SP backbone all C-trees are tunneled through P-tunnels. Thus, we will sometimes talk of a P-tunnel carrying one or more C-trees.

Some of the options for passing multicast control traffic among the PEs do so by sending the control traffic through a PMSI; other options do not send control traffic through a PMSI.

3.2.1. Inclusive and Selective PMSIs

We will distinguish between three different kinds of PMSIs:

- "Multidirectional Inclusive" PMSI (MI-PMSI)

  A Multidirectional Inclusive PMSI is one that enables ANY PE attaching to a particular MVPN to transmit a message such that it will be received by EVERY other PE attaching to that MVPN.

  There is, at most, one MI-PMSI per MVPN. (Though the P-tunnel or P-tunnels that instantiate an MI-PMSI may actually carry the data of more than one PMSI.)

  An MI-PMSI can be thought of as an overlay broadcast network connecting the set of PEs supporting a particular MVPN.

- "Unidirectional Inclusive" PMSI (UI-PMSI)

  A Unidirectional Inclusive PMSI is one that enables a particular PE, attached to a particular MVPN, to transmit a message such that it will be received by all the other PEs attaching to that

MVPN. There is, at most, one UI-PMSI per PE per MVPN, though the P-tunnel that instantiates a UI-PMSI may, in fact, carry the data of more than one PMSI.

- "Selective" PMSI (S-PMSI).

  A Selective PMSI is one that provides a mechanism wherein a particular PE in an MVPN can multicast messages so that they will be received by a subset of the other PEs of that MVPN. There may be an arbitrary number of S-PMSIs per PE per MVPN. The P-tunnel that instantiates a given S-PMSI may carry data from multiple S-PMSIs.

In later sections, we describe the role played by these different kinds of PMSIs. We will use the term "I-PMSI" when we are not distinguishing between "MI-PMSIs" and "UI-PMSIs".

3.2.2. P-Tunnels Instantiating PMSIs

The P-tunnels that are used to instantiate PMSIs will be referred to as "P-tunnels". A number of different tunnel setup techniques can be used to create the P-tunnels that instantiate the PMSIs. Among these are the following:

- PIM

  A PMSI can be instantiated as (a set of) Multicast Distribution trees created by the PIM P-instance ("P-trees").

  The multicast distribution trees that instantiate I-PMSIs may be either shared trees or source-specific trees.

  This document (along with [MVPN-BGP]) specifies procedures for identifying a particular (C-S,C-G) flow and assigning it to a particular S-PMSI. Such an S-PMSI is most naturally instantiated as a source-specific tree.

  The use of shared trees (including bidirectional trees) to instantiate S-PMSIs is outside the scope of this document.

  The use of PIM-DM to create P-tunnels is not supported.

  P-tunnels may be shared by multiple MVPNs (i.e., a given P-tunnel may be the instantiation of multiple PMSIs), as long as the tunnel encapsulation provides some means of demultiplexing the data traffic by MVPN.

- mLDP

  mLDP Point-to-Multipoint (P2MP) LSPs or Multipoint-to-Multipoint (MP2MP) LSPs can be used to instantiate I-PMSIs.

  An S-PMSI or a UI-PMSI could be instantiated as a single mLDP P2MP LSP, whereas an MI-PMSI would have to be instantiated as a set of such LSPs (each PE in the MVPN being the root of one such LSP) or as a single MP2MP LSP.

  Procedures for sharing MP2MP LSPs across multiple MVPNs are outside the scope of this document.

  The use of MP2MP LSPs to instantiate S-PMSIs is outside the scope of this document.

  Section 11.2.3 discusses a way of using a partial mesh of MP2MP LSPs to instantiate a PMSI. However, a full specification of the necessary procedures is outside the scope of this document.

- RSVP-TE

  A PMSI may be instantiated as one or more RSVP-TE Point-to-Multipoint (P2MP) LSPs. An S-PMSI or a UI-PMSI would be instantiated as a single RSVP-TE P2MP LSP, whereas a Multidirectional Inclusive PMSI would be instantiated as a set of such LSPs, one for each PE in the MVPN. RSVP-TE P2MP LSPs can be shared across multiple MVPNs.

- A Mesh of Unicast P-Tunnels.

  If a PMSI is implemented as a mesh of unicast P-tunnels, a PE wishing to transmit a packet through the PMSI would replicate the packet and send a copy to each of the other PEs.

  An MI-PMSI for a given MVPN can be instantiated as a full mesh of unicast P-tunnels among that MVPN's PEs. A UI-PMSI or an S-PMSI can be instantiated as a partial mesh.

It can be seen that each method of implementing PMSIs has its own area of applicability. Therefore, this specification allows for the use of any of these methods. At first glance, this may seem like an overabundance of options. However, the history of multicast development and deployment should make it clear that there is no one option that is always acceptable. The use of segmented inter-AS trees does allow each SP to select the option that it finds most applicable in its own environment, without causing any other SP to choose that same option.

SPECIFYING THE CONDITIONS UNDER WHICH A PARTICULAR TREE-BUILDING METHOD IS APPLICABLE IS OUTSIDE THE SCOPE OF THIS DOCUMENT.

The choice of the tunnel technique belongs to the sender router and is a local policy decision of that router. The procedures defined throughout this document do not mandate that the same tunnel technique be used for all P-tunnels going through a given provider backbone. However, it is expected that any tunnel technique that can be used by a PE for a particular MVPN is also supported by all the other PEs having VRFs for the MVPN. Moreover, the use of ingress replication by any PE for an MVPN implies that all other PEs MUST use ingress replication for this MVPN.

3.3. Use of PMSIs for Carrying Multicast Data

Each PE supporting a particular MVPN must have a way of discovering the following information:

- The set of other PEs in its AS that are attached to sites of that MVPN, and the set of other ASes that have PEs attached to sites of that MVPN. However, if non-segmented inter-AS trees are used (see Section 8.1), then each PE needs to know the entire set of PEs attached to sites of that MVPN.

- If segmented inter-AS trees are to be used, the set of border routers in its AS that support inter-AS connectivity for that MVPN.

- If the MVPN is configured to use an MI-PMSI, the information needed to set up and to use the P-tunnels instantiating the MI-PMSI.

- For each other PE, whether the PE supports Aggregate Trees for the MVPN, and if so, the demultiplexing information that must be provided so that the other PE can determine whether a packet that it received on an Aggregate Tree belongs to this MVPN.

In some cases, the information above is provided by means of the BGP-based auto-discovery procedures discussed in Section 4 of this document and in Section 9 of [MVPN-BGP]. In other cases, this information is provided after discovery is complete, by means of procedures discussed in Section 7.4. In either case, the information that is provided must be sufficient to enable the PMSI to be bound to the identified P-tunnel, to enable the P-tunnel to be created if it does not already exist, and to enable the different PMSIs that may travel on the same P-tunnel to be properly demultiplexed.

If an MVPN uses an MI-PMSI, then the information needed to identify the P-tunnels that instantiate the MI-PMSI has to be known to the PEs attached to the MVPN before any data can be transmitted on the MI-PMSI. This information is either statically configured or auto-discovered (see Section 4). The actual process of constructing the P-tunnels (e.g., via PIM, RSVP-TE, or mLDP) SHOULD occur as soon as this information is known.

When MI-PMSIs are used, they may serve as the default method of carrying C-multicast data traffic. When we say that an MI-PMSI is the "default" method of carrying C-multicast data traffic for a particular MVPN, we mean that it is not necessary to use any special control procedures to bind a particular C-flow to the MI-PMSI; any C-flows that have not been bound to other PMSIs will be assumed to travel through the MI-PMSI.

There is no requirement to use MI-PMSIs as the default method of carrying C-flows. It is possible to adopt a policy in which all C-flows are carried on UI-PMSIs or S-PMSIs. In this case, if an MI-PMSI is not used for carrying routing information, it is not needed at all.

Even when an MI-PMSI is used as the default method of carrying an MVPN’s C-flows, if a particular C-flow has certain characteristics, it may be desirable to migrate it from the MI-PMSI to an S-PMSI. These characteristics, as well as the procedures for migrating a C-flow from an MI-PMSI to an S-PMSI, are discussed in Section 7.

Sometimes a set of C-flows are traveling the same, shared, C-tree (e.g., either unidirectional or bidirectional), and it may be desirable to move the whole set of C-flows as a unit to an S-PMSI. Procedures for doing this are outside the scope of this specification.

Some of the procedures for transmitting C-multicast routing information among the PEs require that the routing information be sent over an MI-PMSI. Other procedures do not use an MI-PMSI to transmit the C-multicast routing information.

For a given MVPN, whether an MI-PMSI is used to carry C-multicast routing information is independent from whether an MI-PMSI is used as the default method of carrying the C-multicast data traffic.

As previously stated, it is possible to send all C-flows on a set of S-PMSIs, omitting any usage of I-PMSIs. This prevents PEs from receiving data that they don’t need, at the cost of requiring additional P-tunnels, and additional signaling to bind the C-flows to P-tunnels. Cost-effective instantiation of S-PMSIs is likely to

require Aggregate P-trees, which, in turn, makes it necessary for the transmitting PE to know which PEs need to receive which multicast streams. This is known as "explicit tracking", and the procedures to enable explicit tracking may themselves impose a cost. This is further discussed in Section 7.4.1.2.

3.4. PE-PE Transmission of C-Multicast Routing

As a PE attached to a given MVPN receives C-Join/Prune messages from its CEs in that MVPN, it must convey the information contained in those messages to other PEs that are attached to the same MVPN.

There are several different methods for doing this. As these methods are not interoperable, the method to be used for a particular MVPN must be either configured or discovered as part of the auto-discovery process.

3.4.1. PIM Peering

3.4.1.1. Full per-MVPN PIM Peering across an MI-PMSI

If the set of PEs attached to a given MVPN are connected via an MI-PMSI, the PEs can form "normal" PIM adjacencies with each other. Since the MI-PMSI functions as a broadcast network, the standard PIM procedures for forming and maintaining adjacencies over a LAN can be applied.

As a result, the C-Join/Prune messages that a PE receives from a CE can be multicast to all the other PEs of the MVPN. PIM "Join suppression" can be enabled and the PEs can send Asserts as needed.

This procedure is fully specified in Section 5.2.

3.4.1.2. Lightweight PIM Peering across an MI-PMSI

The procedure of the previous Section has the following disadvantages:

- Periodic Hello messages must be sent by all PEs.

  Standard PIM procedures require that each PE in a particular MVPN periodically multicast a Hello to all the other PEs in that MVPN. If the number of MVPNs becomes very large, sending and receiving these Hellos can become a substantial overhead for the PE routers.

- Periodic retransmission of C-Join/Prune messages.

  PIM is a "soft-state" protocol, in which reliability is assured through frequent retransmissions (refresh) of control messages. This too can begin to impose a large overhead on the PE routers as the number of MVPNs grows.

The first of these disadvantages is easily remedied. The reason for the periodic PIM Hellos is to ensure that each PIM speaker on a LAN knows who all the other PIM speakers on the LAN are. However, in the context of MVPN, PEs in a given MVPN can learn the identities of all the other PEs in the MVPN by means of the BGP-based auto-discovery procedure of Section 4. In that case, the periodic Hellos would serve no function and could simply be eliminated. (Of course, this does imply a change to the standard PIM procedures.)

When Hellos are suppressed, we may speak of "lightweight PIM peering".

The periodic refresh of the C-Join/Prune messages is not as simple to eliminate. If and when "refresh reduction" procedures are specified for PIM, it may be useful to incorporate them, so as to make the lightweight PIM peering procedures even more lightweight.

Lightweight PIM peering is not specified in this document.

3.4.1.3. Unicasting of PIM C-Join/Prune Messages

PIM does not require that the C-Join/Prune messages that a PE receives from a CE to be multicast to all the other PEs; it allows them to be unicast to a single PE, the one that is upstream on the path to the root of the multicast tree mentioned in the Join/Prune message. Note that when the C-Join/Prune messages are unicast, there is no such thing as "Join suppression". Therefore, PIM Refresh Reduction may be considered to be a prerequisite for the procedure of unicasting the C-Join/Prune messages.

When the C-Join/Prune messages are unicast, they are not transmitted on a PMSI at all. Note that the procedure of unicasting the C-Join/Prune messages is different than the procedure of transmitting the C-Join/Prune messages on an MI-PMSI that is instantiated as a mesh of unicast P-tunnels.

If there are multiple PEs that can be used to reach a given C-source, procedures described in Sections 5.1 and 9 MUST be used to ensure that duplicate packets do not get delivered.

Procedures for unicasting the PIM control messages are not further specified in this document.

3.4.2. Using BGP to Carry C-Multicast Routing

It is possible to use BGP to carry C-multicast routing information from PE to PE, dispensing entirely with the transmission of C-Join/Prune messages from PE to PE. This is discussed in Section 5.3 and fully specified in [MVPN-BGP].

4. BGP-Based Auto-Discovery of MVPN Membership

BGP-based auto-discovery is done by means of a new address family, the MCAST-VPN address family. (This address family also has other uses, as will be seen later.) Any PE that attaches to an MVPN must issue a BGP Update message containing an NLRI ("Network Layer Reachability Information" element) in this address family, along with a specific set of attributes. In this document, we specify the information that must be contained in these BGP Updates in order to provide auto-discovery. The encoding details, along with the complete set of detailed procedures, are specified in a separate document [MVPN-BGP].

This section specifies the intra-AS BGP-based auto-discovery procedures. When segmented inter-AS trees are used, additional procedures are needed, as specified in [MVPN-BGP]. (When segmented inter-AS trees are not used, the inter-AS procedures are almost identical to the intra-AS procedures.)

BGP-based auto-discovery uses a particular kind of MCAST-VPN route known as an "auto-discovery route", or "A-D route". In particular, it uses two kinds of "A-D routes": the "Intra-AS I-PMSI A-D route" and the "Inter-AS I-PMSI A-D route". (There are also additional kinds of A-D routes, such as the Source Active A-D routes, which are used for purposes that go beyond auto-discovery. These are discussed in subsequent sections.)

The Inter-AS I-PMSI A-D route is used only when segmented inter-AS P-tunnels are used, as specified in [MVPN-BGP].

The "Intra-AS I-PMSI A-D route" is originated by the PEs that are (directly) connected to the site(s) of an MVPN. It is distributed to other PEs that attach to sites of the MVPN. If segmented inter-AS P-tunnels are used, then the Intra-AS I-PMSI A-D routes are not distributed outside the AS where they originate; if segmented inter-AS P-tunnels are not used, then the Intra-AS I-PMSI A-D routes are, despite their name, distributed to all PEs attached to the VPN, no matter what AS the PEs are in.

The NLRI of an Intra-AS I-PMSI A-D route must contain the following information:

- The route type (i.e., Intra-AS I-PMSI A-D route).

- The IP address of the originating PE.

- An RD ("Route Distinguisher", [RFC4364]) configured locally for the MVPN. This is an RD that can be prepended to that IP address to form a globally unique VPN-IP address of the PE.

Intra-AS I-PMSI A-D routes carry the following attributes:

- Route Target Extended Communities attribute.

  One or more of these MUST be carried by each Intra-AS I-PMSI A-D route. If any other PE has one of these Route Targets configured for import into a VRF, it treats the advertising PE as a member in the MVPN to which the VRF belongs. This allows each PE to discover the PEs that belong to a given MVPN. More specifically, it allows a PE in the Receiver Sites set to discover the PEs in the Sender Sites set of the MVPN, and the PEs in the Sender Sites set of the MVPN to discover the PEs in the Receiver Sites set of the MVPN. The PEs in the Receiver Sites set would be configured to import the Route Targets advertised in the BGP A-D routes by PEs in the Sender Sites set. The PEs in the Sender Sites set would be configured to import the Route Targets advertised in the BGP A-D routes by PEs in the Receiver Sites set.

- PMSI Tunnel attribute.

  This attribute is present whenever the MVPN uses an MI-PMSI or when it uses a UI-PMSI rooted at the originating router. It contains the following information:

    * tunnel technology, which may be one of the following:

        + Bidirectional multicast tree created by BIDIR-PIM,

        + Source-specific multicast tree created by PIM-SM, supporting the SSM service model,

        + Set of trees (one shared tree and a set of source trees) created by PIM-SM using the ASM service model,

        + Point-to-multipoint LSP created by RSVP-TE,

        + Point-to-multipoint LSP created by mLDP,

+ multipoint-to-multipoint LSP created by mLDP

+ unicast tunnel

* P-tunnel identifier

  Before a P-tunnel can be constructed to instantiate the I-PMSI, the PE must be able to create a unique identifier for the tunnel. The syntax of this identifier depends on the tunnel technology used.

  Each PE attaching to a given MVPN must be configured with information specifying the allowable encapsulations to use for that MVPN, as well as the particular one of those encapsulations that the PE is to identify in the PMSI Tunnel attribute of the Intra-AS I-PMSI A-D routes that it originates.

* Multi-VPN aggregation capability and demultiplexor value.

  This specifies whether the P-tunnel is capable of aggregating I-PMSIs from multiple MVPNs. This will affect the encapsulation used. If aggregation is to be used, a demultiplexor value to be carried by packets for this particular MVPN must also be specified. The demultiplexing mechanism and signaling procedures are described in Section 6.

- PE Distinguisher Labels Attribute

  Sometimes it is necessary for one PE to advertise an upstream-assigned MPLS label that identifies another PE. Under certain circumstances to be discussed later, a PE that is the root of a multicast P-tunnel will bind an MPLS label value to one or more of the PEs that belong to the P-tunnel, and it will distribute these label bindings using Intra-AS I-PMSI A-D routes.

  Specification of when this must be done is provided in Sections 6.4.4 and 11.2.2. We refer to these as "PE Distinguisher Labels".

  Note that, as specified in [MPLS-UPSTREAM-LABEL], PE Distinguisher Label values are unique only in the context of the IP address identifying the root of the P-tunnel; they are not necessarily unique per tunnel.

5. PE-PE Transmission of C-Multicast Routing

As a PE attached to a given MVPN receives C-Join/Prune messages from its CEs in that MVPN, it must convey the information contained in those messages to other PEs that are attached to the same MVPN. This is known as the "PE-PE transmission of C-multicast routing information".

This section specifies the procedures used for PE-PE transmission of C-multicast routing information. Not every procedure mentioned in Section 3.4 is specified here. Rather, this section focuses on two particular procedures:

- Full PIM Peering.

  This procedure is fully specified herein.

- Use of BGP to distribute C-multicast routing

  This procedure is described herein, but the full specification appears in [MVPN-BGP].

Those aspects of the procedures that apply to both of the above are also specified fully herein.

Specification of other procedures is outside the scope of this document.

5.1. Selecting the Upstream Multicast Hop (UMH)

When a PE receives a C-Join/Prune message from a CE, the message identifies a particular multicast flow as belonging either to a source-specific tree (S,G) or to a shared tree (*,G). Throughout this section, we use the term "C-root" to refer to S, in the case of a source-specific tree, or to the Rendezvous Point (RP) for G, in the case of (*,G). If the route to the C-root is across the VPN backbone, then the PE needs to find the "Upstream Multicast Hop" (UMH) for the (S,G) or (*,G) flow. The UMH is either the PE at which (S,G) or (*,G) data packets enter the VPN backbone or the Autonomous System Border Router (ASBR) at which those data packets enter the local AS when traveling through the VPN backbone. The process of finding the upstream multicast hop for a given C-root is known as "upstream multicast hop selection".

5.1.1. Eligible Routes for UMH Selection

In the simplest case, the PE does the upstream hop selection by looking up the C-root in the unicast VRF associated with the PE-CE interface over which the C-Join/Prune message was received. The route that matches the C-root will contain the information needed to select the UMH.

However, in some cases, the CEs may be distributing to the PEs a special set of routes that are to be used exclusively for the purpose of upstream multicast hop selection, and not used for unicast routing at all. For example, when BGP is the CE-PE unicast routing protocol, the CEs may be using Subsequent Address Family Identifier 2 (SAFI 2) to distribute a special set of routes that are to be used for, and only for, upstream multicast hop selection. When OSPF [OSPF] is the CE-PE routing protocol, the CE may use an MT-ID (Multi-Topology Identifier) [OSPF-MT] of 1 to distribute a special set of routes that are to be used for, and only for, upstream multicast hop selection. When a CE uses one of these mechanisms to distribute to a PE a special set of routes to be used exclusively for upstream multicast hop selection, these routes are distributed among the PEs using SAFI 129, as described in [MVPN-BGP]. Whether the routes used for upstream multicast hop selection are (a) the "ordinary" unicast routes or (b) a special set of routes that are used exclusively for upstream multicast hop selection is a matter of policy. How that policy is chosen, deployed, or implemented is outside the scope of this document. In the following, we will simply refer to the set of routes that are used for upstream multicast hop selection, the "Eligible UMH routes", with no presumptions about the policy by which this set of routes was chosen.

5.1.2. Information Carried by Eligible UMH Routes

Every route that is eligible for UMH selection SHOULD carry a VRF Route Import Extended Community [MVPN-BGP]. However, if BGP is used to distribute C-multicast routing information, or if the route is from a VRF that belongs to a multi-AS VPN as described in option b of Section 10 of [RFC4364], then the route MUST carry a VRF Route Import Extended Community. This attribute identifies the PE that originated the route.

If BGP is used for carrying C-multicast routes, OR if "Segmented inter-AS Tunnels" are used, then every UMH route MUST also carry a Source AS Extended Community [MVPN-BGP].

These two attributes are used in the upstream multicast hop selection procedures described below.

5.1.3. Selecting the Upstream PE

The first step in selecting the upstream multicast hop for a given C-root is to select the Upstream PE router for that C-root.

The PE that received the C-Join message from a CE looks in the VRF corresponding to the interfaces over which the C-Join was received. It finds the Eligible UMH route that is the best match for the C-root specified in that C-Join. Call this the "Installed UMH Route".

Note that the outgoing interface of the Installed UMH Route may be one of the interfaces associated with the VRF, in which case the upstream multicast hop is a CE and the route to the C-root is not across the VPN backbone.

Consider the set of all VPN-IP routes that (a) are eligible to be imported into the VRF (as determined by their Route Targets), (b) are eligible to be used for upstream multicast hop selection, and (c) have exactly the same IP prefix (not necessarily the same RD) as the installed UMH route.

For each route in this set, determine the corresponding Upstream PE and Upstream RD. If a route has a VRF Route Import Extended Community, the route's Upstream PE is determined from it. If a route does not have a VRF Route Import Extended Community, the route's Upstream PE is determined from the route's BGP Next Hop. In either case, the Upstream RD is taken from the route's NLRI.

This results in a set of triples of <route, Upstream PE, Upstream RD>.

Call this the "UMH Route Candidate Set". Then, the PE MUST select a single route from the set to be the "Selected UMH Route". The corresponding Upstream PE is known as the "Selected Upstream PE", and the corresponding Upstream RD is known as the "Selected Upstream RD".

There are several possible procedures that can be used by a PE to select a single route from the candidate set.

The default procedure, which MUST be implemented, is to select the route whose corresponding Upstream PE address is numerically highest, where a 32-bit IP address is treated as a 32-bit unsigned integer. Call this the "default Upstream PE selection". For a given C-root, provided that the routing information used to create the candidate set is stable, all PEs will have the same default Upstream PE selection. (Though different default Upstream PE selections may be chosen during a routing transient.)

An alternative procedure that MUST be implemented, but which is disabled by default, is the following. This procedure ensures that, except during a routing transient, each PE chooses the same Upstream PE for a given combination of C-root and C-G.

1. The PEs in the candidate set are numbered from lowest to highest IP address, starting from 0.

2. The following hash is performed:

   - A bytewise exclusive-or of all the bytes in the C-root address and the C-G address is performed.

   - The result is taken modulo n, where n is the number of PEs in the candidate set. Call this result N.

The Selected Upstream PE is then the one that appears in position N in the list of step 1.

Other hashing algorithms are allowed as well, but not required.

The alternative procedure allows a form of "equal cost load balancing". Suppose, for example, that from egress PEs PE3 and PE4, source C-S can be reached, at equal cost, via ingress PE PE1 or ingress PE PE2. The load balancing procedure makes it possible for PE1 to be the ingress PE for (C-S,C-G1) data traffic while PE2 is the ingress PE for (C-S,C-G2) data traffic.

Another procedure, which SHOULD be implemented, is to use the Installed UMH Route as the Selected UMH Route. If this procedure is used, the result is likely to be that a given PE will choose the Upstream PE that is closest to it, according to the routing in the SP backbone. As a result, for a given C-root, different PEs may choose different Upstream PEs. This is useful if the C-root is an anycast address, and can also be useful if the C-root is in a multihomed site (i.e., a site that is attached to multiple PEs). However, this procedure is more likely to lead to steady state duplication of traffic unless (a) PEs discard data traffic that arrives from the "wrong" Upstream PE or (b) data traffic is carried only in non-aggregated S-PMSIs. This issue is discussed at length in Section 9.

General policy-based procedures for selecting the UMH route are allowed but not required, and they are not further discussed in this specification.

5.1.4. Selecting the Upstream Multicast Hop

In certain cases, the Selected Upstream Multicast Hop is the same as the Selected Upstream PE. In other cases, the Selected Upstream Multicast Hop is the ASBR that is the BGP Next Hop of the Selected UMH Route.

If the Selected Upstream PE is in the local AS, then the Selected Upstream PE is also the Selected Upstream Multicast Hop. This is the case if any of the following conditions holds:

- The Selected UMH Route has a Source AS Extended Community, and the Source AS is the same as the local AS,

- The Selected UMH Route does not have a Source AS Extended Community, but the route's BGP Next Hop is the same as the Upstream PE.

Otherwise, the Selected Upstream Multicast Hop is an ASBR. The method of determining just which ASBR it is depends on the particular inter-AS signaling method being used (PIM or BGP) and on whether segmented or non-segmented inter-AS tunnels are used. These details are presented in later sections.

5.2. Details of Per-MVPN Full PIM Peering over MI-PMSI

When an MVPN uses an MI-PMSI, the C-instances of that MVPN can treat the MI-PMSI as a LAN interface and form full PIM adjacencies with each other over that LAN interface.

The use of PIM when an MI-PMSI is not in use is outside the scope of this document.

To form full PIM adjacencies, the PEs execute the standard PIM procedures on the LAN interface, including the generation and processing of PIM Hello, Join/Prune, Assert, DF (Designated Forwarder) election, and other PIM control messages. These are executed independently for each C-instance. PIM "Join suppression" SHOULD be enabled.

5.2.1. PIM C-Instance Control Packets

All IPv4 PIM C-instance control packets of a particular MVPN are addressed to the ALL-PIM-ROUTERS (224.0.0.13) IP destination address and transmitted over the MI-PMSI of that MVPN. While in transit in the P-network, the packets are encapsulated as required for the particular kind of P-tunnel that is being used to instantiate the

MI-PMSI. Thus, the C-instance control packets are not processed by the P routers, and MVPN-specific PIM routes can be extended from site to site without appearing in the P routers.

The handling of IPv6 PIM C-instance control packets will be specified in a follow-on document.

As specified in Section 5.1.2, when PIM is being used to distribute C-multicast routing information, any PE distributing VPN-IP routes that are eligible for use as UMH routes SHOULD include a VRF Route Import Extended Community with each route. For a given VRF, the Global Administrator field of the VRF Route Import Extended Community MUST be set to the same IP address that the PE places in the IP source address field of the PE-PE PIM control messages it originates from that VRF.

Note that BSR (Bootstrap Router Mechanism for PIM) [BSR] messages are treated the same as PIM C-instance control packets, and BSR processing is regarded as an integral part of the PIM C-instance processing.

5.2.2. PIM C-Instance Reverse Path Forwarding (RPF) Determination

Although the MI-PMSI is treated by PIM as a LAN interface, unicast routing is NOT run over it, and there are no unicast routing adjacencies over it. Therefore, it is necessary to specify special procedures for determining when the MI-PMSI is to be regarded as the "RPF Interface" for a particular C-address.

The PE follows the procedures of Section 5.1 to determine the Selected UMH Route. If that route is NOT a VPN-IP route learned from BGP as described in [RFC4364], or if that route’s outgoing interface is one of the interfaces associated with the VRF, then ordinary PIM procedures for determining the RPF interface apply.

However, if the Selected UMH Route is a VPN-IP route whose outgoing interface is not one of the interfaces associated with the VRF, then PIM will consider the RPF interface to be the MI-PMSI associated with the VPN-specific PIM instance.

Once PIM has determined that the RPF interface for a particular C-root is the MI-PMSI, it is necessary for PIM to determine the "RPF neighbor" for that C-root. This will be one of the other PEs that is a PIM adjacency over the MI-PMSI. In particular, it will be the "Selected Upstream PE", as defined in Section 5.1.

5.3. Use of BGP for Carrying C-Multicast Routing

It is possible to use BGP to carry C-multicast routing information from PE to PE, dispensing entirely with the transmission of C-Join/Prune messages from PE to PE. This section describes the procedures for carrying intra-AS multicast routing information. Inter-AS procedures are described in Section 8. The complete specification of both sets of procedures and of the encodings can be found in [MVPN-BGP].

5.3.1. Sending BGP Updates

The MCAST-VPN address family is used for this purpose. MCAST-VPN routes used for the purpose of carrying C-multicast routing information are distinguished from those used for the purpose of carrying auto-discovery information by means of a "route type" field that is encoded into the NLRI. The following information is required in BGP to advertise the MVPN routing information. The NLRI contains the following:

- The type of C-multicast route

  There are two types:

  * source tree join

  * shared tree join

- The C-group address

- The C-source address (In the case of a shared tree join, this is the address of the C-RP.)

- The Selected Upstream RD corresponding to the C-root address (determined by the procedures of Section 5.1).

Whenever a C-multicast route is sent, it must also carry the Selected Upstream Multicast Hop corresponding to the C-root address (determined by the procedures of Section 5.1). The Selected Upstream Multicast Hop must be encoded as part of a Route Target Extended Community to facilitate the optional use of filters that can prevent the distribution of the update to BGP speakers other than the Upstream Multicast Hop. See Section 10.1.3 of [MVPN-BGP] for the details.

There is no C-multicast route corresponding to the PIM function of pruning a source off the shared tree when a PE switches from a (C-*,C-G) tree to a (C-S,C-G) tree. Section 9 of this document

specifies a mandatory procedure that ensures that if any PE joins a (C-S,C-G) source tree, all other PEs that have joined or will join the (C-*,C-G) shared tree will also join the (C-S,C-G) source tree.

This eliminates the need for a C-multicast route that prunes C-S off the (C-*,C-G) shared tree when switching from (C-*,C-G) to (C-S,C-G) tree.

5.3.2. Explicit Tracking

Note that the upstream multicast hop is NOT part of the NLRI in the C-multicast BGP routes. This means that if several PEs join the same C-tree, the BGP routes they distribute to do so are regarded by BGP as comparable routes, and only one will be installed. If a route reflector is being used, this further means that the PE that is used to reach the C-source will know only that one or more of the other PEs have joined the tree, but it won't know which one. That is, this BGP update mechanism does not provide "explicit tracking". Explicit tracking is not provided by default because it increases the amount of state needed and thus decreases scalability. Also, as constructing the C-PIM messages to send "upstream" for a given tree does not depend on knowing all the PEs that are downstream on that tree, there is no reason for the C-multicast route type updates to provide explicit tracking.

There are some cases in which explicit tracking is necessary in order for the PEs to set up certain kinds of P-trees. There are other cases in which explicit tracking is desirable in order to determine how to optimally aggregate multicast flows onto a given aggregate tree. As these functions have to do with the setting up of infrastructure in the P-network, rather than with the dissemination of C-multicast routing information, any explicit tracking that is necessary is handled by sending a particular type of A-D route known as "Leaf A-D routes".

Whenever a PE sends an A-D route with a PMSI Tunnel attribute, it can set a bit in the PMSI Tunnel attribute indicating "Leaf Information Required". A PE that installs such an A-D route MUST respond by generating a Leaf A-D route, indicating that it needs to join (or be joined to) the specified PMSI Tunnel. Details can be found in [MVPN-BGP].

5.3.3. Withdrawing BGP Updates

A PE removes itself from a C-multicast tree (shared or source) by withdrawing the corresponding BGP Update.

If a PE has pruned a C-source from a shared C-multicast tree, and it needs to "unprune" that source from that tree, it does so by withdrawing the route that pruned the source from the tree.

5.3.4. BSR

BGP does not provide a method for carrying the control information of BSR packets received by a PE from a CE. BSR is supported by transmitting the BSR control messages from one PE in an MVPN to all the other PEs in that MVPN.

When a PE needs to transmit a BSR message for a particular MVPN to other PEs, it must put its own IP address into the BSR message as the IP source address. As specified in Section 5.1.2, when a PE distributes VPN-IP routes that are eligible for use as UMH routes, the PE MUST include a VRF Route Import Extended Community with each route. For a given MVPN, a single such IP address MUST be used, and that same IP address MUST be used as the source address in all BSR packets that the PE transmits to other PEs.

The BSR message may be transmitted over any PMSI that will deliver the message to all the other PEs in the MVPN. If no such PMSI has been instantiated yet, then an appropriate P-tunnel must be advertised, and the C-flow whose C-source address is the address of the PE itself, and whose multicast group is ALL-PIM-ROUTERS (224.0.0.13), must be bound to it. This can be done using the procedures described in Sections 7.3 and 7.4. Note that this is NOT meant to imply that the other PIM control packets from the PIM C-instance are to be transmitted to the other PEs.

When a PE receives a BSR message for a particular MVPN from some other PE, the PE accepts the message only if the IP source address in that message is the Selected Upstream PE (see Section 5.1.3) for the IP address of the Bootstrap router. Otherwise, the PE simply discards the packet. If the PE accepts the packet, it does normal BSR processing on it, and it may forward a BSR message to one or more CEs as a result.

6. PMSI Instantiation

This section provides the procedures for using P-tunnels to instantiate a PMSI. It describes the procedures for setting up and maintaining the P-tunnels as well as for sending and receiving C-data and/or C-control messages on the P-tunnels. However, procedures for binding particular C-flows to particular P-tunnels are discussed in Section 7.

PMSIs can be instantiated either by P-multicast trees or by PE-PE unicast tunnels. In the latter case, the PMSI is said to be instantiated by "ingress replication".

This specification supports a number of different methods for setting up P-multicast trees: these are detailed below. A P-tunnel may support a single VPN (a non-aggregated P-multicast tree) or multiple VPNs (an aggregated P-multicast tree).

6.1. Use of the Intra-AS I-PMSI A-D Route

6.1.1. Sending Intra-AS I-PMSI A-D Routes

When a PE is provisioned to have one or more VRFs that provide MVPN support, the PE announces its MVPN membership information using Intra-AS I-PMSI A-D routes, as discussed in Section 4 and detailed in Section 9.1.1 of [MVPN-BGP]. (Under certain conditions, detailed in [MVPN-BGP], the Intra-AS I-PMSI A-D route may be omitted.)

Generally, the Intra-AS I-PMSI A-D route will have a PMSI Tunnel attribute that identifies a P-tunnel that is being used to instantiate the I-PMSI. Section 9.1.1 of [MVPN-BGP] details certain conditions under which the PMSI Tunnel attribute may be omitted (or in which a PMSI Tunnel attribute with the "no tunnel information present" bit may be sent).

As a special case, when (a) C-PIM control messages are to be sent through an MI-PMSI and (b) the MI-PMSI is instantiated by a P-tunnel technique for which each PE needs to know only a single P-tunnel identifier per VPN, then the use of the Intra-AS I-PMSI A-D routes MAY be omitted, and static configuration of the tunnel identifier used instead. However, this is not recommended for long-term use, and in all other cases, the Intra-AS I-PMSI A-D routes MUST be used.

The PMSI Tunnel attribute MAY contain an upstream-assigned MPLS label, assigned by the PE originating the Intra-AS I-PMSI A-D route. If this label is present, the P-tunnel can be carrying data from several MVPNs. The label is used on the data packets traveling through the tunnel to identify the MVPN to which those data packets belong. (The specified label identifies the packet as belonging to the MVPN that is identified by the RTs of the Intra-AS I-PMSI A-D route.)

See Section 12.2 for details on how to place the label in the packet’s label stack.

The Intra-AS I-PMSI A-D route may contain a "PE Distinguisher Labels" attribute. This contains a set of bindings between upstream-assigned labels and PE addresses. The PE that originated the route may use this to bind an upstream-assigned label to one or more of the other PEs that belong to the same MVPN. The way in which PE Distinguisher Labels are used is discussed in Sections 6.4.1, 6.4.3, 11.2.2, and 12.3. Other uses of the PE Distinguisher Labels attribute are outside the scope of this document.

6.1.2. Receiving Intra-AS I-PMSI A-D Routes

The action to be taken when a PE receives an Intra-AS I-PMSI A-D route for a particular MVPN depends on the particular P-tunnel technology that is being used by that MVPN. If the P-tunnel technology requires tunnels to be built by means of receiver-initiated joins, the PE SHOULD join the tunnel immediately.

6.2. When C-flows Are Specifically Bound to P-Tunnels

This situation is discussed in Section 7.

6.3. Aggregating Multiple MVPNs on a Single P-Tunnel

When a P-multicast tree is shared across multiple MVPNs, it is termed an "Aggregate Tree". The procedures described in this document allow a single SP multicast tree to be shared across multiple MVPNs. Unless otherwise specified, P-multicast tree technology supports aggregation.

All procedures that are specific to multi-MVPN aggregation are OPTIONAL and are explicitly pointed out.

Aggregate Trees allow a single P-multicast tree to be used across multiple MVPNs so that state in the SP core grows per set of MVPNs and not per MVPN. Depending on the congruence of the aggregated MVPNs, this may result in trading off optimality of multicast routing.

An Aggregate Tree can be used by a PE to provide a UI-PMSI or MI-PMSI service for more than one MVPN. When this is the case, the Aggregate Tree is said to have an inclusive mapping.

6.3.1. Aggregate Tree Leaf Discovery

BGP MVPN membership discovery (Section 4) allows a PE to determine the different Aggregate Trees that it should create and the MVPNs that should be mapped onto each such tree. The leaves of an Aggregate Tree are determined by the PEs, supporting aggregation, that belong to all the MVPNs that are mapped onto the tree.

If an Aggregate Tree is used to instantiate one or more S-PMSIs, then it may be desirable for the PE at the root of the tree to know which PEs (in its MVPN) are receivers on that tree. This enables the PE to decide when to aggregate two S-PMSIs, based on congruence (as discussed in the next section). Thus, explicit tracking may be required. Since the procedures for disseminating C-multicast routes do not provide explicit tracking, a type of A-D route known as a "Leaf A-D route" is used. The PE that wants to assign a particular C-multicast flow to a particular Aggregate Tree can send an A-D route, which elicits Leaf A-D routes from the PEs that need to receive that C-multicast flow. This provides the explicit tracking information needed to support the aggregation methodology discussed in the next section. For more details on Leaf A-D routes, please refer to [MVPN-BGP].

6.3.2. Aggregation Methodology

This document does not specify the mandatory implementation of any particular set of rules for determining whether or not the PMSIs of two particular MVPNs are to be instantiated by the same Aggregate Tree. This determination can be made by implementation-specific heuristics, by configuration, or even perhaps by the use of offline tools.

It is the intention of this document that the control procedures will always result in all the PEs of an MVPN agreeing on the PMSIs that are to be used and on the tunnels used to instantiate those PMSIs.

This section discusses potential methodologies with respect to aggregation.

The "congruence" of aggregation is defined by the amount of overlap in the leaves of the customer trees that are aggregated on an SP tree. For Aggregate Trees with an inclusive mapping, the congruence depends on the overlap in the membership of the MVPNs that are aggregated on the tree. If there is complete overlap, i.e., all MVPNs have exactly the same sites, aggregation is perfectly congruent. As the overlap between the MVPNs that are aggregated reduces, i.e., the number of sites that are common across all the MVPNs reduces, the congruence reduces.

If aggregation is done such that it is not perfectly congruent, a PE may receive traffic for MVPNs to which it doesn't belong. As the amount of multicast traffic in these unwanted MVPNs increases, aggregation becomes less optimal with respect to delivered traffic. Hence, there is a trade-off between reducing state and delivering unwanted traffic.

An implementation should provide knobs to control the congruence of aggregation. These knobs are implementation dependent. Configuring the percentage of sites that MVPNs must have in common to be aggregated is an example of such a knob. This will allow an SP to deploy aggregation depending on the MVPN membership and traffic profiles in its network. If different PEs or servers are setting up Aggregate Trees, this will also allow a service provider to engineer the maximum amount of unwanted MVPNs for which a particular PE may receive traffic.

6.3.3. Demultiplexing C-Multicast Traffic

If a P-multicast tree is associated with only one MVPN, determining the P-multicast tree on which a packet was received is sufficient to determine the packet's MVPN. All that the egress PE needs to know is the MVPN with which the P-multicast tree is associated.

When multiple MVPNs are aggregated onto one P-multicast tree, determining the tree over which the packet is received is not sufficient to determine the MVPN to which the packet belongs. The packet must also carry some demultiplexing information to allow the egress PEs to determine the MVPN to which the packet belongs. Since the packet has been multicast through the P-network, any given demultiplexing value must have the same meaning to all the egress PEs. The demultiplexing value is a MPLS label that corresponds to the multicast VRF to which the packet belongs. This label is placed by the ingress PE immediately beneath the P-multicast tree header. Each of the egress PEs must be able to associate this MPLS label with the same MVPN. If downstream-assigned labels were used, this would require all the egress PEs in the MVPN to agree on a common label for the MVPN. Instead, the MPLS label is upstream-assigned [MPLS-UPSTREAM-LABEL]. The label bindings are advertised via BGP Updates originated by the ingress PEs.

This procedure requires each egress PE to support a separate label space for every other PE. The egress PEs create a forwarding entry for the upstream-assigned MPLS label, allocated by the ingress PE, in this label space. Hence, when the egress PE receives a packet over an Aggregate Tree, it first determines the tree over which the packet was received. The tree identifier determines the label space in which the upstream-assigned MPLS label lookup has to be performed.

The same label space may be used for all P-multicast trees rooted at the same ingress PE or an implementation may decide to use a separate label space for every P-multicast tree.

A full specification of the procedures to support aggregation on shared trees or on MP2MP LSPs is outside the scope of this document.

The encapsulation format is either MPLS or MPLS-in-something (e.g., MPLS-in-GRE [MPLS-IP]). When MPLS is used, this label will appear immediately below the label that identifies the P-multicast tree. When MPLS-in-GRE is used, this label will be the top MPLS label that appears when the GRE header is stripped off.

When IP encapsulation is used for the P-multicast tree, whatever information that particular encapsulation format uses for identifying a particular tunnel is used to determine the label space in which the MPLS label is looked up.

If the P-multicast tree uses MPLS encapsulation, the P-multicast tree is itself identified by an MPLS label. The egress PE MUST NOT advertise IMPLICIT NULL or EXPLICIT NULL for that tree. Once the label representing the tree is popped off the MPLS label stack, the next label is the demultiplexing information that allows the proper MVPN to be determined.

This specification requires that, to support this sort of aggregation, there be at least one upstream-assigned label per MVPN. It does not require that there be only one. For example, an ingress PE could assign a unique label to each (C-S,C-G). (This could be done using the same technique that is used to assign a particular (C-S,C-G) to an S-PMSI, see Section 7.4.)

When an egress PE receives a C-multicast data packet over a P-multicast tree, it needs to forward the packet to the CEs that have receivers in the packet’s C-multicast group. In order to do this, the egress PE needs to determine the P-tunnel on which the packet was received. The PE can then determine the MVPN that the packet belongs to and, if needed, do any further lookups that are needed to forward the packet.

6.4. Considerations for Specific Tunnel Technologies

While it is believed that the architecture specified in this document places no limitations on the protocols used for setting up and maintaining P-tunnels, the only protocols that have been explicitly considered are PIM-SM (both the SSM and ASM service models are

considered, as are bidirectional trees), RSVP-TE, mLDP, and BGP. (BGP's role in the setup and maintenance of P-tunnels is to "stitch" together the intra-AS segments of a segmented inter-AS P-tunnel.)

6.4.1. RSVP-TE P2MP LSPs

If an I-PMSI is to be instantiated as one or more non-segmented P-tunnels, where the P-tunnels are RSVP-TE P2MP LSPs, then only the PEs that are at the head ends of those LSPs will ever include the PMSI Tunnel attribute in their Intra-AS I-PMSI A-D routes. (These will be the PEs in the "Sender Sites set".)

If an I-PMSI is to be instantiated as one or more segmented P-tunnels, where some of the intra-AS segments of these tunnels are RSVP-TE P2MP LSPs, then only a PE or ASBR that is at the head end of one of these LSPs will ever include the PMSI Tunnel attribute in its Inter-AS I-PMSI A-D route.

Other PEs send Intra-AS I-PMSI A-D routes without PMSI Tunnel attributes. (These will be the PEs that are in the "Receiver Sites set" but not in the "Sender Sites set".) As each "Sender Site" PE receives an Intra-AS I-PMSI A-D route from a PE in the Receiver Sites set, it adds the PE originating that Intra-AS I-PMSI A-D route to the set of receiving PEs for the P2MP LSP. The PE at the head end MUST then use RSVP-TE [RSVP-P2MP] signaling to add the receiver PEs to the P-tunnel.

When RSVP-TE P2MP LSPs are used to instantiate S-PMSIs, and a particular C-flow is to be bound to the LSP, it is necessary to use explicit tracking so that the head end of the LSP knows which PEs need to receive data from the specified C-flow. If the binding is done using S-PMSI A-D routes (see Section 7.4.1), the "Leaf Information Required" bit MUST be set in the PMSI Tunnel attribute.

RSVP-TE P2MP LSPs can optionally support aggregation of multiple MVPNs.

If an RSVP-TE P2MP LSP Tunnel is used for only a single MVPN, the mapping between the LSP and the MVPN can either be configured or be deduced from the procedures used to announce the LSP (e.g., from the RTs in the A-D route that announced the LSP). If the LSP is used for multiple MVPNs, the set of MVPNs using it (and the corresponding MPLS labels) is inferred from the PMSI Tunnel attributes that specify the LSP.

If an RSVP-TE P2MP LSP is being used to carry a set of C-flows traveling along a bidirectional C-tree, using the procedures of Section 11.2, the head end MUST include the PE Distinguisher Labels

attribute in its Intra-AS I-PMSI A-D route or S-PMSI A-D route, and it MUST provide an upstream-assigned label for each PE that it has selected as the Upstream PE for the C-tree's RPA (Rendezvous Point Address). See Section 11.2 for details.

A PMSI Tunnel attribute specifying an RSVP-TE P2MP LSP contains the following information:

- The type of the tunnel is set to RSVP-TE P2MP Tunnel

- The RSVP-TE P2MP Tunnel's SESSION Object.

- Optionally, the RSVP-TE P2MP LSP's SENDER_TEMPLATE Object. This object is included when it is desired to identify a particular P2MP TE LSP.

Demultiplexing the C-multicast data packets at the egress PE follows procedures described in Section 6.3.3. As specified in Section 6.3.3, an egress PE MUST NOT advertise IMPLICIT NULL or EXPLICIT NULL for an RSVP-TE P2MP LSP that is carrying traffic for one or more MVPNs.

If (and only if) a particular RSVP-TE P2MP LSP is possibly carrying data from multiple MVPNs, the following special procedures apply:

- A packet in a particular MVPN, when transmitted into the LSP, must carry the MPLS label specified in the PMSI Tunnel attribute that announced that LSP as a P-tunnel for that for that MVPN.

- Demultiplexing the C-multicast data packets at the egress PE is done by means of the MPLS label that rises to the top of the stack after the label corresponding to the P2MP LSP is popped off.

It is possible that at the time a PE learns, via an A-D route with a PMSI Tunnel attribute, that it needs to receive traffic on a particular RSVP-TE P2MP LSP, the signaling to set up the LSP will not have been completed. In this case, the PE needs to wait for the RSVP-TE signaling to take place before it can modify its forwarding tables as directed by the A-D route.

It is also possible that the signaling to set up an RSVP-TE P2MP LSP will be completed before a given PE learns, via a PMSI Tunnel attribute, of the use to which that LSP will be put. The PE MUST discard any traffic received on that LSP until that time.

In order for the egress PE to be able to discard such traffic, it needs to know that the LSP is associated with an MVPN and that the A-D route that binds the LSP to an MVPN or to a particular a C-flow has not yet been received. This is provided by extending [RSVP-P2MP] with [RSVP-OOB].

6.4.2. PIM Trees

When the P-tunnels are PIM trees, the PMSI Tunnel attribute contains enough information to allow each other PE in the same MVPN to use P-PIM signaling to join the P-tunnel.

If an I-PMSI is to be instantiated as one or more PIM trees, then the PE that is at the root of a given PIM tree sends an Intra-AS I-PMSI A-D route containing a PMSI Tunnel attribute that contains all the information needed for other PEs to join the tree.

If PIM trees are to be used to instantiate an MI-PMSI, each PE in the MVPN must send an Intra-AS I-PMSI A-D route containing such a PMSI Tunnel attribute.

If a PMSI is to be instantiated via a shared tree, the PMSI Tunnel attribute identifies the P-group address. The RP or RPA corresponding to the P-group address is not specified. It must, of course, be known to all the PEs. It is presupposed that the PEs use one of the methods for automatically learning the RP-to-group correspondences (e.g., Bootstrap Router Protocol [BSR]), or else that the correspondence is configured.

If a PMSI is to be instantiated via a source-specific tree, the PMSI Tunnel attribute identifies the PE router that is the root of the tree, as well as a P-group address. The PMSI Tunnel attribute always specifies whether the PIM tree is to be a unidirectional shared tree, a bidirectional shared tree, or a source-specific tree.

If PIM trees are being used to instantiate S-PMSIs, the above procedures assume that each PE router has a set of group P-addresses that it can use for setting up the PIM-trees. Each PE must be configured with this set of P-addresses. If the P-tunnels are source-specific trees, then the PEs may be configured with overlapping sets of group P-addresses. If the trees are not source-specific, then each PE must be configured with a unique set of group P-addresses (i.e., having no overlap with the set configured at any other PE router). The management of this set of addresses is thus greatly simplified when source-specific trees are used, so the use of source-specific trees is strongly recommended whenever unidirectional trees are desired.

Specification of the full set of procedures for using bidirectional PIM trees to instantiate S-PMSIs is outside the scope of this document.

Details for constructing the PMSI Tunnel attribute identifying a PIM tree can be found in [MVPN-BGP].

6.4.3. mLDP P2MP LSPs

When the P-tunnels are mLDP P2MP trees, each Intra-AS I-PMSI A-D route has a PMSI Tunnel attribute containing enough information to allow each other PE in the same MVPN to use mLDP signaling to join the P-tunnel. The tunnel identifier consists of a P2MP Forwarding Equivalence Class (FEC) Element [mLDP].

An mLDP P2MP LSP may be used to carry the traffic of multiple VPNs, if the PMSI Tunnel attribute specifying it contains a non-zero MPLS label.

If an mLDP P2MP LSP is being used to carry the set of flows traveling along a particular bidirectional C-tree, using the procedures of Section 11.2, the root of the LSP MUST include the PE Distinguisher Labels attribute in its Intra-AS I-PMSI A-D route or S-PMSI A-D route, and it MUST provide an upstream-assigned label for the PE that it has selected to be the Upstream PE for the C-tree's RPA. See Section 11.2 for details.

6.4.4. mLDP MP2MP LSPs

The specification of the procedures for assigning C-flows to mLDP MP2MP LSPs that serve as P-tunnels is outside the scope of this document.

6.4.5. Ingress Replication

As described in Section 3, a PMSI can be instantiated using Unicast Tunnels between the PEs that are participating in the MVPN. In this mechanism, the ingress PE replicates a C-multicast data packet belonging to a particular MVPN and sends a copy to all or a subset of the PEs that belong to the MVPN. A copy of the packet is tunneled to a remote PE over a Unicast Tunnel to the remote PE. IP/GRE Tunnels or MPLS LSPs are examples of unicast tunnels that may be used. The same Unicast Tunnel can be used to transport packets belonging to different MVPNs

In order for a PE to use Unicast P-tunnels to send a C-multicast data packet for a particular MVPN to a set of remote PEs, the remote PEs must be able to correctly decapsulate such packets and to assign each

one to the proper MVPN. This requires that the encapsulation used for sending packets through the P-tunnel have demultiplexing information that the receiver can associate with a particular MVPN.

If ingress replication is being used to instantiate the PMSIs for an MVPN, the PEs announce this as part of the BGP-based MVPN membership auto-discovery process, described in Section 4. The PMSI Tunnel attribute specifies ingress replication; it also specifies a downstream-assigned MPLS label. This label will be used to identify that a particular packet belongs to the MVPN that the Intra-AS I-PMSI A-D route belongs to (as inferred from its RTs). If PE1 specifies a particular label value for a particular MVPN, then any other PE sending PE1 a packet for that MVPN through a unicast P-tunnel must put that label on the packet's label stack. PE1 then treats that label as the demultiplexor value identifying the MVPN in question.

Ingress replication may be used to instantiate any kind of PMSI. When ingress replication is done, it is RECOMMENDED, except in the one particular case mentioned in the next paragraph, that explicit tracking be done and that the data packets of a particular C-flow only get sent to those PEs that need to see the packets of that C-flow. There is never any need to use the procedures of Section 7.4 for binding particular C-flows to particular P-tunnels.

The particular case in which there is no need for explicit tracking is the case where ingress replication is being used to create a one-hop ASBR-ASBR inter-AS segment of an segmented inter-AS P-tunnel.

Section 9.1 specifies three different methods that can be used to prevent duplication of multicast data packets. Any given deployment must use at least one of those methods. Note that the method described in Section 9.1.1 ("Discarding Packets from Wrong PE") presupposes that the egress PE of a P-tunnel can, upon receiving a packet from the P-tunnel, determine the identity of the PE that transmitted the packet into the P-tunnel. SPs that use ingress replication to instantiate their PMSIs are cautioned against this use for this purpose of unicast P-tunnel technologies that do not allow the egress PE to identify the ingress PE (e.g., MP2P LSPs for which penultimate-hop-popping is done). Deployment of ingress replication with such P-tunnel technology MUST NOT be done unless it is known that the deployment relies entirely on the procedures of Sections 9.1.2 or 9.1.3 for duplicate prevention.

7. Binding Specific C-Flows to Specific P-Tunnels

As discussed previously, Intra-AS I-PMSI A-D routes may (or may not) have PMSI Tunnel attributes, identifying P-tunnels that can be used as the default P-tunnels for carrying C-multicast traffic, i.e., for carrying C-multicast traffic that has not been specifically bound to another P-tunnel.

If none of the Intra-AS I-PMSI A-D routes originated by a particular PE for a particular MVPN carry PMSI Tunnel attributes at all (or if the only PMSI Tunnel attributes they carry have type "No tunnel information present"), then there are no default P-tunnels for that PE to use when transmitting C-multicast traffic in that MVPN to other PEs. In that case, all such C-flows must be assigned to specific P-tunnels using one of the mechanisms specified in Section 7.4. That is, all such C-flows are carried on P-tunnels that instantiate S-PMSIs.

There are other cases where it may be either necessary or desirable to use the mechanisms of Section 7.4 to identify specific C-flows and bind them to or unbind them from specific P-tunnels. Some possible cases are as follows:

- The policy for a particular MVPN is to send all C-data on S-PMSIs, even if the Intra-AS I-PMSI A-D routes carry PMSI Tunnel attributes. (This is another case where all C-data is carried on S-PMSIs; presumably, the I-PMSIs are used for control information.)

- It is desired to optimize the routing of the particular C-flow, which may already be traveling on an I-PMSI, by sending it instead on an S-PMSI.

- If a particular C-flow is traveling on an S-PMSI, it may be considered desirable to move it to an I-PMSI (i.e., optimization of the routing for that flow may no longer be considered desirable).

- It is desired to change the encapsulation used to carry the C-flow, e.g., because one now wants to aggregate it on a P-tunnel with flows from other MVPNs.

Note that if Full PIM Peering over an MI-PMSI (Section 5.2) is being used, then from the perspective of the PIM state machine, the "interface" connecting the PEs to each other is the MI-PMSI, even if some or all of the C-flows are being sent on S-PMSIs. That is, from

the perspective of the C-PIM state machine, when a C-flow is being sent or received on an S-PMSI, the output or input interface (respectively) is considered to be the MI-PMSI.

Section 7.1 discusses certain general considerations that apply whenever a specified C-flow is bound to a specified P-tunnel using the mechanisms of Section 7.4. This includes the case where the C-flow is moved from one P-tunnel to another as well as the case where the C-flow is initially bound to an S-PMSI P-tunnel.

Section 7.2 discusses the specific case of using the mechanisms of Section 7.4 as a way of optimizing multicast routing by switching specific flows from one P-tunnel to another.

Section 7.3 discusses the case where the mechanisms of Section 7.4 are used to announce the presence of "unsolicited flooded data" and to assign such data to a particular P-tunnel.

Section 7.4 specifies the protocols for assigning specific C-flows to specific P-tunnels. These protocols may be used to assign a C-flow to a P-tunnel initially or to switch a flow from one P-tunnel to another.

Procedures for binding to a specified P-tunnel the set of C-flows traveling along a specified C-tree (or for so binding a set of C-flows that share some relevant characteristic), without identifying each flow individually, are outside the scope of this document.

7.1. General Considerations

7.1.1. At the PE Transmitting the C-Flow on the P-Tunnel

The decision to bind a particular C-flow (designated as (C-S,C-G)) to a particular P-tunnel, or to switch a particular C-flow to a particular P-tunnel, is always made by the PE that is to transmit the C-flow onto the P-tunnel.

Whenever a PE moves a particular C-flow from one P-tunnel, say P1, to another, say P2, care must be taken to ensure that there is no steady state duplication of traffic. At any given time, the PE transmits the C-flow either on P1 or on P2, but not on both.

When a particular PE, say PE1, decides to bind a particular C-flow to a particular P-tunnel, say P2, the following procedures MUST be applied:

- PE1 must issue the required control plane information to signal that the specified C-flow is now bound to P-tunnel P2 (see Section 7.4).

- If P-tunnel P2 needs to be constructed from the root downwards, PE1 must initiate the signaling to construct P2. This is only required if P2 is an RSVP-TE P2MP LSP.

- If the specified C-flow is currently bound to a different P-tunnel, say P1, then:

  * PE1 MUST wait for a "switch-over" delay before sending traffic of the C-flow on P-tunnel P2. It is RECOMMENDED to allow this delay to be configurable.

  * Once the "switch-over" delay has elapsed, PE1 MUST send traffic for the C-flow on P2 and MUST NOT send it on P1. In no case is any C-flow packet sent on both P-tunnels.

When a C-flow is switched from one P-tunnel to another, the purpose of running a switch-over timer is to minimize packet loss without introducing packet duplication. However, jitter may be introduced due to the difference in transit delays between the old and new P-tunnels.

For best effect, the switch-over timer should be configured to a value that is "just long enough" (a) to allow all the PEs to learn about the new binding of C-flow to P-tunnel and (b) to allow the PEs to construct the P-tunnel, if it doesn’t already exist.

If, after such a switch, the "old" P-tunnel P1 is no longer needed, it SHOULD be torn down and the resources supporting it freed. The procedures for "tearing down" a P-tunnel are specific to the P-tunnel technology.

Procedures for binding sets of C-flows traveling along specified C-trees (or sets of C-flows sharing any other characteristic) to a specified P-tunnel (or for moving them from one P-tunnel to another) are outside the scope of this document.

7.1.2. At the PE Receiving the C-flow from the P-Tunnel

Suppose that a particular PE, say PE1, learns, via the procedures of Section 7.4, that some other PE, say PE2, has bound a particular C-flow, designated as (C-S,C-G), to a particular P-tunnel, say P2. Then, PE1 must determine whether it needs to receive (C-S,C-G) traffic from PE2.

If BGP is being used to distribute C-multicast routing information from PE to PE, the conditions under which PE1 needs to receive (C-S,C-G) traffic from PE2 are specified in Section 12.3 of [MVPN-BGP].

If PIM over an MI-PMSI is being used to distribute C-multicast routing from PE to PE, PE1 needs to receive (C-S,C-G) traffic from PE2 if one or more of the following conditions holds:

- PE1 has (C-S,C-G) state such that PE2 is PE1’s Upstream PE for (C-S,C-G), and PE1 has downstream neighbors ("non-null olist") for the (C-S,C-G) state.

- PE1 has (C-*,C-G) state with an Upstream PE (not necessarily PE2) and with downstream neighbors ("non-null olist"), but PE1 does not have (C-S,C-G) state.

- Native PIM methods are being used to prevent steady-state packet duplication, and PE1 has either (C-*,C-G) or (C-S,C-G) state such that the MI-PMSI is one of the downstream interfaces. Note that this includes the case where PE1 is itself sending (C-S,C-G) traffic on an S-PMSI. (In this case, PE1 needs to receive the (C-S,C-G) traffic from PE2 in order to allow the PIM Assert mechanism to function properly.)

Irrespective of whether BGP or PIM is being used to distribute C-multicast routing information, once PE1 determines that it needs to receive (C-S,C-G) traffic from PE2, the following procedures MUST be applied:

- PE1 MUST take all necessary steps to be able to receive the (C-S,C-G) traffic on P2.

  * If P2 is a PIM tunnel or an mLDP LSP, PE1 will need to use PIM or mLDP (respectively) to join P2 (unless it is already joined to P2).

  * PE1 may need to modify the forwarding state for (C-S,C-G) to indicate that (C-S,C-G) traffic is to be accepted on P2. If P2 is an Aggregate Tree, this also implies setting up the demultiplexing forwarding entries based on the inner label as described in Section 6.3.3

- If PE1 was previously receiving the (C-S,C-G) C-flow on another P-tunnel, say P1, then:

  * PE1 MAY run a switch-over timer, and until it expires, SHOULD accept traffic for the given C-flow on both P1 and P2;

    * If, after such a switch, the "old" P-tunnel P1 is no longer needed, it SHOULD be torn down and the resources supporting it freed. The procedures for "tearing down" a P-tunnel are specific to the P-tunnel technology.

- If PE1 later determines that it no longer needs to receive any of the C-multicast data that is being sent on a particular P-tunnel, it may initiate signaling (specific to the P-tunnel technology) to remove itself from that tunnel.

7.2. Optimizing Multicast Distribution via S-PMSIs

Whenever a particular multicast stream is being sent on an I-PMSI, it is likely that the data of that stream is being sent to PEs that do not require it. If a particular stream has a significant amount of traffic, it may be beneficial to move it to an S-PMSI that includes only those PEs that are transmitters and/or receivers (or at least includes fewer PEs that are neither).

If explicit tracking is being done, S-PMSI creation can also be triggered on other criteria. For instance, there could be a "pseudo-wasted bandwidth" criterion: switching to an S-PMSI would be done if the bandwidth multiplied by the number of uninterested PEs (PE that are receiving the stream but have no receivers) is above a specified threshold. The motivation is that (a) the total bandwidth wasted by many sparsely subscribed low-bandwidth groups may be large and (b) there’s no point to moving a high-bandwidth group to an S-PMSI if all the PEs have receivers for it.

Switching a (C-S,C-G) stream to an S-PMSI may require the root of the S-PMSI to determine the egress PEs that need to receive the (C-S,C-G) traffic. This is true in the following cases:

- If the P-tunnel is a source-initiated tree, such as an RSVP-TE P2MP Tunnel, the PE needs to know the leaves of the tree before it can instantiate the S-PMSI.

- If a PE instantiates multiple S-PMSIs, belonging to different MVPNs, using one P-multicast tree, such a tree is termed an Aggregate Tree with a selective mapping. The setting up of such an Aggregate Tree requires the ingress PE to know all the other PEs that have receivers for multicast groups that are mapped onto the tree.

The above two cases require that explicit tracking be done for the (C-S,C-G) stream. The root of the S-PMSI MAY decide to do explicit tracking of this stream only after it has determined to move the stream to an S-PMSI, or it MAY have been doing explicit tracking all along.

If the S-PMSI is instantiated by a P-multicast tree, the PE at the root of the tree must signal the leaves of the tree that the (C-S,C-G) stream is now bound to the S-PMSI. Note that the PE could create the identity of the P-multicast tree prior to the actual instantiation of the P-tunnel.

If the S-PMSI is instantiated by a source-initiated P-multicast tree (e.g., an RSVP-TE P2MP tunnel), the PE at the root of the tree must establish the source-initiated P-multicast tree to the leaves. This tree MAY have been established before the leaves receive the S-PMSI binding, or it MAY be established after the leaves receive the binding. The leaves MUST NOT switch to the S-PMSI until they receive both the binding and the tree signaling message.

7.3. Announcing the Presence of Unsolicited Flooded Data

A PE may receive "unsolicited" data from a CE, where the data is intended to be flooded to the other PEs of the same MVPN and then on to other CEs. By "unsolicited", we mean that the data is to be delivered to all the other PEs of the MVPN, even though those PEs may not have sent any control information indicating that they need to receive that data.

For example, if the BSR [BSR] is being used within the MVPN, BSR control messages may be received by a PE from a CE. These need to be forwarded to other PEs, even though no PE ever issues any kind of explicit signal saying that it wants to receive BSR messages.

If a PE receives a BSR message from a CE, and if the CE’s MVPN has an MI-PMSI, then the PE can just send BSR messages on the appropriate P-tunnel. Otherwise, the PE MUST announce the binding of a particular C-flow to a particular P-tunnel, using the procedures of Section 7.4. The particular C-flow in this case would be (C-IPaddress_of_PE, ALL-PIM-ROUTERS). The P-tunnel identified by the procedures of Section 7.4 may or may not be one that was previously identified in the PMSI Tunnel attribute of an I-PMSI A-D route. Further procedures for handling BSR may be found in Sections 5.2.1 and 5.3.4.

Analogous procedures may be used for announcing the presence of other sorts of unsolicited flooded data, e.g., dense mode data or data from proprietary protocols that presume messages can be flooded. However, a full specification of the procedures for traffic other than BSR traffic is outside the scope of this document.

7.4. Protocols for Binding C-Flows to P-Tunnels

We describe two protocols for binding C-flows to P-tunnels.

These protocols can be used for moving C-flows from I-PMSIs to S-PMSIs, as long as the S-PMSI is instantiated by a P-multicast tree. (If the S-PMSI is instantiated by means of ingress replication, the procedures of Section 6.4.5 suffice.)

These protocols can also be used for other cases in which it is necessary to bind specific C-flows to specific P-tunnels.

7.4.1. Using BGP S-PMSI A-D Routes

Not withstanding the name of the mechanism "S-PMSI A-D routes", the mechanism to be specified in this section may be used any time it is necessary to advertise a binding of a C-flow to a particular P-tunnel.

7.4.1.1. Advertising C-Flow Binding to P-Tunnel

The ingress PE informs all the PEs that are on the path to receivers of the (C-S,C-G) of the binding of the P-tunnel to the (C-S,C-G). The BGP announcement is done by sending an update for the MCAST-VPN address family. An S-PMSI A-D route is used, containing the following information:

1. The IP address of the originating PE.

2. The RD configured locally for the MVPN. This is required to uniquely identify the (C-S,C-G) as the addresses could overlap between different MVPNs. This is the same RD value used in the auto-discovery process.

3. The C-S address.

4. The C-G address.

5. A PE MAY use a single P-tunnel to aggregate two or more S-PMSIs. If the PE already advertised unaggregated S-PMSI A-D routes for these S-PMSIs, then a decision to aggregate them requires the PE to re-advertise these routes. The re-

advertised routes MUST be the same as the original ones, except for the PMSI Tunnel attribute. If the PE has not previously advertised S-PMSI A-D routes for these S-PMSIs, then the aggregation requires the PE to advertise (new) S-PMSI A-D routes for these S-PMSIs. The PMSI Tunnel attribute in the newly advertised/re-advertised routes MUST carry the identity of the P-tunnel that aggregates the S-PMSIs.

If all these aggregated S-PMSIs belong to the same MVPN, and this MVPN uses PIM as its C-multicast routing protocol, then the corresponding S-PMSI A-D routes MAY carry an MPLS upstream-assigned label [MPLS-UPSTREAM-LABEL]. Moreover, in this case, the labels MUST be distinct on a per-MVPN basis, and MAY be distinct on a per-route basis.

If all these aggregated S-PMSIs belong to the MVPN(s) that use mLDP as its C-multicast routing protocol, then the corresponding S-PMSI A-D routes MUST carry an MPLS upstream-assigned label [MPLS-UPSTREAM-LABEL], and these labels MUST be distinct on a per-route (per-mLDP-FEC) basis, irrespective of whether the aggregated S-PMSIs belong to the same or different MVPNs.

When a PE distributes this information via BGP, it must include the following:

1. An identifier for the particular P-tunnel to which the stream is to be bound. This identifier is a structured field that includes the following information:

   * The type of tunnel

   * An identifier for the tunnel. The form of the identifier will depend upon the tunnel type. The combination of tunnel identifier and tunnel type should contain enough information to enable all the PEs to "join" the tunnel and receive messages from it.

2. Route Target Extended Communities attribute. This is used as described in Section 4.

7.4.1.2. Explicit Tracking

If the PE wants to enable explicit tracking for the specified flow, it also indicates this in the A-D route it uses to bind the flow to a particular P-tunnel. Then, any PE that receives the A-D route will

respond with a "Leaf A-D route" in which it identifies itself as a receiver of the specified flow. The Leaf A-D route will be withdrawn when the PE is no longer a receiver for the flow.

If the PE needs to enable explicit tracking for a flow without at the same time binding the flow to a specific P-tunnel, it can do so by sending an S-PMSI A-D route whose NLRI identifies the flow and whose PMSI Tunnel attribute has its tunnel type value set to "no tunnel information present" and its "leaf information required" bit set to 1. This will elicit the Leaf A-D routes. This is useful when the PE needs to know the receivers before selecting a P-tunnel.

7.4.2. UDP-Based Protocol

This procedure carries its control messages in UDP and requires that the MVPN have an MI-PMSI that can be used to carry the control messages.

7.4.2.1. Advertising C-Flow Binding to P-Tunnel

In order for a given PE to move a particular C-flow to a particular P-tunnel, an "S-PMSI Join message" is sent periodically on the MI-PMSI. (Notwithstanding the name of the mechanism, the mechanism may be used to bind a flow to any P-tunnel.) The S-PMSI Join message is a UDP-encapsulated message whose destination address is ALL-PIM-ROUTERS (224.0.0.13) and whose destination port is 3232.

The S-PMSI Join message contains the following information:

- An identifier for the particular multicast stream that is to be bound to the P-tunnel. This can be represented as an (S,G) pair.

- An identifier for the particular P-tunnel to which the stream is to be bound. This identifier is a structured field that includes the following information:

  * The type of tunnel used to instantiate the S-PMSI.

  * An identifier for the tunnel. The form of the identifier will depend upon the tunnel type. The combination of tunnel identifier and tunnel type should contain enough information to enable all the PEs to "join" the tunnel and receive messages from it.

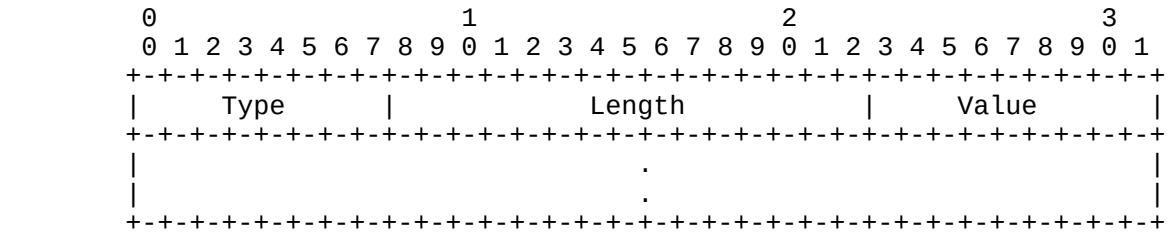
  * If (and only if) the identified P-tunnel is aggregating several S-PMSIs, any demultiplexing information needed by the tunnel encapsulation protocol to identify a particular S-PMSI.

If the policy for the MVPN is that traffic is sent/received by default over an MI-PMSI, then traffic for a particular C-flow can be switched back to the MI-PMSI simply by ceasing to send S-PMSI Joins for that C-flow.

Note that an S-PMSI Join that is not received over a PMSI (e.g., one that is received directly from a CE) is an illegal packet that MUST be discarded.

7.4.2.2. Packet Formats and Constants

The S-PMSI Join message is encapsulated within UDP and has the following type/length/value (TLV) encoding:



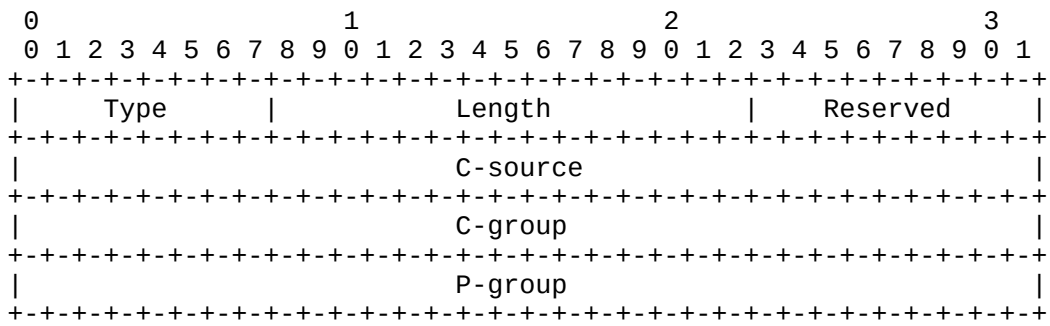
Type (8 bits)

Length (16 bits): the total number of octets in the Type, Length, and Value fields combined

Value (variable length)

In this specification, only one type of S-PMSI Join is defined. A Type 1 S-PMSI Join is used when the S-PMSI tunnel is a PIM tunnel that is used to carry a single multicast stream, where the packets of that stream have IPv4 source and destination IP addresses.

The S-PMSI Join format to use when the C-source and C-group are IPv6 addresses will be defined in a follow-on document.



Type (8 bits): 1

Length (16 bits): 16

Reserved (8 bits): This field SHOULD be zero when transmitted, and MUST be ignored when received.

C-source (32 bits): the IPv4 address of the traffic source in the VPN.

C-group (32 bits): the IPv4 address of the multicast traffic destination address in the VPN.

P-group (32 bits): the IPv4 group address that the PE router is going to use to encapsulate the flow (C-source, C-group).

The P-group identifies the S-PMSI P-tunnel, and the (C-S,C-G) identifies the multicast flow that is carried in the P-tunnel.

The protocol uses the following constants.

[S-PMSI_DELAY]:

Once an S-PMSI Join message has been sent, the PE router that is to transmit onto the S-PMSI will delay this amount of time before it begins using the S-PMSI. The default value is 3 seconds.

[S-PMSI_TIMEOUT]:

If a PE (other than the transmitter) does not receive any packets over the S-PMSI P-tunnel for this amount of time, the PE will prune itself from the S-PMSI P-tunnel, and will expect (C-S,C-G) packets to arrive on an I-PMSI. The default value is 3 minutes.

This value must be consistent among PE routers.

[S-PMSI_HOLDOWN]:

If the PE that transmits onto the S-PMSI does not see any (C-S,C-G) packets for this amount of time, it will resume sending (C-S,C-G) packets on an I-PMSI.

This is used to avoid oscillation when traffic is bursty. The default value is 1 minute.

[S-PMSI_INTERVAL]:

The interval the transmitting PE router uses to periodically send the S-PMSI Join message. The default value is 60 seconds.

7.4.3. Aggregation

S-PMSIs can be aggregated on a P-multicast tree. The S-PMSI to (C-S,C-G) binding advertisement supports aggregation. Furthermore, the aggregation procedures of Section 6.3 apply. It is also possible to aggregate both S-PMSIs and I-PMSIs on the same P-multicast tree.

8. Inter-AS Procedures

If an MVPN has sites in more than one AS, it requires one or more PMSIs to be instantiated by inter-AS P-tunnels. This document describes two different types of inter-AS P-tunnel:

1. "Segmented inter-AS P-tunnels"

   A segmented inter-AS P-tunnel consists of a number of independent segments that are stitched together at the ASBRs. There are two types of segment: inter-AS segments and intra-AS segments. The segmented inter-AS P-tunnel consists of alternating intra-AS and inter-AS segments.

   Inter-AS segments connect adjacent ASBRs of different ASes; these "one-hop" segments are instantiated as unicast P-tunnels.

   Intra-AS segments connect ASBRs and PEs that are in the same AS. An intra-AS segment may be of whatever technology is desired by the SP that administers the that AS. Different intra-AS segments may be of different technologies.

   Note that the intra-AS segments of inter-AS P-tunnels form a category of P-tunnels that is distinct from simple intra-AS P-tunnels; we will rely on this distinction later (see Section 9).

A segmented inter-AS P-tunnel can be thought of as a tree that is rooted at a particular AS, and that has, as its leaves, the other ASes that need to receive multicast data from the root AS.

2. "Non-segmented Inter-AS P-tunnels"

   A non-segmented inter-AS P-tunnel is a single P-tunnel that spans AS boundaries. The tunnel technology cannot change from one point in the tunnel to the next, so all ASes through which the P-tunnel passes must support that technology. In essence, AS boundaries are of no significance to a non-segmented inter-AS P-tunnel.

Section 10 of [RFC4364] describes three different options for supporting unicast inter-AS BGP/MPLS IP VPNs, known as options A, B, and C. We describe below how both segmented and non-segmented inter-AS trees can be supported when options B or C are used. (Option A does not pass any routing information through an ASBR at all, so no special inter-AS procedures are needed.)

8.1. Non-Segmented Inter-AS P-Tunnels

In this model, the previously described discovery and tunnel setup mechanisms are used, even though the PEs belonging to a given MVPN may be in different ASes.

8.1.1. Inter-AS MVPN Auto-Discovery

The previously described BGP-based auto-discovery mechanisms work "as is" when an MVPN contains PEs that are in different Autonomous Systems. However, please note that, if non-segmented inter-AS P-tunnels are to be used, then the Intra-AS I-PMSI A-D routes MUST be distributed across AS boundaries!

8.1.2. Inter-AS MVPN Routing Information Exchange

When non-segmented inter-AS P-tunnels are used, MVPN C-multicast routing information may be exchanged by means of PIM peering across an MI-PMSI or by means of BGP carrying C-multicast routes.

When PIM peering is used to distribute the C-multicast routing information, a PE that sends C-PIM Join/Prune messages for a particular (C-S,C-G) must be able to identify the PE that is its PIM adjacency on the path to S. This is the "Selected Upstream PE" described in Section 5.1.3.

If BGP (rather than PIM) is used to distribute the C-multicast routing information, and if option b of Section 10 of [RFC4364] is in use, then the C-multicast routes will be installed in the ASBRs along the path from each multicast source in the MVPN to each multicast receiver in the MVPN. If option b is not in use, the C-multicast routes are not installed in the ASBRs. The handling of the C-multicast routes in either case is thus exactly analogous to the handling of unicast VPN-IP routes in the corresponding case.

8.1.3. Inter-AS P-Tunnels

The procedures described earlier in this document can be used to instantiate either an I-PMSI or an S-PMSI with inter-AS P-tunnels. Specific tunneling techniques require some explanation.

If ingress replication is used, the inter-AS PE-PE P-tunnels will use the inter-AS tunneling procedures for the tunneling technology used.

Procedures in [RSVP-P2MP] are used for inter-AS RSVP-TE P2MP P-tunnels.

Procedures for using PIM to set up the P-tunnels are discussed in the next section.

8.1.3.1. PIM-Based Inter-AS P-Multicast Trees

When PIM is used to set up a non-segmented inter-AS P-multicast tree, the PIM Join/Prune messages used to join the tree contain the IP address of the Upstream PE. However, there are two special considerations that must be taken into account:

- It is possible that the P routers within one or more of the ASes will not have routes to the Upstream PE. For example, if an AS has a "BGP-free core", the P routers in an AS will not have routes to addresses outside the AS.

- If the PIM Join/Prune message must travel through several ASes, it is possible that the ASBRs will not have routes to he PE routers. For example, in an inter-AS VPN constructed according to "option b" of Section 10 of [RFC4364], the ASBRs do not necessarily have routes to the PE routers.

In either case, "ordinary" PIM Join/Prune messages cannot be routed to the Upstream PE. Therefore, in that case, the PIM Join/Prune messages MUST contain the "PIM MVPN Join attribute". This allows the multicast distribution tree to be properly constructed, even if routes to PEs in other ASes do not exist in the given AS’s IGP and

even if the routes to those PEs do not exist in BGP. The use of a PIM MVPN Join attribute in the PIM messages allows the inter-AS trees to be built.

The PIM MVPN Join attribute adds the following information to the PIM Join/Prune messages: a "proxy address", which contains the address of the next ASBR on the path to the Upstream PE. When the PIM Join/Prune arrives at the ASBR that is identified by the "proxy address", that ASBR must change the proxy address to identify the next hop ASBR.

This information allows the PIM Join/Prune to be routed through an AS, even if the P routers of that AS do not have routes to the Upstream PE. However, this information is not sufficient to enable the ASBRs to route the Join/Prune if the ASBRs themselves do not have routes to the Upstream PE.

However, even if the ASBRs do not have routes to the Upstream PE, the procedures of this document ensure that they will have Intra-AS I-PMSI A-D routes that lead to the Upstream PE. (Recall that if non-segmented inter-AS P-tunnels are being used, the ASBRs and PEs will have Intra-AS I-PMSI A-D routes that have been distributed inter-AS.)

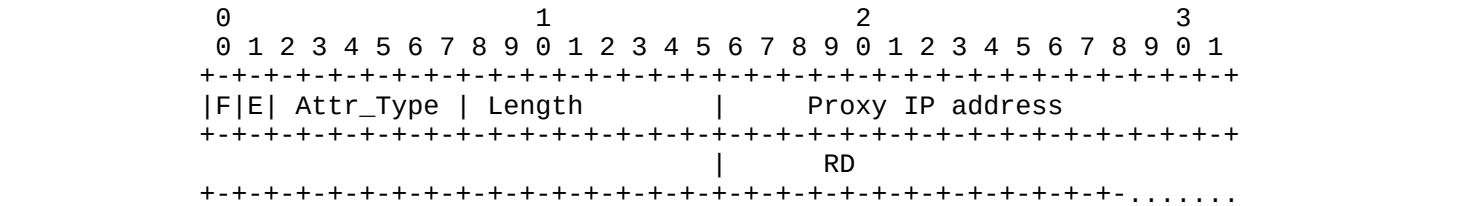
So, rather than having the PIM Join/Prune messages routed by the ASBRs along a route to the Upstream PE, the PIM Join/Prune messages MUST be routed along the path determined by the Intra-AS I-PMSI A-D routes.

The basic format of a PIM Join attribute is specified in [PIM-ATTRIB]. The details of the PIM MVPN Join attribute are specified in the next section.

8.1.3.2. The PIM MVPN Join Attribute

8.1.3.2.1. Definition

In [PIM-ATTRIB], the notion of a "join attribute" is defined, and a format for included join attributes in PIM Join/Prune messages is specified. We now define a new join attribute, which we call the "MVPN Join attribute".



The Attr_Type field of the MVPN Join attribute is set to 1.

The F bit is set to 0.

Two information fields are carried in the MVPN Join attribute:

- Proxy IP address: The IP address of the node towards which the PIM Join/Prune message is to be forwarded. This will be either an IPv4 or an IPv6 address, depending on whether the PIM Join/Prune message itself is IPv4 or IPv6.

- RD: An eight-byte RD. This immediately follows the proxy IP address.

The PIM message also carries the address of the Upstream PE.

In the case of an intra-AS MVPN, the proxy and the Upstream PE are the same. In the case of an inter-AS MVPN, the proxy will be the ASBR that is the exit point from the local AS on the path to the Upstream PE.

8.1.3.2.2. Usage

When a PE router originates a PIM Join/Prune message in order to set up an inter-AS PMSI, it does so as a result of having received a particular Intra-AS I-PMSI A-D route or S-PMSI A-D route. It includes an MVPN Join attribute whose fields are set as follows:

- If the Upstream PE is in the same AS as the local PE, then the proxy field contains the address of the Upstream PE. Otherwise, it contains the address of the BGP Next Hop of the route to the Upstream PE.

- The RD field contains the RD from the NLRI of the Intra-AS A-D route.

- The Upstream PE field contains the address of the PE that originated the Intra-AS I-PMSI A-D route or S-PMSI A-D route (obtained from the NLRI of that route).

When a PIM router processes a PIM Join/Prune message with an MVPN Join attribute, it first checks to see if the proxy field contains one of its own addresses.

If not, the router uses the proxy IP address in order to determine the RPF interface and neighbor. The MVPN Join attribute must be passed upstream unchanged.

If the proxy address is one of the router's own IP addresses, then the router looks in its BGP routing table for an Intra-AS A-D route whose NLRI consists of the Upstream PE address prepended with the RD from the Join attribute. If there is no match, the PIM message is discarded. If there is a match, the IP address from the BGP next hop field of the matching route is used in order to determine the RPF interface and neighbor. When the PIM Join/Prune is forwarded upstream, the proxy field is replaced with the address of the BGP next hop, and the RD and Upstream PE fields are left unchanged.

The use of non-segmented inter-AS trees constructed via BIDIR-PIM is outside the scope of this document.

8.2. Segmented Inter-AS P-Tunnels

The procedures for setting up and maintaining segmented inter-AS Inclusive and Selective P-tunnels may be found in [MVPN-BGP].

9. Preventing Duplication of Multicast Data Packets

Consider the case of an egress PE that receives packets of a particular C-flow, (C-S,C-G), over a non-aggregated S-PMSI. The procedures described so far will never cause the PE to receive duplicate copies of any packet in that stream. It is possible that the (C-S,C-G) stream is carried in more than one S-PMSI; this may happen when the site that contains C-S is multihomed to more than one PE. However, a PE that needs to receive (C-S,C-G) packets only joins one of these S-PMSIs, and so it only receives one copy of each packet. However, if the data packets of stream (C-S,C-G) are carried in either an I-PMSI or an aggregated S-PMSI, then the procedures specified so far make it possible for an egress PE to receive more than one copy of each data packet. Additional procedures are needed to either make this impossible or ensure that the egress PE does not forward duplicates to the CE routers.

This section covers only the situation where the C-trees are unidirectional, in either the ASM or SSM service models. The case where the C-trees are bidirectional is considered separately in Section 11.

There are two cases where the procedures specified so far make it possible for an egress PE to receive duplicate copies of a multicast data packet. These are as follows:

1. The first case occurs when both of the following conditions hold:

a. an MVPN site that contains C-S or C-RP is multihomed to more than one PE, and

b. either an I-PMSI or an aggregated S-PMSI is used for carrying the packets originated by C-S.

In this case, an egress PE may receive one copy of the packet from each PE to which the site is homed. This case is discussed further in Section 9.2.

2. The second case occurs when all of the following conditions hold:

a. the IP destination address of the customer packet, C-G, identifies a multicast group that is operating in ASM mode and whose C-multicast tree is set up using PIM-SM,

b. an MI-PMSI is used for carrying the data packets, and

c. a router or a CE in a site connected to the egress PE switches from the C-RP tree to the C-S tree.

In this case, it is possible to get one copy of a given packet from the ingress PE attached to the C-RP's site and one from the ingress PE attached to the C-S's site. This case is discussed further in Section 9.3.

Additional procedures are therefore needed to ensure that no MVPN customer sees steady state multicast data packet duplication. There are three procedures that may be used:

1. Discarding data packets received from the "wrong" PE

2. Single Forwarder Selection

3. Native PIM methods

These methods are described in Section 9.1. Their applicability to the two scenarios where duplication is possible is discussed in Sections 9.2 and 9.3.

9.1. Methods for Ensuring Non-Duplication

Every MVPN MUST use at least one of the three methods for ensuring non-duplication.

9.1.1. Discarding Packets from Wrong PE

Per Section 5.1.3, an egress PE, say PE1, chooses a specific Upstream PE, for given (C-S,C-G). When PE1 receives a (C-S,C-G) packet from a PMSI, it may be able to identify the PE that transmitted the packet onto the PMSI. If that transmitter is other than the PE selected by PE1 as the Upstream PE, then PE1 can drop the packet. This means that the PE will see a duplicate, but the duplicate will not get forwarded.

The method used by an egress PE to determine the ingress PE for a particular packet, received over a particular PMSI, depends on the P-tunnel technology that is used to instantiate the PMSI. If the P-tunnel is a P2MP LSP, a PIM-SM or PIM-SSM tree, or a unicast P-tunnel that uses IP encapsulation, then the tunnel encapsulation contains information that can be used (possibly along with other state information in the PE) to determine the ingress PE, as long as the P-tunnel is instantiating an intra-AS PMSI or an inter-AS PMSI which is supported by a non-segmented inter-AS tunnel.

Even when inter-AS segmented P-tunnels are used, if an aggregated S-PMSI is used for carrying the packets, the tunnel encapsulation must have some information that can be used to identify the PMSI; in turn, that implicitly identifies the ingress PE.

Consider the case of an I-PMSI that spans multiple ASes and that is instantiated by segmented inter-AS P-tunnels. Suppose it is carrying data that is traveling along a particular C-tree. Suppose also that the C-root of that C-tree is multihomed to two or more PEs, and that each such PE is in a different AS than the others. Then, if there is any duplicate traffic, the duplicates will arrive on a different P-tunnel. Specifically, if the PE was expecting the traffic on a particular inter-AS P-tunnel, duplicate traffic will arrive either on an intra-AS P-tunnel (not an intra-AS segment of an inter-AS P-tunnel) or on some other inter-AS P-tunnel. To detect duplicates, the PE has to keep track of which inter-AS A-D route the PE uses for sending MVPN multicast routing information towards the C-S/C-RP. The PE MUST process received (multicast) traffic originated by C-S/C-RP only from the inter-AS P-tunnel that was carried in the best Inter-AS A-D route for the MVPN and that was originated by the AS that contains C-S/C-RP (where "the best" is determined by the PE). The PE MUST discard, as duplicates, all other multicast traffic originated by the C-S/C-RP, but received on any other P-tunnel.

If, for a given MVPN, (a) an MI-PMSI is used for carrying multicast data packets, (b) the MI-PMSI is instantiated by a segmented inter-AS P-tunnel, (c) the C-S or C-RP is multihomed to different PEs, and (d) at least two such PEs are in the same AS, then, depending on the

tunneling technology used to instantiate the MI-PMSI, it may not always be possible for the egress PE to determine the Upstream PE. In that case, the procedure of Sections 9.1.2 or 9.1.3 must be used.

NB: Section 10 describes an exception case where PE1 has to accept a packet even if it is not from the Selected Upstream PE.

9.1.2. Single Forwarder Selection

Section 5.1 specifies a procedure for choosing a "default Upstream PE selection", such that (except during routing transients) all PEs will choose the same default Upstream PE. To ensure that duplicate packets are not sent through the backbone (except during routing transients), an ingress PE does not forward to the backbone any (C-S,C-G) multicast data packet it receives from a CE, unless the PE is the default Upstream PE selection.

One difference in effect between this procedure and the procedure of Section 9.1.1 is that this procedure sends only one copy of each packet to each egress PE, rather than sending multiple copies and forcing the egress PE to discard all but one.

9.1.3. Native PIM Methods

If PE-PE multicast routing information for a given MVPN is being disseminated by running PIM over an MI-PMSI, then native PIM methods will prevent steady state data packet duplication. The PIM Assert mechanism prevents steady state duplication in the scenario of Section 9.2, even if Single Forwarder Selection is not done. The PIM Prune(S,G,rpt) mechanism addresses the scenario of Section 9.3.

9.2. Multihomed C-S or C-RP

Any of the three methods of Section 9.1 will prevent steady state duplicates in the case of a multihomed C-S or C-RP.

9.3. Switching from the C-RP Tree to the C-S Tree

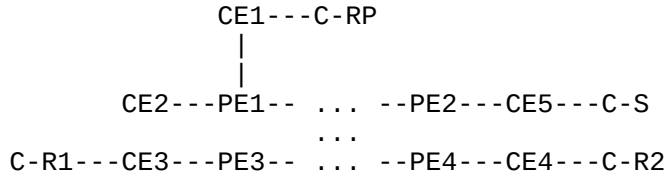
9.3.1. How Duplicates Can Occur

If some PEs are on the C-S tree and some are on the C-RP tree, then a PE may also receive duplicate data traffic after a (C-*,C-G) to (C-S,C-G) switch.

If PIM is being used on an MI-PMSI to disseminate multicast routing information, native PIM methods (in particular, the use of the Prune(S,G,rpt) message) prevent steady state data duplication in this case.

If BGP C-multicast routing is being used, then the procedure of Section 9.1.1, if applicable, can be used to prevent duplication. However, if that procedure is not applicable, then the procedure of Section 9.1.2 is not sufficient to prevent steady state data duplication in all scenarios.

In the scenario in which (a) BGP C-multicast routing is being used, (b) there are inter-site shared C-trees, and (c) there are inter-site source C-trees, additional procedures are needed. To see this, consider the following topology:



Suppose that C-R1 and C-R2 use PIM to join the (C-*,C-G) tree, where C-RP is the RP corresponding to C-G. As a result, CE3 and CE4 will send PIM Join(*,G) messages to PE3 and PE4, respectively. This will cause PE3 and PE4 to originate C-multicast Shared Tree Join Routes, specifying (C-*,C-G). These routes will identify PE1 as the Upstream PE.

Now suppose that C-S is a transmitter for multicast group C-G, and that C-S sends its multicast data packets to C-RP in PIM Register messages. Then, PE1 will receive (C-S,C-G) data packets from CE1, and will forward them over an I-PMSI to PE3 and PE4, who will forward them, in turn, to CE3 and CE4, respectively.

When C-R1 receives (C-S,C-G) data packets, it may decide to join the (C-S,C-G) source tree, by sending a PIM Join(S,G) to CE3. This will, in turn, cause CE3 to send a PIM Join(S,G) to PE3, which will, in turn, cause PE3 to originate a C-multicast Source Tree Join Route, specifying (C-S,C-G) and identifying PE2 as the Upstream PE. As a result, when PE2 receives (C-S,C-G) data packets from CE5, it will forward them on a PMSI to PE3.

At this point, the following situation exists:

- If PE1 receives (C-S,C-G) packets from CE1, PE1 must forward them on the I-PMSI, because PE4 is still expecting to receive the (C-S,C-G) packets from PE1.

- PE3 must continue to receive packets from the I-PMSI, since there may be other sources transmitting C-G traffic and PE3 currently has no other way to receive that traffic.

- PE3 must also receive (C-S,C-G) traffic from PE2.

As a result, PE3 may receive two copies of each (C-S,C-G) packet. The procedure of Section 9.1.2 (Single Forwarder Selection) does not prevent PE3 from receiving two copies, because it does not prevent one PE from forwarding (C-S,C-G) traffic along the shared C-tree while another forwards (C-S,C-G) traffic along a source-specific C-tree.

So if PE3 cannot apply the method of Section 9.1.1 (Discarding Packets from Wrong PE), perhaps because the tunneling technology does not allow the egress PE to identify the ingress PE, then additional procedures are needed.

9.3.2. Solution Using Source Active A-D Routes

The issue described in Section 9.3.1 is resolved through the use of Source Active A-D routes. In the remainder this section, we provide an example of how this works, along with an informal description of the procedures.

A full and precise specification of the relevant procedures can be found in Section 13 of [MVPN-BGP]. In the event of any conflicts or other discrepancies between the description below and the description in [MVPN-BGP], [MVPN-BGP] is to be considered to be the authoritative document.

Please note that the material in this section only applies when inter-site shared trees are being used.

Whenever a PE creates an (C-S,C-G) state as a result of receiving a C-multicast route for (C-S,C-G) from some other PE, and the C-G group is an ASM group, the PE that creates the state MUST originate a Source Active A-D route (see [MVPN-BGP], Section 4.5). The NLRI of the route includes C-S and C-G. By default, the route carries the same set of Route Targets as the Intra-AS I-PMSI A-D route of the MVPN originated by the PE. Using the normal BGP procedures, the

route is propagated to all the PEs of the MVPN. For more details, see Section 13.1 ("Source within a Site - Source Active Advertisement") of [MVPN-BGP].

When, as a result of receiving a new Source Active A-D route, a PE updates its VRF with the route, the PE MUST check if the newly received route matches any (C-*,C-G) entries. If (a) there is a matching entry, (b) the PE does not have (C-S,C-G) state in its MVPN Tree Information Base (MVPN-TIB) for (C-S,C-G) carried in the route, and (c) the received route is selected as the best (using the BGP route selection procedures), then the PE takes the following action:

- If the PE's (C-*,C-G) state has a PMSI as a downstream interface, the PE acts as if all the other PEs had pruned C-S off the (C-*,C-G) tree. That is:

  * If the PE receives (C-S,C-G) traffic from a CE, it does not transmit it to other PEs.

  * Depending on the PIM state of the PE's PE-CE interfaces, the PE may or may not need to invoke PIM procedures to prune C-S off the (C-*,C-G) tree by sending a PIM Prune(S,G,rpt) to one or more of the CEs. This is determined by ordinary PIM procedures. If this does need to be done, the PE SHOULD delay sending the Prune until it first runs a timer; this helps ensure that the source is not pruned from the shared tree until all PEs have had time to receive the Source Active A-D route.

- If the PE's (C-*,C-G) state does not have a PMSI as a downstream interface, the PE sets up its forwarding path to receive (C-S,C-G) traffic from the originator of the selected Source Active A-D route.

Whenever a PE deletes the (C-S,C-G) state that was previously created as a result of receiving a C-multicast route for (C-S,C-G) from some other PE, the PE that deletes the state also withdraws the Source Active A-D route (if there is one) that was advertised when the state was created.

In the example topology of Section 9.3.1, this procedure will cause PE2 to generate a Source Active A-D route for (C-S,C-G). When this route is received, PE4 will set up its forwarding state to expect (C-S,C-G) packets from PE2. PE1 will change its forwarding state so that (C-S,C-G) packets that it receives from CE1 are not forwarded to any other PEs. (Note that PE1 may still forward (C-S,C-G) packets received from CE1 to CE2, if CE2 has receivers for C-G and those

receivers did not switch from the (C-*,C-G) tree to the (C-S,C-G) tree.) As a result, PE3 and PE4 do not receive duplicate packets of the (C-S,C-G) C-flow.

With this procedure in place, there is no need to have any kind of C-multicast route that has the semantics of a PIM Prune(S,G,rpt) message.

It is worth noting that if, as a result of this procedure, a PE sets up its forwarding state to receive (C-S,C-G) traffic from the source tree, the UMH is not necessarily the same as it would be if the PE had joined the source tree as a result of receiving a PIM Join for the same source tree from a directly attached CE.

Note that the mechanism described in Section 7.4.1 can be leveraged to advertise an S-PMSI binding along with the source active messages. This is accomplished by using the same BGP Update message to carry both the NLRI of the S-PMSI A-D route and the NLRI of the Source Active A-D route. (Though an implementation processing the received routes cannot assume that this will always be the case.)

10. Eliminating PE-PE Distribution of (C-*,C-G) State

In the ASM service model, a node that wants to become a receiver for a particular multicast group G first joins a shared tree, rooted at a rendezvous point. When the receiver detects traffic from a particular source, it has the option of joining a source tree, rooted at that source. If it does so, it has to prune that source from the shared tree, to ensure that it receives packets from that source on only one tree.

Maintaining the shared tree can require considerable state, as it is necessary not only to know who the upstream and downstream nodes are, but to know which sources have been pruned off which branches of the share tree.

The BGP-based signaling procedures defined in this document and in [MVPN-BGP] eliminate the need for PEs to distribute to each other any state having to do with which sources have been pruned off a shared C-tree. Those procedures do still allow multicast data traffic to travel on a shared C-tree, but they do not allow a situation in which some CEs receive (S,G) traffic on a shared tree and some on a source tree. This results in a considerable simplification of the PE-PE procedures with minimal change to the multicast service seen within the VPN. However, shared C-trees are still supported across the VPN backbone. That is, (C-*,C-G) state is distributed PE-PE, but (C-*,C-G,rpt) state is not.

In this section, we specify a number of optional procedures that go further and that completely eliminate the support for shared C-trees across the VPN backbone. In these procedures, the PEs keep track of the active sources for each C-G. As soon as a CE tries to join the (*,G) tree, the PEs instead join the (S,G) trees for all the active sources. Thus, all distribution of (C-*,C-G) state is eliminated. These procedures are optional because they require some additional support on the part of the VPN customer and because they are not always appropriate. (For example, a VPN customer may have his own policy of always using shared trees for certain multicast groups.) There are several different options, described in the following sub-sections.

10.1. Co-Locating C-RPs on a PE

[MVPN-REQ] describes C-RP engineering as an issue when PIM-SM (or BIDIR-PIM) is used in Any-Source Multicast (ASM) mode [RFC4607] on the VPN customer site. To quote from [MVPN-REQ]:

> In the case of PIM-SM, when a source starts to emit traffic toward a group (in ASM mode), if sources and receivers are located in VPN sites that are different than that of the RP, then traffic may transiently flow twice through the SP network and the CE-PE link of the RP (from source to RP, and then from RP to receivers). This traffic peak, even short, may not be convenient depending on the traffic and link bandwidth.
>
> Thus, a VPN solution MAY provide features that solve or help mitigate this potential issue.

One of the C-RP deployment models is for the customer to outsource the RP to the provider. In this case, the provider may co-locate the RP on the PE that is connected to the customer site [MVPN-REQ]. This section describes how "anycast-RP" can be used to achieve this. This is described below.

10.1.1. Initial Configuration

For a particular MVPN, at least one or more PEs that have sites in that MVPN, act as an RP for the sites of that MVPN connected to these PEs. Within each MVPN, all of these RPs use the same (anycast) address. All of these RPs use the Anycast RP technique.

10.1.2. Anycast RP Based on Propagating Active Sources

This mechanism is based on propagating active sources between RPs.

10.1.2.1. Receiver(s) within a Site

The PE that receives a C-Join message for (*,G) does not send the information that it has receiver(s) for G until it receives information about active sources for G from an Upstream PE.

On receiving this (described in the next section), the downstream PE will respond with a Join message for (C-S,C-G). Sending this information could be done using any of the procedures described in Section 5. Only the Upstream PE will process this information.

10.1.2.2. Source within a Site

When a PE receives a PIM Register message from a site that belongs to a given VPN, PE follows the normal PIM anycast RP procedures. It then advertises the source and group of the multicast data packet carried in the PIM Register message to other PEs in BGP using the following information elements:

- Active source address
- Active group address
- Route target of the MVPN.

This advertisement goes to all the PEs that belong to that MVPN. When a PE receives this advertisement, it checks whether there are any receivers in the sites attached to the PE for the group carried in the source active advertisement. If there are, then it generates an advertisement for (C-S,C-G) as specified in the previous section.

10.1.2.3. Receiver Switching from Shared to Source Tree

No additional procedures are required when multicast receivers in customer’s site shift from shared tree to source tree.

10.2. Using MSDP between a PE and a Local C-RP

Section 10.1 describes the case where each PE is a C-RP. This enables the PEs to know the active multicast sources for each MVPN, and they can then use BGP to distribute this information to each other. As a result, the PEs do not have to join any shared C-trees, and this results in a simplification of the PE operation.

In another deployment scenario, the PEs are not themselves C-RPs, but use Multicast Source Discovery Protocol (MSDP) [RFC3618] to talk to the C-RPs. In particular, a PE that attaches to a site that contains a C-RP becomes an MSDP peer of that C-RP. That PE then uses BGP to

distribute the information about the active sources to the other PEs. When the PE determines, by MSDP, that a particular source is no longer active, then it withdraws the corresponding BGP Update. Then, the PEs do not have to join any shared C-trees, and they do not have to be C-RPs either.

MSDP provides the capability for a Source Active (SA) message to carry an encapsulated data packet. This capability can be used to allow an MSDP speaker to receive the first (or first several) packet(s) of an (S,G) flow, even though the MSDP speaker hasn't yet joined the (S,G) tree. (Presumably, it will join that tree as a result of receiving the SA message that carries the encapsulated data packet.) If this capability is not used, the first several data packets of an (S,G) stream may be lost.

A PE that is talking MSDP to an RP may receive such an encapsulated data packet from the RP. The data packet should be decapsulated and transmitted to the other PEs in the MVPN. If the packet belongs to a particular (S,G) flow, and if the PE is a transmitter for some S-PMSI to which (S,G) has already been bound, the decapsulated data packet should be transmitted on that S-PMSI. Otherwise, if an I-PMSI exists for that MVPN, the decapsulated data packet should be transmitted on it. (If a MI-PMSI exists, this would typically be used.) If neither of these conditions hold, the decapsulated data packet is not transmitted to the other PEs in the MVPN. The decision as to whether and how to transmit the decapsulated data packet does not affect the processing of the SA control message itself.

Suppose that PE1 transmits a multicast data packet on a PMSI, where that data packet is part of an (S,G) flow, and PE2 receives that packet from that PMSI. According to Section 9, if PE1 is not the PE that PE2 expects to be transmitting (S,G) packets, then PE2 must discard the packet. If an MSDP-encapsulated data packet is transmitted on a PMSI, as specified above, this rule from Section 9 would likely result in the packet being discarded. Therefore, if MSDP-encapsulated data packets being decapsulated and transmitted on a PMSI, we need to modify the rules of Section 9 as follows:

1. If the receiving PE, PE2, has already joined the (S,G) tree, and has chosen PE1 as the Upstream PE for the (S,G) tree, but this packet does not come from PE1, PE2 must discard the packet.

2. If the receiving PE, PE2, has not already joined the (S,G) tree, but is a PIM adjacency to a CE that is downstream on the (*,G) tree, the packet should be forwarded to the CE.

11. Support for PIM-BIDIR C-Groups

In BIDIR-PIM, each multicast group is associated with a Rendezvous Point Address (RPA). The Rendezvous Point Link (RPL) is the link that attaches to the RPA. Usually, it's a LAN where the RPA is in the IP subnet assigned to the LAN. The root node of a BIDIR-PIM tree is a node that has an interface on the RPL.

On any LAN (other than the RPL) that is a link in a BIDIR-PIM tree, there must be a single node that has been chosen to be the DF. (More precisely, for each RPA there is a single node that is the DF for that RPA.) A node that receives traffic from an upstream interface may forward it on a particular downstream interface only if the node is the DF for that downstream interface. A node that receives traffic from a downstream interface may forward it on an upstream interface only if that node is the DF for the downstream interface.

If, for any period of time, there is a link on which each of two different nodes believes itself to be the DF, data forwarding loops can form. Loops in a bidirectional multicast tree can be very harmful. However, any election procedure will have a convergence period. The BIDIR-PIM DF election procedure is very complicated, because it goes to great pains to ensure that if convergence is not extremely fast, then there is no forwarding at all until convergence has taken place.

Other variants of PIM also have a DF election procedure for LANs. However, as long as the multicast tree is unidirectional, disagreement about who the DF is can result only in duplication of packets, not in loops. Therefore, the time taken to converge on a single DF is of much less concern for unidirectional trees and it is for bidirectional trees.

In the MVPN environment, if PIM signaling is used among the PEs, then the standard LAN-based DF election procedure can be used. However, election procedures that are optimized for a LAN may not work as well in the MVPN environment. So, an alternative to DF election would be desirable.

If BGP signaling is used among the PEs, an alternative to DF election is necessary. One might think that the "Single Forwarder Selection" procedures described in Sections 5 and 9 could be used to choose a single PE "DF" for the backbone (for a given RPA in a given MVPN). However, that is still likely to leave a convergence period of at least several seconds during which loops could form, and there could be a much longer convergence period if there is anything disrupting the smooth flow of BGP Updates. So, a simple procedure like that is not sufficient.

The remainder of this section describes two different methods that can be used to support BIDIR-PIM while eliminating the DF election.

11.1. The VPN Backbone Becomes the RPL

On a per-MVPN basis, this method treats the whole service provider(s) infrastructure as a single RPL. We refer to such an RPL as an "MVPN-RPL". This eliminates the need for the PEs to engage in any "DF election" procedure because BIDIR-PIM does not have a DF on the RPL.

However, this method can only be used if the customer is "outsourcing" the RPL/RPA functionality to the SP.

An MVPN-RPL could be realized either via an I-PMSI (this I-PMSI is on a per-MVPN basis and spans all the PEs that have sites of a given MVPN), via a collection of S-PMSIs, or even via a combination of an I-PMSI and one or more S-PMSIs.

11.1.1. Control Plane

Associated with each MVPN-RPL is an address prefix that is unambiguous within the context of the MVPN associated with the MVPN-RPL.

For a given MVPN, each VRF connected to an MVPN-RPL of that MVPN is configured to advertise to all of its connected CEs the address prefix of the MVPN-RPL.

Since, in BIDIR-PIM, there is no Designated Forwarder on an RPL, in the context of MVPN-RPL, there is no need to perform the Designated Forwarder election among the PEs (note it is still necessary to perform the Designated Forwarder election between a PE and its directly attached CEs, but that is done using plain BIDIR-PIM procedures).

For a given MVPN, a PE connected to an MVPN-RPL of that MVPN should send multicast data (C-S,C-G) on the MVPN-RPL only if at least one other PE connected to the MVPN-RPL has a downstream multicast state for C-G. In the context of MVPN, this is accomplished by requiring a PE that has a downstream state for a particular C-G of a particular VRF present on the PE to originate a C-multicast route for (C-*,C-G). The RD of this route should be the same as the RD associated with the VRF. The RTs carried by the route should be such as to ensure that the route gets distributed to all the PEs of the MVPN.

11.1.2. Data Plane

A PE that receives (C-S,C-G) multicast data from a CE should forward this data on the MVPN-RPL of the MVPN the CE belongs to only if the PE receives at least one C-multicast route for (C-*, C-G). Otherwise, the PE should not forward the data on the RPL/I-PMSI.

When a PE receives a multicast packet with (C-S,C-G) on an MVPN-RPL associated with a given MVPN, the PE forwards this packet to every directly connected CE of that MVPN, provided that the CE sends Join (C-*,C-G) to the PE (provided that the PE has the downstream (C-*,C-G) state). The PE does not forward this packet back on the MVPN-RPL. If a PE has no downstream (C-*,C-G) state, the PE does not forward the packet.

11.2. Partitioned Sets of PEs

This method does not require the use of the MVPN-RPL, and it does not require the customer to outsource the RPA/RPL functionality to the SP.

11.2.1. Partitions

Consider a particular C-RPA, call it C-R, in a particular MVPN. Consider the set of PEs that attach to sites that have senders or receivers for a BIDIR-PIM group C-G, where C-R is the RPA for C-G. (As always, we use the "C-" prefix to indicate that we are referring to an address in the VPN's address space rather than in the provider's address space.)

Following the procedures of Section 5.1, each PE in the set independently chooses some other PE in the set to be its "Upstream PE" for those BIDIR-PIM groups with RPA C-R. Optionally, they can all choose the "default selection" (described in Section 5.1) to ensure that each PE to choose the same Upstream PE. Note that if a PE has a route to C-R via a VRF interface, then the PE may choose itself as the Upstream PE.

The set of PEs can now be partitioned into a number of subsets. We'll say that PE1 and PE2 are in the same partition if and only if there is some PE3 such that PE1 and PE2 have each chosen PE3 as the Upstream PE for C-R. Note that each partition has exactly one Upstream PE. So it is possible to identify the partition by identifying its Upstream PE.

Consider packet P, and let PE1 be its ingress PE. PE1 will send the packet on a PMSI so that it reaches the other PEs that need to receive it. This is done by encapsulating the packet and sending it

on a P-tunnel. If the original packet is part of a PIM-BIDIR group (its ingress PE determines this from the packet's destination address C-G), and if the VPN backbone is not the RPL, then the encapsulation MUST carry information that can be used to identify the partition to which the ingress PE belongs.

When PE2 receives a packet from the PMSI, PE2 must determine, by examining the encapsulation, whether the packet's ingress PE belongs to the same partition (relative to the C-RPA of the packet's C-G) to which the PE2 itself belongs. If not, PE2 discards the packet. Otherwise, PE2 performs the normal BIDIR-PIM data packet processing. With this rule in place, harmful loops cannot be introduced by the PEs into the customer's bidirectional tree.

Note that if there is more than one partition, the VPN backbone will not carry a packet from one partition to another. The only way for a packet to get from one partition to another is for it to go up towards the RPA and then down another path to the backbone. If this is not considered desirable, then all PEs should choose the same Upstream PE for a given C-RPA. Then, multiple partitions will only exist during routing transients.

11.2.2. Using PE Distinguisher Labels

If a given P-tunnel is to be used to carry packets traveling along a bidirectional C-tree, then, EXCEPT for the case described in Sections 11.1 and 11.2.3, the packets that travel on that P-tunnel MUST carry a PE Distinguisher Label (defined in Section 4), using the encapsulation discussed in Section 12.3.

When a given PE transmits a given packet of a bidirectional C-group to the P-tunnel, the packet will carry the PE Distinguisher Label corresponding to the partition, for the C-group's C-RPA, that contains the transmitting PE. This is the PE Distinguisher Label that has been bound to the Upstream PE of that partition; it is not necessarily the label that has been bound to the transmitting PE.

Recall that the PE Distinguisher Labels are upstream-assigned labels that are assigned and advertised by the node that is at the root of the P-tunnel. The information about PE Distinguisher Labels is distributed with Intra-AS I-PMSI A-D routes and/or S-PMSI A-D routes by encoding it into the PE Distinguisher Labels attribute carried by these routes.

When a PE receives a packet with a PE label that does not identify the partition of the receiving PE, then the receiving PE discards the packet.

Note that this procedure does not necessarily require the root of a P-tunnel to assign a PE Distinguisher Label for every PE that belongs to the tunnel. If the root of the P-tunnel is the only PE that can transmit packets to the P-tunnel, then the root needs to assign PE Distinguisher Labels only for those PEs that the root has selected to be the UMHs for the particular C-RPAs known to the root.

11.2.3. Partial Mesh of MP2MP P-Tunnels

There is one case in which support for BIDIR-PIM C-groups does not require the use of a PE Distinguisher Label. For each C-RPA, suppose a distinct MP2MP LSP is used as the P-tunnel serving that C-RPA's partition. Then, for a given packet, a PE receiving the packet from a P-tunnel can infer the partition from the tunnel. So, PE Distinguisher Labels are not needed in this case.

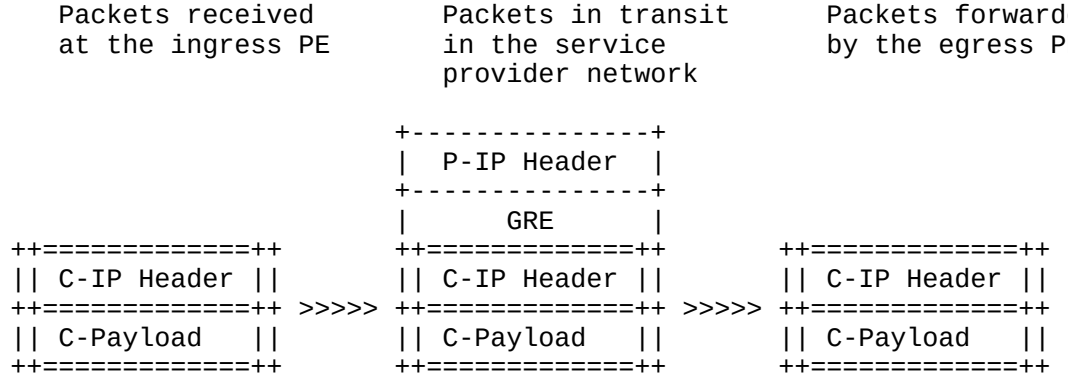
12. Encapsulations

The BGP-based auto-discovery procedures will ensure that the PEs in a single MVPN only use tunnels that they can all support, and for a given kind of tunnel, that they only use encapsulations that they can all support.

12.1. Encapsulations for Single PMSI per P-Tunnel

12.1.1. Encapsulation in GRE

GRE encapsulation can be used for any PMSI that is instantiated by a mesh of unicast P-tunnels, as well as for any PMSI that is instantiated by one or more PIM P-tunnels of any sort.



The IP Protocol Number field in the P-IP header MUST be set to 47. The Protocol Type field of the GRE header is set to either 0x800 or 0x86dd, depending on whether the C-IP header is IPv4 or IPv6, respectively.

When an encapsulated packet is transmitted by a particular PE, the source IP address in the P-IP header must be the same address that the PE uses to identify itself in the VRF Route Import Extended Communities that it attaches to any of VPN-IP routes eligible for UMH determination that it advertises via BGP (see Section 5.1).

If the PMSI is instantiated by a PIM tree, the destination IP address in the P-IP header is the group P-address associated with that tree. The GRE key field value is omitted.

If the PMSI is instantiated by unicast P-tunnels, the destination IP address is the address of the destination PE, and the optional GRE key field is used to identify a particular MVPN. In this case, each PE would have to advertise a key field value for each MVPN; each PE would assign the key field value that it expects to receive.

[RFC2784] specifies an optional GRE checksum and [RFC2890] specifies an optional GRE sequence number fields.

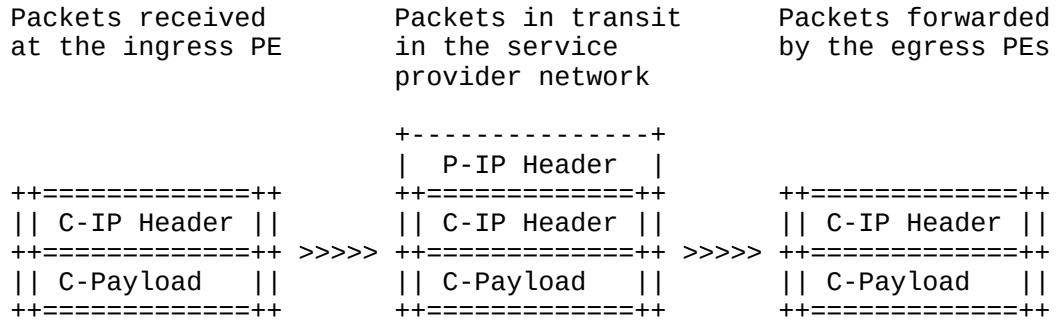
The GRE sequence number field is not needed because the transport layer services for the original application will be provided by the C-IP header.

The use of the GRE checksum field must follow [RFC2784].

To facilitate high speed implementation, this document recommends that the ingress PE routers encapsulate VPN packets without setting the checksum or sequence fields.

12.1.2. Encapsulation in IP

IP-in-IP [RFC2003] is also a viable option. The following diagram shows the progression of the packet as it enters and leaves the service provider network.



When the P-IP header is an IPv4 header, its Protocol Number field is set to either 4 or 41, depending on whether the C-IP header is an IPv4 header or an IPv6 header, respectively.

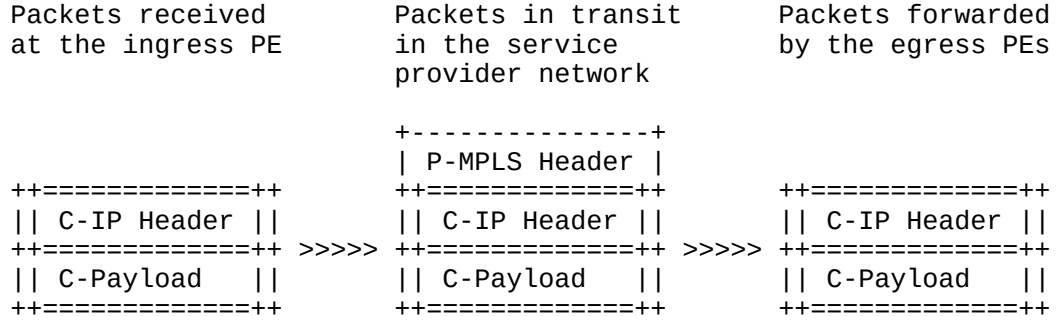
When the P-IP header is an IPv6 header, its Next Header field is set to either 4 or 41, depending on whether the C-IP header is an IPv4 header or an IPv6 header, respectively.

When an encapsulated packet is transmitted by a particular PE, the source IP address in the P-IP header must be the same address that the PE uses to identify itself in the VRF Route Import Extended Communities that it attaches to any of VPN-IP routes eligible for UMH determination that it advertises via BGP (see Section 5.1).

12.1.3. Encapsulation in MPLS

If the PMSI is instantiated as a P2MP MPLS LSP or a MP2MP LSP, MPLS encapsulation is used. Penultimate-hop-popping MUST be disabled for the LSP.

If other methods of assigning MPLS labels to multicast distribution trees are in use, these multicast distribution trees may be used as appropriate to instantiate PMSIs, and appropriate additional MPLS encapsulation procedures may be used.



12.2. Encapsulations for Multiple PMSIs per P-Tunnel

The encapsulations for transmitting multicast data messages when there are multiple PMSIs per P-tunnel are based on the encapsulation for a single PMSI per P-tunnel, but with an MPLS label used for demultiplexing.

The label is upstream-assigned and distributed via BGP as specified in Section 4. The label must enable the receiver to select the proper VRF and may enable the receiver to select a particular multicast routing entry within that VRF.

12.2.1. Encapsulation in GRE

Rather than the IP-in-GRE encapsulation discussed in Section 12.1.1, we use the MPLS-in-GRE encapsulation. This is specified in [MPLS-IP]. The GRE protocol type MUST be set to 0x8847. (The reason for using the unicast rather than the multicast value is specified in [MPLS-MCAST-ENCAPS]).

12.2.2. Encapsulation in IP

Rather than the IP-in-IP encapsulation discussed in Section 12.1.2, we use the MPLS-in-IP encapsulation. This is specified in [MPLS-IP]. The IP protocol number field MUST be set to the value identifying the payload as an MPLS unicast packet. (There is no "MPLS multicast packet" protocol number.)

12.3. Encapsulations Identifying a Distinguished PE

12.3.1. For MP2MP LSP P-Tunnels

As discussed in Section 9, if a multicast data packet is traveling on a unidirectional C-tree, it is highly desirable for the PE that receives the packet from a PMSI to be able to determine the identity of the PE that transmitted the data packet onto the PMSI. The encapsulations of the previous sections all provide this information, except in one case. If a PMSI is being instantiated by an MP2MP LSP, then the encapsulations discussed so far do not allow one to determine the identity of the PE that transmitted the packet onto the PMSI.

Therefore, when a packet traveling on a unidirectional C-tree is traveling on a MP2MP LSP P-tunnel, it MUST carry, as its second label, a label that has been bound to the packet’s ingress PE. This label is an upstream-assigned label that the LSP’s root node has bound to the ingress PE and has distributed via the PE Distinguisher

Labels attribute of a PMSI A-D route (see Section 4). This label will appear immediately beneath the labels that are discussed in Sections 12.1.3 and 12.2.

A full specification of the procedures for advertising and for using the PE Distinguisher Labels attribute in this case is outside the scope of this document.

12.3.2. For Support of PIM-BIDIR C-Groups

As was discussed in Section 11, when a packet belongs to a PIM-BIDIR multicast group, the set of PEs of that packet's VPN can be partitioned into a number of subsets, where exactly one PE in each partition is the Upstream PE for that partition. When such packets are transmitted on a PMSI, unless the procedures of Section 11.2.3 are being used, it is necessary for the packet to carry information identifying a particular partition. This is done by having the packet carry the PE Distinguisher Label corresponding to the Upstream PE of one partition. For a particular P-tunnel, this label will have been advertised by the node that is the root of that P-tunnel. (A full specification of the procedures for advertising PE Distinguisher Labels is out of the scope of this document.)

This label needs to be used whenever a packet belongs to a PIM-BIDIR C-group, no matter what encapsulation is used by the P-tunnel. Hence, the encapsulations of Section 12.2 MUST be used. If the P-tunnel contains only one PMSI, the PE label replaces the label discussed in Section 12.2. If the P-tunnel contains multiple PMSIs, the PE label follows the label discussed in Section 12.2.

In general, PE Distinguisher Labels can be carried if the encapsulation is MPLS, MPLS-in-IP, or MPLS-in-GRE. However, procedures for advertising and using PE Distinguisher Labels when the encapsulation is LDP-based MP2P MPLS is outside the scope of this specification.

12.4. General Considerations for IP and GRE Encapsulations

These apply also to the MPLS-in-IP and MPLS-in-GRE encapsulations.

12.4.1. MTU (Maximum Transmission Unit)

It is the responsibility of the originator of a C-packet to ensure that the packet is small enough to reach all of its destinations, even when it is encapsulated within IP or GRE.

When a packet is encapsulated in IP or GRE, the router that does the encapsulation MUST set the DF bit in the outer header. This ensures that the decapsulating router will not need to reassemble the encapsulating packets before performing decapsulation.

In some cases, the encapsulating router may know that a particular C-packet is too large to reach its destinations. Procedures by which it may know this are outside the scope of the current document. However, if this is known, then:

- If the DF bit is set in the IP header of a C-packet that is known to be too large, the router will discard the C-packet as being "too large" and follow normal IP procedures (which may require the return of an ICMP message to the source).

- If the DF bit is not set in the IP header of a C-packet that is known to be too large, the router MAY fragment the packet before encapsulating it and then encapsulate each fragment separately. Alternatively, the router MAY discard the packet.

If the router discards a packet as too large, it should maintain Operations, Administration, and Maintenance (OAM) information related to this behavior, allowing the operator to properly troubleshoot the issue.

Note that if the entire path of the P-tunnel does not support an MTU that is large enough to carry the a particular encapsulated C-packet, and if the encapsulating router does not do fragmentation, then the customer will not receive the expected connectivity.

12.4.2. TTL (Time to Live)

The ingress PE should not copy the TTL field from the payload IP header received from a CE router to the delivery IP or MPLS header. The setting of the TTL of the delivery header is determined by the local policy of the ingress PE router.

12.4.3. Avoiding Conflict with Internet Multicast

If the SP is providing Internet multicast, distinct from its VPN multicast services, and using PIM based P-multicast trees, it must ensure that the group P-addresses that it used in support of MVPN services are distinct from any of the group addresses of the Internet multicasts it supports. This is best done by using administratively scoped addresses [ADMIN-ADDR].

The group C-addresses need not be distinct from either the group P-addresses or the Internet multicast addresses.

12.5. Differentiated Services

The setting of the DS (Differentiated Services) field in the delivery IP header should follow the guidelines outlined in [RFC2983]. Setting the Traffic Class field [RFC5462] in the delivery MPLS header should follow the guidelines in [RFC3270]. An SP may also choose to deploy any of additional Differentiated Services mechanisms that the PE routers support for the encapsulation in use. Note that the type of encapsulation determines the set of Differentiated Services mechanisms that may be deployed.

13. Security Considerations

This document describes an extension to the procedures of [RFC4364], and hence shares the security considerations described in [RFC4364] and [RFC4365].

When GRE encapsulation is used, the security considerations of [MPLS-IP] are also relevant. Additionally, the security considerations of [RFC4797] are relevant as it discusses implications on packet spoofing in the context of BGP/MPLS IP VPNs.

The security considerations of [MPLS-HDR] apply when MPLS encapsulation is used.

This document makes use of a number of control protocols: PIM [PIM-SM], BGP [MVPN-BGP], mLDP [MLDP], and RSVP-TE [RSVP-P2MP]. Security considerations relevant to each protocol are discussed in the respective protocol specifications.

If one uses the UDP-based protocol for switching to S-PMSI (as specified in Section 7.4.2), then an S-PMSI Join message (i.e., a UDP packet with destination port 3232 and destination address ALL-PIM-ROUTERS) that is not received over a PMSI (e.g., one received directly from a CE router) is an illegal packet and MUST be dropped.

The various procedures for P-tunnel construction have security issues that are specific to the way that the P-tunnels are used in this document. When P-tunnels are constructed via such techniques as PIM, mLDP, or RSVP-TE, each P or PE router receiving a control message MUST ensure that the control message comes from another P or PE router, not from a CE router. (Interpreting an mLDP or PIM or RSVP-TE control message from a CE router as referring to a P-tunnel would be a bug.)

A PE MUST NOT accept BGP routes of the MCAST-VPN address family from a CE.

If BGP is used as a CE-PE routing protocol, then when a PE receives an IP route from a CE, if this route carries the VRF Route Import Extended Community, the PE MUST remove this Community from the route before turning it into a VPN-IP route. Routes that a PE advertises to a CE MUST NOT carry the VRF Route Import Extended Community.

An ASBR may receive, from one SP's domain, an mLDP, PIM, or RSVP-TE control message that attempts to extend a P-tunnel from one SP's domain into another SP's domain. This is perfectly valid if there is an agreement between the SPs to jointly provide an MVPN service. In the absence of such an agreement, however, this could be an illegitimate attempt to intercept data packets. By default, an ASBR MUST NOT allow P-tunnels to extend beyond AS boundaries. However, it MUST be possible to configure an ASBR to allow this on a specified set of interfaces.

Many of the procedures in this document cause the SP network to create and maintain an amount of state that is proportional to customer multicast activity. If the amount of customer multicast activity exceeds expectations, this can potentially cause P and PE routers to maintain an unexpectedly large amount of state, which may cause control and/or data plane overload. To protect against this situation, an implementation should provide ways for the SP to bound the amount of state it devotes to the handling of customer multicast activity.

In particular, an implementation SHOULD provide mechanisms that allow an SP to place limitations on the following:

- total number of (C-*,C-G) and/or (C-S,C-G) states per VRF

- total number of P-tunnels per VRF used for S-PMSIs

- total number of P-tunnels traversing a given P router

A PE implementation MAY also provide mechanisms that allow an SP to limit the rate of change of various MVPN-related states on PEs, as well as the rate at which MVPN-related control messages may be received by a PE from the CEs and/or sent from the PE to other PEs.

An implementation that provides the procedures specified in Sections 10.1 or 10.2 MUST provide the capability to impose an upper bound on the number of Source Active A-D routes generated and on how frequently they may be originated. This MUST be provided on a per-PE, per-MVPN granularity.

Lack of the mechanisms that allow an SP to limit the rate of change of various MVPN-related states on PEs, as well as the rate at which MVPN-related control messages may be received by a PE from the CEs and/or sent from the PE to other PEs may result in the control plane overload on the PE, which in turn would adversely impact all the customers connected to that PE, as well as to other PEs.

See also the Security Considerations section of [MVPN-BGP].

14. IANA Considerations

Section 7.4.2 defines the "S-PMSI Join message", which is carried in a UDP datagram whose port number is 3232. This port number had already been assigned by IANA to "MDT port". The reference has been updated to this document.

IANA has created a registry for the "S-PMSI Join message Type Field". Assignments are to be made according to the policy "IETF Review" as defined in [RFC5226]. The value 1 has been registered with a reference to this document. The description reads "PIM IPv4 S-PMSI (unaggregated)".

[PIM-ATTRIB] establishes a registry for "PIM Join attribute Types". IANA has assigned the value 1 to the "MVPN Join Attribute" with a reference to this document.

IANA has assigned SAFI 129 to "Multicast for BGP/MPLS IP Virtual Private Networks (VPNs)" with a reference to this document and [MVPN-BGP].

15. Acknowledgments

Significant contributions were made Arjen Boers, Toerless Eckert, Adrian Farrel, Luyuan Fang, Dino Farinacci, Lenny Giuliano, Shankar Karuna, Anil Lohiya, Tom Pusateri, Ted Qian, Robert Raszuk, Tony Speakman, Dan Tappan.

16. References

16.1. Normative References

[MLDP] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

[MPLS-HDR] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

[MPLS-IP] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[MPLS-MCAST-ENCAPS] Eckert, T., Rosen, E., Ed., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.

[MPLS-UPSTREAM-LABEL] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.

[MVPN-BGP] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

[OSPF] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[OSPF-MT] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.

[PIM-ATTRIB] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008.

[PIM-SM] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, September 2006.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

[RSVP-OOB] Ali, Z., Swallow, G., and R. Aggarwal, "Non-Penultimate Hop Popping Behavior and Out-of-Band Mapping for RSVP-TE Label Switched Paths", RFC 6511, February 2012.

[RSVP-P2MP] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

16.2. Informative References

[ADMIN-ADDR] Meyer, D., "Administratively Scoped IP Multicast", BCP 23, RFC 2365, July 1998.

[BIDIR-PIM] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[BSR] Bhaskar, N., Gall, A., Lingard, J., and S. Venaas, "Bootstrap Router (BSR) Mechanism for Protocol Independent Multicast (PIM)", RFC 5059, January 2008.

[MVPN-REQ] Morin, T., Ed., "Requirements for Multicast in Layer 3 Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4834, April 2007.

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, September 2000.

[RFC2983] Black, D., "Differentiated Services and Tunnels", RFC 2983, October 2000.

[RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[RFC3618] Fenner, B., Ed., and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, October 2003.

[RFC4365] Rosen, E., "Applicability Statement for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4365, February 2006.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

[RFC4797] Rekhter, Y., Bonica, R., and E. Rosen, "Use of Provider Edge to Provider Edge (PE-PE) Generic Routing Encapsulation (GRE) or IP in BGP/MPLS IP Virtual Private Networks", RFC 4797, January 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Contributing Authors

Sarveshwar Bandi
Motorola
Vanenburg IT park, Madhapur,
Hyderabad, India
EMail: sarvesh@motorola.com

Yiqun Cai
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
EMail: ycai@cisco.com

Thomas Morin
France Telecom R & D
2, avenue Pierre-Marzin
22307 Lannion Cedex
France
EMail: thomas.morin@francetelecom.com

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
EMail: yakov@juniper.net

IJsbrand Wijnands
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
EMail: ice@cisco.com

Seisho Yasukawa
NTT Corporation
9-11, Midori-Cho 3-Chome
Musashino-Shi, Tokyo 180-8585,
Japan
Phone: +81 422 59 4769
EMail: yasukawa.seisho@lab.ntt.co.jp

Editors' Addresses

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
EMail: erosen@cisco.com

Rahul Aggarwal
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
EMail: raggarwa_1@yahoo.com