Network Working Group
Internet Draft
Expiration Date: October 2007

Eric C. Rosen (Editor)
Cisco Systems, Inc.

Rahul Aggarwal (Editor)
Juniper Networks

April 2007

Multicast in MPLS/BGP IP VPNs

draft-ietf-l3vpn-2547bis-mcast-04.txt

Status of this Memo

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.

Table of Contents

1. Specification of requirements

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. Introduction

[RFC4364] specifies the set of procedures which 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 also provides detailed specification of some of the protocols. The detailed specification of some of the other protocols is found in pre-existing documents or in companion documents.

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 which is specific to that flow. 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 CE sends to a PE, such that the IP destination address of the packets is a multicast group address, or 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 which 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 which are specific to that flow. Each such tree requires that state be maintained in all the P routers that are in the tree.

This would potentially 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. Nor does the SP have any control over the number of transmitters in each group, nor of 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 tradeoffs 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 which 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 which 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 which 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 carry 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. Overview

2.2.1. 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.

The multicast routing protocol on the PE-CE link is presumed to be PIM. The Sparse Mode, Dense Mode, Single Source Mode, and Bidirectional Modes are supported. A CE router exchanges "ordinary" PIM control messages with the PE router to which it is attached.

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.

To support the "Carrier's Carrier" model of [RFC4364], mLDP or BGP can be used on the PE-CE interface. This will be described in subsequent versions of this document.

2.2.2. MVPN Definition

An MVPN is defined by two sets of sites, Sender Sites set and 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, 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 Sender Sites set and Receiver Site 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, with extensions, as necessary.

2.2.3. 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 autodiscovery:

- 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]) which 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 multicast data packets of a particular MVPN are to be transmitted (at least, by default) through a non-aggregated Inclusive Tree which is to be set up by PIM-SM or PIM-Bidir, 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.

2.2.4. 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 routing adjacency with other PE routers to exchange MVPN control information.

The solution in [RFC4364] uses BGP to exchange VPN routing information between PE routers. This document describes various

solutions for exchanging MVPN control information. One option is the use of BGP, providing reliable transport. Another option is the use of the currently existing, "soft state" PIM standard [PIM-SM].

2.2.5. 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 MPLS or IP/GRE encapsulation), multicast distribution trees created by PIM-SSM, PIM-SM, or PIM-Bidir (using IP/GRE encapsulation), point-to-multipoint LSPs created by RSVP-TE or mLDP, and multipoint-to-multipoint LSPs created by mLDP. (However, techniques for aggregating the traffic of multiple MVPNs onto a single multipoint-to-multipoint LSP or onto a single bidirectional multicast distribution tree are for further study.) For selective trees, only unicast PE-PE tunnels (using MPLS or IP/GRE encapsulation) and unidirectional single-source trees are supported, and the supported tree creation protocols are PIM-SSM (using IP/GRE encapsulation), RSVP-TE, and 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.

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 for further study.

2.2.6. Inter-AS MVPNs

[RFC4364] describes different options for supporting Inter-AS BGP/MPLS unicast VPNs. 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, which 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.

2.2.7. Optional Deployment Models

The document also discusses an optional MVPN deployment model in which PEs take on all or part of the role of a PIM RP (Rendezvous Point). The necessary protocol extensions to support this are defined.

3. Concepts and Framework

3.1. PE-CE Multicast Routing

Support of multicast in BGP/MPLS IP VPNs is modeled closely after 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" ("Virtual 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 will 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)

- PIM Single Source Mode (PIM-SSM)

- PIM Bidirectional Mode (PIM-Bidir)

- PIM Dense Mode (PIM-DM)

In order to support the "Carrier's Carrier" model of [RFC4364], mLDP or BGP will also be supported on the PE-CE interface; however, this is not described in this revision.

As the document only supports 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 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 which is processed by the PIM P-instance, and a C-Join would be a PIM Join which 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.

3.2. P-Multicast Service Interfaces (PMSIs)

Multicast data packets received by a PE over a PE-CE interface must be forwarded 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 a PMSI connecting those PE routers. 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 such a packet will be able to tell which MVPN the packet belongs to.

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 "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 which is used to instantiate a PMSI may allow a single "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 "tunnel" is used to refer only to the transport mechanism that instantiates a service.

[This is a significant change from previous drafts on the topic of MVPN, which have used the term "Multicast Tunnel" to refer both to the multicast service (what we call here the PMSI) and to its instantiation.]

3.2.1. Inclusive and Selective PMSIs

We will distinguish between three different kinds of PMSI:

- "Multidirectional Inclusive" PMSI (MI-PMSI)

  A Multidirectional Inclusive PMSI is one which 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 tunnel which instantiates 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.

  [The "Default MDTs" of rosen-08 provide the transport service of MI-PMSIs, in this terminology.]

- "Unidirectional Inclusive" PMSI (UI-PMSI)

  A Unidirectional Inclusive PMSI is one which 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 "tunnel" which instantiates a UI-PMSI may in fact carry the data of more than one PMSI.

- "Selective" PMSI (S-PMSI).

  A Selective PMSI is one which 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. Again, the "tunnel" which instantiates a given S-PMSI may carry data from multiple S-PMSIs.

  [The "Data MDTs" of earlier drafts provide the transport service of "Selective PMSIs" in the terminology of this draft.]

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

3.2.2. Tunnels Instantiating PMSIs

A number of different tunnel setup techniques can be used to create the tunnels that instantiate the PMSIs. Among these are:

- PIM

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

  PIM-SSM, PIM-Bidir, or PIM-SM can be used to create P-trees. (PIM-DM is not supported for this purpose.)

  A single MI-PMSI can be instantiated by a single shared P-tree, or by a number of source P-trees (one for each PE of the MI-PMSI). P-trees may be shared by multiple MVPNs (i.e., a given P-tree may be the instantiation of multiple PMSIs), as long as the encapsulation provides some means of demultiplexing the data traffic by MVPN.

  Selective PMSIs are most instantiated by source P-trees, and are most naturally created by PIM-SSM, since by definition only one

PE is the source of the multicast data on a Selective PMSI.

[The "Default MDTs" of [rosen-08] are MI-PMSIs instantiated as PIM trees. The "data MDTs" of [rosen-08] are S-PMSIs instantiated as PIM trees.]

- MLDP

A PMSI may be instantiated as one or more mLDP Point-to-Multipoint (P2MP) LSPs, or as an mLDP Multipoint-to-Point(MP2MP) LSP. A Selective PMSI or a Unidirectional Inclusive PMSI would be instantiated as a single mLDP P2MP LSP, whereas a Multidirectional Inclusive PMSI could be instantiated either as a set of such LSPs (one for each PE in the MVPN) or as a single M2PMP LSP.

MLDP P2MP LSPs can be shared across multiple MVPNs.

- RSVP-TE

A PMSI may be instantiated as one or more RSVP-TE Point-to-Multipoint (P2MP) LSPs. A Selective PMSI or a Unidirectional Inclusive 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 Tunnels.

If a PMSI is implemented as a mesh of unicast 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 tunnels among that MVPN's PEs. A UI-PMSI or an S-PMSI can be instantiated as a partial mesh.

- Unicast Tunnels to the Root of a P-Tree.

Any type of PMSI can be instantiated through a method in which there is a single P-tree (created, for example, via PIM-SSM or via RSVP-TE), and a PE transmits a packet to the PMSI by sending it in a unicast tunnel to the root of that P-tree. All PEs in the given MVPN would need to be leaves of the tree.

When this instantiation method is used, the transmitter of the

multicast data may receive its own data back. Methods for

avoiding this are for further study.

It can be seen that each method of implementing PMSIs has its own area of applicability. This specification therefore 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 which is always acceptable. The use of segmented inter-AS trees does allow each SP to select the option which 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 the router. The procedures defined throughout this document do not mandate that the same tunnel technique be used for all PMSI tunnels going through a same provider backbone. It is however expected that any tunnel technique that can be subject to being used by a PE for a particular MVPN is also supported by other PE 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 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 segmented inter-AS trees are not used (see section 8.2), 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 a default MI-PMSI, the information needed to set up and to use the tunnels instantiating the default MI-PMSI,

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

In some cases this information is provided by means of the BGP-based auto-discovery procedures detailed in section 4. In other cases, this information is provided after discovery is complete, by means of procedures defined in section 6.1.2. In either case, the information which is provided must be sufficient to enable the PMSI to be bound to the identified tunnel, to enable the tunnel to be created if it does not already exist, and to enable the different PMSIs which may travel on the same tunnel to be properly demultiplexed.

3.3.1. MVPNs with Default MI-PMSIs

If an MVPN uses an MI-PMSI, then the MI-PMSI for that MVPN will be created as soon as the necessary information has been obtained. Creating a PMSI means creating the tunnel which carries it (unless that tunnel already exists), as well as binding the PMSI to the tunnel. The MI-PMSI for that MVPN is then used as the default method of transmitting multicast data packets for that MVPN. In effect, all the multicast streams for the MVPN are, by default, aggregated onto the MI-MVPN.

If a particular multicast stream from a particular source PE has certain characteristics, it can be desirable to migrate it from the MI-PMSI to an S-PMSI. Procedures for migrating a stream from an MI-PMSI to an S-PMSI are discussed in section 7.

3.3.2. When MI-PMSIs are Required

MI-PMSIs are required under the following conditions:

- The MVPN is using PIM-DM, or some other protocol (such as BSR) which relies upon flooding. Only with an MI-PMSI can the C-data

(or C-control-packets) received from any CE be flooded to all PEs.

- If the procedure for carrying C-multicast routes from PE to PE involves the multicasting of P-PIM control messages among the PEs (see sections 5.2.1, 5.2.2, and 5.2.4).

3.3.3. MVPNs That Do Not Use MI-PMSIs

If a particular MVPN does not use a default MI-PMSI, then its multicast data may be sent by default on a UI-PMSI.

It is also possible to send all the multicast data on an S-PMSI, omitting any usage of I-PMSIs. This prevents PEs from receiving data

which they don't need, at the cost of requiring additional tunnels. However, 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 7.2.2.2.

4. BGP-Based Autodiscovery of MVPN Membership

BGP-based autodiscovery 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 which attaches to an MVPN must issue a BGP update message containing an NLRI in this address family, along with a specific set of attributes. In this document, we specify the information which 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 autodiscovery procedures. When segmented inter-AS trees are used, additional procedures are needed, as specified in section 8. Further detail may be found 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 autodiscovery uses a particular kind of MCAST-VPN route known as an "auto-discovery routes", or "A-D route".

An "intra-AS A-D route" is a particular kind of A-D route that is never distributed outside its AS of origin. Intra-AS A-D routes are originated by the PEs that are (directly) connected to the site(s) of that MVPN.

For the purpose of auto-discovery, each PE attached to a site in a given MVPN must originate an intra-AS auto-discovery route. The NLRI of that route must the following information:

- The route type (i.e., intra-AS A-D route)

- IP address of the originating PE

- An RD configured locally for the MVPN. This is an RD which can be prepended to that IP address to form a globally unique VPN-IP address of the PE.

The A-D route must also carry the following attributes:

- One or more Route Target attributes. 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 Auto-Discovery 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 Auto-Discovery routes by PEs in the receiver sites set.

* PMSI tunnel attribute. This attribute is present if and only if a default MI-PMSI is to be used for the MVPN. It contains the following information:

whether the MI-PMSI is instantiated by

+ A PIM-Bidir tree,

+ a set of PIM-SSM trees,

+ a set of PIM-SM trees

+ a set of RSVP-TE point-to-multipoint LSPs

+ a set of mLDP point-to-multipoint LSPs

+ an mLDP multipoint-to-multipoint LSP

+ a set of unicast tunnels

+ a set of unicast tunnels to the root of a shared tree (in this case the root must be identified)

* If the PE wishes to setup a default tunnel to instantiate the I-PMSI, a unique identifier for the tunnel used to instantiate the I-PMSI.

  All the PEs attaching to a given MVPN (within a given AS) must have been configured with the same PMSI tunnel attribute for that MVPN. They are also expected to know the encapsulation to use.

Note that a default tunnel can be identified at discovery time only if the tunnel already exists (e.g., it was constructed by means of configuration), or if it can be constructed without each PE knowing the the identities of all the others (e.g., it is constructed by a receiver-initiated join technique such as PIM or mLDP).

In other cases, a default tunnel cannot be identified until the PE has discovered one or more of the other PEs. This will be the case, for example, if the tunnel is an RSVP-TE P2MP LSP, which must be set up from the head end. In these cases, a PE will first send an A-D route without a tunnel

identifier, and then will send another one with a tunnel identifier after discovering one or more of the other PEs.

* Whether the tunnel used to instantiate the I-PMSI for this MVPN is 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.

Further details of the use of this information are provided in subsequent sections.

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.

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

5.1. RPF Information for Unicast VPN-IP Routes

When a PE receives a C-Join/Prune message from a CE, the message identifies a particular multicast flow as belong either to a source tree (S,G) or to a shared tree (*,G). We use the term C-source to refer to S, in the case of a source tree, or to the Rendezvous Point (RP) for G, in the case of (*,G). The PE needs to find the "upstream multicast hop" for the (S,G) or (*,G) flow, and it does this by looking up the C-source in the unicast VRF associated with the PE-CE interfaces over which the C-Join/Prune was received. To facilitate

this, all unicast VPN-IP routes from an MVPN will carry RPF information, which identifies the PE that originated the route, as well as identifying the Autonomous System containing that PE. This information is consulted when a PE does an "RPF lookup" of the C-source as part of processing the C-Join/Prune messages. This RPF

information contains the following:

- Source AS Extended Community

  To support MVPN a PE that originates a (unicast) route to VPN-IPv4 addresses MUST include in the BGP Update message that carries this route the Source AS extended community, except if it is known a priori that none of these addresses will act as multicast sources and/or RP, in which case the (unicast) route need not carry the Source AS extended community. The Global Administrator field of this community MUST be set to the autonomous system number of the PE. The Local Administrator field of this community SHOULD be set to 0. This community is described further in [MVPN-BGP].

- Route Import Extended Community

  To support MVPN in addition to the import/export Route Target(s) used by the unicast routing, each VRF on a PE MUST have an import Route Target that is unique to this VRF, except if it is known a priori that none of the (local) MVPN sites associated with the VRF contain multicast source(s) and/or RP, in which case the VRF need not have this import Route Target. This Route Target MUST be IP address specific, and is constructed as follows:

  + The Global Administrator field of the Route Target MUST be set to an IP address of the PE. This address MUST be a routable IP address. This address MAY be common for all the VRFs on the PE (e.,g., this address may be PE's loopback address).

  + The Local Administrator field of the Route Target associated with a given VRF contains a 2 octets long number that uniquely identifies that VRF within the PE that contains the VRF (procedures for assigning such numbers are purely local to the PE, and outside the scope of this document).

A PE that originates a (unicast) route to VPN-IPv4 addresses MUST include in the BGP Updates message that carries this route the Route Import extended community that has the value of this Route Target, except if it is known a priori that none of these addresses will act as multicast sources and/or RP, in which case the (unicast) route need not carry the Route Import extended community.

The Route Import Extended Community is described further in [MVPN-BGP].

5.2. PIM Peering

5.2.1. Full Per-MVPN PIM Peering Across a MI-PMSI

If the set of PEs attached to a given MVPN are connected via a 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 which 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 is the procedure specified in [rosen-08].]

5.2.2. Lightweight PIM Peering Across a 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/Prunes is not as simple to eliminate. The L3VPN WG has asked the PIM WG to specify "refresh reduction" procedures for PIM, so as to eliminate the need for the periodic refreshes. If and when such procedures have been specified, it will be very useful to incorporate them, so as to make the lightweight PIM peering procedures even more lightweight.

5.2.3. Unicasting of PIM C-Join/Prune Messages

PIM does not require that the C-Join/Prune messages which 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 which 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 pre-requisite for the procedure of unicasting the C-Join/Prune messages.

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

If there are multiple PEs that can be used to reach a given C-source, procedures described in section 9 MUST be used to ensue that, at least within a single AS, all PEs choose the same PE to reach the C-source.

5.2.4. Details of Per-MVPN PIM Peering over MI-PMSI

In this section, we assume that inter-AS MVPNs will be supported by means of non-segmented inter-AS trees. Support for segmented inter-AS trees with PIM peering is for further study.

When an MVPN uses an MI-PMSI, the C-instances of that MVPN can treat

the MI-PMSI as a LAN interface, and form either full PIM adjacencies or lightweight PIM adjacencies with each other over that "LAN interface".

To form a full PIM adjacency, the PEs execute the PIM LAN procedures, including the generation and processing of PIM Hello, Join/Prune,

Assert, DF election and other PIM control packets. These are executed independently for each C-instance. PIM "join suppression" SHOULD be enabled.

If it is known that all C-instances of a particular MVPN can support lightweight adjacencies, then lightweight adjacencies MUST be used. If it is not known that all such C-instances support lightweight instances, then full adjacencies MUST be used. Whether all the C-instances support lightweight adjacencies is known by virtue of the BGP-based auto-discovery procedures (combined with configuration). This knowledge might change over time, so the PEs must be able to switch in real time between the use of full adjacencies and lightweight adjacencies.

The difference between a lightweight adjacency and a full adjacency is that no PIM Hellos are sent or received on a lightweight adjacency. The function which Hellos usually provide in PIM can be provided in MVPN by the BGP-based auto-discovery procedures, so the Hellos become superfluous.

Whether or not Hellos are sent, if PIM Refresh Reduction procedures are available, and all the PEs supporting the MVPN are known to support these procedures, then the refresh reduction procedures MUST be used.

5.2.4.1. PIM C-Instance Control Packets

All 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 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.

5.2.4.2. PIM C-instance 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. It is therefore necessary to specify special procedures for determining when the MI-PMSI is to be regarded as the "RPF Interface" for a particular C-address.

When a PE needs to determine the RPF interface of a particular C-address, it looks up the C-address in the VRF. If the route matching

it (call this the "RPF route") is not a VPN-IP route learned from MP-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 RPF route is a VPN-IP route whose outgoing interface is not one of the interfaces associated with the VRF, then PIM will consider the outgoing 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-address is the MI-PMSI, it is necessary for PIM to determine the RPF neighbor for that C-address. This will be one of the other PEs that is a PIM adjacency over the MI-PMSI.

When a PE distributes a given VPN-IP route via BGP, the PE must determine whether that route might possibly be regarded, by another PE, as an RPF route. (If a given VRF is part of an MVPN, it may be simplest to regard every route exported from that VRF to be a potential RPF route.) If the given VPN-IP route is a potential RPF route, then when the VPN-IP route is distributed by BGP, it SHOULD be accompanied by a VRF Route Import Extended Community (see [MVPN-BGP]).

The VRF Route Import Extended Community contains an embedded IP address. If a PE advertises a route with a VRF Route Import Extended Community, then the PE MUST use that the IP address embedded therein as its Source IP address in any PIM control messages which it

transmits to other PEs in the same MVPN. If a VRF Route Import Extended Community is not present, then the source IP address in any PIM control messages which it transmits to other PEs in the same MVPN MUST be be the same as the address carried in the BGP Next Hop of the route.

When a PE has determined that the RPF interface for a particular C-address is the MI-PMSI, it must look up the RPF information that was distributed along with the VPN-IP address corresponding to that C-address. The IP address in this RPF information will be considered to be the IP address of the RPF adjacency for the C-address.

If the RPF information is not present, but the "BGP Next Hop" for the C-address is one of the PEs that is a PIM adjacency over the MI-PMSI, then that PE should be treated as the RPF adjacency for that C-address. However, if the MVPN spans multiple Autonomous Systems, the BGP Next Hop might not be a PIM adjacency, and if that is the case the RPF check will not succeed unless the RPF information is used.

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.

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 which is encoded into the NLRI. The following information is required in BGP to advertise the MVPN routing information. The NLRI contains:

- The type of C-multicast route.

There are two types:

  * source tree join

  * shared tree join

- The RD configured, for the MVPN, on the PE that is advertising the information. This is required to uniquely identify the <C-Source, C-Group> as the addresses could overlap between different MVPNs.

- The C-Source address. (Omitted if the route type is "shared tree join")

- The C-Group address.

- The RD from the VPN-IP route to the C-source.

  That is, the route to the C-source is looked up in the local unicast VRF associated with the CE-PE interface over which the C-multicast control packet arrived. The corresponding VPN-IP route is then examined, and the RD from that route is placed into the C-multicast route.

  Note that this RD is NOT necessarily one which is configured on the local PE. Rather it is one which is configured on the remote PE that is on the path to the C-source.

The following attribute must also be included:

- The upstream multicast hop.

  If a PE receives a C-Join (*, G) from a CE, the C-source is considered to be the C-RP for the particular C-G. When the C-multicast route represents a "shared tree join", it is presumed that the root of the tree (e.g., the RP) is determined by some means outside the scope of this specification.

  When the PE processes a C-PIM Join/Prune message, the route to the C-source is looked up in the local unicast VRF associated with the CE-PE interface over which the C-multicast control packet arrived. The corresponding VPN-IP route is then examined.

If the AS specified therein is the local AS, or if no AS is specified therein, then the PE specified therein becomes the upstream multicast hop. If the AS specified therein is a remote AS, the BGP next hop on the route to the MVPN Auto-Discovery route advertised by the remote AS, becomes the upstream multicast hop.

N.B.: It is possible that here is more than one unicast VPN-IP route to the C-source. In this case, the route that was installed in the VRF is not necessarily the route that must be chosen by the PE. In order to choose the proper route, the procedures followed in section 9 MUST be followed.

The upstream multicast hop is identified in an Extended Communities attribute to facilitate the optional use of filters which can prevent the distribution of the update to BGP speakers other than the upstream multicast hop.

When a PE distributes this information via BGP, it must include a Route Import Extended Communities attribute learned from the RPF information.

Note that for these procedures to work the VPN-IP route MUST contain the RPF information.

Note that 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 which 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 the "source active" A-D routes, that are described in sections 9 and 10. Detailed procedures for turning on explicit tracking 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.

6. I-PMSI Instantiation

This section describes how tunnels in the SP network can be used to instantiate an I-PMSI for an MVPN on a PE. When C-multicast data is delivered on an I-PMSI, the data will go to all PEs that are on the path to receivers for that C-group, but may also go to PEs that are not on the path to receivers for that C-group.

The tunnels which instantiate I-PMSIs can be either PE-PE unicast tunnels or P-multicast trees. When PE-PE unicast tunnels are used the PMSI is said to be instantiated using ingress replication. The instantiation of a tunnel for an I-PMSI is a matter of local policy decision and is not mandatory. Even for a site attached to multicast sources, transport of customer multicast traffic can be accommodated with S-PMSI-bound tunnels only

[Editor's Note: MD trees described in [ROSEN-8], [MVPN-BASE]] are an example of P-multicast trees. Also Aggregate Trees described in [RAGGARWA-MCAST] are an example of P-multicast trees.]

6.1. MVPN Membership and Egress PE Auto-Discovery

As described in section 4 a PE discovers the MVPN membership information of other PEs using BGP auto-discovery mechanisms or using a mechanism that instantiates a MI-PMSI interface. When a PE supports only a UI-PMSI service for an MVPN, it MUST rely on the BGP auto-discovery mechanisms for discovering this information. This information also results in a PE in the sender sites set discovering the leaves of the P-multicast tree, which are the egress PEs that have sites in the receiver sites set in one or more MVPNs mapped onto the tree.

6.1.1. Auto-Discovery for Ingress Replication

In order for a PE to use Unicast 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 tunnel have demultiplexing information which the receiver can associate with a particular MVPN.

If ingress replication is being used 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. The demultiplexor value is a downstream-assigned MPLS label (i.e., assigned by the PE that originated the A-D route, to be

used by other PEs when they send multicast packets on a unicast tunnel to that PE).

Other demultiplexing procedures for unicast are under consideration.

6.1.2. Auto-Discovery for P-Multicast Trees

A PE announces the P-multicast technology it supports for a specified MVPN, as part of the BGP MVPN membership discovery. This allows other PEs to determine the P-multicast technology they can use for building P-multicast trees to instantiate an I-PMSI. If a PE has a default tree instantiation of an I-PMSI, it also announces the tree identifier as part of the auto-discovery, as well as announcing its aggregation capability.

The announcement of a tree identifier at discovery time is only possible if the tree already exists (e.g., a preconfigured "traffic engineered" tunnel), or if the tree can be constructed dynamically without any PE having to know in advance all the other PEs on the tree (e.g., the tree is created by receiver-initiated joins).

6.2. C-Multicast Routing Information Exchange

When a PE doesn't support the use of a MI-PMSI for a given MVPN, it MUST either unicast MVPN routing information using PIM or else use BGP for exchanging the MVPN routing information.

6.3. Aggregation

A P-multicast tree can be used to instantiate a PMSI service for only one MVPN or for more than one MVPN. When a P-multicast tree is shared across multiple MVPNs it is termed an Aggregate Tree [RAGGARWA-MCAST]. The procedures described in this document allow a single SP multicast tree to be shared across multiple MVPNs. The procedures that are specific to aggregation are optional and are explicitly pointed out. Unless otherwise specified a P-multicast tree technology supports aggregation.

Aggregate Trees allow a single P-multicast tree to be used across multiple MVPNs and hence 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 an 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 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 which 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.

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 to agree on the PMSIs which 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 a 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 tradeoff 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 a 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 hat a particular PE may receive traffic for.

6.3.3. Encapsulation of the Aggregate Tree

An Aggregate Tree may use an IP/GRE encapsulation or an MPLS encapsulation. The protocol type in the IP/GRE header in the former case and the protocol type in the data link header in the latter need further explanation. This will be specified in a separate document.

6.3.4. Demultiplexing C-multicast traffic

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 label assignment 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 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 that the packet was received over. 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.

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,G). (This could be done using the same technique this is used to assign a particular C-(S,G) to an S-PMSI, see section 7.3.)

6.4. Mapping Received Packets to MVPNs

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. It also needs to determine the RPF interface for the C-multicast data packet. In order to do this the egress PE needs to determine the tunnel that the packet was received on. 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.

6.4.1. Unicast Tunnels

When ingress replication is used, the MVPN to which the received C-multicast data packet belongs can be determined by the MPLS label that was allocated by the egress. This label is distributed by the egress. This also determines the RPF interface for the C-multicast data packet.

6.4.2. Non-Aggregated P-Multicast Trees

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 the P-multicast tree is associated with.

There are different ways in which the egress PE can learn this association:

a) Configuration. The P-multicast tree that a particular MVPN belongs to is configured on each PE.

   [Editor's Note: PIM-SM Default MD trees in [ROSEN-8] and [MVPN-BASE] are examples of configuring the P-multicast tree and MVPN association]

b) BGP based advertisement of the P-multicast tree - MPVN mapping after the root of the tree discovers the leaves of the tree. The root of the tree sets up the tree after discovering each of the PEs that belong to the MVPN. It then advertises the P-multicast tree - MVPN mapping to each of the leaves. This mechanism can be used with both source initiated trees [e.g. RSVP-TE P2MP LSPs] and receiver initiated trees [e.g. PIM trees].

   [Editor's Note: Aggregate tree advertisements in [RAGGARWA-MCAST] are examples of this.]

c) BGP based advertisement of the P-multicast tree - MVPN mapping as part of the MVPN membership discovery. The root of the tree advertises, to each of the other PEs that belong to the MVPN, the P-multicast tree that the MVPN is associated with. This implies that the root doesn't need to know the leaves of the tree beforehand. This is possible only for receiver initiated trees e.g. PIM based trees.

   [Editor's Note: PIM-SSM discovery in [ROSEN-8] is an example of the above]

Both of the above require the BGP based advertisement to contain the P-multicast tree identifier. This identifier is encoded as a BGP attribute and contains the following elements:

- Tunnel Type.

- Tunnel identifier. The semantics of the identifier is determined by the tunnel type.

6.4.3. Aggregate P-Multicast Trees

Once a PE sets up an Aggregate Tree it needs to announce the C-multicast groups being mapped to this tree to other PEs in the network. This procedure is referred to as Aggregate Tree discovery. For an Aggregate Tree with an inclusive mapping this discovery implies announcing:

- The mapping of all MVPNs mapped to the Tree.

- For each MVPN mapped onto the tree the inner label allocated for it by the ingress PE. The use of this label is explained in the demultiplexing procedures of section 6.3.4.

- The P-multicast tree Identifier

The egress PE creates a logical interface corresponding to the tree identifier. This interface is the RPF interface for all the <C-Source, C-Group> entries mapped to that tree.

When PIM is used to setup P-multicast trees, the egress PE also Joins the P-Group Address corresponding to the tree. This results in setup of the PIM P-multicast tree.

6.5. I-PMSI Instantiation Using 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 an Unicast Tunnel to the remote PE. IP/GRE Tunnels or MPLS LSPs are examples of unicast tunnels that may be used. Note

that the same Unicast Tunnel can be used to transport packets

belonging to different MVPNs.

Ingress replication can be used to instantiate a UI-PMSI. The PE sets up unicast tunnels to each of the remote PEs that support ingress replication. For a given MVPN all C-multicast data packets are sent to each of the remote PEs in the MVPN that support ingress replication. Hence a remote PE may receive C-multicast data packets for a group even if it doesn't have any receivers in that group.

Ingress replication can also be used to instantiate a MI-PMSI. In this case each PE has a mesh of unicast tunnels to every other PE in that MVPN.

However when ingress replication is used it is recommended that only S-PMSIs be used. Instantiation of S-PMSIs with ingress replication is described in section 7.2. Note that this requires the use of explicit tracking, i.e., a PE must know which of the other PEs have receivers for each C-multicast tree.

6.6. Establishing P-Multicast Trees

It is believed that the architecture outlined in this document places no limitations on the protocols used to instantiate P-multicast trees. However, the only protocols being explicitly considered are PIM-SM, PIM-SSM, PIM-Bidir, RSVP-TE, and mLDP.

A P-multicast tree can be either a source tree or a shared tree. A source tree is used to carry traffic only for the multicast VRFs that exist locally on the root of the tree i.e. for which the root has local CEs. The root is a PE router. Source P-multicast trees can be instantiated using PIM-SM, PIM-SSM, RSVP-TE P2MP LSPs, and mLDP P2MP LSPs.

A shared tree on the other hand can be used to carry traffic belonging to VRFs that exist on other PEs as well. The root of a shared tree is not necessarily one of the PEs in the MVPN. All PEs that use the shared tree will send MVPN data packets to the root of the shared tree; if PIM is being used as the control protocol, PIM control packets also get sent to the root of the shared tree. This may require an unicast tunnel between each of these PEs and the root. The root will then send them on the shared tree and all the PEs that are leaves of the shared tree will receive the packets. For example a RP based PIM-SM tree would be a shared tree. Shared trees can be instantiated using PIM-SM, PIM-SSM, PIM-Bidir, RSVP-TE P2MP LSPs, mLDP P2MP LSPs, and mLDP MP2MP LSPs.. Aggregation support for bidirectional P-trees (i.e., PIM-Bidir trees or mLDP MP2MP trees) is

for further study. Shared trees require all the PEs to discover the root of the shared tree for a MVPN. To achieve this the root of a shared tree advertises as part of the BGP based MVPN membership discovery:

- The capability to setup a shared tree for a specified MVPN.

- A downstream assigned label that is to be used by each PE to encapsulate a MVPN data packet, when they send this packet to the root of the shared tree.

- A downstream assigned label that is to be used by each PE to encapsulate a MVPN control packet, when they send this packet to the root of the shared tree.

Both a source tree and a shared tree can be used to instantiate an I-PMSI. If a source tree is used to instantiate an UI-PMSI for a MVPN, all the other PEs that belong to the MVPN, must be leaves of the source tree. If a shared tree is used to instantiate a UI-PMSI for a MVPN, all the PEs that are members of the MVPN must be leaves of the shared tree.

6.7. RSVP-TE P2MP LSPs

This section describes procedures that are specific to the usage of RSVP-TE P2MP LSPs for instantiating a UI-PMSI. The RSVP-TE P2MP LSP can be either a source tree or a shared tree. Procedures in [RSVP-P2MP] are used to signal the LSP. The LSP is signaled after the root of the LSP discovers the leaves. The egress PEs are discovered using the MVPN membership procedures described in section 4. RSVP-TE P2MP LSPs can optionally support aggregation.

6.7.1. P2MP TE LSP Tunnel - MVPN Mapping

P2MP TE LSP Tunnel to MVPN mapping can be learned at the egress PEs using either option (a) or option (b) described in section 6.4.2. Option (b) i.e. BGP based advertisements of the P2MP TE LSP Tunnel - MPVN mapping require that the root of the tree include the P2MP TE LSP Tunnel identifier as the tunnel identifier in the BGP advertisements. This identifier contains the following information elements:

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

- RSVP-TE P2MP Tunnel's SESSION Object

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

6.7.2. Demultiplexing C-Multicast Data Packets

Demultiplexing the C-multicast data packets at the egress PE follow procedures described in section 6.3.4. The RSVP-TE P2MP LSP Tunnel must be signaled with penultimate-hop-popping (PHP) off. Signaling the P2MP TE LSP Tunnel with PHP off requires an extension to RSVP-TE which will be described later.

7. 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 which 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" criteria: 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 tunnel is a source initiated tree, such as a 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 to the S-PMSI. Note that the PE could create the identity of the P-multicast tree prior to the actual instantiation of the 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 MAY be established after the leaves receives the binding. The leaves MUST not switch to the S-PMSI until they receive both the binding and the tree signaling message.

7.1. S-PMSI Instantiation Using Ingress Replication

As described in section 6.1.1, ingress replication can be used to instantiate a UI-PMSI. However this can result in a PE receiving

packets for a multicast group for which it doesn't have any receivers. This can be avoided if the ingress PE tracks the remote PEs which have receivers in a particular C-multicast group. In order to do this it needs to receive C-Joins from each of the remote PEs. It then replicates the C-multicast data packet and sends it to only those egress PEs which are on the path to a receiver of that C-group. It is possible that each PE that is using ingress replication instantiates only S-PMSIs. It is also possible that some PEs instantiate UI-PMSIs while others instantiate only S-PMSIs. In both these cases the PE MUST either unicast MVPN routing information using PIM or use BGP for exchanging the MVPN routing information. This is because there may be no MI-PMSI available for it to exchange MVPN routing information.

Note that the use of ingress replication doesn't require any extra procedures for signaling the binding of the S-PMSI from the ingress

PE to the egress PEs. The procedures described for I-PMSIs are sufficient.

7.2. Protocol for Switching to S-PMSIs

We describe two protocols for switching to S-PMSIs. These protocols can be used when the tunnel that instantiates the S-PMSI is a P-multicast tree.

7.2.1. A UDP-based Protocol for Switching to S-PMSIs

This procedure can be used for any MVPN which has an MI-PMSI. Traffic from all multicast streams in a given MPVN is sent, by default, on the MI-PMSI. Consider a single multicast stream within a given MVPN, and consider a PE which is attached to a source of multicast traffic for that stream. The PE can be configured to move the stream from the MI-PMSI to an S-PMSI if certain configurable conditions are met. To do this, it needs to inform all the PEs which attach to receivers for stream. These PEs need to start listening for traffic on the S-PMSI, and the transmitting PE may start sending traffic on the S-PMSI when it is reasonably certain that all receiving PEs are listening on the S-PMSI.

7.2.1.1. Binding a Stream to an S-PMSI

When a PE which attaches to a transmitter for a particular multicast stream notices that the conditions for moving the stream to an S-PMSI are met, it begins to periodically send an "S-PMSI Join Message" on the MI-PMSI. The S-PMSI Join 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 which is to be bound to the S-PMSI. This can be represented as an (S,G) pair.

- An identifier for the particular S-PMSI to which the stream is to be bound. This identifier is a structured field which 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.

    * Any demultiplexing information needed by the tunnel encapsulation protocol to identify the particular S-PMSI. This allows a single tunnel to aggregate multiple S-PMSIs. If a particular tunnel is not aggregating multiple S-PMSIs, then no demultiplexing information is needed.

A PE router which is not connected to a receiver will still receive the S-PMSI Joins, and MAY cache the information contained therein. Then if the PE later finds that it is attached to a receiver, it can immediately start listening to the S-PMSI.

Upon receiving the S-PMSI Join, PE routers connected to receivers for

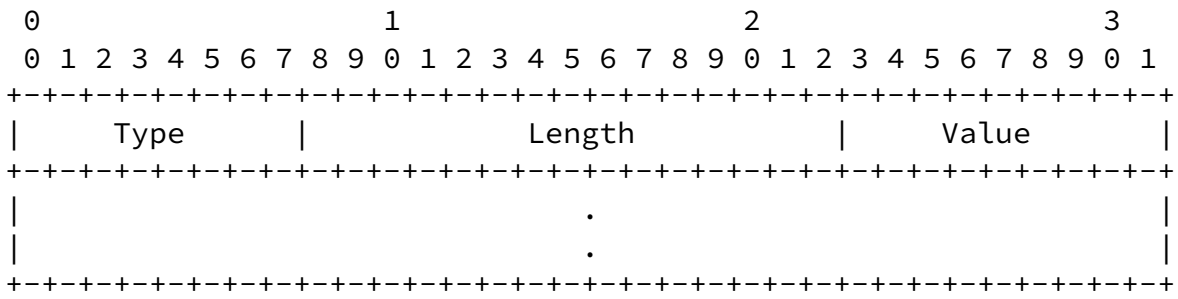
the specified stream will take whatever action is necessary to start receiving multicast data packets on the S-PMSI. The precise action taken will depend upon the tunnel type.

After a configurable delay, the PE router which is sending the S-PMSI Joins will start transmitting the stream's data packets on the S-PMSI.

When the pre-configured conditions are no longer met for a particular stream, e.g. the traffic stops, the PE router connected to the source stops announcing S-PMSI Joins for that stream. Any PE that does not receive, over a configurable interval, an S-PMSI Join for a particular stream will stop listening to the S-PMSI.

7.2.1.2. Packet Formats and Constants

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

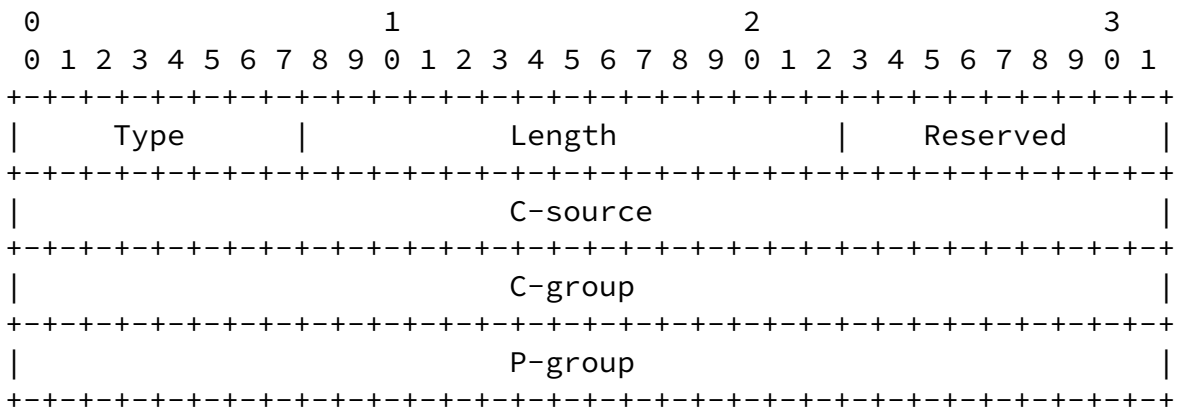


Type (8 bits)

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

Value (variable length)

Currently 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 which is used to carry a single multicast stream, where the packets of that stream have IPv4 source and destination IP addresses.



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

The protocol uses the following constants.

[S-PMSI_DELAY]:

the PE router which 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 tunnel for this amount of time, the PE will prune itself from the S-PMSI 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.

7.2.2. A BGP-based Protocol for Switching to S-PMSIs

This procedure can be used for a MVPN that is using either a UI-PMSI or a MI-PMSI. Consider a single multicast stream for a C-(S, G) within a given MVPN, and consider a PE which is attached to a source of multicast traffic for that stream. The PE can be configured to move the stream from the MI-PMSI or UI-PMSI to an S-PMSI if certain configurable conditions are met. Once a PE decides to move the C-(S, G) for a given MVPN to a S-PMSI, it needs to instantiate the S-PMSI using a tunnel and announce to all the egress PEs, that are on the path to receivers of the C-(S, G), of the binding of the S-PMSI to the C-(S, G). The announcement is done using BGP. Depending on the tunneling technology used, this announcement may be done before or after setting up the tunnel. The source and egress PEs have to switch to using the S-PMSI for the C-(S, G).

7.2.2.1. Advertising C-(S, G) Binding to a S-PMSI using BGP

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

a) IP address of the originating PE

b) The RD configured locally for the MVPN. This is required to uniquely identify the <C-Source, C-Group> as the addresses could overlap between different MVPNs. This is the same RD value used in the auto-discovery process.

c) The C-Source address. This address can be a prefix in order to allow a range of C-Source addresses to be mapped to an Aggregate Tree.

d) The C-Group address. This address can be a range in order to allow a range of C-Group addresses to be mapped to an Aggregate Tree.

e) A PE MAY aggregate two or more S-PMSIs originated by the PE onto the same P-Multicast tree. If the PE already advertises S-PMSI auto-discovery routes for these S-PMSIs, then aggregation 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 auto-discovery routes for these S-PMSIs, then the aggregation requires the PE to advertise (new) S-PMSI auto-discovery routes for these S-PMSIs. The PMSI Tunnel attribute in the newly advertised/re-advertised routes MUST carry the identity of the P- Multicast tree that aggregates the S-PMSIs. If at least some of the S-PMSIs aggregated onto the same P-Multicast tree belong to different MVPNs, then all these routes MUST carry an MPLS upstream assigned label [MPLS-UPSTREAM-LABEL, section 6.3.4]. If all these aggregated S-PMSIs belong to the same MVPN, then the routes MAY carry an MPLS upstream assigned label [MPLS-UPSTREAM-LABEL]. The labels MUST be distinct on a per MVPN basis, and MAY be distinct on a per route basis.

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

1. An identifier for the particular S-PMSI to which the stream is to be bound. This identifier is a structured field which 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.

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

7.2.2.2. 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 S-PMSI. Then any PE which 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 before binding the flow to an S-PMSI, it can do so by sending an A-D route identifying the flow but not specifying an S-PMSI. This will elicit the Leaf A-D Routes. This is useful when the PE needs to know the receivers before selecting an S-PMSI.

7.2.2.3. Switching to S-PMSI

After the egress PEs receive the announcement they setup their forwarding path to receive traffic on the S-PMSI if they have one or more receivers interested in the <C-S, C-G> bound to the S-PMSI. This involves changing the RPF interface for the relevant <C-S, C-G> entries to the interface that is used to instantiate the S-PMSI. If an Aggregate Tree is used to instantiate a S-PMSI this also implies setting up the demultiplexing forwarding entries based on the inner label as described in section 6.3.4. The egress PEs may perform the switch to the S-PMSI once the advertisement from the ingress PE is received or wait for a preconfigured timer to do so.

A source PE may use one of two approaches to decide when to start transmitting data on the S-PMSI. In the first approach once the source PE instantiates the S-PMSI, it starts sending multicast packets for <C-S, C-G> entries mapped to the S-PMSI on both that as well as on the I-PMSI, which is currently used to send traffic for

the <C-S, C-G>. After some preconfigured timer the PE stops sending multicast packets for <C-S, C-G> on the I-PMSI. In the second approach after a certain pre-configured delay after advertising the <C-S, C-G> entry bound to a S-PMSI, the source PE begins to send traffic on the S-PMSI. At this point it stops to send traffic for the <C-S, C-G> on the I-PMSI. This traffic is instead transmitted on the S-PMSI.

7.3. Aggregation

S-PMSIs can be aggregated on a P-multicast tree. The S-PMSI to C-(S, 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.

7.4. Instantiating the S-PMSI with a PIM Tree

The procedures of section 7.3 tell a PE when it must start listening and stop listening to a particular S-PMSI. Those procedures also specify the method for instantiating the S-PMSI. In this section, we provide the procedures to be used when the S-PMSI is instantiated as a PIM tree. The PIM tree is created by the PIM P-instance.

If a single PIM tree is being used to aggregate multiple S-PMSIs, then the PIM tree to which a given stream is bound may have already been joined by a given receiving PE. If the tree does not already exist, then the appropriate PIM procedures to create it must be executed in the P-instance.

If the S-PMSI for a particular multicast stream is instantiated as a PIM-SM or PIM-Bidir tree, the S-PMSI identifier will specify the RP and the group P-address, and the PE routers which have receivers for that stream must build a shared tree toward the RP.

If the S-PMSI is instantiated as a PIM-SSM tree, the PE routers build a source tree toward the PE router that is advertising the S-PMSI Join. The IP address root of the tree is the same as the source IP address which appears in the S-PMSI Join. In this case, the tunnel identifier in the S-PMSI Join will only need to specify a group P-address.

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 PIM-SSM is used to set up the tunnels, then the PEs may be with overlapping sets of group P-addresses. If PIM-SSM is not used, 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 PIM-SSM is used, so the use of PIM-SSM is strongly recommended whenever PIM trees are used to instantiate S-PMSIs.

If it is known that all the PEs which need to receive data traffic on a given S-PMSI can support aggregation of multiple S-PMSIs on a single PIM tree, then the transmitting PE, may, at its discretion,

decide to bind the S-PMSI to a PIM tree which is already bound to one or more other S-PMSIs, from the same or from different MVPNs. In this case, appropriate demultiplexing information must be signaled.

7.5. Instantiating S-PMSIs using RSVP-TE P2MP Tunnels

RSVP-TE P2MP Tunnels can be used for instantiating S-PMSIs. Procedures described in the context of I-PMSIs in section 6.7 apply.

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 tunnels. This document describes two different types of inter-AS tunnel:

1. "Segmented Inter-AS tunnels"

   A segmented inter-AS tunnel consists of a number of independent segments which are stitched together at the ASBRs. There are two types of segment, inter-AS segments and intra-AS segments. The segmented inter-AS 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 tunnels.

Intra-AS segments connect ASBRs and PEs which 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 an intra-AS segment of an inter-AS tunnel is distinct from any intra-AS tunnel in the AS.

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

2. "Non-segmented Inter-AS tunnels"

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

tunnel.

[Editor's Note: This is the model in [ROSEN-8] and [MVPN-BASE].]

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 option B or option C is 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 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. The ASBRs play no special role, but function merely as P routers.

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.

8.1.2. Inter-AS MVPN Routing Information Exchange

MVPN routing information exchange can be done by PIM peering (either lightweight or full) across an MI-PMSI, or by unicasting PIM messages. The method of using BGP to send MVPN routing information can also be used.

If any form of PIM peering is used, a PE that sends C-PIM Join/Prune messages for a particular C-(S,G) must be able to identify the PE which is its PIM adjacency on the path to S. The identity of the PIM adjacency is determined from the RPF information associated with the VPN-IP route to S.

If no RPF information is present, then the identity of the PIM adjacency is taken from the BGP Next Hop attribute of the VPN-IP route to S. Note that this will not give the correct result if option b of section 10 of [RFC4364] is used. To avoid this possibility of error, the RPF information SHOULD always be present if MVPN routing information is to be distributed by PIM.

If BGP (rather than PIM) is used to distribute the MVPN routing information, and if option b of section 10 of [RFC4364] is in use, then the MVPN 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 MVPN routes are not installed in the ASBRs. The handling of MVPN 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 I-PMSI

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

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

2. Inter-AS PIM-SM or PIM-SSM based trees rely on a PE joining a (P-S, P-G) tuple where P-S is the address of a PE in another AS. This (P-S, P-G) tuple is learned using the MVPN membership and BGP MVPN-tunnel binding procedures described earlier. However, if the source of the tree is in a different AS than a particular P router, it is possible that the P router will not have a route to the source. For example, the remote AS may be using BGP to distribute a route to the source, but a particular P router may be part of a "BGP-free core", in which the P routers are not aware of BGP-distributed routes.

   In such a case it is necessary for a PE to to tell PIM to construct the tree through a particular BGP speaker, the "BGP next hop" for the tree source. This can be accomplished with a PIM extension, in which the P-PIM Join/Prune messages carry a new "proxy" field which contains the address of that BGP next hop. As the P-multicast tree is constructed, it is built towards the proxy (the BGP next hop) rather than towards P-S, so the P routers will not need to have a route to P-S.

   Support for inter-AS trees using PIM-Bidir are for further study.

   When the BGP-based discovery procedures for MVPN are in place, one can distinguish two different inter-AS routes to a particular P-S:

- BGP will install a unicast route to P-S along a particular path, using the IP AFI/SAFI ;

- A PE's MVPN auto-discovery information is advertised by sending a BGP update whose NLRI is in a special address

family (AFI/SAFI) used for this purpose. The NLRI of the address family contains the IP address of the PE, as well as an RD. If the NLRI contains the IP address of P-S, this in effect creates a second route to P-S. This route might follow a different path than the route in the unicast IP family.

When building a PIM tree towards P-S, it may be desirable to build it along the route on which the MVPN auto-discovery AFI/SAFI is installed, rather than along the route on which the IP AFI/SAFI is installed. This enables the inter-AS portion of the tree to follow a path which is specifically chosen for multicast (i.e., it allows the inter-AS multicast topology to be "non-congruent" to the inter-AS unicast topology).

In order for P routers to send P-Join/Prune messages along this path, they need to make use of the "proxy" field extension discussed above. The PIM message must also contain the full NLRI in the MVPN auto-discovery family, so that the BGP speakers can look up that NLRI to find the BGP next hop.

3. Procedures in [RSVP-P2MP] are used for inter-AS RSVP-TE P2MP Tunnels.

8.1.4. Inter-AS S-PMSI

The leaves of the tunnel are discovered using the MVPN routing information. Procedures for setting up the tunnel are similar to the ones described in section 8.2.3 for an inter-AS I-PMSI.

8.2. Segmented Inter-AS Tunnels

8.2.1. Inter-AS MVPN Auto-Discovery Routes

The BGP based MVPN membership discovery procedures of section 4 are used to auto-discover the intra-AS MVPN membership. This section describes the additional procedures for inter-AS MVPN membership discovery. It also describes the procedures for constructing segmented inter-AS tunnels.

In this case, for a given MVPN in an AS, the objective is to form a

spanning tree of MVPN membership, rooted at the AS. The nodes of this tree are ASes. The leaves of this tree are only those ASes that have at least one PE with a member in the MVPN. The inter-AS tunnel used to instantiate an inter-AS PMSI must traverse this spanning tree. A given AS needs to announce to another AS only the fact that it has membership in a given MVPN. It doesn't need to announce the membership of each PE in the AS to other ASes.

This section defines an inter-AS auto-discovery route as a route that carries information about an AS that has one or more PEs (directly) connected to the site(s) of that MVPN. Further it defines an inter-AS leaf auto-discovery route (leaf auto-discovery route) as a route used to inform the root of an intra-AS segment, of an inter-AS tunnel, of a leaf of that intra-AS segment.

8.2.1.1. Originating Inter-AS MVPN A-D Information

A PE in a given AS advertises its MVPN membership to all its IBGP peers. This IBGP peer may be a route reflector which in turn advertises this information to only its IBGP peers. In this manner all the PEs and ASBRs in the AS learn this membership information.

An Autonomous System Border Router (ASBR) may be configured to support a particular MVPN. If an ASBR is configured to support a particular MVPN, the ASBR MUST participate in the intra-AS MVPN auto-discovery/binding procedures for that MVPN within the AS that the ASBR belongs to, as defined in this document.

Each ASBR then advertises the "AS MVPN membership" to its neighbor ASBRs using EBGP. This inter-AS auto-discovery route must not be advertised to the PEs/ASBRs in the same AS as this ASBR. The advertisement carries the following information elements:

a. A Route Distinguisher for the MVPN. For a given MVPN each ASBR in the AS must use the same RD when advertising this information to other ASBRs. To accomplish this all the ASBRs within that AS, that are configured to support the MVPN, MUST be configured with the same RD for that MVPN. This RD MUST be of Type 0, MUST embed the autonomous system number of the AS.

b. The announcing ASBR's local address as the next-hop for the above information elements.

c. By default the BGP Update message MUST carry export Route Targets used by the unicast routing of that VPN. The default could be modified via configuration by having a set of Route Targets used for the inter-AS auto-discovery routes being

distinct from the ones used by the unicast routing of that VPN.

8.2.1.2. Propagating Inter-AS MVPN A-D Information

As an inter-AS auto-discovery route originated by an ASBR within a given AS is propagated via BGP to other ASes, this results in creation of a data plane tunnel that spans multiple ASes. This tunnel is used to carry (multicast) traffic from the MVPN sites connected to the PEs of the AS to the MVPN sites connected to the PEs that are in the other ASes. Such tunnel consists of multiple intra-AS segments (one per AS) stitched at ASBRs' boundaries by single hop <ASBR-ASBR> LSP segments.

An ASBR originates creation of an intra-AS segment when the ASBR receives an inter-AS auto-discovery route from an EBGP neighbor. Creation of the segment is completed as a result of distributing via IBGP this route within the ASBR's own AS.

For a given inter-AS tunnel each of its intra-AS segments could be constructed by its own independent mechanism. Moreover, by using upstream labels within a given AS multiple intra-AS segments of different inter-AS tunnels of either the same or different MVPNs may share the same P-Multicast Tree.

Since (aggregated) inter-AS auto-discovery routes have granularity of <AS, MVPN>, an MVPN that is present in N ASes would have total of N inter-AS tunnels. Thus for a given MVPN the number of inter-AS tunnels is independent of the number of PEs that have this MVPN.

The following sections specify procedures for propagation of (aggregated) inter-AS auto-discovery routes across ASes.

8.2.1.2.1. Inter-AS Auto-Discovery Route received via EBGP

When an ASBR receives from one of its EBGP neighbors a BGP Update message that carries the inter-AS auto-discovery route if (a) at least one of the Route Targets carried in the message matches one of the import Route Targets configured on the ASBR, and (b) the ASBR determines that the received route is the best route to the

destination carried in the NLRI of the route, the ASBR:

a) Re-advertises this inter-AS auto-discovery route within its own AS.

   If the ASBR uses ingress replication to instantiate the intra-AS segment of the inter-AS tunnel, the re-advertised route SHOULD carry a Tunnel attribute with the Tunnel Identifier set to Ingress Replication, but no MPLS labels.

   If a P-Multicast Tree is used to instantiate the intra-AS segment of the inter-AS tunnel, and in order to advertise the P-Multicast tree identifier the ASBR doesn't need to know the leaves of the tree beforehand, then the advertising ASBR SHOULD advertise the P-Multicast tree identifier in the Tunnel Identifier of the Tunnel attribute. This, in effect, creates a binding between the inter-AS auto-discovery route and the P-Multicast Tree.

   If a P-Multicast Tree is used to instantiate the intra-AS segment of the inter-AS tunnel, and in order to advertise the P-Multicast tree identifier the advertising ASBR needs to know the leaves of the tree beforehand, the ASBR first discovers the leaves using the Auto-Discovery procedures, as specified further down. It then advertises the binding of the tree to the inter-AS auto-discovery route using the the original auto-discovery route with the addition of carrying in the route the Tunnel attribute that contains the type and the identity of the tree (encoded in the Tunnel Identifier of the attribute).

b) Re-advertises the received inter-AS auto-discovery route to its EBGP peers, other than the EBGP neighbor from which the best inter-AS auto-discovery route was received.

c) Advertises to its neighbor ASBR, from which it received the best inter-AS autodiscovery route to the destination carried in the NRLI of the route, a leaf auto-discovery route that carries

an ASBR-ASBR tunnel binding with the tunnel identifier set to ingress replication. This binding as described in section 6 can be used by the neighbor ASBR to send traffic to this ASBR.

8.2.1.2.2. Leaf Auto-Discovery Route received via EBGP

When an ASBR receives via EBGP a leaf auto-discovery route, the ASBR finds an inter-AS auto-discovery route that has the same RD as the leaf auto-discovery route. The MPLS label carried in the leaf auto-discovery route is used to stitch a one hop ASBR-ASBR LSP to the tail of the intra-AS tunnel segment associated with the inter-AS auto-

discovery route.

8.2.1.2.3. Inter-AS Auto-Discovery Route received via IBGP

If a given inter-AS auto-discovery route is advertised within an AS by multiple ASBRs of that AS, the BGP best route selection performed by other PE/ASBR routers within the AS does not require all these PE/ASBR routers to select the route advertised by the same ASBR - to the contrary different PE/ASBR routers may select routes advertised by different ASBRs.

Further when a PE/ASBR receives from one of its IBGP neighbors a BGP Update message that carries a AS MVPN membership tree , if (a) the route was originated outside of the router's own AS, (b) at least one of the Route Targets carried in the message matches one of the import Route Targets configured on the PE/ASBR, and (c) the PE/ASBR determines that the received route is the best route to the destination carried in the NLRI of the route, if the router is an ASBR then the ASBR propagates the route to its EBGP neighbors. In addition the PE/ASBR performs the following.

If the received inter-AS auto-discovery route carries the Tunnel attribute with the Tunnel Identifier set to LDP P2MP LSP, or PIM-SSM tree, or PIM-SM tree, the PE/ASBR SHOULD join the P-Multicast tree whose identity is carried in the Tunnel Identifier.

If the received source auto-discovery route carries the Tunnel

attribute with the Tunnel Identifier set to RSVP-TE P2MP LSP, then the ASBR that originated the route MUST signal the local PE/ASBR as one of leaf LSRs of the RSVP-TE P2MP LSP. This signaling MAY have been completed before the local PE/ASBR receives the BGP Update message.

If the NLRI of the route does not carry a label, then this tree is an intra-AS LSP segment that is part of the inter-AS Tunnel for the MVPN advertised by the inter-AS auto-discovery route. If the NLRI carries a (upstream) label, then a combination of this tree and the label identifies the intra-AS segment.

If this is an ASBR, this intra-AS segment may further be stitched to ASBR-ASBR inter-AS segment of the inter-AS tunnel. If the PE/ASBR has local receivers in the MVPN, packets received over the intra-AS segment must be forwarded to the local receivers using the local VRF.

If the received inter-AS auto-discovery route either does not carry the Tunnel attribute, or carries the Tunnel attribute with the Tunnel Identifier set to ingress replication, then the PE/ASBR originates a

new auto-discovery route to allow the ASBR from which the auto-discovery route was received, to learn of this ASBR as a leaf of the intra-AS tree.

Thus the AS MVPN membership information propagates across multiple ASes along a spanning tree. BGP AS-Path based loop prevention mechanism prevents loops from forming as this information propagates.

8.2.2. Inter-AS MVPN Routing Information Exchange

All of the MVPN routing information exchange methods specified in section 5 can be supported across ASes.

The objective in this case is to propagate the MVPN routing information to the remote PE that originates the unicast route to C-S/C-RP, in the reverse direction of the AS MVPN membership information announced by the remote PE's origin AS. This information is processed by each ASBR along this reverse path.

To achieve this the PE that is generating the MVPN routing

advertisement, first determines the source AS of the unicast route to C-S/C-RP. It then determines from the received AS MVPN membership information, for the source AS, the ASBR that is the next-hop for the best path of the source AS MVPN membership. The BGP MVPN routing update is sent to this ASBR and the ASBR then further propagates the BGP advertisement. BGP filtering mechanisms ensure that the BGP MVPN routing information updates flow only to the upstream router on the reverse path of the inter-AS MVPN membership tree. Details of this filtering mechanism and the relevant encoding will be specified in a separate document.

8.2.3. Inter-AS I-PMSI

All PEs in a given AS, use the same inter-AS heterogeneous tunnel, rooted at the AS, to instantiate an I-PMSI for an inter-AS MVPN service. As explained earlier the intra-AS tunnel segments that comprise this tunnel can be built using different tunneling technologies. To instantiate an MI-PMSI service for a MVPN there must be an inter-AS tunnel rooted at each AS that has at least one PE that is a member of the MVPN.

A C-multicast data packet is sent using an intra-AS tunnel segment by the PE that first receives this packet from the MVPN customer site. An ASBR forwards this packet to any locally connected MVPN receivers for the multicast stream. If this ASBR has received a tunnel binding for the AS MVPN membership that it advertised to a neighboring ASBR,

it also forwards this packet to the neighboring ASBR. In this case the packet is encapsulated in the downstream MPLS label received from the neighboring ASBR. The neighboring ASBR delivers this packet to any locally connected MVPN receivers for that multicast stream. It also transports this packet on an intra-AS tunnel segment, for the inter-AS MVPN tunnel, and the other PEs and ASBRs in the AS then receive this packet. The other ASBRs then repeat the procedure followed by the ASBR in the origin AS and the packet traverses the overlay inter-AS tunnel along a spanning tree.

8.2.3.1. Support for Unicast VPN Inter-AS Methods

The above procedures for setting up an inter-AS I-PMSI can be

supported for each of the unicast VPN inter-AS models described in [RFC4364]. These procedures do not depend on the method used to exchange unicast VPN routes. For Option B and Option C they do require MPLS encapsulation between the ASBRs.

8.2.4. Inter-AS S-PMSI

An inter-AS tunnel for an S-PMSI is constructed similar to an inter-AS tunnel for an I-PMSI. Namely, such a tunnel is constructed as a concatenation of tunnel segments. There are two types of tunnel segments: an intra-AS tunnel segment (a segment that spans ASBRs within the same AS), and inter-AS tunnel segment (a segment that spans adjacent ASBRs in adjacent ASes). ASes that are spanned by a tunnel are not required to use the same tunneling mechanism to construct the tunnel - each AS may pick up a tunneling mechanism to construct the intra-AS tunnel segment of the tunnel on its

The PE that decides to set up a S-PMSI, advertises the S-PMSI tunnel binding using procedures in section 7.3.2 to the routers in its own AS. The <C-S, C-G> membership for which the S-PMSI is instantiated, is propagated along an inter-AS spanning tree. This spanning tree traverses the same ASBRs as the AS MVPN membership spanning tree. In addition to the information elements described in section 7.3.2 (Origin AS, RD, next-hop) the C-S and C-G is also advertised.

An ASBR that receives the AS <C-S, C-G> information from its upstream ASBR using EBGP sends back a tunnel binding for AS <C-S, C-G> information if a) at least one of the Route Targets carried in the message matches one of the import Route Targets configured on the ASBR, and (b) the ASBR determines that the received route is the best route to the destination carried in the NLRI of the route. If the ASBR instantiates a S-PMSI for the AS <C-S, C-G> it sends back a downstream label that is used to forward the packet along its intra-

AS S-PMSI for the <C-S, C-G>. However the ASBR may decide to use an AS MVPN membership I-PMSI instead, in which case it sends back the same label that it advertised for the AS MVPN membership I-PMSI. If the downstream ASBR instantiates a S-PMSI, it further propagates the <C-S, C-G> membership to its downstream ASes, else it does not.

An AS can instantiate an intra-AS S-PMSI for the inter-AS S-PMSI

tunnel only if the upstream AS instantiates a S-PMSI. The procedures allow each AS to determine whether it wishes to setup a S-PMSI or not and the AS is not forced to setup a S-PMSI just because the upstream AS decides to do so.

The leaves of an intra-AS S-PMSI tunnel will be the PEs that have local receivers that are interested in <C-S, C-G> and the ASBRs that have received MVPN routing information for <C-S, C-G>. Note that an AS can determine these ASBRs as the MVPN routing information is propagated and processed by each ASBR on the AS MVPN membership spanning tree.

The C-multicast data traffic is sent on the S-PMSI by the originating PE. When it reaches an ASBR that is on the spanning tree, it is delivered to local receivers, if any, and is also forwarded to the neighbor ASBR after being encapsulated in the label advertised by the neighbor. The neighbor ASBR either transports this packet on the S-PMSI for the multicast stream or an I-PMSI, delivering it to the ASBRs in its own AS. These ASBRs in turn repeat the procedures of the origin AS ASBRs and the multicast packet traverses the spanning tree.

9. Duplicate Packet Detection and Single Forwarder PE

An egress PE may receive duplicate multicast data packets, from more than one ingress PE, for a MVPN when a a site that contains C-S or C-RP is multihomed to more than one PE. An egress PE may also receive duplicate data packets for a MVPN, from two different ingress PEs, when the CE-PE routing protocol is PIM-SM and a router or a CE in a site switches from the C-RP tree to C-S tree.

For a given <C-S, C-G> a PE, say PE1, expects to receive C-data packets from the upstream PE, say PE2, which PE1 identified as the upstream multicast hop in the C-Multicast Routing Update that PE1 sent in order to join <C-S, C-G>. If PE1 can determine that a data packet for <C-S, C-G> was received from the expected upstream PE, PE2, PE1 will accept the packet. Otherwise, PE1 will drop the packet. (But see section 10 for an exception case where PE1 will accept a packet even if it is from an unexpected upstream PE.) This determination can be performed only if the PMSI on which the packets are being received and the tunneling technology used to instantiate

the PMSI allows the PE to determine the source PE that sent the packet. However this determination may not always be possible.

Therefore, procedures are needed to ensure that packets are received at a PE only from a single upstream PE. This is called single forwarder PE selection.

Single forwarder PE selection is achieved by the following set of procedures:

a. If there is more than one PE within the same AS through which C-S or C-RP of a given MVPN could be reached, and in the case of C-S not every such PE advertises an S-PMSI for <C-S, C-G>, all PEs that have this MVPN MUST send the MVPN routing information update for <C-S, C-G> or <C-*, C-G> to the same upstream PE. This is achieved using the following procedure:

   Using the procedure for "RPF determination" specified in section 5.1, find (a) the upstream multicast hop for the C-S or C-RP, and (b) the route used to reach the upstream multicast hop. Call this route the "installed RPF route" for C-S or C-RP.

   If the next-hop interface of the installed RPF route for C-S or C-RP is a VRF interface of the PE, then the PE uses that route to reach the C-S or C-RP.

   Otherwise, consider the set of all VPN-IP routes that are (a) eligible to be imported into the VRF (as determined by their Route Targets), (b) are eligible to be used for RPF determination (i.e., if RPF determination is done via a non-congruent multicast topology, this would include only the routes that are part of that topology), and (c) have exactly the same IP prefix as the installed RPF route.

   For each route in this set, determine the corresponding upstream PE. 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 attribute.

   This results in a set of pairs of <route, upstream PE>. The PE will 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 "selected RPF route". The PE will use the selected RPF route to reach the C-S or C-

RP.

b. The above procedure ensures that if C-S or C-RP is multi-homed to PEs within a single AS, a PE will not receive duplicate traffic as long as all the PEs in that AS are on either the C-S or C-RP tree.

   However the PE may receive duplicate traffic if C-S or C-RP is multi-homed to different ASes. In this case the PE can detect duplicate traffic as such duplicate traffic will arrive on a different tunnel - if the PE was expecting the traffic on an inter-AS tunnel, duplicate traffic will arrive on an intra-AS tunnel [this is not an intra-AS tunnel segment, of an inter-AS tunnel] and vice-versa.

   To achieve the above the PE has to keep track of which (inter-AS) auto-discovery route the PE uses for sending MVPN multicast routing information towards C-S/C-RP. Then the PE should receive (multicast) traffic originated by C-S/C-RP only from the (inter-AS) tunnel that was carried in the best source auto-discovery route for the MVPN and was originated by the AS that contains C-S/C-RP (where "the best" is determined by the PE). All other multicast traffic originated by C-S/C-RP, but received on any other tunnel should be discarded as duplicated.

   The PE may also receive duplicate traffic during a <C-*, C-G> to <C-S, C-G> switch. The issue and the solution are described next.

c. If the tunneling technology in use for a particular MVPN does not allow the egress PEs to identify the ingress PE, then having all the PEs select the same PE to be the upstream multicast hop is not sufficient to prevent packet duplication. The reason is that a single tunnel may be carrying traffic on both the (C-*, C-G) tree and the (C-S, C-G) tree. If some of the egress PEs have joined the source tree, but others expect to receive (S,G) packets from the shared tree, then two copies of data packet will travel on the tunnel, and the egress PEs will have no way to determine that only one copy should be accepted.

   To avoid this, it is necessary to ensure that once any PE joins the (C-S, C-G) tree, any other PE that has joined the (C-*, C-

G) tree also switches to the (C-S, C-G) tree (selecting, of course, the same upstream multicast hop, as specified above).

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 a Sparse Mode group, the PE that creates the state MUST originate an auto-discovery route as specified below. The route is being advertised using the same procedures as the MVPN auto-discovery/binding (both intra-AS and inter-AS) specified in this document with the following modifications:

1. The Multicast Source field MUST be set to C-S. The Multicast Source Length field is set appropriately to reflect this.

2. The Multicast Group field MUST be set to C-G. The Multicast Group Length field is set appropriately to reflect this.

The route goes to all the PEs of the MVPN. When a PE receives this route, it checks whether there are any receivers in the MVPN sites attached to the PE for the group carried in the route. If yes, then it generates a C-multicast route indicating Join for <C-S, C-G>. This forces all the PEs (in all ASes) to switch to the C-S tree for <C-S, C-G> from the C-RP tree.

This is the same type of A-D route used to report active sources in the scenarios described in section 10.

Note that when a PE thus joins the <C-S, C-G> tree, it may need to send a PIM (S,G,RPT-bit) prune to one of its CE PIM neighbors, as determined by ordinary PIM procedures..

Whenever the 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 auto-discovery route that was advertised when the state was created.

N.B.: SINCE ALL PES WITH RECEIVERS FOR GROUP C-G WILL JOIN THE C-S SOURCE TREE IF ANY OF THEM DO, IT IS NEVER NECESSARY TO DISTRIBUTE A BGP C-MULTICAST ROUTE FOR THE PURPOSE OF PRUNING SOURCES FROM THE SHARED TREE.

In summary when the CE-PE routing protocol for all PEs that belong to a MVPN is not PIM-SM, selection of a consistent upstream PE to reach C-S is sufficient to eliminate duplicates when C-S is multi-homed to a single AS. When C-S is multi-homed to multiple ASes, duplicate packet detection can be performed as the receiver PE can always determine whether packets arrived on the wrong tunnel. When the CE-PE

routing protocol is PIM-SM, additional procedures as described above are required to force all PEs within all ASes to switch to the C-S tree from the C-RP tree when any PE switches to the C-S tree.

10. Deployment Models

This section describes some optional deployment models and specific procedures for those deployment models.

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 ASM mode on the VPN customer site. To quote from [MVPN-REQ]:

"In some cases this engineering problem is not trivial: for instance, if sources and receivers are located in VPN sites that are different than that of the RP, then traffic may 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 is obviously not ideal. A multicast VPN solution SHOULD propose a way to help on solving this RP engineering 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 model is introduced in [RP-MVPN]. This section describes how

anycast-RP can be used for achieving this by advertising active sources. 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 these RPs use the same (anycast) address. All 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.

[Editor's Note: This is derived from the model in [RP-MVPN].]

10.1.2.1. Receiver(s) Within a Site

The PE which receives C-Join for (*,G) or (S,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 Join for C-(S,G). Sending this information could be done using any of the procedures described in section 5. If BGP is used, the ingress address is set to the upstream PE's address which has triggered the source active information. Only the upstream PE will process this information. If unicast PIM is used then a unicast PIM message will have to be sent to the PE upstream PE that has triggered the source active information. If a MI-PMSI is used than further clarification is needed on the upstream neighbor address of the PIM message and will be provided in a future revision.

10.1.2.2. Source Within a Site

When a PE receives PIM-Register 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 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 yes, then it generates an advertisement for C-(S,G) as specified in the previous section.

Note that the mechanism described in section 7.3.2. can be leveraged to advertise a S-PMSI binding along with the source active messages.

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 MSDP to talk to the C-RPs. In particular, a PE which 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, but they do not have to be C-RPs either.

MSDP provides the capability for a Source Active 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 which 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 which 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 default 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 effect 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 form that PMSI. According to section 9, 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's getting 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, PE1, has already joined the (S,G) tree, and has chosen PE2 as the upstream PE for the (S,G) tree, but this packet does not come from PE2, PE1 must discard the packet.

2. If the receiving PE, PE1, has not already joined the (S,G) tree, but is a PIM adjacency to a CE which is downstream on the (*,G) tree, the packet should be forwarded to the CE.

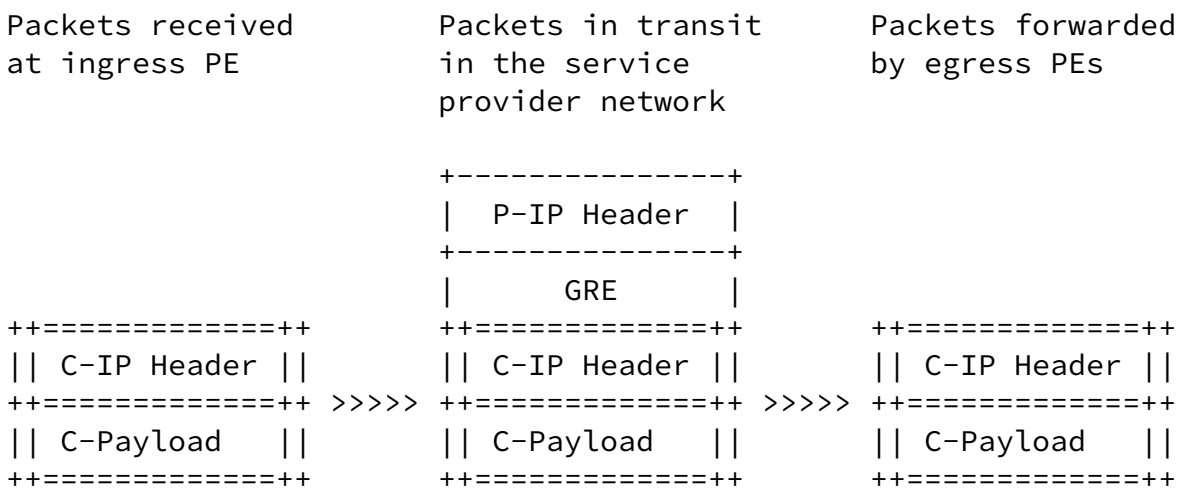
11. Encapsulations

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.

11.1. Encapsulations for Single PMSI per Tunnel

11.1.1. Encapsulation in GRE

GRE encapsulation can be used for any PMSI that is instantiated by a mesh of unicast tunnels, as well as for any PMSI that is instantiated by one or more PIM 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 must be set to 0x800.

When an encapsulated packet is transmitted by a particular PE, the source IP address in the P-IP header must be the same address as is advertised by that PE in the RPF information.

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 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.

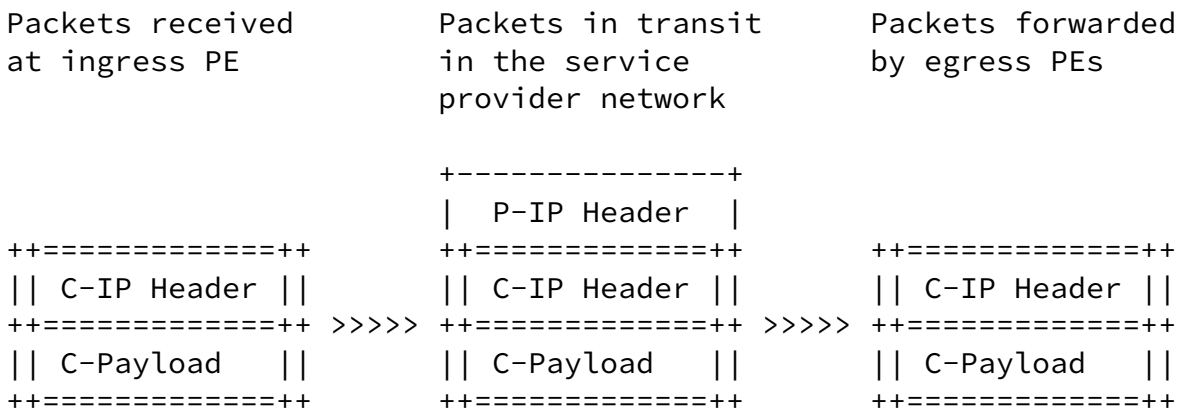
The GRE sequence number field is not needed because the transport layer services for the original application will be provided by the C-IP Header.

The use of 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.

11.1.2. Encapsulation in IP

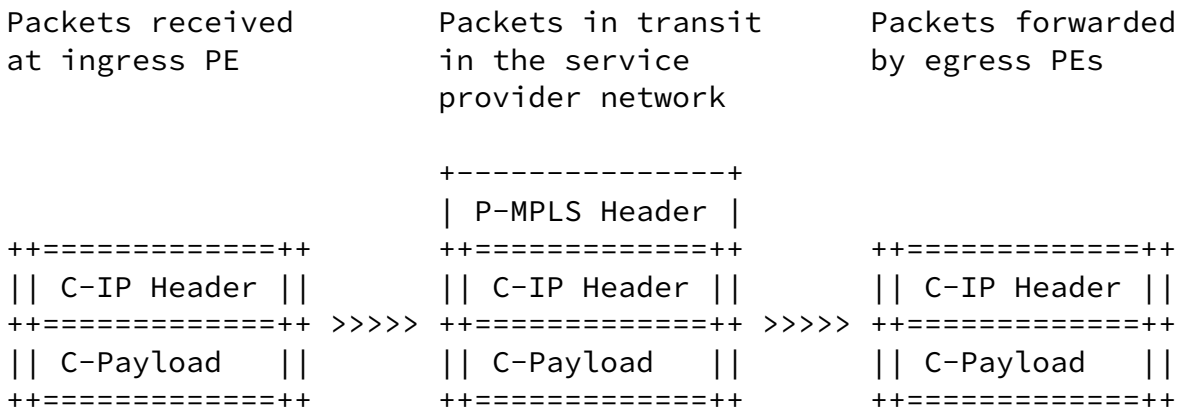
IP-in-IP [RFC1853] is also a viable option. When it is used, the IPv4 Protocol Number field is set to 4. The following diagram shows the progression of the packet as it enters and leaves the service provider network.



11.1.3. Encapsulation in MPLS

If the PMSI is instantiated as a P2MP MPLS LSP, MPLS encapsulation is used. Penultimate-hop-popping must be disabled for the P2MP MPLS LSP. If the PMSI is instantiated as an RSVP-TE P2MP LSP, additional MPLS encapsulation procedures are used, as specified in [RSVP-P2MP].

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 any additional MPLS encapsulation procedures may be used.



11.2. Encapsulations for Multiple PMSIs per Tunnel

The encapsulations for transmitting multicast data messages when there are multiple PMSIs per tunnel are based on the encapsulation for a single PMSI per 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.

11.2.1. Encapsulation in GRE

Rather than the IP-in-GRE encapsulation discussed in section 11.1.1, 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].

11.2.2. 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 MUST be set to the value identifying the payload as an MPLS unicast packet. [There is no "MPLS multicast packet" protocol number.]

11.3. Encapsulations for Unicasting PIM Control Messages

When PIM control messages are unicast, rather than being sent on an MI-PMSI, the the receiving PE needs to determine the particular MVPN whose multicast routing information is being carried in the PIM message. One method is to use a downstream-assigned MPLS label which the receiving PE has allocated for this specific purpose. The label would be distributed via BGP. This can be used with an MPLS, MPLS-

in-GRE, or MPLS-in-IP encapsulation.

A possible alternative to modify the PIM messages themselves so that they carry information which can be used to identify a particular MVPN, such as an RT.

This area is still under consideration.

11.4. General Considerations for IP and GRE Encaps

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

11.4.1. MTU

It is the responsibility of the originator of a C-packet to ensure that the packet 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 which 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 which 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 OAM

information related to this behavior, allowing the operator to properly troubleshoot the issue.

Note that if the entire path of the tunnel does not support an MTU which 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.

11.4.2. TTL

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.

11.4.3. Differentiated Services

The setting of the DS field in the delivery IP header should follow the guidelines outlined in [RFC2983]. Setting the EXP field in the delivery MPLS header should follow the guidelines in [RFC3270]. An SP may also choose to deploy any of the additional mechanisms the PE routers support.

11.4.4. 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 which it used in support of MPVN 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. Security Considerations

To be supplied.

13. IANA Considerations

To be supplied.

14. Other Authors

Sarveshwar Bandi, Yiqun Cai, Thomas Morin, Yakov Rekhter, IJsbrands Wijnands, Seisho Yasukawa

15. Other Contributors

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

16. Authors' Addresses

Rahul Aggarwal (Editor)
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: rahul@juniper.net

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
E-mail: 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

Eric C. Rosen (Editor)
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
E-mail: erosen@cisco.com

IJsbrand Wijnands
Cisco Systems, Inc.

170 Tasman Drive
San Jose, CA, 95134
E-mail: 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

17. Normative References

[MVPN-BGP], R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, C. Kodeboniya, "BGP Encodings for Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-bgp-02.txt, March 2007

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

[MPLS-MCAST-ENCAPS] T. Eckert, E. Rosen, R. Aggarwal, Y. Rekhter, "MPLS Multicast Encapsulations", draft-ietf-mpls-multicast-encaps-04.txt, April 2007

[MPLS-UPSTREAM-LABEL] R. Aggarwal, Y. Rekhter, E. Rosen, "MPLS Upstream Label Assignment and Context Specific Label Space", draft-ietf-mpls-upstream-label-02.txt, March 2007

[PIM-SM] "Protocol Independent Multicast - Sparse Mode (PIM-SM)", Fenner, Handley, Holbrook, Kouvelas, August 2006, RFC 4601

[RFC2119] "Key words for use in RFCs to Indicate Requirement

Levels.", Bradner, March 1997

[RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006

[RSVP-P2MP] R. Aggarwal, et. al., "Extensions to RSVP-TE for Point to Multipoint TE LSPs", draft-ietf-mpls-rsvp-te-p2mp-07.txt, January 2007

18. Informative References

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

[MVPN-REQ] T. Morin, Ed., "Requirements for Multicast in L3 Provider-Provisioned VPNs", RFC 4834, April 2007

[MVPN-BASE] R. Aggarwal, A. Lohiya, T. Pusateri, Y. Rekhter, "Base Specification for Multicast in MPLS/BGP VPNs", draft-raggarwa-l3vpn-2547-mvpn-00.txt

[RAGGARWA-MCAST] R. Aggarwal, et. al., "Multicast in BGP MPLS VPNs and VPLS", draft-raggarwa-l3vpn-mvpn-vpls-mcast-01.txt".

[ROSEN-8] E. Rosen, Y. Cai, I. Wijnands, "Multicast in MPLS/BGP IP VPNs", draft-rosen-vpn-mcast-08.txt

[RP-MVPN] S. Yasukawa, et. al., "BGP/MPLS IP Multicast VPNs", draft-yasukawa-l3vpn-p2mp-mcast-01.txt

[RFC1853] W. Simpson, "IP in IP Tunneling", October 1995

[RFC2784] D. Farinacci, et. al., "Generic Routing Encapsulation", March 2000

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

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

[RFC3270] F. Le Faucheur, et. al., "MPLS Support of Differentiated Services", May 2002

19. Full Copyright Statement

20. Intellectual Property