Network Working Group
Internet Draft
Intended Status: Standards Track
Expires: July 14, 2008

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

Rahul Aggarwal (Editor)
Juniper Networks

January 14, 2008

Multicast in MPLS/BGP IP VPNs

draft-ietf-l3vpn-2547bis-mcast-06.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. A multicast flow is identified by the (source, group) tuple where the source is the IP address of the sender and the group is the IP multicast group address of the destination. Scalability would be poor if the amount of state in the P routers were proportional to the number of multicast flows in the VPNs. Therefore, when supporting multicast service for a BGP/MPLS IP VPN, the optimality of the multicast routing must be traded off against the scalability of the P routers. We explain this below in more detail.

If a particular VPN is transmitting "native" multicast traffic over the backbone, we refer to it as an "MVPN". By "native" multicast traffic, we mean packets that a 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 ("Protocol Independent Multicast") [PIM-SM]. 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 BIDIR-PIM, and if the PEs attaching to that MVPN are configured with the group address corresponding to that tree, then the PEs can auto-discover each other simply by joining the tree and then multicasting PIM Hellos over the tree.

2.2.4. 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 BGP routing adjacency with other PE routers to exchange MVPN control information. BGP also provides reliable transport and uses incremental updates. Another option is the use of the currently existing, "soft state" PIM standard [PIM-SM] that uses periodic complete updates.

2.2.5. 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 BIDIR-PIM (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 BGP/MPLS IP unicast VPNs whose provider backbones contain more than one Autonomous System (AS). These are know as Inter-AS VPNs. In an Inter-AS VPN, the ASes may belong to the same provider or to different providers. This document describes how Inter-AS MVPNs can be supported for each of the unicast BGP/MPLS VPN Inter-AS options. This document also specifies a model where Inter-AS MVPN service can be offered without requiring a single SP multicast tree to span multiple ASes. In this model, an inter-AS multicast tree consists of a number of "segments", one per AS, 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. Optionally Eliminating Shared Tree State

The document also discusses some options and protocol extensions which can be used to eliminate the need for the PE routers to distribute to each other the (*, G) and (*, G, RPT-bit) states when there are PIM Sparse Mode multicast groups in the VPNs.

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 (BIDIR-PIM)

- 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. The use of mLDP on the PE-CE interface is described in [MVPN-BGP]. The use of BGP on the PE-CE interface is not 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.

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 or tunnels that instantiate an MI-PMSI may actually carry the data of more than one PMSI.)

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

- "Unidirectional Inclusive" PMSI (UI-PMSI)

  A Unidirectional Inclusive PMSI is one 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.

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

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

- PIM

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

  PIM-SSM, BIDIR-PIM, 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 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.

- MLDP

  A PMSI may be instantiated as one or more mLDP Point-to-Multipoint (P2MP) LSPs, or as an mLDP Multipoint-to-MultiPoint(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 MP2MP 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 given provider backbone. It is however expected that any tunnel technique that can be 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 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 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. These characteristics and 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 3.4.1.1, 3.4.1.2, and 5.2).

3.3.3. MVPNs That Do Not Use MI-PMSIs

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

It is also possible to send all the multicast data on a set of S-PMSIs, 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.

3.4. PE-PE Transmission of C-Multicast Routing

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

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

3.4.1. PIM Peering

3.4.1.1. Full Per-MVPN PIM Peering Across 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 procedure is fully specified in section 5.2.

3.4.1.2. 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. If and when "refresh reduction" procedures are specified for PIM, it may be useful to incorporate them, so as to make the lightweight PIM peering procedures even more lightweight.

Lightweight PIM peering is not specified in this document.

3.4.1.3. Unicasting of PIM C-Join/Prune Messages

PIM does not require that the C-Join/Prune messages 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.

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

3.4.2. Using BGP to Carry C-Multicast Routing

It is possible to use BGP to carry C-multicast routing information from PE to PE, dispensing entirely with the transmission of C-Join/Prune messages from PE to PE. This is specified in section 5.3. Inter-AS procedures are described in section 8.

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". In particular, it uses two kinds of "A-D routes", the "Intra-AS A-D Route" and the "Inter-AS A-D Route". (There are also additional kinds of A-D routes, such as the Source Active A-D routes which are used for purposes that go beyond auto-discovery. These are discussed in subsequent sections.)

The Inter-AS A-D Route is used only when segmented inter-AS tunnels are used, as specified in section 8.

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

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

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

- The 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 either MI-PMSI is to be used for the MVPN, or UI-PMSI is to be used for the MVPN on the PE that originates the intra-AS A-D route. It contains the following information:

  * whether the MI-PMSI is instantiated by

    + A BIDIR-PIM 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 tunnel to instantiate the I-PMSI, a unique identifier for the tunnel used to instantiate the I-PMSI. This identifier depends on the tunnel technology used.

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 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. This is obviously the case when the tunnel is constructed by a receiver-initiated join technique such as PIM or mLDP. It is also the case when the tunnel is an RSVP-TE P2MP LSP as the tunnel identifier can be constructed without the head end learning the identities of the other PEs.

In other cases, a tunnel cannot be identified until the PE has discovered one or more of the other PEs. 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.

All the PEs attaching to a given MVPN must be configured with information specifying the encapsulation to use.

* 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.

Sometimes it is necessary for one PE to advertise an upstream-assigned MPLS label that identifies another PE. Under certain circumstances to be discussed later, a PE which is the root of a multicast P-tunnel will bind an MPLS label value to one or more of the PEs that belong to the P-tunnel, and will distribute these label bindings using A-D routes. The precise details of this label distribution will be included in the next revision of this document. We will refer to these as "PE Labels". A packet traveling on the P-tunnel may carry one of these labels as an indication that the PE corresponding to that label is special. See section 11.3 for more details.

5. PE-PE Transmission of C-Multicast Routing

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

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

- Full PIM Peering.

  This procedure is fully specified herein.

- Use of BGP to distribute C-multicast routing

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

Those aspect of the procedures which apply to both of the above are also specified fully herein.

Specification of other procedures is for future study.

5.1. Selecting the Upstream Multicast Hop (UMH)

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

5.1.1. Eligible Routes for UMH Selection

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

However, in some cases, the CEs may be distributing to the PEs a special set of routes that are to be used exclusively for the purpose of upstream multicast hop selection, and not used for unicast routing at all. For example, when BGP is the CE-PE unicast routing protocol, the CEs may be using SAFI 2 to distribute a special set of routes that are to be used for, and only for, upstream multicast hop selection. When OSPF is the CE-PE routing protocol, the CE may use an MT-ID of 1 to distribute a special set of routes that are to be used for, and only for, upstream multicast hop selection . When a CE uses one of these mechanisms to distribute to a PE a special set of routes to be used exclusively for upstream multicast hop selection, these routes are distributed among the PEs using SAFI 129, as described in [MVPN-BGP].

Whether the routes used for upstream multicast hop selection are (a) the "ordinary" unicast routes or (b) a special set of routes that are used exclusively for upstream multicast hop selection, is a matter of policy. How that policy is chosen, deployed, or implemented is outside the scope of this document. In the following, we will simply refer to the set of routes that are used for upstream multicast hop selection, the "Eligible UMH routes", with no presumptions about the policy by which this set of routes was chosen.

5.1.2. Information Carried by Eligible UMH Routes

Every route which is eligible for UMH selection MUST carry a VRF Route Import Extended Community [MVPN-BGP]. This attribute identifies the PE that originated the route.

If BGP is used for carrying C-multicast routes, OR if "Segmented Inter-AS Tunnels" (see section 8.2) are used, then every UMH route MUST also carry a Source AS Extended Community [MVPN-BGP].

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

5.1.3. Selecting the Upstream PE

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

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

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

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 upstream multicast hop selection, and (c) have exactly the same IP prefix (not necessarily the same RD) as the installed UMH route.

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

This results in a set of pairs of <route, upstream PE, upstream RD>.

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

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

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

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

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

2. The following hash is performed:

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

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

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

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

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

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

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

5.1.4. Selecting the Upstream Multicast Hop

In certain cases, the selected upstream multicast hop is the same as the selected upstream PE. In other cases, the selected upstream multicast hop is the ASBR which is the "BGP next hop" of the Selected UMH Route.

If the selected upstream PE is in the local AS, then the selected upstream PE is also the selected upstream multicast hop. This is the case if any of the following conditions holds:

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

- The selected UMH route does not have a Source AS Extended Community, but the route's BGP next hop is the same as the upstream PE.

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

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

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 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.

5.2.1. 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.

As specified in section 5.1.2, when a PE distributes VPN-IP routes which are eligible for use as UMH routes, the PE MUST include a VRF Route Import Extended Community with each route. For a given MVPN, a single such IP address MUST be used, and that same IP address MUST be used as the source address in all PIM control packets for that MVPN.

5.2.2. 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.

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

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

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

5.2.3. Backwards Compatibility

There are older implementations which do not use the VRF Route Import Extended Community or any explicit mechanism for carrying information to identify the originating PE of a selected UMH route.

For backwards compatibility, when the selected UMH route does not have any such mechanism, the IP address from the "BGP Next Hop" field of the selected UMH route will be used as the selected UMH address, and will be treated as the address of the upstream PE. There is no selected upstream RD in this case. However, use of this backwards compatibility technique presupposes that:

- The PE which originated the selected UMH route placed the same IP address in the BGP Next Hop field that it is using as the source address of the PE-PE PIM control packets for this MVPN.

- The MVPN is not an Inter-AS MVPN that uses option b from section 10 of [RFC4364].

Should either of these conditions fail, interoperability with the older implementations will not be achieved.

5.3. Use of BGP for Carrying C-Multicast Routing

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

5.3.1. Sending BGP Updates

The MCAST-VPN address family is used for this purpose. MCAST-VPN routes used for the purpose of carrying C-multicast routing information are distinguished from those used for the purpose of carrying auto-discovery information by means of a "route type" field 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. The RD is required in order to uniquely identify the <C-Source, C-Group> when different MVPNs have overlapping address spaces.

- The C-Group address.

- The C-Source address.

  This field is omitted if the route type is "shared tree join". In the case of a shared tree join, the C-source is a C-RP. The address of the C-RP corresponding to the C-group address is presumed to be already known (or automatically determinable) be the other PEs, though means that are outside the scope of this specification.

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

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

There is no C-multicast route corresponding to the PIM function of pruning a source off the shared tree when a PE switches from a <C-*, C-G> tree to a <C-S, C-G> tree. Section 9 of this document specifies a mandatory procedure that ensures that if any PE joins a <C-S, C-G> source tree, all other PEs that have joined or will join the <C-*, C-G> shared tree will also join the <C-S, C-G> source tree. This eliminates the need for a C-multicast route that prunes C-S off the <C-*, C-G> shared tree when switching from <C-*, C-G> to <C-S, C-G> tree.

5.3.2. Explicit Tracking

Note that the upstream multicast hop is NOT part of the NLRI in the C-multicast BGP routes. This means that if several PEs join the same C-tree, the BGP routes they distribute to do so are regarded by BGP as comparable routes, and only one will be installed. If a route reflector is being used, this further means that the PE 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

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 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". 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. For more details on Leaf A-D routes please refer to [MVPN-BGP].

6.3.2. Aggregation Methodology

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

It is the intention of this document that the control procedures will always result in all the PEs of an MVPN 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 support of aggregation for shared trees and MP2MP trees is discussed in section 6.6.

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. 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.

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.

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].

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.

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 7.1. 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, BIDIR-PIM, 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, BIDIR-PIM, RSVP-TE P2MP LSPs, mLDP P2MP LSPs, and mLDP MP2MP LSPs.. Aggregation support for bidirectional P-trees (i.e., BIDIR-PIM 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:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |            Length             |     Value     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |           Length              |    Reserved   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           C-source                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           C-group                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           P-group                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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.

d) The C-Group address.

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

   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.

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.

8.1.1. Inter-AS MVPN Auto-Discovery

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

8.1.2. Inter-AS MVPN Routing Information Exchange

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

When PIM peering is used to distribute the C-multicast routing information, a PE that sends C-PIM Join/Prune messages for a particular C-(S,G) must be able to identify the PE which is its PIM adjacency on the path to S. This is the "selected upstream PE" described in section 5.1.

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

8.1.3. Inter-AS P-Tunnels

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

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

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

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

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

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

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

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

If either of these two conditions obtains, then "ordinary" PIM Join/Prune messages cannot be routed to the upstream PE. Thus the following information needs to be added to the PIM Join/Prune messages: a "Proxy Address", which contains the address of the next ASBR on the path to the upstream PE. When the PIM Join/Prune arrives at the ASBR which is identified by the "proxy address", that ASBR must change the proxy address to identify the next hop ASBR.

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

However, even if the ASBRs do not have routes to the upstream PE, the procedures of this draft ensure that they will have A-D routes that lead to the upstream PE. If non-segmented inter-AS MVPNs are being used, the ASBRs (and PEs) will have Intra-AS A-D routes which have been distributed inter-AS.

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

If the only intra-AS A-D route for a given MVPN is the "Intra-AS I-PMSI Route", the PIM Join/Prunes will be routed along that. However, if the PIM Join/Prune message is for a particular P-group address, and there is an "Intra-AS S-PMSI Route" specifying that particular P-group address as the P-tunnel for a particular S-PMSI, then the PIM Join/Prunes MUST be routed along the path determined by those intra-AS A-D routes.

The next revision of this document will provide the following details:

- encoding of the proxy address in the PIM message (the PIM Join Attribute [PIM-ATTRIB] will be used)

- encoding of any other information which may be needed in order to enable the correct intra-AS route to be chosen.

Support for non-segmented inter-AS trees using BIDIR-PIM is for further study.

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 in the following way:

- Consider a node which is the root of an an intra-AS segment of an inter-AS tunnel. An inter-AS leaf autodiscovery route is used to inform such a node 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 tunnel 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 and PEs 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 own.

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

Consider the case of an egress PE that receives packets of a customer multicast stream (C-S, C-G) over a non-aggregated S-PMSI. The procedures described so far will never cause the PE to receive duplicate copies of any packet in that stream. It is possible that the (C-S, C-G) stream is carried in more than one S-PMSI; this may happen when the site that contains C-S is multihomed to more than one PE. However, a PE that needs to receive (C-S, C-G) packets only joins one of these S-PMSIs, and so only receives one copy of each packet.

However, if the data packets of stream (C-S, C-G) are carried in either an I-PMSI or in an aggregated S-PMSI, then it the procedures specified so far make it possible for an egress PE to receive more than one copy of each data packet. In this section, we define additional procedures to that an MVPN customer sees no multicast data packet duplication.

This section covers the situation where the customer multicast tree is unidirectional, i.e. with the C-G is either a "Sparse Mode" or a "Single Source Mode" group. The case where the customer multicast tree is bidirectional (the C-G is a BIDIR-PIM group) is considered separately in section 12.

The first case when an egress PE may receive duplicate multicast data packets, is the case where both (a) an MVPN site that contains C-S or C-RP is multihomed to more than one PE, and (b) either an I-PMSI, or an aggregated S-PMSI is used for carrying the packets originated by

C-S. In this case, an egress PE may receive one copy of the packet from each PE to which the site is homed.

The second case when an egress PE may receive duplicate multicast data packets is when all of the following is true: (a) the IP destination address of the customer packet is a C-G that is operating in ASM mode, and whose C-multicast tree is set up using PIM-SM, (b) an MI-PMSI is used for carrying the packets, and (c) a router or a CE in a site connected to the egress PE switches from the C-RP tree to C-S tree. In this case, it is possible to get one copy of a given packet from the ingress PE attached to the C-RP's site, and one from the ingress PE attached to the C-S's site.

9.1. Multihomed C-S or C-RP

In the first case for a given <C-S, C-G> an egress 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 and forward the packet. Otherwise, PE1 will drop the packet; this means that the PE will see a duplicate, but the duplicate will not get forwarded. (But see section 10 for an exception case where PE1 will accept a packet even if it is from an unexpected upstream PE.)

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

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

If an I-PMSI is used for carrying the packets, the I-PMSI spans multiple ASes, and the I-PMSI is realized via segmented inter-AS tunnels, if C-S or C-RP is multi-homed to different PEs, as long as each such PE is in a different AS, the egress PE can detect duplicate traffic as such duplicate traffic will arrive on a different (inter-AS) tunnel. Specifically, if the PE was expecting the traffic on an

particular inter-AS tunnel, duplicate traffic will arrive either on an intra-AS tunnel [this is not an intra-AS tunnel segment, of an inter-AS tunnel], or on some other inter-AS tunnel. Therefore, to detect duplicates 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 Inter-AS 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). The PE should discard, as duplicated, all other multicast traffic originated by C-S/C-RP, but received on any other tunnel.

9.1.1. Single forwarder PE selection

When for a given MVPN (a) MI-PMSI is used for carrying multicast data packets, (b) C-S or C-RP is multi-homed to different PEs, and (c) at least two of such PEs are in the same AS, then depending on the tunneling technology used by the MI-PMSI it may not always be possible for the egress PE to determine the upstream PE. Therefore, when this determination may not be possible procedures are needed to ensure that packets are received on an MI-PMSI at an egress PE only from a single upstream PE. Furthermore, even if the determination is possible, it may be preferable to send only one copy of each packet to each egress PE, rather than sending multiple copies and having the egress PE discard all but one.

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

This procedure is optional whenever the P-tunnel technology that is being used to carry the multicast stream in question allows the egress PEs to determine the identity of the ingress PE. This procedure is mandatory if the P-tunnel technology does not make this determination possible.

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 are on either the C-S or C-RP tree. If some PEs are on the C-S tree and some on the C-RP tree, however, packet duplication is still possible. This is discussed in the next section.

9.2. Switching from the C-RP tree to C-S tree

If some PEs are on the C-S tree and some on the R-RP tree then a 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.

When for a given MVPN (a) MI-PMSI is used for carrying multicast data packets, (b) C-S and C-RP are connected to PEs within the same AS, and (c) the MI-PMSI tunneling technology in use 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 for C-S or C-RP is not sufficient to prevent packet duplication.

The reason is that a single tunnel used by MI-PMSI 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 (C-S, C-G) packets from the shared tree, then two copies of data packet will travel on the tunnel, and since due to the choice of the tunneling technology the egress PEs have no way to identify the ingress PE, 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 a Source Active auto-discovery route (see [MVPN-BGP] section 4.5) as specified below. The route is 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 as a result of receiving a new Source Active auto-discovery route a PE updates its VRF with the route, the PE MUST check if the newly received route matches any <C-*, C-G> entries. If (a) there is a matching entry, (b) the PE does not have (C-S, C-G) state in its MVPN-TIB for (C-S, C-G) carried in the route, and (c) the received route is selected as the

best (using the BGP route selection procedures), then the PE sets up its forwarding path to receive (C-S, C-G) traffic from the tunnel the originator of the selected Source Active auto-discovery route uses for sending (C-S, C-G). This procedures forces all the PEs (in all ASes) to switch from the C-RP tree to the C-S tree for <C-S, C-G>.

(Additional uses of the Source Active A-D route are discussed 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. (This will be the case if the incoming interface for the (C-*, C-G) tree is one of the VRF interfaces.) However, before doing this, it SHOULD run a timer to help ensure that the source is not pruned from the shared tree until all PEs have had time to receive the Source Active route.

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.

It is worth nothing that if a PE joins a source tree as a result of this procedure, the UMH is not necessarily the same as it would be if the PE had joined the source tree as a result of receiving a PIM Join for the same source tree from a directly attached CE.

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

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

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

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

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

10.1. Co-locating C-RPs on a PE

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

"In 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 section describes how anycast-RP can be used for achieving this. This is described below.

10.1.1. Initial Configuration

For a particular MVPN, at least one or more PEs that have sites in that MVPN, act as an RP for the sites of that MVPN connected to these PEs. Within each MVPN all 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.

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 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 from that PMSI. According to section 9, if PE1 is not the PE that PE2 expects to be transmitting (S,G) packets, then PE2 must discard the packet. If an MSDP-encapsulated data packet is transmitted on a PMSI as specified above, this rule from section 9 would likely result in the packet'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, PE2, has already joined the (S,G) tree, and has chosen PE1 as the upstream PE for the (S,G) tree, but this packet does not come from PE1, PE2 must discard the packet.

2. If the receiving PE, PE2, has not already joined the (S,G) tree, but is a PIM adjacency to a CE 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.

```
   Packets received        Packets in transit      Packets forwarded
   at ingress PE           in the service          by egress PEs
                           provider network

                           +---------------+
                           |  P-IP Header  |
                           +---------------+
                           |      GRE      |
   ++=============++       ++=============++       ++=============++
   || C-IP Header ||       || C-IP Header ||       || C-IP Header ||
   ++=============++ >>>>> ++=============++ >>>>> ++=============++
   || C-Payload   ||       || C-Payload   ||       || C-Payload   ||
   ++=============++       ++=============++       ++=============++
```

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 that the PE uses to identify itself in the VRF Route Import Extended Communities that it attaches to any of VPN-IP routes eligible for UMH determination that it advertises via BGP (see section 5.1).

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

If the PMSI is instantiated by unicast 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.

```
Packets received        Packets in transit      Packets forwarded
at ingress PE           in the service          by egress PEs
                        provider network

                        +---------------+
                        |  P-IP Header  |
++=============++       ++=============++       ++=============++
|| C-IP Header ||       || C-IP Header ||       || C-IP Header ||
++=============++ >>>>> ++=============++ >>>>> ++=============++
|| C-Payload   ||       || C-Payload   ||       || C-Payload   ||
++=============++       ++=============++       ++=============++
```

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

11.1.3. Encapsulation in MPLS

If the PMSI is instantiated as a P2MP MPLS LSP or MP2MP 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 appropriate additional MPLS encapsulation procedures may be used.

```
   Packets received        Packets in transit      Packets forwarded
   at ingress PE           in the service          by egress PEs
                           provider network

                           +---------------+
                           | P-MPLS Header |
   ++=============++       ++=============++       ++=============++
   || C-IP Header ||       || C-IP Header ||       || C-IP Header ||
   ++=============++ >>>>> ++=============++ >>>>> ++=============++
   || C-Payload   ||       || C-Payload   ||       || C-Payload   ||
   ++=============++       ++=============++       ++=============++
```

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 Identifying a Distinguished PE

11.3.1. For MP2MP LSP P-tunnels

As discussed in section 9, if a multicast data packet belongs to a Sparse Mode or Single Source Mode multicast group, it is highly desirable for the PE that receives the packet from a PMSI to be able to determine the identity of the PE that transmitted the data packet onto the PMSI. The encapsulations of the previous sections all provide this information, except in one case. If a PMSI is being instantiated by a MP2MP LSP, then the encapsulations discussed so far do not allow one to determine the identity of the PE that transmitted the packet onto the PMSI.

Therefore, when a packet that belongs to a Sparse Mode or Single Source Mode multicast group is traveling on a MP2MP LSP P-tunnel, it MUST carry, as its second label, a label which has been bound to the packet's ingress PE. This label is an upstream-assigned label that the LSP's root node has bound to the ingress PE and has distributed via an A-D Route (see section 4; precise details of this distribution procedure will be included in the next revision of this document). This label will appear immediately beneath the labels that are discussed in sections 11.1.3 and 11.2.

11.3.2. For Support of PIM-BIDIR C-Groups

As will be discussed in section 12, when a packet belongs to a PIM-BIDIR multicast group, the set of PEs of that packet's VPN can be partitioned into a number of subsets, where exactly one PE in each partition is the upstream PE for that partition. When such packets are transmitted on a PMSI, then unless the procedures of section 12.2.3 are being used, it is necessary for the packet to carry information identifying a particular partition. This is done by having the packet carry the PE label corresponding to the upstream PE of one partition. For a particular P-tunnel, this label will have been advertised by the node which is the root of that P-tunnel. (Details of the procedure by which the PE labels are advertised will be included in the next revision of this document.)

This label needs to be used whenever a packet belongs to a PIM-BIDIR C-group, no matter what encapsulation is used by the P-tunnel. Hence the encapsulations of section 11.2 MUST be used. If the tunnel contains only one PMSI, the PE label replaces the label discussed in section 11.2 If the tunnel contains multiple PMSIs, the PE label follows the label discussed in section 11.2

11.4. Encapsulations for Unicasting PIM Control Messages

When PIM control messages are unicast, rather than being sent on an MI-PMSI, then 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.

11.5. General Considerations for IP and GRE Encaps

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

11.5.1. MTU

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

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

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

- If the DF bit is set in the IP header of a C-packet 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.

11.5.2. 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.

11.5.3. Avoiding Conflict with Internet Multicast

If the SP is providing Internet multicast, distinct from its VPN multicast services, and using PIM based P-multicast trees, it must ensure that the group P-addresses 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.

11.6. 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 additional Differentiated Services mechanisms that the PE routers support for the encapsulation in use. Note that the type of encapsulation determines the set of Differentiated Services mechanisms that may be deployed.

12. Support for PIM-BIDIR C-Groups

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

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

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

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

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

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

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

12.1. The VPN Backbone Becomes the RPL

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

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

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

12.1.1. Control Plane

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

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

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

For a given MVPN a PE connected to an MVPN-RPL of that MVPN should send multicast data (C-S,C-G) on the MVPN-RPL only if at least one other PE connected to the MVPN-RPL has a downstream multicast state for C-G. In the context of MVPN this is accomplished by requring a PE that has a downstream state for a particular C-G of a particular VRF present on the PE to originate a C-multicast route for (*, C-G). The RD of this route should be the same as the RD associated with the VRF. The RT(s) carried by the route should be the same as the one(s) used for VPN-IPv4 routes. This route will be distributed to all the PEs of the MVPN.

12.1.2. Data Plane

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

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

12.2. Partitioned Sets of PEs

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

12.2.1. Partitions

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

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

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

Consider packet P, and let PE1 be its ingress PE. PE1 will send the

packet on a PMSI so that it reaches the other PEs that need to receive it. This is done by encapsulating the packet and sending it on a P-tunnel. If the original packet is part of a PIM-BIDIR group (its ingress PE determines this from the packet's destination address C-G), and if the VPN backbone is not the RPL, then the encapsulation MUST carry information that can be used to identify the partition to which the ingress PE belongs.

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

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

12.2.2. Using PE Labels

If a given P-tunnel is to be used to carry packets belonging to a bidirectional C-group, then, EXCEPT for the case described in section 12.2.3 the packets that travel on that P-tunnel MUST carry a PE label (defined in section 4), using the encapsulation discussed in section 11.3.

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

Recall that the PE labels are upstream-assigned labels that are assigned and advertised by the node which is at the root of the P-tunnel. (Procedures for PE label assignment when the P-tunnel is not a multicast tree will be given is later revisions of this document.)

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

Note that this procedure does not require the root of a P-tunnel to assign a PE label for every PE that belongs to the tunnel, but only for those PEs that might become the upstream PEs of some partition.

12.2.3. Mesh of MP2MP P-Tunnels

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

13. Security Considerations

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

When GRE encapsulation is used, the security considerations of [MPLS-IP] are also relevant. The security considerations of [RFC4797] are also relevant as it discusses implications on packet spoofing in the context of 2547 VPNs.

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

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

If one uses the UDP-based protocol for switching to S-PMSI (as specified in Section 7.2.1), then by default each PE router MUST install packet filters that would result in discarding all UDP packets with the destination port 3232 that the PE router receives from the CE routers connected to the PE router.

The various procedures for P-tunnel construction have security issues that are specific to the way in which the P-tunnels are used in this document. When P-tunnels are constructed via such techniques as as PIM, mLDP, or RSVP-TE, it is important for each P or PE router receiving a control message to be sure that the control message comes from another P or PE router, not from a CE router. This should not be a problem, because mLDP or PIM or RSVP-TE control messages from CE routers will never be interpreted as referring to P-tunnels.

An ASBR may receive, from one SP's domain, an mLDP, PIM, or RSVP-TE control message that attempts to extend a multicast distribution tree from one SP's domain into another SP's domain. The ASBR should not allow this unless explicitly configured to do so.

14. IANA Considerations

Section 7.2.1.1 defines the "S-PMSI Join Message", which is carried in a UDP datagram whose port number is 3232. This port number is already assigned by IANA to "MDT port". IANA should now have that assignment reference this document.

IANA should create a registry for the "S-PMSI Join Message Type Field". The value 1 should be registered with a reference to this document. The description should read "PIM IPv4 S-PMSI (unaggregated)".

15. Other Authors

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

16. 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.

17. 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

18. Normative References

[MLDP] I. Minei, K., Kompella, I. Wijnands, B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-03, July 2007

[MPLS-HDR] E. Rosen, et. al., "MPLS Label Stack Encoding", RFC 3032, January 2001

[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-06.txt, July 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

[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-04.txt, November 2007

[PIM-ATTRIB], A. Boers, IJ. Wijnands, E. Rosen, "Format for Using TLVs in PIM Messages", draft-ietf-pim-join-attributes-03, May 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, D. Papadimitriou, S. Yasukawa, et. al., "Extensions to RSVP-TE for Point-to-Multipoint TE LSPs", RFC 4875,

May 2007

19. 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

[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

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

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

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

20. Full Copyright Statement

21. Intellectual Property