Network Working Group
Internet Draft
Expiration Date: November 2005

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

Rahul Aggarwal (Editor)
Juniper Networks

May 2005

Multicast in MPLS/BGP IP VPNs

draft-ietf-l3vpn-2547bis-mcast-00.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

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

[RFC2547bis] specifies a set of procedures which must be implemented for a Service Provider (SP) to provide an IP unicast VPN service. However [RFC2547bis] does not provide a way for a SP to offer IP multicast VPN (MVPN) service. In particular it does not provide mechanisms for IP multicast data or control traffic to travel from one VPN site to another. This document specifies the protocols and procedures that enable a SP to provide multicast service in a VPN.

2.1. Optimality vs Scalability

In a "BGP/MPLS VPN" [RFC2547bis], 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"). VPN routing information is maintained only by the PEs participating in the VPN service. This allows the SP to increase the number of VPNs it supports, without requiring additional state to be kept in the P routers, and it still allows the site-to site routing to be optimal.

However, when supporting multicast routing in a BGP/MPLS VPN, the optimality of the multicast must be traded off against scalability.

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.

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.

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 a SP to provide multicast VPN service without requiring an unbounded amount of state in the P routers. The amount of state is traded off against the optimality of the multicast routing. The procedures provide flexibility so that a given SP can make his own tradeoffs between scalability and optimality. The procedures also allow for some multicast groups in some VPNs to receive optimal routing, while others do not.

One supported option to carry MVPN data traffic is to setup 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 makes no attempt to optimize the multicast routing.

Another supported option is to use 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". With this option, the P routers do not maintain state on a per-multicast-group basis, but only on a per-set-of-MVPNs basis. These multicast distribution trees can be set up to carry the traffic of a single MVPN, or to carry the traffic of multiple MVPNs. The term "Aggregate Inclusive Tree" will be used to specifically refer to such a tree when used to carry the traffic of multiple VPNs. The tree will include every PE that is a member of any of the MVPNs that are using the tree. This enables the SP to place a bound on the amount of multicast routing state which the P routers must have. 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 by which a single multicast distribution tree in the backbone can 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. So 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.

2.2. Overview

In BGP MPLS VPNs [RFC2547bis], each CE 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 PIM adjacency of a PE router, but CE routers at different sites do not become PIM adjacencies of each other. A CE router exchanges "ordinary" PIM control messages with the PE router to which it is attached. The set of PE routers attaching to a given MVPN then exchange MVPN control information with each other.

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. One of the options in this document for achieving this is the use of BGP for discovering

and maintaining MVPN membership. In this solution 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 [2547bis]) which is configured to contain the multicast routing information of that MVPN. This discovery option can be used with any multicast transport technology.

This document also describes an option for discovering MVPN members that relies on the presence of a PIM based Inclusive Tree for each PE router that belongs to the MVPN.

The BGP/MPLS VPNs [RFC2547bis] specification requires a PE to maintain at most one BGP peering with every other PE in the network. This peering is used to exchange VPN routing information. The use of Route Reflectors further reduces the number of BGP adjacencies maintained by a PE to exchange VPN routing information with other PEs. This document describes various options for exchanging MVPN control information between PE routers based on the use of PIM or BGP. These options have different overheads with respect to the number of routing adjacencies that a PE router needs to maintain to exchange MVPN control information with other PE routers. Some of these options allow the retention of the unicast BGP/MPLS VPN model letting a PE maintain at most one routing adjacency with other PE routers to exchange MVPN control information.

The option of using PIM for exchanging MVPN control information along with BGP to discover the MVPN members closely mimics the unicast BGP/MPLS PE-PE routing adjacency model. The option of using BGP for exchanging MVPN control information comes even closer to the unicast VPN model as it allows the use of Route Reflectors.

The solution in [RFC2547bis] uses BGP to exchange VPN routing information between PE routers and BGP provides a reliable transport. This document describes various solutions for providing a reliable transport for exchanging MVPN control information. One option is the use of PIM with reliability extensions and the other is the use of BGP. The use of the currently existing, "soft state" PIM standard [PIM-SM], is also supported.

Like [RFC2547bis], this document decouples procedures for exchanging routing information from the procedures for transmitting data traffic. Hence a variety of transport technologies may be used in the backbone, including unicast PE-PE tunnels using MPLS or IP/GRE encapsulation, PIM trees (created by PIM-SSM, PIM-SM, or PIM-Bidir), and point-to-multipoint LSPs created by RSVP-TE. (Techniques for aggregating the traffic of multiple MVPNs onto a single PIM-bidir tree are for further study.) Selective trees are always set up with

PIM-SSM or with RSVP-TE.

In order to aggregate traffic from multiple MVPNs onto a single multicast distribution tree, the root of the tree must be able to discover the MVPN membership of all the PEs and/or the set of multicast groups in which each PE has receivers. This document describes the procedures for achieving this. Aggregation also requires a mechanism which enables the egresses of the tree to demultiplex the multicast traffic received over the tree. This document specifies a mechanism whereby upstream label allocation [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 encapsulation for supporting the various SP multicast transport options.

[RFC2547bis] describes different options for supporting Inter-AS BGP/MPLS unicast VPNs. This document describes how Inter-AS MVPNs can be supported for each of the unicast BGP/MPLS VPN Inter-AS options. This document introduces a model where Inter-AS MVPN service can be offered without requiring a SP multicast tree to span multiple ASs. Each AS can have a different multicast tree, possibly with a different MVPN transport technology in each AS. It is also possible to support Inter-AS MVPNs with source trees that extend across AS boundaries.

The document also discusses different deployment models for MVPNs and any protocol extensions that are required for supporting the specific models.

This document assumes that when SP multicast trees are used, traffic for a particular multicast group can be transmitted by a 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.

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", defined in [RFC2547bis], 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. Rather, we consider multicast C-instances only for the following multicast routing protocols:

- PIM Sparse Mode (PIM-SM)

- PIM Single Source Mode (PIM-SSM)

- PIM Bidirectional Mode (PIM-Bidir)

- PIM Dense Mode (PIM-DM)

As the document only support 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 an "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 are making a clear distinction between the multicast service (the PMSI) and its instantiation. This allows us to separate the discussion of different services from the discussion of different instantiations of each service. The term "tunnel" is used to refer only to the transport mechanism that instantiates a service.

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

3.2.1. Inclusive and Selective PMSIs

We will distinguish between three different kinds of PMSI:

- "Multidirectional Inclusive" PMSI (MI-PMSI)

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

  There is at most one MI-PMSI per MVPN. (Though the tunnel which instantiates an MI-PMSI may actually carry the data of more than one PMSI.)

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

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

- "Unidirectional Inclusive" PMSI (UI-PMSI)

  A Unidirectional Inclusive PMSI is one which enables a particular PE, attached to a particular MVPN, to transmit a message such that it will be received by all the other PEs attaching to that MVPN. There is at most one UI-PMSI per PE per MVPN, though the "tunnel" which instantiates a UI-PMSI may in fact carry the data of more than one PMSI.

- "Selective" PMSI (S-PMSI).

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

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

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

3.2.2. Tunnels Instantiating PMSIs

A number of different tunneling techniques can be used to instantiate PMSIs. Among these are:

- PIM Trees.

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

  PIM-SSM, PIM-Bidir, or PIM-SM can be used to create P-trees.

  A single MI-PMSI can be instantiated by a single PIM-SM shared tree or a PIM-Bidir tree. Using PIM-SSM to instantiate an MI-PMSI requires one P-tree for each of the MI-PMSI's PEs. This P-tree may be shared across multiple MVPNs.

  Selective PMSIs are most naturally created with PIM-SSM, since by definition only one PE is the source of the multicast data on a Selective PMSI.

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

- RSVP-TE Point-to-Multipoint LSP.

  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.

- Meshed 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.

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. Further discussion of the applicability of these methods is outside the scope of this document.

3.3. Use of PMSIs for Carrying Multicast Data

We presuppose that each PE supporting a particular MVPN has a way of discovering:

- The set of other PEs supporting that same MVPN

- For each of the other PEs in the MVPN, whether they require a MI-PMSI.

- The set of UI-PMSIs that the MVPN requires, if any; each UI-PMSI would be identified by its transmitting PE

- The tunneling technique, announced by each PE, to be used for instantiating the I-PMSIs

- A "tunnel identifier", which identifies the particular tunnel or tunnels to be used to instantiate the PMSIs; this will be specific to the particular tunneling technology used;

- Whether a PE supports Aggregation for that MVPN.

- Any demultiplexing information which is needed in order to enable a specified PMSI to be carried in the specified tunnel (this is needed if a given tunnel is capable of carrying multiple PMSIs).

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 a separate set of BGP-based tunnel binding procedures. 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.

If an MVPN requires the use of one or more I-PMSIs, then I-PMSIs for that MVPN will be "created" as soon as the necessary information has

been discovered. Creating a PMSI means creating the tunnel which carries it (unless that tunnel already exists), as well as binding the PMSI to the tunnel.

3.3.1. MVPNs that Use 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. The MI-PMSI for that MVPN is then used as the default method of transmitting multicast data packets for that MVPN. In effect, all the multicast streams for the MVPN are, by default, aggregated onto the MI-MVPN.

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

3.3.2. MVPNs That Do Not Use MI-PMSIs

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

It is also possible to send all the multicast data on an S-PMSI, omitting any usage of I-PMSIs. This prevents PEs from receiving data which they don't need, but requires that the transmitting PE knows which PEs need to receive which multicast streams.

4. BGP-Based Autodiscovery of MVPN Membership

BGP-based autodiscovery is done by means of a new address family. The encoding details are to specified in a separate document. Any PE which attaches to an MVPN must issue a BGP update message containing NLRI for this address family, as well as specific attributes. In this document, we specify the information which must be contained in these BGP updates.

Each such BGP update must contain the following information:

- IPv4 address of the originating PE

- An RD which can be prepended to that IPv4 address to form a globally unique VPN-IPv4 address of the PE.

- One or more Route Target attributes. If any other PE has one of these Route Targets configured for 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.

- An "RPF Attribute". This is an IP address of the PE which originally generated the update.

- I-PMSI attribute. This attribute specifies:

  * Whether an MI-PMSI is to be used to support the MVPN,

  * whether the tunnel instantiating the the MVPN should be set up and/or joined immediately

  * whether the I-PMSI is instantiated by

    + A PIM-Bidir tree,

    + a set of PIM-SSM trees,

    + a set of PIM-SM trees (in this case the RP must be identified)

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

    + a set of unicast tunnels

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

  * If the PE wishes to setup a default tunnel to instantiate the I-PMSI, a unique identifier for the tunnel used to instantiate the I-PMSI. If the tunnel is a PIM-SSM tree, this identifier is a group P-address. Together with the RPF attribute, this provides sufficient information to enable the tree to be created. If the tunnel is a set of RSVP-TE point-to-multipoint LSPs, this will be the tunnel identifier used in the RSVP-TE messages used to create the tunnel from the originating PE.

    Note that a default tunnel can be identified at discovery time only if the tunnel already exists (e.g., it was constructed by means of configuration), or if it can be

constructed without each PE knowing the the identities of all the others (e.g., it is constructed by a receiver-initiated join technique such as PIM).

If a default tunnel is not identified at discovery time, the PE uses the tunnel to MVPN binding advertisements to signal the tunnel identifier.

* The type of encapsulation used on the tunnel, e.g. GRE or MPLS.

* 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.

* Whether multicast control information is to be sent on the MI-PMSI, whether it is to be unicast, or whether it is to be transmitted via BGP. Procedures for transmitting multicast control information are specified in section 5.

  + If multicast control information is to be sent on the MI-PMSI, whether this PE requires to receive periodic PIM Hello messages or not.

  + If multicast control information is to be sent on the MI-PMSI, or by using unicast PIM, whether this PE can perform PIM Refresh Reduction procedures.

  + If multicast control information is to be unicast, rather than sent on an MI-PMSI, then a demultiplexor value must be carried by the control messages in order to identify the particular MVPN to which the control message belongs. The form of this demultiplexor is not yet agreed upon, but one possibility is a downstream-assigned MPLS label. If this procedure is adopted, this label would be carried in the BGP update message. The PE originating the BGP update message would specify the label which it expects to see on received control packets for the specified MVPN.

The information specified here should be sufficient to enable a PE attached to a given MVPN:

- to discover the identities of all the other PEs attached to that same MVPN,

- to learn the procedure used to transmit multicast control messages,

- to learn the default procedure used to transmit multicast data messages, and

- if an I-PMSI is to be used, to identify any default tunnels, if possible. The cases when a default tunnel can be identified are described above and also discussed in further detail in section 6.

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

5. PE-PE Transmission of Multicast Routing

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

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

5.1. PIM Peering

5.1.1. 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 is the procedure specified in [rosen-08].]

5.1.2. 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 a BGP-based auto-discovery procedure. So in MVPN the periodic Hellos serve no function, and can simply be eliminated. (Of course, this does imply a change to the standard PIM procedures.)

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

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

5.1.3. 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, all PEs in the MVPN must send the C-Join/Prune for the given C-Source to the same PE. This is achieved using the following procedure:

If the next-hop interface on a PE's route to C-S, that is installed in the VRF, is a VRF interface than the PE should use that route to reach C-S. Else: from all the VPN-IPv4 routes that could be imported into the VRF and have exactly the same IPv4 prefix as the route in the VRF the PE uses the one with the highest next-hop address to determine the upstream PE to send C-Joins/C-Prunes for C-S.

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

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

To form a full PIM adjacency, the PEs execute the PIM LAN procedures, including the generation and processing of PIM Hello, Join/Prune, Assert, DF election and other PIM control packets. These are executed independently for each C-instance. PIM "join suppression" MAY be enabled.

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

The difference between a lightweight adjacency and a full adjacency is that no PIM Hellos are sent or received on a lightweight adjacency. The function which Hellos usually provide in PIM can be

provided in MVPN by the BGP-based auto-discovery procedures, so the Hellos become superfluous.

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

5.1.4.1. PIM C-Instance Control Packets

All PIM C-Instance control packets of a particular MVPN are addressed to the ALL-PIM-ROUTERS (224.0.0.13) IP destination address, and transmitted over the MI-PMSI of that MVPN. While in transit in the P-network, the packets are encapsulated as required for the particular kind of tunnel that is being used to instantiate the MI-PMSI. Thus the C-instance control packets are not processed by the P routers, and MVPN-specific PIM routes can be extended from site to site without appearing in the P routers.

5.1.4.2. PIM C-instance RPF Determination

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

When a PE needs to determine the RPF interface of a particular C-address, it looks up the C-address in the VRF. If the route matching it (call this the "RPF route") is not a VPN-IP route learned from MP-BGP as described in [RFC2547bis], or if that route's outgoing interface is one of the interfaces associated with the VRF, then ordinary PIM procedures for determining the RPF interface apply.

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

Once PIM has determined that the RPF interface for a particular C-address is the MI-PMSI, it is necessary for PIM to determine the RPF neighbor for that C-address. This will be one of the other PEs that is a PIM adjacency over the MI-PMSI.

When a PE distributes a given VPN-IP route via BGP, the PE must determine whether that route might possibly be regarded, by another

PE, as an RPF route. (If a given VRF is part of an MVPN, it may be simplest to regard every route exported from that VRF to be a potential RPF route.) If the given VPN-IP route is a potential RPF route, then when the VPN-IP route is distributed by BGP, it SHOULD be accompanied by an "RPF attribute".

The value of the RPF attribute is an IP address which the PE will use as its Source IP address in any PIM control messages which it transmits to other PEs in the same MVPN.

When a PE has determined that the RPF interface for a particular C-address is the MI-PMSI, it must look up the RPF attribute that was distributed along with the VPN-IPv4 address corresponding to that C-address. The value of this RPF attribute will be considered to be the IP address of the RPF adjacency for the C-address.

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

5.2. Use of BGP for Carrying PE-PE Multicast Routing

It may be possible to use BGP to carry multicast routing information from PE to PE, dispensing entirely with the transmission of C-Join/Prune messages from PE to PE.

A new SAFI is used for this purpose. The following information is required in BGP to advertise the MVPN routing information:

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

2. The address of the advertising PE. This is the same address that is used by the PE in the BGP MVPN auto-discovery procedures.

3. The C-Source address. This can be a prefix.

4. The C-Group address. This can be a prefix.

5. The upstream PE for which the message is intended. This address is either taken from the RPF attribute of the route matching the C-Source, or, if the RPF attribute is not present, is the BGP next-hop on the path to the C-Source. If the C-Source is a wildcard (e.g. (*, G) information), the upstream PE is the BGP next-hop on the path to C-RP. When a PE receives a C-Join (S,G) from a CE, the PE checks its VRF to find the address of the (upstream) PE that is used to reach the C-Source (we call this Ingress-PE address). A procedure must be specified to ensure that if there is more than one PE through which S could be reached, all PEs in the MVPN will send the C-Join/Prunes for the given C-source to the same PE. This is achieved using the following procedure:

   If the next-hop interface on a PE's route to C-S, that is installed in the VRF, is a VRF interface than the PE should use that route to reach C-S. Else: from all the VPN-IPv4 routes that could be imported into the VRF and have exactly the same IPv4 prefix as the route in the VRF the PE uses the one with the highest next-hop address to determine the upstream PE to send C-Joins/C-Prunes for C-S.

When a PE distributes this information via BGP, it must include a Route Target Extended Communities attribute. This RT must be an "Import RT" [RFC2547bis] of each VRF in the MVPN. The BGP distribution procedures used by [RFC2547bis] will then ensure that this information gets associated with the right VRFs. A PE will process this information only if the upstream PE address carried in the advertisement is the PE's own address, or if the PE is configured to be an anycast C-RP (section 9).

The use of BGP to propagate MVPN routing information allows the use of Route Reflectors, and has the same characteristics, with respect to the routing adjacencies maintained, as the use of BGP for distributing unicast VPN routing information as in [RFC2547bis].

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.

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

6.1. MVPN Membership and Egress PE Auto-Discovery

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

Thus an ingress PE must not transmit C-multicast data through a particular kind of unicast tunnel to a particular remote PE unless it knows that the remote PE has supports that particular tunnel encapsulation, and unless there is some agreement between the two PEs as to what values of the demultiplexing information represent which MVPNs. If a PE is capable of ingress replication for an MVPN, it announces this as part of the BGP based MVPN membership auto-discovery process, described in section 4. The following information elements need to be advertised by a PE:

1. The ability to support ingress replication for a particular MVPN, using a particular encapsulation format (e.g., MPLS, MPLS-in-GRE).

2. A mapping of demultiplexing values to MVPNs.

   If the encapsulation format is MPLS or MPLS-in-something, the demultiplexing values may be downstream-assigned MPLS labels. The encapsulation procedures are described further in section 11.

Other demultiplexing procedures for unicast are under consideration.

6.1.2. Auto-Discovery for P-Multicast Trees

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

The announcement of a tree identifier at discovery time is only possible if the tree already exists (e.g., a preconfigured "traffic engineered" tunnel which is know to include all the necessary PEs), 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., PIM-created trees).

6.2. MVPN 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. This is because there may be no MI-PMSI available for it to exchange MVPN routing information.

6.3. Aggregation

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

Aggregate Trees allow a single P-multicast tree to be used across multiple MVPNs and hence state in the SP core grows per-set-of-MVPNs and not per MVPN. Depending on the congruency 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.

6.3.2. Aggregation Methodology

This document does not specify any 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 "congruency" 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 congruency 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 congruency 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 congruency 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 a MPLS encapsulation. The protocol type in the IP/GRE header in the former case and the protocol type in the data link header in the latter need further explanation. This will be specified in a separate document.

6.3.4. Demultiplexing C-multicast traffic

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

This procedure requires each egress PE to support a separate label space for every other PE. The egress PEs create a forwarding entry for the upstream assigned MPLS label, allocated by the ingress PE, in this label space. Hence when the egress PE receives a packet over an Aggregate Tree, it first determines the tree that the packet was received over. The tree identifier determines the label space in which the upstream assigned MPLS label lookup has to be performed. The same label space may be used for all P-multicast trees rooted at the same ingress PE, or an implementation may decide to use a separate label space for every P-multicast tree.

The encapsulation format is either MPLS or MPLS-in-something (e.g.

MPLS-in-GRE). 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.)

6.4. Mapping Received Packets to MVPNs

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

6.4.1. Unicast Tunnels

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

6.4.2. Non-Aggregated P-Multicast Trees

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

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

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

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

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

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

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

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

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

- Tunnel Type.

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

6.4.3. Aggregate P-Multicast Trees

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

- The mapping of all MVPNs mapped to the Tree.

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

- The P-multicast tree Identifier

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

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

6.5. I-PMSI Instantiation Using Ingress Replication

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

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

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

However when ingress replication is used it is recommended that only S-PMSIs be used. Instantiation of S-PMSIs with ingress replication is described in section 7.2.

6.6. Establishing P-Multicast Trees

The architecture outlined in this document places no limitations on the protocols used to instantiate P-multicast trees. However any protocol specific procedures, are described only for PIM-SM, PIM-SSM, PIM-Bidir and RSVP-TE P2MP LSPs.

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 and RSVP-TE 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 or control packets to the root of the shared tree. This may require an unicast tunnel between each of these PEs and the root. The root will then send them on the shared tree and all the PEs that are leaves of the shared tree will receive the packets. For example a RP based PIM-SM tree would be a shared tree. Shared trees can be instantiated using PIM-SM, PIM-SSM, PIM-Bidir and RSVP-TE P2MP LSPs. Aggregation support for PIM-Bidir based P-multicast 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-TE-P2MP] are used to signal the LSP. The LSP is signalled 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. Signalling 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).

S-PMSI creation can be triggered on other criteria than mere bandwidth once Join suppression is disabled. For instance there could be a "pseudo wasted bandwidth" criteria : switching to a 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 many sparsely subscribed low-bandwidth groups may waste much bandwidth, and using an S-PMSI for a high bandwidth multicast stream for which all PEs have receivers is of no use either

Switching to a S-PMSI requires the root of the S-PMSI to a) Discover the egress PEs that will receive traffic using the S-PMSI b) Setup the S-PMSI and c) If required, signal the binding of the multicast stream to the S-PMSI to the leaves of the tunnel used to instantiate the S-PMSI.

Step (c) is required only when the tunnel is a P-multicast tree and we specify two methods for achieving this.

7.1. Egress PE Discovery

S-PMSI instantiation using a tunnel may imply using an existing tunnel or creating a new tunnel. Depending on the type of tunnel used the PE may or may not need to know the PEs that have receivers in the C-(S, G) bound to the S-PMSI. If the PE needs to know the egress PEs before instantiating the S-PMSI, it MUST receive C-Joins from each of the egress PEs. To achieve this either unicast PIM or BGP or MI-PMSI with Join suppression disabled MUST be used to is used to transmit C-Joins. There are two cases in which the PE needs to know the egress PEs:

- 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. This is learned from the C-Joins received by the ingress PE. Hence the leaves of the Aggregate Tree are discovered using C-Joins.

7.2. 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 it should 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.

7.3. 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.

7.3.1. 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.

7.3.1.1. 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.

7.3.1.2. Packet Formats and Constants

To be included.

7.3.2. 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).

7.3.2.1. 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 using an update with a new <AFI, SAFI>. The update contains the following:

a) The originating PE's address.

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

c) If the S-PMSI is instantiated using an Aggregate Tree with a selective mapping, an inner label allocated by the Aggregate Tree root for the <C-Source, C-Group>. This allows a single tunnel to aggregate multiple S-PMSIs. This is the upstream label the usage of which is described in section 6.3.4.

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

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

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.

7.3.2.2. 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.

7.4. 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.

7.5. Instantiating the S-PMSI with a PIM Tree

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

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

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

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

The above procedures assume that each PE router has a set of group

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

If it is known that all the PEs which need to receive data traffic on a given S-PMSI can support aggregation of multiple S-PMSIs on a single PIM tree, then the transmitting PE, may, at its discretion, decide to bind the S-PMSI to a PIM tree which is already bound to one or more other S-PMSIs, from the same or from different MVPNs. In this case, appropriate demultiplexing information must be signaled.

7.6. 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

This document describes two PMSI instantiation models for supporting an Inter-AS MVPN service:

1. A model where Inter-AS MVPN service is delivered over a PMSI that is instantiated using a tunnel that spans multiple ASs.

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

2. A model where Inter-AS MVPN service is delivered over a PMSI that is instantiated without requiring a tunnel to span multiple ASs. Each AS can have a different tunnel. The PMSI is instantiated using an "overlay tunnel" which is overlayed on each of the intra-AS tunnels. This allows an Inter-AS MVPN service to be provided with each AS potentially supporting a different MVPN transport technology.

[RFC2547bis] describes different options for supporting Inter-AS BGP/MPLS unicast VPNs. Inter-AS MVPNs can be supported for each of these unicast BGP/MPLS VPN Inter-AS options, for both the PMSI instantiation models mentioned above. No further clarifications are required for option A. We describe these two instantiation models in

the sub-sections below.

8.1. Tunnel Spans Multiple ASs

In this model, the previously described discovery and tunnel setup mechanisms are used, even though the PEs belonging to a given MVPN may be in different ASes. The ASBRs play no special role, but function merely as P routers.

8.1.1. Inter-AS MVPN Auto-Discovery

The previously described BGP-based auto-discovery mechanisms work "as is" when an MVPN contains PEs that are in different Autonomous Systems.

8.1.2. Inter-AS MVPN Routing Information Exchange

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

If any form of PIM peering is used, a PE that sends C-PIM Join/Prune messages for a particular C-(S,G) must be able to identify the PE which is its PIM adjacency on the path to S. The identity of the PIM adjacency is determined from the RPF attribute associated with the VPN-IPv4 route to S.

If no RPF attribute is present, then the identity of the PIM adjacency is taken from the BGP Next Hop attribute of the VPN-IPv4 route to S. Note that this will not give the correct result if option b of section 10 of [rfc2547bis] is used. To avoid this possibility of error, the RPF attribute SHOULD always be present if MVPN routing information is to be distributed by PIM.

If BGP (rather than PIM) is used to distribute the MVPN routing information, and if option b of section 10 of [rfc2547bis] is in use, then the MVPN routes will be installed in the ASBRs along the path from each multicast source in the MVPN to each multicast receiver in the MVPN. If option b is not in use, the MVPN routes are not installed in the ASBRs. The handling of MVPN routes in either case is thus exactly analogous to the handling of unicast VPN-IPv4 routes in the corresponding case.

8.1.3. Inter-AS I-PMSI

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

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

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

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

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

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

   - BGP will install a unicast route to P-S along a particular path, using the IPv4 AFI/SAFI ;

   - A PE's MVPN auto-discovery information is advertised by sending a BGP update whose NLRI is in a special address family (AFI/SAFI) used for this purpose. The NLRI of the address family contains the IPv4 address of the PE, as well as an RD. If the NLRI contains the IPv4 address of P-S, this in effect creates a second route to P-S. This route might follow a different path than the route in the unicast IPv4 family.

When building a PIM tree towards P-S, it may be desirable to build it along the route on which the MVPN auto-discovery AFI/SAFI is installed, rather than along the route on which the IPv4 AFI/SAFI is installed. This enables the inter-AS portion of the tree to follow a path which is specifically chosen for multicast (i.e., it allows the inter-AS multicast topology to be "non-congruent" to the inter-AS unicast topology).

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

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

8.1.4. Inter-AS S-PMSI

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

8.2. Overlay Inter-AS Tunnel

8.2.1. Inter-AS MVPN Auto-Discovery

The BGP based MVPN membership discovery procedures of section 4 are used to auto-discover the inter-AS MVPN membership.

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 ASs. The leaves of this tree are only those ASs that have at least one PE with a member in the MVPN. The overlay 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 ASs.

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.

Each ASBR then advertises the "AS MVPN membership" to its neighbor ASBRs using EBGP. This advertisement must not be advertised to the

PEs/ASBRs in the the same AS as this ASBR. This advertisement in effect may take place using Route Reflectors. The advertisement carries the following information elements:

a. A Route Distinguisher for the MVPN. Each ASBR in the AS must use the same RD when advertising this information to other ASBRs. To accomplish this either the MVPN Route Target can be used as the RD or each ASBR can pick the RD advertised by a PE with the lowest BGP next-hop address.

b. Origin AS number. This AS number must be encoded in an IP address. As defined in [RFC1940] an IP address from network 128.0.0.0 is used to encode a next hop that is a domain. The least significant two octets contain the DI, which is an Internet Autonomous System number.

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

An ASBR in an AS that receives the above information from its EBGP peers, changes the next-hop to self, and announces it to its IBGP peers. It also runs BGP path selection on the information elements (a, b) described above.

If the ASBR is the next-hop for the best path it advertises:

a) To its neighbor ASBR, from which it received the information element, an ASBR-ASBR tunnel binding. This binding as described in section 6 can be used by the neighbor ASBR to send traffic to this ASBR.

b) To the PEs in its AS an I-PMSI tunnel binding, for the AS MVPN membership received from the other AS, as described in section 6.

The ASBR announces the best path to its EBGP peers, with next-hop as self. Thus the AS MVPN membership information propagates across multiple ASs 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 ASs.

The objective in this case is to propagate the MVPN routing information to the remote PE, in the reverse direction of the AS MVPN

routing information announced by the remote PE's origin AS. This is to build S-PMSIs that utilize overlay inter-AS tunnels as described in section 8.2.4. 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 remote PE. It then determines from the received AS MVPN membership information, the ASBR that is the next-hop for the best path of the source AS MVPN membership. The MVPN membership information is sent to this ASBR, if unicast PIM is used. If BGP is used the upstream address is set to this ASBR and the ASBR then further propagates the BGP advertisement.

8.2.3. Inter-AS I-PMSI

One option for instantiating an inter-AS I-PMSI is to use different tunnels in each AS and stitch these tunnels together using MPLS label switching. This results in an overlay inter-AS tunnel. This allows each AS to use a different tunneling technology. If P-multicast trees are used, it allows each AS to use a different multicast tree protocol.

The overlay inter-AS tunnel, for an I-PMSI is rooted at the router that instantiates the I-PMSI. It traverses the AS MVPN membership spanning tree. Each ASBR on the spanning tree advertises a tunnel binding for the MVPN to its upstream ASBR. For a given AS only one ASBR advertises this binding for a given AS MVPN membership to its upstream ASBR. This ensures that there is only one exit point in an AS to reach another AS for the inter-AS PMSI. It also ensures that there is only one entry point in an AS to receive traffic from another AS for the inter-AS PMSI. This tunnel binding carries the following information elements:

a. The AS MVPN membership for which the tunnel binding is being advertised.

b. A downstream label that the ASBR allocates to receive traffic from its upstream ASBR on the inter-AS overlay tunnel. This is the tunnel binding.

Each downstream ASBR that is on the spanning tree (as determined by the AS MVPN information best path selection) also instantiates an I-PMSI using an intra-AS tunnel, for the MVPN.

A C-multicast data packet is sent using an intra-AS tunnel by the PE that first receives this packet. 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 for the MVPN 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 2547 Unicast VPN Inter-AS Methods

The above procedures for setting up an inter-AS I-PMSI can be supported for each of the 2547 unicast VPN inter-AS models. 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

One option for instantiating an inter-AS S-PMSI is to use different tunnels in each AS and stitch these tunnels together using MPLS label switching. This results in an inter-AS overlay tunnel for the S-PMSI. The procedures are conceptually similar to the ones described for an I-PMSI inter-AS overlay tunnel.

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. A downstream ASBR on the spanning tree sends back a tunnel binding for AS <C-S, C-G> information. If the downstream 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 downstream ASBR may decide to use an I-PMSI instead, in which case it sends back the same label that it advertised for the I-PMSI. If the downstream ASBR instantiates a S-PMSI, it further propagates the <C-S, C-G> membership to its downstream ASs, else it does not.

An AS can instantiate an intra-AS S-PMSI for the inter-AS S-PMSI overlay 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 leaves <C-S, C-G> 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. Deployment Models

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

9.1. Co-locating C-RPs on a PE

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

"In some cases this engineering problem is not trivial: for instance, if sources and receivers are located in VPN sites that are different than that of the RP, then traffic may flow twice through the SP network and the CE-PE link of the RP (from source to RP, and then from RP to receivers) ; this is obviously not ideal. A multicast VPN solution SHOULD propose a way to help on solving this RP engineering issue."

One of the C-RP deployment models is for the customer to outsource the RP to the provider. In this case the provider may co-locate the RP on the PE that is connected to the customer site [MVPN-REQ]. This model is introduced in [RP-MVPN]. This section describes how anycast-RP can be used for achieving this. It can either be done without advertising the active sources or by also advertising the active sources. This is described below.

9.1.1. Initial Configuration

For each VPN site connected to a PE the PE acts as an RP. Within each VPN all these RPs use the same (anycast) address. All these RPs use the Anycast RP technique.

9.1.2. Anycast RP Based on C-(*, G) Advertisements

9.1.2.1. Receiver(s) Within a Site

When a host within a VPN site connected to a given PE joins a particular multicast group G, the designated router connected to the host would send C-Join for (*,G) to the PE (as the PE acts as an RP within the site). If the site is multi-homed (means the site is connected to more than one PE), then all the PEs connected to the site act as (anycast) RP. So, the Join would arrive at one of these PEs (the one that is closest from the routing point of view to the designated router that originated Join (*,G)).

When a PE receives a Join (*,G) from a CE, the PE sends to all other PEs that are in the same VPN as the CE, the information that it has receiver(s) for G. Sending this information could be done using any of the procedures described in section 5. If BGP is used the upstream PE address is set to 0. A PE will process the message after it imports it into the VRF only if it is configured to be an anycast C-RP.If unicast PIM is used then a unicast PIM message will have to be sent to each of the other PEs. The upstream neighbor address for a unicast PIM message sent to a PE is the destination PE's address. 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.

9.1.2.2. Source within a Site

When a host S within a VPN site starts sending (multicast) packets to a particular group G, the designated router connected to the host would encapsulate these messages into PIM-Register and will send it to the PE connected to the site (as the PE acts as an (anycast) RP for that site).

When a PE receives PIM-Register from a site that belongs to a given VPN, PE follows the normal PIM procedures. In addition, if the PE receives information from other PEs that other sites of that VPN have receivers for the group G, carried in the PIM-Register, the PE sends the multicast data on an I-PMSI or a S-PMSI.

9.1.2.3. Receiver Switching from Shared to Source Tree

When a receivers within a VPN site switches from shared to source tree the receiver sends Join C-(S,G) towards C-S, and also Prune (S,G,RPT) towards its (proxy) RP.

9.1.2.3.1. Handling Join (S,G)

If PE already has an existing Join(*,G) state, then when the PE receives Join(S,G) the PE does not need to advertise this information to the other PEs.

If Join(*,G) is expired, but Join(S,G) is still present, the PE has to advertise replacement of C-(*,G) with C-(S,G).

Likewise, if PE does not have an existing Join(*,G) state, when PE receives Join(S,G) the PE has to advertise C-(S,G).

The PE determines the upstream PE address as specified in section 5.2. Once this is determined, the PE sends <S,G, Ingress-PE> information to all other PEs that are in the same VPN as the CE. Sending this information could be done with either of the methods described in section 5.

9.1.2.3.2. Handling Prune (S,G, RPT)

When PE receives Prune(S,G,RPT) from CE, and the PE is receiving traffic for the (S,G) on an I-PMSI the PE does nothing; and if the PE receives (S,G) on an S-PMSI associated with (S,G) the PE informs the root of that S-PMSI that the PE is no longer interested in (S,G). This can be accomplished using any of the mechanisms of section 5.

9.1.2.3.3. Receiving information from other PEs

When a PE receives <S, G, Ingress-PE> information from some other PE, the local PE checks whether it is the Ingress-PE, and if yes, then the PE sends Join (S,G) towards S. Otherwise, if PE has been sending Join (S,G) to one of its CEs, the PE now sends Prune (S,G) to that CE. Otherwise, the PE does nothing.

9.1.3. Anycast RP Based on Propagating Active Sources

The second mechanism is based on propagating active sources between RPs.

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

The essential difference between this mechanism and the mechanism specified in the previous section is that when a PE discovers that there is a source in a site attached to that PE, the PE advertises this source in BGP.

9.1.3.1. 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 source active information from an upstream PE.

On receiving source active information (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.

9.1.3.2. Source Within a Site

When a PE receives PIM-Register from a site that belongs to a given VPN, PE follows the normal PIM procedures. It then advertises this source and group to other PEs in BGP using the following information elements:

- Active source address

- Active group address

- Ingress PE address

- Route target of the MVPN.

This advertisement goes to all the PEs. 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.

9.1.3.3. 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. BGP Advertisements

10.1. Functions and Information Elements

The information elements signalled in BGP and the functions of these information elements has been described in detail earlier.

10.2. Encoding

The encoding will be described later.

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 IPv4 Protocol Number field in the P-IP Header must be set to 47. The Protocol Type field of the GRE Header must be set to 0x800.

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

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.

[GRE2784] specifies an optional GRE checksum, and [GRE2890] 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 [GRE2784].

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 [IPIP1853] 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   ||
++=============++       ++=============++       ++=============++
```

11.1.3. Encapsulation in MPLS

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

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

```
   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 [draft-mpls-in-ip-or-gre]. 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 11.1.2, we use the MPLS-in-IP encapsulation. This is specified in [draft-mpls-in-ip-or-gre]. The IP protocol number MUST be set to the value identifying the payload as an MPLS unicast packet. [There is no "MPLS multicast packet" protocol number.]

11.3. Encapsulations for Unicasting PIM Control Messages

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

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

This area is still under consideration.

11.4. General Considerations for IP and GRE Encaps

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

11.4.1. MTU

Path MTU discovery cannot be relied upon to ensure that the transmitter sends packets which are small enough to reach all the destinations. This requires that:

a. The ingress PE router (one that does the encapsulation) must not set the DF bit in the outer header, and

b. If the "DF" bit is cleared in the IP header of the C-Packet, fragment the C-Packet before encapsulation if appropriate. This is very important in practice due to the fact that the performance of reassembly function is significantly lower than that of decapsulating and forwarding packets on today's router implementations.

11.4.2. TTL

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

11.4.3. Differentiated Services

By default, the setting of the DS field in the delivery IP header should follow the guidelines outlined in [DIFF2983]. Setting the EXP field in the delivery MPLS header should follow the guidelines in [REF]. An SP may also choose to deploy any of the additional mechanisms the PE routers support.

11.4.4. Avoiding Conflict with Internet Multicast

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

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

12. Security Considerations

To be supplied.

13. IANA Considerations

To be supplied.

14. Other Authors

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

15. Other Contributors

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

16. Authors' Addresses

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

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

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

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

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

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

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

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

17. Normative References

[MVPN-REQ] T. Morin, Ed., "Requirements for Multicast in L3 Provider-Provisioned VPNs", draft-ietf-l3vpn-ppvpn-mcast-reqts-00.txt

[2547bis] "BGP/MPLS VPNs", Rosen, Rekhter, et. al., September 2003, draft-ietf-l3vpn-rfc2547bis-01.txt

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels.", Bradner, March 1997

[PIM-SM] "Protocol Independent Multicast - Sparse Mode (PIM-SM)", Fenner, Handley, Holbrook, Kouvelas, October 2003, draft-ietf-pim-sm-v2-new-08.txt

[RSVP-P2MP] R. Aggarwal, et. al., "Extensions to RSVP-TE for Point to Multipoint TE LSPs", draft-ietf-mpls-rsvp-te-p2mp-01.txt

[RFC3107] Y. Rekhter, E. Rosen, "Carrying Label Information in BGP-4", RFC3107.

[MPLS-IP] T. Worster, Y. Rekhter, E. Rosen, "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", draft-ietf-mpls-in-ip-or-gre-08.txt

[MPLS-MCAST-ENCAPS] T. Eckert, E. Rosen, R. Aggarwal, Y. Rekhter, "MPLS Multicast Encapsulations", draft-rosen-mpls-multicast-encaps-00.txt, April 2005

[MPLS-UPSTREAM-LABEL] R. Aggarwal, Y. Rekhter, E. Rosen, draft-raggarwa-mpls-upstream-label-00.txt, "MPLS Upstream Label Assignment and Context Specific Label Space", draft-raggarwa-mpls-upstream-label-00.txt, January 2005

18. Informative References

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

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

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

[RP-MVPN] S. Yasukawa, et. al., "BGP/MPLS IP Multicast VPNs", draft-yasukawa-l3vpn-p2mp-mcast-00.txt

19. Full Copyright Statement

Copyright (C) The Internet Society (2005).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

20. Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.