Network Working Group
Internet Draft
Intended Status: Historic
Expires: December 9, 2010

Eric C. Rosen
Yiqun Cai
IJsbrand Wijnands
Cisco Systems, Inc.

June 9, 2010

Cisco Systems' Solution for Multicast in MPLS/BGP IP VPNs

draft-rosen-vpn-mcast-15.txt

Status of this Memo

Copyright and License Notice

Abstract

This draft describes the MVPN (Multicast in BGP/MPLS IP VPNs) solution designed and deployed by Cisco Systems. The procedures specified in this draft are largely a subset of the generalized MVPN framework recently standardized by the IETF. However, as the deployment of the procedures specified herein pre-dates the publication of IETF standards (in some cases by over five years), an implementation based on these procedures differs in some respects from a fully standards-compliant implementation. These differences are pointed out in the 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

This draft describes the MVPN (Multicast in BGP/MPLS IP VPNs) solution designed and deployed by Cisco Systems. This document is being made available for the record and as a reference for interoperating with deployed implementations. This document is a technical specification and should not be used to infer the current or future plans of Cisco Systems.

The procedures specified in this draft differ are largely a subset of the generalized MVPN framework defined in [MVPN]. However, as this draft specifies an implementation that precedes the standardization of [MVPN] by several years, it does differ in a few respects from a fully standards-compliant implementation. These differences are pointed out where they occur.

The base specification for BGP/MPLS IP VPNs [RFC4364] does not provide a way for IP multicast data or control traffic to travel from one VPN site to another. This document extends that specification by specifying the necessary protocols and procedures for support of IP multicast.

This specification presupposes that:

1. PIM [PIMv2], running over either IPv4 or IPv6, is the multicast routing protocol used within the VPN,

2. PIM, running over IPv4, is the multicast routing protocol used within the SP network, and

3. the SP network supports native IPv4 multicast forwarding.

Familiarity with the terminology and procedures of [RFC4364] is presupposed. Familiarity with [PIMv2] is also presupposed.

2.1. Scaling Multicast State Info. in the Network Core

The BGP/MPLS IP VPN service of [RFC4364] provides a VPN with "optimal" unicast routing through the SP backbone, in that a packet follows the "shortest path" across the backbone, as determined by the backbone's own routing algorithm. This optimal routing is provided without requiring the P routers to maintain any routing information that is specific to a VPN; indeed, the P routers do not maintain any per-VPN state at all.

Unfortunately, optimal MULTICAST routing cannot be provided without requiring the P routers to maintain some VPN-specific state information. Optimal multicast routing would require that one or more multicast distribution trees be created in the backbone for each multicast group that is in use. If a particular multicast group from within a VPN is using source-based distribution trees, optimal routing requires that there be one distribution tree for each transmitter of that group. If shared trees are being used, one tree for each group is still required. Each such tree requires state in some set of the P routers, with the amount of state being proportional to the number of multicast transmitters. The reason there needs to be at least one distribution tree per multicast group is that each group may have a different set of receivers; multicast routing algorithms generally go to great lengths to ensure that a multicast packet will not be sent to a node that is not on the path to a receiver.

Given that an SP generally supports many VPNs, where each VPN may have many multicast groups, and each multicast group may have many transmitters, it is not scalable to have one or more distribution trees for each multicast group. The SP has no control whatsoever over the number of multicast groups and transmitters that exist in the VPNs, and it is difficult to place any bound on these numbers.

In order to have a scalable multicast solution for MPLS/BGP IP VPNs, the amount of state maintained by the "P routers" (routers in the provider backbone, other than the "provider edge" or "PE" routers) needs to be proportional to something that IS under the control of the SP. This specification describes such a solution. In this solution, the amount of state maintained in the P routers is proportional only to the number of VPNs that run over the backbone; the amount of state in the P routers is NOT sensitive to the number of multicast groups or to the number of multicast transmitters within the VPNS. To achieve this scalability, the optimality of the multicast routes is reduced. A PE that is not on the path to any receiver of a particular multicast group may still receive multicast packets for that group, and if so, will have to discard them. The SP does however have control over the tradeoff between optimal routing

and scalability.

2.2. Overview

An SP determines whether a particular VPN is multicast-enabled. If it is, it corresponds to a "Multicast Domain". A PE that attaches to a particular multicast-enabled VPN is said to belong to the corresponding Multicast Domain. For each Multicast Domain, there is a default "Multicast Distribution Tree (MDT)" through the backbone, connecting ALL of the PEs that belong to that Multicast Domain. A given PE may be in as many Multicast Domains as there are VPNs attached to that PE. However, each Multicast Domain has its own MDT. The MDTs are created by running PIM in the backbone, and in general an MDT also includes P routers on the paths between the PE routers.

In a departure from the usual multicast tree distribution procedures, the Default MDT for a Multicast Domain is constructed automatically as the PEs in the domain come up. Construction of the Default MDT does not depend on the existence of multicast traffic in the domain; it will exist before any such multicast traffic is seen. Default MDTs correspond to the "MI-PMSIs" of [MVPN].

In BGP/IP MPLS VPNs, each CE ("Customer Edge", see [RFC4364]) 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. Multicast packets from within a VPN are received from a CE router by an ingress PE router. The ingress PE encapsulates the multicast packets and (initially) forwards them along the Default MDT tree to all the PE routers connected to sites of the given VPN. Every PE router attached to a site of the given VPN thus receives all multicast packets from within that VPN. If a particular PE routers is not on the path to any receiver of that multicast group, the PE simply discards that packet.

If a large amount of traffic is being sent to a particular multicast group, but that group does not have receivers at all the VPN sites, it can be wasteful to forward that group's traffic along the Default MDT. Therefore, we also specify a method for establishing individual MDTs for specific multicast groups. We call these "Data MDTs". A Data MDT delivers VPN data traffic for a particular multicast group only to those PE routers that are on the path to receivers of that multicast group. Using a Data MDT has the benefit of reducing the amount of multicast traffic on the backbone, as well reducing the load on some of the PEs; it has the disadvantage of increasing the

amount of state that must be maintained by the P routers. The SP has complete control over this tradeoff. Data MDTs correspond to the S-PMSIs of [MVPN].

This solution requires the SP to deploy appropriate protocols and procedures, but is transparent to the SP's customers. An enterprise that uses PIM-based multicasting in its network can migrate from a private network to a BGP/MPLS IP VPN service, while continuing to use whatever multicast router configurations it was previously using; no changes need be made to CE routers or to other routers at customer sites. For instance, any dynamic RP-discovery procedures that are already in use may be left in place.

3. Multicast VRFs

The notion of a "VRF", defined in [RFC4364], is extended to include multicast routing entries as well as unicast routing entries.

Each VRF has its own multicast routing table. When a multicast data or control packet is received from a particular CE device, multicast routing is done in the associated VRF.

Each PE router runs a number of instances of PIM-SM, as many as one per VRF. In each instance of PIM-SM, the PE maintains a PIM adjacency with each of the PIM-capable CE routers associated with that VRF. The multicast routing table created by each instance is specific to the corresponding VRF. We will refer to these PIM instances as "VPN-specific PIM instances", or "PIM C-instances".

Each PE router also runs a "provider-wide" instance of PIM-SM (a "PIM P-instance"), in which it has a PIM adjacency with each of its IGP neighbors (i.e., with P routers), but NOT with any CE routers, and not with other PE routers (unless they happen to be adjacent in the SP's network). The P routers also run the P-instance of PIM, but do NOT run a C-instance.

In order to help clarify when we are speaking of the PIM P-instance and when we are speaking of a a PIM C-instance, we will also apply the prefixes "P-" and "C-" respectively to control messages, addresses, etc. Thus a P-Join would be a PIM Join that is processed by the PIM P-instance, and a C-Join would be a PIM Join that is processed by a C-instance. A P-group address would be a group address in the SP's address space, and a C-group address would be a group address in a VPN's address space.

4. Multicast Domains

4.1. Model of Operation

A "Multicast Domain (MD)" is essentially a set of VRFs associated with interfaces that can send multicast traffic to each other. From the standpoint of PIM C-instance, a multicast domain is equivalent to a multi-access interface. The PE routers in a given MD become PIM adjacencies of each other in the PIM C-instance.

Each multicast VRF is assigned to one MD. Each MD is configured with a distinct, multicast P-group address, called the "Default MDT group address". This address is used to build the Default MDT for the MD.

When a PE router needs to send PIM C-instance control traffic to the other PE routers in the MD, it encapsulates the control traffic, with its own IPv4 address as source IP address and the Default MDT group address as destination IP address. Note that the Default MDT is part of the PIM P-instance, whereas the PEs that communicate over the Default MDT are PIM adjacencies in a C-instance. Within the C-instance, the Default MDT appears to be a multi-access network to which all the PEs are attached. This is discussed in more detail in section 5.

The Default MDT does not only carry the PIM control traffic of the MD's PIM C-instance. It also, by default, carries the multicast data traffic of the C-instance. In some cases though, multicast data traffic in a particular MD will be sent on a Data MDT rather than on the Default MDT. The use of Data MDTs is described in section 7.

Note that, if an MDT (Default or Data) is set up using the ASM ("Any Source Multicast") Service Model, the MDT (Default or Data) must have a P-group address that is "globally unique" (more precisely, unique over the set of SP networks carrying the multicast traffic of the corresponding MD). If the MDT is set up using the SSM ("Single Source Multicast") model, the P-group address of an MDT only needs to be unique relative to the source of the MDT (though see section 5.4). However, some implementations require the same SSM group address to be assigned to all the PEs. Interoperability with those implementations requires conformance to this restriction.

5. Multicast Tunnels

An MD can be thought of as a set of PE routers connected by a "multicast tunnel (MT)". From the perspective of a VPN-specific PIM instance, an MT is a single multi-access interface. In the SP network, a single MT is realized as a Default MDT combined with zero or more Data MDTs.

5.1. Ingress PEs

An ingress PE is a PE router that is either directly connected to the multicast sender in the VPN, or via a CE router. When the multicast sender starts transmitting, and if there are receivers (or PIM RP) behind other PE routers in the common MD, the ingress PE becomes the transmitter of either the Default MDT group or a Data MDT group in the SP network.

5.2. Egress PEs

A PE router with a VRF configured in an MD becomes a receiver of the Default MDT group for that MD. A PE router may also join a Data MDT group if if it has a VPN-specific PIM instance in which it is forwarding to one of its attached sites traffic for a particular C-group, and that particular C-group has been associated with that particular Data MDT. When a PE router joins any P-group used for encapsulating VPN multicast traffic, the PE router becomes one of the endpoints of the corresponding MT.

When a packet is received from an MT, the receiving PE derives the MD from the destination address, which is a P-group address, of the received packet. The packet is then passed to the corresponding Multicast VRF and VPN-specific PIM instance for further processing.

5.3. Tunnel Destination Address(es)

An MT is an IP tunnel for which the destination address is a P-group address. However an MT is not limited to using only one P-group address for encapsulation. Based on the payload VPN multicast traffic, it can choose to use the Default MDT group address, or one of the Data MDT group addresses (as described in section 7 of this document), allowing the MT to reach a different set of PE routers in the common MD.

5.4. Auto-Discovery

Any of the variants of PIM may be used to set up the Default MDT: PIM-SM, Bidirectional PIM [BIDIR], or PIM-SSM [SSM]. Except in the case of PIM-SSM, the PEs need only know the proper P-group address in order to begin setting up the Default MDTs. The PEs will then discover each others' addresses by virtue of receiving PIM control traffic, e.g., PIM Hellos, sourced (and encapsulated) by each other.

However, in the case of PIM-SSM, the necessary MDTs for an MD cannot be set up until each PE in the MD knows the source address of each of the other PEs in that same MD. This information needs to be auto-discovered.

A new BGP Address Family, MDT-SAFI is defined. The NLRI for this address family consists of an RD, an IPv4 unicast address, and a multicast group address. A given PE router in a given MD constructs an NLRI in this family from:

- Its own IPv4 address. If it has several, it uses the one that it will be placing in the IP source address field of multicast packets that it will be sending over the MDT.

- An RD that has been assigned to the MD.

- The P-group address, an IPv4 multicast address that is to be used as the IP destination address field of multicast packets that will be sent over the MDT.

When a PE distributes this NLRI via BGP, it may include a Route Target (RT) Extended Communities attribute. This RT must be an "Import RT" [RFC4364] of each VRF in the MD. The ordinary BGP distribution procedures used by [RFC4364] will then ensure that each PE learns the MDT-SAFI "address" of each of the other PEs in the MD, and that the learned MDT-SAFI addresses get associated with the right VRFs.

If a PE receives an MDT-SAFI NLRI that does not have an RT attribute, the P-group address from the NLRI has to be used to associate the NLRI with a particular VRF. In this case, each multicast domain must be associated with a unique P-address, even if PIM-SSM is used. However, finding a unique P-address for a multi-provider multicast group may be difficult.

In order to facilitate the deployment of multi-provider multicast domains, this specification REQUIRES the use of the MDT-SAFI NLRI (even if PIM-SSM is not used to set up the default MDT). This specification also REQUIRES that an implementation be capable of

using PIM-SSM to set up the default MDT.

In [MVPN], the MDT-SAFI is replaced by the "Intra-AS I-PMSI A-D Route." The latter is a generalized version of the MDT-SAFI, which allows the "default MDTs" and "data MDTs" to be implemented as MPLS P2MP LSPs ("Point-to-Multipoint Label Switched Paths") or MP2MP ("Multipoint-to-Multipoint Label Switched Paths") LSPs, as well as by PIM-created multicast distribution trees. In the latter case, the Intra-AS A-D routes carry the same information that the MDT-SAFI does, though with a different encoding.

The Intra-AS A-D Routes also carry Route Targets, and so may be distributed in the same manner as unicast routes, including being distributed inter-AS. (Despite their name, the inter-AS distribution of Intra-AS I-PMSI A-D routes is sometimes necessary in [MVPN].)

The encoding of the MDT-SAFI is specified in the following subsection:

5.4.1. MDT-SAFI

BGP messages in which AFI=1 and SAFI=66 are "MDT-SAFI" messages.

The NLRI format is 8-byte-RD:IPv4-address followed by the MDT group address. i.e. The MP_REACH attribute for this SAFI will contain one or more tuples of the following form :

```
+-------------------------------+
|                               |
|  RD:IPv4-address (12 octets)  |
|                               |
+-------------------------------+
|    Group Address (4 octets)   |
+-------------------------------+
```

The IPv4 address identifies the PE that originated this route, and the RD identifies a VRF in that PE. The group address MUST be an IPv4 multicast group address, and is used to build the P-tunnels. All PEs attached to a given MVPN MUST specify the same group-address, even if the group is an SSM group. MDT-SAFI routes do not carry RTs, and the group address is used to associate a received MDT-SAFI route with a VRF.

5.5. Which PIM Variant to Use

To minimize the amount of multicast routing state maintained by the P routers, the Default MDTs should be realized as shared trees, such as PIM Bidirectional trees. However, the operational procedures for assigning P-group addresses may be greatly simplified, especially in the case of multi-provider MDs, if PIM-SSM is used.

Data MDTs are best realized as source trees, constructed via PIM-SSM.

5.6. Inter-AS MDT Construction

Standard PIM techniques for the construction of source trees presuppose that every router has a route to the source of the tree. 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.

What is needed in this case is a way for a PE 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.

If the PE has selected the source of the tree from the MDT SAFI address family, then it may be desirable to build the tree along the route to the MDT SAFI address, rather than along the route to the corresponding IPv4 address. This enables the inter-AS portion of the tree to follow a path that is specifically chosen for multicast (i.e., it allows the inter-AS multicast topology to be "non-congruent" to the inter-AS unicast topology). This too requires a PIM extension.

The necessary PIM extension is the PIM MVPN Join Attribute described in in the following sub-section.

5.6.1. The PIM MVPN Join Attribute

5.6.1.1. Definition

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

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |F|E|   Type    | Length        |     Proxy IP address
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                                    |      RD
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-.......
```

The 6-bit Type field of the MVPN Join Attribute is set to 1.

The F bit is set to 0, indicating that the attribute is non-transitive.

Rules for setting the E bit are given in [PIM-ATTRIB].

Two information fields are carried in the MVPN Join attribute:

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

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

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

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

5.6.1.2. Usage

When a PE router creates a PIM Join/Prune message in order to set up an inter-AS default MDT, it does so as a result of having received a particular MDT-SAFI route. It includes an MVPN Join attribute whose fields are set as follows:

- If the upstream PE is in the same AS as the local PE, then the proxy field contains the address of the upstream PE. Otherwise, it contains the address of the BGP next hop on the route to the upstream PE.

- The RD field contains the RD from the NLRI of the MDT-SAFI route.

- The upstream PE field contains the address of the PE that originated the MDT-SAFI route (obtained from the NLRI of that route).

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

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

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

5.7. Encapsulation in GRE

GRE [GRE1701] encapsulation is used when sending multicast traffic through an MDT. 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  |
                           +---------------+
                           |      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 0x0800 if C-IP header is an IPv4 header; it MUST be set to 0x86dd if the C-IP header is an IPv6 header.

[GRE2784] specifies an optional GRE checksum, and [GRE2890] specifies optional GRE key and sequence number fields.

The GRE key field is not needed because the P-group address in the delivery IP header already identifies the MD, and thus associates the VRF context, for the payload packet to be further processed.

The GRE sequence number field is also 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, key or sequence field.

5.8. MTU

Because multicast group addresses are used as tunnel destination addresses, existing Path MTU discovery mechanisms can not be used. This requires that:

1. The ingress PE router (one that does the encapsulation) MUST NOT set the DF ("Don't Fragment") bit in the outer header, and

2. 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.

5.9. TTL

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

5.10. Differentiated Services

By default, the setting of the DS ("Differentiated Services") field in the delivery IP header should follow the guidelines outlined in [DIFF2983]. An SP may also choose to deploy any of the additional mechanisms the PE routers support.

5.11. Avoiding Conflict with Internet Multicast

If the SP is providing Internet multicast, distinct from its VPN multicast services, it must ensure that the P-group addresses that correspond to its MDs 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 C-group addresses need not be distinct from either the P-group addresses or the Internet multicast addresses.

6. The PIM C-Instance and the MT

If a particular VRF is in a particular MD, the corresponding MT is treated by that VRF's VPN-specific PIM instances as a LAN interface. As a result, the PEs that are adjacent on the MT will generate and process PIM control packets, such as Hello, Join/Prune, and Assert. DF election occurs just as it would on an actual LAN interface.

6.1. PIM C-Instance Control Packets

The PIM protocol packets are sent to ALL-PIM-ROUTERS (224.0.0.13 for IPv4 or ff02::d for IPv6) in the context of that VRF, but when in transit in the provider network, they are encapsulated using the Default MDT group configured for that MD. This allows VPN-specific PIM routes to be extended from site to site without appearing in the P routers.

If a PIM C-Instance control packet is an IPv6 packet, its source address is the IPv4-mapped IPv6 address corresponding to the IPv4 address of the PE router sending the packet.

6.2. PIM C-Instance RPF Determination

Although the MT is treated as a PIM-enabled 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 MT 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 is not a VPN-IP route learned from MP-BGP as described in [RFC4364], or if that route's outgoing interface is one of the interfaces associated with the VRF, then ordinary PIM procedures for determining the RPF interface apply.

However, if the route matching the C-address 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 MT associated with the VPN-specific PIM instance.

Once PIM has determined that the RPF interface for a particular C-address is the MT, 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 MT.

The BGP "Connector" attribute is defined. Whenever a PE router distributes a VPN-IP address from a VRF that is part of an MD, it SHOULD distribute a Connector attribute along with it. The Connector attribute specifies the MDT address family, and its value is the IP address that the PE router is using as its source IP address for the multicast packets that are encapsulated and sent over the MT. When a PE has determined that the RPF interface for a particular C-address is the MT, it looks up the Connector attribute that was distributed along with the VPN-IP address corresponding to that C-address. The

value of this Connector attribute is considered to be the RPF adjacency for the C-address.

There are older implementations in which the Connector attribute is not present. In this case, as long as "BGP Next Hop" for the C-address is one of the PEs that is a PIM adjacency, then that PE is treated as the RPF adjacency for that C-address.

However, if the MD spans multiple Autonomous Systems, and an "option b" interconnect ([RFC4364], section 10) is used, the BGP Next Hop might not be a PIM adjacency, and the RPF check will not succeed unless the Connector attribute is used.

In [MVPN], the connector attribute is replaced by the "VRF Route Import Extended Community" attribute. The latter is a generalized version, but carries the same information as the connector attribute does; the encoding however is different.

The connector attribute is defined in the following sub-section.

6.2.1. Connector Attribute

The Connector Attribute is an optional transitive attribute. Its value field is formatted as follows:

```
  0                   1
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1|
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                               |
 |  IPv4 Address of PE           |
 |                               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

7. Data MDT: Optimizing Flooding

7.1. Limitation of Multicast Domain

While the procedure specified in the previous section requires the P routers to maintain multicast state, the amount of state is bounded by the number of supported VPNs. The P routers do NOT run any VPN-specific PIM instances.

In particular, the use of a single bidirectional tree per VPN scales well as the number of transmitters and receivers increases, but not so well as the amount of multicast traffic per VPN increases.

The multicast routing provided by this scheme is not optimal, in that a packet of a particular multicast group may be forwarded to PE routers that have no downstream receivers for that group, and hence which may need to discard the packet.

In the simplest configuration model, only the Default MDT group is configured for each MD. The result of the configuration is that all VPN multicast traffic, control or data, will be encapsulated and forwarded to all PE routers that are part of the MD. While this limits the number of multicast routing states the provider network has to maintain, it also requires PE routers to discard multicast C-packets if there are no receivers for those packets in the corresponding sites. In some cases, especially when the content involves high bandwidth but only a limited set of receivers, it is desirable that certain C-packets only travel to PE routers that do have receivers in the VPN to save bandwidth in the network and reduce load on the PE routers.

7.2. Signaling Data MDT Trees

A simple protocol is proposed to signal additional P-group addresses to encapsulate VPN traffic. These P-group addresses are called data MDT groups. The ingress PE router advertises a different P-group address (as opposed to always using the Default MDT group) to encapsulate VPN multicast traffic. Only the PE routers on the path to eventual receivers join the P-group, and therefore form an optimal multicast distribution tree in the service provider network for the VPN multicast traffic. These multicast distribution trees are called Data MDT trees because they do not carry PIM control packets exchanged by PE routers.

The following documents the procedures of the initiation and teardown of the Data MDT trees. The definition of the constants and timers can be found in section 8.

- The PE router connected to the source of the content initially uses the Default MDT group when forwarding the content to the MD.

- When one or more pre-configured conditions are met, it starts to periodically announce MDT Join TLV at the interval of [MDT_INTERVAL]. The MDT Join TLV is forwarded to all the PE routers in the MD.

  If a PE in a particular MD transmits a C-multicast data packet to the backbone, by transmitting it through an MD, every other PE in that MD will receive it. Any of those PEs that are not on a C-multicast distribution tree for the packet's C-multicast destination address (as determined by applying ordinary PIM procedures to the corresponding multicast VRF) will have to discard the packet.

  A commonly used condition is the bandwidth. When the VPN traffic exceeds certain threshold, it is more desirable to deliver the flow to the PE routers connected to receivers in order to optimize the performance of PE routers and the resource of the provider network. However, other conditions can also be devised and they are purely implementation specific.

- The MDT Join TLV is encapsulated in UDP.

  UDP over IPv4 is used if the multicast stream being assigned to a data-MDT is an IPv4 stream. In this case the UDP datagram is addressed to ALL-PIM-ROUTERS (224.0.0.13).

  UDP over IPv6 is used if the multicast stream being assigned to a data-MDT is an IPv6 stream. In this case the UDP datagram is addressed to ALL-PIM-ROUTERS (ff02::d).

  The destination UDP port is 3232.

  The UDP datagram is sent on the Default MDT. This allows all PE routers to receive the information. Any MDT Join that is not received over a Default MDT MUST be dropped.

- Upon receiving MDT Join TLV, PE routers connected to receivers will join the Data MDT group announced by the MDT Join TLV in the global table. When the Data MDT group is in PIM-SM or bidirectional PIM mode, the PE routers build a shared tree toward the RP. When the data MDT group is setup using PIM-SSM, the PE routers build a source tree toward the PE router that is advertising the MDT Join TLV. The IP address of the source address is the same as the source IP address used in the IP packet advertising the MDT Join TLV.

PE routers that are not connected to receivers may wish to cache the states in order to reduce the delay when a receiver comes up in the future.

- After [MDT_DATA_DELAY], the PE router connected to the source starts encapsulating traffic using the Data MDT group.

- When the pre-configured conditions are no longer met, e.g. the traffic stops, the PE router connected to the source stops announcing MDT Join TLV.

- If the MDT Join TLV is not received over [MDT_DATA_TIMEOUT], PE routers connected to the receivers just leave the Data MDT group in the global instance.

7.3. Use of SSM for Data MDTs

The use of Data MDTs requires that a set of multicast P-addresses be pre-allocated and dedicated for use as the destination addresses for the Data MDTs.

If SSM is used to set up the Data MDTs, then each MD needs to be assigned a set of these of multicast P-addresses. Each VRF in the MD needs to be configured with this set (i.e., all VRFs in the MD are configured with the same set). If there are n addresses in this set, then each PE in the MD can be the source of n Data MDTs in that MD.

If SSM is not used for setting up Data MDTs, then each VRF needs to be configured with a unique set of multicast P-addresses; two VRFs in the same MD cannot be configured with the same set of addresses. This requires the pre-allocation of many more multicast P-addresses, and the need to configure a different set for each VRF greatly complicates the operations and management. Therefore the use of SSM for Data MDTs is very strongly recommended.

8. Packet Formats and Constants

8.1. MDT TLV

"MDT TLV" has the following format.

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

Type (8 bits):

the type of the MDT TLV. In this specification, types 1 and 4 are defined.

Length (16 bits):

the total number of octets in the TLV for this type, including both the Type and Length field.

Value (variable length):

the content of the TLV.

8.2. MDT Join TLV for IPv4 streams

"MDT Join TLV for IPv4 streams" has the following format.

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

Type (8 bits):

Must be set to 1.

Length (16 bits):

Must be set to 16.

Reserved (8 bits):

for future use.

C-Source (32 bits):

the IPv4 address of the traffic source in the VPN.

C-Group (32 bits):

the IPv4 address of the multicast traffic destination address in the VPN.

P-Group (32 bits):

the IPv4 group address that the PE router is going to use to encapsulate the flow (C-Source, C-Group).

8.3. MDT Join TLV for IPv6 streams

"MDT Join TLV for IPv6 streams" has the following format.

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

Type (8 bits):

Must be set to 4.

Length (16 bits):

Must be set to 40.

Reserved (8 bits):

for future use.

C-Source (128 bits):

the IPv6 address of the traffic source in the VPN.

C-Group (128 bits):

the IPv6 address of the multicast traffic destination address in the VPN.

P-Group (32 bits):

the IPv4 group address that the PE router is going to use to encapsulate the flow (C-Source, C-Group).

8.4. Multiple MDT Join TLVs per Datagram

A single UDP datagram MAY carry multiple MDT Join TLVs, as many as can fit entirely within it. If there are multiple MDT Join TLVs in a UDP datagram, they MUST be of the same type. The end of the last MDT Join TLV (as determined by the MDT Join TLV length field) MUST coincide with the end of the UDP datagram, as determined by the UDP length field. When processing a received UDP datagram that contains one or more MDT Join TLVs, a router MUST be able to process all the MDT Join TLVs that fit into the datagram.

8.5. Constants

[MDT_DATA_DELAY]:

the interval before the PE router connected to the source to switch to the Data MDT group. The default value is 3 seconds.

[MDT_DATA_TIMEOUT]:

the interval before which the PE router connected to the receivers to time out MDT JOIN TLV received and leave the data MDT group. The default value is 3 minutes. This value must be consistent among PE routers.

[MDT_DATA_HOLDOWN]:

the interval before which the PE router will switch back to the Default MDT tree after it started encapsulating packets using the Data MDT group. This is used to avoid oscillation when traffic is bursty. The default value is 1 minute.

[MDT_INTERVAL]
the interval the source PE router uses to periodically send MDT_JOIN_TLV message. The default value is 60 seconds.

9. IANA Considerations

As no new codepoints allocations are requested, this section should be removed by the RFC Editor before publication.

The codepoint for the connector attribute is defined in IANA's registry of BGP attributes. The reference should be changed to refer to this document. On the IANA web page, the codepoint is denoted as "deprecated". This document does not request any change in that status. However, IANA may wish to note that there are a large number of deployments using this codepoint, and this is likely to be the case for a number of years.

The codepoint for MDT-SAFI is defined in IANA's registry of BGP SAFI assignments. The reference should be changed to refer to this document.

10. Security Considerations

[RFC4364] discusses in general the security considerations that pertain to when the RFC4364 type of VPN is deployed.

[PIMv2] discusses the security considerations that pertain to the use of PIM.

The security considerations of [RFC4023] and [RFC4797] apply whenever VPN traffic is carried through IP or GRE tunnels.

11. Acknowledgments

Major contributions to this work have been made by Dan Tappan and Tony Speakman.

The authors also wish to thank Arjen Boers, Robert Raszuk, Toerless Eckert and Ted Qian for their help and their ideas.

12. Normative References

[GRE2784] "Generic Routing Encapsulation (GRE)", Farinacci, Li, Hanks, Meyer, Traina, March 2000, RFC 2784

[PIMv2] "Protocol Independent Multicast - Sparse Mode (PIM-SM)", Fenner, Handley, Holbrook, Kouvelas, August 2006, RFC 4601

[PIM-ATTRIB] "The PIM Join Attribute Format" A. Boers, IJ. Wijnands, E. Rosen, November 2008, RFC 5384

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

[RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, February 2006, RFC 4364

13. Informative References

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

[BIDIR] "Bidirectional Protocol Independent Multicast", Handley, Kouvelas, Speakman, Vicisano, October 2007, RFC 5015

[DIFF2983] "Differentiated Services and Tunnels", Black, October 2000, RFC2983.

[GRE1701] "Generic Routing Encapsulation (GRE)", Farinacci, Li, Hanks, Traina, October 1994, RFC 1701

[GRE2890] "Key and Sequence Number Extensions to GRE", Dommety, September 2000, RFC 2890

[MVPN] "Multicast in MPLS/BGP IP VPNs", Rosen, Aggarwal, draft-ietf-l3vpn-2547bis-mcast-10.txt, January 2010

[SSM] "Source-Specific Multicast for IP", Holbrook, Cain, August 2006, RFC 4607

[RFC4023] " Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", T. Worster, Y. Rekhter, E. Rosen, Ed.. March 2005, RFC 4023

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

14. Authors' Addresses

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

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