Base Specification for Multicast in BGP/MPLS VPNs

draft-raggarwa-l3vpn-2547-mvpn-00.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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

This document describes the minimal set of procedures required to build multi-vendor inter-operable implementations of multicast for BGP/MPLS VPNs. It is based on prior specifications of multicast for BGP/MPLS VPN specifications that have been implemented and deployed. The procedures described herein require PIM-SM as the multicast routing protocol in the SP network.
Conventions used in this document

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 RFC-2119 [KEYWORDS].

Table of Contents

1. Motivation.............................................. 2
2. Terminology............................................ 3
3. Introduction........................................... 3
3.1. Efficient vs. Scalable Solution..................... 4
4. Basic Concepts........................................ 5
4.1. Multicast Domains.................................... 5
4.2. Provider and VPN PIM Instances....................... 5
4.3. Multicast Tunnels.................................... 6
5. Procedures.............................................. 6
5.1. Multicast VPN PIM-SM Join/Prune/Assert Propagation.. 6
5.1.1. C-Join/Prune/Assert RPF Interface.................... 6
5.1.2. C-Join/Prune/Assert PIM Neighbor Address......... 7
5.1.3. Switching from Shared to Source Specific MD Trees in the SP Network.................................. 7
5.2. Multicast VPN Data Forwarding........................ 8
5.3. Operation.............................................. 9
5.3.1. PIM Neighbor Discovery in a MD..................... 9
5.3.2. Handling a PIM-Join/Prune received from a CE...... 10
6. Inter-AS Considerations................................ 10
7. Security Considerations................................ 11
8. Acknowledgments....................................... 11
9. Normative References.................................. 11
10. Informative References................................ 12

1. Motivation

This document describes the minimal set of procedures required to build inter-operable implementations of multicast support for BGP/MPLS VPNs (MVPNs). It is based on prior multicast in BGP/MPLS VPN specifications [MVPN-6] that have been implemented and deployed. Procedures presented herein are not new. However the intent of this document is to clearly define the base set of procedures required to build inter-operable implementations of multicast support for BGP/MPLS VPNs.

This document requires PIM-SM as multicast routing protocol in the SP network.
This document does not preclude various optional optimizations of multicast support for BGP/MPLS VPNs - it assumes that procedures for such optimizations will be specified in separate documents.

2. Terminology

In addition to the terminology used in [2547] and [PIM-SM] this document introduces the following terms:

Multicast Domain (MD): A set of VRFs on different PEs, belonging to a given VPN, associated with interfaces that can send multicast traffic to each other.

Provider PIM Instance: PIM instance in the SP network

VPN PIM Instance: PIM instance in the VPN

P-Join: PIM Join message in the Provider PIM instance.

C-Join: PIM Join message in the VPN PIM instance.

Multicast Tunnel (MT): Tunnel created for each MD, in the provider PIM instance. The MT is used to carry multicast customer packets, both data and control, among the PE routers in a common MD.

3. Introduction

[2547] specifies a set of procedures which must be implemented for a SP to provide a unicast VPN service. [MVPN-6] describes various methods that can be used to extend [2547] to enable a SP to provide multicast service in a VPN. However it does not specify the minimal and the exact set of procedures required for inter-operability. This has lead to non inter-operable implementations.

This document specifies the minimal set of procedures required for an inter-operable solution that enable a SP to provide multicast service in a VPN. The procedures specified herein require a SP to use PIM-SM as the multicast routing protocol in the SP network. Use of other multicast routing protocols (PIM-SSM, PIM-BIDIR, PIM-DM) in the SP network for the purpose of providing multicast service in a VPN is optional and is not part of the minimal set of required procedures discussed here.

Within a VPN, any of PIM-SM, PIM-DM, PIM-SSM, PIM-BIDIR can be used as the multicast routing protocol.
3.1. Efficient vs. Scalable Solution

In the context of this document we define "efficient multicast routing" as follows. When a PE router receives a multicast data packet of a particular multicast group from a CE router, the packet must reach every other PE router which is on the path to a receiver of that group. It should not reach any PEs that aren't on the path to a receiver. It should not be unnecessarily replicated. Efficient multicast routing requires a source-tree for the multicast group, which would mean that the P routers would have to maintain state for each transmitter of each multicast group in each VPN.

Note that efficient multicast routing, as defined above, requires potentially an unbounded amount of state in the SP routers, since the SP has no control on 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.

However, even if the amount of state was possible, the same multicast group address can be used in multiple VPNs to carry different traffic. This traffic cannot be mixed or delivered to the wrong VPN. This dictates the need for a tunneling mechanism to keep the traffic with the same destination IP address separated for each VPN.

One option 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. Hence this solution uses ingress replication but requires minimal state in the SP network.

This documents specifies a solution that aims at achieving a compromise between the the amount of multicast state required to be maintained in the SP network and the efficiency of multicast routing. It uses a PIM-SM shared tree multicast tunnel for each VPN. That allows it to bound the total amount of multicast state in the SP network solely by the number of VPNs. PIM-SM provides a way to improve efficiency of multicast routing (albeit at the cost of additional multicast state in the SP network) by switching from the shared tree to source trees, rooted at each PE. The PIM-SM source tree is shared by all the multicast sources within a VPN that are behind that PE.
4. Basic Concepts

This section describes terminology used in the remainder of this document.

4.1. Multicast Domains

A "Multicast Domain (MD)" is a set of VRFs on different PEs, belonging to the same VPN, associated with interfaces that can send multicast traffic to each other. Each MD is assigned a MD P-Group address. It is required that each MD be assigned a unique address across all the domains that are part of the MVPN service. Each VRF has its own multicast routing table. When a multicast control packet is received from a particular CE device, PIM RPF lookup and PIM Join propagation is done in the associated VRF. Similarly, when a multicast data packet is received from a particular CE device, multicast data forwarding is done in the associated VRF. The goal of this is to send the multicast control or data packet to all other VRFs in that MD. This is achieved by building one or more multicast distribution tree for a given MD in the SP network.

4.2. Provider and VPN PIM Instances

Each PE router runs an instance of PIM per VRF. In each VRF instance of PIM, 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. These PIM instances are referred as "VPN-specific PIM instances". These PIM instances can support any flavor of PIM, for instance PIM-SM or PIM-DM.

Each PE router also runs a "provider-wide" instance of PIM-SM, in which it has a PIM adjacency with each of its IGP neighbors (i.e., with P and directly connected PE routers), but NOT with any CE routers. The provider PIM instance MUST support PIM-SM.

In order to help refer to provider-wide PIM instance and to VPN-specific PIM instance, the prefixes "P-" and "C-" are used respectively. Thus a P-Join would be a PIM Join which is processed by the provider-wide PIM-SM instance, and a C-Join would be a PIM Join which is processed by a VPN-specific PIM instance. A P-group address would be a group address in the SP’s address space.
4.3. Multicast Tunnels

Each MD is assigned a unique multicast P-group address across the provider network. As part of normal PIM-SM procedures the provider wide PIM-SM instance has to know the RP for each P-group.

Each PE sends the traffic for an MD encapsulated to the P-group for that MD. This is called a Multicast Tunnel (MT). The MT is treated like an interface and normal PIM Hellos are sent through the tunnel. This leads to all PEs discovering each other as PIM neighbors over that MT interface in the given MD. The details are described in section 5.3.1.

The MT is used to carry multicast C-packets, both data and control packets, among the PE routers in a common MD. Data forwarding is described in section 5.2.

When a packet is received by a PE from another router in the SP network, the receiving PE can determine the MT (and hence the MD) from which the packet was received as the destination address of the packet is the MD P-group address. The decapsulated packet is then passed to the corresponding Multicast VRF and VPN-specific PIM instance for further processing.

5. Procedures

5.1. Multicast VPN PIM-SM Join/Prune/Assert Propagation

For a VRF in a particular MD, the corresponding MT is treated by that VRF’s VPN-specific PIM instance as an interface. The PEs which are adjacent on the MT must execute the PIM interface procedures, including the generation and processing of Assert packets. The VPN PIM instances can send C-Join messages through the MT. These messages are received by all PEs in the MD. This allows VPN-specific PIM Join/Prune messages to be extended from site to site, without appearing in the P routers. Note that a C-Join message carries the address of the neighbor for which the C-Join message is meant. This message is processed by the corresponding PIM neighbor on the MT interface.

5.1.1. C-Join/Prune/Assert RPF Interface

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 [2547], 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.

5.1.2. C-Join/Prune/Assert PIM Neighbor Address

Determination of the C-Join PIM neighbor address i.e. the RPF neighbor address needs to be further explained. This depends on the procedure used to assign an address to the MT interface. The address of this interface MUST be the BGP next-hop address of the unicast VPN routes advertised by the MD VRF. This will typically be a PE loopback address in the provider address space.

To determine the C-Join neighbor address, the PE does a route lookup on the C-Source address. This address is a VPN unicast route learnt from the PE sitting in front of the multicast source. The route lookup results in the BGP next-hop of the C-source VPN unicast route. This BGP next-hop is the neighbor address to use while sending the PIM-Join.

5.1.3. Switching from Shared to Source Specific MD Trees in the SP Network

By default the generation of VPN instance PIM control messages on a MT by a PE results in all the other PEs in that MD to switch from the shared MD tree in the SP network to a source specific MD tree rooted at the PE that is generating the control messages. This is the case even though there may not be any multicast sources transmitting in that given VRF on that PE. This results in a different source specific tree for a given MD for each PE that belongs to that MD.

To reduce the number of source specific trees in the SP network an implementation SHOULD provide the following knobs to control switching from the shared MD tree in the SP network:

a) A knob on the RP so that it sends the source specific MD P-Group Join to the source PE (after receiving Register messages) only after the multicast traffic being received for that MD from the source PE exceeds a certain threshold.
b) A knob on a PE so that it sends the source specific MD P-Group Join to the source PE only after the multicast traffic being received for that MD from the source PE exceeds a certain threshold.

Note that this is a local implementation choice and does not impact inter-operability.

5.2. Multicast VPN Data Forwarding

A PE in a particular MD transmits a C-multicast data packet through the SP network by transmitting it through the MT corresponding to the MD. The MT is installed as the outgoing interface for the C-multicast data packets when C-Join messages corresponding to the data packet’s source and group are received on the MT interface.

An implementation MUST support GRE encapsulation.

The following diagram shows the progression of the packet using GRE encapsulation as it enters and leaves the service provider network.

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

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

The destination address in the P-IP header is the MD address corresponding to the MT. This enables the P routers to forward this packet along the multicast distribution tree corresponding to the MD.

The IPv4 Protocol Number field in the P-IP Header must be set to GRE (47).

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 which 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.

5.3. Operation

5.3.1. PIM Neighbor Discovery in a MD

MTs are described in section 4.3. A MT is a pseudo interface unique to a VPN that can be created when PIM-SM initializes. Unicast routing is not run over this interface. This interface is used to carry PIM control and multicast data traffic. PIM sends Hello messages on the MT interfaces using the PE loopback address in the provider address space as the source address. Each PE router needs to join the MD P-Groups associated with all the MDs it belongs to. Discovery of remote PEs in the same VPN is done by sending PIM Hello messages over MT tunnels as follows:

- When PIM-SM initializes in a MD, the PE originates a PIM Join message for the MD P-Group address towards the RP in the SP space. This is done for each MD that is configured on the PE.

- Since an MT interface belongs to a VPN, sending a Hello message on this interface does the following:
  - The PIM Hello message has the source address of PE’s loopback interface in the SP address space and the destination of ALL-PIM-ROUTERS group.
  - This PIM Hello gets encapsulated in a GRE header with the source address as the PE’s loopback interface and the destination as the MD P-Group address. After the encapsulation, the original PIM-SM Hello travels as the data packet in a PIM-SM Register towards the SP RP.
  - RP in the SP network knows about all the receivers (the PEs) because of the earlier PIM Join for the MD P-Group address that it received from all the PEs when they initialized. So, when the RP receives the above PIM-SM register, it decapsulates it and forwards it down to all the PEs. So, all the remote PEs (including the one who sent the packet) receives this data packet which has the source address of the originating PE.
  - This PIM Hello packet originated within the VRF travels as the data packet (due to encapsulation) in the SP network towards the RP.
  - The above procedure is repeated on all the PEs. Hence, all the PEs receive each other’s data packets which contain PIM Hello messages and discover one another. PEs can decide to send the source Join directly to the remote PEs at this point.
5.3.2. Handling a PIM-Join/Prune received from a CE

When a PE receives a PIM Join/Prune for a group in the VPN space in its VRF, it processes this message exactly as per PIM procedures. Then it forwards this message to the upstream PIM neighbor in the path to the VPN-RP or the VPN source. The neighbor address in the PIM message is set as described in section 5.1.2. The PIM message is encapsulated in a GRE header with the source address as the PE’s loopback interface and the destination as the MD P-Group address.

The original PIM control message in the VPN instance PIM now becomes a data packet within the SP space and gets sent either as the PIM-SM Register to the SP-RP or natively through the SP network. It is sent to all PEs that had sent a PIM-SM Join for the MD P-Group address earlier. The packet finally reaches all the PEs in the MD. The PE for which the "upstream neighbor address" matches forwards the original PIM control message towards the RP or source behind the CE.

6. Inter-AS Considerations

[2547] describes three methods for creating inter-AS VPNs:

Option A: VRF-to-VRF connections at the AS border routers.

Option B: EBGP redistribution of labeled VPN-IP routes from AS to neighboring AS.

Option C: Multihop EBGP distribution of labeled VPN-IP routes between source and destination ASes, with EBGP redistribution of labeled IP routes from AS to neighboring AS.

The mechanisms described in this draft support multi-AS VPN multicast when either Option A or C is used. However, they are not sufficient when Option B is used. This is because the BGP Next-hop of the VPN routes is re-written in Option B at the ASBRs. As a result of this the PIM neighbor and the BGP next-hop do not match and the procedures described in section 5.1.2 cannot be used for determining the RPF neighbor. Solution to this issue is outside the scope of this document.

It is possible that Option C is used with a ‘BGP free SP network’. In this case the P routers in one AS do not know how to route to the PE addresses in another AS. As a result of this they will not be able to forward the P-Join messages towards the egress PE. Solution to this issue is outside the scope of this document.

For inter-AS VPNs that require multicast service if the involved ASs
are all under a single provider, these ASs can share RPs, and MSDP is not required. Even if the ASs are under control of multiple service providers, the level of cooperation required to offer even plain unicast 2547 VPN service is high enough, which means that one more issue (ownership of RP) may not be a significant addition to what is already required. And if that is the case, the providers can share RPs, and MSDP is not required. If each provider insists on having its own local RP, MSDP can be used between the RPs that belong to the different providers. However, in many cases, this will not be necessary.

If there are inter-AS VPNs that span multiple SPs and require multicast service, then MDs (and MTs) for these VPNs will cross provider boundaries. The assignment of the multicast group addresses associated with the MDs for such VPNs must then be coordinated upon by the providers.

7. Security Considerations

Security considerations discussed in [2547] and [PIM-SM] apply to this document.

8. Acknowledgment

As mentioned earlier, this draft is based on [MVPN-6]. The authors of [MVPN-6] are Eric Rosen, Yiqun Cai, Dan Tappan, IJsbrand Wijnands, Yakov Rekhter and Dino Farinacci. We would like to thank them for their tremendous contribution to this technology.

We would also like to thank Paras Trivedi for his detailed review of this document.

9. Normative References


[RFC2119] "Key words for use in RFCs to Indicate Requirement
10. Informative References


Author Information

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

Anil Lohiya
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: alohiya@juniper.net

Tom Pusateri
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: pusateri@juniper.net

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

The IETF takes no position regarding the validity or scope of any intellectual property 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; neither does it represent that it has made any effort to identify any such rights. Information on the IETF’s procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication 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 implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

Full Copyright Notice

Copyright (C) The Internet Society (2003). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS 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."

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.