MVPN: Extranets, Anycast-Sources, ‘Hub & Spoke’, with PIM Control Plane

draft-rosen-l3vpn-mvpn-extranet-03.txt

Abstract

This document provides the specification for using the PIM control plane of to provide Multicast Virtual Private Network support for extranets, for anycast sources, and for "hub and spoke” configurations.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and 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.
Copyright and License Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1 Specification of requirements .......................... 3
2 Introduction ........................................... 3
3 Extranets using PIM as the MVPN Control Plane ....... 3
  3.1 Default PMSI ....................................... 4
  3.2 Red method ........................................ 4
  3.2.1 Control Plane RPF Check ....................... 5
  3.2.2 Data Plane RPF Check ............................ 5
  3.3 Blue method ...................................... 5
  3.4 Binding Specific Extranet C-Flows to S-PMSIs ...... 6
  3.5 Two VRFs on One PE ................................ 7
4 Supporting Anycast Sources with PIM Control Plane .... 7
5 Hub and Spoke MVPNs ................................ 8
  5.1 Unicast Hub and Spoke VPNs ......................... 8
  5.2 Multicast Hub and Spoke VPNs ...................... 10
6 IANA Considerations ................................... 13
7 Security Considerations ................................. 13
8 Acknowledgments ........................................ 13
9 Authors’ Addresses .................................... 13
10 Normative References .................................. 14
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 document extends the PIM control plane of [MVPN] to provide support for the following features:

- MVPN Extranets

  In an MVPN "extranet", the transmitter of a multicast traffic flow is in a different VPN than the receivers. Additional procedures are defined to determine how the traffic is associated with a particular MI-PMSI [MVPN] or MS-PMSI [MVPN_MSPMSI], and how the RPF checks are done.

- Support for Anycast Sources

- Support for "Hub and Spoke" VPNs

3. Extranets using PIM as the MVPN Control Plane

Suppose there are two VPNs. VPN1 consists of a set of VRFs, each of which has been configured with RT1 as it export and import Route Target. VPN2 consists of a set of VRFs, each of which has been configured with RT2 as it export and import Route Target. For convenience, we will use the term "blue" instead of "RT1" and the term "red" instead of "RT2". Thus we will call VPN1 the "blue VPN" and VPN2 the "red VPN". Similarly, the blue VPN consists of a number of "blue sites" containing "blue systems"; these sites are attached to PEs via VRF interfaces that are associated with "blue VRFs".

We want to create an MVPN extranet in which blue receivers can join multicast groups whose sources and/or RPs are red.

The first step is to ensure that the blue VRFs (or the subset of blue VRFs whose attached sites are allowed to receive multicasts from red sources) import routes to the red sources. This is done as follows:

- The red VRFs are configured so that the subset of red routes that are to be part of the extranet are exported with a seconds RT value (call it RT3), as well as with RT2. For convenience, we will call RT3 "violet".
- The blue VRFs are configured so that they import violet routes as well as blue routes.

There are two different methods of providing the extranets, which will shall call the "red method" and the "blue method". (Remember that the red VPN contains the transmitter, and the blue VPN contains the receivers.)

This document assumes that in the case of non-SSM extranet multicast groups, the mapping between a group address and an RP is pre-configured in the PEs.

This document does not provide support for bidirectional C-trees in extranets.

3.1. Default PMSI

Some of the procedures subsequently specified in this section are largely independent of whether PIM is used with (a) an MI-PMSI or (b) with an MS-PMSI that has been bound to the double wildcard. We will use the term "default PMSI" as a general term to mean either (a) or (b), depending upon which technique is actually being used in a given network.

3.2. Red method

In the "red method", extranet multicasts are carried by default in the default PMSI of the red VPN, which we will of course call the "red PMSI".

To use this method, blue VRFs must be configured to import "red" I-PMSI A-D routes and red S-PMSI A-D routes. If MI-PMSIs are being used, the blue VRFs must immediately join the P-tunnels specified in the red I-PMSI A-D routes. If MS-PMSIs are being used, a blue VRF need not join the MS-PMSI P-tunnel rooted at a particular PE unless a PIM Join needs to be sent to that PE.

The PIM C-instance associated with a blue VRF will treat the red and blue default PMSIs as two different PIM interfaces.

The blue VRFs must also be configured to "associate" violet unicast routes with the red default PMSI. What this means is that the red default PMSI will be considered to be the RPF interface for the violet unicast routes. The RPF interface for the blue unicast routes remains, as usual, the blue default PMSI.
All that remains to be specified is how the control plane and data plane RPF checks are done. Apart from these MVPN-specific procedures for the RPF check, ordinary PIM procedures are used.

3.2.1. Control Plane RPF Check

Suppose a PE receives a PIM Join(S,G) from a CE, over a VRF interface that is associated with a blue VRF. The PE does the RPF check for S by looking up S in the blue VRF. If the route matching S is a blue route (i.e., carries the blue RT but not the violet RT), then a Join is sent over the blue default PMSI. However, if the route matching S is a violet route (i.e., carries the violet RT), a Join is sent over the red default PMSI.

If the PE receives a PIM Join(*,G) from a CE, the RPF check is done against the address of the corresponding RP; otherwise the procedure is the same.

3.2.2. Data Plane RPF Check

Suppose a red default PMSI has been associated with a blue VRF, as specified above, and an (S,G) multicast data packet is received from the red default PMSI. Then S is looked up in the (blue) VRF. If it matches a violet route, the packet is forwarded normally. However, if it matches a blue route, the packet is discarded as having failed the RPF check.

This prevents the blue sites from receiving packets from red transmitters, except in the case where routes to the red receivers have been explicitly imported into the blue VRF.

3.3. Blue method

In the "blue method", extranet multicasts are carried by default in the default PMSI of the blue VPN.

In the blue method, the red VRFs must be configured to import "blue" I-PMSI and S-PMSI A-D routes. If MI-PMSIs are being used the P-tunnels specified therein must be joined immediately. If MS-PMSIs are being used, the P-tunnels need not be joined unless and until it is necessary to send a PIM Join to the root of the P-tunnel.

The PIM C-instance associated with a red VRF will treat the red default PMSI and the blue default PMSI as two different PIM interfaces.
PIM Joins from blue receivers are then received at the red VRF over the blue PMSI, whereas PIM Joins from red receivers are received at the red VRF over the red PMSI. As a result, PIM may add one or the other or both PMSIs to a particular multicast tree's olist.

In this method, the blue VRFs are associated with only one default PMSI, so the RPF check for both blue and violet sources (and RPs) always resolves to that PMSI. Hence the special RPF check procedures of the red method are not necessary. However, a PE with a red VRF may need to transmit multicast traffic on more than one MI-PMSI.

Note that since the data plane RPF check of section 3.2.2 is not needed, one does not really need a "violet" RT value. Rather, one may simply configure certain routes from the red VRF to be exported with both the red and the blue RTs.

3.4. Binding Specific Extranet C-Flows to S-PMSIs

If the procedure of [MVPN] section 7.4.2 is used, the S-PMSI Join message MUST be sent on whatever default PMSI or default PMSIs are used to carry the C-flow identified in the message.

If the procedure of [MVPN] section 7.4.1 is used, then procedures differ slightly depending upon whether the red method or the blue method is in use.

If the red method is in use, and if a C-flow whose target source is exported from a red VRF is bound to an S-PMSI, then the S-PMSI A-D route that specifies the binding must carry both the red RT and the violet RT. Blue VRFs must be configured to import the violet S-PMSI A-D routes.

If the blue method is in use, and if a C-flow whose target source is exported from a red VRF is bound to an S-PMSI, then the S-PMSI A-D route that specifies the binding:

- must carry the red RT if the C-flow has any receivers on the red default PMSI, and
- must carry the blue RT if the C-flow has any receivers on the blue default PMSI.
3.5. Two VRFs on One PE

It is possible that a red VRF and a blue VRF will exist on the same PE. Then by the above procedures, one of these VRFs will need to join a PMSI that it can use for sending control packets to and receiving data packets from the other. However, the protocol used to construct the P-tunnels instantiating the PMSI may not provide a mechanism by which a given PE can join a P-tunnel of which it is the root. In this case, the PE implementation MUST support a local function whereby a given VRF, say VRF1, can "join" a P-tunnel whose root is another VRF, say VRF2, on the same PE. The PE MUST also support a local function whereby packets can be transmitted from one VRF to another just as if the VRFs had been on separate PEs.

4. Supporting Anycast Sources with PIM Control Plane

Suppose that some customer site contains router C-R1 and some other customer site in the same VPN contains router C-R2. And that each sends a PIM Join(C-S,C-G) messages towards C-S. Ordinarily, the result will be to create a single C-tree whose root is C-S and whose leaves include C-R1 and C-R2.

However, in some deployment scenarios, C-S may be an anycast address that belongs to two or more different sources, say C-S1 and C-S2. Let’s suppose that these two sources attach to the VPN backbone through two different PEs, and let’s further suppose that C-S1 is "close" to C-R1, and C-S2 is "close" to C-R2. Then even though both C-R1 and C-R2 send Join(S,G) messages, what is really desired is to create two C-trees, one rooted at C-S1 (with C-R1 as a leaf) and one rooted at C-S2 (with C-R2 as a leaf).

If the data traffic traveling along both C-trees is carried on a single MI-PMSI, it is important that a (C-S,C-G) data packet is forwarded towards C-R1 only if the packet is actually traveling on the C-tree rooted at C-S1, and not on the C-tree rooted as C-S2.

To ensure this, if a particular MVPN is providing anycast service, its PEs MUST use the procedure described in section 9.1.1 of [MVPN], and MUST NOT use the procedures described in sections 9.1.2 and 9.1.3 of [MVPN].

This also enables the use of C-RPs that have anycast addresses.

Furthermore, if anycast source support is provided for a particular multicast group C-G, all PEs MUST execute the procedure described in section 4.2.1 of [PIM], and MUST act as if SwitchToSPTDesired(S,G) (defined in [PIM] section 4.2.1) is true when the first (S,G) packet
(from any PE) is received. (This procedure MUST be executed by each PE even if the PE is not the "last hop" of the C-tree.) This will ensure that each PE receives and forwards (C-S,C-G) traffic from the appropriate source C-tree, even if PE has received only Join(C-* ,C-G) messages but not Join(C-S,C-G) messages from its directly attached CEs.

5. Hub and Spoke MVPNs

The Layer 3 Virtual Private Network (L3VPN) technology of [RFC4364] generally provides an "any-to-any" network service, where any system at one site of a VPN can send traffic to and receive traffic from a system at any other site. Or more precisely, nothing in the procedures governing the distribution of routing information in the VPN prevents any-to-any communication.

In some deployments, however, it has been convenient to distinguish between two kinds of VPN site, the "hub site" and the "spoke sites". In this section, we first describe how the "hub and spoke" configuration affects the distribution of unicast routing. We then specify a means of providing multicast VPN service in the hub and spoke configuration.

5.1. Unicast Hub and Spoke VPNs

In a unicast hub and spoke VPN:

- any system in a hub site can send traffic to and receive traffic from any other system in a hub site;
- any system in a hub site can send traffic to and receive traffic from any system in a spoke site;
- any system in a spoke site can send traffic to and receive traffic from any system in a hub site;
- a system in one spoke site cannot send traffic to and cannot receive traffic from a system in a different spoke site.

Using the technology of [RFC4364], it is possible to create this sort of "hub and spoke" VPN by suitable restricting the flow of routing information among the sites. One way to construct a hub and spoke VPN is as follows:
- Within a given VPN, every site is denoted as either a hub site or a spoke site.

- On a given PE, every spoke site is attached to a distinct VRF (i.e., all interfaces of that VRF lead to the same spoke site). We will call these "Spoke VRFs".

- On a given PE, any number of hub sites can be attached to a single "Hub VRF".

- Each Hub VRF is configured with an export-RT that we shall call "Hub_Route", and with a pair of import-RTs, one of which is "Hub_Route", and the other of which we shall call "Spoke_Route". (Of course, each hub and spoke VPN has its unique Hub_Route RT and its unique Spoke_Route RT.)

- Each Spoke VRF is configured with export-RT "Spoke_Route" and import-RT "Hub_Route".

With this configuration, the Spoke VRFs will contain only routes to systems at hub sites, whereas the Hub VRFs will contain routes to systems at both hub and spoke sites. Even if two spoke sites attach to the same PE, they cannot communicate directly, because they are associated with different VRFs, and their respective VRFs do not import each others’ routes. (There are implementation techniques that can eliminate the need to configure a separate VRF for each spoke site on a PE, but these are out of scope of this document.)

There are several different variations on this theme. For example, in a particular VPN, spoke-to-spoke communication may be allowed, but only if the spoke-to-spoke traffic first enters a hub site. Some system at the hub site would be responsible for "turning the traffic around", i.e., sending it back to VPN backbone for delivery to the target spoke site. This can be useful if the "turnaround system" at the hub site performs some sort of inspection of the spoke-to-spoke traffic and then applies authorization policies of some sort. To provide this sort of Hub and Spoke VPN:

- The total set of routes exported by the Hub VRFs must include routes that "summarize" all the routes exported by the Spoke VRFs. For example, one or more Hub VRFs may export a default route. In the Hub VRFs, each of these summary routes will have one of the VRF interfaces as its next hop interface.

- When such a summary route is exported as a VPN-IP route, it MUST be advertised with a label for which the Next Hop Label Forwarding Entry (see section 3.10 of [RFC3031]) specifies on of the VRF interfaces as the next hop interface.
In this scenario, if a PE receives traffic from a spoke site, and the IP destination address of that traffic is a system in another spoke site, the traffic will be tunneled to a PE that attaches to a hub, and then sent over one of the Hub VRF’s "VRF interfaces", i.e., sent to a Hub CE router. The Hub PE, when it receives the tunneled packet, does not look up the packet’s IP destination address in the Hub VRF, but rather forwards based on the MPLS label. If the Hub CE decides (possibly after inspecting the packet and authorizing the transmission) to "turn the packet around", sending it back to the PE, the PE will look up the IP destination address in the Hub VRF, find that it matches one of the routes imported from a spoke VRF, and tunnel the packet to the PE attaches to the corresponding spoke site.

Note that setting up a hub and spoke VPN is just a matter of proper configuration. There are no protocol differences between a Hub and Spoke VPN and any other kind of RFC 4364 VPN.

5.2. Multicast Hub and Spoke VPNs

Sometimes it is necessary to support multicast service over a Hub and Spoke VPN. In this scenario, it is generally desired to provide an MVPN service with the following properties:

- A receiver at a hub site may receive multicast traffic from a transmitter at a spoke site (including the case where the RP is at a spoke site)

- A receiver at a spoke site may receive multicast traffic from a transmitter at a hub site (including the case where the RP is at a hub site)

- A receiver at a spoke site must not be allowed to join a shared tree (i.e., a (C-*,C-G) tree whose root (i.e., the RP) is at a different spoke site.

- A receiver at a spoke site must not be allowed to receive multicast traffic from a transmitter at a different spoke site, except possibly in the case where the traffic traverses a hub site on its path from one spoke site to the other.

This type of MVPN service can be provided by using a variation of the "PIM over MS-PMSI" model described in [MVPN_MSPMSI]. In this model, each PE advertises an MS-PMSI for each VRF. If these advertisements are made using BGP S-PMSI A-D routes, the A-D route originating at a Hub VRF carries the "Hub_Route" RT; an A-D route originating at a spoke VRF carries the "Spoke_Route" RT. That is, the S-PMSI A-D routes originating at a given VRF carry the same RT as the unicast
routes originating at that VRF.

To support Hub and Spoke functionality, the MS-PMSIs originating at the spoke VRFS may all specify the same P-tunnel identifier. Similarly, the MS-PMSIs originating at the hub VRFS may all specify the same P-tunnel identifier, but this must be a different P-tunnel identifier than the one specified for the MS-PMSIs originating from the spoke VRFS. In this case, it is convenient to speak of the Hub and Spoke infrastructure as consisting of two MS-PMSIs, a "spoke-rooted" MS-PMSI and a "hub-rooted" MS-PMSI.

As discussed in [MVPN_MSPMSI], it is possible to instantiate an MS-PMSI as a set of PIM-SM trees. This means of instantiation can be useful in Hub and Spoke scenarios when GRE/PIM tunneling is used. In this case, for a given VPN, there may be a single sparse mode group address associated with the MS-PMSIs rooted at the spoke VRFS, and a second sparse mode group address associated with the MS-PMSIs rooted at the hub VRFS. The result is the creation of two distinct sets of P-tunnels for the VPN, one set used to carry data traffic from spoke sites to hub sites (and PIM control traffic in the opposite direction), and the other set used to carry data traffic from hub sites to spoke sites (and PIM control traffic in the opposite direction).

Suppose that a spoke VRF and a hub VRF are on the same PE, and that an MS-PMSI advertisement exported by one of those VRFS is imported by the other. The PE implementation MUST support a local function whereby the importing VRF can "join" the MS-PMSI exported by the other VRF, and MUST support a local function whereby packets transmitted from one VRF onto the MS-PMSI are received by the other VRF (if and only if the latter VRF has joined the MS-PMSI exported by the former).

Since spoke VRFS do not import each others’ S-PMSI A-D routes, and do not import each other’s unicast routes, and since there is no MI-PMSI, there is no way for a C-Join to be transmitted directly from one spoke VRF to another. If a CE at a spoke site sends a Join(S,G) to its PE, the PE will forward it on the hub-rooted MS-PMSI advertised by the hub site that is the BGP next hop for S; no spoke VRF can receive PIM control packets on that MS-PMSI.

In this scheme, each hub VRF joins two MS-PMSIs, the one spoke-rooted MS-PMSI and the hub-rooted MS-PMSI. Normal PIM procedures would see these as two PIM interfaces. If a hub VRF at PE1 receives a Join(S,G) from the hub-rooted MS-PMSI, where S is at a spoke site, normal PIM/MVPN procedures would cause PE1 to send a Join(S,G) over the spoke-rooted PMSI towards a PE that attaches to S’s site. If these procedures are followed, a receiver at a spoke site could get
multicast data from a different spoke site; the data would get "turned around" at a PE that attaches to a hub site. Since this violates the requirements as stated above, a PE providing Hub and Spoke MVPN service MUST NOT send a Join message on one MS-PMSI as a result of having received a Join message over another.

Note that this does not completely prevent a receiver in a spoke site from being able to receive multicast data from a transmitter in a different spoke site. This can happen in the following situation:

- A receiver R1 at a hub site, Site1, joins a (C-*,C-G) tree,
- The RP for (C-*,C-G) is at a different hub site, Site2,
- A system S1 at a spoke site, Site3, transmits multicast traffic to group C-G.

In this situation, the PE attached to Site 2 may need to turn the (S1,G) data around and transmit it on the hub-rooted PMSI, so that R1 can receive it. This also allows spoke sites to receive it. However, turnaround at a PE is never a desirable traffic pattern, and implementations are NOT required to support it. An alternative procedure which enables R1 to receive the (S1,G) traffic is for the PE at Site3 to generate a BGP Source Active A-D route, carrying the "spoke route" RT, when it receives a Join(S1,G) on the spoke-rooted MS-PMSI. This route would be withdrawn when the PE no longer has the corresponding (S1,G) state. The PE attached to Site1 will see this SA route, and if it has (*,G) state, will then generate (S1,G) state and expect to receive (S1,G) traffic from the spoke-rooted MS-PMSI.

Another situation in which a receiver in a spoke site may be able to receive multicast data from a transmitter in a different spoke site is the following:

- A receiver R1 at a spoke site, Site1, joins a (C-*,C-G) tree,
- The RP for (C-*,C-G) is at a hub site
- A system S2 at a different spoke site, Site2, transmits multicast traffic to group C-G,
- The hub site containing the RP is multiply connected to the SP backbone,
- The best path from R1 to the RP enters the RP’s hub site via a particular PE-CE link, link1,
- The best path from S2 to the RP enters the RP’s hub site via a different PE-CE link, link2.

In this case, it is possible for multicast data traffic to travel from S2 to link1 to the RP to link2 to R1. In some situations, a SP and its customer may wish to explicitly set up this scenario in order to allow spoke sites to receive selected multicast traffic from other spoke sites.

The procedures described in this section are compatible with the procedures of section 4.

6. IANA Considerations

This document does not specify any actions for IANA.

7. Security Considerations

There are no additional security considerations beyond those of [MVPN] and [MVPN-BGP].

8. Acknowledgments

The authors wish to thank DP Ayadevara, Rayen Mohanty, Maria Napierala, and Karthik Subramanian.

9. Authors’ Addresses

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

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
E-mail: erosen@cisco.com
10. Normative References


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