L3VPN Working Group
Internet Draft
Intended Status: Proposed Standard
Expires: August 6, 2012

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

Maria Napierala
AT&T

Arjen Boers

February 6, 2012

MVPN: Optimized use of PIM via MS-PMSIs

draft-rosen-l3vpn-mvpn-mspmsi-10.txt

Abstract

This document specifies an optimized method that a service provider can use to provide MVPN service when using PIM as the MVPN control protocol. As in prior MVPN methods, PIM control messages are sent over multicast tunnels through the provider network. However, unlike older MVPN methods, the tunnels are only created if they are needed to carry multicast data traffic; no tunnels are used only for control traffic.

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.

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

[MVPN] specifies how to run PIM [PIM] as the multicast routing protocol of a particular MVPN, by running it over an MI-PMSI for that MVPN. In this specification, we provide a specification for running PIM over an MS-PMSI. When PIM is run over an MI-PMSI, there may need to be P-tunnels that only carry PIM messages, but do not carry multicast data. However, when PIM is run over an MS-PMSI, there is never any need to create a P-tunnel just for control messages; the only P-tunnels needed are those that carry multicast data.

2.1. Terminology

In the following, we will sometimes talk of a PE receiving traffic from a PMSI and then discarding it. If PIM is being used as the multicast control protocol between PEs, this always implies that the discarded traffic will not be seen by PIM on the receiving PE.

In the following, we will sometimes speak of an S-PMSI A-D route being "ignored". When we say the route is "ignored", we do not mean that its normal BGP processing is not done, but that the route is not considered when determining which P-tunnel to use when sending multicast data, and that the MPLS label values it conveys are not used. We will generally use "ignore" in quotes to indicate this meaning.

3. MS-PMSI: Multidirectional Selective PMSI

[MVPN] defines three kinds of PMSI:

- "Multidirectional Inclusive" PMSI (MI-PMSI)

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

- "Unidirectional Inclusive" PMSI (UI-PMSI)

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

- "Selective" PMSI (S-PMSI).

  A Selective PMSI is one that provides a mechanism wherein a particular PE in an MVPN can multicast messages so that they will be received by a subset of the other PEs of that MVPN. There may be an arbitrary number of S-PMSIs per PE per MVPN.

In this document we add the notion of a "Multidirectional Selective PMSI" (MS-PMSI). An MS-PMSI provides a mechanism that enables a subset of PEs in a given MVPN to multicast messages so that they will be received by the other PEs that are in the subset. There may be an arbitrary number of MS-PMSIs per PE per MVPN.

According to the definition of S-PMSI in [MVPN], only a single PE can transmit onto a given S-PMSI. An MS-PMSI may be thought of as a collection of S-PMSIs, each of which has the same subset of PEs as transmitters or receivers. Although each individual S-PMSI in the set has a single PE as transmitter, the collection of S-PMSIs has all members of the subset as transmitters, and all members of the subset as receivers.

3.1. A PE's Primary MS-PMSI

Although a PE may belong to many MS-PMSIs, we allow one MS-PMSI per PE to be distinguished as the MS-PMSI that is that PE's "primary MS-PMSI". A PE is considered to be advertising its primary MS-PMSI in a BGP S-PMSI A-D route if that route has the following properties:

- the double wild card selector (C-*,C-*) [MVPN_WILD] is specified

- the advertised S-PMSI is instantiated using one of the set of techniques described in the next section.

3.2. Instantiating MS-PMSIs

There are a number of ways to instantiate MS-PMSIs. These are specified in in the follow sub-sections. Additional methods of instantiation may be added in the future.

3.2.1. Bidirectional P-Tunnels

An MS-PMSI be instantiated as a bidirectional P-tunnel. See [MVPN_BIDIR] for the details of advertising bidirectional P-tunnels.

[MVPN_BIDIR] specifies two kinds of bidirectional P-tunnels (P-tunnels that are BIDIR-PIM [BIDIR-PIM] multicast trees, or that are MP2MP LSPs [MLDP] without PE distinguisher labels) that may only be advertised by their "roots" (as defined in that document). It follows that a PE may advertise such a P-tunnel as the instantiation of its primary MS-PMSI only if that PE is the root of the P-tunnel.

If PE1, PE2, ..., PEn are using a MP2MP LSP with PE Distinguisher labels to instantiate an MS-PMSI, the MP2MP LSP should be thought of as instantiating n MS-PMSIs, each one being the primary MS-PMSI of one of the PEs. A packet traveling on the MP2MP LSP is said to be traveling PEi's primary MS-PMSI if it is carrying the PE Distinguisher label that the root of the LSP has assigned to PEi.

3.2.2. Unidirectional P-Tunnels

For best efficiency, MS-PMSIs should be instantiated by bidirectional P-tunnels. However, it is possible to instantiate MS-PMSIs as unidirectional P-tunnels, and this can be useful in certain circumstances.

3.2.2.1. PPMP LSPs

An MS-PMSI can be implemented as a Point-to-Point-to-Multipoint (PPMP) LSP. (See, e.g, the "shared P2MP LSP" of [mLDP] section 3.) The procedures for advertising a PPMP LSP in an S-PMSI A-D route are as follows.

A new BGP attribute is defined, the "PPMP Label" attribute. This is an optional transitive attribute defined as follows:

This attribute may be carried by a BGP S-PMSI A-D route that is advertising a primary MS-PMSI instantiated as a P2MP LSP. The PPMP label is a downstream-assigned MPLS label assigned by the PE that originated the route carrying this attribute.

A PPMP label MUST NOT be added to an S-PMSI A-D route UNLESS the route contains a PTA identifying a P2MP LSP, and the route is originated by the root of the LSP.

The rules for transmitting packets on a PPMP LSP are as follows:

- The root of the LSP transmits normally, without using the PPMP label.

- A PE which is not the root of the LSP transmits a packet on the LSP as follows:

  * it pushes the PPMP label onto the packet's label stack, then

  * it unicasts the packet to the PE that is the root of the LSP; this requires pushing another label onto the packet's label stack.

When the packet is received (as a unicast) by the PE at root of the LSP, the PPMP label will either be at the top of the label stack (if penultimate hop popping is in use), or else will rise to the . The PPMP label is then popped from the stack, and that PE processes packet's label stack, recognizes the PPMP label, and as a result retransmits the packet on the corresponding P2MP LSP. In addition, the PE at the root of the P2MP processes the received packet as a multicast packet in the context of the VPN corresponding to the PPMP label. (The relationship between a PPMP label and a VPN is established by the RTs carried by the S-PMSI A-D route that advertised the PPMP label.)

Note that when an MS-PMSI is instantiated as a PPMP LSP, the PE that transmits a given packet may receive it back. A PE MUST discard, without processing, any packet it receives from the PPMP LSP if it transmitted that packet to the PPMP LSP. As a result, the procedure of instantiating an MS-PMSI as a PPMP LSP MUST NOT be used UNLESS there is a method by which a PE can identify the packets it transmitted. It is recommended to use this method only for transmitting PIM control packets, rather than multicast data packets.

3.2.2.2. Sparse Mode ASM Groups

One way to instantiate an MS-PMSI is to use a set of PIM sparse mode ASM groups. A PE advertises its primary MS-PMSI by sending an S-PMSI A-D route whose PTA identifies a "PIM-SM Tree". Every PE would have to advertise a PIM-SM tree with a distinct ASM ("Any Source Multicast") group address. To transmit a packet on the primary MS-PMSI of a particular PE, the packet would be encapsulated in GRE, with the GRE header's IP source address being the IP address of the transmitting PE, and the GRE header's IP destination address being the group address that was advertised for that MS-PMSI.

Generally speaking, this is not an efficient method of instantiating an MS-PMSI. However, it can be useful in certain circumstances, such as the "hub and spoke" MVPN discussed in [MVPN_EXTRANET]. It can also be useful as a transitional method of instantiating MS-PMSIs, allowing MS-PMSIs to be used in a network even if that network has not (yet) deployed native MPLS multicast techniques.

4. PIM over MS-PMSI

[MVPN] provides two alternative means of distributing C-multicast routing information: PIM or BGP. Procedures for running PIM over MI-PMSI are specified in that document. However, a number of efficiencies can be obtained by running PIM instead over MS-PMSI. The procedures for this are as follows.

Each PE that attaches to a given MVPN MUST originate an Intra-AS I-PMSI A-D route that does NOT contain a PTA. Each such PE MUST also advertise a primary MS-PMSI instantiated by one of the methods described in the previous section.

If PE1 needs to direct a PIM Join/Prune message to PE2, PE1 MUST join the PE2's primary MS-PMSI by joining the P-tunnel advertised in PE2's corresponding S-PMSI A-D route. The PIM J/P messages MUST be sent over that MS-PMSI.

If PE1 does not need to direct a PIM Join/Prune message to PE2, then PE1 SHOULD NOT join the P-tunnel advertised in PE2's S-PMSI A-D route, as PE1 will not be receiving any multicast data on that LSP. In the "PIM over MI-PMSI" scheme of [MVPN], if PE1 attaches to a given MVPN, the number of P-tunnels that it has to join for that MVPN is on the order of the total number of PEs attached to that MVPN. In the "PIM over MS-PMSI" scheme, the number of P-tunnels that PE1 has to join is on the order of the total number of PEs attached to those sites of the MVPN that contain multicast transmitters. In many deployments, only a small proportion of a VPN's sites contain

multicast transmitters, in which case the MS-PMSI scheme can result in a considerable reduction in the total number of P-tunnels.

Note that if PE1 and PE3 both need to send PIM Join/Prune messages to PE2, then PE1 and PE3 both join PE2's primary MS-PMSI. As a result, PE1 will see the Join/Prune messages that PE3 sends to PE2, and PE3 will see the one that PE1 sends to PE2. This allows such PIM procedures as "join suppression" and "prune override" to work normally, without impacting any PEs that do not send PIM Join/Prune messages to PE2.

At some time after PE1 has joined the P-tunnel instantiating PE2's primary MS-PMSI, PE1 may find that it no longer has any need to send PIM J/P messages to PE2. PE1 SHOULD NOT immediately prune itself from the P-tunnel, but SHOULD instead remain joined to the P-tunnel for a configurable amount of time. An implementation MUST provide a configurable parameter determining how long a PE remains joined to the P-tunnel instantiating an MS-PMSI when that PE no longer has any need to send PIM J/P messages on that tunnel.

Any PE that sends a PIM Join/Prune message on a given P-tunnel is automatically considered to be a PIM adjacency of every PE that receives the message on that P-tunnel. This implies that any PE receiving the LSP MUST accept a PIM Join/Prune message on that P-tunnel from any other PE, even if the PE that transmitted the Join/Prune messages has not previously transmitted a PIM Hello. That is, the "adjacency relationship" does not depend on the reception of PIM Hellos.

However, there is no harm if Hellos are sent, as the suppression of Hellos is only an optimization. This optimization, allowing a PIM Join/Prune message from a given router to be accepted even if no Hello from that router has been received, is often deployed in non-MVPN scenarios, and would work even when MI-PMSIs are being used. Also, PIM Hellos may be useful in certain environments for OAM purposes. Therefore, it MUST be possible to configure a PE to allow Hellos to be sent. Any PIM Hellos that PE1 sends MUST be sent on the P-tunnel advertised in PE1's S-PMSI A-D route above.

Standard PIM procedures are used, except for:

- The above change in the adjacency maintenance procedures.

- Changes in the "RPF determination" or "RPF checking" procedures as may be defined in [MVPN] or other documents extending or enhancing MVPN procedures, such as [MVPN_EXTRANET].

If an MS-PMSI is instantiated as a bidirectional P-tunnel, then the

data handling procedures of [MVPN_BIDIR] will prevent PIM from ever seeing any packets that come from the wrong transmitter or that are in the wrong partition; when such packets are received they are discarded, rather than being passed to PIM's state machinery. As a result, such packets do not cause Asserts to be generated. Other standard PIM procedures, such as Join Suppression and Prune Override may come into play, however.

If an MS-PMSI is instantiated as a PPMP tree, a PE that transmits a Join/Prune message will receive it back. Any such message is easily identified by its source address, and MUST be discarded. A PE only transmits data packets on its primary MS-PMSI, and hence does not receive them back.

All other MVPN-specific PIM procedures are as specified in [MVPN].

5. IANA Considerations

This document specifies a new BGP optional transitive attribute, "PPMP Label". A value must be assigned from the "BGP Path Attributes Registry".

6. Security Considerations

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

7. Acknowledgments

The "PPMP" mechanism is similar to a mechanism that appeared in earlier drafts of [MVPN], known as "unicasting to the root of a shared tree"; this mechanism was discussed among the authors of [MVPN].

The possibility of using of sparse mode groups to instantiate MS-PMSIs arose from a discussion with Yakov Rekhter.

8. Authors' Addresses

Arjen Boers

E-mail: arjen@boers.com

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

Maria Napierala
AT&T Labs
200 Laurel Avenue, Middletown, NJ 07748
E-mail: mnapierala@att.com

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

IJsbrand Wijnands
Cisco Systems, Inc.
De kleetlaan 6a Diegem 1831
Belgium
E-mail: ice@cisco.com

9. Normative References

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

[MLDP] "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", Wijnands, Minei, Kompella, Thomas, RFC 6388, November 2011

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

[MVPN-BGP] "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", Aggarwal, Rosen, Morin, Rekhter, draft-ietf-l3vpn-2547bis-mcast-bgp-08.txt, October 2009

[MVPN_BIDIR] "MVPN: Using Bidirectional P-Tunnels", Cai, Rosen, Wijnands, Boers, draft-ietf-l3vpn-mvpn-bidir-01.txt, February 2012

[MVPN_WILD] "Wildcards in Multicast VPN Auto-Discovery Routes", Rosen, Rekhter, Hendrickx, Qiu, draft-ietf-l3vpn-mvpn-wildcards-01.txt, January 2012

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

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

10. Informative References

[MVPN_EXTRANET] "MVPN: Extranets, Anycast-Sources, 'Hub & Spoke', with PIM Control Plane", Cai, Rosen, Sharma, Wijnands, draft-rosen-l3vpn-mvpn-extranet-04.txt, February 2012