Internet Engineering Task Force (IETF)
Request for Comments: 7740
Category: Standards Track
ISSN: 2070-1721

Z. Zhang
Y. Rekhter
Juniper Networks
A. Dolganow
Alcatel-Lucent

January 2016

Simulating Partial Mesh of Multipoint-to-Multipoint (MP2MP) Provider Tunnels with Ingress Replication

Abstract

RFC 6513 ("Multicast in MPLS/BGP IP VPNs") describes a method to support bidirectional customer multicast flows using a partial mesh of Multipoint-to-Multipoint (MP2MP) tunnels. This document specifies how a partial mesh of MP2MP tunnels can be simulated using Ingress Replication. This solution enables a service provider to use Ingress Replication to offer transparent bidirectional multicast service to its VPN customers.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7740.

Table of Contents

1. Introduction 3
1.1. Terminology 3
1.2. Requirements Language 4
2. Operation 4
2.1. Control State 4
2.2. Forwarding State 6
3. Security Considerations 7
4. References 7
4.1. Normative References 7
4.2. Informative References 8
Acknowledgements 8
Authors' Addresses 8

1. Introduction

Section 11.2 of RFC 6513 ("Partitioned Sets of PEs") describes two methods of carrying Bidirectional PIM (BIDIR-PIM) [RFC5015] C-flow traffic over a provider core without using the core as the Rendezvous Point Link (RPL) or requiring Designated Forwarder election.

With these two methods, all Provider Edges (PEs) of a particular VPN are separated into partitions, with each partition being all the PEs that elect the same PE as the Upstream PE with respect to the C-RPA (the Rendezvous Point Address in the customer's address space). A PE must discard bidirectional C-flow traffic from PEs that are not in the same partition as the PE itself.

In particular, Section 11.2.3 of RFC 6513 ("Partial Mesh of MP2MP P-Tunnels") guarantees the above discard behavior without using an extra PE Distinguisher Label by having all PEs in the same partition join a single MP2MP tunnel dedicated to that partition and use it to transmit traffic. All traffic arriving on the tunnel will be from PEs in the same partition, so it will be always accepted.

RFC 6514 specifies BGP encodings and procedures used to implement Multicast VPN (MVPN) as specified in RFC 6513, while the details related to MP2MP tunnels are specified in [RFC7582].

RFC 7582 assumes that an MP2MP P-tunnel is realized either via BIDIR-PIM [RFC5015] or via MP2MP mLDP (Multipoint extensions for LDP) [RFC6388]. Each would require signaling and state not just on PEs, but on the P routers as well. This document describes how the MP2MP tunnel can be simulated with a mesh of P2MP tunnels, each of which is instantiated by Ingress Replication (IR) [RFC6513] [RFC6514]. The procedures in this document are different from the procedures that are used to set up the mesh of Ingress Replication tunnels as described in RFC 6514; the procedures in this document do not require each PE on the MP2MP tunnel to send a Selective P-Multicast Service Interface (S-PMSI) auto-discovery route (A-D route) for the P2MP tunnel that the PE is the root for, nor do they require each PE to send a Leaf A-D route to the root of each P2MP tunnel in the mesh.

Because it uses Ingress Replication, this scheme has both the advantages and the disadvantages of Ingress Replication in general.

1.1. Terminology

This document uses terminology from [RFC5015], [RFC6513], [RFC6514], and [RFC7582].

1.2. Requirements Language

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. Operation

In the following sections, the originator of an S-PMSI A-D route or Leaf A-D route is determined from the "originating router's IP address" field of the corresponding route.

2.1. Control State

If a PE, say PEx, is connected to a site of a given VPN and PEx's next-hop interface to some C-RPA is a VPN Routing and Forwarding (VRF) interface, then PEx MUST advertise a (C-*,C-*-BIDIR) S-PMSI A-D route, regardless of whether it has any local BIDIR-PIM join states corresponding to the C-RPA learned from its Customer Edges (CEs). It MAY also advertise one or more (C-*,C-G-BIDIR) S-PMSI A-D routes, if selective distribution trees are needed for those C-G-BIDIR groups and the corresponding C-RPA is in the site that the PEx connects to. For example, the (C-*,C-G-BIDIR) S-PMSI A-D routes could be triggered when the (C-*,C-G-BIDIR) traffic rate goes above a threshold (this may require measuring the traffic in both directions, due to the nature of BIDIR-PIM), and fan-out could also be taken into account.

The S-PMSI A-D routes include a PMSI Tunnel Attribute (PTA) with tunnel type set to Ingress Replication, with the Leaf Information Required flag set, with a downstream allocated MPLS label that other PEs in the same partition MUST use when sending relevant C-BIDIR flows to this PE, and with the Tunnel Identifier field in the PTA set to a routable address of the originator. This specification does not prevent sharing of labels between P-tunnels, such as a label being shared by a (C-*,C-*-BIDIR) and a (C-*,C-G-BIDIR) S-PMSI A-D route originated by a given PE (note that other specifications put constraints on how that can be done, e.g., [MVPN-EXTRANET]).

If some other PE, PEy, receives and imports into one of its VRFs any (C-*,C-*-BIDIR) S-PMSI A-D route whose PTA specifies an IR P-tunnel and the VRF has any local BIDIR-PIM join state that PEy has received from its CEs and if PEy chooses PEx as its Upstream PE with respect to the C-RPA for those states, PEy MUST advertise a Leaf A-D route in response. Or, if PEy has received and imported into one of its VRFs a (C-*,C-*-BIDIR) S-PMSI A-D route from PEx before, then upon receiving in the VRF any local BIDIR-PIM join state from its CEs with PEx being the Upstream PE for those states' C-RPA, PEy MUST advertise a Leaf A-D route.

The encoding of the Leaf A-D route is as specified in RFC 6514, except that the Route Targets are set to the same value as in the corresponding S-PMSI A-D route so that the Leaf A-D route will be imported by all VRFs that import the corresponding S-PMSI A-D route. This is irrespective of whether or not the originator of the S-PMSI A-D route is the Upstream PE from a receiving PE's perspective. The label in the PTA of the Leaf A-D route originated by PEy MUST be allocated specifically for PEx, so that when traffic arrives with that label, the traffic can associate with the partition (represented by the PEx). This specification does not prevent sharing of labels between P-tunnels, such as a label being shared by a (C-*,C-*-BIDIR) and a (C-*,C-G-BIDIR) Leaf A-D route originated by a given PE (note that other specifications put constraints on how that can be done, e.g., [MVPN-EXTRANET]).

Note that RFC 6514 requires that a PE or an ASBR (Autonomous System Border Router) take no action with regard to a Leaf A-D route unless that Leaf A-D route carries an IP-address-specific Route Target identifying the PE/ASBR. This document removes that requirement when the route key of a Leaf A-D route identifies a (C-*,C-*-BIDIR) or a (C-*,C-G-BIDIR) S-PMSI.

To speed up convergence (so that PEy starts receiving traffic from its new Upstream PE immediately instead of waiting until the new Leaf A-D route corresponding to the new Upstream PE is received by sending PEs), PEy MAY advertise a Leaf A-D route even if it does not choose PEx as its Upstream PE with respect to the C-RPA. With that, it will receive traffic from all PEs, but some will arrive with the label corresponding to its choice of Upstream PE while some will arrive with a different label; the traffic in the latter case will be discarded.

Similar to the (C-*,C-*-BIDIR) case, if PEy receives and imports into one of its VRFs any (C-*,C-G-BIDIR) S-PMSI A-D route whose PTA specifies an IR P-tunnel, PEy chooses PEx as its Upstream PE with respect to the C-RPA, and it has corresponding local (C-*,C-G-BIDIR) join state that it has received from its CEs in the VRF, PEy MUST advertise a Leaf A-D route in response. Or, if PEy has received and imported into one of its VRFs a (C-*,C-G-BIDIR) S-PMSI A-D route before, then upon receiving its local (C-*,C-G-BIDIR) join state from its CEs in the VRF, it MUST advertise a Leaf A-D route.

The encoding of the Leaf A-D route is similar to the (C-*,C-*-BIDIR) case. Similarly, PEy MAY advertise a Leaf A-D route even if it does not choose PEx as its Upstream PE with respect to the C-RPA.

PEy MUST withdraw the corresponding Leaf A-D route if any of the following conditions are true:

- the (C-*,C-*-BIDIR) or (C-*,C-G-BIDIR) S-PMSI A-D route is withdrawn.
- PEy no longer chooses the originator PEx as its Upstream PE with respect to C-RPA and PEy only advertises Leaf A-D routes in response to its Upstream PE's S-PMSI A-D route.
- if relevant local join state is pruned.

2.2. Forwarding State

The specification regarding forwarding state in this section matches the "When an S-PMSI is a 'Match for Transmission'" and "When an S-PMSI is a 'Match for Reception'" rules for the "Flat Partitioning" method in [RFC7582], except that the rules about (C-*,C-*) are not applicable, because this document requires that (C-*,C-*-BIDIR) S-PMSI A-D routes are always originated for a VPN that supports C-BIDIR flows.

For the (C-*,C-G-BIDIR) S-PMSI A-D route that a PEy receives and imports into one of its VRFs from its Upstream PE with respect to the C-RPA, if PEy itself advertises the S-PMSI A-D route in the VRF, PEy maintains a (C-*,C-G-BIDIR) forwarding state in the VRF, with the Ingress Replication provider tunnel leaves being the originators of the S-PMSI A-D route and all relevant Leaf A-D routes. The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN Network Layer Reachability Information (NLRI) of the (C-*,C-G-BIDIR) S-PMSI A-D route advertised by the Upstream PE.

For the (C-*,C-*-BIDIR) S-PMSI A-D route that a PEy receives and imports into one of its VRFs from its Upstream PE with respect to a C-RPA, if PEy itself advertises the S-PMSI A-D route in the VRF, it maintains appropriate forwarding states in the VRF for the ranges of bidirectional groups for which the C-RPA is responsible. The provider tunnel leaves are the originators of the S-PMSI A-D route and all relevant Leaf A-D routes. The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-*,C-*-BIDIR) S-PMSI A-D route advertised by the Upstream PE. This is for the so-called "Sender Only Branches" where a router only has data to send upstream towards C-RPA but no explicit join state for a particular bidirectional group. Note that the traffic must be sent to all PEs (not just the Upstream PE) in the

partition, because they may have specific (C-*,C-G-BIDIR) join states that this PEy is not aware of, while there are no corresponding (C-*,C-G-BIDIR) S-PMSI A-D and Leaf A-D routes.

For a (C-*,C-G-BIDIR) join state that a PEy has received from its CEs in a VRF, if there is no corresponding (C-*,C-G-BIDIR) S-PMSI A-D route from its Upstream PE in the VRF, PEy maintains a corresponding forwarding state in the VRF, with the provider tunnel leaves being the originators of the (C-*,C-*-BIDIR) S-PMSI A-D route and all relevant Leaf A-D routes (same as the "Sender Only Branches" case above). The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-*,C-*-BIDIR) S-PMSI A-D route originated by the Upstream PE. If there is also no (C-*,C-*-BIDIR) S-PMSI A-D route from its Upstream PE, then the provider tunnel has an empty set of leaves, and PEy does not forward relevant traffic across the provider network.

3. Security Considerations

This document raises no new security issues. Security considerations for the base protocol are covered in [RFC6513] and [RFC6514].

4. References

4.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <http://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <http://www.rfc-editor.org/info/rfc6514>.

[RFC7582] Rosen, E., Wijnands, IJ., Cai, Y., and A. Boers, "Multicast Virtual Private Network (MVPN): Using Bidirectional P-Tunnels", RFC 7582, DOI 10.17487/RFC7582, July 2015, <http://www.rfc-editor.org/info/rfc7582>.

4.2. Informative References

[MVPN-EXTRANET]
Rekhter, Y., Ed., Rosen, E., Ed., Aggarwal, R., Cai, Y., and T. Morin, "Extranet Multicast in BGP/IP MPLS VPNs", Work in Progress, draft-ietf-bess-mvpn-extranet-06, January 2016.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <http://www.rfc-editor.org/info/rfc5015>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <http://www.rfc-editor.org/info/rfc6388>.

Acknowledgements

We would like to thank Eric Rosen for his comments and suggestions for some text used in the document.

Authors' Addresses

Zhaohui Zhang
Juniper Networks
10 Technology Park Dr.
Westford, MA 01886
United States

Email: zzhang@juniper.net

Yakov Rekhter
Juniper Networks

Andrew Dolganow
Alcatel-Lucent
600 March Rd.
Ottawa, ON K2K 2E6
Canada

Email: andrew.dolganow@alcatel-lucent.com