

Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: January 7, 2017

V. Govindan  
M. Mudigonda  
A. Sajassi  
Cisco Systems  
G. Mirsky  
Ericsson  
July 6, 2016

Fault Management for EVPN networks  
draft-gsm-bess-evpn-bfd-00

Abstract

This document proposes a proactive, in-band network OAM mechanism to detect loss of continuity and miss-connection faults that affect unicast and multi-destination paths, used by Broadcast, unknown Unicast and Multicast traffic, in an EVPN network. The mechanisms proposed in the draft use the principles of the widely adopted Bidirectional Forwarding Detection protocol.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on January 7, 2017.

Copyright Notice

Copyright (c) 2016 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. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
2. Scope of the Document . . . . .	3
3. Motivation for running BFD at the network layer of EVPN . . . . .	3
4. Fault Detection of unicast traffic . . . . .	4
5. Fault Detection of BUM traffic using ingress replication (MP2P) . . . . .	5
6. Fault Detection of BUM traffic using P2MP tunnels (LSM) . . . . .	5
7. BFD packet encapsulation . . . . .	5
7.1. Using GAL/G-ACh encapsulation without IP headers . . . . .	5
7.1.1. Ingress replication . . . . .	5
7.1.1.1. Alternative encapsulation format . . . . .	5
7.1.2. LSM . . . . .	6
7.1.3. Unicast . . . . .	6
7.1.3.1. Alternative encapsulation format . . . . .	6
7.2. Using IP headers . . . . .	7
8. Scalability Considerations . . . . .	7
9. IANA Considerations . . . . .	7
10. Security Considerations . . . . .	8
11. References . . . . .	8
11.1. Normative References . . . . .	8
11.2. Informative References . . . . .	10
Authors' Addresses . . . . .	10

## 1. Introduction

[I-D.salam-l2vpn-evpn-oam-req-frmwk] and [I-D.ooamdt-rtgwg-ooam-requirement] outlines the OAM requirements of Ethernet VPN networks [RFC7432]. This document proposes mechanisms for proactive fault detection at the network(overlay) OAM layer of EVPN. EVPN fault detection mechanisms need to consider unicast and Broadcast and unknown Unicast (BUM) traffic separately since they map to different FECs in EVPN, hence this document proposes different fault detection mechanisms to suit each type using the principles of [RFC5880], [RFC5884] and Point-to-multipoint BFD [I-D.ietf-bfd-multipoint] and [I-D.ietf-bfd-multipoint-active-tail].

### 1.1. 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 RFC 2119 [RFC2119].

## 2. Scope of the Document

This document proposes proactive fault detection for EVPN [RFC7432] using BFD mechanisms for:

- o Unicast traffic.
- o BUM traffic using Multi-point-to-Point (MP2P) tunnels (ingress replication).
- o BUM traffic using Point-to-Multipoint (P2MP) tunnels (LSM).

This document does not discuss BFD mechanisms for:

- o EVPN variants like PBB-EVPN [RFC7623]. This will be addressed in future versions.
- o IRB solution based on EVPN [I-D.ietf-bess-evpn-inter-subnet-forwarding]. This will be addressed in future versions.
- o EVPN using other encapsulations like VxLAN, NVGRE and MPLS over GRE [I-D.ietf-bess-evpn-overlay].
- o BUM traffic using MP2MP tunnels will also be addressed in a future version of this document.

This specification describes procedures only for BFD asynchronous mode. BFD demand mode is outside the scope of this specification. Further, the use of the Echo function is outside the scope of this specification.

## 3. Motivation for running BFD at the network layer of EVPN

The choice of running BFD at the network layer of the OAM model for EVPN [I-D.salam-l2vpn-evpn-oam-req-frmwk] and [I-D.ooamdt-rtgwg-ooam-requirement] was made after considering the following:

- o In addition to detecting link failures in the EVPN network, BFD sessions at the network layer can be used to monitor the successful programming of labels used for setting up MP2P and P2MP

EVPN tunnels transporting Unicast and BUM traffic. The scope of reachability detection covers the ingress and the egress EVPN PE nodes and the network connecting them.

- o Monitoring a representative set of path(s) or a particular path among the multiple paths available between two EVPN PE nodes could be done by exercising the entropy labels when they are used. However paths that cannot be realized by entropy variations cannot be monitored. Fault monitoring requirements outlined by [I-D.salam-l2vpn-evpn-oam-req-frmwk] are addressed by the mechanisms proposed by this draft.

Successful establishment and maintenance of BFD sessions between EVPN PE nodes does not fully guarantee that the EVPN service is functioning. For example, an egress EVPN-PE can understand the EVPN label but could switch data to incorrect interface. However, once BFD sessions in the EVPN Network Layer reach UP state, it does provide additional confidence that data transported using those tunnels will reach the expected egress node. When the BFD session in EVPN overlay goes down that can be used as indication of the Loss-of-Connectivity defect in the EVPN underlay that would cause EVPN service failure.

#### 4. Fault Detection of unicast traffic

The mechanisms specified in BFD for MPLS LSPs [RFC5884] [RFC7726] can be applied to bootstrap and maintain BFD sessions for unicast EVPN traffic. The discriminators required for de-multiplexing the BFD sessions MUST be exchanged using EVPN LSP ping specifying the Unicast EVPN FEC [I-D.jain-bess-evpn-lsp-ping] before establishing the BFD session. This is needed since the MPLS label stack does not contain enough information to disambiguate the sender of the packet. The usage of MPLS entropy labels take care of addressing the requirement of monitoring various paths of the multi-path server layer network [RFC6790]. Each unique realizable path between the participating PE routers MAY be monitored separately when entropy labels are used. The multi-path connectivity between two PE routers MUST be tracked by at least one representative BFD session, in which case the granularity of fault-detection would be coarser. The PE node receiving the EVPN LSP ping MUST allocate BFD discriminators using the procedures defined in [RFC7726]. Note that once the BFD session for the EVPN label is UP, either end of the BFD session MUST NOT change the local discriminator values of the BFD Control packets it generates, unless it first brings down the session as specified in [RFC5884].

## 5. Fault Detection of BUM traffic using ingress replication (MP2P)

Ingress replication uses separate MP2P tunnels for transporting BUM traffic from the ingress PE (head) to a set of one or more egress PEs (tails). The fault detection mechanism proposed by this document takes advantage of the fact that a unique copy is made by the head for each tail. Another key aspect to be considered in EVPN is the advertisement of the inclusive multicast route. The BUM traffic flows from a head node to a particular tail only after the head receives the inclusive multicast route containing the BUM EVPN label (downstream allocated) corresponding to the MP2P tunnel. The head-end PE performing ingress replication MUST initiate an EVPN LSP ping using the inclusive multicast FEC [I-D.jain-bess-evpn-lsp-ping] upon receiving an inclusive multicast route from a tail to bootstrap the BFD session. There MAY exist multiple BFD sessions between a head PE and an individual tail due to the usage of entropy labels [RFC6790] for an inclusive multicast FEC. The PE node receiving the EVPN LSP ping MUST allocate BFD discriminators using the procedures defined in [RFC7726]. Note that once the BFD session for the EVPN label is UP, either end of the BFD session MUST NOT change the local discriminator values of the BFD Control packets it generates, unless it first brings down the session as specified in [RFC5884].

## 6. Fault Detection of BUM traffic using P2MP tunnels (LSM)

TBD.

## 7. BFD packet encapsulation

### 7.1. Using GAL/G-ACh encapsulation without IP headers

#### 7.1.1. Ingress replication

The packet contains the following labels: LSP label (transport) when not using PHP, the optional entropy label, the BUM label and the SH label [RFC7432] (where applicable). The G-ACh type is set to TBD. The G-ACh payload of the packet MUST contain the L2 header (in overlay space) followed by the IP header encapsulating the BFD packet. The MAC address of the inner packet is used to validate the <EVI, MAC> in the receiving node. The discriminator values of BFD are obtained through negotiation through the out-of-band EVPN LSP ping.

##### 7.1.1.1. Alternative encapsulation format

A new TLV can be defined as proposed in Sec 3 of [RFC6428] to include the EVPN FEC information as a TLV following the BFD Control packet.

The format of the TLV can be reused from the EVPN Inclusive Multicast sub-TLV proposed by Fig 2 of [I-D.jain-bess-evpn-lsp-ping].

A new type (TBD3) to indicate the EVPN Inclusive Multicast SubTLV is requested from the "CC/ CV MEP-ID TLV" registry [RFC6428].

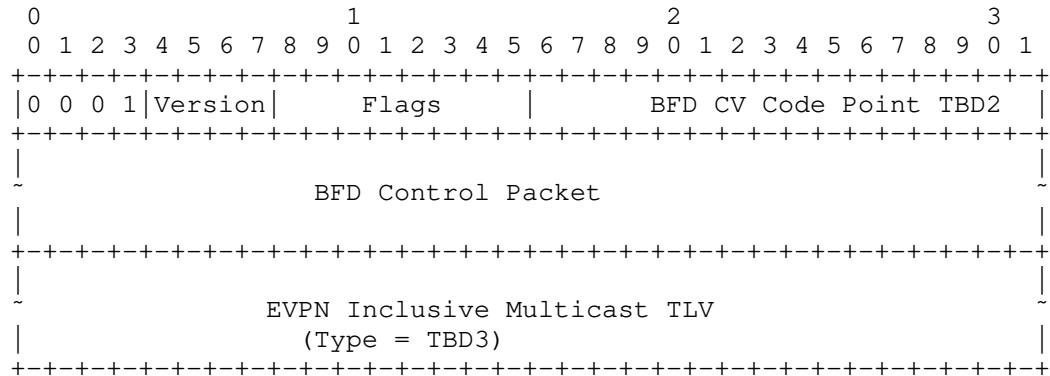


Figure 1: BFD-EVPN CV Message for EVPN Multicast  
(Ingress Replication)

#### 7.1.2. LSM

TBD.

#### 7.1.3. Unicast

The packet contains the following labels: LSP label (transport) when not using PHP, the optional entropy label and the EVPN Unicast label. The G-ACh type is set to TBD. The G-ACh payload of the packet MUST contain the L2 header (in overlay space) followed by the IP header encapsulating the BFD packet. The MAC address of the inner packet is used to validate the <EVI, MAC> in the receiving node. The discriminator values of BFD are obtained through negotiation through the out-of-band EVPN ping.

##### 7.1.3.1. Alternative encapsulation format

A new TLV can be defined as proposed in Sec 3 of [RFC6428] to include the EVPN FEC information as a TLV following the BFD Control packet. The format of the TLV can be reused from the EVPN MAC sub-TLV proposed by Fig 1 of [I-D.jain-bess-evpn-lsp-ping]. A new type (TBD4) to indicate the EVPN MAC SubTLV is requested from the "CC/ CV MEP-ID TLV" registry [RFC6428].

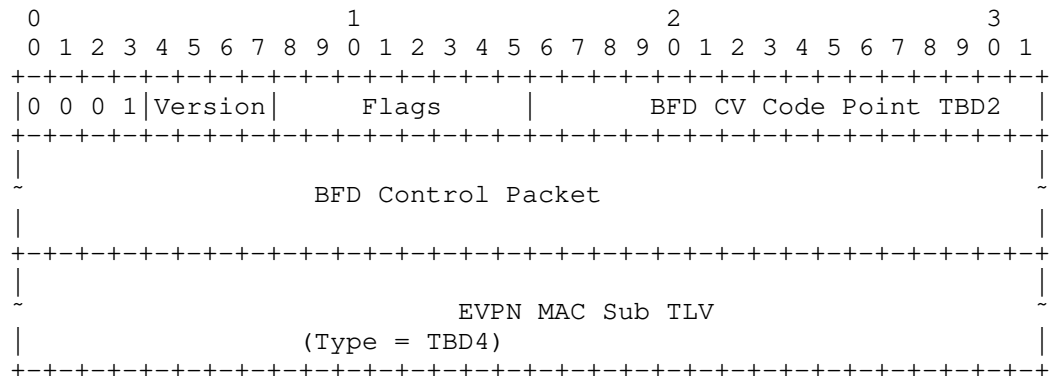


Figure 2: BFD-EVPN CV Message for EVPN Unicast

## 7.2. Using IP headers

The encapsulation option using IP headers will not be suited for EVPN, as using different values in the destination IP address for data and OAM (BFD) packets could cause the BFD packets to follow a different path than that of data packets. Hence this option MUST NOT be used for EVPN.

## 8. Scalability Considerations

The mechanisms proposed by this draft could affect the packet load on the network and its elements especially when supporting configurations involving a large number of EVIs. The option of slowing down or speeding up BFD timer values can be used by an administrator or a network management entity to maintain the overhead incurred due to fault monitoring at an acceptable level.

## 9. IANA Considerations

IANA is requested for two channel types from the "Pseudowire Associated Channel Types" registry in [RFC4385].

TBD1 BFD-EVPN CC message

TBD2 BFD-EVPN CV message

Ed Note: Do we need a CC code point? TBD

IANA is requested to allocate the following code-points from the "CC/ CV MEP-ID TLV" registry [RFC6428]. The parent registry is the "Pseudowire Associated Channel Types" registry of [RFC4385]. All

code points within this registry shall be allocated according to the "Standards Action" procedures as specified in [RFC5226]. The items tracked in the registry will be the type, associated name, and reference. The requested values are:

TBD3 - CV code-point for BFD EVPN Inclusive multicast.

TBD4 - CV code-point for BFD EVPN Unicast.

## 10. Security Considerations

TBD.

## 11. References

### 11.1. Normative References

- [I-D.ietf-bess-evpn-inter-subnet-forwarding]  
Sajassi, A., Salam, S., Thoria, S., Rekhter, Y., Drake, J., Yong, L., and L. Dunbar, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-01 (work in progress), October 2015.
- [I-D.ietf-bess-evpn-overlay]  
Sajassi, A., Drake, J., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-04 (work in progress), June 2016.
- [I-D.ietf-bfd-multipoint]  
Katz, D., Ward, D., and J. Networks, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-08 (work in progress), April 2016.
- [I-D.ietf-bfd-multipoint-active-tail]  
Katz, D., Ward, D., and J. Networks, "BFD Multipoint Active Tails.", draft-ietf-bfd-multipoint-active-tail-02 (work in progress), May 2016.
- [I-D.jain-bess-evpn-lsp-ping]  
Jain, P., Boutros, S., and S. Salam, "LSP-Ping Mechanisms for EVPN and PBB-EVPN", draft-jain-bess-evpn-lsp-ping-03 (work in progress), May 2016.
- [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>>.



- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <<http://www.rfc-editor.org/info/rfc4385>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<http://www.rfc-editor.org/info/rfc5880>>.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <<http://www.rfc-editor.org/info/rfc5884>>.
- [RFC6428] Allan, D., Ed., Swallow, G., Ed., and J. Drake, Ed., "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, DOI 10.17487/RFC6428, November 2011, <<http://www.rfc-editor.org/info/rfc6428>>.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <<http://www.rfc-editor.org/info/rfc6790>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.
- [RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <<http://www.rfc-editor.org/info/rfc7623>>.
- [RFC7726] Govindan, V., Rajaraman, K., Mirsky, G., Akiya, N., and S. Aldrin, "Clarifying Procedures for Establishing BFD Sessions for MPLS Label Switched Paths (LSPs)", RFC 7726, DOI 10.17487/RFC7726, January 2016, <<http://www.rfc-editor.org/info/rfc7726>>.

## 11.2. Informative References

- [I-D.ooamdt-rtgwg-ooam-requirement]  
Kumar, N., Pignataro, C., Kumar, D., Mirsky, G., Chen, M., Nordmark, E., Networks, J., and D. Mozes, "Overlay OAM Requirements", draft-ooamdt-rtgwg-ooam-requirement-00 (work in progress), March 2016.
- [I-D.salam-l2vpn-evpn-oam-req-frmwk]  
Salam, S., Sajassi, A., Aldrin, S., and J. Drake, "E-VPN Operations, Administration and Maintenance Requirements and Framework", draft-salam-l2vpn-evpn-oam-req-frmwk-02 (work in progress), January 2014.

## Authors' Addresses

Vengada Prasad Govindan  
Cisco Systems  
  
Email: venggovi@cisco.com

Mudigonda Mallik  
Cisco Systems  
  
Email: mmudigon@cisco.com

Ali Sajassi  
Cisco Systems  
  
Email: sajassi@cisco.com

Gregory Mirsky  
Ericsson  
  
Email: gregory.mirsky@ericsson.com

BESS Working Group  
Internet-Draft  
Intended Status: Standards Track

Ali Sajassi  
Samir Thoria  
Cisco  
Keyur Patel  
Derek Yeung  
Arrcus  
John Drake  
Wen Lin  
Juniper

Expires: December 24, 2018

June 24, 2018

IGMP and MLD Proxy for EVPN  
draft-ietf-bess-evpn-igmp-mld-proxy-02

Abstract

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

This draft describes how to support efficiently endpoints running IGMP for the above services over an EVPN network by incorporating IGMP proxy procedures on EVPN PEs.

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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2018 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	Introduction . . . . .	4
1.1	Terminology . . . . .	5
2	IGMP Proxy . . . . .	6
2.1	Proxy Reporting . . . . .	6
2.1.1	IGMP Membership Report Advertisement in BGP . . . . .	6
2.1.1	IGMP Leave Group Advertisement in BGP . . . . .	8
2.2	Proxy Querier . . . . .	9
3	Operation . . . . .	10
3.1	PE with only attached hosts/VMs for a given subnet . . . . .	10
3.2	PE with mixed of attached hosts/VMs and multicast source . . . . .	11
3.3	PE with mixed of attached hosts/VMs, multicast source and router . . . . .	11
4	All-Active Multi-Homing . . . . .	11
4.1	Local IGMP Join Synchronization . . . . .	12
4.2	Local IGMP Leave Group Synchronization . . . . .	13
4.2.1	Remote Leave Group Synchronization . . . . .	13
4.2.2	Common Leave Group Synchronization . . . . .	14
5	Single-Active Multi-Homing . . . . .	14
6	Selective Multicast Procedures for IR tunnels . . . . .	14
7	BGP Encoding . . . . .	15
7.1	Selective Multicast Ethernet Tag Route . . . . .	15
7.1.1	Constructing the Selective Multicast Ethernet Tag route . . . . .	17
7.2	IGMP Join Synch Route . . . . .	18
7.2.1	Constructing the IGMP Join Synch Route . . . . .	19

7.3 IGMP Leave Synch Route . . . . .	20
7.3.1 Constructing the IGMP Leave Synch Route . . . . .	22
7.4 Multicast Flags Extended Community . . . . .	23
7.5 EVI-RT Extended Community . . . . .	24
7.6 Rewriting of RT ECs and EVI-RT ECs by ASBRs . . . . .	26
8 Acknowledgement . . . . .	26
9 Security Considerations . . . . .	26
10 IANA Considerations . . . . .	26
11 References . . . . .	27
11.1 Normative References . . . . .	27
11.2 Informative References . . . . .	27
Authors' Addresses . . . . .	28

## 1 Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

In DC applications, a POD can consist of a collection of servers supported by several TOR and Spine switches. This collection of servers and switches are self contained and may have their own control protocol for intra-POD communication and orchestration. However, EVPN is used as way of standard inter-POD communication for both intra-DC and inter-DC. A subnet can span across multiple PODs and DCs. EVPN provides robust multi-tenant solution with extensive multi-homing capabilities to stretch a subnet (e.g., VLAN) across multiple PODs and DCs. There can be many hosts/VMs (e.g., several hundreds) attached to a subnet that is stretched across several PODs and DCs.

These hosts/VMs express their interests in multicast groups on a given subnet/VLAN by sending IGMP membership reports (Joins) for their interested multicast group(s). Furthermore, an IGMP router (e.g., IGMPv1) periodically sends membership queries to find out if there are hosts on that subnet still interested in receiving multicast traffic for that group. The IGMP/MLD Proxy solution described in this draft has three objectives to accomplish:

- 1) Reduce flooding of IGMP messages: just like ARP/ND suppression mechanism in EVPN to reduce the flooding of ARP messages over EVPN, it is also desired to have a mechanism to reduce the flood of IGMP messages (both Queries and Reports) in EVPN.
- 2) Distributed anycast multicast proxy: it is desired for the EVPN network to act as a distributed anycast multicast router with respect to IGMP/MLD proxy function for all the hosts attached to that subnet.
- 3) Selective Multicast: to forward multicast traffic over EVPN network such that it only gets forwarded to the PEs that have interest in the multicast group(s) - i.e., multicast traffic will not be forwarded to the PEs that have no receivers attached to them for that multicast group. This draft shows how this objective may be achieved when Ingress Replication is used to distribute the multicast traffic among the PEs. Procedures for supporting selective multicast using P2MP tunnels can be found in [bum-procedure-updates]

The first two objectives are achieved by using IGMP/MLD proxy on the

PE and the third objective is achieved by setting up a multicast tunnel (e.g., ingress replication) only among the PEs that have interest in that multicast group(s) based on the trigger from IGMP/MLD proxy processes. The proposed solutions for each of these objectives are discussed in the following sections.

## 1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

POD: Point of Delivery

ToR: Top of Rack

NV: Network Virtualization

NVO: Network Virtualization Overlay

VNI: Virtual Network Identifier (for VXLAN)

EVPN: Ethernet Virtual Private Network

IGMP: Internet Group Management Protocol

MLD: Multicast Listener Discovery

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

PE: Provider Edge device.

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single

or multiple BDs. In case of VLAN-bundle and VLAN-based service models VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms.

**Ethernet Tag:** An Ethernet tag identifies a particular broadcast domain, e.g., a VLAN. An EVPN instance consists of one or more broadcast domains.

**Single-Active Redundancy Mode:** When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

**All-Active Redundancy Mode:** When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

## 2 IGMP Proxy

IGMP Proxy mechanism is used to reduce the flooding of IGMP messages over EVPN network similar to ARP proxy used in reducing the flooding of ARP messages over EVPN. It also provides triggering mechanism for the PEs to setup their underlay multicast tunnels. IGMP Proxy mechanism consist of two components: a) Proxy for IGMP Reports and b) Proxy for IGMP Queries.

### 2.1 Proxy Reporting

When IGMP protocol is used between host/VMs and its first hop EVPN router (EVPN PE), Proxy-reporting is used by the EVPN PE to summarize (when possible) reports received from downstream hosts and propagate it in BGP to other PEs that are interested in the info. This is done by terminating IGMP Reports in the first hop PE, translating and exchanging the relevant information among EVPN BGP speakers. The information is again translated back to IGMP message at the recipient EVPN speaker. Thus it helps create an IGMP overlay subnet using BGP. In order to facilitate such an overlay, this document also defines a new EVPN route type NLRI, EVPN Selective Multicast Ethernet Tag route, along with its procedures to help exchange and register IGMP multicast groups [section 5].

#### 2.1.1 IGMP Membership Report Advertisement in BGP

When a PE wants to advertise an IGMP membership report (Join) using the BGP EVPN route, it follows the following rules:

- 1) When the first hop PE receives several IGMP membership reports



(Joins) , belonging to the same IGMP version, from different attached hosts/VMs for the same (\*,G) or (S,G), it only sends a single BGP message corresponding to the very first IGMP Join. This is because BGP is a statefull protocol and no further transmission of the same report is needed. If the IGMP Join is for (\*,G), then multicast group address along with the corresponding version flag (v1, v2, or v3) are set. In case of IGMPv3, exclude flag also needs to be set to indicate that no source IP address to be excluded (e.g., include all sources "\*"). If the IGMP Join is for (S,G), then besides setting multicast group address along with the version flag v3, the source IP address and the include/exclude flag must be set. It should be noted that when advertising the EVPN route for (S,G), the only valid version flag is v3 (i.e., v1 and v2 flags must be set to zero).

2) When the first hop PE receives an IGMPv3 Join for (S,G) on a given BD, it advertises the corresponding EVPN Selective Multicast Ethernet Tag (SMET) route regardless of whether the source (S) is attached to itself or not in order to facilitate the source move in the future.

3) When the first hop PE receives an IGMP version-X Join first for (\*,G) and then later it receives an IGMP version-Y Join for the same (\*,G), then it will re-advertise the same EVPN SMET route with flag for version-Y set in addition to any previously-set version flag(s). In other words, the first hop PE does not withdraw the EVPN route before sending the new route because the flag field is not part of BGP route key processing.

4) When the first hop PE receives an IGMP version-X Join first for (\*,G) and then later it receives an IGMPv3 Join for the same multicast group address but for a specific source address S, then the PE will advertise a new EVPN SMET route with v3 flag set (and v1 and v2 reset). Include/exclude flag also need to be set accordingly. Since source IP address is used as part of BGP route key processing, it is considered as a new BGP route advertisement.

5) When a PE receives an EVPN SMET route with more than one version flag set, it will generate the corresponding IGMP report for (\*,G) for each version specified in the flag field. With multiple version flags set, there should be no source IP address in the receive EVPN route. If there is, then an error should be logged. If v3 flag is set (in addition to v1 or v2), then the include/exclude flag needs to indicate "exclude". If not, then an error should be logged. The PE MUST generate an IGMP membership report (Join) for that (\*,G) and each IGMP version in the version flag.

6) When a PE receives a list of EVPN SMET NLRIs in its BGP update message, each with a different source IP address and the multicast group address, and the version flag is set to v3, then the PE generates an IGMPv3 membership report with a record corresponding to the list of source IP addresses and the group address along with the proper indication of inclusion/exclusion.

7) Upon receiving EVPN SMET route(s) and before generating the corresponding IGMP Join(s), the PE checks to see whether it has any CE multicast router for that BD on any of its ES's. The PE provides such check by listening for PIM hellos on that AC (i.e., <ES,BD>). If it has router's ACs, then the generated IGMP Join(s) are sent to those ACs. If it doesn't have any router's AC, then no IGMP Join(s) needs to be generated because sending IGMP Joins to other hosts can result in unintentionally preventing a host from joining a specific multicast group for IGMPv1 and IGMPv2 - i.e., if the PE does not receive a join from the host it will not forward multicast data to it. Per [RFC4541], when an IGMPv1 or IGMPv2 host receives a membership report for a group address that it intends to join, the host will suppress its own membership report for the same group. In other words, an IGMPv1 or IGMPv2 Join MUST NOT be sent on an AC that does not lead to a CE multicast router. This message suppression is a requirement for IGMPv1 and IGMPv2 hosts. This is not a problem for hosts running IGMPv3 because there is no suppression of IGMP Membership reports.

#### 2.1.1 IGMP Leave Group Advertisement in BGP

When a PE wants to withdraw an EVPN SMET route corresponding to an IGMPv2 Leave Group (Leave) or IGMPv3 "Leave" equivalent message, it follows the following rules:

1) For IGMPv1, there is no explicit membership leave; therefore, the PE needs to periodically send out an IGMP membership query to determine whether there is any host left who is interested in receiving traffic directed to this multicast group (this proxy query function will be described in more details in section 2.2). If there is no host left, then the PE re-advertises EVPN SMET route with the v1 version flag reset. If this is the last version flag to be reset, then instead of re-advertising the EVPN route with all version flags reset, the PE withdraws the EVPN route for that (\*,G).

2) When a PE receives an IGMPv2 Leave Group or its "Leave" equivalent message for IGMPv3 from its attached host, it checks to see if this host is the last host who is interested in this multicast group by sending a query for the multicast group. If the host was indeed the last one, then the PE re-advertises EVPN SMET Multicast route with the corresponding version flag reset. If this is the last version

flag to be reset, then instead of re-advertising the EVPN route with all version flags reset, the PE withdraws the EVPN route for that (\*,G).

3) When a PE receives an EVPN SMET route for a given (\*,G), it compares the received version flags from the route with its per-PE stored version flags. If the PE finds that a version flag associated with the (\*,G) for the remote PE is reset, then the PE generates IGMP Leave for that (\*,G) toward its local interface (if any) attached to the multicast router for that multicast group. It should be noted that the received EVPN route should at least have one version flag set. If all version flags are reset, it is an error because the PE should have received an EVPN route withdraw for the last version flag. If the PE receives an EVPN SMET route withdraw, then it must remove the remote PE from the OIF list associated with that multicast group.

4) When a PE receives an EVPN SMET route withdraw, it removes the remote PE from its OIF list for that multicast group and if there are no more OIF entries for that multicast group (either locally or remotely), then the PE MUST stop responding to queries from the locally attached router (if any). If there is a source for that multicast group, the PE stops sending multicast traffic for that source.

## 2.2 Proxy Querier

As mentioned in the previous sections, each PE need to have proxy querier functionality for the following reasons:

1) To enable the collection of EVPN PEs providing L2VPN service to act as distributed multicast router with Anycast IP address for all attached hosts/VMs in that subnet.

2) To enable suppression of IGMP membership reports and queries over MPLS/IP core.

3) To enable generation of query messages locally to their attached host. In case of IGMPv1, the PE needs to send out an IGMP membership query to verify that at least one host on the subnet is still interested in receiving traffic directed to that group. When there is no reply to three consecutive IGMP membership queries, the PE times out the group, stops forwarding multicast traffic to the attached hosts for that (\*,G), and sends a EVPN SMET route associated with that (\*,G) with the version-1 flag reset or withdraws that route.

### 3 Operation

Consider the EVPN network of figure-1, where there is an EVPN instance configured across the PEs shown in this figure (namely PE1, PE2, and PE3). Lets consider that this EVPN instance consist of a single bridge domain (single subnet) with all the hosts, sources and the multicast router shown in this figure connected to this subnet. PE1 only has hosts connected to it. PE2 has a mix of hosts and multicast source. PE3 has a mix of hosts, multicast source, and multicast router. Further more, lets consider that for (S1,G1), R1 is used as the multicast router. The following subsections describe the IGMP proxy operation in different PEs with regard to whether the locally attached devices for that subnet are:

- only hosts/VMs
- mix of hosts/VMs and multicast source
- mix of hosts/VMs, multicast source, and multicast router

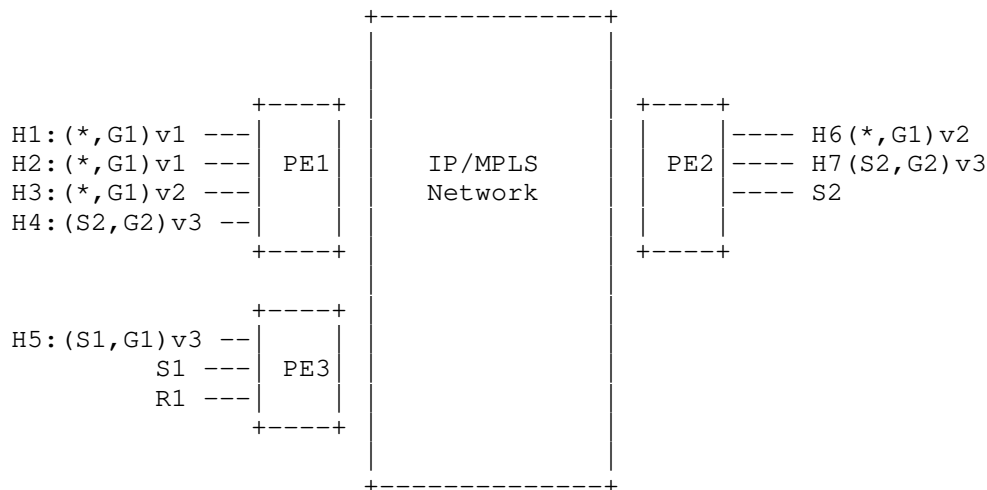


Figure 1:

#### 3.1 PE with only attached hosts/VMs for a given subnet

When PE1 receives an IGMPv1 Join Report from H1, it does not forward this join to any of its other ports (for this subnet) because all these local ports are associated with the hosts/VMs. PE1 sends an

EVPN Multicast Group route corresponding to this join for (\*,G1) and setting v1 flag. This EVPN route is received by PE2 and PE3 that are the member of the same BD (i.e., same EVI in case of VLAN-based service or <EVI,VLAN> in case of VLAN-aware bundle service). PE3 reconstructs IGMPv1 Join Report from this EVPN BGP route and only sends it to the port(s) with multicast routers attached to it (for that subnet). In this example, PE3 sends the reconstructed IGMPv1 Join Report for (\*,G1) to only R1. Furthermore, PE2 although receives the EVPN BGP route, it does not send it to any of its port for that subnet - namely ports associated with H6 and H7.

When PE1 receives the second IGMPv1 Join from H2 for the same multicast group (\*,G1), it only adds that port to its OIF list but it doesn't send any EVPN BGP route because there is no change in information. However, when it receives the IGMPv2 Join from H3 for the same (\*,G1), besides adding the corresponding port to its OIF list, it re-advertises the previously sent EVPN SMET route with the version-2 flag set.

Finally when PE1 receives the IMGMPv3 Join from H4 for (S2,G2), it advertises a new EVPN SMET route corresponding to it.

### 3.2 PE with mixed of attached hosts/VMs and multicast source

The main difference in here is that when PE2 receives IGMPv3 Join from H7 for (S2,G2), it does not advertises it in BGP because PE2 knows that S2 is attached to its local AC. PE2 adds the port associated with H7 to its OIF list for (S2,G2). The processing for IGMPv2 received from H6 is the same as the v2 Join described in previous section.

### 3.3 PE with mixed of attached hosts/VMs, multicast source and router

The main difference in here relative to the previous two sections is that Join messages received locally needs to be sent to the port associated with router R1. Furthermore, the Joins received via BGP need to be passed to the R1 port but filtered for all other ports.

## 4 All-Active Multi-Homing

Because a CE's LAG flow hashing algorithm is unknown, in an All-Active redundancy mode it must be assumed that the CE can send a given IGMP message to any one of the multi-homed PEs, either DF or non-DF - i.e., different IGMP Join messages can arrive at different PEs in the redundancy group and furthermore their corresponding Leave messages can arrive at PEs that are different from the ones received

the Join messages. Therefore, all PEs attached to a given ES must coordinate IGMP Join and Leave Group (x, G) state, where x may be either '\*' or a particular source S, for each BD on that ES. This allows the DF for that [ES, BD] to correctly advertise or withdraw a Selective Multicast Ethernet Tag (SMET) route for that (x, G) group in that BD when needed.

All-Active multihoming PEs for a given ES MUST support IGMP synch procedures described in this section if they want to perform IGMP proxy for hosts connects to that ES.

#### 4.1 Local IGMP Join Synchronization

When a PE, either DF or non-DF, receives, on a given multihomed ES operating in All-Active redundancy mode, an IGMP Membership Report for (x, G), it determines the BD to which the IGMP Membership Report belongs. If the PE doesn't already have local IGMP Join (x, G) state for that BD on that ES, it instantiates local IGMP Join (x, G) state and advertises a BGP IGMP Join Synch route for that [ES, BD]. Local IGMP Join (x, G) state refers to IGMP Join (x, G) state that is created as the result of processing an IGMP Membership Report for (x, G).

The IGMP Join Synch route carries the ES-Import RT for the ES on which the IGMP Membership Report was received. Thus it may only go to the PEs attached to that ES (and not any other PEs).

When a PE, either DF or non-DF, receives an IGMP Join Synch route it installs that route and if it doesn't already have IGMP Join (x, G) state for that [ES, BD], it instantiates that IGMP Join (x, G) state - i.e., IGMP Join (x, G) state is the union of local IGMP Join (x, G) state and installed IGMP Join Synch route. If the DF is not currently advertising (originating) a SMET route for that (x, G) group in that BD, it does so now.

When a PE, either DF or non-DF, deletes its local IGMP Join (x, G) state for that [ES, BD], it withdraws its BGP IGMP Join Synch route for that [ES, BD].

When a PE, either DF or non-DF, receives the withdrawal of an IGMP Join Synch route from another PE it removes that route. When a PE has no local IGMP Join (x, G) state and it has no installed IGMP Join Synch routes, it removes IGMP Join (x, G) state for that [ES, BD]. If the DF no longer has IGMP Join (x, G) state for that BD on any ES for which it is DF, it withdraws its SMET route for that (x, G) group in that BD.

I.e., A PE advertises an SMET route for that (x, G) group in that BD

when it has IGMP Join (x, G) state in that BD on at least one ES for which it is DF and it withdraws that SMET route when it does not have IGMP Join (x, G) state in that BD on any ES for which it is DF.

#### 4.2 Local IGMP Leave Group Synchronization

When a PE, either DF or non-DF, receives, on a given multihomed ES operating in All-Active redundancy mode, an IGMP Leave Group message for (x, G) from the attached CE, it determines the BD to which the IGMPv2 Leave Group belongs. Regardless of whether it has IGMP Join (x, G) state for that [ES, BD], it initiates the (x, G) leave group synchronization procedure, which consists of the following steps:

- 1) It computes the Maximum Response Time, which is the duration of (x, G) leave group synchronization procedure. This is the product of two locally configured values, Last Member Query Count and Last Member Query Interval (described in Section 3 of [RFC2236]), plus delta, the time it takes for a BGP advertisement to propagate between the PEs attached to the multihomed ES (delta is a consistently configured value on all PEs attached to the multihomed ES).
- 2) It starts the Maximum Response Time timer. Note that the receipt of subsequent IGMP Leave Group messages or BGP Leave Synch routes for (x, G) do not change the value of a currently running Maximum Response Time timer and are ignored by the PE.
- 3) It initiates the Last Member Query procedure described in Section 3 of [RFC2236]; viz, it sends a number of Group-Specific Query (x, G) messages (Last Member Query Count) at a fixed interval (Last Member Query Interval) to the attached CE.
- 4) It advertises an IGMP Leave Synch route for that that [ES, BD]. This route notifies the other multihomed PEs attached to the given multihomed ES that it has initiated an (x, G) leave group synchronization procedure; i.e., it carries the ES-Import RT for the ES on which the IGMP Leave Group was received. It also contains the Maximum Response Time and the Leave Group Synchronization Procedure Sequence number. The latter identifies the specific (x, G) leave group synchronization procedure initiated by the advertising PE, which increments the value whenever it initiates a procedure.
- 5) When the Maximum Response Timer expires, the PE that has advertised the IGMP Leave Synch route withdraws it.

##### 4.2.1 Remote Leave Group Synchronization

When a PE, either DF or non-DF, receives an IGMP Leave Synch route it

installs that route and it starts a timer for (x, G) on the specified [ES, BD] whose value is set to the Maximum Response Time in the received IGMP Leave Synch route. Note that the receipt of subsequent IGMPv2 Leave Group messages or BGP Leave Synch routes for (x, G) do not change the value of a currently running Maximum Response Time timer and are ignored by the PE.

#### 4.2.2 Common Leave Group Synchronization

If a PE attached to the multihomed ES receives an IGMP Membership Report for (x, G) before the Maximum Response Time timer expires, it advertises a BGP IGMP Join Synch route for that [ES, BD]. If it doesn't already have local IGMP Join (x, G) state for that [ES, BD], it instantiates local IGMP Join (x, G) state. If the DF is not currently advertising (originating) a SMET route for that (x, G) group in that BD, it does so now.

If a PE attached to the multihomed ES receives an IGMP Join Synch route for (x, G) before the Maximum Response Time timer expires, it installs that route and if it doesn't already have IGMP Join (x, G) state for that BD on that ES, it instantiates that IGMP Join (x, G) state. If the DF is not currently advertising (originating) a SMET route for that (x, G) group in that BD, it does so now.

When the Maximum Response Timer expires a PE that has advertised an IGMP Leave Synch route, withdraws it. Any PE attached to the multihomed ES, that started the Maximum Response Time and has no local IGMP Join (x, G) state and no installed IGMP Join Synch routes, it removes IGMP Join (x, G) state for that [ES, BD]. If the DF no longer has IGMP Join (x, G) state for that BD on any ES for which it is DF, it withdraws its SMET route for that (x, G) group in that BD.

### 5 Single-Active Multi-Homing

Note that to facilitate state synchronization after failover, the PEs attached to a multihomed ES operating in Single-Active redundancy mode should also coordinate IGMP Join (x, G) state. In this case all IGMP Join messages are received by the DF and distributed to the non-DF PEs using the procedures described above.

### 6 Selective Multicast Procedures for IR tunnels

If an ingress PE uses ingress replication, then for a given (x, G) group in a given BD:

- 1) It sends (x, G) traffic to the set of PEs not supporting IGMP



Proxy. This set consists of any PE that has advertised an Inclusive Multicast Tag route for the BD without the "IGMP Proxy Support" flag.

2) It sends (x, G) traffic to the set of PEs supporting IGMP Proxy and having listeners for that (x, G) group in that BD. This set consists of any PE that has advertised an Inclusive Multicast Tag route for the BD with the "IGMP Proxy Support" flag and that has advertised an SMET route for that (x, G) group in that BD.

If an ingress PE's Selective P-Tunnel for a given BD uses P2MP and all of the PEs in the BD support that tunnel type and IGMP, then for a given (x, G) group in a given BD it sends (x, G) traffic using the Selective P-Tunnel for that (x, G) group in that BD. This tunnel will include those PEs that have advertised an SMET route for that (x, G) group on that BD (for Selective P-tunnel) but it may include other PEs as well (for Aggregate Selective P-tunnel).

## 7 BGP Encoding

This document defines three new BGP EVPN routes to carry IGMP membership reports. This route type is known as:

- + 6 - Selective Multicast Ethernet Tag Route
- + 7 - IGMP Join Synch Route
- + 8 - IGMP Leave Synch Route

The detailed encoding and procedures for this route type is described in subsequent section.

### 7.1 Selective Multicast Ethernet Tag Route

An Selective Multicast Ethernet Tag route type specific EVPN NLRI consists of the following:

RD (8 octets)
Ethernet Tag ID (4 octets)
Multicast Source Length (1 octet)
Multicast Source Address (variable)
Multicast Group Length (1 octet)
Multicast Group Address (Variable)
Originator Router Length (1 octet)
Originator Router Address (variable)
Flags (1 octets) (optional)

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the one-octet optional flag field (if included). The Flags fields are defined as follows:

0	1	2	3	4	5	6	7
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+	-	+	-	+	-	+	-
+</							

The least significant bit, bit 7 indicates support for IGMP version 1.

The second least significant bit, bit 6 indicates support for IGMP version 2.

The third least significant bit, bit 5 indicates support for IGMP version 3.

The forth least significant bit, bit 4 indicates whether the (S, G) information carried within the route-type is of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

This EVPN route type is used to carry tenant IGMP multicast group information. The flag field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version

bits help associate IGMP version of receivers participating within the EVPN domain.

The include/exclude bit helps in creating filters for a given multicast route.

#### 7.1.1 Constructing the Selective Multicast Ethernet Tag route

This section describes the procedures used to construct the Selective Multicast Ethernet Tag (SMET) route. Support for this route type is optional.

The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE.

The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0  
EVI is VLAN-Aware Bundle service without translation - set to the customer VID for that BD  
EVI is VLAN-Aware Bundle service with translation - set to the normalized Ethernet Tag ID - e.g., normalized VID

The Multicast Source length MUST be set to length of multicast source address in bits. In case of a (\*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership report. In case of a (\*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group address in bits.

The Multicast Group is the Group address of the IGMP membership report.

The Originator Router Length is the length of the Originator Router address in bits.

The Originator Router Address is the IP address of Router Originating the prefix. It should be noted that using the "Originating Router's IP address" field is needed for local-bias procedures and may be needed for building inter-AS multicast underlay tunnels where BGP next hop can get over written.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

IGMP protocol is used to receive group membership information from hosts/VMs by TORs. Upon receiving the hosts/VMs expression of interest of a particular group membership, this information is then forwarded using Ethernet Multicast Source Group Route NLRI. The NLRI also keeps track of receiver's IGMP protocol version and any "source filtering" for a given group membership. All EVPN SMET routes are announced with per-EVI Route Target extended communities.

## 7.2 IGMP Join Synch Route

This EVPN route type is used to coordinate IGMP Join (x,G) state for a given BD between the PEs attached to a given ES operating in All-Active (or Single-Active) redundancy mode and it consists of following:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
Multicast Source Length (1 octet)
Multicast Source Address (variable)
Multicast Group Length (1 octet)
Multicast Group Address (Variable)
Originator Router Length (1 octet)
Originator Router Address (variable)
Flags (1 octet)

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the one-octet Flags field, whose fields are defined as follows:

```

      0  1  2  3  4  5  6  7
+---+---+---+---+---+---+---+
| reserved | IE|v3|v2|v1|
+---+---+---+---+---+---+---+

```

The least significant bit, bit 7 indicates support for IGMP version 1. The second least significant bit, bit 6 indicates support for IGMP version 2. The third least significant bit, bit 5 indicates support for IGMP version 3. The fourth least significant bit, bit 4 indicates whether the (S, G) information carried within the route-type is of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

The Flags field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version bits help associate IGMP version of receivers participating within the EVPN domain. The include/exclude bit helps in creating filters for a given multicast route.

#### 7.2.1 Constructing the IGMP Join Synch Route

This section describes the procedures used to construct the IGMP Join Synch route. Support for this route type is optional. If a PE does not support this route, then it MUST not indicate that it supports 'IGMP proxy' in Multicast Flag extended community for the EVIs corresponding to its multi-homed Ethernet Segments.

An IGMP Join Synch route MUST carry exactly one ES-Import Route Target extended community, the one that corresponds to the ES on which the IGMP Join was received. It MUST also carry exactly one EVI-RT EC, the one that corresponds to the EVI on which the IGMP Join was received. See Section 7.5 for details on how to form the EVI-RT EC.

The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE.

The Ethernet Segment Identifier (ESI) MUST be set to the 10-octet value defined for the ES.

The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0  
EVI is VLAN-Aware Bundle service without translation - set to the customer VID for the BD  
EVI is VLAN-Aware Bundle service with translation - set to the normalized Ethernet Tag ID - e.g., normalized VID

The Multicast Source length MUST be set to length of multicast source address in bits. In case of a (\*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership report. In case of a (\*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group address in bits.

The Multicast Group is the Group address of the IGMP membership report.

The Originator Router Length is the length of the Originator Router address in bits.

The Originator Router Address is the IP address of Router Originating the prefix.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

7.3 IGMP Leave Synch Route This EVPN route type is used to coordinate IGMP Leave Group (x,G) state for a given BD between the PEs attached to a given ES operating in All-Active (or Single-Active) redundancy mode and it consists of following:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
Multicast Source Length (1 octet)
Multicast Source Address (variable)
Multicast Group Length (1 octet)
Multicast Group Address (Variable)
Originator Router Length (1 octet)
Originator Router Address (variable)
Leave Group Synchronization # (4 octets)
Maximum Response Time (1 octet)
Flags (1 octet)

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the Maximum Response Time and the one-octet Flags field, whose fields are defined as follows:

0	1	2	3	4	5	6	7
+	-	+	-	+	-	+	-
	reserved				IE		v3
+	-	+	-	+	-	+	-
					v2		v1
+	-	+	-	+	-	+	-

The least significant bit, bit 7 indicates support for IGMP version 1. The second least significant bit, bit 6 indicates support for IGMP version 2. The third least significant bit, bit 5 indicates support for IGMP version 3. The fourth least significant bit, bit 4 indicates whether the (S, G) information carried within the route-type is of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

The Flags field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version bits help associate IGMP version of receivers participating within the EVPN domain. The include/exclude bit helps in creating filters for a given multicast route.

### 7.3.1 Constructing the IGMP Leave Synch Route

This section describes the procedures used to construct the IGMP Leave Synch route. Support for this route type is optional. If a PE does not support this route, then it MUST not indicate that it supports 'IGMP proxy' in Multicast Flag extended community for the EVIs corresponding to its multi-homed Ethernet Segments.

An IGMP Leave Synch route MUST carry exactly one ES-Import Route Target extended community, the one that corresponds to the ES on which the IGMP Leave was received. It MUST also carry exactly one EVI-RT EC, the one that corresponds to the EVI on which the IGMP Leave was received. See Section 7.5 for details on how to form the EVI-RT EC.

The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE.

The Ethernet Segment Identifier (ESI) MUST be set to the 10-octet value defined for the ES.

The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0  
EVI is VLAN-Aware Bundle service without translation - set to the customer VID for the BD  
EVI is VLAN-Aware Bundle service with translation - set to the normalized Ethernet Tag ID - e.g., normalized VID

The Multicast Source length MUST be set to length of multicast source address in bits. In case of a (\*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership report. In case of a (\*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group address in bits.



The Multicast Group is the Group address of the IGMP membership report.

The Originator Router Length is the length of the Originator Router address in bits.

The Originator Router Address is the IP address of Router Originating the prefix.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

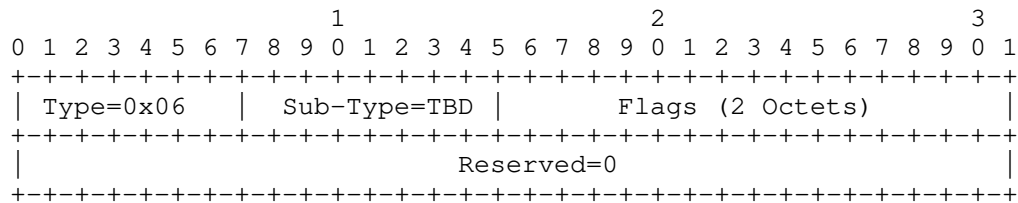
#### 7.4 Multicast Flags Extended Community

The 'Multicast Flags' extended community is a new EVPN extended community. EVPN extended communities are transitive extended communities with a Type field value of 6. IANA will assign a Sub-Type from the 'EVPN Extended Community Sub-Types' registry.

A PE that supports IGMP proxy on a given BD MUST attach this extended community to the Inclusive Multicast Ethernet Tag (IMET) route it advertises for that BD and it Must set the IGMP Proxy Support flag to 1. Note that an [RFC7432] compliant PE will not advertise this extended community so its absence indicates that the advertising PE does not support IGMP Proxy.

The advertisement of this extended community enables more efficient multicast tunnel setup from the source PE specially for ingress replication - i.e., if an egress PE supports IGMP proxy but doesn't have any interest in a given (x, G), it advertises its IGMP proxy capability using this extended community but it does not advertise any SMET route for that (x, G). When the source PE (ingress PE) receives such advertisements from the egress PE, it does not replicate the multicast traffic to that egress PE; however, it does replicate the multicast traffic to the egress PEs that don't advertise such capability even if they don't have any interests in that (x, G).

A Multicast Flags extended community is encoded as an 8-octet value, as follows:



The low-order bit of the Flags is defined as the "IGMP Proxy Support" bit. A value of 1 means that the PE supports IGMP Proxy as defined in this document, and a value of 0 means that the PE does not support IGMP proxy. The absence of this extended community also means that the PE doesn't support IGMP proxy.

### 7.5 EVI-RT Extended Community

In EVPN, every EVI is associated with one or more Route Targets (RTs). These Route Targets serve two functions:

- Distribution control: RTs control the distribution of the routes. If a route carries the RT associated with a particular EVI, it will be distributed to all the PEs on which that EVI exists.
- EVI Identification: Once a route has been received by a particular PE, the RT is used to identify the EVI to which it applies.

An IGMP Join Synch or IGMP Leave Synch route is associated with a particular combination of ES and EVI. These routes need to be distributed only to PEs that are attached to the associated ES. Therefore these routes carry the ES-Import RT for that ES.

Since an IGMP Join Synch or IGMP Leave Synch route does not need to be distributed to all the PEs on which the associated EVI exists, these routes cannot carry the RT associated with that EVI. Therefore, when such a route arrives at a particular PE, the route's RTs cannot be used to identify the EVI to which the route applies. Some other means of associating the route with an EVI must be used.

This document specifies four new Extended Communities (EC) that can be used to identify the EVI with which a route is associated, but which do not have any effect on the distribution of the route. These new ECs are known as the "Type 0 EVI-RT EC", the "Type 1 EVI-RT EC", the "Type 2 EVI-RT EC", and the "Type 3 EVI-RT EC".

A Type 0 EVI-RT EC is an EVPN EC (type 6) of sub-type 0xA.

A Type 1 EVI-RT EC is an EVPN EC (type 6) of sub-type 0xB.

A Type 2 EVI-RT EC is an EVPN EC (type 6) of sub-type 0xC.

A Type 3 EVI-RT EC is an EVPN EC (type 6) of sub-type TBD.

Each IGMP Join Synch or IGMP Leave Synch route MUST carry exactly one EVI-RT EC. The EVI-RT EC carried by a particular route is constructed as follows. Each such route is the result of having received an IGMP Join or an IGMP Leave message from a particular BD. We will say that the route is associated with that BD. For each BD, there is a corresponding RT that is used to ensure that routes "about" that BD are distributed to all PEs attached to that BD. So suppose a given IGMP Join Synch or Leave Synch route is associated with a given BD, say BD1, and suppose that the corresponding RT for BD1 is RT1. Then:

0. If RT1 is a Transitive Two-Octet AS-specific EC, then the EVI-RT EC carried by the route is a Type 0 EVI-RT EC. The value field of the Type 0 EVI-RT EC is identical to the value field of RT1.

1. If RT1 is a Transitive IPv4-Address-specific EC, then the EVI-RT EC carried by the route is a Type 1 EVI-RT EC. The value field of the Type 1 EVI-RT EC is identical to the value field of RT1.

2. If RT1 is a Transitive Four-Octet-specific EC, then the EVI-RT EC carried by the route is a Type 2 EVI-RT EC. The value field of the Type 2 EVI-RT EC is identical to the value field of RT1.

3. If RT1 is a Transitive IPv6-Address-specific EC, then the EVI-RT EC carried by the route is a Type 3 EVI-RT EC. The value field of the Type 3 EVI-RT EC is identical to the value field of RT1.

An IGMP Join Synch or Leave Synch route MUST carry exactly one EVI-RT EC.

Suppose a PE receives a particular IGMP Join Synch or IGMP Leave Synch route, say R1, and suppose that R1 carries an ES-Import RT that is one of the PE's Import RTs. If R1 has no EVI-RT EC, or has more than one EVI-RT EC, the PE MUST apply the "treat-as-withdraw" procedure of [RFC7606].

Note that an EVI-RT EC is not a Route Target Extended Community, is not visible to the RT Constrain mechanism [RFC4684], and is not intended to influence the propagation of routes by BGP.

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=0x06										Sub-Type=n										RT associated with EVI											
										RT associated with the EVI (cont.)																					

Where the value of 'n' is 0x0A, 0x0B, 0x0C, or 0x0D corresponding to EVI-RT type 0, 1, 2, or 3 respectively.

#### 7.6 Rewriting of RT ECs and EVI-RT ECs by ASBRs

There are certain situations in which an ES is attached to a set of PEs that are not all in the same AS, or not all operated by the same provider. In some such situations, the RT that corresponds to a particular EVI may be different in each AS. If a route is propagated from AS1 to AS2, an ASBR at the AS1/AS2 border may be provisioned with a policy that removes the RTs that are meaningful in AS1 and replaces them with the corresponding (i.e., RTs corresponding to the same EVIs) RTs that are meaningful in AS2. This is known as RT-rewriting.

Note that if a given route's RTs are rewritten, and the route carries an EVI-RT EC, the EVI-RT EC needs to be rewritten as well.

#### 8 Acknowledgement

#### 9 Security Considerations

Same security considerations as [RFC7432].

#### 10 IANA Considerations

IANA has allocated the following codepoints from the EVPN Extended Community sub-types registry.

0x09	Multicast Flags Extended Community	[this document]
0x0A	EVI-RT Type 0	[this document]
0x0B	EVI-RT Type 1	[this document]
0x0C	EVI-RT Type 2	[this document]

IANA is requested to allocate a new codepoint from the EVPN Extended Community sub-types registry for the following.

0x0D      EVI-RT Type 3      [this document]

IANA has allocated the following EVPN route types from the EVPN Route Type registry.

- 6 - Selective Multicast Ethernet Tag Route
- 7 - IGMP Join Synch Route
- 8 - IGMP Leave Synch Route

IANA is requested to create a registry, "Multicast Flags Extended Community Flags", in the BGP registry.

The Multicast Flags Extended Community contains a 16-bit Flags field. The bits are numbered 0-15, from low-order to high-order.

The registry should be initialized as follows:

0      : IGMP Proxy Support      [this document]  
1-15 : unassigned

The registration policy should be "Standards Action".

## 11 References

### 11.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4360] S. Sangli et al, "'BGP Extended Communities Attribute", February, 2006.
- [RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", February, 2015.

### 11.2 Informative References

- [ETREE-FMWK] Key et al., "A Framework for E-Tree Service over MPLS Network", draft-ietf-l2vpn-etree-frwk-03, work in progress, September 2013.
- [PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-05.txt, work in progress, October, 2013.
- [RFC4541] Christensen, M., Kimball, K., and F. Solensky,

"Considerations for IGMP and MLD snooping PEs", RFC 4541, 2006.

Authors' Addresses

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

Samir Thoria  
Cisco  
Email: sthoria@cisco.com

Keyur Patel  
Arrcus  
Email: keyur@arrcus.com

Derek Yeung  
Arrcus  
Email: derek@arrcus.com

John Drake  
Juniper  
Email: jdrake@juniper.net

Wen Lin  
Juniper  
Email: wlin@juniper.net

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: November 5, 2018

T. Morin, Ed.  
Orange  
R. Kebler, Ed.  
Juniper Networks  
G. Mirsky, Ed.  
ZTE Corp.  
May 4, 2018

Multicast VPN fast upstream failover  
draft-ietf-bess-mvpn-fast-failover-03

Abstract

This document defines multicast VPN extensions and procedures that allow fast failover for upstream failures, by allowing downstream PEs to take into account the status of Provider-Tunnels (P-tunnels) when selecting the upstream PE for a VPN multicast flow, and extending BGP MVPN routing so that a C-multicast route can be advertised toward a standby upstream PE.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on November 5, 2018.

## Copyright Notice

Copyright (c) 2018 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 (<https://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. Introduction . . . . .	3
2. Terminology . . . . .	3
3. UMH Selection based on tunnel status . . . . .	3
3.1. Determining the status of a tunnel . . . . .	4
3.1.1. mVPN tunnel root tracking . . . . .	5
3.1.2. PE-P Upstream link status . . . . .	5
3.1.3. P2MP RSVP-TE tunnels . . . . .	5
3.1.4. Leaf-initiated P-tunnels . . . . .	6
3.1.5. ((S, G)) counter information . . . . .	6
3.1.6. BFD Discriminator . . . . .	6
3.1.7. Per PE-CE link BFD Discriminator . . . . .	9
4. Standby C-multicast route . . . . .	10
4.1. Downstream PE behavior . . . . .	11
4.2. Upstream PE behavior . . . . .	12
4.3. Reachability determination . . . . .	13
4.4. Inter-AS . . . . .	13
4.4.1. Inter-AS procedures for downstream PEs, ASBR fast failover . . . . .	14
4.4.2. Inter-AS procedures for ASBRs . . . . .	14
5. Hot leaf standby . . . . .	14
6. Duplicate packets . . . . .	15
7. IANA Considerations . . . . .	15
8. Security Considerations . . . . .	15
9. Acknowledgments . . . . .	16
10. Contributor Addresses . . . . .	16
11. References . . . . .	18
11.1. Normative References . . . . .	18
11.2. Informative References . . . . .	18
Authors' Addresses . . . . .	19



## 1. Introduction

In the context of multicast in BGP/MPLS VPNs, it is desirable to provide mechanisms allowing fast recovery of connectivity on different types of failures. This document addresses failures of elements in the provider network that are upstream of PEs connected to VPN sites with receivers.

Section 3 describes local procedures allowing an egress PE (a PE connected to a receiver site) to take into account the status of P-tunnels to determine the Upstream Multicast Hop (UMH) for a given (C-S, C-G). This method does not provide a "fast failover" solution when used alone, but can be used with the following sections for a "fast failover" solution.

Section 4 describes protocol extensions that can speed up failover by not requiring any multicast VPN routing message exchange at recovery time.

Moreover, section 5 describes a "hot leaf standby" mechanism, that uses a combination of these two mechanisms. This approach has similarities with the solution described in [RFC7431] to improve failover times when PIM routing is used in a network given some topology and metric constraints.

## 2. Terminology

The terminology used in this document is the terminology defined in [RFC6513] and [RFC6514].

x-PMSI: I-PMSI or S-PMSI

## 3. UMH Selection based on tunnel status

Current multicast VPN specifications [RFC6513], section 5.1, describe the procedures used by a multicast VPN downstream PE to determine what the upstream multicast hop (UMH) is for a given (C-S,C-G).

The procedure described here is an OPTIONAL procedure that consists of having a downstream PE take into account the status of P-tunnels rooted at each possible upstream PEs, for including or not including each given PE in the list of candidate UMHs for a given (C-S,C-G) state. The result is that, if a P-tunnel is "down" (see Section 3.1), the PE that is the root of the P-tunnel will not be considered for UMH selection, which will result in the downstream PE to failover to the upstream PE which is next in the list of candidates.

A downstream PE monitors the status of the tunnels of UMHs that are ahead of the current one. Whenever the downstream PE determines that one of these tunnels is no longer "known to down", the PE selects the UMH corresponding to that as the new UMH.

More precisely, UMH determination for a given (C-S,C-G) will consider the UMH candidates in the following order:

- o first, the UMH candidates that either (a) advertise a PMSI bound to a tunnel, where the specified tunnel is not known to be down or (b) do not advertise any x-PMSI applicable to the given (C-S,C-G) but have associated a VRF Route Import BGP attribute to the unicast VPN route for S (this is necessary to avoid incorrectly invalidating an UMH PE that would use a policy where no I-PMSI is advertised for a given VRF and where only S-PMSI are used, the S-PMSI advertisement being possibly done only after the upstream PE receives a C-multicast route for (C-S, C-G)/(C-\*, C-G) to be carried over the advertised S-PMSI)
- o second, the UMH candidates that advertise a PMSI bound to a tunnel that is "down" -- these will thus be used as a last resort to ensure a graceful fallback to the basic MVPN UMH selection procedures in the hypothetical case where a false negative would occur when determining the status of all tunnels

For a given downstream PE and a given VRF, the P-tunnel corresponding to a given upstream PE for a given (C-S,C-G) state is the S-PMSI tunnel advertised by that upstream PE for this (C-S,C-G) and imported into that VRF, or if there isn't any such S-PMSI, the I-PMSI tunnel advertised by that PE and imported into that VRF.

Note that this document assumes that if a site of a given MVPN that contains C-S is dual-homed to two PEs, then all the other sites of that MVPN would have two unicast VPN routes (VPN-IPv4 or VPN-IPv6) routes to C-S, each with its own RD.

### 3.1. Determining the status of a tunnel

Different factors can be considered to determine the "status" of a P-tunnel and are described in the following sub-sections. The procedure proposed here also allows that all downstream PEs don't apply the same rules to define what the status of a P-tunnel is (please see Section 6), and some of them will produce a result that may be different for different downstream PEs. Thus what is called the "status" of a P-tunnel in this section, is not a characteristic of the tunnel in itself, but is the status of the tunnel, \*as seen from a particular downstream PE\*. Additionally, some of the following methods determine the ability of downstream PE to receive

traffic on the P-tunnel and not specifically on the status of the P-tunnel itself. This could be referred to as "P-tunnel reception status", but for simplicity, we will use the terminology of P-tunnel "status" for all of these methods.

Depending on the criteria used to determine the status of a P-tunnel, there may be an interaction with another resiliency mechanism used for the P-tunnel itself, and the UMH update may happen immediately or may need to be delayed. Each particular case is covered in each separate sub-section below.

#### 3.1.1. mVPN tunnel root tracking

A condition to consider that the status of a P-tunnel is up is that the root of the tunnel, as determined in the PMSI tunnel attribute, is reachable through unicast routing tables. In this case, the downstream PE can immediately update its UMH when the reachability condition changes.

This is similar to BGP next-hop tracking for VPN routes, except that the address considered is not the BGP next-hop address, but the root address in the PMSI tunnel attribute.

If BGP next-hop tracking is done for VPN routes and the root address of a given tunnel happens to be the same as the next-hop address in the BGP auto-discovery route advertising the tunnel, then this mechanisms may be omitted for this tunnel, as it will not bring any specific benefit.

#### 3.1.2. PE-P Upstream link status

A condition to consider a tunnel status as Up can be that the last-hop link of the P-tunnel is up.

This method should not be used when there is a fast restoration mechanism (such as MPLS FRR [RFC4090]) in place for the link.

#### 3.1.3. P2MP RSVP-TE tunnels

For P-tunnels of type P2MP MPLS-TE, the status of the P-tunnel is considered up if one or more of the P2MP RSVP-TE LSPs, identified by the P-tunnel Attribute, are in Up state. The determination of whether a P2MP RSVP-TE LSP is in Up state requires Path and Resv state for the LSP and is based on procedures in [RFC4875]. In this case, the downstream PE can immediately update its UMH when the reachability condition changes.

When signaling state for a P2MP TE LSP is removed (e.g. if the ingress of the P2MP TE LSP sends a PathTear message) or the P2MP TE LSP changes state from Up to Down as determined by procedures in [RFC4875], the status of the corresponding P-tunnel SHOULD be re-evaluated. If the P-tunnel transitions from up to Down state, the upstream PE, that is the ingress of the P-tunnel, SHOULD NOT be considered a valid UMH.

#### 3.1.4. Leaf-initiated P-tunnels

A PE can be removed from the UMH candidate list for a given ((S, G)) if the P-tunnel for this (S, G) (I or S, depending) is leaf triggered (PIM, mLDP), but for some reason internal to the protocol the upstream one-hop branch of the tunnel from P to PE cannot be built. In this case, the downstream PE can immediately update its UMH when the reachability condition changes.

#### 3.1.5. ((S, G)) counter information

In cases, where the downstream node can be configured so that the maximum inter-packet time is known for all the multicast flows mapped on a P-tunnel, the local per-(C-S,C-G) traffic counter information for traffic received on this P-tunnel can be used to determine the status of the P-tunnel.

When such a procedure is used, in the context where fast restoration mechanisms are used for the P-tunnels, downstream PEs should be configured to wait before updating the UMH, to let the P-tunnel restoration mechanism happen. A configurable timer MUST be provided for this purpose, and it is recommended to provide a reasonable default value for this timer.

This method can be applicable, for instance, when a ((S, G)) flow is mapped on an S-PMSI.

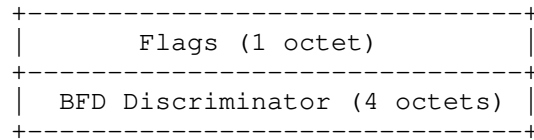
In cases where this mechanism is used in conjunction with Hot leaf standby, then no prior knowledge of the rate of the multicast streams is required; downstream PEs can compare reception on the two P-tunnels to determine when one of them is down.

#### 3.1.6. BFD Discriminator

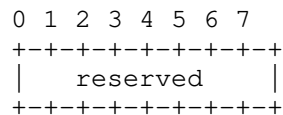
P-tunnel status can be derived from the status of a multipoint BFD session [I-D.ietf-bfd-multipoint] whose discriminator is advertised along with an x-PMSI A-D route.

This document defines the format and ways of using a new BGP attribute called the "BGP- BFD attribute". This is an optional

transitive BGP attribute. The format of this attribute is defined as follows:



The Flags field has the following format:



### 3.1.6.1. Upstream PE Procedures

When it is desired to track the P-tunnel status using p2mp BFD session, the Upstream PE:

- o MUST initiate BFD session and set `bfd.SessionType = MultipointHead` as described in [I-D.ietf-bfd-multipoint];
- o MUST use [Ed.note] address as destination IP address when transmitting BFD control packets;
- o MUST use the IP address of the Upstream PE as source IP address when transmitting BFD control packets;
- o MUST include the BGP-BFD Attribute in the x-PMSI A-D Route with BFD Discriminator value set to My Discriminator value.

If tracking of the P-tunnel by using a p2mp BFD session is to be enabled after the P-tunnel has been already signaled, the the procedure described above MUST be followed. Note that x-PMSI A-D Route MUST be re-sent with exactly the same attributes as before and the BGP-BFD Attribute included.

If P-tunnel is already signaled, and P-tunnel status tracked using the p2mp BFD session and it is desired to stop tracking P-tunnel status using BFD, then:

- o x-PMSI A-D Route MUST be re-sent with exactly the same attributes as before, but the BGP-BFD Attribute MUST be excluded;
- o the p2mp BFD session SHOULD be deleted.

#### 3.1.6.2. Downstream PE Procedures

On receiving the BGP-BFD Attribute in the x-PMSI A-D Route, the Downstream PE:

- o MUST associate the received BFD discriminator value with the P-tunnel originating from the Root PE;
- o MUST create p2mp BFD session and set `bfd.SessionType = MultipointTail` as described in [I-D.ietf-bfd-multipoint];
- o MUST use the source IP address of a BFD control packet, the value of BFD Discriminator from the BGP-BFD Attribute to properly demultiplex BFD sessions;

After the state of the p2mp BFD session is up, i.e. `bfd.SessionState = Up`, the session state will then be used to track the health of the P-tunnel.

According to [I-D.ietf-bfd-multipoint], if the Downstream PE receives Down or AdminDown in the State field of the BFD control packet or associated with the BFD session Detection Timer expires, the BFD session state is down, i.e. `bfd.SessionState = Down`. When the BFD session state is Down, then the P-tunnel associated with the BFD session as down MUST be declared down. Then The Downstream PE MAY initiate a switchover of the traffic from the Primary Upstream PE to the Standby Upstream PE.

If the Downstream PE's P-tunnel is already up when the Downstream PE receives the new x-PMSI A-D Route with BGP-BFD Attribute, the Downstream PE MUST accept the x-PMSI A-D Route and associate the value of BFD Discriminator field with the P-tunnel. The Upstream PE MUST follow procedures listed above in this section to bring the p2mp BFD session up and use it to monitor the state of the associated P-tunnel.

If the Downstream PE's P-tunnel is already up, its state being monitored by the p2mp BFD session, and the Downstream PE receives the

new x-PMSI A-D Route without the BGP-BFD Attribute, the Downstream PE:

- o MUST accept the x-PMSI A-D Route;
- o MUST stop receiving BFD control packets for this p2mp BFD session;
- o SHOULD delete the p2mp BFD session associated with the P-tunnel;
- o SHOULD NOT switch the traffic to the Standby Upstream PE.

When such a procedure is used, in the context where fast restoration mechanisms are used for the P-tunnels, leaf PEs should be configured to wait before updating the UMH, to let the P-tunnel restoration mechanism happen. A configurable timer MUST be provided for this purpose, and it is recommended to provide a reasonable default value for this timer.

### 3.1.7. Per PE-CE link BFD Discriminator

The following approach is defined for the fast failover in response to the detection of PE-CE link failures, in which UMH selection for a given C-multicast route takes into account the state of the BFD session associated with the state of the upstream PE-CE link.

#### 3.1.7.1. Upstream PE Procedures

For each protected PE-CE link, the upstream PE initiates a multipoint BFD session [I-D.ietf-bfd-multipoint] as MultipointHead toward downstream PEs. A downstream PE monitors the state of the p2mp session as MultipointTail and MAY interpret transition of the BFD session into Down state as the indication of the associated PE-CE link being down.

For SSM groups, the upstream PE advertises an ((S, G)) S-PMSI A-D route or wildcard (S,\*) S-PMSI A-D route for each received SSM ((S, G)) C-multicast route for which protection is desired. For each ASM ((S, G)) C-multicast route for which protection is desired, the upstream PE advertises a ((S, G)) S-PMSI A-D route. For each ASM (\*,G) C-Multicast route for which protection is desired, the upstream PE advertises a wildcard (\*,G) S-PMSI A-D route. Note that all S-PMSI A-D routes can signal the same P-tunnel, so there is no need for a new P-tunnel for each S-PMSI A-D route. Multicast flows for which protection is desired is controlled by configuration/policy on the upstream PE. The protected link is the RPF PE-CE interface towards the src/RP. The upstream PE advertises the BFD discriminator of the protected link in the S-PMSI A-D route. If the route to the src/RP changes such that the RPF interface is changed to be a new PE-

CE interface, then the upstream PE will update the S-PMSI A-D route with included BGP-BFD Attribute so that value of the BFD Discriminator is associated with the new RPF link.

#### 3.1.7.2. Downstream PE Procedures

If an S-PMSI A-D route bound to a given C-multicast is signaled with a multipoint BFD session, then the upstream PE is considered during UMH selection for the C-multicast if and only if the corresponding BFD session is not in state Down, i.e `bfd.SessionState != Down`. Whenever the state of the BFD session changes to Down the Provider Tunnel will be considered down, and the downstream PE will switch to the backup Provider Tunnel. Note that the Provider Tunnel is considered down only for the C-multicast states that match to an S-PMSI A-D route which included BGP-BFD Attribute with the BFD Discriminator of the p2mp BFD session which is down.

#### 4. Standby C-multicast route

The procedures described below are limited to the case where the site that contains C-S is connected to exactly two PEs. The procedures require all the PEs of that MVPN to follow the single forwarder PE selection, as specified in [RFC6513]. The procedures assume that if a site of a given MVPN that contains C-S is dual-homed to two PEs, then all the other sites of that MVPN would have two unicast VPN routes (VPN-IPv4 or VPN-IPv6) routes to C-S, each with its own RD.

As long as C-S is reachable via both PEs, a given downstream PE will select one of the PEs connected to C-S as its Upstream PE with respect to C-S. We will refer to the other PE connected to C-S as the "Standby Upstream PE". Note that if the connectivity to C-S through the Primary Upstream PE becomes unavailable, then the PE will select the Standby Upstream PE as its Upstream PE with respect to C-S. When the Primary PE later becomes available, then the PE will select the Primary Upstream PE again as its Upstream PE. This is referred to as "revertive" behavior and MUST be supported. Non-revertive behavior would refer to the behavior of continuing to select the backup PE as the UMH even after the Primary has come up. This non-revertive behavior can also be optionally supported by an implementation and would be enabled through some configuration.

For readability, in the following sub-sections, the procedures are described for BGP C-multicast Source Tree Join routes, but they apply equally to BGP C-multicast Shared Tree Join routes failover for the case where the customer RP is dual-homed (substitute "C-RP" to "C-S").



#### 4.1. Downstream PE behavior

When a (downstream) PE connected to some site of an MVPN needs to send a C-multicast route (C-S, C-G), then following the procedures specified in Section "Originating C-multicast routes by a PE" of [RFC6514] the PE sends the C-multicast route with RT that identifies the Upstream PE selected by the PE originating the route. As long as C-S is reachable via the Primary Upstream PE, the Upstream PE is the Primary Upstream PE. If C-S is reachable only via the Standby Upstream PE, then the Upstream PE is the Standby Upstream PE.

If C-S is reachable via both the Primary and the Standby Upstream PE, then in addition to sending the C-multicast route with an RT that identifies the Primary Upstream PE, the PE also originates and sends a C-multicast route with an RT that identifies the Standby Upstream PE. This route, that has the semantics of being a 'standby' C-multicast route, is further called a "Standby BGP C-multicast route", and is constructed as follows:

- o the NLRI is constructed as the original C-multicast route, except that the RD is the same as if the C-multicast route was built using the standby PE as the UMH (it will carry the RD associated to the unicast VPN route advertised by the standby PE for S)
- o SHOULD carry the "Standby PE" BGP Community (this is a new BGP Community, see Section 7)

The normal and the standby C-multicast routes must have their Local Preference attribute adjusted so that, if two C-multicast routes with same NLRI are received by a BGP peer, one carrying the "Standby PE" attribute and the other one *\*not\** carrying the "Standby PE" community, then preference is given to the one *\*not\** carrying the "Standby PE" attribute. Such a situation can happen when, for instance, due to transient unicast routing inconsistencies, two different downstream PEs consider different upstream PEs to be the primary one; in that case, without any precaution taken, both upstream PEs would process a standby C-multicast route and possibly stop forwarding at the same time. For this purpose, routes that carry the "Standby PE" BGP Community MUST have the LOCAL\_PREF attribute set to zero.

Note that, when a PE advertises such a Standby C-multicast join for an ((S, G)) it must join the corresponding P-tunnel.

If at some later point the local PE determines that C-S is no longer reachable through the Primary Upstream PE, the Standby Upstream PE becomes the Upstream PE, and the local PE re-sends the C-multicast route with RT that identifies the Standby Upstream PE, except that

now the route does not carry the Standby PE BGP Community (which results in replacing the old route with a new route, with the only difference between these routes being the presence/absence of the Standby PE BGP Community).

#### 4.2. Upstream PE behavior

When a PE receives a C-multicast route for a particular (C-S, C-G), and the RT carried in the route results in importing the route into a particular VRF on the PE, if the route carries the Standby PE BGP Community, then the PE performs as follows:

when the PE determines that C-S is not reachable through some other PE, the PE SHOULD install VRF PIM state corresponding to this Standby BGP C-multicast route (the result will be that a PIM Join message will be sent to the CE towards C-S, and that the PE will receive (C-S,C-G) traffic), and the PE SHOULD forward (C-S, C-G) traffic received by the PE to other PEs through a P-tunnel rooted at the PE.

Furthermore, irrespective of whether C-S carried in that route is reachable through some other PE:

- a) based on local policy, as soon as the PE receives this Standby BGP C-multicast route, the PE MAY install VRF PIM state corresponding to this BGP Source Tree Join route (the result will be that Join messages will be sent to the CE toward C-S, and that the PE will receive (C-S,C-G) traffic)
- b) based on local policy, as soon as the PE receives this Standby BGP C-multicast route, the PE MAY forward (C-S, C-G) traffic to other PEs through a P-tunnel independently of the reachability of C-S through some other PE. [note that this implies also doing (a)]

Doing neither (a) or (b) for a given (C-S,C-G) is called "cold root standby".

Doing (a) but not (b) for a given (C-S,C-G) is called "warm root standby".

Doing (b) (which implies also doing (a)) for a given (C-S,C-G) is called "hot root standby".

Note that, if an upstream PE uses an S-PMSI only policy, it shall advertise an S-PMSI for an ((S, G)) as soon as it receives a C-multicast route for ((S, G)), normal or Standby; i.e. it shall not wait for receiving a non-Standby C-multicast route before advertising the corresponding S-PMSI.

Section 9.3.2 of [RFC6514], describes the procedures of sending a Source-Active A-D result as a result of receiving the C-multicast route. These procedures should be followed for both the normal and Standby C-multicast routes.

#### 4.3. Reachability determination

The standby PE can use the following information to determine that C-S can or cannot be reached through the primary PE:

- o presence/absence of a unicast VPN route toward C-S
- o supposing that the standby PE is an egress of the tunnel rooted at the Primary PE, the standby PE can determine the reachability of C-S through the Primary PE based on the status of this tunnel, determined thanks to the same criteria as the ones described in Section 3.1 (without using the UMH selection procedures of Section 3)
- o other mechanisms MAY be used

#### 4.4. Inter-AS

If the non-segmented inter-AS approach is used, the procedures in section 4 can be applied.

When multicast VPNs are used in an inter-AS context with the segmented inter-AS approach described in section 8.2 of [RFC6514], the procedures in this section can be applied.

A pre-requisite for the procedures described below to be applied for a source of a given MVPN is:

- o that any PE of this MVPN receives two Inter-AS I-PMSI auto-discovery routes advertised by the AS of the source (or more)
- o that these Inter-AS I-PMSI auto-discovery routes have distinct Route Distinguishers (as described in item "(2)" of section 9.2 of [RFC6514]).

As an example, these conditions will be satisfied when the source is dual-homed to an AS that connects to the receiver AS through two ASBR using auto-configured RDs.

#### 4.4.1. Inter-AS procedures for downstream PEs, ASBR fast failover

The following procedure is applied by downstream PEs of an AS, for a source S in a remote AS.

Additionally, to choosing an Inter-AS I-PMSI auto-discovery route advertised from the AS of the source to construct a C-multicast route, as described in section 11.1.3 [RFC6514] a downstream PE will choose a second Inter-AS I-PMSI auto-discovery route advertised from the AS of the source and use this route to construct and advertise a Standby C-multicast route (C-multicast route carrying the Standby extended community) as described in Section 4.1.

#### 4.4.2. Inter-AS procedures for ASBRs

When an upstream ASBR receives a C-multicast route, and at least one of the RTs of the route matches one of the ASBR Import RT, the ASBR locates an Inter-AS I-PMSI A-D route whose RD and Source AS matches the RD and Source AS carried in the C-multicast route. If the match is found, and C-multicast route carries the Standby PE BGP Community, then the ASBR performs as follows:

- o if the route was received over iBGP; the route is expected to have a LOCAL\_PREF attribute set to zero and it should be re-advertised in eBGP with a MED attribute (MULTI\_EXIT\_DISC) set to the highest possible value (0xffff)
- o if the route was received over eBGP; the route is expected to have a MED attribute set of 0xffff and should be re-advertised in iBGP with a LOCAL\_PREF attribute set to zero

Other ASBR procedures are applied without modification.

### 5. Hot leaf standby

The mechanisms defined in sections Section 4 and Section 3 can be used together as follows.

The principle is that, for a given VRF (or possibly only for a given C-S,C-G):

- o downstream PEs advertise a Standby BGP C-multicast route (based on Section 4)
- o upstream PEs use the "hot standby" optional behavior and thus will forward traffic for a given multicast state as soon as they have whether a (primary) BGP C-multicast route or a Standby BGP C-multicast route for that state (or both)

- o downstream PEs accept traffic from the primary or standby tunnel, based on the status of the tunnel (based on Section 3)

Other combinations of the mechanisms proposed in Section 4) and Section 3 are for further study.

Note that the same level of protection would be achievable with a simple C-multicast Source Tree Join route advertised to both the primary and secondary upstream PEs (carrying as Route Target extended communities, the values of the VRF Route Import attribute of each VPN route from each upstream PEs). The advantage of using the Standby semantic for is that, supposing that downstream PEs always advertise a Standby C-multicast route to the secondary upstream PE, it allows to choose the protection level through a change of configuration on the secondary upstream PE, without requiring any reconfiguration of all the downstream PEs.

## 6. Duplicate packets

Multicast VPN specifications [RFC6513] impose that a PE only forwards to CEs the packets coming from the expected upstream PE (Section 9.1).

We highlight the reader's attention to the fact that the respect of this part of multicast VPN specifications is especially important when two distinct upstream PEs are susceptible to forward the same traffic on P-tunnels at the same time in the steady state. This will be the case when "hot root standby" mode is used (Section 4), and which can also be the case if procedures of Section 3 are used and (a) the rules determining the status of a tree are not the same on two distinct downstream PEs or (b) the rule determining the status of a tree depend on conditions local to a PE (e.g. the PE-P upstream link being up).

## 7. IANA Considerations

Allocation is expected from IANA for the BGP "Standby PE" community. (TBC)

[Note to RFC Editor: this section may be removed on publication as an RFC.]

## 8. Security Considerations

## 9. Acknowledgments

The authors want to thank Greg Reaume, Eric Rosen, and Jeffrey Zhang for their review and useful feedback.

## 10. Contributor Addresses

Below is a list of other contributing authors in alphabetical order:

Rahul Aggarwal  
Arktan

Email: raggarwa\_1@yahoo.com

Nehal Bhau  
Alcatel-Lucent, Inc.  
701 E Middlefield Rd  
Mountain View, CA 94043  
USA

Email: Nehal.Bhau@alcatel-lucent.com

Clayton Hassen  
Bell Canada  
2955 Virtual Way  
Vancouver  
CANADA

Email: Clayton.Hassen@bell.ca

Wim Henderickx  
Alcatel-Lucent  
Copernicuslaan 50  
Antwerp 2018  
Belgium

Email: wim.henderickx@alcatel-lucent.com

Pradeep Jain  
Alcatel-Lucent, Inc.

701 E Middlefield Rd  
Mountain View, CA 94043  
USA

Email: [pradeep.jain@alcatel-lucent.com](mailto:pradeep.jain@alcatel-lucent.com)

Jayant Kotalwar  
Alcatel-Lucent, Inc.  
701 E Middlefield Rd  
Mountain View, CA 94043  
USA

Email: [Jayant.Kotalwar@alcatel-lucent.com](mailto:Jayant.Kotalwar@alcatel-lucent.com)

Praveen Muley  
Alcatel-Lucent  
701 East Middlefield Rd  
Mountain View, CA 94043  
U.S.A.

Email: [praveen.muley@alcatel-lucent.com](mailto:praveen.muley@alcatel-lucent.com)

Ray (Lei) Qiu  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, CA 94089  
U.S.A.

Email: [rqiujuniper.net](mailto:rqiujuniper.net)

Yakov Rekhter  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, CA 94089  
U.S.A.

Email: [yakov@juniper.net](mailto:yakov@juniper.net)

Kanwar Singh

Alcatel-Lucent, Inc.  
701 E Middlefield Rd  
Mountain View, CA 94043  
USA

Email: kanwar.singh@alcatel-lucent.com

## 11. References

### 11.1. Normative References

- [I-D.ietf-bfd-multipoint]  
Katz, D., Ward, D., Networks, J., and G. Mirsky, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-16 (work in progress), April 2018.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <<https://www.rfc-editor.org/info/rfc4875>>.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <<https://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, <<https://www.rfc-editor.org/info/rfc6514>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

### 11.2. Informative References



- [RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <<https://www.rfc-editor.org/info/rfc4090>>.
- [RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <<https://www.rfc-editor.org/info/rfc7431>>.

## Authors' Addresses

Thomas Morin (editor)  
Orange  
2, avenue Pierre Marzin  
Lannion 22307  
France

Email: [thomas.morin@orange-ftgroup.com](mailto:thomas.morin@orange-ftgroup.com)

Robert Kebler (editor)  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, CA 94089  
U.S.A.

Email: [rkebler@juniper.net](mailto:rkebler@juniper.net)

Greg Mirsky (editor)  
ZTE Corp.

Email: [gregimirsky@gmail.com](mailto:gregimirsky@gmail.com)

BESS Workgroup  
Internet-Draft  
Intended status: Standards Track  
Expires: December 28, 2018

P. Jain, Ed.  
S. Salam  
A. Sajassi  
Cisco Systems, Inc.  
S. Boutros  
VmWare, Inc.  
G. Mirsky  
ZTE Corporation.  
June 26, 2018

LSP-Ping Mechanisms for EVPN and PBB-EVPN  
draft-jain-bess-evpn-lsp-ping-07

Abstract

LSP-Ping is a widely deployed Operation, Administration, and Maintenance (OAM) mechanism in MPLS networks. This document describes mechanisms for detecting data-plane failures using LSP Ping in MPLS based EVPN and PBB-EVPN networks.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on December 28, 2018.

Copyright Notice

Copyright (c) 2018 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 (<https://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. Introduction . . . . .	2
2. Specification of Requirements . . . . .	3
3. Terminology . . . . .	3
4. Proposed Target FEC Stack Sub-TLVs . . . . .	3
4.1. EVPN MAC Sub-TLV . . . . .	4
4.2. EVPN Inclusive Multicast Sub-TLV . . . . .	4
4.3. EVPN Auto-Discovery Sub-TLV . . . . .	5
4.4. EVPN IP Prefix Sub-TLV . . . . .	6
5. Encapsulation of OAM Ping Packets . . . . .	7
6. Operations . . . . .	7
6.1. Unicast Data-plane connectivity checks . . . . .	7
6.2. Inclusive Multicast Data-plane Connectivity Checks . . . . .	8
6.2.1. Ingress Replication . . . . .	9
6.2.2. Using P2MP P-tree . . . . .	10
6.2.3. Controlling Echo Responses when using P2MP P-tree . . . . .	11
6.3. EVPN Aliasing Data-plane connectivity check . . . . .	11
6.4. EVPN IP Prefix (RT-5) Data-plane connectivity check . . . . .	11
7. Security Considerations . . . . .	12
8. IANA Considerations . . . . .	12
8.1. Sub-TLV Type . . . . .	12
8.2. Proposed new Return Codes . . . . .	12
9. Acknowledgments . . . . .	12
10. References . . . . .	13
10.1. Normative References . . . . .	13
10.2. Informative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Introduction

[RFC7432] describes MPLS based Ethernet VPN (EVPN) technology. An EVPN comprises CE(s) connected to PE(s). The PEs provide layer 2 EVPN among the CE(s) over the MPLS core infrastructure. In EVPN networks, PEs advertise the MAC addresses learned from the locally connected CE(s), along with MPLS Label, to remote PE(s) in the control plane using multi-protocol BGP. EVPN enables multi-homing of CE(s) connected to multiple PEs and load balancing of traffic to and from multi-homed CE(s).

[RFC7623] describes the use of Provider Backbone Bridging [802.1ah] with EVPN. PBB-EVPN maintains the C-MAC learning in data plane and

only advertises Provider Backbone MAC (B-MAC) addresses in control plane using BGP.

Procedures for simple and efficient mechanisms to detect data-plane failures using LSP Ping in MPLS network are well defined in [RFC8029][RFC6425]. This document defines procedures to detect data-plane failures using LSP Ping in MPLS networks deploying EVPN and PBB-EVPN. This draft defines 4 new Sub-TLVs for Target FEC Stack TLV with the purpose of identifying the FEC on the Peer PE.

## 2. 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].

## 3. Terminology

AD: Auto Discovery

B-MAC: Backbone MAC Address

CE: Customer Edge Device

C-MAC: Customer MAC Address

DF: Designated Forwarder

ESI: Ethernet Segment Identifier

EVI: EVPN Instance Identifier that globally identifies the EVPN Instance

EVPN: Ethernet Virtual Private Network

MPLS-OAM: MPLS Operations, Administration, and Maintenance

P2MP: Point-to-Multipoint

PBB: Provider Backbone Bridge

PE: Provider Edge Device

## 4. Proposed Target FEC Stack Sub-TLVs

This document introduces four new Target FEC Stack sub-TLVs that are included in the LSP-Ping Echo Request packet sent for detecting

faults in data-plane connectivity in EVPN and PBB-EVPN networks. These Target FEC Stack sub-TLVs are described next.

#### 4.1. EVPN MAC Sub-TLV

The EVPN MAC sub-TLV is used to identify the MAC for an EVI under test at a peer PE.

The EVPN MAC sub-TLV fields are derived from the MAC/IP advertisement route defined in [RFC7432] Section 7.2 and have the format as shown in Figure 1. This TLV is included in the Echo Request sent to the Peer PE by the PE that is the originator of the request.

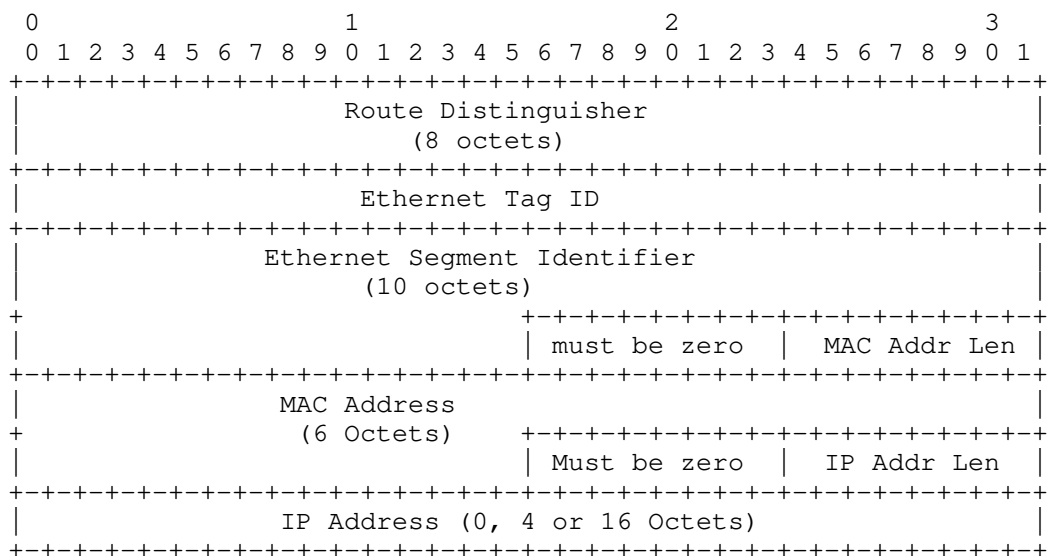


Figure 1: EVPN MAC sub-TLV format

The LSP Ping echo request is sent using the EVPN MPLS label(s) associated with the MAC route announced by a remote PE and the MPLS transport label(s) to reach the remote PE.

#### 4.2. EVPN Inclusive Multicast Sub-TLV

The EVPN Inclusive Multicast sub-TLV fields are based on the EVPN Inclusive Multicast route defined in [RFC7432] Section 7.3.

The EVPN Inclusive Multicast sub-TLV has the format as shown in Figure 2. This TLV is included in the echo request sent to the EVPN

peer PE by the originator of request to verify the multicast connectivity state on the peer PE(s) in EVPN and PBB-EVPN.

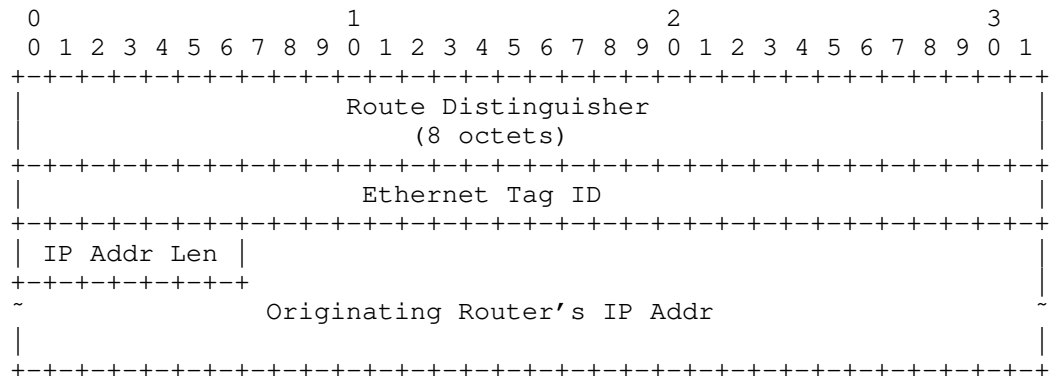


Figure 2: EVPN Inclusive Multicast sub-TLV format

Broadcast, multicast, and unknown unicast traffic can be sent using ingress replication or P2MP P-tree in EVPN and PBB-EVPN network. In case of ingress replication, the Echo Request is sent using a label stack of [Transport label, Inclusive Multicast label] to each remote PE participating in EVPN or PBB-EVPN. The inclusive multicast label is the downstream assigned label announced by the remote PE to which the Echo Request is being sent. The Inclusive Multicast label is the inner label in the MPLS label stack.

When using P2MP P-tree in EVPN or PBB-EVPN, the Echo Request is sent using P2MP P-tree transport label for inclusive P-tree arrangement or using a label stack of [P2MP P-tree transport label, upstream assigned EVPN Inclusive Multicast label] for the aggregate inclusive P2MP P-tree arrangement as described in Section 6.

In case of EVPN, an additional, EVPN Auto-Discovery sub-TLV and ESI MPLS label as the bottom label, may also be included in the Echo Request as is described in Section 6.

#### 4.3. EVPN Auto-Discovery Sub-TLV

The EVPN Auto-Discovery (AD) sub-TLV fields are based on the Ethernet AD route advertisement defined in [RFC7432] Section 7.1. EVPN AD sub-TLV applies to only EVPN.

The EVPN AD sub-TLV has the format shown in Figure 3.

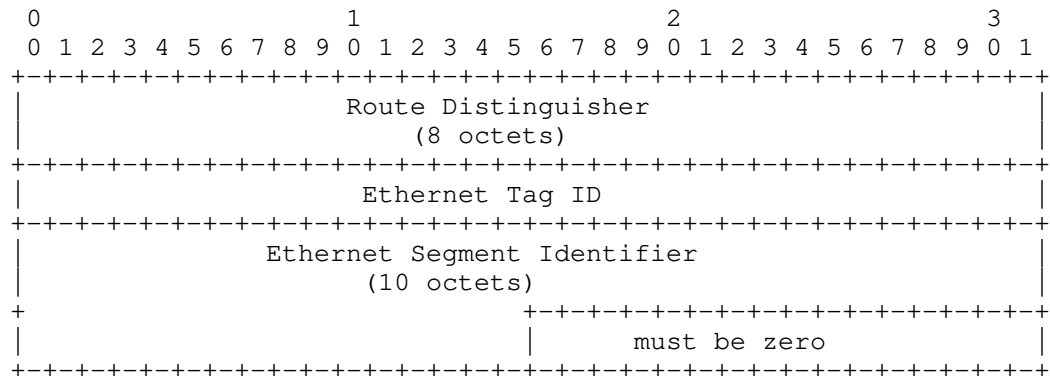


Figure 3: EVPN Auto-Discovery sub-TLV format

#### 4.4. EVPN IP Prefix Sub-TLV

The EVPN IP Prefix sub-TLV is used to identify the IP Prefix for an EVI under test at a peer PE.

The EVPN IP Prefix sub-TLV fields are derived from the IP Prefix Route (RT-5) advertisement defined in [I-D.ietf-bess-evpn-prefix-advertisement] and has the format as shown in Figure 4. This TLV is included in the Echo Request sent to the Peer PE by the PE that is the originator of the request.

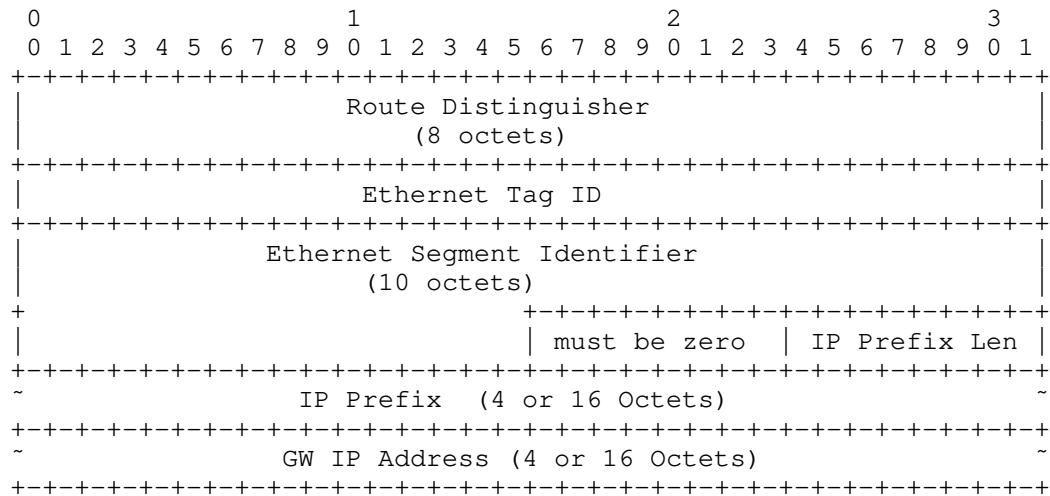


Figure 4: EVPN IP Prefix sub-TLV format

The LSP Ping echo request is sent using the EVPN MPLS label(s) associated with the IP Prefix route announced by a remote PE and the MPLS transport label(s) to reach the remote PE.

## 5. Encapsulation of OAM Ping Packets

The LSP Ping Echo request IPv4/UDP packets are encapsulated with the Transport and EVPN Label(s) followed by the Generic Associated Channel Label (GAL) [RFC6426] which is the bottom most label. The GAL label is followed by IPv4(0x0021) or IPv6(0x0057) Associated Channel Header (ACH) [RFC4385].

## 6. Operations

### 6.1. Unicast Data-plane connectivity checks

Figure 5 is an example of a PBB-EVPN network. CE1 is dual-homed to PE1 and PE2. Assume, PE1 announced a MAC route with RD 1.1.1.1:00 and B-MAC 00aa.00bb.00cc and with MPLS label 16001 for EVI 10. Similarly, PE2 announced a MAC route with RD 2.2.2.2:00 and B-MAC 00aa.00bb.00cc and with MPLS label 16002.

On PE3, when an operator performs a connectivity check for the B-MAC address 00aa.00bb.00cc on PE1, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN MAC sub-TLV in the Echo Request packet. The Echo Request packet is sent with the {Transport Label(s) to reach PE1 + EVPN Label = 16001 + GAL} MPLS



label stack and IP ACH Channel header. Once the echo request packet reaches PE1, PE1 will use the GAL label and the IP ACH Channel header to determine that the packet is IPv4 OAM Packet. The PE1 will process the packet and perform checks for the EVPN MAC sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC8029] and respond according to [RFC8029] processing rules.

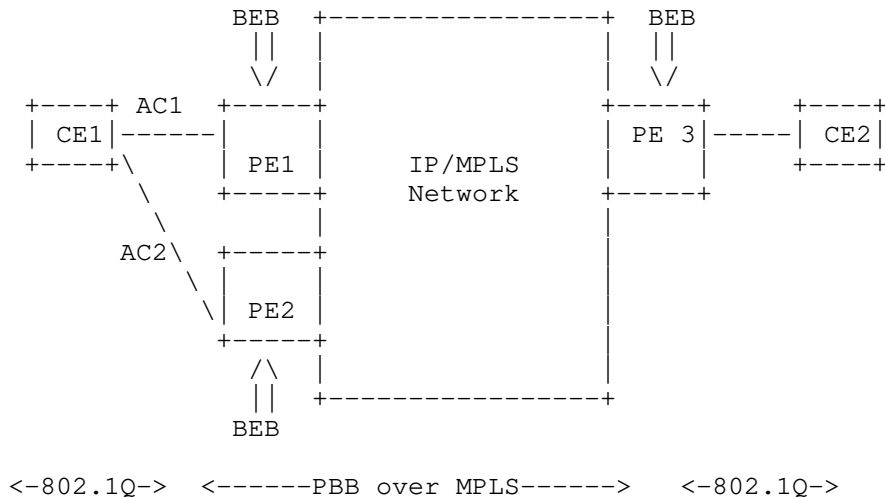


Figure 5: PBB EVPN network

Similarly, on PE3, when an operator performs a connectivity check for the B-MAC address 00aa.00bb.00cc on PE2, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN MAC sub-TLV in the echo request packet. The echo request packet is sent with the {MPLS transport Label(s) to reach PE2 + EVPN Label = 16002 + GAL} MPLS label stack and IP ACH Channel header.

LSP Ping operation for unicast data-plane connectivity checks in E-VPN, are similar to those described above for PBB-EVPN except that the checks are for C-MAC addresses instead of B-MAC addresses.

## 6.2. Inclusive Multicast Data-plane Connectivity Checks

### 6.2.1. Ingress Replication

Assume PE1 announced an Inclusive Multicast route for EVI 10, with RD 1.1.1.1:00, Ethernet Tag (ISID 10), PMSI tunnel attribute Tunnel type set to ingress replication and downstream assigned inclusive multicast MPLS label 17001. Similarly, PE2 announced an Inclusive Multicast route for EVI 10, with RD 2.2.2.2:00, Ethernet Tag (ISID 10), PMSI tunnel attribute Tunnel type set to ingress replication and downstream assigned inclusive multicast MPLS label 17002.

Given CE1 is dual-homed to PE1 and PE2, assume that PE1 is the DF for ISID 10 for the port corresponding to the ESI 11aa.22bb.33cc.44dd.5500.

When an operator at PE3 initiates a connectivity check for the inclusive multicast on PE1, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN Inclusive Multicast sub-TLV in the Echo Request packet. The Echo Request packet is sent with the {Transport Label(s) to reach PE1 + EVPN Incl. Multicast Label = 17001 + GAL} MPLS label stack and IP ACH Channel header. Once the echo request packet reaches PE1, PE1 will use the GAL label and the IP ACH Channel header to determine that the packet is IPv4 OAM Packet. The packet will have EVPN Inclusive multicast label. PE1 will process the packet and perform checks for the EVPN Inclusive Multicast sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC8029] and respond according to [RFC8029] processing rules.

An operator at PE3, may similarly also initiate an LSP Ping to PE2 with the target FEC stack TLV containing EVPN Inclusive Multicast sub-TLV in the echo request packet. The echo request packet is sent with the {transport Label(s) to reach PE2 + EVPN Incl. Multicast Label = 17002 + GAL} MPLS label stack and IP ACH Channel header. Once the echo request packet reaches PE2, PE2 will use the GAL label and the IP ACH Channel header to determine that the packet is IPv4 OAM Packet. Since PE2 is not the DF for ISID 10 for the port corresponding to the ESI value in the Inclusive Multicast sub-TLV in the Echo Request, PE2 will reply with the special code indicating that FEC exists on the router and the behavior is to drop the packet because of not DF as described in Section 8.

In case of EVPN, in the Echo Request packet, an Ethernet AD sub-TLV and the associated MPLS Split Horizon Label above the GAL label in the MPLS label stack, may be added to emulate traffic coming from a MH site, this label is used by leaf PE(s) attached to the same MH site not to forward packets back to the MH site. If the behavior on a leaf PE is to drop the packet because of Split Horizon filtering, the PE2 will reply with the special code indicating that FEC exists

on the router and the behavior is to drop the packet because of Split Horizon Filtering as described in Section 8.

#### 6.2.2. Using P2MP P-tree

Both inclusive P-Tree and aggregate inclusive P-tree can be used in EVPN or PBB-EVPN networks.

When using an inclusive P-tree arrangement, p2mp p-tree transport label itself is used to identify the L2 service associated with the Inclusive Multicast Route, this L2 service could be a customer Bridge, or a Provider Backbone Bridge.

For an Inclusive P-tree arrangement, when an operator performs a connectivity check for the multicast L2 service, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN Inclusive Multicast sub-TLV in the echo request packet. The echo request packet is sent over P2MP LSP with the {P2MP P-tree label, GAL} MPLS label stack and IP ACH Channel header.

When using Aggregate Inclusive P-tree, a PE announces an upstream assigned MPLS label along with the P-tree ID, in that case both the p2mp p-tree MPLS transport label and the upstream MPLS label can be used to identify the L2 service.

For an Aggregate Inclusive P-tree arrangement, when an operator performs a connectivity check for the multicast L2 service, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN Inclusive Multicast sub-TLV in the echo request packet. The echo request packet is sent over P2MP LSP using the IP-ACH Control channel with the {P2MP P-tree label, EVPN Upstream assigned Multicast Label, GAL} MPLS label stack and IP ACH Channel header.

The Leaf PE(s) of the p2mp tree will process the packet and perform checks for the EVPN Inclusive Multicast sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC8029] and respond according to [RFC8029] processing rules. A PE that is not the DF for the EVI on the ESI in the Inclusive Multicast sub-TLV, will reply with a special code indicating that FEC exists on the router and the behavior is to drop the packet because of not DF as described in Section 8.

In case of EVPN, in the Echo Request packet, an Ethernet AD sub-TLV and the associated MPLS Split Horizon Label above the GAL Label in MPLS label stack, may be added to emulate traffic coming from a MH site, this label is used by leaf PE(s) attached to the same MH site not to forward packets back to the MH site. If the behavior on a

leaf PE is to drop the packet because of Split Horizon filtering, the PE2 will reply with special code indicating that FEC exists on the router and the behavior is to drop the packet because of Split Horizon Filtering as described in Section 8.

#### 6.2.3. Controlling Echo Responses when using P2MP P-tree

The procedures described in [RFC6425] for preventing congestion of Echo Responses (Echo Jitter TLV) and limiting the echo reply to a single egress node (Node Address P2MP Responder Identifier TLV) can be applied to LSP Ping in PBB EVPN and EVPN when using P2MP P-trees for broadcast, multicast, and unknown unicast traffic.

#### 6.3. EVPN Aliasing Data-plane connectivity check

Assume PE1 announced an Ethernet Auto discovery Route with the ESI set to CE1 system ID and MPLS label 19001, and PE2 an Ethernet Auto discovery Route with the ESI set to CE1 system ID and MPLS label 19002.

When an operator performs at PE3 a connectivity check for the aliasing aspect of the Ethernet AD route to PE1, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN Ethernet AD sub-TLV in the echo request packet. The echo request packet is sent with the {Transport label(s) to reach PE1 + EVPN Ethernet AD Label 19001 + GAL} MPLS label stack and IP ACH Channel header.

When PE1 receives the packet it will process the packet and perform checks for the EVPN Ethernet AD sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC8029] and respond according to [RFC8029] processing rules.

#### 6.4. EVPN IP Prefix (RT-5) Data-plane connectivity check

Assume PE1 in Figure 5, announced an IP Prefix Route (RT-5) with an IP prefix reachable behind CE1 and MPLS label 20001. When an operator on PE3 performs a connectivity check for the IP prefix on PE1, the operator initiates an LSP Ping request with the target FEC stack TLV containing EVPN IP Prefix sub-TLV in the echo request packet. The echo request packet is sent with the {Transport label(s) to reach PE1 + EVPN IP Prefix Label 20001 } MPLS label stack.

When PE1 receives the packet it will process the packet and perform checks for the EVPN IP Prefix sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC8029] and respond according to [RFC8029] processing rules.

## 7. Security Considerations

The proposal introduced in this document does not introduce any new security considerations beyond that already apply to [RFC7432], [RFC7623] and [RFC6425].

## 8. IANA Considerations

### 8.1. Sub-TLV Type

This document defines 4 new sub-TLV type to be included in Target FEC Stack TLV (TLV Type 1) [RFC8029] in LSP Ping.

IANA is requested to assign a sub-TLV type value to the following sub-TLV from the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry:

- o EVPN MAC route sub-TLV
- o EVPN Inclusive Multicast route sub-TLV
- o EVPN Auto-Discovery Route sub-TLV
- o EVPN IP Prefix Route sub-TLV

### 8.2. Proposed new Return Codes

[RFC8029] defines values for the Return Code field of Echo Reply. This document proposes two new Return Codes, which SHOULD be included in the Echo Reply message by a PE in response to LSP Ping Echo Request message:

1. The FEC exists on the PE and the behavior is to drop the packet because of not DF.
2. The FEC exists on the PE and the behavior is to drop the packet because of Split Horizon Filtering.

## 9. Acknowledgments

The authors would like to thank Patrice Brissette and Weiguo Hao for their comments.

## 10. References

### 10.1. Normative References

- [I-D.ietf-bess-evpn-prefix-advertisement]  
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.
- [RFC6425] Saxena, S., Ed., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, DOI 10.17487/RFC6425, November 2011, <<https://www.rfc-editor.org/info/rfc6425>>.
- [RFC6426] Gray, E., Bahadur, N., Boutros, S., and R. Aggarwal, "MPLS On-Demand Connectivity Verification and Route Tracing", RFC 6426, DOI 10.17487/RFC6426, November 2011, <<https://www.rfc-editor.org/info/rfc6426>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <<https://www.rfc-editor.org/info/rfc7623>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <<https://www.rfc-editor.org/info/rfc8029>>.

### 10.2. Informative References

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

- [RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <<https://www.rfc-editor.org/info/rfc4875>>.
- [RFC5085] Nadeau, T., Ed. and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, DOI 10.17487/RFC5085, December 2007, <<https://www.rfc-editor.org/info/rfc5085>>.
- [RFC6338] Giralt, V. and R. McDuff, "Definition of a Uniform Resource Name (URN) Namespace for the Schema for Academia (SCHAC)", RFC 6338, DOI 10.17487/RFC6338, August 2011, <<https://www.rfc-editor.org/info/rfc6338>>.

## Authors' Addresses

Parag Jain (editor)  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, ON K2K 3E8  
Canada

Email: [paragj@cisco.com](mailto:paragj@cisco.com)

Samer Salam  
Cisco Systems, Inc.  
595 Burrard Street, Suite 2123  
Vancouver, BC V7X 1J1  
Canada

Email: [ssalam@cisco.com](mailto:ssalam@cisco.com)

Ali Sajassi  
Cisco Systems, Inc.  
USA

Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Sami Boutros  
VmWare, Inc.  
USA

Email: sboutros@vmware.com

Greg Mirsky  
ZTE Corporation.  
USA

Email: gregmirsky@gmail.com>



INTERNET-DRAFT

Intended Status: Proposed Standard

Expires: Jul 19, 2019

N. Malhotra, Ed.  
(Arrcus)  
A. Sajassi  
A. Pattekar  
(Cisco)  
A. Lingala  
(AT&T)  
J. Rabadan  
(Nokia)  
J. Drake  
(Juniper Networks)

Jan 15, 2019

Extended Mobility Procedures for EVPN-IRB  
draft-malhotra-bess-evpn-irb-extended-mobility-04

Abstract

The procedure to handle host mobility in a layer 2 Network with EVPN control plane is defined as part of RFC 7432. EVPN has since evolved to find wider applicability across various IRB use cases that include distributing both MAC and IP reachability via a common EVPN control plane. MAC Mobility procedures defined in RFC 7432 are extensible to IRB use cases if a fixed 1:1 mapping between VM IP and MAC is assumed across VM moves. Generic mobility support for IP and MAC that allows these bindings to change across moves is required to support a broader set of EVPN IRB use cases, and requires further consideration. EVPN all-active multi-homing further introduces scenarios that require additional consideration from mobility perspective. Intent of this draft is to enumerate a set of design considerations applicable to mobility across EVPN IRB use cases and define generic sequence number assignment procedures to address these IRB use cases.

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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

#### Copyright and License Notice

Copyright (c) 2017 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	Introduction . . . . .	4
1.1	Terminology . . . . .	5
2.	Optional MAC only RT-2 . . . . .	5
3.	Mobility Use Cases . . . . .	6
3.1	VM MAC+IP Move . . . . .	6
3.2	VM IP Move to new MAC . . . . .	6
3.2.1	VM Reload . . . . .	6
3.2.2	MAC Sharing . . . . .	6
3.2.3	Problem . . . . .	7
3.3	VM MAC move to new IP . . . . .	8
3.3.1	Problem . . . . .	8
4.	EVPN All Active multi-homed ES . . . . .	10
5.	Design Considerations . . . . .	11
6.	Solution Components . . . . .	12
6.1	Sequence Number Inheritance . . . . .	12
6.2	MAC Sharing . . . . .	13
6.3	Multi-homing Mobility Synchronization . . . . .	14

7.	Requirements for Sequence Number Assignment . . . . .	14
7.1	LOCAL MAC-IP learning . . . . .	14
7.2	LOCAL MAC learning . . . . .	15
7.3	Remote MAC OR MAC-IP Update . . . . .	15
7.4	REMOTE (SYNC) MAC update . . . . .	15
7.5	REMOTE (SYNC) MAC-IP update . . . . .	16
7.6	Inter-op . . . . .	16
8.	Routed Overlay . . . . .	16
9.	Duplicate Host Detection . . . . .	18
9.1	Scenario A . . . . .	18
9.2	Scenario B . . . . .	18
9.2.1	Duplicate IP Detection Procedure for Scenario B . . . . .	19
9.3	Scenario C . . . . .	19
9.4	Duplicate Host Recovery . . . . .	20
9.4.1	Route Un-freezing Configuration . . . . .	20
9.4.2	Route Clearing Configuration . . . . .	21
10.	Security Considerations . . . . .	21
11.	IANA Considerations . . . . .	21
12.	References . . . . .	21
12.1	Normative References . . . . .	21
12.2	Informative References . . . . .	22
13.	Acknowledgements . . . . .	22
	Authors' Addresses . . . . .	22
	Appendix A . . . . .	22

## 1 Introduction

EVPN-IRB enables capability to advertise both MAC and IP routes via a single MAC+IP RT-2 advertisement. MAC is imported into local bridge MAC table and enables L2 bridged traffic across the network overlay. IP is imported into the local ARP table in an asymmetric IRB design OR imported into the IP routing table in a symmetric IRB design, and enables routed traffic across the layer 2 network overlay. Please refer to [EVPN-INTER-SUBNET] more background on EVPN IRB forwarding modes.

To support EVPN mobility procedure, a single sequence number mobility attribute is advertised with the combined MAC+IP route. A single sequence number advertised with the combined MAC+IP route to resolve both MAC and IP reachability implicitly assumes a 1:1 fixed mapping between IP and MAC. While a fixed 1:1 mapping between IP and MAC is a common use case that could be addressed via existing MAC mobility procedure, additional IRB scenarios need to be considered, that don't necessarily adhere to this assumption. Following IRB mobility scenarios are considered:

- o VM move results in VM IP and MAC moving together
- o VM move results in VM IP moving to a new MAC association
- o VM move results in VM MAC moving to a new IP association

While existing MAC mobility procedure can be leveraged for MAC+IP move in the first scenario, subsequent scenarios result in a new MAC-IP association. As a result, a single sequence number assigned independently per-[MAC, IP] is not sufficient to determine most recent reachability for both MAC and IP, unless the sequence number assignment algorithm is designed to allow for changing MAC-IP bindings across moves.

Purpose of this draft is to define additional sequence number assignment and handling procedures to adequately address generic mobility support across EVPN-IRB overlay use cases that allow MAC-IP bindings to change across VM moves and can support mobility for both MAC and IP components carried in an EVPN RT-2 for these use cases.

In addition, for hosts on an ESI multi-homed to multiple GW devices, additional procedure is proposed to ensure synchronized sequence number assignments across the multi-homing devices.

Content presented in this draft is independent of data plane encapsulation used in the overlay being MPLS or NVO Tunnels. It is also largely independent of the EVPN IRB solution being based on

symmetric OR asymmetric IRB design as defined in [EVPN-INTER-SUBNET]. In addition to symmetric and asymmetric IRB, mobility solution for a routed overlay, where traffic to an end host in the overlay is always IP routed using EVPN RT-5 is also presented in section 8.

To summarize, this draft covers mobility mobility for the following independent of the overlay encapsulation being MPLS or an NVO Tunnel:

- o Symmetric EVPN IRB overlay
- o Asymmetric EVPN IRB overlay
- o Routed EVPN overlay

### 1.1 Terminology

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 [RFC2119].

- o ARP is widely referred to in this document. This is simply for ease of reading, and as such, these references are equally applicable to ND (neighbor discovery) as well.
- o GW: used widely in the document refers to an IRB GW that is doing routing and bridging between an access network and an EVPN enabled overlay network.
- o RT-2: EVPN route type 2 carrying both MAC and IP reachability
- o RT-5: EVPN route type 5 carrying IP prefix reachability
- o ES: EVPN Ethernet Segment
- o MAC-IP: IP association for a MAC, referred to in this document may be IPv4, IPv6 or both.

### 2. Optional MAC only RT-2

In an EVPN IRB scenario, where a single MAC+IP RT-2 advertisement carries both IP and MAC routes, a MAC only RT-2 advertisement is redundant for host MACs that are advertised via MAC+IP RT-2. As a result, a MAC only RT-2 is an optional route that may not be advertised from or received at an IRB GW. This is an important consideration for mobility scenarios discussed in subsequent sections.

MAC only RT-2 may still be advertised for non-IP host MACs that are

not advertised via MAC+IP RT-2.

### 3. Mobility Use Cases

This section describes the IRB mobility use cases considered in this document. Procedures to address them are covered later in section 6 and section 7.

- o VM move results in VM IP and MAC moving together
- o VM move results in VM IP moving to a new MAC association
- o VM move results in VM MAC moving to a new IP association

#### 3.1 VM MAC+IP Move

This is the baseline case, wherein a VM move results in both VM MAC and IP moving together with no change in MAC-IP binding across a move. Existing MAC mobility defined in RFC 7432 may be leveraged to apply to corresponding MAC+IP route to support this mobility scenario.

#### 3.2 VM IP Move to new MAC

This is the case, where a VM move results in VM IP moving to a new MAC binding.

##### 3.2.1 VM Reload

A VM reload or an orchestrated VM move that results in VM being re-spawned at a new location may result in VM getting a new MAC assignment, while maintaining existing IP address. This results in a VM IP move to a new MAC binding:

IP-a, MAC-a ---> IP-a, MAC-b

##### 3.2.2 MAC Sharing

This takes into account scenarios, where multiple hosts, each with a unique IP, may share a common MAC binding, and a host move results in a new MAC binding for the host IP.

As an example, host VMs running on a single physical server, each with a unique IP, may share the same physical server MAC. In yet another scenario, an L2 access network may be behind a firewall, such that all hosts IPs on the access network are learnt with a common firewall MAC. In all such "shared MAC" use cases, multiple local MAC-

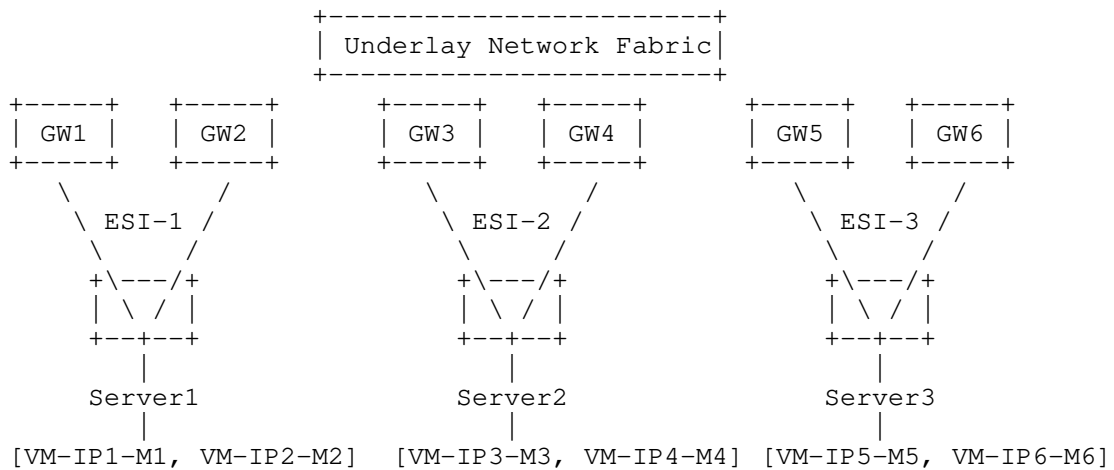


### 3.3 VM MAC move to new IP

This is a scenario where host move or re-provisioning behind a new gateway location may result in the same VM MAC getting a new IP address assigned.

#### 3.3.1 Problem

Complication with this scenario is that MAC reachability could be carried via a combined MAC+IP route while a MAC only route may not be advertised at all. A single sequence number association with the MAC+IP route again implicitly assumes a fixed mapping between MAC and IP. A MAC move resulting in a new IP association for the host MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route independently assumes a new sequence number, this mobility attribute can no longer be used to determine most recent host MAC reachability as opposed to the older existing MAC reachability.



As an example, IP1-M1 is learnt locally at [GW1, GW2] and currently advertised to remote hosts with a sequence number N. Consider a scenario where a VM with MAC M1 is re-provisioned at server 2, however, as part of this re-provisioning, assigned a different IP address say IP7. [IP7, M1] is learnt as a new route at [GW3, GW4] and advertised to remote GWs with a sequence number of 0. As a result, L3 reachability to IP7 would be established across the overlay, however, MAC mobility procedure for MAC1 will not trigger as a result of this MAC-IP route advertisement. If an optional MAC only route is also advertised, sequence number associated with the MAC only route would



trigger MAC mobility as per [RFC7432]. However, in the absence of an additional MAC only route advertisement, a single sequence number advertised with a combined MAC+IP route would not be sufficient to update MAC reachability across the overlay.

A MAC-IP sequence number assignment procedure needs to be defined to unambiguously determine the most recent MAC reachability in such a scenario without a MAC only route being advertised.

Further, GW1/GW2, on learning new reachability for [IP7, M1] via GW3/GW4 MUST probe and delete any local IPs associated with MAC M1, such as [IP1, M1] in the above example.

Arguably, MAC mobility sequence number defined in [RFC7432], could be interpreted to apply only to the MAC part of MAC-IP route, and would hence cover this scenario. It could hence be interpreted as a clarification to [RFC7432] and one of the considerations for a common sequence number assignment procedure across all MAC-IP mobility scenarios detailed in this document.

## 4. EVPN All Active multi-homed ES

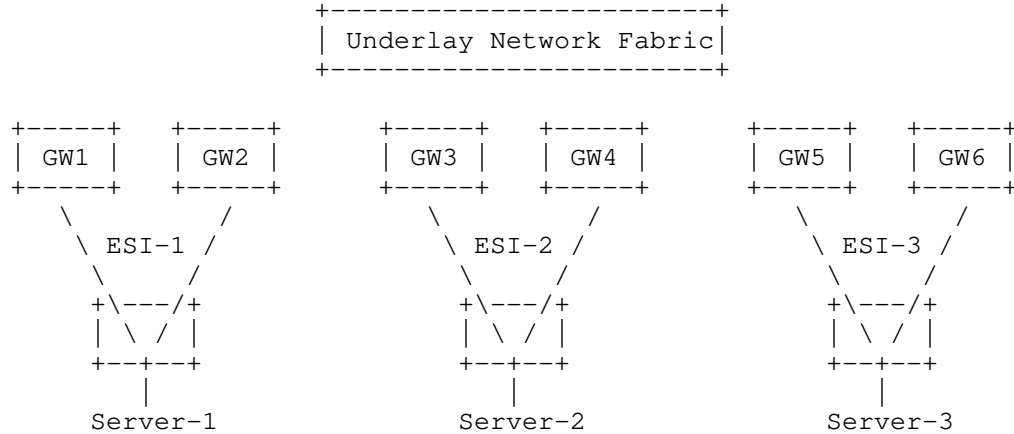


Figure 2

Consider an EVPN-IRB overlay network shown in Figure 2, with hosts multi-homed to two or more leaf GW devices via an all-active multi-homed ES. MAC and ARP entries learnt on a local ESI may also be synchronized across the multi-homing GW devices sharing this ESI. This MAC and ARP SYNC enables local switching of intra and inter subnet ECMP traffic flows from remote hosts. In other words, local MAC and ARP entries on a given Ethernet segment (ES) may be learnt via local learning and / or sync from another GW device sharing the same ES.

For a host that is multi-homed to multiple GW devices via an all-active ES interface, local learning of host MAC and MAC-IP at each GW device is an independent asynchronous event, that is dependent on traffic flow and or ARP / ND response from the host hashing to a directly connected GW on the MC-LAG interface. As a result, sequence number mobility attribute value assigned to a locally learnt MAC or MAC-IP route (as per RFC 7432) at each device may not always be the same, depending on transient states on the device at the time of local learning.

As an example, consider a host VM that is deleted from ESI-2 and moved to ESI-1. It is possible for host to be learnt on say, GW1 following deletion of the remote route from [GW3, GW4], while being learnt on GW2 prior to deletion of remote route from [GW3, GW4]. If so, GW1 would process local host route learning as a new route and assign a sequence number of 0, while GW2 would process local host

route learning as a remote to local move and assign a sequence number of  $N+1$ ,  $N$  being the existing sequence number assigned at [GW3, GW4]. Inconsistent sequence numbers advertised from multi-homing devices introduces ambiguity with respect to sequence number based mobility procedures across the overlay.

- o Ambiguity with respect to how the remote ToRs should handle paths with same ESI and different sequence numbers. A remote ToR may not program ECMP paths if it receives routes with different sequence numbers from a set of multi-homing GWs sharing the same ESI.
- o Breaks consistent route versioning across the network overlay that is needed for EVPN mobility procedures to work.

As an example, in this inconsistent state, GW2 would drop a remote route received for the same host with sequence number  $N$  (as its local sequence number is  $N+1$ ), while GW1 would install it as the best route (as its local sequence number is 0).

There is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

In order to support mobility for multi-homed hosts using the sequence number mobility attribute, local MAC and MAC-IP routes MUST be advertised with the same sequence number by all GW devices that the ESI is multi-homed to. In other words, there is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

## 5. Design Considerations

To summarize, sequence number assignment scheme and implementation must take following considerations into account:

- o MAC+IP may be learnt on an ESI multi-homed to multiple GW devices, hence requires sequence numbers to be synchronized across multi-homing GW devices.
- o MAC only RT-2 is optional in an IRB scenario and may not necessarily be advertised in addition to MAC+IP RT-2
- o Single MAC may be associated with multiple IPs, i.e., multiple host IPs may share a common MAC
- o Host IP move could result in host moving to a new MAC, resulting in a new IP to MAC association and a new MAC+IP route.

- o Host MAC move to a new location could result in host MAC being associated with a different IP address, resulting in a new MAC to IP association and a new MAC+IP route
- o LOCAL MAC-IP learn via ARP would always accompanied by a LOCAL MAC learn event resulting from the ARP packet. MAC and MAC-IP learning, however, could happen in any order
- o Use cases discussed earlier that do not maintain a constant 1:1 MAC-IP mapping across moves could potentially be addressed by using separate sequence numbers associated with MAC and IP components of MAC+IP route. Maintaining two separate sequence numbers however adds significant overhead with respect to complexity, debugability, and backward compatibility. It is therefore goal of solution presented here to address these requirements via a single sequence number attribute.

## 6. Solution Components

This section goes over main components of the EVPN IRB mobility solution proposed in this draft. Later sections will go over exact sequence number assignment procedures resulting from concepts described in this section.

### 6.1 Sequence Number Inheritance

Main idea presented here is to view a LOCAL MAC-IP route as a child of the corresponding LOCAL MAC only route that inherits the sequence number attribute from the parent LOCAL MAC only route:

Mx-IPx -----> Mx (seq# = N)

As a result, both parent MAC and child MAC-IP routes share one common sequence number associated with the parent MAC route. Doing so ensures that a single sequence number attribute carried in a combined MAC+IP route represents sequence number for both a MAC only route as well as a MAC+IP route, and hence makes the MAC only route truly optional. As a result, optional MAC only route with its own sequence number is not required to establish most recent reachability for a MAC in the overlay network. Specifically, this enables a MAC to assume a different IP address on a move, and still be able to establish most recent reachability to the MAC across the overlay network via mobility attribute associated with the MAC+IP route advertisement. As an example, when Mx moves to a new location, it would result in LOCAL Mx being assigned a higher sequence number at its new location as per RFC 7432. If this move results in Mx assuming a different IP address, IPz, LOCAL Mx+IPz route would inherit the new

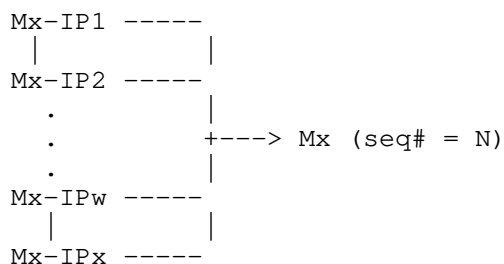
sequence number from Mx.

LOCAL MAC and LOCAL MAC-IP routes would typically be sourced from data plane learning and ARP learning respectively, and could get learnt in control plane in any order. Implementation could either replicate inherited sequence number in each MAC-IP entry OR maintain a single attribute in the parent MAC by creating a forward reference LOCAL MAC object for cases where a LOCAL MAC-IP is learnt before the LOCAL MAC.

Arguably, this inheritance may be assumed from RFC 7432, in which case, the above may be interpreted as a clarification with respect to interpretation of a MAC sequence number in a MAC-IP route.

## 6.2 MAC Sharing

Further, for the shared MAC scenario, this would result in multiple LOCAL MAC-IP siblings inheriting sequence number attribute from a common parent MAC route:



In such a case, a host-IP move to a different physical server would result in IP moving to a new MAC binding. A new MAC-IP route resulting from this move must now be advertised with a sequence number that is higher than the previous MAC-IP route for this IP, advertised from the prior location. As an example, consider a route Mx-IPx that is currently advertised with sequence number N from GW1. IPx moving to a new physical server behind GW2 results in IPx being associated with MAC Mz. A new local Mz-IPx route resulting from this move at GW2 must now be advertised with a sequence number higher than N. This is so that GW devices, including GW1, GW2, and other remote GW devices that are part of the overlay can clearly determine and program the most recent MAC binding and reachability for the IP. GW1, on receiving this new Mz-IPx route with sequence number say, N+1, for symmetric IRB case, would update IPx reachability via GW2 in forwarding, for asymmetric IRB case, would update IPx's ARP binding to Mz. In addition, GW1 would clear and withdraw the stale Mx-IPx route with the lower sequence number.

This also implies that sequence number associated with local MAC Mz and all local MAC-IP children of Mz at GW2 must now be incremented to N+1, and re-advertised across the overlay. While this re-advertisement of all local MAC-IP children routes affected by the parent MAC route is an overhead, it avoids the need for two separate sequence number attributes to be maintained and advertised for IP and MAC components of MAC+IP RT-2. Implementation would need to be able to lookup MAC-IP routes for a given IP and update sequence number for it's parent MAC and its MAC-IP children.

### 6.3 Multi-homing Mobility Synchronization

In order to support mobility for multi-homed hosts, local MAC and MAC-IP routes learnt on the shared ESI MUST be advertised with the same sequence number by all GW devices that the ESI is multi-homed to. This also applies to local MAC only routes. LOCAL MAC and MAC-IP may be learnt natively via data plane and ARP/ND respectively as well as via SYNC from another multi-homing GW to achieve local switching. Local and SYNC route learning can happen in any order. Local MAC-IP routes advertised by all multi-homing GW devices sharing the ESI must carry the same sequence number, independent of the order in which they are learnt. This implies:

- o On local or sync MAC-IP route learning, sequence number for the local MAC-IP route MUST be compared and updated to the higher value.
- o On local or sync MAC route learning, sequence number for the local MAC route MUST be compared and updated to the higher value.

If an update to local MAC-IP sequence number is required as a result of above comparison with sync MAC-IP route, it would essentially amount to a sequence number update on the parent local MAC, resulting in the inherited sequence number update on the MAC-IP route.

## 7. Requirements for Sequence Number Assignment

Following sections summarize sequence number assignment procedure needed on local and sync MAC and MAC-IP route learning events in order to accomplish the above.

### 7.1 LOCAL MAC-IP learning

A local Mx-IPx learning via ARP or ND should result in computation OR re-computation of parent MAC Mx's sequence number, following which the MAC-IP route Mx-IPx would simply inherit parent MAC's sequence number. Parent MAC Mx Sequence number should be computed as follows:

- o MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.
- o MUST be at least equal to corresponding SYNC MAC sequence number if one is present.
- o If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number

Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

## 7.2 LOCAL MAC learning

Local MAC Mx Sequence number should be computed as follows:

- o MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.
- o MUST be at least equal to corresponding SYNC MAC sequence number if one is present.
- o Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

Note that the local MAC sequence number might already be present if there was a local MAC-IP learnt prior to the local MAC, in which case the above may not result in any change in local MAC's sequence number.

## 7.3 Remote MAC OR MAC-IP Update

On receiving a remote MAC OR MAC-IP route update associated with a MAC Mx with a sequence number that is higher than a LOCAL route for MAC Mx:

- o GW MUST trigger probe and deletion procedure for all LOCAL IPs associated with MAC Mx
- o GW MUST trigger deletion procedure for LOCAL MAC route for Mx

## 7.4 REMOTE (SYNC) MAC update

Corresponding local MAC Mx (if present) Sequence number should be re-computed as follows:

- o If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.
- o If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

#### 7.5 REMOTE (SYNC) MAC-IP update

If this is a SYNCed MAC-IP on a local ESI, it would also result in a derived SYNC MAC Mx route entry, as MAC only RT-2 advertisement is optional. Corresponding local MAC Mx (if present) Sequence number should be re-computed as follows:

- o If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.
- o If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

#### 7.6 Inter-op

In general, if all GW nodes in the overlay network follow the above sequence number assignment procedure, and the GW is advertising both MAC+IP and MAC routes, sequence number advertised with the MAC and MAC+IP routes with the same MAC would always be the same. However, an inter-op scenario with a different implementation could arise, where a GW implementation non-compliant with this document or with RFC 7432 assigns and advertises independent sequence numbers to MAC and MAC+IP routes. To handle this case, if different sequence numbers are received for remote MAC+IP and corresponding remote MAC routes from a remote GW, sequence number associated with the remote MAC route should be computed as:

- o Highest of the all received sequence numbers with remote MAC+IP and MAC routes with the same MAC.
- o MAC sequence number would be re-computed on a MAC or MAC+IP route withdraw as per above.

A MAC and / or IP move to the local GW would now result in the MAC (and hence all MAC-IP) sequence numbers incremented from the above computed remote MAC sequence number.

#### 8. Routed Overlay



An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- o A host MAC is never advertised in EVPN overlay control plane
- o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI
- o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged
- o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Host mobility within the stretched subnet would still need to be supported for this use. In the absence of any host MAC routes, sequence number mobility EXT-COMM specified in [RFC7432], section 7.7 may be associated with a /32 OR /128 host IP prefix advertised via EVPN route type 5. MAC mobility procedures defined in RFC 7432 can now be applied as is to host IP prefixes:

- o On LOCAL learning of a host IP, on a new ESI, host IP MUST be advertised with a sequence number attribute that is higher than what is currently advertised with the old ESI
- o on receiving a host IP route advertisement with a higher sequence number, a PE MUST trigger ARP/ND probe and deletion procedure on any LOCAL route for that IP with a lower sequence number. A PE would essentially move the forwarding entry to point to the remote route with a higher sequence number and send an ARP/ND PROBE for the local IP route. If the IP has indeed moved, PROBE would timeout and the local IP host route would be deleted.

Note that there is still only one sequence number associated with a host route at any time. For earlier use cases where a host MAC is advertised along with the host IP, a sequence number is only associated with a MAC. Only if the MAC is not advertised at all, as in this use case, is a sequence number associated with a host IP.

Note that this mobility procedure would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

## 9. Duplicate Host Detection

Duplicate host detection scenarios across EVPN IRB can be classified as follows:

- o Scenario A: where two hosts have the same MAC (host IPs may or may not be duplicate)
- o Scenario B: where two hosts have the same IP but different MACs
- o Scenario C: where two hosts have the same IP and host MAC is not advertised at all

Duplicate detection procedures for scenario B and C would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

### 9.1 Scenario A

For all use cases where duplicate hosts have the same MAC, MAC is detected as duplicate via duplicate MAC detection procedure described in RFC 7432. Corresponding MAC-IP routes with the same MAC do not require duplicate detection and MUST simply inherit the DUPLICATE property from the corresponding MAC route. In other words, if a MAC route is in DUPLICATE state, all corresponding MAC-IP routes MUST also be treated as DUPLICATE. Duplicate detection procedure need only be applied to MAC routes.

### 9.2 Scenario B

Due to misconfiguration, a situation may arise where hosts with different MACs are configured with the same IP. This scenario would not be detected by existing duplicate MAC detection procedure and would result in incorrect forwarding of routed traffic destined to this IP.

Such a situation, on LOCAL MAC-IP learning, would be detected as a move scenario via the following local MAC sequence number computation procedure described earlier in section 5.1:

- o If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number

Such a move that results in sequence number increment on local MAC because of a remote MAC-IP route associated with a different MAC MUST be counted as an "IP move" against the "IP" independent of MAC. Duplicate detection procedure described in RFC 7432 can now be applied to an "IP" entity independent of MAC. Once an IP is detected

as DUPLICATE, corresponding MAC-IP route should be treated as DUPLICATE. Associated MAC routes and any other MAC-IP routes associated with this MAC should not be affected.

#### 9.2.1 Duplicate IP Detection Procedure for Scenario B

Duplicate IP detection procedure for such a scenario is specified in [EVPN-PROXY-ARP]. What counts as an "IP move" in this scenario is further clarified as follows:

- o On learning a LOCAL MAC-IP route Mx-IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different MAC association, say, Mz-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC
- o On learning a REMOTE MAC-IP route Mz-IPx, check if there is an existing LOCAL route for IPx with a different MAC association, say, Mx-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC

A MAC-IP route SHOULD be treated as DUPLICATE if either of the following two conditions are met:

- o Corresponding MAC route is marked as DUPLICATE via existing duplicate detection procedure
- o Corresponding IP is marked as DUPLICATE via extended procedure described above

#### 9.3 Scenario C

For a purely routed overlay scenario described in section 8, where only a host IP is advertised via EVPN RT-5, together with a sequence number mobility attribute, duplicate MAC detection procedures specified in RFC 7432 can be intuitively applied to IP only host routes for the purpose of duplicate IP detection.

- o On learning a LOCAL host IP route IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx.
- o On learning a REMOTE host IP route IPx, check if there is an existing LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx
- o With configurable parameters "N" and "M", If "N" IP moves are detected within "M" seconds for IPx, treat IPx as DUPLICATE

#### 9.4 Duplicate Host Recovery

Once a MAC or IP is marked as DUPLICATE and FROZEN, corrective action must be taken to un-provision one of the duplicate MAC or IP. Un-provisioning a duplicate MAC or IP in this context refers to a corrective action taken on the host side. Once one of the duplicate MAC or IP is un-provisioned, normal operation would not resume until the duplicate MAC or IP ages out, following this correction, unless additional action is taken to speed up recovery.

This section lists possible additional corrective actions that could be taken to achieve faster recovery to normal operation.

##### 9.4.1 Route Un-freezing Configuration

Unfreezing the DUPLICATE OR FROZEN MAC or IP via a CLI can be leveraged to recover from DUPLICATE and FROZEN state following corrective un-provisioning of the duplicate MAC or IP.

Unfreezing the frozen MAC or IP via a CLI at a GW should result in that MAC OR IP being advertised with a sequence number that is higher than the sequence number advertised from the other location of that MAC or IP.

Two possible corrective un-provisioning scenarios exist:

- o Scenario A: A duplicate MAC or IP may have been un-provisioned at the location where it was NOT marked as DUPLICATE and FROZEN
- o Scenario B: A duplicate MAC or IP may have been un-provisioned at the location where it was marked as DUPLICATE and FROZEN

Unfreezing the DUPLICATE and FROZEN MAC or IP, following the above corrective un-provisioning scenarios would result in recovery to steady state as follows:

- o Scenario A: If the duplicate MAC or IP was un-provisioned at the location where it was NOT marked as DUPLICATE, unfreezing the route at the FROZEN location will result in the route being advertised with a higher sequence number. This would in-turn result in automatic clearing of local route at the GW location, where the host was un-provisioned via ARP/ND PROBE and DELETE procedure specified earlier in section 8 and in [RFC 7432].
- o Scenario B: If the duplicate host is un-provisioned at the location where it was marked as DUPLICATE, unfreezing the route will trigger an advertisement with a higher sequence number to the other location. This would in-turn trigger re-learning of

local route at the remote location, resulting in another advertisement with a higher sequence number from the remote location. Route at the local location would now be cleared on receiving this remote route advertisement, following the ARP/ND PROBE.

#### 9.4.2 Route Clearing Configuration

In addition to the above, route clearing CLIs may also be leveraged to clear the local MAC or IP route, to be executed AFTER the duplicate host is un-provisioned:

- o clear mac CLI: A clear MAC CLI can be leveraged to clear a DUPLICATE MAC route, to recover from a duplicate MAC scenario
- o clear ARP/ND: A clear ARP/ND CLI may be leveraged to clear a DUPLICATE IP route to recover from a duplicate IP scenario

Note that the route unfreeze CLI may still need to be run if the route was un-provisioned and cleared from the NON-DUPLICATE / NON-FROZEN location. Given that unfreezing of the route via the un-freeze CLI would any ways result in auto-clearing of the route from the "un-provisioned" location, as explained in the prior section, need for a route clearing CLI for recovery from DUPLICATE / FROZEN state is truly optional.

### 10. Security Considerations

### 11. IANA Considerations

### 12. References

#### 12.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.

[EVPN-PROXY-ARP] Rabadan et al., "Operational Aspects of Proxy-ARP/ND in EVPN Networks", draft-ietf-bess-evpn-proxy-arp-nd-02, work in progress, April 2017, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-proxy-arp-nd-02>>.

[EVPN-INTER-SUBNET] Sajassi et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-03,

work in progress, Feb 2017,  
<<https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-03>>.

[RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., Fee, B.,  
"Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension  
Solution", RFC 7814, March 2016,  
<<https://tools.ietf.org/html/rfc7814>>.

## 12.2 Informative References

## 13. Acknowledgements

Authors would like to thank Vibov Bhan and Patrice Brisset for feedback and comments through the process.

### Authors' Addresses

Neeraj Malhotra (Editor)  
Arrcus  
EMail: [neeraj.ietf@gmail.com](mailto:neeraj.ietf@gmail.com)

Ali Sajassi  
Cisco  
EMail: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Aparna Pattekar  
Cisco  
Email: [apjoshi@cisco.com](mailto:apjoshi@cisco.com)

Jorge Rabadan  
Nokia  
Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

Avinash Lingala  
AT&T  
Email: [ar977m@att.com](mailto:ar977m@att.com)

John Drake  
Juniper Networks  
EMail: [jdrake@juniper.net](mailto:jdrake@juniper.net)

## Appendix A

An alternative approach considered was to associate two independent

sequence number attributes with MAC and IP components of a MAC-IP route. However, the approach of enabling IRB mobility procedures using a single sequence number associated with a MAC, as specified in this document was preferred for the following reasons:

- o Procedural overhead and complexity associated with maintaining two separate sequence numbers all the time, only to address scenarios with changing MAC-IP bindings is a big overhead for topologies where MAC-IP bindings never change.
- o Using a single sequence number associated with MAC is much simpler and adds no overhead for topologies where MAC-IP bindings never change.
- o Using a single sequence number associated with MAC is aligned with existing MAC mobility implementations. On other words, it is an easier implementation extension to existing MAC mobility procedure.

INTERNET-DRAFT

Intended Status: Proposed Standard

N. Malhotra, Ed.  
Arcus  
A. Sajassi  
Cisco  
J. Rabadan  
Nokia  
J. Drake  
Juniper  
A. Lingala  
AT&T  
S. Thoria  
Cisco

Expires: Jan 17, 2019

July 16, 2018

Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing  
draft-malhotra-bess-evpn-unequal-lb-04

Abstract

In an EVPN-IRB based network overlay, EVPN LAG enables all-active multi-homing for a host or CE device connected to two or more PEs via a LAG bundle, such that bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

- o provide greater flexibility, with respect to adding or removing individual PE-CE links within the access LAG
- o handle PE-CE LAG member link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs

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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2017 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	Introduction . . . . .	4
1.1	PE CE Link Provisioning . . . . .	5
1.2	PE CE Link Failures . . . . .	6
1.3	Design Requirement . . . . .	7
1.4	Terminology . . . . .	7
2.	Solution Overview . . . . .	8
3.	Weighted Unicast Traffic Load-balancing . . . . .	8
3.1	LOCAL PE Behavior . . . . .	8
3.1	Link Bandwidth Extended Community . . . . .	8
3.2	REMOTE PE Behavior . . . . .	9
4.	Weighted BUM Traffic Load-Sharing . . . . .	10
4.1	The BW Capability in the DF Election Extended Community . .	10
4.2	BW Capability and Default DF Election algorithm . . . . .	11
4.3	BW Capability and HRW DF Election algorithm (Type 1 and 4) . . . . .	11
4.3.1	BW Increment . . . . .	11
4.3.2	HRW Hash Computations with BW Increment . . . . .	12
4.3.3	Cost-Benefit Tradeoff on Link Failures . . . . .	13

4.4 BW Capability and Preference DF Election algorithm . . . .	14
5. Real-time Available Bandwidth . . . . .	15
6. Routed EVPN Overlay . . . . .	15
7. EVPN-IRB Multi-homing with non-EVPN routing . . . . .	16
7. References . . . . .	17
7.1 Normative References . . . . .	17
7.2 Informative References . . . . .	17
8. Acknowledgements . . . . .	18
Authors' Addresses . . . . .	18

## 1 Introduction

In an EVPN-IRB based network overlay, with an access CE multi-homed via a LAG interface, bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs:

- o ECMP Load-balancing for bridged unicast traffic is enabled via aliasing and mass-withdraw procedures detailed in RFC 7432.
- o ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.
- o Load-sharing of bridged BUM traffic on local ports is enabled via EVPN DF election procedure detailed in RFC 7432

All of the above load-balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of multi-homing PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all PEs, ALL remote traffic is equally load balanced across the multi-homing PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of multi-homing EVPN PEs is proposed in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and to optimally handle PE-CE link failures.

## 1.1 PE CE Link Provisioning

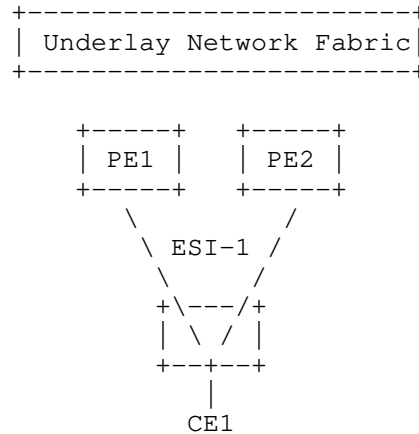


Figure 1

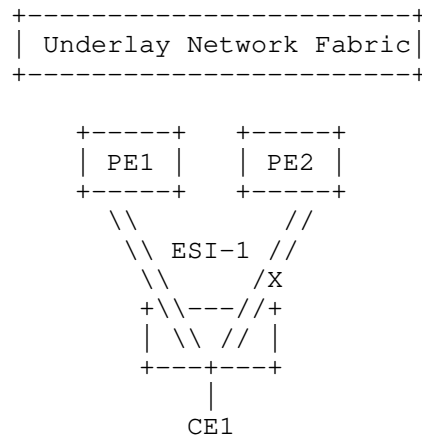
Consider a CE1 that is dual-homed to PE1 and PE2 via EVPN-LAG with single member links of equal bandwidth to each PE (aka, equal access bandwidth distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it MUST add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load-balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth OR number of links in ANY link increments. As an example, for CE1 dual-homed to PE1 and PE2 in all-active mode, provider may wish to add a third link to ONLY PE1 to increase total bandwidth for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load-balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to attract equal amount of CE1 destined traffic from remote PEs, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1

link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner.

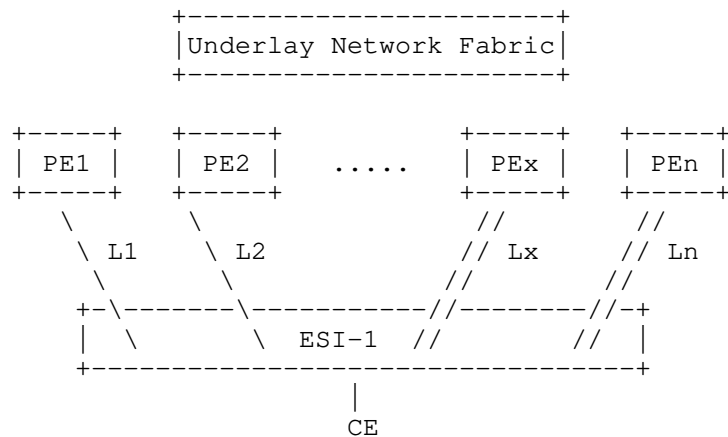
## 1.2 PE CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.



Consider a CE1 that is multi-homed to PE1 and PE2 via a link bundle with two member links to each PE. On a PE2-CE1 physical link failure, link bundle represented by ESI-1 on PE2 stays up, however, it's bandwidth is cut in half. With the existing ECMP procedures, both PE1 and PE2 will continue to attract equal amount of traffic from remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG.

## 1.3 Design Requirement



To generalize, if total link bandwidth to a CE is distributed across "n" multi-homing PEs, with Lx being the number of links / bandwidth to PEx, traffic from remote PEs to this CE MUST be load-balanced unequally across [PE1, PE2, ....., PEn] such that, fraction of total unicast and BUM flows destined for CE that are serviced by PEx is:

$$Lx / [L1+L2+.....+Ln]$$

Solution proposed below includes extensions to EVPN procedures to achieve the above.

## 1.4 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

"LOCAL PE" in the context of an ESI refers to a provider edge switch OR router that physically hosts the ESI.

"REMOTE PE" in the context of an ESI refers to a provider edge switch OR router in an EVPN overlay, who's overlay reachability to the ESI is via the LOCAL PE.

## 2. Solution Overview

In order to achieve weighted load balancing for overlay unicast traffic, Ethernet A-D per-ES route (EVPN Route Type 1) is leveraged to signal the ESI bandwidth to remote PEs. Using Ethernet A-D per-ES route to signal the ESI bandwidth provides a mechanism to be able to react to changes in access bandwidth in a service and host independent manner. Remote PEs computing the MAC path-lists based on global and aliasing Ethernet A-D routes now have the ability to setup weighted load-balancing path-lists based on the ESI access bandwidth received from each PE that the ESI is multi-homed to. If Ethernet A-D per-ES route is also leveraged for IP path-list computation, as per [EVPN-IP-ALIASING], it also provides a method to do weighted load-balancing for IP routed traffic.

In order to achieve weighted load-balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ESI bandwidth to PEs within an ESI's redundancy group to influence per-service DF election. PEs in an ESI redundancy group now have the ability to do service carving in proportion to each PE's relative ESI bandwidth.

Procedures to accomplish this are described in greater detail next.

## 3. Weighted Unicast Traffic Load-balancing

### 3.1 LOCAL PE Behavior

A PE that is part of an ESI's redundancy group would advertise a additional "link bandwidth" EXT-COMM attribute with Ethernet A-D per-ES route (EVPN Route Type 1), that represents total bandwidth of PE's physical links in an ESI. BGP link bandwidth EXT-COMM defined in [BGP-LINK-BW] is re-used for this purpose.

### 3.1 Link Bandwidth Extended Community

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs is re-used here to signal local ES link bandwidth to remote PEs. link-bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. In inter-AS scenarios, link-bandwidth may need to be signaled to an eBGP neighbor along with next-hop unchanged. It is work in progress with authors of [BGP-LINK-BW] to allow for this attribute to be used as transitive in inter-AS scenarios.

### 3.2 REMOTE PE Behavior

A receiving PE should use per-ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE, per-ES, as shown below.

if,

$L(x,y)$  : link bandwidth advertised by PE-x for ESI-y

$W(x,y)$  : normalized weight assigned to PE-x for ESI-y

$H(y)$  : Highest Common Factor (HCF) of  $[L(1,y), L(2,y), \dots, L(n,y)]$

then, the normalized weight assigned to PE-x for ESI-y may be computed as follows:

$$W(x,y) = L(x,y) / H(y)$$

For a MAC+IP route (EVPN Route Type 2) received with ESI-y, receiving PE MUST compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE dual-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a link bundle represented by ESI-10:

$$L(1, 10) = 2000 \text{ Mbps}$$

$$L(2, 10) = 1000 \text{ Mbps}$$

$$L(3, 10) = 1000 \text{ Mbps}$$

$$H(10) = 1000$$

Normalized weights assigned to each PE for ESI-10 are as follows:

$$W(1, 10) = 2000 / 1000 = 2.$$

$$W(2, 10) = 1000 / 1000 = 1.$$

$$W(3, 10) = 1000 / 1000 = 1.$$

For a remote MAC+IP host route received with ESI-10, forwarding load-balancing path-list must now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load-balancing of all traffic destined for ESI-10 across the three multi-homing PEs in



proportion to ESI-10 bandwidth at each PE.

Above weighted path-list computation MUST only be done for an ESI, IF a link bandwidth attribute is received from ALL of the PE's advertising reachability to that ESI via Ethernet A-D per-ES Route Type 1. In the event that link bandwidth attribute is not received from one or more PEs, forwarding path-list would be computed using regular ECMP semantics.

#### 4. Weighted BUM Traffic Load-Sharing

Optionally, load sharing of per-service DF role, weighted by individual PE's link-bandwidth share within a multi-homed ES may also be achieved.

In order to do that, a new DF Election Capability [EVPN-DF-ELECT-FRAMEWORK] called "BW" (Bandwidth Weighted DF Election) is defined. BW may be used along with some DF Election Types, as described in the following sections.

##### 4.1 The BW Capability in the DF Election Extended Community

[EVPN-DF-ELECT-FRAMEWORK] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests a bit in the DF Election extended community Bitmap:

Bit 28: BW (Bandwidth Weighted DF Election)

ES routes advertised with the BW bit set will indicate the desire of the advertising PE to consider the link-bandwidth in the DF Election algorithm defined by the value in the "DF Type".

As per [EVPN-DF-ELECT-FRAMEWORK], all the PEs in the ES MUST advertise the same Capabilities and DF Type, otherwise the PEs will fall back to Default [RFC7432] DF Election procedure.

The BW Capability MAY be advertised with the following DF Types:

- o Type 0: Default DF Election algorithm, as in [RFC7432]
- o Type 1: HRW algorithm, as in [EVPN-DF-ELECT-FRAMEWORK]
- o Type 2: Preference algorithm, as in [EVPN-DF-PREF]
- o Type 4: HRW per-multicast flow DF Election, as in [XXX]

The following sections describe how the DF Election procedures are modified for the above DF Types when the BW Capability is used.

#### 4.2 BW Capability and Default DF Election algorithm

When all the PEs in the ES agree to use the BW Capability with DF Type 0, the Default DF Election procedure is modified as follows:

- o Each PE advertises a "Link Bandwidth" EXT-COMM attribute along with the ES route to signal the PE-CE link bandwidth (LBW) for the ES.
- o A receiving PE MUST use the ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE.
- o The DF Election procedure MUST now use this weighted list of PEs to compute the per-VLAN Designated Forwarder, such that the DF role is distributed in proportion to this normalized weight.

Considering the same example as in Section 3, the candidate PE list for DF election is:

[PE-1, PE-1, PE-2, PE-3].

The DF for a given VLAN-a on ES-10 is now computed as  $(\text{VLAN-a} \% 4)$ . This would result in the DF role being distributed across PE1, PE2, and PE3 in portion to each PE's normalized weight for ES-10.

#### 4.3 BW Capability and HRW DF Election algorithm (Type 1 and 4)

[EVPN-DF-ELECT-FRAMEWORK] introduces Highest Random Weight (HRW) algorithm (DF Type 1) for DF election in order to solve potential DF election skew depending on Ethernet tag space distribution. [EVPN-PER-MCAST-FLOW-DF] further extends HRW algorithm for per-multicast flow based hash computations (DF Type 4). This section describes extensions to HRW Algorithm for EVPN DF Election specified in [EVPN-DF-ELECT-FRAMEWORK] and in [EVPN-PER-MCAST-FLOW-DF] in order to achieve DF election distribution that is weighted by link bandwidth.

##### 4.3.1 BW Increment

A new variable called "bandwidth increment" is computed for each [PE, ES] advertising the ES link bandwidth attribute as follows:

In the context of an ES,

$L(i)$  = Link bandwidth advertised by PE(i) for this ES

$L(\min)$  = lowest link bandwidth advertised across all PEs for this ES

Bandwidth increment, " $b(i)$ " for a given PE(i) advertising a link bandwidth of  $L(i)$  is defined as an integer value computed as:

$$b(i) = L(i) / L(\min)$$

As an example,

with  $PE(1) = 10$ ,  $PE(2) = 10$ ,  $PE(3) = 20$

bandwidth increment for each PE would be computed as:

$$b(1) = 1, b(2) = 1, b(3) = 2$$

with  $PE(1) = 10$ ,  $PE(2) = 10$ ,  $PE(3) = 10$

bandwidth increment for each PE would be computed as:

$$b(1) = 1, b(2) = 1, b(3) = 1$$

Note that the bandwidth increment must always be an integer, including, in an unlikely scenario of a PE's link bandwidth not being an exact multiple of  $L(\min)$ . If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

#### 4.3.2 HRW Hash Computations with BW Increment

HRW algorithm as described in [EVPN-DF-ELECT-FRAMEWORK] and in [EVPN-PER-MCAST-FLOW-DF] compute a random hash value (referred to as affinity here) for each  $PE(i)$ , where,  $(0 < i \leq N)$ ,  $PE(i)$  is the PE at ordinal  $i$ , and  $Address(i)$  is the IP address of PE at ordinal  $i$ .

For ' $N$ ' PEs sharing an Ethernet segment, this results in ' $N$ ' candidate hash computations. PE that has the highest hash value is selected as the DF.

Affinity computation for each  $PE(i)$  is extended to be computed one per-bandwidth increment associated with  $PE(i)$  instead of a single affinity computation per  $PE(i)$ .

$PE(i)$  with  $b(i) = j$ , results in  $j$  affinity computations:

affinity( $i, x$ ), where  $1 < x \leq j$

This essentially results in number of candidate HRW hash computations for each PE that is directly proportional to that PE's relative bandwidth within an ES and hence gives  $PE(i)$  a probability of being DF in proportion to it's relative bandwidth within an ES.

As an example, consider an ES that is multi-homed to two PEs,  $PE1$  and  $PE2$ , with equal bandwidth distribution across  $PE1$  and  $PE2$ . This would

result in a total of two candidate hash computations:

affinity(PE1, 1)

affinity(PE2, 1)

Now, consider a scenario with PE1's link bandwidth as 2x that of PE2. This would result in a total of three candidate hash computations to be used for DF election:

affinity(PE1, 1)

affinity(PE1, 2)

affinity(PE2, 1)

which would give PE1 2/3 probability of getting elected as a DF, in proportion to its relative bandwidth in the ES.

Depending on the chosen HRW hash function, affinity function MUST be extended to include bandwidth increment in the computation.

For e.g.,

affinity function specified in [EVPN-PER-MCAST-FLOW-DF] MAY be extended as follows to incorporate bandwidth increment j:

$$\text{affinity}(S,G,V, \text{ESI}, \text{Address}(i,j)) = \\ (1103515245 \cdot ((1103515245 \cdot \text{Address}(i).j + 12345) \text{ XOR} \\ D(S,G,V,\text{ESI})) + 12345) \pmod{2^{31}}$$

affinity or random function specified in [EVPN-DF-ELECT-FRAMEWORK] MAY be extended as follows to incorporate bandwidth increment j:

$$\text{affinity}(v, \text{Es}, \text{Address}(i,j)) = (1103515245 \cdot ((1103515245 \cdot \text{Address}(i).j \\ + 12345) \text{ XOR} D(v,\text{Es})) + 12345) \pmod{2^{31}}$$

#### 4.3.3 Cost-Benefit Tradeoff on Link Failures

While incorporating link bandwidth into the DF election process provides optimal BUM traffic distribution across the ES links, it also implies that affinity values for a given PE are re-computed, and DF elections are re-adjusted on changes to that PE's bandwidth increment that might result from link failures or link additions. If the operator does not wish to have this level of churn in their DF election, then they should not advertise the BW capability. Not advertising BW capability may result in less than optimal BUM traffic

distribution while still retaining the ability to allow a remote ingress PE to do weighted ECMP for its unicast traffic to a set of multi-homed PEs, as described in section 3.2.

Same also applies to use of BW capability with service carving (DF Type 0), as specified in section 4.2.

#### 4.4 BW Capability and Preference DF Election algorithm

This section applies to ES'es where all the PEs in the ES agree use the BW Capability with DF Type 2. The BW Capability modifies the Preference DF Election procedure [EVPN-DF-PREF], by adding the LBW value as a tie-breaker as follows:

- o Section 4.1, bullet (f) in [EVPN-DF-PREF] now considers the LBW value:
  - f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP bit, the LBW value and the lowest IP PE in that order. For instance:
    - o If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=1, LBW=2000] in PE2, PE2 would be elected due to the DP bit.
    - o If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=0, LBW=2000] in PE2, PE2 would be elected due to a higher LBW, even if PE1's IP address is lower.
    - o The LBW exchanged value has no impact on the Non-Revertive option described in [EVPN-DF-PREF].

## 5. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE-CE links within a multi-homed ESI due to various factors such as flow based hashing combined with fat flows and unbalanced hashing. Reacting to real-time available bandwidth is at this time outside the scope of this document. Procedures described in this document are strictly based on static link bandwidth parameter.

## 6. Routed EVPN Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- o A host MAC is never advertised in EVPN overlay control plane
- o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI
- o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged
- o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Weighted multi-path procedure described in this document may be used together with procedures described in [EVPN-IP-ALIASING] for this use case. Ethernet A-D per-ES route advertised with Layer 3 VRF RTs would be used to signal ES link bandwidth attribute instead of the Ethernet A-D per-ES route with Layer 2 VRF RTs. All other procedures described earlier in this document would apply as is.

If [EVPN-IP-ALIASING] is not used for routed fast convergence, link bandwidth attribute may still be advertised with IP routes (RT-5) to achieve PE-CE link bandwidth based load-balancing as described in this document. In the absence of [EVPN-IP-ALIASING], re-balancing of traffic following changes in PE-CE link bandwidth will require all IP routes from that CE to be re-advertised in a prefix dependent manner.

## 7. EVPN-IRB Multi-homing with non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and NOT as an overlay VPN control plane. EVPN control plane in this case enables:

- o DF election via EVPN RT-4 based procedures described in [RFC7432]
- o LOCAL MAC sync across multi-homing PEs via EVPN RT-2
- o LOCAL ARP and ND sync across multi-homing PEs via EVPN RT-2

Applicability of weighted ECMP procedures proposed in this document to these set of use cases will be addressed in subsequent revisions.

## 7. References

### 7.1 Normative References

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.
- [BGP-LINK-BW] Mohapatra, P., Fernando, R., "BGP Link Bandwidth Extended Community", January 2013, <<https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-06>>.
- [EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <<https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>>.
- [EVPN-DF-PREF] Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., and S. Mohanty, "Preference-based EVPN DF Election", internet-draft ietf-bess-evpn-pref-df-01.txt, April 2018.
- [EVPN-PER-MCAST-FLOW-DF] Sajassi, et al., "Per multicast flow Designated Forwarder Election for EVPN", March 2018, <<https://tools.ietf.org/html/draft-sajassi-bess-evpn-per-mcast-flow-df-election-00>>.
- [EVPN-DF-ELECT-FRAMEWORK] Rabadan, Mohanty, et al., "Framework for EVPN Designated Forwarder Election Extensibility", March 2018, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-df-election-framework-03>>.
- [RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, <<https://tools.ietf.org/html/rfc2119>>.
- [RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", May 2017, <<https://tools.ietf.org/html/rfc8174>>.

### 7.2 Informative References



## 8. Acknowledgements

Authors would like to thank Satya Mohanty for valuable review and inputs with respect to HRW algorithm refinements proposed in this document.

### Authors' Addresses

Neeraj Malhotra, Ed.  
Arrcus  
Email: neeraj.ietf@gmail.com

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

Jorge Rabadan  
Nokia  
Email: jorge.rabadan@nokia.com

John Drake  
Juniper  
EMail: jdrake@juniper.net

Avinash Lingala  
AT&T  
Email: ar977m@att.com

Samir Thoria  
Cisco  
Email: sthoria@cisco.com

BESS Workgroup  
Internet Draft  
Intended status: Standards Track

J. Rabadan, Ed.  
Nokia  
A. Sajassi, Ed.  
Cisco

E. Rosen  
J. Drake  
W. Lin  
Juniper

J. Uttaro  
AT&T

A. Simpson  
Nokia

Expires: January 3, 2019

July 2, 2018

EVPN Interworking with IPVPN  
draft-rabadan-sajassi-bess-evpn-ipvpn-interworking-01

Abstract

EVPN is used as a unified control plane for tenant network intra and inter-subnet forwarding. When a tenant network spans not only EVPN domains but also domains where IPVPN provides inter-subnet forwarding, there is a need to specify the interworking aspects between both EVPN and IPVPN domains, so that the end to end tenant connectivity can be accomplished. This document specifies how EVPN should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families for inter-subnet forwarding.

Status of this Memo

This Internet-Draft is submitted 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/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on January 3, 2019.

#### Copyright Notice

Copyright (c) 2018 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. Introduction and Problem Statement . . . . .	3
2. Terminology and Interworking PE Components . . . . .	3
3. Domain Path Attribute (D-PATH) . . . . .	9
3.1. D-PATH and Loop Prevention . . . . .	11
4. BGP Path Attribute Propagation across ISF SAFIs . . . . .	12
4.1. No-Propagation-Mode . . . . .	12
4.2. Uniform-Propagation-Mode . . . . .	12
4.3. Aggregation of Routes and Path Attribute Propagation . . . . .	13
5. Route Selection Process between EVPN and other ISF SAFIs . . . . .	14
6. Composite PE Procedures . . . . .	15
7. Gateway PE Procedures . . . . .	17
8. Interworking Use-Cases . . . . .	19
9. Conclusion . . . . .	21
10. Conventions used in this document . . . . .	21
11. Security Considerations . . . . .	21
12. IANA Considerations . . . . .	21
13. References . . . . .	21

13.1. Normative References . . . . .	21
13.2. Informative References . . . . .	22
14. Acknowledgments . . . . .	22
15. Contributors . . . . .	22
16. Authors' Addresses . . . . .	22

## 1. Introduction and Problem Statement

EVPN is used as a unified control plane for tenant network intra and inter-subnet forwarding. When a tenant network spans not only EVPN domains but also domains where IPVPN provides inter-subnet forwarding, there is a need to specify the interworking aspects between both EVPN and IPVPN domains, so that the end to end tenant connectivity can be accomplished. This document specifies how EVPN should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families for inter-subnet forwarding.

EVPN supports the advertisement of IPv4 or IPv6 prefixes in two different route types:

- o Route Type 2 - MAC/IP route (only for /32 and /128 host routes), as described by [INTER-SUBNET].
- o Route Type 5 - IP Prefix route, as described by [IP-PREFIX].

When interworking with other BGP address families (AFIs/SAFIs) for inter-subnet forwarding, the IP prefixes in those two EVPN route types must be propagated to other domains using different SAFIs. Some aspects of that propagation must be clarified. Examples of these aspects or procedures across BGP families are: route selection, loop prevention or BGP Path attribute propagation. The Interworking PE concepts are defined in section 2, and the rest of the document describes the interaction between Interworking PEs and other PEs for end-to-end inter-subnet forwarding.

## 2. Terminology and Interworking PE Components

This section summarizes the terminology related to the "Interworking PE" concept that will be used throughout the rest of the document.

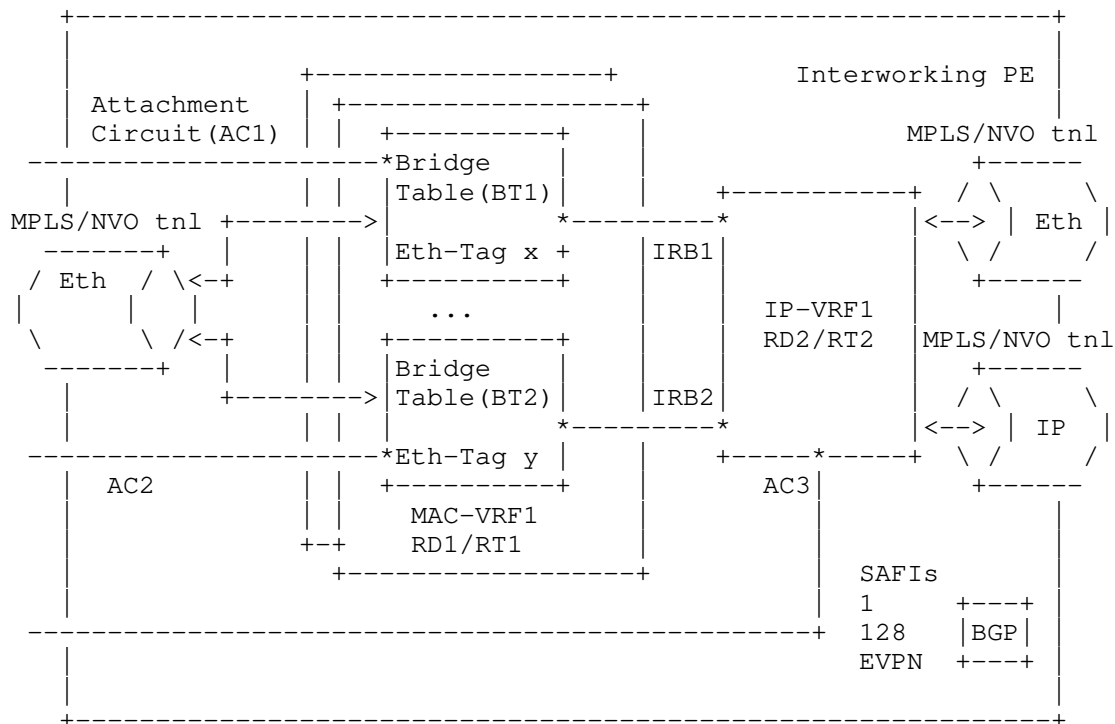


Figure 1 EVPN-IPVPN Interworking PE

- o ISF SAFI: Inter-Subnet Forwarding (ISF) SAFI is a MP-BGP Sub-Address Family that advertises reachability for IP prefixes and can be used for inter-subnet forwarding within a given tenant network. The ISF SAFIs are 1 (including IPv4 and IPv6 AFIs), 128 (including IPv4 and IPv6 AFIs) and 70 (EVPN, including only AFI 25).
- o ISF route: a route for a given prefix whose ISF SAFI may change as it transits different domains.
- o IP-VRF: an IP Virtual Routing and Forwarding table, as defined in [RFC4364]. It is also the instantiation of an IPVPN in a PE. Route Distinguisher and Route Target(s) are required properties of an IP-VRF.
- o MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [RFC7432]. It is also the instantiation of an EVI (EVPN Instance) in a PE. Route Distinguisher and Route Target(s) are required properties and they are normally different than the ones defined in the associated IP-VRF.

- o BT: a Bridge Table, as defined in [RFC7432]. A BT is the instantiation of a Broadcast Domain in a PE. When there is a single Broadcast Domain in a given EVI, the MAC-VRF in each PE will contain a single BT. When there are multiple BTs within the same MAC-VRF, each BT is associated to a different Ethernet Tag. The EVPN routes specific to a BT, will indicate which Ethernet Tag the route corresponds to.

Example: In Figure 1, MAC-VRF1 has two BTs: BT1 and BT2. Ethernet Tag x is defined in BT1 and Ethernet Tag y in BT2.

- o AC: Attachment Circuit or logical interface associated to a given BT or IP-VRF. To determine the AC on which a packet arrived, the PE will examine the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).

Example: In Figure 1, AC1 is associated to BT1, AC2 to BT2 and AC3 to IP-VRF1.

- o IRB: Integrated Routing and Bridging interface. It refers to the logical interface that connects a BT to an IP-VRF and allows to forward packets with destination in a different subnet.
- o MPLS/NVO tnl: It refers to a tunnel that can be MPLS or NVO-based (Network Virtualization Overlays) and it is used by MAC-VRFs and IP-VRFs. Irrespective of the type, the tunnel may carry an Ethernet or an IP payload. MAC-VRFs can only use tunnels with Ethernet payloads (setup by EVPN), whereas IP-VRFs can use tunnels with Ethernet (setup by EVPN) or IP payloads (setup by EVPN or IPVPN). IPVPN-only PEs have IP-VRFs but they cannot send or receive traffic on tunnels with Ethernet payloads.

Example: Figure 1 shows an MPLS/NVO tunnel that is used to transport Ethernet frames to/from MAC-VRF1. The PE determines the MAC-VRF and BT the packets belong to based on the EVPN label (MPLS or VNI). Figure 1 also shows two MPLS/NVO tunnels being used by IP-VRF1, one carrying Ethernet frames and the other one carrying IP packets.

- o RT-2: Route Type 2 or MAC/IP route, as per [RFC7432].
- o RT-5: Route Type 5 or IP Prefix route, as per [IP-PREFIX].
- o Domain: Two PEs are in the same domain if they are attached to the same tenant and the packets between them do not require a data path IP lookup (in the tenant space) in any intermediate router. A gateway PE is always configured with multiple Domain-IDs.

Example 1: Figure 4 depicts an example where TS1 and TS2 belong to the same tenant, and they are located in different Data Centers that are connected by gateway PEs (see the gateway PE definition later). These gateway PEs use IPVPN in the WAN. When TS1 sends traffic to TS2, the intermediate routers between PE1 and PE2 require a tenant IP lookup in their IP-VRFs so that the packets can be forwarded. In this example there are three different domains. The gateway PEs connect the EVPN domains to the IPVPN domain.

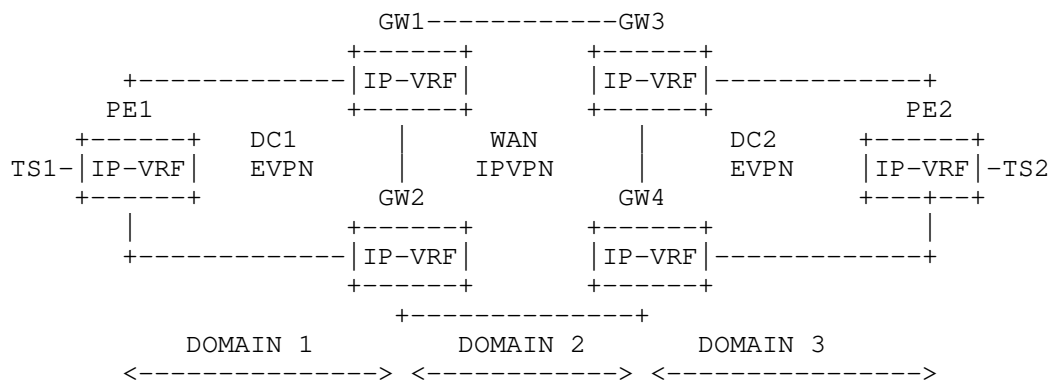


Figure 4 Multiple domain DCI example

Example 2: Figure 5 illustrates a similar example, but PE1 and PE2 are now connected by a BGP-LU (BGP Labeled Unicast) tunnel, and they have a BGP peer relationship for EVPN. Contrary to Example 1, there is no need for tenant IP lookups on the intermediate routers in order to forward packets between PE1 and PE2. Therefore, there is only one domain in the network and PE1/PE2 belong to it.

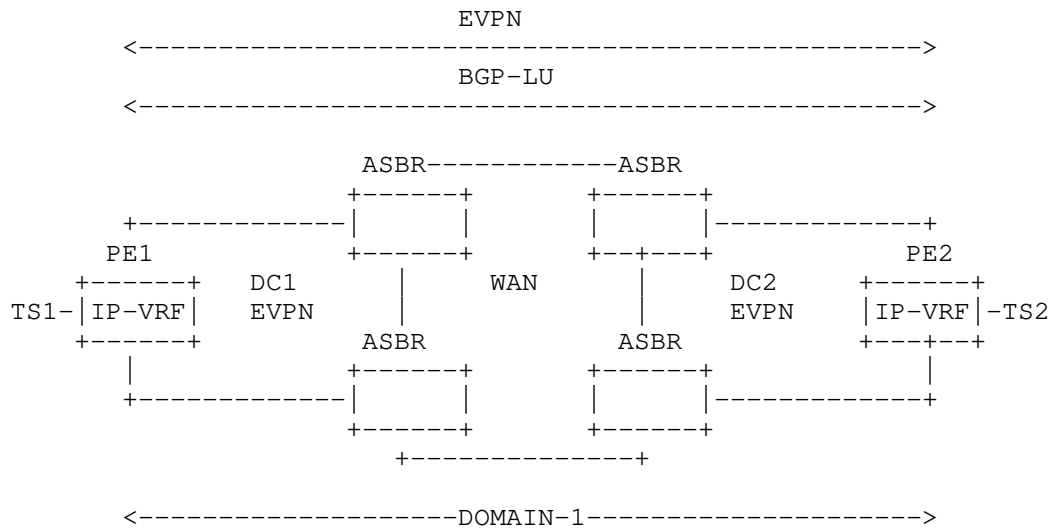


Figure 5 Single domain DCI example

- o Regular Domain: a domain in which a single control plane, IPVPN or EVPN, is used and which is composed of regular PEs, see below. In Figures 4 and 5, above, all domains are regular domains.
- o Composite Domain: a domain in which multiple control planes, IPVPN and EVPN, are used and which is composed of regular PEs, see below, and composite PEs, see below.
- o Regular PE: a PE that is attached to a domain, either regular or composite, and which uses one of the control plane protocols (IPVPN or EVPN) operating in the domain.
- o Interworking PE: a PE that may advertise a given prefix with an EVPN ISF route (RT-2 or RT-5) and/or an IPVPN ISF route. An interworking PE has one IP-VRF per tenant, and one or multiple MAC-VRFs per tenant. Each MAC-VRF may contain one or more BTs, where each BT may be attached to that IP-VRF via IRB. There are two types of Interworking PEs: composite PEs and gateway PEs. Both PE functions can be independently implemented per tenant and they may both be implemented for the same tenant.

Example: Figure 1 shows an interworking PE of type gateway, where ISF SAFIs 1, 128 and 70 are enabled. IP-VRF1 and MAC-VRF1 are instantiated on the PE, and together provide inter-subnet forwarding for the tenant.



- o Composite PE: an interworking PE that is attached to a composite domain and which advertises a given prefix to an IPVPN peer with an IPVPN ISF route, to an EVPN peer with an EVPN ISF route, and to a route reflector with both an IPVPN and EVPN ISF route. A composite PE performs the procedures of Sections 5 and 6.

Example: Figure 2 shows an example where PE1 is a composite PE since PE1 has EVPN and another ISF SAFI enabled to the same route-reflector, and PE1 advertises a given IP prefix IPn/x twice, one using EVPN and another one using ISF SAFI 128. PE2 and PE3 are not composite PEs.

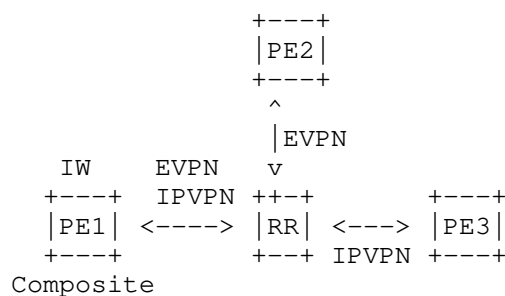


Figure 2 Interworking composite PE example

- o Gateway PE: an interworking PE that is attached to two domains, each either regular or composite, and which, based on configuration, does one of the following:
  - Propagates the same control plane protocol, either IPVPN or EVPN, between the two domains.
  - Propagates an ISF route with different ISF SAFIs between the two domains. E.g., propagate an EVPN ISF route in one domain as an IPVPN ISF route in the other domain and vice versa. A gateway PE performs the procedures of Sections 3, 4, 5 and 7.

A gateway PE is always configured with multiple Domain-IDs. The Domain-ID is encoded in the Domain Path Attribute (D-PATH), and advertised along with EVPN and other ISF SAFI routes. Section 3 describes the D-PATH attribute.

Example: Figure 3 illustrates an example where PE1 is a gateway PE since the EVPN and IPVPN SAFIs are enabled on different BGP peers, and a given local IP prefix IPn/x is sent to both BGP

peers for the same tenant. PE2 and PE1 are in one domain and PE3 and PE1 are in another domain.

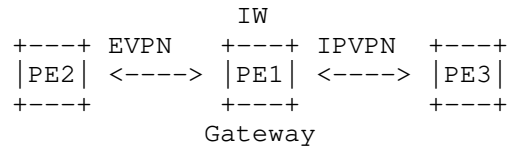


Figure 3 Interworking gateway PE example

- o Composite/Gateway PE: an interworking PE that is both a composite PE and a gateway PE that is attached to two domains, one regular and one composite, and which does the following:
  - Propagates an ISF route, either IPVPN or EVPN, from the regular domain into the composite domain. Within the composite domain it acts as a composite PE.
  - Propagates an ISF route, either IPVPN or EVPN, from the composite domain into the regular domain. Within the regular domain it is propagated as an ISF route using the ISF SAFI for that domain.

This is particularly useful when a tenant network is attached to both IPVPN and EVPN domains, any-to-any connectivity is required, and end-to-end control plane consistency, when possible, is desired.

It would be instantiated by attaching the disparate, regular IPVPN and EVPN domains via these PEs to a central composite domain.

### 3. Domain Path Attribute (D-PATH)

The BGP Domain Path (D-PATH) attribute is an optional and transitive BGP path attribute.

Similar to AS\_PATH, D-PATH is composed of a length field followed by a sequence of Domain segments, where each domain segment is represented by <DOMAIN-ID:ISF\_SAFI\_TYPE>.

- o The length field is a 1-octet field, containing the number of domain segments.

- o DOMAIN-ID is a 6-octet field that represents a domain. It is composed of a 4-octet Global Administrator sub-field and a 2-octet Local Administrator sub-field. The Global Administrator sub-field MAY be filled with an Autonomous System Number (ASN), an IPv4 address, or any value that guarantees the uniqueness of the DOMAIN-ID when the tenant network is connected to multiple Operators.
- o ISF\_SAFI\_TYPE is a 1-octet field that indicates the Inter-Subnet Forwarding SAFI type in which a route was advertised in the DOMAIN. The following types are valid in this document:

Value	Type
1	SAFI 1
70	EVPN
128	SAFI 128

About the BGP D-PATH attribute:

- a) Identifies the sequence of domains, each identified by a <DOMAIN-ID:ISF\_SAFI\_TYPE> through which a given ISF route has passed.
  - This attribute list may contain zero, one or more entries.
  - The first entry in the list (leftmost) is the <DOMAIN-ID:ISF\_SAFI\_TYPE> from which a gateway PE is propagating an ISF route. The last entry in the list (rightmost) is the <DOMAIN-ID:ISF\_SAFI\_TYPE> from which a gateway PE received an ISF route without a D-PATH attribute. Intermediate entries in the list are domains that the ISF route has transited.
  - As an example, an ISF route received with a D-PATH attribute of {<6500:2:IPVPN>,<6500:1:EVPN>} indicates that the ISF route was originated in EVPN domain 6500:1, and propagated into IPVPN domain 6500:2.
- b) It is added/modified by a gateway PE when propagating an update to a different domain:
  - A gateway PE's IP-VRF, that connects two domains, belongs to two DOMAIN-IDs, e.g. 6500:1 for EVPN and 6500:2 for IPVPN.
  - Whenever a prefix arrives at a gateway PE in a particular ISF SAFI route, if the gateway PE needs to export that prefix to a BGP peer, the gateway PE will prepend a <DOMAIN-ID:ISF\_SAFI\_TYPE> segment to the list of segments in the

received D-PATH.

- For instance, in an IP-VRF configured with DOMAIN-IDs 6500:1 for EVPN and 6500:2 for IPVPN, if an EVPN route for prefix P is received and P installed in the IP-VRF, the IPVPN route for P that is exported to an IPVPN peer will prepend the segment <6500:1:EVPN> to the previously received D-PATH attribute. Likewise, IP-VRF prefixes that are received from IP-VPN, will be exported to EVPN peers with the additional segment <6500:2:IPVPN>.
  - In the above example, if the EVPN route is received without D-PATH, the gateway PE will add the D-PATH attribute with segment <6500:1:EVPN> when re-advertising to domain 6500:2.
  - Within the originating domain, the update does not contain a D-PATH attribute because the update has not passed through a gateway PE yet.
- c) The gateway PE MUST NOT add the D-PATH attribute to ISF routes generated for IP-VRF prefixes that are not learned via any ISF SAFI, for instance, local prefixes.
- d) An ISF route received by a gateway PE with a D-PATH attribute that contains one or more of its locally configured domains for the IP-VRF is considered to be a looped ISF route and MUST be dropped.
- e) The number of domain segments in the D-PATH attribute indicates the number of gateway PEs that the ISF route update has transited.

### 3.1. D-PATH and Loop Prevention

The D-PATH attribute is used to prevent loops in interworking PE networks. For instance, in the example of Figure 4, gateway GW1 receives TS1 prefix in two different ISF routes:

- o In an EVPN RT-5 with next-hop PE1 and no D-PATH attribute.
- o In a SAFI 128 route with next-hop GW2 and D-PATH = (6500:1:EVPN), assuming that DOMAIN-ID for domain 1 is 6500:1.

Gateway GW1 flags the SAFI 128 route as a loop, and does not re-advertise it to the EVPN neighbors since the route includes the GW1's local domain.

In general, any interworking PE that imports an ISF route MUST flag the route as "looped" if its D-PATH contains a <DOMAIN-

ID:ISF\_SAFI\_TYPE> segment, where DOMAIN-ID matches a local DOMAIN-ID in the tenant IP-VRF.

#### 4. BGP Path Attribute Propagation across ISF SAFIs

Based on configurations a gateway PE is required to propagate an ISF route with different ISF SAFIs between two domains. This requires a definition of what a gateway PE is to do with Path attributes attached to the ISF route that it is propagating.

##### 4.1. No-Propagation-Mode

This is the default mode of operation. In this mode, the gateway PE will simply re-initialize the Path Attributes when propagating an ISF route, as though it would for direct or local IP prefixes. This model may be enough in those use-cases where the EVPN domain is considered an "abstracted" CE and remote IPVPN/IP PEs don't need to consider the original EVPN Attributes for path calculations.

Since this mode of operation does not propagate the D-PATH attribute either, redundant gateway PEs are exposed to routing loops. Those loops may be resolved by policies and the use of other attributes, such as the Route Origin extended community [RFC4360], however not all the loop situations may be solved.

##### 4.2. Uniform-Propagation-Mode

In this mode, the gateway PE simply keeps accumulating or mapping certain key commonly used Path Attributes when propagating an ISF route. This mode is typically used in networks where EVPN and IPVPN SAFIs are used seamlessly to distribute IP prefixes.

The following rules MUST be observed by the gateway PE when propagating Path Attributes:

- o The gateway PE imports an ISF route in the IP-VRF and stores the original Path Attributes. The following set of Path Attributes SHOULD be propagated by the gateway PE to other ISF SAFIs (other Path Attributes SHOULD NOT be propagated):
  - AS\_PATH
  - D-PATH
  - IBGP-only Path Attributes: LOCAL\_PREF, ORIGINATOR\_ID, CLUSTER\_ID
  - MED
  - AIGP
  - Communities, (non-EVPN) Extended Communities and Large Communities

- o When propagating an ISF route to a different ISF SAFI and IBGP peer, the gateway PE SHOULD copy the AS\_PATH of the originating family and add it to the destination family without any modification. When re-advertising to a different ISF SAFI and EBGp peer, the gateway PE SHOULD copy the AS\_PATH of the originating family and prepend the IP-VRF's AS before sending the route.
- o When propagating an ISF route to IBGP peers, the gateway PE SHOULD copy the IBGP-only Path Attributes from the originating SAFI to the re-advertised route.
- o Communities, non-EVPN Extended Communities and Large Communities SHOULD be copied by the gateway PE from the originating SAFI route.

#### 4.3. Aggregation of Routes and Path Attribute Propagation

Instead of propagating a high number of (host) ISF routes between ISF SAFIs, a gateway PE that receives multiple ISF routes of one ISF SAFI MAY choose to propagate a single ISF aggregate route with a different ISF SAFI. In this document, aggregation is used to combine the characteristics of multiple ISF routes of the same ISF SAFI in such way that a single aggregate ISF route of a different ISF SAFI can be propagated. Aggregation of multiple ISF routes of one ISF SAFI into an aggregate ISF route of a different ISF SAFI is only done by a gateway PE.

Aggregation on gateway PEs may use either the No-Propagation-Mode or the Uniform-Propagation-Mode explained in Sections 4.1. and 4.2, respectively.

When using Uniform-Propagation-Mode, Path Attributes of the same type code MAY be aggregated according to the following rules:

- o AS\_PATH is aggregated based on the rules in [RFC4271]. The gateway PEs SHOULD NOT receive AS\_PATH attributes with path segments of type AS\_SET [RFC6472]. Routes received with AS\_PATH attributes including AS\_SET path segments MUST NOT be aggregated.
- o ISF routes that have different attributes of the following type codes MUST NOT be aggregated: D-PATH, LOCAL\_PREF, ORIGINATOR\_ID, CLUSTER\_ID, MED or AIGP.
- o The Community, Extended Community and Large Community attributes of the aggregate ISF route MUST contain all the Communities/Extended Communities/Large Communities from all of the aggregated ISF routes.

Assuming the aggregation can be performed (the above rules are applied), the operator should consider aggregation to deal with scaled tenant networks where a significant number of host routes exists. For a example, large Data Centers.

#### 5. Route Selection Process between EVPN and other ISF SAFIs

A PE may receive an IP prefix in ISF routes with different ISF SAFIs, from the same or different BGP peer. It may also receive the same IP prefix (host route) in an EVPN RT-2 and RT-5. A route selection algorithm across all ISF SAFIs is needed so that:

- o Different gateway and composite PEs have a consistent and deterministic view on how to reach a given prefix.
- o Prefixes advertised in EVPN and other ISF SAFIs can be compared based on path attributes commonly used by operators across networks.
- o Equal Cost Multi-Path (ECMP) is allowed across EVPN and other ISF SAFI routes.

For a given prefix advertised in one or more non-EVPN ISF routes, the BGP best path selection procedure will produce a set of "non-EVPN best paths". For a given prefix advertised in one or more EVPN ISF routes, the BGP best path selection procedure will produce a set of "EVPN best paths". To support IP/EVPN interworking, it is then necessary to run a tie-breaking selection algorithm on the union of these two sets. This tie-breaking algorithm begins by considering all EVPN and other ISF SAFI routes, equally preferable routes to the same destination, and then selects routes to be removed from consideration. The process terminates as soon as only one route remains in consideration.

The route selection algorithm must remove from consideration the routes following the rules and the order defined in [RFC4271], with the following exceptions and in the following order:

- 1- Immediately after removing from consideration all routes that are not tied for having the highest Local Preference, any routes that do not have the shortest D-PATH are also removed from consideration. Routes with no D-PATH are considered to have a zero-length D-PATH.
- 2- Then regular [RFC4271] selection criteria is followed.
- 3- At the end of the selection algorithm, if at least one route still

under consideration is an RT-2 route, remove from consideration any RT-5 routes.

- 4- Steps 1-3 could possibly leave Equal Cost Multi-Path (ECMP) between IP and EVPN paths. By default, the EVPN path is considered (and the IP path removed from consideration). However, if ECMP across ISF SAFIs is enabled by policy, and an "IP path" and an "EVPN path" remain at the end of step 3, both path types will be used.

Example 1 - PE1 receives the following routes for IP1/32, that are candidate to be imported in IP-VRF-1:

```
{SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=(100,200)}  
{SAFI=EVPN, RT-5, Local-Pref=100, AS-Path=(100,200)}  
{SAFI=128, Local-Pref=100, AS-Path=(100,200)}
```

Selected route: {SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=100,200}  
(due to step 3, and no ECMP)

Example 2 - PE1 receives the following routes for IP2/24, that are candidate to be imported in IP-VRF-1:

```
{SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200),  
MED=10}  
{SAFI=128, D-PATH=(6500:1:EVPN,6500:2:IPVPN), AS-Path=(200),  
MED=200}
```

Selected route: {SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200), MED=10} (due to step 1)

## 6. Composite PE Procedures

As described in Section 2, composite PEs are typically used in tenant networks where EVPN and IPVPN are both used to provide inter-subnet forwarding within the same composite domain.

Figure 6 depicts an example of a composite domain, where PE1/PE2/PE4 are composite PEs (they support EVPN and IPVPN ISF SAFIs on their peering to the Route Reflector), and PE3 is a regular IPVPN PE.



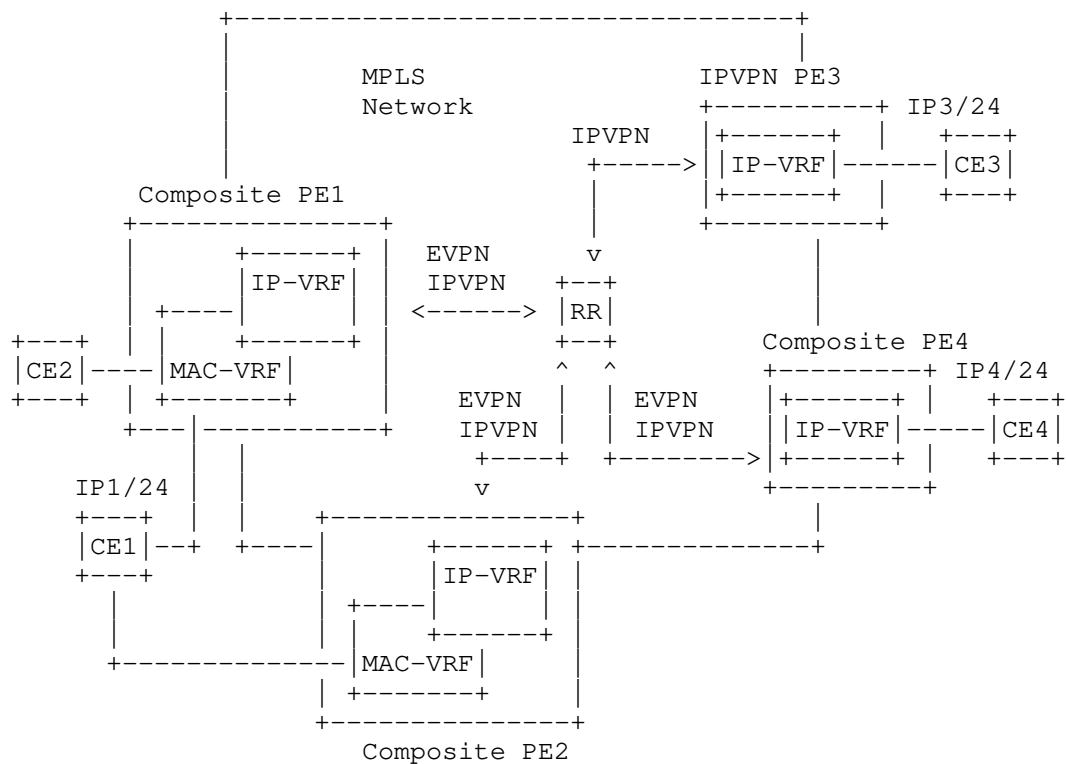


Figure 6 Composite PE example

In a composite domain with composite and regular PEs:

- o The composite PEs advertise the same IP prefixes in each ISF SAFI to the RR. For example, in Figure 6, the prefix IP1/24 is advertised by PE1 and PE2 to the RR in two separate NLRIs, one for AFI/SAFI 1/128 and another one for EVPN.
- o The RR does not forward EVPN routes to PE3 (since the RR does not have the EVPN SAFI enabled on its BGP session to PE3), whereas the IPVPN routes are forwarded to all the PEs.
- o PE3 receives only the IPVPN route for IP1/24 and resolves the BGP next-hop to an MPLS tunnel (with IP payload) to PE1 and/or PE2.
- o Composite PE4 receives IP1/24 encoded in EVPN and another ISF SAFI route (EVPN RT-5 and IPVPN). The route selection follows the procedures in Section 5. Assuming an EVPN route is selected, PE4

resolves the BGP next-hop to an MPLS tunnel (with Ethernet or IP payload) to PE1 and/or PE2. As described in Section 2, two EVPN PEs may use tunnels with Ethernet or IP payloads to connect their IP-VRFs, depending on the [IP-PREFIX] model implemented. If some attributes are modified so that the route selection process (Section 5) results in PE4 selecting the IPVPN path instead of the EVPN path, the operator should be aware that the EVPN advanced forwarding features, e.g. recursive resolution to overlay indexes, will be lost for PE4.

- o The other composite PEs (PE1 and PE2) receive also the same IP prefix via EVPN and IPVPN SAFIs and they also follow the route selection in Section 5.
- o When a given route has been selected as the route for a particular packet, the transmission of the packet is done according to the rules for that route's AFI/SAFI.
- o It is important to note that in composite domains, such as the one in Figure 6, the EVPN advanced forwarding features will only be available to composite and EVPN PEs (assuming they select an RT-5 to forward packets for a given IP prefix), and not to IPVPN PEs. For example, assuming PE1 sends IP1/24 in an EVPN and an IPVPN route and the EVPN route is the best one in the selection, the recursive resolution of the EVPN RT-5s can only be used in PE2 and PE4 (composite PEs), and not in PE3 (IPVPN PE). As a consequence of this, the indirection provided by the RT5's recursive resolution and its benefits in a scaled network, will not be available in all the PEs in the network.

## 7. Gateway PE Procedures

Section 2 defines a gateway PE as an Interworking PE that advertises IP prefixes to different BGP peers, using EVPN to one BGP peer and another ISF SAFI to another BGP peer. Examples of gateway PEs are Data Center gateways connecting domains that make use of EVPN and other ISF SAFIs for a given tenant. Figure 7 illustrates this use-case, in which PE1 and PE2 (and PE3/PE4) are gateway PEs interconnecting domains for the same tenant.

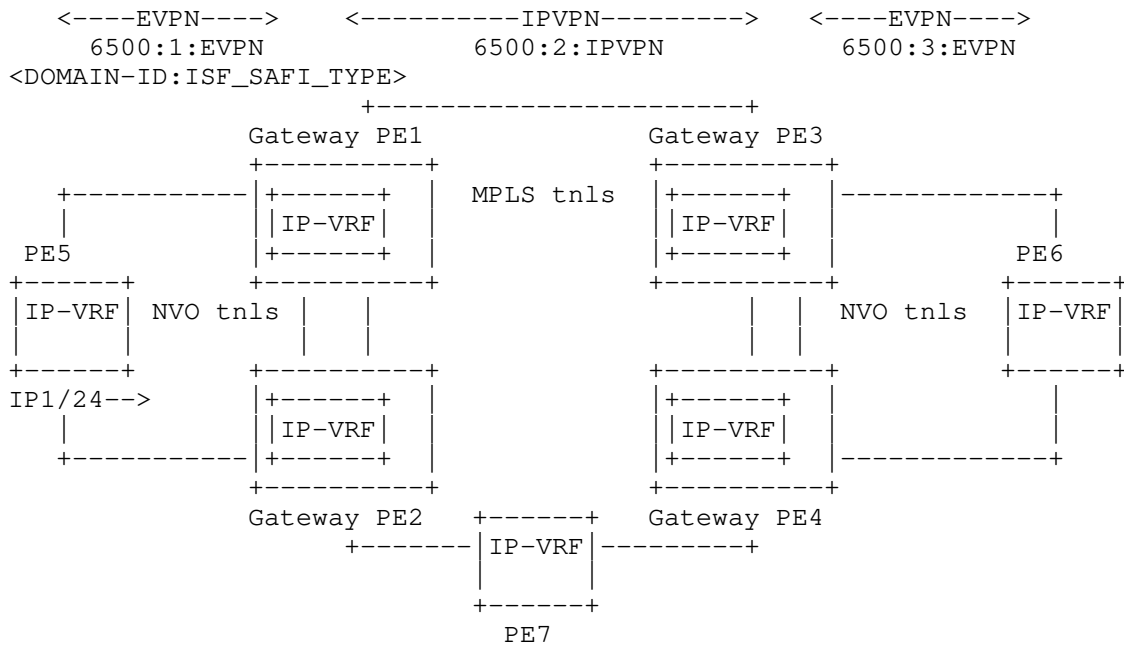


Figure 7 Gateway PE example

The gateway PE procedures are described as follows:

- o A gateway PE that imports an ISF SAFI-x route to prefix P in an IP-VRF, MUST export P in ISF SAFI-y if:
  1. P is installed in the IP-VRF (hence the SAFI-x route is the best one for P) and
  2. PE has a BGP peer for SAFI-y (enabled for the same IP-VRF) and
  3. Either x or y is EVPN.

In the example of Figure 7, gateway PE1 and PE2 receive an EVPN RT-5 with IP1/24, install the prefix in the IP-VRF and re-advertise it using SAFI 128.

- o ISF SAFI routes advertised by a gateway PE MUST include a D-PATH attribute, so that loops can be detected in remote gateway PEs. When a gateway PE propagates an IP prefix between EVPN and another ISF SAFI, it MUST prepend a <DOMAIN-ID:ISF\_SAFI\_TYPE> to the received D-PATH attribute. The DOMAIN-ID and ISF\_SAFI\_TYPE fields refer to the domain over which the gateway PE received the IP prefix and the ISF SAFI of the route, respectively. If the received

IP prefix route did not include any D-PATH attribute, the gateway IP MUST add the D-PATH when readvertising. The D-PATH in this case will have only one segment on the list, the <DOMAIN-ID:ISF\_SAFI\_TYPE> of the received route.

In the example of Figure 7, gateway PE1/PE2 receive the EVPN RT-5 with no D-PATH attribute since the route is originated at PE5. Therefore PE1 and PE2 will add the D-PATH attribute including <DOMAIN-ID:ISF\_SAFI\_TYPE> = <6500:1:EVPN>. Gateways PE3/PE4 will propagate the route again, now prepending their <DOMAIN-ID:ISF\_SAFI\_TYPE> = <6500:2:IPVPN>. PE6 receives the EVPN RT-5 routes with D-PATH = {<6500:2:IPVPN>, <6500:1:EVPN>} and can use that information to make BGP path decisions.

- o The gateway PE MAY use the Route Distinguisher of the IP-VRF to readvertise IP prefixes in EVPN or the other ISF SAFI.
- o The label allocation used by each gateway PE is a local implementation matter. The IP-VRF advertising IP prefixes for EVPN and another ISF SAFI may use a label per-VRF, per-prefix, etc.
- o The gateway PE MUST be able to use the same or different set of Route Targets per ISF SAFI on the same IP-VRF. In particular, if different domains use different set of Route Targets for the same tenant, the gateway PE MUST be able to import and export routes with the different sets.
- o Even though Figure 7 only shows two domains per gateway PE, the gateway PEs may be connected to more than two domains.
- o There is no limitation of gateway PEs that a given IP prefix can pass through until it reaches a given PE.
- o It is worth noting that an IP prefix that was originated in an EVPN domain but traversed a different ISF SAFI domain, will lose EVPN-specific attributes that are used in advanced EVPN procedures. For example, even if PE1 advertises IP1/24 along with a given non-zero ESI (for recursive resolution to that ESI), when PE6 receives the IP prefix in an EVPN route, the ESI value will be zero. This is because the route traverses an ISF SAFI domain that is different than EVPN.

## 8. Interworking Use-Cases

While Interworking PE networks may well be similar to the examples described in Sections 6 and 7, in some cases a combination of both functions may be required. Figure 8 illustrates an example where the

gateway PEs are also composite PEs, since not only they need to re-advertise IP prefixes from EVPN routes to another ISF SAFI routes, but they also need to interwork with IPVPN-only PEs in a domain with a mix of composite and IPVPN-only PEs.

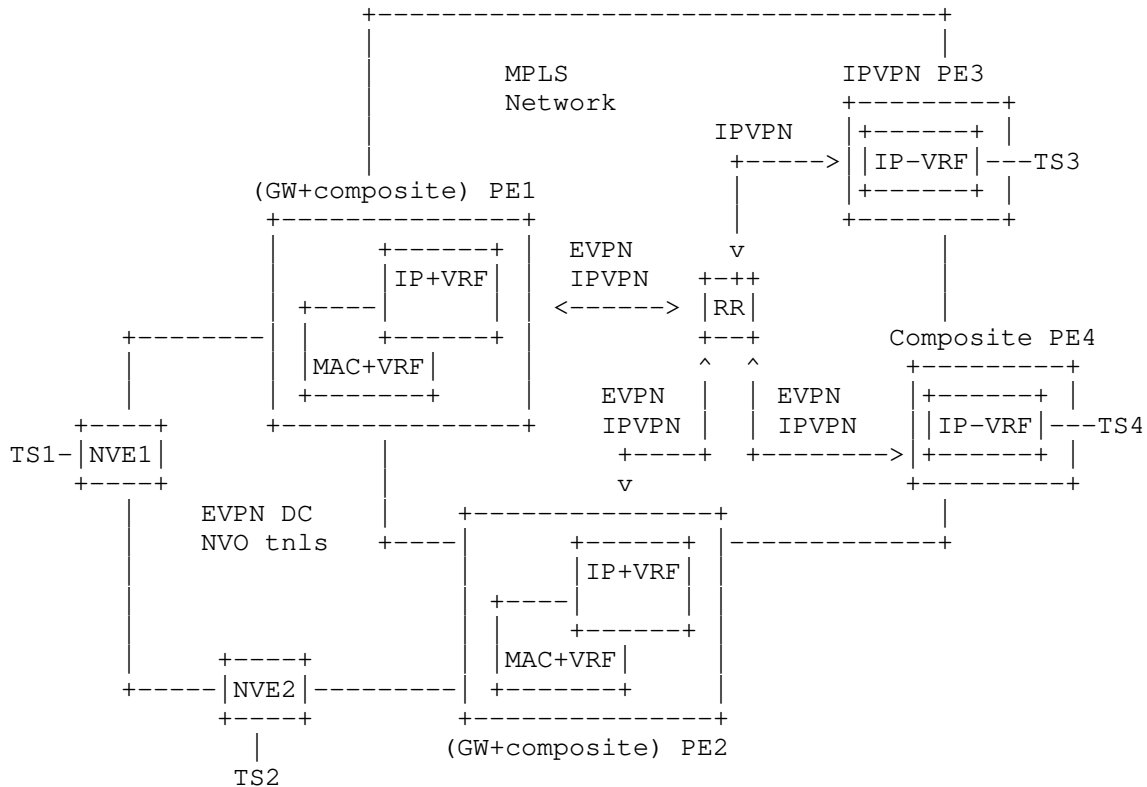


Figure 8 Gateway and composite combined functions - example

In the example above, PE1 and PE2 MUST follow the procedures described in Sections 6 and 7. Compared to section 7, PE1 and PE2 now need to also propagate prefixes from EVPN to EVPN, in addition to propagating prefixes from EVPN to IPVPN.

It is worth noting that PE1 and PE2 will receive TS4's IP prefix via IPVPN and RT-5 routes. When readvertising to NVE1 and NVE2, PE1 and PE2 will consider the D-PATH rules and attributes of the selected route for TS4 (Section 5 describes the Route Selection Process).

## 9. Conclusion

This document describes the procedures required in PEs that use EVPN and another Inter-Subnet Forwarding SAFI to import and export IP prefixes for a given tenant. In particular, this document defines:

- o A route selection algorithm so that a PE can determine what path to choose between EVPN paths and other ISF SAFI paths.
- o A new BGP Path attribute called D-PATH that provides loop protection and visibility on the domains a particular route has traversed.
- o The way Path attributes should be propagated between EVPN and another ISF SAFI.
- o The procedures that must be followed on Interworking PEs that behave as composite PEs, gateway PEs or a combination of both.

The above procedures provide an operator with the required tools to build large tenant networks that may span multiple domains, use different ISF SAFIs to handle IP prefixes, in a deterministic way and with routing loop protection.

## 10. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 11. Security Considerations

This section will be added in future versions.

## 12. IANA Considerations

This document defines a new BGP path attribute known as the BGP Domain Path (D-PATH) attribute and requests IANA to assign a new attribute code type from the "BGP Path Attributes" subregistry under the "Border Gateway Protocol (BGP) Parameters" registry.

## 13. References

### 13.1. Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.

### 13.2. Informative References

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<http://www.rfc-editor.org/info/rfc4360>>.

[IP-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11, May, 2018.

[INTER-SUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-03.txt, work in progress, February, 2017

[ENCAP-ATT] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-09.txt, work in progress, February, 2018.

[RFC6472] Kumari, W. and K. Sriram, "Recommendation for Not Using AS\_SET and AS\_CONFED\_SET in BGP", BCP 172, RFC 6472, DOI 10.17487/RFC6472, December 2011, <<https://www.rfc-editor.org/info/rfc6472>>.

### 14. Acknowledgments

### 15. Contributors

### 16. Authors' Addresses

Jorge Rabadan (editor)  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

Ali Sajassi (editor)  
Cisco  
170 West Tasman Drive  
San Jose, CA 95134, US  
EMail: sajassi@cisco.com

Eric C. Rosen  
Juniper Networks, Inc.  
EMail: erosen@juniper.net

John Drake  
Juniper Networks, Inc.  
EMail: jdrake@juniper.net

Wen Lin  
Juniper Networks, Inc.  
EMail: wlin@juniper.net

Jim Uttaro  
AT&T  
Email: jul738@att.com

Adam Simpson  
Nokia  
Email: adam.1.simpson@nokia.com



Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: December 22, 2018

E. Rosen, Ed.  
R. Bonica  
Juniper Networks, Inc.  
June 20, 2018

Augmenting RFC 4364 Technology to  
Provide Secure Layer L3VPNs over Public Infrastructure  
draft-rosen-bess-secure-l3vpn-01

Abstract

The Layer 3 Virtual Private Network (VPN) technology described in RFC 4364 is focused on the scenario in which a network Service Provider (SP) maintains a secure backbone network and offers VPN service over that network to its customers. Customers access the SP's network by attaching "Customer Edge" (CE) routers to "Provider Edge" (PE) routers, which exchange cleartext IP packets. PE routers generally serve multiple customers, and prevent unauthorized communication among customers. Customer data sent across the backbone (from one PE to another) is encapsulated in MPLS, using an MPLS label to associate a given packet with a given customer. The labeled packets are then sent across the backbone network in the clear, using MPLS transport. However, many customers want a VPN service that is secure enough to run over the public Internet, and which does not require them to send cleartext IP packets to a service provider. Often they want to connect directly to edge nodes of the public Internet, which does not provide MPLS support. Each customer may itself have multiple tenants who are not allowed to intercommunicate with each other freely. In this case, the customer may need to provide a VPN service for the tenants. This document describes a way in which this can be achieved using the technology of RFC 4364. The functionality assigned therein to a PE router can be placed instead in Customer Premises Equipment. This functionality can be augmented by transmitting MPLS packets through IPsec Security Associations. The BGP control plane sessions can also be protected by IPsec. This allows a customer to use RFC 4364 technology to provide VPN service to its internal departments, while sending only IPsec-protected packets to the Internet or other backbone network, and eliminating the need for MPLS transport in the backbone.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute

working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on December 22, 2018.

#### Copyright Notice

Copyright (c) 2018 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 (<https://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. Introduction . . . . .	3
1.1. Review of L3VPN Concepts and Terminology . . . . .	3
1.2. Secured L3VPN . . . . .	4
1.3. Terminology . . . . .	5
2. Model of Operation . . . . .	7
3. How the C-PEs Advertise Red Routes . . . . .	10
3.1. Red and Black C-PE Loopback Addresses . . . . .	10
3.2. Setting Up Red BGP Sessions Between C-PEs and RRs . . . . .	11
3.3. Routes Transmitted by the C-PE on Red BGP Sessions . . . . .	13
3.4. Propagating Red Routes . . . . .	13
4. Resolving the Next Hop of a Red VPN-IP Route . . . . .	14
5. MPLS-in-IPsec . . . . .	16
6. Security Handle . . . . .	17
7. Data Plane Security Procedures . . . . .	17
8. Implementation Challenges . . . . .	18
9. Security Considerations . . . . .	18
10. IANA Considerations . . . . .	19
11. Acknowledgments . . . . .	19
12. References . . . . .	19
12.1. Normative References . . . . .	19
12.2. Informational References . . . . .	20

Authors' Addresses . . . . .	21
------------------------------	----

## 1. Introduction

### 1.1. Review of L3VPN Concepts and Terminology

In conventional Virtual Private Networks (L3VPNs) based on the technology of [RFC4364], a Service Provider (SP) maintains a secure private network (known as the "SP backbone"). An SP maintains a number of "Provider Edge" (PE) routers to which customers may attach. A customer router that attaches to a PE router is known as a "Customer Edge" (CE) router.

Multiple customers may connect to a single PE router. Within a given PE, each customer is associated with a routing context of its own (known as a Virtual Routing and Forwarding table, or VRF). A particular customer attaches to the PE via a set of one or more interfaces or Virtual LANs (VLANs) that are not shared with other customers. (In the subsequent text, the term "interface" will include VLANs and other "virtual" interfaces.) Each such interface is associated with a particular customer's VRF; thus such interfaces are known as "VRF interfaces". These are the PE's "customer-facing" interfaces. The VRF interfaces carry IP datagrams, either IPv4 or IPv6 or both.

A given customer's VRF is automatically populated with, and only with:

- o routes that lead out the local VRF interfaces, and
- o routes that lead to remote VRF interfaces of the same customer.

Routes leading outside a customer's VPN are excluded from that customer's VRF unless explicitly allowed by policy. Thus two customers can attach to the same PE even if they are not allowed to communicate with each other through that PE.

The PE at which a customer data packet enters the SP backbone network is known as the packet's "ingress PE". The PE at which a customer data packet leaves the SP backbone is known as the packet's "egress PE". Generally, the ingress PE pushes two MPLS labels onto each data packet. The top label (sometimes known as the "transport label") directs the packet to its egress PE. The second label (sometimes known as the "VPN label") is used at the egress PE to associate a given customer's packets with that customer's VRF at the egress PE.

These labeled packets travel across the SP backbone "in the clear" (i.e., with no cryptographic protection to provide privacy,

authentication, or integrity), as the SP backbone is presumed to be adequately secure.

The control plane protocol for this type of VPN is BGP. A given customer's routes are distributed among the PEs to which that customer attaches by means of a BGP address family known as "VPN-IP" (either VPN-IPv4 or VPN-IPv6). Distribution of these routes is controlled in such a way as to ensure that a given customer's routes, exported from one of that customer's VRFs, are imported only by other VRFs associated with the same customer.

## 1.2. Secured L3VPN

For security reasons, the L3VPN technology summarized in Section 1.1 is not generally used in the following scenarios:

- o Some or all of the customer sites need to be reached over a network that is untrusted (e.g., the public Internet).
- o The customer wants to be very sure that its SP is not able to read or modify its data.
- o The customer does not want to expose any of its routing control information to the SP, and/or wishes to hide his internal IP addressing structure from the SP.

In such situations, the customer needs to use cryptographic methods in order to ensure privacy, integrity, and authentication for the IP datagrams sent over the backbone network; the cryptography must be applied before the datagrams are sent to the SP backbone network or Internet. (It is presumed of course that the customer's own sites and systems have been satisfactorily secured; how that is achieved is outside the scope of this document.)

In these use cases, the customer may still want some of the benefits of the L3VPN service, e.g.:

- o The customer may itself be providing a VPN service to multiple "tenants". E.g.,
  - \* The customer may be an enterprise or governmental agency that consists of multiple internal departments or organizations that are not allowed to communicate freely with each other, and that may even have independent IP address spaces. We will use the term "tenant" to refer to such a department or organization.
  - \* The customer may be a Data Center operator that is providing a virtual network to each of multiple Data Center tenants, and

needs to extend some or all of those virtual networks over a non-secured backbone network.

(Of course, the same technology works when there is only a single tenant.)

- o In L3VPN, a CE router at one customer site does not have to be provisioned with the addresses of CE routers at other sites. Rather, these are auto-discovered via BGP. This sort of auto-discovery is just as valuable when the customer needs more security than is provided by conventional L3VPN. Auto-discovery also allows some or all of the CE routers to be mobile, changing their IP addresses from time to time; for some customers, this is a mission-critical need.

It is possible to adapt the L3VPN technology to handle use cases where cryptographic methods must be applied before a packet is sent to an SP or to a backbone network. This document describes a way in which this may be done. We will refer to this adaptation as a "Secured L3VPN". Section 2 outlines the way this adaptation works. Subsequent sections of this document specify the necessary procedures in more detail.

Secured L3VPN makes use of IPsec technology. This document does not discuss the details of IPsec. A roadmap through the set of RFCs describing IPsec can be found in [RFC6071]. Of particular importance are [RFC4303] (IPsec Encapsulating Security Payload), [RFC7296] (Internet Key Exchange Protocol version 2), and [RFC8221] (Cryptographic Algorithm Implementation Requirements and Usage Guidance).

### 1.3. Terminology

In this document we shall use the following terminology:

- o SP:

A network service provider (possibly an Internet service provider).

- o customer:

An organization or other entity that obtains network service (private network service or Internet service) from an SP.

- o tenant:

An organization or other entity that obtains VPN service from the customer. For example, if the customer is a governmental agency, its tenants might be the various departments of the agency. If the customer is an enterprise, its tenants might be the various organizations within the enterprise. If the customer is a Data Center provider, its tenants might be organizations to which it sells Data Center services.

- o C-PE router:

A router that performs the functions of an L3VPN PE router ([RFC4364]), but that is operated and managed not by a network service provider, but rather by a customer of the network service provider. The customer may use the C-PE to provide a Secured L3VPN service to one or more of its tenants.

- o Red interface:

A tenant-facing interface of a C-PE device, where the tenant in question is receiving Secured L3VPN service.

- o Black interface:

A C-PE interface that is not a red interface. This may be an interface to the Internet, to an SP backbone network, or to a tenant that is not receiving the Secure L3VPN service.

- o Red BGP session:

A BGP session, protected by IPsec, between two C-PEs, or between a C-PE and a BGP Route Reflector.

- o Black BGP session:

A BGP session other than a red BGP session.

- o Black Network:

The part of the communications infrastructure that is not trusted or not regarded as adequately secure. E.g., the public Internet is a black network.

- o Red Route:

- \* Local red route:

A route whose next hop interface is a local red interface.

\* Remote red route:

A route received, as a VPN-IPv4 or VPN-IPv6 route, over a red BGP session.

\* Red Loopback Route: This term is defined in Section 3.3.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Model of Operation

In a Secured L3VPN, the functions conventionally performed by an L3VPN PE router (as detailed in [RFC4364]) are instead performed by a router that is operated and managed by the customer, rather than by the SP. Since such a router is part of the customer's network, but has the functionality of an L3VPN PE router, we will refer to it as a "C-PE router". The customer is responsible for ensuring that the C-PE itself is properly secured. The C-PE provides L3VPN functionality to the customer's tenants.

Each interface of a C-PE is either a "red interface" or a "black interface":

- o A red interface is a tenant-facing interface that attaches to a tenant who is receiving Secured L3VPN service from the customer.
- o A black interface is any interface that is not a red interface. Black interfaces may be backbone-facing interfaces (attached to an SP backbone), or may be tenant-facing interfaces attached to tenants that are not receiving any L3VPN from the customer.

We assume in this document that a C-PE that provides the Secured L3VPN service to one or more tenants does not provide a conventional (unsecured) L3VPN service to any of the tenants.

(Note that a black interface could also be attached to an "Internet Gateway", owned by a single tenant or shared by multiple tenants, that provides controlled access to the Internet for that tenant or tenants. This scenario is not further discussed in this document.)

A C-PE has one or more VRFs, one per tenant. Each VRF is associated with a distinct set of red interfaces, the ones that lead to the network(s), VLAN(s), or virtual network(s) that is (are) specific to

the given tenant. Standard L3VPN techniques then prevent communication among the different tenants unless explicitly allowed by policy. In simpler scenarios, the customer may have sites with only a single tenant. The C-PEs at those sites require only a single VRF, and all the red interfaces will be associated with that VRF.

The black interfaces of a C-PE can attach to an access router of the public internet, or to a conventional L3VPN PE router belonging to an SP, or to any other router that provides IP connectivity among the customer's C-PE routers. (If a C-PE attaches to a conventional L3VPN PE router, then the C-PE appears to the conventional PE to be a CE router.)

As in any L3VPN, the VRFs are populated with a combination of local routes and remote routes:

- o The local routes in a given VRF are those routes whose "next hop interface" is a local red interface associated with that VRF.
- o The remote routes in a given VRF are those routes learned via BGP from the customer's other C-PEs. These routes may be learned directly via BGP sessions to those other C-PEs, or indirectly via one or more secure BGP Route Reflectors (RRs).

In this document, we will use the term "red routes" to refer to routes within a VRF. These are distinguished from the "black routes" that exist in a C-PE's global routing table.

A conventional PE router sends and receives MPLS packets over its backbone-facing interfaces. A C-PE, on the other hand, sends "MPLS-in-IPsec" packets (see [RFC4023] and Section 5 of this document) over its backbone-facing black interfaces. Since an MPLS-in-IPsec packet is an IP datagram, there is no need for the backbone network to support MPLS transport. IPsec is used to provide privacy, integrity and authentication for the packets sent by the C-PE to the backbone network.

In a Secured L3VPN, protection of the control plane is just as important as is protection of the data plane. It is therefore necessary to ensure that the BGP messages used to disseminate the red routes also have privacy, integrity, and authentication. In order to ensure this, the BGP sessions used to disseminate information about red routes will be protected by IPsec. We will refer to such BGP sessions as "red BGP sessions". It is recommended to use IPsec Transport mode to protect these BGP sessions. This means that a C-PE MUST NOT send or receive VPN-IP routes over any BGP session that is not protected by IPsec. (A VPN-IP route is a route whose BGP Address



Family (AFI) is 1 (IPv4) or 2 (IPv6) and whose Subsequent Address Family (SAFI) is 128, 129, or 5.)

Note that if RRs are used, the RRs must be as secure as the C-PEs. This likely means that they are managed by the customer and located at sites regarded by the customer as adequately secure.

Thus in a Secured L3VPN, red routes are propagated only among trusted systems, and only via red BGP sessions. The propagation of red routes on red BGP sessions is controlled by attaching Route Targets to those routes, as with any [RFC4364]-based technology.

As in any L3VPN, BGP uses the VPN-IPv4 and/or VPN-IPv6 address families when disseminating information about VPN routes from one VRF to another. Each such route carries an MPLS label that is to be pushed on the label stack of any tenant packet for which the address prefix in the route's NLRI is the best match to the packet's IP destination address.

In Secured L3VPNs, these routes MUST also carry a Tunnel Encapsulation attribute ([TUNNEL\_ENCAPS]) specifying the "MPLS-in-IPsec" tunnel type (see Section 5, as well as Sections 3 and 8.1 of [RFC4023]). This indicates that before a tenant's MPLS packet is sent to the backbone network, it must be encapsulated in IP and then sent on an IPsec Transport Mode Security Association (SA).

A C-PE may have unprotected (black) BGP sessions, e.g., to gather public Internet routes. However, the black BGP sessions MUST NOT be enabled for the VPN-IP AFI/SAFIs. This prevents any routes learned over the black BGP sessions from being imported into the VRFs.

As we shall see in Section 3.3, there are also some IP routes (as distinguished from VPN-IP routes) that MUST NOT be transmitted on black BGP sessions, and that MUST be ignored if received on black BGP sessions. These are known as the "red loopback routes".

The procedures of this document result in a network overlay whose control plane consists of red BGP sessions, and whose data plane consists of MPLS-in-IPsec Security Associations. This allows an SP's customer to provide Secured L3VPN service to its tenants.

When RRs are used, C-PEs "register" with the RRs by setting up BGP sessions to them, running the BGP sessions through IPsec SAs. The procedures for setting up IPsec SAs between a C-PE and an RR will authenticate the C-PE to the RR, and vice versa. One C-PE learns of another other C-PE's presence when the RR propagates routes from the latter C-PE to the former.

The procedures specified in this document result in one MPLS-in-IPsec SA between a given pair of C-PEs. This one SA will carry the traffic of all the tenants that are attached to both C-PEs. That should provide adequate security, as the tenants' data is already exposed to the C-PEs. If for some reason it is desired to have a distinct SA for each tenant, a method of doing so is mentioned in Section 4.

### 3. How the C-PEs Advertise Red Routes

#### 3.1. Red and Black C-PE Loopback Addresses

To support the Secured L3VPN control plane, each C-PE MUST have two loopback addresses. One of these will be known as its "red loopback", the other as its "black loopback".

The black loopbacks MUST be addresses that are globally routable. That is, they are public addresses. (Strictly speaking, the black loopback only needs to be routable in any network that might be used to carry traffic between two C-PEs. But we will assume that traffic between two C-PEs might need to traverse the public Internet.) Typically a C-PE's black loopback will be in the address space administered by the network service provider to which the C-PE attaches. The service provider may assign it dynamically, or it may be assigned statically and configured in the C-PE by the customer.

In addition to having a globally routable black loopback, a C-PE will of course have globally routable interface addresses for each of its black interfaces.

Interface addresses of the red interfaces SHOULD NOT be globally routable.

If the C-PE attaches to multiple service providers, the black loopback is likely to be a provider-independent address. However, it MUST be routable in the backbone network of both providers, and most likely will need to be globally routable.

The C-PE may have one or more (black) BGP sessions with service provider peers, in which case it may advertise the black loopback; the next hop field of such an advertisement would be the interface address of the interface over which that BGP session runs. In some scenarios, it may be sufficient to advertise the black loopback via an IGP.

Each C-PE of a given customer MUST be provisioned with a red loopback that is unique among the set of C-PEs of that customer. The red loopback SHOULD NOT be a routable address in the public Internet or in the backbone networks of any service provider to which any of the

C-PEs is attached. If a C-PE has a (black) BGP session with a service provider peer, it MUST NOT advertise a route to its red loopback over that session. That is, any IP route to a red loopback is considered to be a red route, and MUST NOT be advertised or received on a black BGP session. These "red loopback routes" can thus be considered to be "red routes", even though they are IP rather than VPN-IP routes.

### 3.2. Setting Up Red BGP Sessions Between C-PEs and RRs

The customer is expected to have two or more BGP Route Reflectors (red RRs). The red RRs are presumed to be secure; making them so is the responsibility of the customer. As with the C-PEs, each red RR has a black loopback and a red loopback. If the RR is not also a C-PE, it will have only black interfaces, each of course with a globally routable interface address.

A customer's red RRs will form BGP sessions with that customer's C-PEs. These BGP sessions MUST be protected by IPsec. The use of IPsec transport mode is RECOMMENDED. If the RR's red loopback is an IPv4 address, it may be used as the RR's BGP Identifier (see [RFC4271] and [RFC6286]).

When a C-PE device comes up, it attempts to set up an IPsec-protected BGP session with the red RRs. This requires first setting up an IPsec SA with each red RR, and then using IPsec Transport Mode to protect the BGP session.

If the C-PE's red loopback is an IPv4 address, the C-PE's BGP Identifier (see [RFC4271] and [RFC6286]) may be the red loopback.

The endpoint addresses of the IPsec SA are the black loopbacks of the endpoint systems.

Therefore, in order to initiate a BGP session to a red RR, a C-PE must be provisioned to know a publicly routable address (i.e., the black loopback) of the RR. A C-PE must also be provisioned with whatever additional information is needed in order to set up an IPsec SA with each of the red RRs. Each C-PE will attempt to continuously maintain live BGP sessions (protected by IPsec) with each red RR. Note that the source and destination IP address fields of the IP datagrams carrying the IPsec-encapsulated BGP messages will be publicly routable addresses.

In some scenarios, it may be desirable to provision each red RR with the publicly routable address and pre-shared secret of every C-PE. This makes it easy for the C-PEs to authenticate themselves to the

RR, but requires each RR to be reprovisioned every time a new C-PE is added to the network.

In other scenarios, it may be considered desirable to allow the RRs to auto-discover the C-PEs, without the need for any per-C-PE pre-provisioning of the RRs. In this case, a certificate-based authentication method can be used when setting up the IPsec SAs that carry the BGP sessions.

In either type of scenario, the C-PE SHOULD NOT be assumed to have a fixed black loopback address or fixed black interface addresses; rather, it SHOULD be assumed that a C-PE might be a mobile device whose globally routable addresses change from time to time.

If a customer's C-PEs support multiple VPNs (for multiple tenants), that customer's red RRs will receive and disseminate the VPN-IP routes of all those VPNs.

Note that according to the above procedures, the C-PEs will only have red BGP sessions to the red RRs; the C-PEs will not have BGP sessions to each other. Thus it is not necessary for the C-PEs to know of each other in advance. Of course, if a particular customer deems it desirable for the C-PEs to have red BGP sessions to each other, each C-PE can be provisioned with a publicly routable address of each other C-PE, along with any additional information needed to set up an IPsec SA to each other C-PE.

It is RECOMMENDED that, for the purpose of setting up the red BGP sessions, all the RRs and C-PEs be considered to be in the same Autonomous System (AS). Then the red BGP sessions will all be IBGP sessions, and the next hop field of a red route will not be modified as the route is propagated. Note that if an implementation allows a given router to be part of two different ASes, this does not require that all the C-PEs and red RRs attach to the Internet via the same AS. The "red overlay" may appear to be within a single AS, but the "black underlay" need not be within a single AS.

If it is necessary to use an EBGP session between a C-PE and an RR (perhaps because the implementation does not allow one router to be part of two different ASes), the RR SHOULD have a configured policy to leave the next hop unchanged when propagating red VPN-IP routes on an EBGP session. See Section 3.4.

In some scenarios, the C-PEs may set up red BGP sessions to Autonomous System Border Routers (ASBRs), rather than to RRs, creating what is sometimes known as an "option B interconnect" (Section 10 of [RFC4364]). This is transparent to the C-PE.

### 3.3. Routes Transmitted by the C-PE on Red BGP Sessions

A C-PE MUST propagate its local VPN-IP routes on the red BGP sessions, and only on the red BGP sessions. The next hop of each local VPN-IP route MUST be set to the red loopback of the C-PE. The choice to transmit a particular VPN-IP route on a particular session may of course be influenced by the route's Route Targets.

A C-PE MUST NOT transmit its local VPN-IP routes on black BGP sessions. VPN-IP routes MUST NOT be accepted from black BGP sessions.

In all other respects, the handling of VPN-IP routes is done by normal L3VPN procedures.

Each C-PE MUST also transmit the following IP (IPv4 or IPv6) route on the red BGP sessions. We refer to this route as the C-PE's "red loopback route":

- o The address prefix field of the route's Network Layer Reachability Information (NLRI) contains the C-PE's red loopback as a host address.
- o The Next Hop of the route is the C-PE's black loopback.
- o The route carries a Tunnel Encapsulation Attribute [TUNNEL\_ENCAPS] with the the following parameters:
  - \* Tunnel Type = "MPLS-in-IPsec" (see Section 5.)
  - \* Remote Endpoint = the C-PE's black loopback
  - \* An optional "Security Handle" (see Section 6). This provides any information needed by another C-PE to set up an MPLS-in-IPsec Security Association with the advertising C-PE.

A C-PE MUST NOT transmit, on any black BGP session, an IP route whose NLRI contains its red loopback.

A given C-PE's red loopback route must be propagated to all other the C-PEs belonging to the same customer. Therefore, such routes SHOULD NOT carry Route Targets.

### 3.4. Propagating Red Routes

A route that is received over a red BGP session may need to be propagated to other red BGP sessions. A route that is received over a red BGP session MUST NOT be propagated over a black BGP session.

Similarly, a route that is received over a black BGP session MUST NOT be propagated over a red BGP session.

When a route is propagated from one red BGP session to another, its next hop SHOULD be left unchanged. As specified in Section 4, this will ensure that a data packet sent on the path advertised by that route are sent on an IPsec SA between its ingress C-PE and its egress C-PE. Changing the next hop would change the IPsec SA endpoint.

Changing the next hop may be useful in certain deployments. For instance, the path from an ingress C-PE to an egress C-PE may traverse several ASBRs. If these ASBRs are secure, it may be desirable to set up a sequence of IPsec SAs, (e.g., C-PE1--ASBR1, ASBR1--ASBR2, ASBR2--C-PE2) instead of using a single IPsec SA between C-PE1 and C-PE2. (This reduces the number of IPsec sessions supported by a C-PE, at the cost of requiring secure ASBRs along the path.) If this is not the intention, the red BGP sessions MUST leave the next hop unchanged, even if those sessions are EBGP sessions.

In all other respects, propagation of red routes is governed by the normal procedures for propagating routes. If the route carries one or more Route Targets, these may affect its propagation. However, note that propagation of a route between a red BGP session and a black BGP session MUST NOT be done, irrespective of the Route Targets.

#### 4. Resolving the Next Hop of a Red VPN-IP Route

Suppose a C-PE, say C-PE1, receives a packet, say packet P, on one of its local red interfaces. Suppose that packet P is addressed to a system that is reached via one of the red interfaces of another C-PE, say C-PE2. C-PE1 looks up packet P's destination address in the VRF associated with P's incoming interface. The matching route will be a "Labeled VPN-IP route" [RFC4364] originated by C-PE2, and disseminated to C-PE1 over a red BGP session. Per Section 3.3, the next hop of that route will be C-PE2's red loopback.

The labeled VPN-IP route matched by packet P's destination address will contain an MPLS label, the "VPN label". C-PE1 pushes the VPN label onto packet P's MPLS label stack. Then C-PE1 needs to determine how to transmit the resulting MPLS packet to the next hop of the VPN-IP route. The next hop of the labeled VPN-IP route will be the red loopback address C-PE2. So C-PE1 looks for the route to that red loopback address. This will be the red loopback route (i.e., the red IP route, see Section 3.3) originated by C-PE2.

C-PE2's red loopback will then be resolved through C-PE2's red loopback route. By virtue of the Tunnel Encapsulation attribute

carried by the latter route, C-PE1 will realize that to send packet P, it must set up an MPLS-in-IPsec SA (see Section 5, as well as Sections 3 and 8.1 of [RFC4023]) with C-PE2. Per the route's Tunnel Encapsulation attribute, the remote endpoint of this IPsec SA will be C-PE2's black loopback, and the Security Handle in the Tunnel Encapsulation attribute will carry any other information needed to set up the Security Association.

Note that the remote endpoint of the IPsec SA is determined by the Tunnel Encapsulation attribute of the red loopback route, rather than by the next hop field of that route. This ensures that the SA is made to the proper endpoint, even if the next hop field of the red loopback route was modified while the route was propagated.

**IMPORTANT:** The next hop of a VPN-IP route **MUST NOT** be resolved through an IP route that was not received over a red BGP session.

If a VPN-IP route's next hop resolves to a route that was not received over a red BGP session, the existence of the latter route **MUST** be regarded as the result of an attempt to spoof the location of the egress C-PE. That is, the latter route **MUST** be considered to be a spoofed route. The next hop of a VPN-IP route should always be a red loopback. However, since the full set of red loopbacks is not necessarily known in advance, it may not be possible to detect this spoofing attack until the attempt is made to resolve the VPN-IP route's next hop. Implementors should take special care to ensure that their implementations are not vulnerable to this sort of spoofing attack. Implementors should also take care to consider various corner cases, such as:

- o There is a black route to the next hop of a VPN-IP route, but no red route to that next hop. In this case the next hop **MUST** be considered to be unreachable.
- o There is both a black route and a red route to the next hop of a VPN-IP route. In this case, the red route **MUST** be preferred to the black route for the purpose of resolving the next hop.

When packet P is transmitted, it is transmitted through an MPLS-in-IPsec SA. Thus the only information that appears in the clear is the IP header needed to get the packet across the network. The IP source and destination addresses of that packet will be the black loopbacks of C-PE1 and C-PE2 respectively. The red loopback addresses do not appear in the packets at all, and no part of the payload packet (neither the VPN label nor the IP datagram following the VPN label) appears in the clear.

The MPLS-in-IPsec SA between C-PE1 and C-PE2 may be initiated by C-PE1 as soon as it receives a red loopback route originated by C-PE2. Alternatively, the initiation of the setup of the Security Association may be delayed until the SA is actually needed for transmitting packets.

These procedures will result in a single IPsec SA between a pair of C-PEs, with the data of multiple tenants carried on that single SA. If for some reason it is considered preferable to have an SA per tenant, the following procedures can be used:

- o On each C-PE, provision a distinct red loopback for each tenant.
- o Each C-PE will originate a red loopback route for each red loopback.
- o Each red loopback route will have its own Tunnel Encapsulation attribute. The respective Security Handle sub-TLVs (if present) MUST be distinct.

Note that this section is not intended to describe an implementation strategy.

## 5. MPLS-in-IPsec

Packets traveling from one C-PE to another travel through "MPLS-in-IPsec" tunnels. To transmit an MPLS packet through an MPLS-in-IPsec tunnel, one does the following:

- o Encapsulate the MPLS packet in IP, as specified in Section 3 of [RFC4023].
- o Use an IPsec transport mode Security Association to send the MPLS-in-IP packet from one C-PE to the other. This is specified in Section 8.1 of [RFC4023].

The result of encapsulating MPLS in IP and then transmitting the MPLS-in-IP packet on an IPsec transport mode Security Association is known as an MPLS-in-IPsec packet.

On the wire, an MPLS-in-IPsec packet consists of a cleartext IP header followed by a payload. The IP source and destination addresses of an MPLS-in-IPsec packet will be the black loopbacks of the source and destination C-PEs. The payload will be an MPLS packet. If the IPsec Security Association is providing privacy, authentication, and integrity, the payload is protected from inspection or alteration.



When the packet arrives at the destination C-PE, any necessary decryption is done, and packet appears to be an MPLS-in-IP packet addressed to the black address of the destination C-PE. The IP encapsulation is removed, yielding an MPLS packet. Per the usual L3VPN procedures, the label at the top of the MPLS label stack will be used to govern the further disposition of the packet. However, if a packet received over a black interface was not received through an IPsec SA, the packet MUST NOT be sent out any VRF interface.

MPLS-in-IP packets received in the clear (i.e., not received over an IPsec SA) MUST be discarded.

Note that this section is not intended to describe an implementation strategy.

## 6. Security Handle

This document defines a new BGP Tunnel Encapsulation attribute sub-TLV, the "Security Handle". This sub-TLV has a one-octet length field. It is intended for use in the Tunnel Encapsulation attribute carried by the red loopback routes. Its use is deployment specific.

As an example, in some deployments, this sub-TLV might be used to carry the IPsec Security Parameters Index (SPI). When setting up an SA to the originator of a particular Tunnel Encapsulation attribute, the SPI would be used as part of the SA setup procedure.

In deployments where the C-PEs auto-discover each other through RRs, and authenticate via certificate-based mechanisms, the Security Handle may not be needed at all. If a given deployment does not make use of the Security Handle, the sub-TLV SHOULD be omitted from the Tunnel Encapsulation attribute.

## 7. Data Plane Security Procedures

If a C-PE receives data over one of its local red interfaces, it may forward the data out another of its local red interfaces, as long as those two interfaces are associated with the same VRF, or if there is policy allowing communication ("extranet") between those two interfaces.

However, data received by a C-PE over one of its red interfaces MUST NOT be forwarded out a black interface, unless that data is being sent over the black interface through an IPsec SA.

Similarly, data received by a C-PE over one of its black interfaces MUST NOT be forwarded out a red interface unless the data arrived through an IPsec SA.

Typically an IPsec implementation has procedures to prevent unauthorized red-to-black or black-to-red forwarding. However, the conventional procedures are based on filtering of IP addresses, and hence do not apply directly if MPLS-in-IPsec is used. Implementors should take care to ensure that unauthorized red-to-black or black-to-red forwarding is prohibited.

Note that this rule has an important side-effect. A C-PE will not be able to forward packets received on a red interface to destinations that are outside the VPN, such as destinations on the public Internet. However, nothing prevents the customer from having an "Internet Gateway" at one or more sites, and attaching the Gateways to the C-PEs via black interfaces. If the routing at the customer site is such that intra-VPN traffic goes to the C-PE via a red interface, but traffic to the Internet goes via the Gateway, the C-PE can serve as an Internet access point without compromising the VPN's security (assuming of course that the Gateway provides the necessary security for traffic to/from the Internet).

## 8. Implementation Challenges

This document specifies an architecture for Secured L3VPNs, but a successful implementation faces a number of challenges.

This document specifies a route resolution process that makes use of the Tunnel Encapsulation attribute. This is a new feature.

This document specifies that during resolution of the next hop of a VPN-IP route, routes received over black BGP sessions must be disregarded. This is a new feature that may present challenges.

Ultimately, success will require a highly scalable IPsec implementation, that can set up SAs dynamically based on information disseminated by BGP. This presents a number of implementation challenges.

## 9. Security Considerations

Security considerations are discussed throughout this document.

As long as the C-PE devices and the Route Reflectors are physically secure, and not compromised, the techniques of this document provide privacy, integrity, and authentication for customer data and customer routing information.

The techniques of this document do not protect against attacks on the network backbone that make the black addresses unreachable, or that spoof the routes to the black addresses. Such attacks can disrupt or

disable the customer's ability to communicate over the unsecured network infrastructure. However, such attacks cannot expose the customer's routing or data.

Proper security depends on the correct implementation of such policies as "do not forward packets between red and black interfaces unless the packets are protected by IPsec on the black interfaces" and "do not resolve the next hop of a VPN-IP red route using a route that was not received over a red BGP session".

Attacks that are based on traffic analysis are not prevented by the techniques of this document.

## 10. IANA Considerations

IANA is requested to create a new entry in the "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry, "Security Handle". This sub-TLV is defined in Section 6 to have a one-octet length field. Thus it needs to be assigned a codepoint in the range 0-127 inclusive.

## 11. Acknowledgments

We wish to thank John Scudder for his ideas and contributions to this work.

The idea of integrating IPsec into L3VPN is not new to this document. This document has been influenced by earlier work such as [PE-PE\_IPsec], and we wish to thank the authors of the earlier work.

## 12. References

### 12.1. 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, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005, <<https://www.rfc-editor.org/info/rfc4023>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.

- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <<https://www.rfc-editor.org/info/rfc4303>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <<https://www.rfc-editor.org/info/rfc6286>>.
- [RFC7296] Kaufman, C., Hoffman, P., Nir, Y., Eronen, P., and T. Kivinen, "Internet Key Exchange Protocol Version 2 (IKEv2)", STD 79, RFC 7296, DOI 10.17487/RFC7296, October 2014, <<https://www.rfc-editor.org/info/rfc7296>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8221] Wouters, P., Migault, D., Mattsson, J., Nir, Y., and T. Kivinen, "Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 8221, DOI 10.17487/RFC8221, October 2017, <<https://www.rfc-editor.org/info/rfc8221>>.
- [TUNNEL\_ENCAPS]  
Rosen, E., Patel, K., and G. Van de Velde, "The BGP Tunnel Encapsulation Attribute VPN", internet-draft draft-ietf-idr-tunnel-encaps-09, February 2018.

## 12.2. Informational References

- [PE-PE\_IPsec]  
Rosen, E., De Clercq, J., Paridaens, O., T'Joens, Y., and C. Sargor, "Architecture for the Use of PE-PE IPsec Tunnels in BGP/MPLS IP VPNs", internet-draft draft-ietf-l3vpn-ipsec-2547-05, August 2005.
- [RFC6071] Frankel, S. and S. Krishnan, "IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap", RFC 6071, DOI 10.17487/RFC6071, February 2011, <<https://www.rfc-editor.org/info/rfc6071>>.

Authors' Addresses

Eric C. Rosen (editor)  
Juniper Networks, Inc.  
10 Technology Park Drive  
Westford, Massachusetts 01886  
United States

Email: [erosen@juniper.net](mailto:erosen@juniper.net)

Ron Bonica  
Juniper Networks, Inc.  
2251 Corporate Park Drive  
Herndon, Virginia 20171  
United States

Email: [rbonica@juniper.net](mailto:rbonica@juniper.net)

BESS Working Group  
Internet Draft  
Category: Standard Track

A. Sajassi  
S. Thoria  
Cisco  
A. Gupta  
Avi Networks

Expires: October 26, 2018

April 26, 2018

Seamless Multicast Interoperability between EVPN and MVPN PEs  
draft-sajassi-bess-evpn-mvpn-seamless-interop-01.txt

#### Abstract

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive for Network Virtualization Overlay (NVO) services in data center (DC) networks and as the next generation VPN services in service provider (SP) networks.

As service providers transform their networks in their COs toward next generation data center with Software Defined Networking (SDN) based fabric and Network Function Virtualization (NFV), they want to be able to maintain their offered services including multicast VPN (MVPN) service between their existing network and their new Service Provider Data Center (SPDC) network seamlessly without the use of gateway devices. They want to have such seamless interoperability between their new SPDCs and their existing networks for a) reducing cost, b) having optimum forwarding, and c) reducing provisioning. This document describes a unified solution based on RFCs 6513 & 6514 for seamless interoperability of multicast VPN between EVPN and MVPN PEs. Furthermore, it describes how the proposed solution can be used as a routed multicast solution in data centers with only EVPN PEs.

#### 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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2015 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. Introduction . . . . .	4
2. Requirements Language . . . . .	5
3. Terminology . . . . .	5
4. Requirements . . . . .	6
4.1. Optimum Forwarding . . . . .	6
4.2. Optimum Replication . . . . .	6
4.3. All-Active and Single-Active Multi-Homing . . . . .	7
4.4. Inter-AS Tree Stitching . . . . .	7
4.5. EVPN Service Interfaces . . . . .	7
4.6. Distributed Anycast Gateway . . . . .	7
4.7. Selective & Aggregate Selective Tunnels . . . . .	8
4.8. Tenants' (S,G) or (*,G) states . . . . .	8
4.9. Zero Disruption upon BD/Subnet Addition . . . . .	8
4.10. No Changes to Existing EVPN Service Interface Models . . . . .	8
5. IRB Unicast versus IRB Multicast . . . . .	8
5.1. Emulated Virtual LAN Service . . . . .	9
6. Solution Overview . . . . .	9
6.1. Operational Model for EVPN IRB PEs . . . . .	9
6.2. Unicast Route Advertisements for IP multicast Source . . . . .	12
6.3. Multi-homing of IP Multicast Source and Receivers . . . . .	13
6.3.1. Single-Active Multi-Homing . . . . .	13
6.3.2. All-Active Multi-Homing . . . . .	14
6.4. Mobility for Tenant's Sources and Receivers . . . . .	16

6.5. Intra-Subnet BUM Traffic Handling . . . . .	17
7. Control Plane Operation . . . . .	17
7.1. Intra-subnet/Intra-ES IP multicast tunnel . . . . .	17
7.2. Intra-subnet BUM tunnel . . . . .	18
7.3. Inter-subnet IP Multicast tunnel . . . . .	18
7.4. IGMP Hosts as TSes . . . . .	19
7.5. TS PIM Routers . . . . .	20
8 Data Plane Operation . . . . .	20
8.1 Intra-Subnet L2 Switching . . . . .	21
8.2 Inter-Subnet L3 Routing . . . . .	21
9. DCs with only EVPN PEs . . . . .	22
9.1. Setup of overlay multicast delivery . . . . .	22
9.2. Handling of different encapsulations . . . . .	24
9.2.1. MPLS Encapsulation . . . . .	24
9.2.2 VxLAN Encapsulation . . . . .	24
9.2.3. Other Encapsulation . . . . .	24
10. DCI with MPLS in WAN and VxLAN in DCs . . . . .	24
10.1. Control plane inter-connect . . . . .	25
10.2. Data plane inter-connect . . . . .	26
11. IANA Considerations . . . . .	27
12. Security Considerations . . . . .	27
13. Acknowledgements . . . . .	27
14. References . . . . .	27
14.1. Normative References . . . . .	27
14.2. Informative References . . . . .	27
15. Authors' Addresses . . . . .	28
Appendix A. Use Cases . . . . .	29
A.1. DCs with only IGMP/MLD hosts w/o tenant router . . . . .	29
A.2. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-SSM . . . . .	30
A.3. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-ASM . . . . .	30
A.4. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-Bidir . . . . .	30



## 1. Introduction

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive for Network Virtualization Overlay (NVO) services in data center (DC) networks and as the next generation VPN services in service provider (SP) networks.

As service providers transform their networks in their COs toward next generation data center with Software Defined Networking (SDN) based fabric and Network Function Virtualization (NFV), they want to be able to maintain their offered services including multicast VPN (MVPN) service between their existing network and their new SPDC network seamlessly without the use of gateway devices. There are several reasons for having such seamless interoperability between their new DCs and their existing networks:

- Lower Cost: gateway devices need to have very high scalability to handle VPN services for their DCs and as such need to handle large number of VPN instances (in tens or hundreds of thousands) and very large number of routes (e.g., in tens of millions). For the same speed and feed, these high scale gateway boxes are relatively much more expensive than the edge devices (e.g., PEs and TORs) that support much lower number of routes and VPN instances.
- Optimum Forwarding: in a given CO, both EVPN PEs and MVPN PEs can be connected to the same fabric/network (e.g., same IGP domain). In such scenarios, the service providers want to have optimum forwarding among these PE devices without the use of gateway devices. Because if gateway devices are used, then the IP multicast traffic between an EVPN and MVPN PEs can no longer be optimum and in some case, it may even get tromboned. Furthermore, when an SPDC network spans across multiple LATA (multiple geographic areas) and gateways are used between EVPN and MVPN PEs, then with respect to IP multicast traffic, only one GW can be designated forwarder (DF) between EVPN and MVPN PEs. Such scenarios not only results in non-optimum forwarding but also it can result in tromboing of IP multicast traffic between the two LATAs when both source and destination PEs are in the same LATA and the DF gateway is elected to be in a different LATA.
- Less Provisioning: If gateways are used, then the operator need to configure per-tenant info on the gateways. In other words, for each tenant that is configured, one (or maybe two) additional touch points are needed.

This document describes a unified solution based on [RFC6513] and [RFC6514] for seamless interoperability of multicast VPN between EVPN and MVPN PEs. Furthermore, it describes how the proposed solution can be used as a routed multicast solution in data centers with only EVPN

PEs (e.g., routed multicast VPN only among EVPN PEs).

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without any normative meaning.

## 3. Terminology

Most of the terminology used in this documents comes from [RFC8365]

**Broadcast Domain:** In a bridged network, the broadcast domain corresponds to a Virtual LAN (VLAN), where a VLAN is typically represented by a single VLAN ID (VID) but can be represented by several VIDs where Shared VLAN Learning (SVL) is used per [802.1Q].

**Bridge Table:** An instantiation of a broadcast domain on a MAC-VRF.

**VXLAN:** Virtual Extensible LAN

**POD:** Point of Delivery

**NV:** Network Virtualization

**NVO:** Network Virtualization Overlay

**NVE:** Network Virtualization Endpoint

**VNI:** Virtual Network Identifier (for VXLAN)

**EVPN:** Ethernet VPN

**EVI:** An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN

**MAC-VRF:** A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE

**IP-VRF:** A Virtual Routing and Forwarding table for Internet Protocol (IP) addresses on a PE

**Ethernet Segment (ES):** When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that

set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

Ethernet Tag: An Ethernet tag identifies a particular broadcast domain, e.g., a VLAN. An EVPN instance consists of one or more broadcast domains.

PE: Provider Edge device.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

PIM-SM: Protocol Independent Multicast - Sparse-Mode

PIM-SSM: Protocol Independent Multicast - Source Specific Multicast

Bidir PIM: Bidirectional PIM

#### 4. Requirements

This section describes the requirements specific in providing seamless multicast VPN service between MVPN and EVPN capable networks.

##### 4.1. Optimum Forwarding

The solution SHALL support optimum multicast forwarding between EVPN and MVPN PEs within a network. The network can be confined to a CO or it can span across multiple LATAs. The solution SHALL support optimum multicast forwarding with both ingress replication tunnels and P2MP tunnels.

##### 4.2. Optimum Replication

For EVPN PEs with IRB capability, the solution SHALL use only a

single multicast tunnel among EVPN and MVPN PEs for IP multicast traffic. Multicast tunnels can be either ingress replication tunnels or P2MP tunnels. The solution MUST support optimum replication for both Intra-subnet and Inter-subnet IP multicast traffic:

- Non-IP traffic SHALL be forwarded per EVPN baseline [RFC7432] or [RFC8365]
- If a Multicast VPN spans across both Intra and Inter subnets, then for Ingress replication regardless of whether the traffic is Intra or Inter subnet, only a single copy of IP multicast traffic SHALL be sent from the source PE to the destination PE.
- If a Multicast VPN spans across both Intra and Inter subnets, then for P2MP tunnels regardless of whether the traffic is Intra or Inter subnet, only a single copy of multicast data SHALL be transmitted by the source PE. Source PE can be either EVPN or MVPN PE and receiving PEs can be a mix of EVPN and MVPN PEs - i.e., a multicast VPN can be spread across both EVPN and MVPN PEs.

#### 4.3. All-Active and Single-Active Multi-Homing

The solution MUST support multi-homing of source devices and receivers that are sitting in the same subnet (e.g., VLAN) and are multi-homed to EVPN PEs. The solution SHALL allow for both Single-Active and All-Active multi-homing. The solution MUST prevent loop during steady and transient states just like EVPN baseline solution [RFC7432] and [RFC8365] for all multi-homing types.

#### 4.4. Inter-AS Tree Stitching

The solution SHALL support multicast tree stitching when the tree spans across multiple Autonomous Systems.

#### 4.5. EVPN Service Interfaces

The solution MUST support all EVPN service interfaces listed in section 6 of [RFC7432]:

- VLAN-based service interface
- VLAN-bundle service interface
- VLAN-aware bundle service interface

#### 4.6. Distributed Anycast Gateway

The solution SHALL support distributed anycast gateways for tenant workloads on NVE devices operating in EVPN-IRB mode.

#### 4.7. Selective & Aggregate Selective Tunnels

The solution SHALL support selective and aggregate selective P-tunnels as well as inclusive and aggregate inclusive P-tunnels. When selective tunnels are used, then multicast traffic SHOULD only be forwarded to the remote PE which have receivers - i.e., if there are no receivers at a remote PE, the multicast traffic SHOULD NOT be forwarded to that PE and if there are no receivers on any remote PEs, then the multicast traffic SHOULD NOT be forwarded to the core.

#### 4.8. Tenants' (S,G) or (\*,G) states

The solution SHOULD store (C-S,C-G) and (C-\*,C-G) states only on PE devices that have interest in such states hence reducing memory and processing requirements - i.e., PE devices that have sources and/or receivers interested in such multicast groups.

#### 4.9. Zero Disruption upon BD/Subnet Addition

In DC environments, various Bridge Domains are provisioned and removed on regular basis due to host mobility, policy and tenant changes. Such change in BD configuration should not affect existing flows within the same BD or any other BD in the network.

#### 4.10. No Changes to Existing EVPN Service Interface Models

VLAN-aware bundle service as defined in [RFC7432] typically does not require any VLAN ID translation from one tenant site to another - i.e., the same set of VLAN IDs are configured consistently on all tenant segments. In such scenarios, EVPN-IRB multicast service MUST maintain the same mode of operation and SHALL NOT require any VLAN ID translation.

### 5. IRB Unicast versus IRB Multicast

[EVPN-IRB] describes the operation for EVPN PEs in IRB mode for unicast traffic. The same IRB model for a PE described in [EVPN-IRB], where an IP-VRF is attached to one or more bridge tables (BTs) via virtual IRB interfaces, is also applicable here. However, there are some noticeable differences between the IRB operation for unicast traffic described in [EVPN-IRB] versus for multicast traffic described in this document. For unicast traffic, the intra-subnet traffic, is bridged within the MAC-VRF associated with that subnet (i.e., a lookup based on MAC-DA is performed); whereas, the inter-subnet traffic is routed in the corresponding IP-VRF (ie, a lookup based on IP-DA is performed). A given tenant can have one or more IP-VRFs; however, without loss of generality, this document assumes one

IP-VRF per tenant. In context of a given tenant's multicast traffic, the intra-subnet traffic is bridged for non-IP traffic and it is Layer-2 switched for IP traffic. Whereas, the tenants's inter-subnet multicast traffic is always routed in the corresponding IP-VRF. The difference between bridging and L2-switching for multicast traffic is that the former uses MAC-DA lookup for forwarding the multicast traffic; whereas, the latter uses IP-DA lookup for such forwarding where the forwarding states are built in the MAC-VRF using IGMP/MLD or PIM snooping.

#### 5.1. Emulated Virtual LAN Service

EVPN does not provide a Virtual LAN (VLAN) service per [IEEE802.1Q] but rather an emulated VLAN service. This VLAN service emulation is not only done for unicast traffic but also is extended for intra-subnet multicast traffic described in [EVPN-IGMP-PROXY] and [EVPN-PIM-PROXY]. For intra-subnet multicast, an EVPN PE builds multicast forwarding states in its bridge table (BT) based on snooping of IGMP/MLD and/or PIM messages and the forwarding is performed based on destination IP multicast address of the Ethernet frame rather than destination MAC address as noted above. In order to enable seamless integration of EVPN and MVPN PEs, this document extends the concept of an emulated VLAN service for multicast IRB applications such that the intra-subnet IP multicast traffic can get treated same as inter-subnet IP multicast traffic which means intra-subnet IP multicast traffic can get routed instead of being L2-switched - i.e., TTL value gets decremented and the Ethernet header of the L2 frame is de-capsulated and encapsulated at both ingress and egress PEs. It should be noted that the non-IP multicast or broadcast traffic still gets bridged and frames get forwarded based on their destination MAC addresses.

### 6. Solution Overview

This section describes a multicast VPN solution based on [RFC6513] and [RFC6514] for EVPN PEs operating in IRB mode that want to perform seamless interoperability with their counterparts MVPN PEs.

#### 6.1. Operational Model for EVPN IRB PEs

Without the loss of generality, this section assumes that all EVPN PEs have IRB capability and operating in IRB mode for both unicast and multicast traffic (e.g., all EVPN PEs are homogenous in terms of their capabilities and operational modes). As it will be seen later, an EVPN network can consist of a mix of PEs where some are capable of multicast IRB and some are not and the multicast operation of such heterogeneous EVPN network will be an extension of an EVPN homogenous

network. Therefore, we start with the multicast IRB solution description for the EVPN homogenous network.

The EVPN PEs terminate IGMP/MLD messages from tenant host devices or PIM messages from tenant routers on their IRB interfaces, thus avoid sending these messages over MPLS/IP core. A tenant virtual/physical router (e.g., CE) attached to an EVPN PE becomes a multicast routing adjacency of that PE. Furthermore, the PE uses MVPN BGP protocol and procedures per [RFC6513] and [RFC6514]. With respect to multicast routing protocol between tenant's virtual/physical router and the PE that it is attached to, any of the following PIM protocols is supported per [RFC6513]: PIM-SM with Any Source Multicast (ASM) mode, PIM-SM with Source Specific Multicast (SSM) mode, and PIM Bidirectional (BIDIR) mode. Support of PIM-DM (Dense Mode) is excluded in this document per [RFC6513].

The EVPN PEs use MVPN BGP routes defined in [RFC6514] to convey tenant (S,G) or (\*,G) states to other MVPN or EVPN PEs and to set up overlay trees (inclusive or selective) for a given MVPN instance. The root or a leaf of such an overlay tree is terminated on an EVPN or MVPN PE. Furthermore, this inclusive or selective overlay tree is terminated on a single IP-VRF of the EVPN or MVPN PE. In case of EVPN PE, these overlay trees never get terminated on MAC-VRFs of that PE. Overlay trees are instantiated by underlay provider tunnels (P-tunnels) - e.g., P2MP, MP2MP, or unicast tunnels per [RFC 6513]. When there are several overlay trees mapped to a single underlay P-tunnel, the tunnel is referred to as an aggregate tunnel.

Figure-1 below depicts a scenario where a tenant's MVPN spans across both EVPN and MVPN PEs; where all EVPN PEs have multicast IRB capability. An EVPN PE (with multicast IRB capability) can be modeled as a MVPN PE where the virtual IRB interface of an EVPN PE (virtual interface between a BT and IP-VRF) can be considered a routed interface for the MVPN PE.

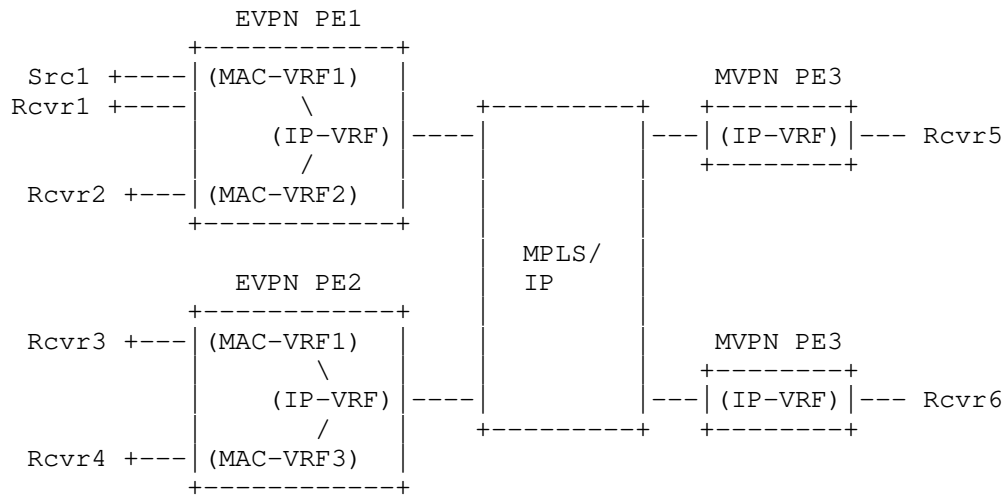


Figure-1: EVPN &amp; MVPN PEs Seamless Interop

Figure 2 depicts the modeling of EVPN PEs based on MVPN PEs where an EVPN PE can be modeled as a PE that consists of a MVPN PE whose routed interfaces (e.g., attachment circuits) are replaced with IRB interfaces connecting each IP-VRF of the MVPN PE to a set of BTs. Similar to a MVPN PE where an attachment circuit serves as a routed multicast interface for an IP-VRF associated with a MVPN instance, an IRB interface serves as a routed multicast interface for the IP-VRF associated with the MVPN instance. Since EVPN PEs run MVPN protocols (e.g., [RFC6513] and [RFC6514]), for all practical purposes, they look just like MVPN PEs to other PE devices. Such modeling of EVPN PEs, transforms the multicast VPN operation of EVPN PEs to that of MVPN and thus simplifies the interoperability between EVPN and MVPN PEs to that of running a single unified solution based on MVPN.



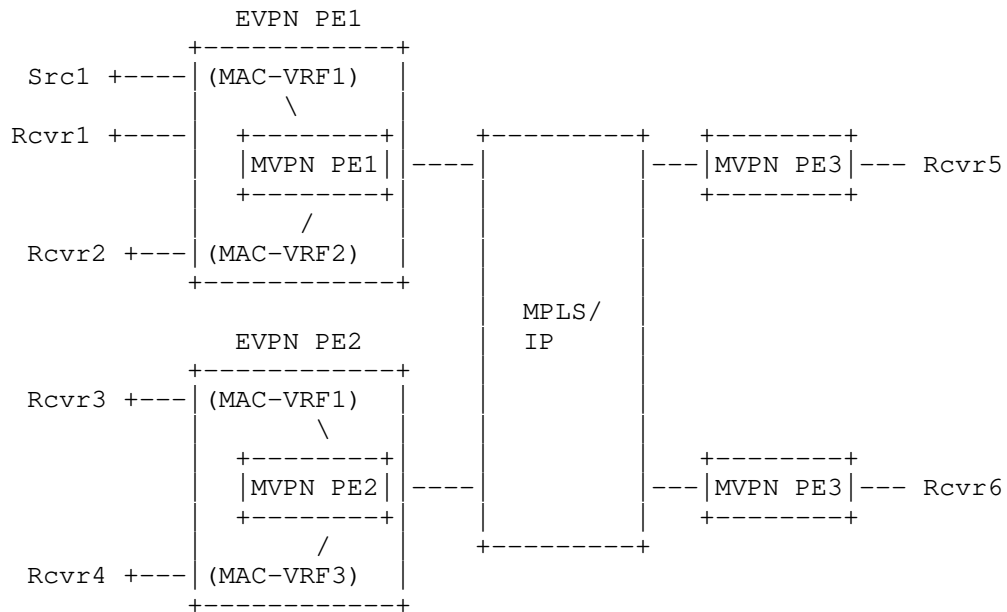


Figure-2: Modeling EVPN PEs as MVPN PEs

Although modeling an EVPN PE as a MVPN PE, conceptually simplifies the operation to that of a solution based on MVPN, the following operational aspects of EVPN need to be factored in when considering seamless integration between EVPN and MVPN PEs.

- 1) Unicast route advertisements for IP multicast source
- 2) Multi-homing of IP multicast sources and receivers
- 3) Mobility for Tenant's sources and receivers
- 4) non-IP multicast traffic handling

## 6.2. Unicast Route Advertisements for IP multicast Source

When an IP multicast source is attached to an EVPN PE, the unicast route for that IP multicast source needs to be advertised. When the source is attached to a Single-Active multi-homed ES, then the EVPN DF PE is the PE that advertises a unicast route corresponding to the source IP address with VRF Route Import extended community which in turn is used as the Route Target for Join (S,G) messages sent toward the source PE by the remote PEs. The EVPN PE advertises this unicast route using EVPN route type 2 (or 5) and IPVPN unicast route along with VRF Route Import extended community. EVPN route type 2 (or 5) is advertised with the Route Targets corresponding to both IP-VRF and

MAC-VRF/BT; whereas, IPVPN unicast route is advertised with RT corresponding to the IP-VRF. When unicast routes are advertised by MVPN PEs, they are advertised using IPVPN unicast route along with VRF Route Import extended community per [RFC6514].

When the source is attached to an All-Active multi-homed ES, then the PE that learns the source advertises the unicast route for that source using EVPN route type 2 (or 5) and IPVPN unicast route along with VRF Route Import extended community. EVPN route type 2 (or 5) is advertised with the Route Targets corresponding to both IP-VRF and MAC-VRF/BT; whereas, IPVPN unicast route is advertised with RT corresponding to the IP-VRF. When the other multi-homing EVPN PEs for that ES receive this unicast EVPN route, they import the route and check to see if they have learned the route locally for that ES, if they have, then they do nothing. But if they have not, then they add the IP and MAC addresses to their IP-VRF and MAC-VRF/BT tables respectively with the local interface corresponding to that ES as the corresponding route adjacency. Furthermore, these PEs advertise an IPVPN unicast route along with VRF Route Import extended community and Route Target corresponding to IP-VRF to other remote PEs for that MVPN. Therefore, the remote PEs learn the unicast route corresponding to the source from all multi-homing PEs associated with that All-Active Ethernet Segment even though one of the multi-homing PEs may only have directly learned the IP address of the source.

### 6.3. Multi-homing of IP Multicast Source and Receivers

EVPN [RFC7432] has extensive multi-homing capabilities that allows TSes to be multi-homed to two or more EVPN PEs in Single-Active or All-Active mode. In Single-Active mode, only one of the multi-homing EVPN PEs can receive/transmit traffic for a given subnet (a given BD) for that multi-homed Ethernet Segment (ES). In All-Active mode, any of the multi-homing EVPN PEs can receive/transmit unicast traffic but only one of them (the DF PE) can send BUM traffic to the multi-homed ES for a given subnet.

The multi-homing mode (Single-Active versus All-Active) of a TS source can impact the MVPN procedures as described below.

#### 6.3.1. Single-Active Multi-Homing

When a TS source reside on an ES that is multi-homed to two or more EVPN PEs operating in Single-Active mode, only one of the EVPN PEs can be active for the source subnet on that ES. Therefore, only one of the multi-homing PE learns the unicast route of the TS source and advertises that using EVPN and IPVPN to other PEs as described previously.

A downstream PE that receives a Join/Prune message from a TS host/router, selects a Upstream Multicast Hop (UMH) which is the upstream PE that receives the IP multicast flow in case of Single-Active multi-homing. An IP multicast flow belongs to either a source-specific tree (S,G) or to a shared tree (\*,G). We use the notation (X,G) to refer to either (S,G) or (\*,G); where X refers to S in case of (S,G) and X refers to the Rendezvous Point (RP) for G in case of (\*,G). Since the active PE (which is also the UMH PE) has advertised unicast route for X along with the VRF Route Import EC, the downstream PEs select the UMH without any ambiguity based on MVPN procedures described in section 5.1 of [RFC6513]. Any of the three algorithms described in that section works fine.

The multi-homing PE that receives the IP multicast flow on its local AC, performs the following tasks:

- L2 switches the multicast traffic in its BT associated with the local AC over which it received the flow if there are any interested receivers for that subnet.
- L3 routes the multicast traffic to other BTs for other subnets if there are any interested receivers for those subnets.
- L3 routes the multicast traffic to other PEs per MVPN procedures.

The multicast traffic can be sent on Inclusive, Selective, or Aggregate-Selective tree. Regardless what type of tree is used, only a single copy of the multicast traffic is received by the downstream PE.

### 6.3.2. All-Active Multi-Homing

When a TS source resides on an ES that is multi-homed to two or more EVPN PEs operating in All-Active mode, then any of the multi-homing PEs can learn the TS source's unicast route; however, that PE may not be the same PE that receives the IP multicast flow. Therefore, the procedures for Single-Active Multi-homing need to be augmented for All-Active scenario as below.

The multi-homing EVPN PE that receives the IP multicast flow on its local AC, needs to do the following task in addition to the ones listed in the previous section for Single-Active multi-homing: L2 switch the multicast traffic to other multi-homing EVPN PEs for that ES via an intra-subnet overlay tunnel. There will be a dedicated intra-subnet tunnel for this purpose which is different from inter-subnet overlay tunnel setup by MVPN procedures.

When the multi-homing EVPN PEs receive the IP multicast flow via this

intra-subnet tunnel, they treat it as if they receive the flow via their local ACs and thus perform the tasks mentioned in the previous section for Single-Active multi-homing. The tunnel type for this intra-subnet tunnel can be any of the supported tunnel types such as ingress-replication, P2MP tunnel, BIER, and Assisted Replication; however, given that vast majority of multi-homing ESes are just dual-homing, a simple ingress replication tunnel will serve well. For a given ES, since multicast traffic that is locally received by one multi-homing PE is sent to other multi-homing PEs via this intra-subnet tunnel, there is no need for sending the multicast tunnel via MVPN tunnel to these multi-homing PEs - i.e., MVPN multicast tunnels are used only for remote EVPN and MVPN PEs. Multicast traffic sent over this intra-subnet tunnel to other multi-homing PEs (only one other in case of dual-homing) for a given ES, is sent regardless of whether there is a receiver on these multi-homing PEs.

By feeding IP multicast flow received on one of the EVPN multi-homing PEs to the rest of the EVPN PEs in the multi-homing group, we have essentially enabled all the PEs in the multi-homing group to serve as UMH for that IP multicast flow. Each of these UMH PEs advertises unicast route for X in (X,G) along with the VRF Route Import EC to all PEs for that MVPN instance. The downstream PEs build a candidate UMH set based on procedures described in section 5.1 of [RFC6513] and pick a UMH from the set. It should be noted that both the default UMH selection procedure based on highest UMH PE IP address and the UMH selection algorithm based on hash function specified in section 5.1.3 of [RFC6513] (which is also a MUST implement algorithm) result in the same UMH PE be selected by all downstream PEs running the same algorithm. However, in order to allow a form of "equal cost load balancing", the hash algorithm is recommended to be used among all EVPN and MVPN PEs. This hash algorithm distributes UMH selection for different IP multicast flows among the multi-homing PEs for a given ES.

Since all downstream PEs (EVPN and MVPN) use the same hash-based algorithm for UMH determination, they all choose the same upstream PE as their UMH for a given (X,G) flow and thus they all send their (X,G) join message via BGP to the same upstream PE. This results in one of the multi-homing PEs to receive the join message and thus send the IP multicast flow for (X,G) over its associated overlay tree even though all of the multi-homing PEs in the All-Active redundancy group have received the IP multicast flow (one of them directly via its local AC and the rest indirectly via the associated intra-subnet tunnel). Therefore, only a single copy of routed IP multicast flow is sent over the network regardless of overlay tree type supported by the PEs - i.e., the overlay tree type can selective or aggregate selective or inclusive tree. This gives the network operator the maximum flexibility of choosing any overlay tree type that is

suitable for its network operation and still be able to deliver only a single copy of the IP multicast flows to the egress PEs. In other words, an egress PE only receives a single copy of the IP multicast flow over the network, because it either receives it via the EVPN intra-subnet tunnel or MVPN inter-subnet tunnel. Furthermore, if it receives it via MVPN inter-subnet tunnel, then only one of the multi-homing PEs associated with the source ES, sends the IP multicast traffic.

Since the network of interest for seamless interoperability between EVPN and MVPN PEs is MPLS, the EVPN handling of BUM traffic for MPLS network needs to be considered. EVPN [RFC7432] uses ESI MPLS label for split-horizon filtering of Broadcast/Unknown unicast/multicast (BUM) traffic from an All-Active multi-homing Ethernet Segment to ensure that BUM traffic doesn't get loop back to the same Ethernet Segment that it came from. This split-horizon filtering mechanism applies as-is for multicast IRB scenario because of using the intra-subnet tunnel among multi-homing PEs. Since the multicast traffic received from a TS source on an All-Active ES by a multi-homing PE is bridged to all other multi-homing PEs in that group, the standard EVPN split-horizon filtering described in [RFC7432] applies as-is. Split-horizon filtering for non-MPLS encapsulations such as VxLAN is described in section 9.2.2 that deals with a DC network that consists of only EVPN PEs.

#### 6.4. Mobility for Tenant's Sources and Receivers

When a tenant system (TS), source or receiver, is multi-homed behind a group of multi-homing EVPN PEs, then TS mobility SHALL be supported among EVPN PEs. Furthermore, such TS mobility SHALL only cause an temporary disruption to the related multicast service among EVPN and MVPN PEs. If a source is moved from one EVPN PE to another one, then the EVPN mobility procedure SHALL discover this move and a new unicast route advertisement (using both EVPN and IP-VPN routes) is made by the EVPN PE where the source has moved to per section 6.3 above and unicast route withdraw (for both EVPN and IP-VPN routes) is performed by the EVPN PE where the source has moved from.

The move of a source results in disruption of the IP multicast flow for the corresponding (S,G) flow till the new unicast route associated with the source is advertised by the new PE along with the VRF Route Import EC, the join messages sent by the egress PEs are received by the new PE, the multicast state for that flow is installed in the new PE and a new overlay tree is built for that source from the new PE to the egress PEs that are interested in receiving that IP multicast flow.

The move of a receiver results in disruption of the IP multicast flow

to that receiver only till the new PE for that receiver discovers the source and joins the overlay tree for that flow.

#### 6.5. Intra-Subnet BUM Traffic Handling

Link local IP multicast traffic consists IPv4 traffic with a destination address prefix of 224/8 and IPv6 traffic with a destination address prefix of FF02/16. Such IP multicast traffic as well as non-IP multicast/broadcast traffic are sent per EVPN [RF7432] BUM procedures and does not get routed via IP-VRF for multicast addresses. So, such BUM traffic will be limited to a given EVI/VLAN (e.g., a give subnet); whereas, IP multicast traffic, will be locally switched for local interfaces attached on the same subnet and will be routed for local interfaces attached on a different subnet or for forwarding traffic to other EVPN PEs (refer to section 5.1.1 for data plane operation).

### 7. Control Plane Operation

In seamless interop between EVPN and MVPN PEs, the control plane may need to setup the following three types of multicast tunnels. The first two are among EVPN PEs only but the third one is among EVPN and MVPN PEs.

- 1) Intra-subnet/Intra-ES IP multicast tunnel
- 2) Intra-subnet BUM tunnel
- 3) Inter-subnet IP multicast tunnel

#### 7.1. Intra-subnet/Intra-ES IP multicast tunnel

As described in section 6.3.2, when a multicast source is sitting behind an All-Active ES, then an intra-subnet multicast tunnel is needed among EVPN PEs for that ES to carry multicast flow received by one of the multi-homing PEs to the other PEs in that ES. Vast majority of All-Active multi-homing for TOR devices in DC networks are just dual-homing which means the multicast flow received by one of the dual-homing PE only needs to be sent to the other dual-homing PE. Therefore, a simple ingress replication tunnel is all that is needed. In case of multi-homing to three or more EVPN PEs, then other tunnel types such as P2MP, MP2MP, BIER, and Assisted Replication can be considered. It should be noted that this intra-subnet/intra-ES tunnel is only needed for All-Active multi-homing and it is not required for Single-Active multi-homing.

The EVPN PEs belonging to a given All-Active ES discover each other using EVPN Ethernet Segment route per procedures described in [RFC7432]. These EVPN PEs perform DF election per [RFC7432], [EVPN-DF-Framework], or other DF election algorithms to decide who is a DF for a given BD. If the BD belongs to a tenant that has IRB multicast enabled for it, then each PE sets up an intra-subnet/intra-ES tunnel to forward IP multicast traffic received locally on that BD to other PE(s) for that ES. Therefore, IP multicast traffic received via a local attachment circuit is sent on this tunnel and on the associated IRB interface for that BT and other local attachment circuits if there are interested receivers for them. The other multi-homing EVPN PEs treat this intra-subnet/intra-ES tunnel just like their local ACs - i.e., the multicast traffic received over this tunnel is treated as if it is received via its local AC. Thus, the multi-homing PEs cannot receive the same IP multicast flow from an MVPN tunnel (e.g., over an IRB interface for that BD) because between a source behind a local AC versus a source behind a remote PE, the PE always chooses its local AC.

When ingress replication is used for intra-subnet/intra-ES tunnel, every PE in the All-Active multi-homing ES has all the information to setup these tunnels - i.e., a) each PE knows what are the other multi-homing PEs for that ES via EVPN Ethernet Segment route and b) each PE already knows what MPLS label to use for multicast traffic to every other PE for that ES via EVPN IMET route. Both EVPN ES and IMET routes are composed and advertised per [RFC7432].

## 7.2. Intra-subnet BUM tunnel

As the name implies, this tunnel is setup to carry BUM traffic for a given subnet/BD among EVNP PEs. In [RFC7432], this overlay tunnel is used for transmission of all BUM traffic including user IP multicast traffic. However, for multicast traffic handling in EVPN-IRB PEs, this tunnel is used for all broadcast, unknown-unicast, non-IP multicast traffic, and link-local IP multicast traffic - i.e., it is used for all BUM traffic except user IP multicast traffic. This tunnel is setup using IMET route for a given EVI/BD. The composition and advertisement of IMET routes are exactly per [RFC7432]. It should be noted that when an EVPN All-Active multi-homing PE uses both this tunnel as well as intra-subnet/intra-ES tunnel, there SHALL be no duplication of multicast traffic over the network because they carry different types of multicast traffic - i.e., intra-subnet/intra-ES tunnel carries only user IP multicast traffic; whereas, intra-subnet tunnel carries link-local IP multicast traffic and BUM traffic (w/ non-IP multicast).

## 7.3. Inter-subnet IP Multicast tunnel

As its name implies, this tunnel is setup to carry IP-only multicast traffic for a given tenant across all its subnets (BDs) among EVPN and MVPN PEs.

The following NLRIs from [RFC6514] is used for setting up this inter-subnet tunnel in the network.

Intra-AS I-PMSI A-D route is used to form default underlay tunnel (also called inclusive tunnel) for a tenant IP-VRF. The tunnel attributes are indicated using PMSI attribute with this route.

S-PMSI A-D route is used to form Customer flow specific underlay tunnels. This enables selective delivery of data to PEs having active receivers and optimizes fabric bandwidth utilization. The tunnel attributes are indicated using PMSI attribute with this route.

Each EVPN PE supporting a specific MVPN instance discovers the set of other PEs in its AS that are attached to sites of that MVPN using Intra-AS I-PMSI A-D route (route type 1) per [RFC6514]. It can also discover the set of other ASes that have PEs attached to sites of that MVPN using Inter-AS I-PMSI A-D route (route type 2) per [RFC6514]. After the discovery of PEs that are attached to sites of the MVPN, an inclusive overlay tree (I-PMSI) can be setup for carrying tenant multicast flows for that MVPN; however, this is not a requirement per [RFC6514] and it is possible to adopt a policy in which all tenant flows are carried on S-PMSIs.

An EVPN-IRB PE sends a user IP multicast flow to other EVPN and MVPN PEs over this inter-subnet tunnel that is instantiated using MVPN I-PMSI or S-PMSI. This tunnel can be considered as being originated and terminated from/to among IP-VRFs of EVPN/MVPN PEs; whereas, intra-subnet tunnel is originated/terminated among MAC-VRFs of EVPN PEs.

#### 7.4. IGMP Hosts as TSes

If a tenant system which is an IGMP host is multi-homed to two or more EVPN PEs using All-Active multi-homing, then IGMP join and leave messages are synchronized between these EVPN PEs using EVPN IGMP Join Synch route (route type 7) and EVPN IGMP Leave Synch route (route type 8) per [IGMP-PROXY]. IGMP states are built in the corresponding BDs of the multi-homing EVPN PEs. In [IGMP-PROXY] the DF PE for that BD originates an EVPN Selective Multicast Tag route (SMET route) route to other EVPN PEs. However, in here there is no need to use SMET because the IGMP messages are terminated by the EVPN-IRB PE and



tenant (\*,G) or (S,G) join messages are sent via MVPN Shared Tree Join route (route type 6) or Source Tree Join route (route type 7) respectively of MCAST-VPN NLRI per [RFC6514]. In case of a network with only IGMP hosts, the preferred mode of operation is that of SPT-only per section 14 of [RFC6514]. This mode is only supported for PIM-SM and avoids the RP configuration overhead. Such mode is chosen by provisioning/ configuration.

#### 7.5. TS PIM Routers

Just like a MVPN PE, an EVPN PE runs a separate tenant multicast routing instance (VPN-specific) per MVPN instance and the following tenant multicast routing instances are supported:

- PIM Sparse Mode (PIM-SM) with the ASM service model
- PIM Sparse Mode with the SSM service model
- PIM Bidirectional Mode (BIDIR-PIM), which uses bidirectional tenant-trees to support the ASM service model

A given tenant's PIM join messages for (\*,G) or (S, G) are processed by the corresponding tenant multicast routing protocol and they are advertised over MPLS/IP network using Shared Tree Join route (route type 6) and Source Tree Join route (route type 7) respectively of MCAST-VPN NLRI per [RFC6514].

### 8 Data Plane Operation

When an EVPN-IRB PE receives an IGMP/MLD join message over one of its Attachment Circuits (ACs), it adds that AC to its Layer-2 (L2) OIF list. This L2 OIF list is associated with the MAC-VRF/BT corresponding to the subnet of the tenant device that sent the IGMP/MLD join. Therefore, tenant (S,G) or (\*,G) forwarding entries are created/updated for the corresponding MAC-VRF/BT based on these source and group IP addresses. Furthermore, the IGMP/MLD join message is propagated over the corresponding IRB interface and it is processed by the tenant multicast routing instance which creates the corresponding tenant (S,G) or (\*,G) Layer-3 (L3) forwarding entries. It adds this IRB interface to the L3 OIF list. An IRB is removed as a L3 OIF when all L2 tenant (S,G) or (\*,G) forwarding states is removed for the MAC-VRF/BT associated with that IRB. Furthermore, tenant (S,G) or (\*,G) L3 forwarding state is removed when all of its L3 OIFs are removed - i.e., all the IRB and L3 interfaces associated with that tenant (S,G) or (\*,G) are removed.

When an EVPN PE receives IP multicast traffic from one of its AC, if it has any attached receivers for that subnet, it performs L2 switching of the intra-subnet traffic within the BT attached to that

AC. If the multicast flow is received over an AC that belongs to an All-Active ES, then the multicast flow is also sent over the intra-subnet/intra-ES tunnel. The EVPN PE then sends the multicast traffic over the corresponding IRB interface. The multicast traffic then gets routed in the corresponding IP-VRF and it gets forwarded to interfaces in the L3 OIF list which can include other IRB interfaces, other L3 interfaces directly connected to TSes, and the MVPN inter-subnet tunnel which is instantiated by an I-PMSI or S-PMSI tunnel. When the multicast packet is routed within the IP-VRF of the EVPN PE, its Ethernet header is stripped and its TTL gets decremented as the result of this IP routing. When the multicast traffic is received on an IRB interface by the BT corresponding to that interface, it gets L2 switched and sent over ACs that belong to the L2 OIF list.

### 8.1 Intra-Subnet L2 Switching

Rcvr1 in Figure 1 is connected to PE1 in MAC-VRF1 (same as Src1) and sends IGMP join for (C-S, C-G), IGMP snooping will record this state in local bridging entry. A routing entry will be formed as well which will point to MAC-VRF1 as RPF for Src1. We assume that Src1 is known via ARP or similar procedures. Rcvr1 will get a locally bridged copy of multicast traffic from Src1. Rcvr3 is also connected in MAC-VRF1 but to PE2 and hence would send IGMP join which will be recorded at PE2. PE2 will also form routing entry and RPF will be assumed as Tenant Tunnel "Tenant1" formed beforehand using MVPN procedures. Also this would cause multicast control plane to initiate a BGP MCAST-VPN type 7 route which would include VRI for PE1 and hence be accepted on PE1. PE1 will include Tenant1 tunnel as Outgoing Interface (OIF) in the routing entry. Now, since it has knowledge of remote receivers via MVPN control plane it will encapsulate original multicast traffic in Tenant1 tunnel towards core.

### 8.2 Inter-Subnet L3 Routing

Rcvr2 in Figure 1 is connected to PE1 in MAC-VRF2 and hence PE1 will record its membership in MAC-VRF2. Since MAC-VRF2 is enabled with IRB, it gets added as another OIF to routing entry formed for (C-S, C-G). Rcvr2 and Rcvr4 are also in different MAC-VRFs than multicast speaker Src1 and hence need Inter-subnet forwarding. PE2 will form local bridging entry in MAC-VRF2 due to IGMP joins received from Rcvr3 and Rcvr4 respectively. PE2 now adds another OIF 'MAC-VRF2' to its existing routing entry. But there is no change in control plane states since its already sent MVPN route and no further signaling is required. Also since Src1 is not part of MAC-VRF2 subnet, it is treated as routing OIF and hence MAC header gets modified as per normal procedures for routing. PE3 forms routing entry very similar

to PE2. It is to be noted that PE3 does not have MAC-VRF1 configured locally but still can receive the multicast data traffic over Tenant1 tunnel formed due to MVPN procedures

## 9. DCs with only EVPN PEs

As mentioned earlier, the proposed solution can be used as a routed multicast solution in data center networks with only EVPN PEs (e.g., routed multicast VPN only among EVPN PEs). It should be noted that the scope of intra-subnet forwarding for the solution described in this document, is limited to a single EVPN PE for Single-Active multi-homing and to multi-homing PEs for All-Active multi-homing. In other words, the IP multicast traffic that needs to be forwarded from the source PE to remote PEs is routed to remote PEs regardless of whether the traffic is intra-subnet or inter-subnet. As the result, the TTL value for intra-subnet traffic that spans across two or more PEs get decremented. Based on past experiences with MVPN over last dozen years for supported IP multicast applications, layer-3 forwarding of intra-subnet multicast traffic should be fine. However, if there are applications that require intra-subnet multicast traffic to be L2 forwarded (e.g., without decrementing TTL value), then [EVPN-IRB-MCAST] proposes a solution to accommodate such applications.

### 9.1. Setup of overlay multicast delivery

It must be emphasized that this solution poses no restriction on the setup of the tenant BDs and that neither the source PE, nor the receiver PEs do not need to know/learn about the BD configuration on other PEs in the MVPN. The Reverse Path Forwarder (RPF) is selected per the tenant multicast source and the IP-VRF in compliance with the procedures in [RFC6514], using the incoming EVPN route type 2 or 5 NLRI per [RFC7432].

The VRF Route Import (VRI) extended community that is carried with the IP-VPN routes in [RFC6514] MUST be carried via the EVPN unicast routes instead. The construction and processing of the VRI are consistent with [RFC6514]. The VRI MUST uniquely identify the PE which is advertising a multicast source and the IP-VRF it resides in.

VRI is constructed as following:

- The 4-octet Global Administrator field MUST be set to an IP address of the PE. This address SHOULD be common for all the IP-VRFs on the PE (e.g., this address may be the PE's loopback

address).

- The 2-octet Local Administrator field associated with a given IP-VRF contains a number that uniquely identifies that IP-VRF within the PE that contains the IP-VRF.

Every PE which detects a local receiver via a local IGMP join or a local PIM join for a specific source (overlay SSM mode) MUST terminate the IGMP/PIM signaling at the IP-VRF and generate a (C-S,C-G) via the BGP MCAST-VPN route type 7 per [RFC6514] if and only if the RPF for the source points to the fabric. If the RPF points to a local multicast source on the same MAC-VRF or a different MAC-VRF on that PE, the MCAST-VPN MUST NOT be advertised and data traffic will be locally routed/bridged to the receiver as detailed in section 6.2.

The VRI received with EVPN route type 2 or 5 NLRI from source PE will be appended as an export route-target extended community. More details about handling of various types of local receivers are in section 10. The PE which has advertised the unicast route with VRI, will import the incoming MCAST-VPN NLRI in the IP-VRF with the same import route-target extended-community and other PEs SHOULD ignore it. Following such procedure the source PE learns about the existence of at least one remote receiver in the tenant overlay and programs data plane accordingly so that a single copy of multicast data is forwarded into the core VRF using tenant VRF tunnel.

If the multicast source is unknown (overlay ASM mode), the MCAST-VPN route type 6 (C-\*,C-G) join SHOULD be targeted towards the designated overlay Rendezvous Point (RP) by appending the received RP VRI as an export route-target extended community. Every PE which detects a local source, registers with its RP PE. That is how the RP learns about the tenant source(s) and group(s) within the MVPN. Once the overlay RP PE receives either the first remote (C-RP,C-G) join or a local IGMP/PIM join, it will trigger an MCAST-VPN route type 7 (C-S,C-G) towards the actual source PE for which it has received PIM register message in full compliance with regular PIM procedures. This involves the source PE to advertise the MCAST-VPN Source Active A-D route (MCAST-VPN route-type 5) towards all PEs. The Source Active A-D route is used to inform all PEs in a given MVPN about the active multicast source for switching from RPT to SPT when MVPNs use tenant RP-shared trees (i.e., rooted at tenant's RP) per section 13 of [RFC6514]. This is done in order to choose a single forwarder PE and to suppress receiving duplicate traffic. In such scenarios, the active multicast source is used by the receiver PEs to join the SPT if they have not received tenant (S,G) joins and by the RPT PEs to prune off the tenant (S,G) state from the RPT. The Source Active A-D route is also used for MVPN scenarios without tenant RP-shared trees. In such scenarios, the receiver PEs with tenant (\*,G) states use the Source Active A-D route to know which upstream PEs with sources

behind them to join per section 14 of [RFC6514] - i.e., to suppress joining Overlay shared tree.

## 9.2. Handling of different encapsulations

Just as in [RFC6514] the MVPN I-PMSI and S-PMSI A-D routes are used to form the overlay multicast tunnels and signal the tunnel type using the P-Multicast Service Interface Tunnel (PMSI Tunnel) attribute.

### 9.2.1. MPLS Encapsulation

The [RFC6514] assumes MPLS/IP core and there is no modification to the signaling procedures and encoding for PMSI tunnel formation therein. Also, there is no need for a gateway to inter-operate with non-EVPN PEs supporting [RFC6514] based MVPN over IP/MPLS.

### 9.2.2 VxLAN Encapsulation

In order to signal VXLAN, the corresponding BGP encapsulation extended community [TUNNEL-ENCAP] SHOULD be appended to the MVPN I-PMSI and S-PMSI A-D routes. The MPLS label in the PMSI Tunnel Attribute MUST be the Virtual Network Identifier (VNI) associated with the customer MVPN. The supported PMSI tunnel types with VXLAN encapsulation are: PIM-SSM Tree, PIM-SM Tree, BIDIR-PIM Tree, Ingress Replication [RFC6514]. Further details are in [RFC8365].

In this case, a gateway is needed for inter-operation between the EVPN PEs and non-EVPN MVPN PEs. The gateway should re-originate the control plane signaling with the relevant tunnel encapsulation on either side. In the data plane, the gateway terminates the tunnels formed on either side and performs the relevant stitching/re-encapsulation on data packets.

### 9.2.3. Other Encapsulation

In order to signal a different tunneling encapsulation such as NVGRE, GPE, or GENEVE the corresponding BGP encapsulation extended community [TUNNEL-ENCAP] SHOULD be appended to the MVPN I-PMSI and S-PMSI A-D routes. If the Tunnel Type field in the encapsulation extended-community is set to a type which requires Virtual Network Identifier (VNI), e.g., VXLAN-GPE or NVGRE [TUNNEL-ENCAP], then the MPLS label in the PMSI Tunnel Attribute MUST be the VNI associated with the customer MVPN. Same as in VXLAN case, a gateway is needed for inter-operation between the EVPN-IRB PEs and non-EVPN MVPN PEs.

## 10. DCI with MPLS in WAN and VxLAN in DCs

This section describes the inter-operation between MVPN PEs in WAN using MPLS encapsulation with EVPN PEs in a DC network using VxLAN encapsulation. Since the tunnel encapsulation between these networks are different, we must have at least one gateway in between. Usually, two or more are required for redundancy and load balancing purpose. In such scenarios, a DC network can be represented as a customer network that is multi-homed to two or more MVPN PEs via L3 interfaces and thus standard MVPN multi-homing procedures are applicable here. It should be noted that a MVPN overlay tunnel over the DC network is terminated on the IP-VRF of the gateway and not the MAC-VRF/BTs. Therefore, the considerations for loop prevention and split-horizon filtering described in [INTERCON-EVPN] are not applicable here. Some aspects of the multi-homing between VxLAN DC networks and MPLS WAN is in common with [INTERCON-EVPN].

#### 10.1. Control plane inter-connect

The gateway(s) MUST be setup with the inclusive set of all the IP-VRFs that span across the two domains. On each gateway, there will be at least two BGP sessions: one towards the DC side and the other towards the WAN side. Usually for redundancy purpose, more sessions are setup on each side. The unicast route propagation follows the exact same procedures in [INTERCON-EVPN]. Hence, a multicast host located in either domain, is advertised with the gateway IP address as the next-hop to the other domain. As a result, PEs view the hosts in the other domain as directly attached to the gateway and all inter-domain multicast signaling is directed towards the gateway(s). Received MVPN routes type 1-7 from either side of the gateway(s), MUST NOT be reflected back to the same side but processed locally and re-advertised (if needed) to the other side:

- Intra-AS I-PMSI A-D Route: these are distributed within each domain to form the overlay tunnels which terminate at gateway(s). They are not passed to the other side of the gateway(s).
- C-Multicast Route: joins are imported into the corresponding IP-VRF on each gateway and advertised as a new route to the other side with the following modifications (the rest of NLRI fields and path attributes remain on-touched):
  - \* Route-Distinguisher is set to that of the IP-VRF
  - \* Route-target is set to the exported route-target list on IP-VRF
  - \* The PMSI tunnel attribute and BGP Encapsulation extended community will be modified according to section 8
  - \* Next-hop will be set to the IP address which represents the gateway on either domain
- Source Active A-D Route: same as joins
- S-PMSI A-D Route: these are passed to the other side to form selective PMSI tunnels per every (C-S,C-G) from the gateway to the PEs in the other domain provided it contains receivers for the given (C-S, C-G). Similar modifications made to joins are made to the newly originated S-PMSI.

In addition, the Originating Router's IP address is set to GW's IP address. Multicast signaling from/to hosts on local ACs on the gateway(s) are generated and propagated in both domains (if needed) per the procedures in section 7 in this document and in [RFC6514] with no change. It must be noted that for a locally attached source, the gateway will program an OIF per every domain from which it receives a remote join in its forwarding plane and different encapsulation will be used on the data packets.

## 10.2. Data plane inter-connect

Traffic forwarding procedures on gateways are same as those described for PEs in section 5 and 6 except that, unlike a non-border leaf PE, the gateway will not only route the incoming traffic from one side to its local receivers, but will also send it to the remote receivers in the the other domain after de-capsulation and appending the right encapsulation. The OIF and IIF are programmed in FIB based on the received joins from either side and the RPF calculation to the source or RP. The de-capsulation and encapsulation actions are programmed based on the received I-PMSI or S-PMSI A-D routes from either sides.

If there are more than one gateway between two domains, the multi-homing procedures described in the following section must be considered so that incoming traffic from one side is not looped back to the other gateway.

The multicast traffic from local sources on each gateway flows to the other gateway with the preferred WAN encapsulation.

#### 11. IANA Considerations

There is no additional IANA considerations for PBB-EVPN beyond what is already described in [RFC7432].

#### 12. Security Considerations

All the security considerations in [RFC7432] apply directly to this document because this document leverages [RFC7432] control plane and their associated procedures.

#### 13. Acknowledgements

The authors would like to thank Niloofar Fazlollahi, Aamod Vyavaharkar, Kesavan Thiruvengatasamy, and Swadesh Agrawal for their discussions and contributions.

#### 14. References

##### 14.1. Normative References

- [RFC7432] A. Sajassi, et al., "BGP MPLS Based Ethernet VPN", RFC 7432, February 2015.
- [RFC8365] A. Sajassi, et al., "A Network Virtualization Overlay Solution using EVPN", RFC 8365, February 2018.
- [RFC6513] E. Rosen, et al., "Multicast in MPLS/BGP IP VPNs", RFC6513, February 2012.
- [RFC6514] R. Aggarwal, et al., "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC6514, February 2012.

##### 14.2. Informative References



- [RFC7080] A. Sajassi, et al., "Virtual Private LAN Service (VPLS) Interoperability with Provider Backbone Bridges", RFC 7080, December 2013.
- [RFC7209] D. Thaler, et al., "Requirements for Ethernet VPN (EVPN)", RFC 7209, May 2014.
- [RFC4389] A. Sajassi, et al., "Neighbor Discovery Proxies (ND Proxy)", RFC 4389, April 2006.
- [RFC4761] K. Kompella, et al., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [INTERCON-EVPN] J. Rabadan, et al., "Interconnect Solution for EVPN Overlay networks", <https://tools.ietf.org/html/draft-ietf-bess-dci-evpn-overlay-04>, September 2016
- [TUNNEL-ENCAPS] E. Rosen, et al. "The BGP Tunnel Encapsulation Attribute", <https://tools.ietf.org/html/draft-ietf-idr-tunnel-encaps-06>, work in progress, June 2017.
- [EVPN-IGMP-PROXY] A. Sajassi, et. al., "IGMP and MLD Proxy for EVPN", <https://tools.ietf.org/html/draft-ietf-bess-evpn-igmp-mlt-proxy-01>, work in progress, March 2018.
- [EVPN-PIM-PROXY] J. Rabadan, et. al., "PIM Proxy in EVPN Networks", <https://tools.ietf.org/html/draft-skr-bess-evpn-pim-proxy-00>, work in progress, July 3, 2017.

15. Authors' Addresses

Ali Sajassi  
Cisco  
170 West Tasman Drive  
San Jose, CA 95134, US  
Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Samir Thoria  
Cisco  
170 West Tasman Drive  
San Jose, CA 95134, US  
Email: [sthoria@cisco.com](mailto:sthoria@cisco.com)

Ashutosh Gupta  
Avi Networks

Email: ashutosh@avinetworks.com

## Appendix A. Use Cases

### A.1. DCs with only IGMP/MLD hosts w/o tenant router

In a EVPN network consisting of only IGMP/MLD hosts, PE's will receive IGMP (\*, G) or (S, G) joins from their locally attached host and would originate MVPN C-Multicast Route Type 6 and 7 NLRI's respectively. As described in RFC 6514 these NLRI's are directed towards RP-PE for Type 6 or Source-PE for Type 7. In case of (\*, G) join a Shared-Path Tree will be built in the core from RP-PE towards all Receiver-PE's. Once a Source starts to send Multicast data to specified multicast-group, the PE directly connected to Source will do PIM-registration with RP. Since there are existing receivers for the Group, RP will originate a PIM (S, G) join towards Source. This will be converted to MVPN Type 7 NLRI by RP-PE. Please note that the router RP-PE would be the PE configured as RP (e.g., using static configuration or by using BSR or Auto-RP procedures). The detailed working of such protocols is beyond the scope of this document. Upon receiving Type 7 NLRI, Source-PE will include MVPN Tunnel in its Outgoing Interface List. Furthermore, Source-PE will follow the procedures in RFC-6514 to originate MVPN SA-AD route (RT 5) to avoid duplicate traffic and allow all Receiver-PE's to shift from Share-Tree to Shortest-Path-Tree rooted at Source-PE. Section 13 of [RFC6514] describes it.

However a network operator can chose to have only Shortest-Path-Tree built in MVPN core as described in section 14 of [RFC6514]. One way to achieve this, is for all PE's act as RP for its locally connected hosts and thus avoid sending any Shared-Tree Join (MVPN Type 6) into the core. In this scenario, there will be no PIM registration needed since all PE's are first-hop router as well as acting RP. Once a source starts to send multicast data, the PE directly connected to it originates Source-Active AD (RT 5) to all other PE's in network. Upon Receiving Source-Active AD route a PE must cache it in its local database and also look for any matching interest for (\*, G) where G is the multicast group described in received Source-Active AD route. If it finds any such matching entry, it must originate a C-Multicast route (RT 7) in order to start receiving traffic from Source-PE.

This procedure must be repeated on reception of any further Source-Active AD routes.

A.2. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-SSM

This scenario has multicast routers which can send PIM SSM (S, G) joins. Upon receiving these joins and if source described in join is learnt to be behind a MVPN peer PE, local PE will originate C-Multicast Join (RT 7) towards Source-PE. It is expected that PIM SSM group ranges are kept separate from ASM range for which IGMP hosts can send (\*, G) joins. Hence both ASM and SSM groups shall operate without any overlap. There is no RP needed for SSM range groups and Shortest Path tree rooted at Source is built once a receiver interest is known.

A.3. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-ASM

This scenario includes reception of PIM (\*, G) joins on PE's local AC. These joins are handled similar to IGMP (\*, G) join as explained in sections above. Another interesting case can arise here is when one of the tenant routers can act as RP for some of the ASM Groups. In such scenario, a Upstream Multicast Hop (UMH) will be elected by other PE's in order to send C-Multicast Routes (RT 6). All procedures described in RFC 6513 with respect to UMH should be used to avoid traffic duplication due to incoherent selection of RP-PE by different Receiver-PE's.

A.4. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-Bidir

Creating Bidirectional (\*, G) trees is useful when a customer wants least amount of control state in network. But on downside all receivers for a particular multicast group receive traffic from all sources sending to that group. However for the purpose of this document, all procedures as described in RFC 6513 and RFC 6514 apply when PIM-Bidir is used.

BESS WorkGroup  
Internet-Draft  
Intended status: Standards Track  
Expires: December 30, 2018

Ali. Sajassi  
Mankamana. Mishra  
Samir. Thoria  
Cisco Systems  
Jorge. Rabadan  
Nokia  
John. Drake  
Juniper Networks  
June 28, 2018

Per multicast flow Designated Forwarder Election for EVPN  
draft-sajassi-bess-evpn-per-mcast-flow-df-election-01

## Abstract

[RFC7432] describes mechanism to elect designated forwarder (DF) at the granularity of (ESI, EVI) which is per VLAN (or per group of VLANs in case of VLAN bundle or VLAN-aware bundle service). However, the current level of granularity of per-VLAN is not adequate for some applications. [I-D.ietf-bess-evpn-df-election-framework] improves base line DF election by introducing HRW DF election. [I-D.ietf-bess-evpn-igmp-mld-proxy] introduces applicability of EVPN to Multicast flows, routes to sync them and a default DF election. This document is an extension to HRW base draft [I-D.ietf-bess-evpn-df-election-framework] and further enhances HRW algorithm for the Multicast flows to do DF election at the granularity of (ESI, VLAN, Mcast flow).

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on December 30, 2018.

## Copyright Notice

Copyright (c) 2018 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 (<https://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. Introduction . . . . .	2
2. Terminology . . . . .	4
3. The DF Election Extended Community . . . . .	4
4. HRW base per multicast flow EVPN DF election . . . . .	6
4.1. DF election for IGMP (S,G) membership request . . . . .	6
4.2. DF election for IGMP (*,G) membership request . . . . .	7
4.3. Default DF election procedure . . . . .	7
5. Procedure to use per multicast flow DF election algorithm . . . . .	8
6. Triggers for DF re-election . . . . .	9
7. Security Considerations . . . . .	10
8. IANA Considerations . . . . .	10
9. Acknowledgement . . . . .	10
10. Normative References . . . . .	10
Authors' Addresses . . . . .	11

## 1. Introduction

EVPN based All-Active multi-homing is becoming the basic building block for providing redundancy in next generation data center deployments as well as service provider access/aggregation networks. [RFC7432] defines the role of a designated forwarder as the node in the redundancy group that is responsible to forward Broadcast, Unknown unicast, Multicast (BUM) traffic on that Ethernet Segment (CE device or network) in All-Active multi-homing.

The default DF election mechanism allows selecting a DF at the granularity of (ES, VLAN) or (ES, VLAN bundle) for BUM traffic. While [I-D.ietf-bess-evpn-df-election-framework] improve on the default DF election procedure, some service provider residential applications require a finer granularity, where whole multicast flows are delivered on a single VLAN.

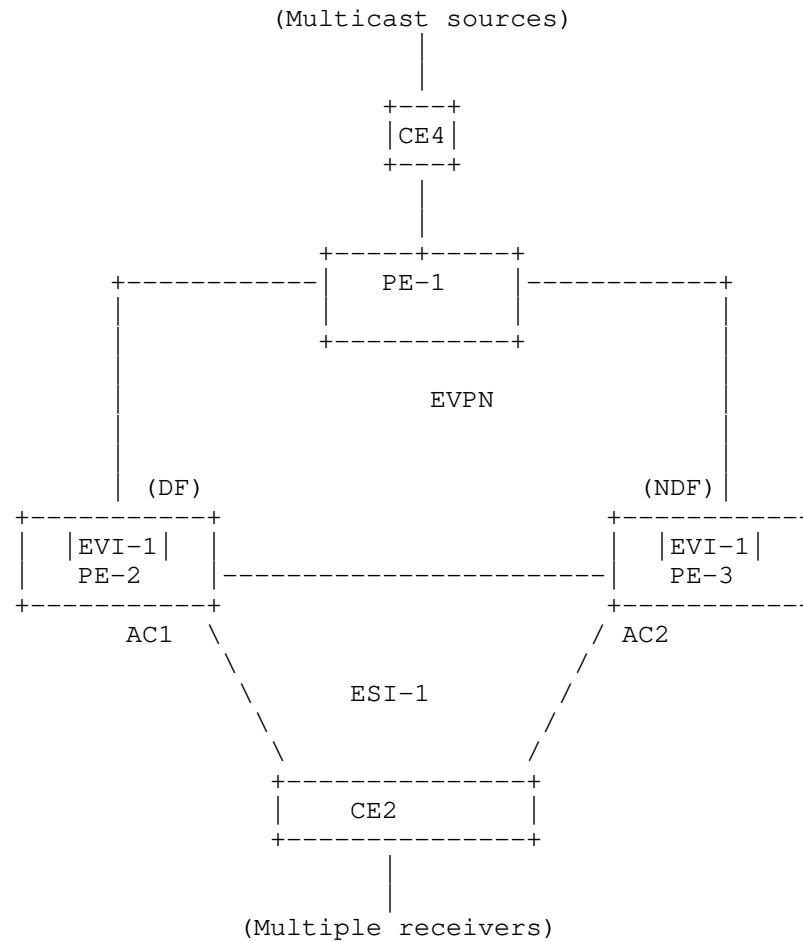


Figure 1: Multi-homing Network of EVPN  
for IPTV deployments

Consider the above topology, which shows a typical residential deployment scenario, where multiple receivers are behind an all-active multihoming segments. All of the multicast traffic is provisioned on EVI-1. Assume PE-2 get elected as DF. According to [RFC7432], PE-2 will be responsible for forwarding multicast traffic to that Ethernet segment.

- o Forcing sole data plane forwarding responsibility on PE-2 is a limitation in the current DF election mechanism. The topology at Figure 1 would always have only one of the PE to be elected as DF irrespective of which current DF election mechanism is in use

defined in [RFC7432] or  
[I-D.ietf-bess-evpn-df-election-framework].

- o The problem may also manifest itself in a different way. For example, AC1 happens to use 80% of its available bandwidth to forward unicast data. And now there is need to serve multicast receivers where it would require more than 20% of AC1 bandwidth. In this case, AC1 becomes oversubscribed and multicast traffic drop would be observed even though there is already another link (AC2) present in network which can be used more efficiently load balance the multicast traffic.

In this document, we propose an extension to the HRW base draft to allow DF election at the granularity of (ESI, VLAN, Mcast flow) which would allow multicast flows to be better distributed among redundancy group PEs to share the load.

## 2. Terminology

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] .

With respect to EVPN, this document follows the terminology that has been defined in [RFC7432] and [RFC4601] for multicast terminology.

## 3. The DF Election Extended Community

[I-D.ietf-bess-evpn-df-election-framework] defines an extended community, which would be used for PEs in redundancy group to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs in redundancy group about its willingness to support Per multicast flow base DF election capability by signaling a DF election extended community along with Ethernet-Segment Route (Type-4). The current proposal extends the existing extended community defined in [I-D.ietf-bess-evpn-df-election-framework]. This draft defines new a DF type.

- o DF type (1 octet) - Encodes the DF Election algorithm values (between 0 and 255) that the advertising PE desires to use for the ES.
  - \* Type 0: Default DF Election algorithm, or modulus-based algorithms in [RFC7432].
  - \* Type 1: HRW algorithm defined in [I-D.ietf-bess-evpn-df-election-framework]

- \* Type 2: Handshake defines in [I-D.ietf-bess-evpn-fast-df-recovery]
  - \* Type 3: Time-Synch defined in [I-D.ietf-bess-evpn-fast-df-recovery]
  - \* Type 4: HRW base per (S,G) multicast flow DF election (explained in this document)
  - \* Type 5: HRW base per (\*,G) multicast flow DF election (explained in this document)
  - \* Type 6 – 254: Unassigned
  - \* Type 255: Reserved for Experimental Use.
- o The [I-D.ietf-bess-evpn-df-election-framework] describes encoding of capabilities associated to the DF election algorithm using Bitmap field. When these capabilities bits are set along with the DF type-4 and type-5, they need to be interpreted in context of this new DF type-4 and type-5. For example, consider a scenario where all PEs in the same redundancy group (same ES) can support both AC-DF, DF type-4 and DF type-5 and receive such indications from the other PEs in the ES. In this scenario, if a VLAN is not active in a PE, then the DF election procedure on all PEs in the ES should factor that in and exclude that PE in the DF election per multicast flow.
  - o A PE SHOULD attach the DF election Extended Community to ES route and Extended Community MUST be sent if the ES is locally configured for DF type Per Multicast flow DF election. Only one DF Election Extended community can be sent along with an ES route.
  - o When a PE receives the ES Routes from all the other PEs for the ES, it checks if all of other PEs have advertised their desire to proceed by Per multicast flow DF election. If all peering PEs have done so, it performs DF election based on Per multicast flow procedure. But if:
    - \* There is at least one PE which advertised route-4 ( AD per ES Route) which does not indicate its capability to perform Per multicast flow DF election. OR
    - \* There is at least one PE signaling single active in the AD per ES route



it MUST be considered as an indication to support of only Default DF election [RFC7432] and DF election procedure in [RFC7432] MUST be used.

#### 4. HRW base per multicast flow EVPN DF election

This document is an extension of [I-D.ietf-bess-evpn-df-election-framework], so this draft does not repeat the description of HRW algorithm itself.

EVPN PE does the discovery of redundancy groups based on [RFC7432]. If redundancy group consists of N peering EVPN PE nodes, after the discovery all PEs build an unordered list of IP address of all the nodes in the redundancy group. The procedure defined in this draft does not require the list of PEs to be ordered. Address [i] denotes the IP address of the [i]th EVPN PE in redundancy group where  $(0 < i \leq N)$ .

##### 4.1. DF election for IGMP (S,G) membership request

The DF is the PE who has maximum weight for (S, G, V, Es) where

- o S - Multicast Source
- o G - Multicast Group
- o V - VLAN ID.
- o Es - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

##### 1. Weight

- \* The weight of PE(i) to (S,G,VLAN ID, Es) is calculated by function,  $\text{weight}(S, G, V, Es, \text{Address}(i))$ , where  $(0 < i \leq N)$ , PE(i) is the PE at ordinal i.
- \*  $\text{Weight}(S, G, V, Es, \text{Address}(i)) = (1103515245 \cdot ((1103515245 \cdot \text{Address}(i) + 12345) \text{ XOR } D(S, G, V, \text{ESI})) + 12345) \pmod{2^{31}}$
- \* In case of tie, the PE whose IP address is numerically least is chosen.

##### 2. Digest

- \*  $D(S, G, V, Es) = CRC\_32(S, G, V, Es)$
- \* Here  $D(S, G, V, Es)$  is the 31-bit digest (CRC\_32 and discarding the MSB) of the Source IP, Group IP, Vlan ID and Es. The CRC MUST proceed as if the architecture is in network byte order (big-endian).

#### 4.2. DF election for IGMP (\*,G) membership request

The DF is the PE who has maximum weight for (G, V, Es) where

- o G - Multicast Group
- o V - VLAN ID.
- o Es - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

##### 1. Weight

- \* The weight of PE(i) to (G, VLAN ID, Es) is calculated by function,  $weight(G, V, Es, Address(i))$ , where  $(0 < i \leq N)$ , PE(i) is the PE at ordinal i.
- \*  $Weight(G, V, Es, Address(i)) = (1103515245 \cdot ((1103515245 \cdot Address(i) + 12345) \text{ XOR } D(G, V, Es)) + 12345) \pmod{2^{31}}$
- \* In case of tie, the PE whose IP address is numerically least is chosen.

##### 2. Digest

- \*  $D(G, V, Es) = CRC\_32(G, V, Es)$
- \* Here  $D(G, V, Es)$  is the 31-bit digest (CRC\_32 and discarding the MSB) of the Group IP, Vlan ID and Es. The CRC MUST proceed as if the architecture is in network byte order (big-endian).

#### 4.3. Default DF election procedure

Per multicast DF election procedure would be applicable only when host behind Attachment Circuit (of the Es) start sending IGMP membership requests. Membership requests are synced using procedure defined in [I-D.ietf-bess-evpn-igmp-mld-proxy], and each of the PE in redundancy group can use per flow DF election and create DF state per

multicast flow. The HRW DF election "Type 1" procedure defined in [I-D.ietf-bess-evpn-df-election-framework] MUST be used for the Es DF election and SHOULD be performed on Es even before learning multicast membership request state. This default election procedure MUST be used at port level but will be overwritten by Per flow DF election as and when new membership request state are learnt.

##### 5. Procedure to use per multicast flow DF election algorithm

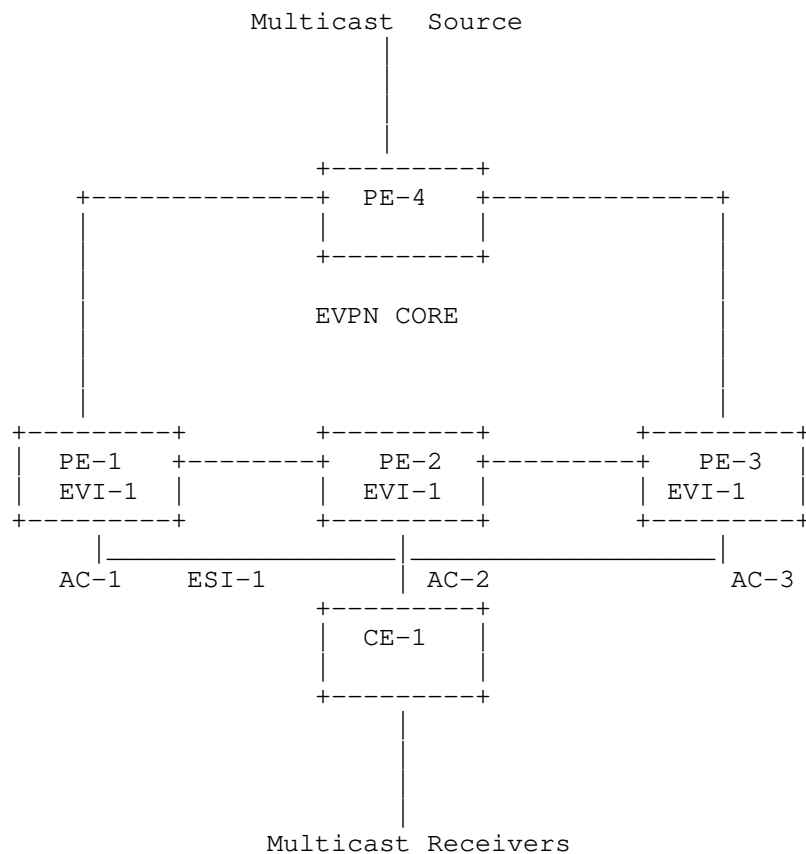


Figure-2 : Multihomed network

Figure-2 shows multihomed network. Where EVPN PE-1, PE-2, PE-3 are multihomed to CE-1. Multiple multicast receivers are behind all active multihoming segment.

1. PEs connected to the same Ethernet segment can automatically discover each other through exchange of the Ethernet Segment

Route. This draft does not change any of this procedure, it still uses the procedure defined in [RFC7432].

2. Each of the PEs in redundancy group advertise Ethernet segment route with extended community indicating their ability to participate in per multicast flow DF election procedure. Since Per multicast flow would not be applicable unless PE learns about membership request from receiver, there is a need to have the default DF election among PEs in redundancy group for BUM traffic. Until multicast membership state are learnt, we use the the DF election procedure in Section 4.3, namely HRW per (v,Es) as defined in [I-D.ietf-bess-evpn-df-election-framework] .
3. When a receiver starts sending membership requests for (s1,g1), where s1 is multicast source address and g1 is multicast group address, CE-1 could hash membership request (IGMP join) to any of the PEs in redundancy group. Let's consider it is hashed to PE-2. [I-D.ietf-bess-evpn-igmp-mld-proxy] defines a procedure to sync IGMP join state among redundancy group of PEs. Now each of the PE would have information about membership request (s1,g1) and each of them run DF election procedure Section 4.1 to elect DF among participating PEs in redundancy group. Consider PE-2 gets elected as DF for multicast flow (s1,g1).
  1. PE-1 forwarding state would be nDF for flow (s1,g1) and DF for rest other BUM traffic.
  2. PE-2 forwarding state would be DF for flow (s1,g1) and nDF for rest other BUM traffic.
  3. PE-3 forwarding state would be nDF for flow (s1,g1) and rest other BUM traffic.
4. As and when new multicast membership request comes, same procedure as above would continue.
5. If Section 3 has DF type 4, For membership request (S,G) it MUST use Section 4.1 to elect DF among participating PEs. And membership request (\*,G) MUST use Section 4.2 to elect DF among participating PEs.
6. Triggers for DF re-election

There are multiple triggers which can cause DF re-election. Some of the triggers could be

  1. Local ES going down due to physical failure or configuration change triggers DF re-election at peering PE.

2. Detection of new PE through ES route.
3. AC going up / down
4. ESI change
5. Remote PE removed / Down
6. Local configuration change of DF election Type and peering PE consensus on new DF Type

This document does not provide any new mechanism to handle DF re-election procedure. It uses the existing mechanism defined in [RFC7432]. Whenever either of the triggers occur, a DF re-election would be done. and all of the flows would be redistributed among existing PEs in redundancy group for ES.

## 7. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

## 8. IANA Considerations

Allocation of DF type in DF extended community for EVPN.

## 9. Acknowledgement

Authors would like to acknowledge helpful comments and contributions of Luc Andre Burdet.

## 10. Normative References

[HRW1999]    IEEE, "Using name-based mappings to increase hit rates",  
IEEE HRW, February 1998.

[I-D.ietf-bess-evpn-df-election-framework]  
Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake,  
J., Nagaraj, K., and S. Sathappan, "Framework for EVPN  
Designated Forwarder Election Extensibility", draft-ietf-  
bess-evpn-df-election-framework-03 (work in progress), May  
2018.

[I-D.ietf-bess-evpn-fast-df-recovery]  
Sajassi, A., Badoni, G., Rao, D., Brissette, P., Drake,  
J., and J. Rabadan, "Fast Recovery for EVPN DF Election",  
draft-ietf-bess-evpn-fast-df-recovery-00 (work in  
progress), June 2018.

- [I-D.ietf-bess-evpn-igmp-mld-proxy]  
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J.,  
and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-  
bess-evpn-igmp-mld-proxy-00 (work in progress), March  
2017.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", BCP 14, RFC 2119,  
DOI 10.17487/RFC2119, March 1997,  
<<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,  
"Protocol Independent Multicast - Sparse Mode (PIM-SM):  
Protocol Specification (Revised)", RFC 4601,  
DOI 10.17487/RFC4601, August 2006,  
<<https://www.rfc-editor.org/info/rfc4601>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A.,  
Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based  
Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February  
2015, <<https://www.rfc-editor.org/info/rfc7432>>.

Authors' Addresses

Ali Sajassi  
Cisco Systems  
821 Alder Drive,  
MILPITAS, CALIFORNIA 95035  
UNITED STATES

Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Mankamana Mishra  
Cisco Systems  
821 Alder Drive,  
MILPITAS, CALIFORNIA 95035  
UNITED STATES

Email: [mankamis@cisco.com](mailto:mankamis@cisco.com)

Samir Thoria  
Cisco Systems  
821 Alder Drive,  
MILPITAS, CALIFORNIA 95035  
UNITED STATES

Email: sthoria@cisco.com

Jorge Rabadan  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043  
UNITED STATES

Email: jorge.rabadan@nokia.com

John Drake  
Juniper Networks

Email: jdrake@juniper.net

INTERNET-DRAFT  
Intended Status: Informational

Samer Salam  
Ali Sajassi  
Cisco  
Sam Aldrin  
Google  
John E. Drake  
Juniper  
Donald Eastlake  
Huawei  
May 29, 2018

Expires: November 28, 2018

EVPN Operations, Administration and Maintenance  
Requirements and Framework  
draft-salam-bess-evpn-oam-req-frmwk-00

## Abstract

This document specifies the requirements and reference framework for Ethernet VPN (EVPN) Operations, Administration and Maintenance (OAM). The requirements cover the OAM aspects of EVPN and PBB-EVPN. The framework defines the layered OAM model encompassing the EVPN service layer, network layer and underlying Packet Switched Network (PSN) transport layer.

## 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/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2018 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. Introduction.....	4
1.1 Relationship to Other OAM Work.....	4
1.2 Specification of Requirements.....	5
1.3 Terminology.....	5
2. EVPN OAM Framework.....	6
2.1 OAM Layering.....	6
2.2 EVPN Service OAM.....	7
2.3 EVPN Network OAM.....	7
2.4 Transport OAM for EVPN.....	9
2.5 Link OAM.....	9
2.6 OAM Inter-working.....	9
3. EVPN OAM Requirements.....	11
3.1 Fault Management Requirements.....	11
3.1.1 Proactive Fault Management Functions.....	11
3.1.1.1 Fault Detection (Continuity Check).....	11
3.1.1.2 Defect Indication.....	12
3.1.1.2.1 Forward Defect Indication.....	12
3.1.1.2.2 Reverse Defect Indication (RDI).....	12
3.1.2 On-Demand Fault Management Functions.....	13
3.1.2.1 Connectivity Verification.....	13
3.1.2.2 Fault Isolation.....	14
3.2 Performance Management.....	14
3.2.1 Packet Loss.....	14
3.2.2 Packet Delay.....	15
4. Security Considerations.....	16
5. Acknowledgements.....	16
6. IANA Considerations.....	16
Normative References.....	17
Informative References.....	18

## 1. Introduction

This document specifies the requirements and defines a reference framework for Ethernet VPN (EVPN) Operations, Administration and Maintenance (OAM, [RFC6291]). In this context, we use the term EVPN OAM to loosely refer to the OAM functions required for and/or applicable to [RFC7432] and [RFC7623].

EVPN is an L2VPN solution for multipoint Ethernet services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the core MPLS/IP network.

PBB-EVPN combines Provider Backbone Bridging (PBB) [802.1Q] with EVPN in order to reduce the number of BGP MAC advertisement routes, provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes.

This document focuses on the fault management and performance management aspects of EVPN OAM.

### 1.1 Relationship to Other OAM Work

This document leverages concepts and draws upon elements defined and/or used in the following documents:

[RFC6136] specifies the requirements and a reference model for OAM as it relates to L2VPN services, pseudowires and associated Packet Switched Network (PSN) tunnels. This document focuses on VPLS and VPWS solutions and services.

[RFC8029] defines mechanisms for detecting data plane failures in MPLS LSPs, including procedures to check the correct operation of the data plane, as well as mechanisms to verify the data plane against the control plane.

[802.1Q] specifies the Ethernet Connectivity Fault Management (CFM) protocol, which defines the concepts of Maintenance Domains, Maintenance Associations, Maintenance End Points, and Maintenance Intermediate Points.

[Y.1731] extends Connectivity Fault Management in the following areas: it defines fault notification and alarm suppression functions for Ethernet. It also specifies mechanisms for Ethernet performance management, including loss, delay, jitter, and throughput measurement.

## 1.2 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] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 1.3 Terminology

This document uses the following terminology defined in [RFC6136]:

- MA Maintenance Association is a set of MEPs belonging to the same Maintenance Domain, established to verify the integrity of a single service instance.
- MEP Maintenance End Point is responsible for origination and termination of OAM frames for a given MA.
- MIP Maintenance Intermediate Point is located between peer MEPs and can process and respond to certain OAM frames but does not initiate them.
- MD Maintenance Domain, an OAM Domain that represents a region over which OAM frames can operate unobstructed.

## 2. EVPN OAM Framework

### 2.1 OAM Layering

Multiple layers come into play for implementing an L2VPN service using the EVPN family of solutions:

- The Service Layer runs end to end between the sites, or Ethernet Segments, that are being interconnected by the EVPN solution.
- The Network Layer extends in between the EVPN PE nodes and is mostly transparent to the core nodes (except where Flow Entropy comes into play). It leverages MPLS for service (i.e. EVI) multiplexing and Split-Horizon functions.
- The Transport Layer is dictated by the networking technology of the PSN. It may be either based on MPLS LSPs or IP.
- The Link Layer is dependent upon the physical technology used. Ethernet is a popular choice for this layer, but other alternatives are deployed (e.g. POS, DWDM etc.).

This layering extends to the set of OAM protocols that are involved in the ongoing maintenance and diagnostics of EVPN networks. The figure below depicts the OAM layering, and shows which devices have visibility into what OAM layer(s).

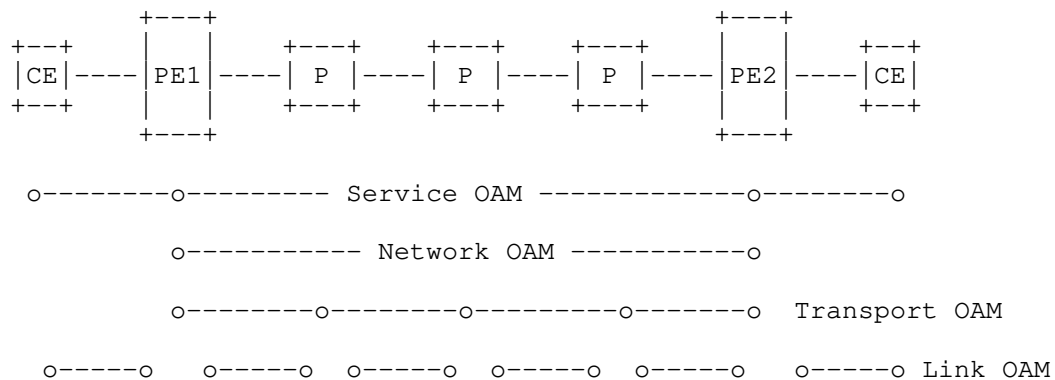


Figure 1: EVPN OAM Layering

Figure 2 below shows an example network where native Ethernet domains are interconnected via EVPN, and the OAM mechanisms applicable at each layer. The details of the layers are described in the sections that follow.

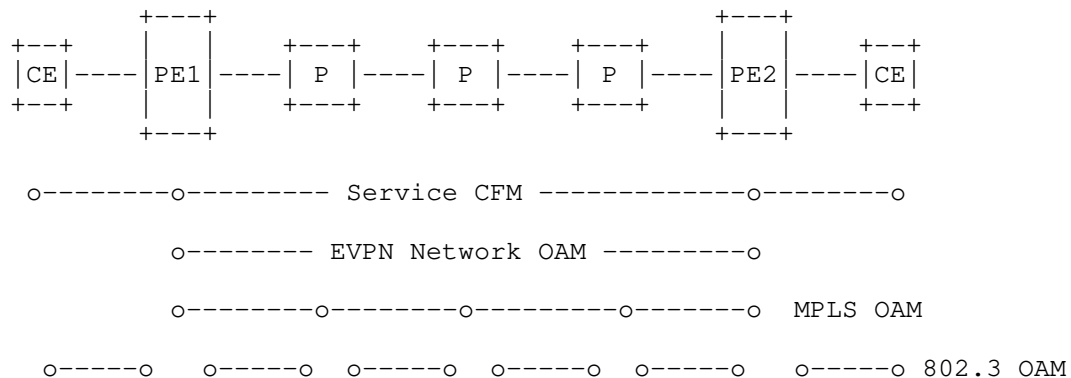


Figure 2: EVPN OAM Example

## 2.2 EVPN Service OAM

The EVPN Service OAM protocol depends on what service layer technology is being interconnected by the EVPN solution. In case of [RFC7432] and [RFC7623], the service layer is Ethernet; hence, the corresponding service OAM protocol is Ethernet Connectivity Fault Management (CFM) [802.1Q].

EVPN service OAM is visible to the CEs and EVPN PEs, but not to the core (P) nodes. This is because the PEs operate at the Ethernet MAC layer in [RFC7432] [RFC7623] whereas the P nodes do not.

The EVPN PE MUST support MIP functions in the applicable service OAM protocol, for example Ethernet CFM.

The EVPN PE SHOULD support MEP functions in the applicable service OAM protocol. This includes both Up and Down MEP functions.

## 2.3 EVPN Network OAM

EVPN Network OAM is visible to the PE nodes only. This OAM layer is analogous to VCCV [RFC5085] in the case of VPLS/VPWS. It provides mechanisms to check the correct operation of the data plane, as well as a mechanism to verify the data plane against the control plane. This includes the ability to perform fault detection and diagnostics on:

- the MP2P tunnels used for the transport of unicast traffic between PEs. EVPN allows for three different models of unicast label assignment: label per EVI, label per <ESI, Ethernet Tag> and label

per MAC address. In all three models, the label is bound to an EVPN Unicast FEC.

EVPN Network OAM MUST provide mechanisms to check the operation of the data plane and verify that operation against the control plane view for the EVPN Unicast FEC.

- the MP2P tunnels used for aliasing unicast traffic destined to a multi-homed Ethernet Segment. The three label assignment models, discussed above, apply here as well. In all three models, the label is bound to an EVPN Aliasing FEC. EVPN Network OAM MUST provide mechanisms to check the operation of the data plane and verify that operation against the control plane view for the EVPN Aliasing FEC.
- the multicast tunnels (either MP2P or P2MP) used for the transport of broadcast, unknown unicast and multicast traffic between PEs. In the case of ingress replication, a label is allocated per EVI or per <EVI, Ethernet Tag> and is bound to an EVPN Multicast FEC. In the case of LSM, and more specifically aggregate inclusive trees, again a label may be allocated per EVI or per <EVI, Ethernet Tag> and is bound to an EVPN Multicast FEC.

EVPN Network OAM MUST provide mechanisms to check the operation of the data plane and verify that operation against the control plane view for the EVPN Multicast FEC.

- the correct operation of the ESI split-horizon filtering function. In EVPN, a label is allocated per multi-homed Ethernet Segment for the purpose of performing the access split-horizon enforcement. The label is bound to an EVPN Ethernet Segment FEC.

EVPN Network OAM MUST provide mechanisms to check the operation of the data plane and verify that operation against the control plane view for the EVPN Ethernet Segment FEC.

- the correct operation of the DF filtering function.

EVPN Network OAM MUST provide mechanisms to check the operation of the data plane and verify that operation against the control plane view for the DF filtering function.

EVPN network OAM mechanisms MUST provide in-band management capabilities. As such, OAM messages MUST be encoded so that they exhibit identical entropy characteristics to data traffic.

EVPN network OAM SHOULD provide both proactive and on-demand mechanisms of monitoring the data plane operation and data plane conformance to the state of the control plane.

## 2.4 Transport OAM for EVPN

The transport OAM protocol depends on the nature of the underlying transport technology in the PSN. MPLS OAM mechanisms [RFC8029] [RFC6425] as well as ICMP [RFC792] are applicable, depending on whether the PSN employs MPLS or IP transport, respectively. Furthermore, BFD mechanisms per [RFC5880], [RFC5881], [RFC5883] and [RFC5884] apply. Also, the BFD mechanisms pertaining to MPLS-TP LSPs per [RFC6428] are applicable.

## 2.5 Link OAM

Link OAM depends on the data link technology being used between the PE and P nodes. For example, if Ethernet links are employed, then Ethernet Link OAM [802.3] Clause 57 may be used.

## 2.6 OAM Inter-working

When inter-working two networking domains, such as native Ethernet and EVPN to provide an end-to-end emulated service, there is a need to identify the failure domain and location, even when a PE supports both the Service OAM mechanisms and the EVPN Network OAM mechanisms. In addition, scalability constraints may not allow running proactive monitoring, such as Ethernet Continuity Check Messages (CCMs), at a PE to detect the failure of an EVI across the EVPN domain. Thus, the mapping of alarms generated upon failure detection in one domain (e.g. native Ethernet or EVPN network domain) to the other domain is needed. There are also cases where a PE may not be able to process Service OAM messages received from a remote PE over the PSN even when such messages are defined, as in the Ethernet case, thereby necessitating support for fault notification message mapping between the EVPN Network domain and the Service domain.

OAM inter-working is not limited though to scenarios involving disparate network domains. It is possible to perform OAM inter-working across different layers in the same network domain. In general, alarms generated within an OAM layer, as a result of proactive fault detection mechanisms, may be injected into its client layer OAM mechanisms. This allows the client layer OAM to trigger event-driven (i.e. asynchronous) fault notifications. For example, alarms generated by the Link OAM mechanisms may be injected into the Transport OAM layer, and alarms generated by the Transport OAM mechanism may be injected into the Network OAM mechanism, and so on.

EVPN OAM MUST support inter-working between the Network OAM and Service OAM mechanisms. EVPN OAM MAY support inter-working among



other OAM layers.

### 3. EVPN OAM Requirements

This section discusses the EVPN OAM requirements pertaining to Fault Management and Performance Management.

#### 3.1 Fault Management Requirements

##### 3.1.1 Proactive Fault Management Functions

The network operator configures proactive fault management functions to run periodically without a time bound. Certain actions, for example protection switchover or alarm indication signaling, can be associated with specific events, such as entering or clearing fault states.

###### 3.1.1.1 Fault Detection (Continuity Check)

Proactive fault detection is performed by periodically monitoring the reachability between service endpoints, i.e. MEPs in a given MA, through the exchange of Continuity Check messages. The reachability between any two arbitrary MEPs may be monitored for:

- in-band per-flow monitoring. This enables per flow monitoring between MEPs. EVPN Network OAM MUST support fault detection with per user flow granularity. EVPN Service OAM MAY support fault detection with per user flow granularity.
- a representative path. This enables liveness check of the nodes hosting the MEPs assuming that the loss of continuity to the MEP is interpreted as a failure of the hosting node. This, however, does not conclusively indicate liveness of the path(s) taken by user data traffic. This enables node failure detection but not path failure detection, through the use of a test flow. EVPN Network OAM and Service OAM MUST support fault detection using test flows.
- all paths. For MPLS/IP networks with ECMP, monitoring of all unicast paths between MEPs (on non-adjacent nodes) may not be possible, since the per-hop ECMP hashing behavior may yield situations where it is impossible for a MEP to pick flow entropy characteristics that result in exercising the exhaustive set of ECMP paths. Monitoring of all ECMP paths between MEPs (on non-adjacent nodes) is not a requirement for EVPN OAM.

The fact that MPLS/IP networks do not enforce congruency between

unicast and multicast paths means that the proactive fault detection mechanisms for EVPN networks MUST provide procedures to monitor the unicast paths independently of the multicast paths. This applies to EVPN Service OAM and Network OAM.

### 3.1.1.2 Defect Indication

EVPN Service OAM MUST support event-driven defect indication upon the detection of a connectivity defect. Defect indications can be categorized into two types: forward and reverse defect indications.

#### 3.1.1.2.1 Forward Defect Indication

This is used to signal a failure that is detected by a lower layer OAM mechanism. A server MEP (i.e. an actual or virtual MEP) transmits a Forward Defect Indication in a direction that is away from the direction of the failure (refer to Figure 3 below).

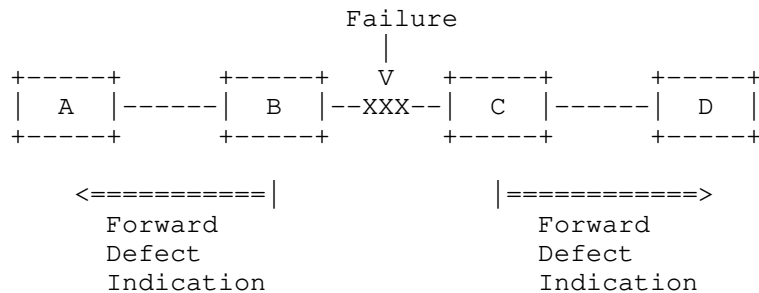


Figure 3: Forward Defect Indication

Forward defect indication may be used for alarm suppression and/or for purpose of inter-working with other layer OAM protocols. Alarm suppression is useful when a transport/network level fault translates to multiple service or flow level faults. In such a scenario, it is enough to alert a network management station (NMS) of the single transport/network level fault in lieu of flooding that NMS with a multitude of Service or Flow granularity alarms. EVPN PEs SHOULD support Forward Defect Indication in the Service OAM mechanisms.

#### 3.1.1.2.2 Reverse Defect Indication (RDI)

RDI is used to signal that the advertising MEP has detected a loss of continuity (LoC) defect. RDI is transmitted in the direction of the

failure (refer to Figure 4).

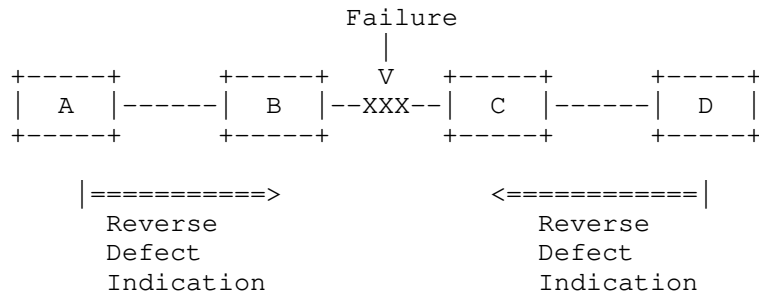


Figure 4: Reverse Defect Indication

RDI allows single-sided management, where the network operator can examine the state of a single MEP and deduce the overall health of a monitored service. EVPN PEs SHOULD support Reverse Defect Indication in the Service OAM mechanisms. This includes both the ability to signal LoC defect to a remote MEP, as well as the ability to recognize RDI from a remote MEP. It is worth noting that, in a multipoint MA, RDI is not a useful indicator of unidirectional fault. This is because RDI carries no indication of the affected MEP(s) with which the sender had detected a LoC defect.

### 3.1.2 On-Demand Fault Management Functions

On-demand fault management functions are initiated manually by the network operator and continue for a time bound period. These functions enable the operator to run diagnostics to investigate a defect condition.

#### 3.1.2.1 Connectivity Verification

EVPN Network OAM MUST support on-demand connectivity verification mechanisms for unicast and multicast destinations. The connectivity verification mechanisms SHOULD provide a means for specifying and carrying in the messages:

- variable length payload/padding to test MTU related connectivity problems.
- test frame formats as defined in Appendix C of [RFC2544] to detect potential packet corruption.

EVPN Network OAM MUST support connectivity verification at per flow

granularity. This includes both user flows (to test a specific path between PEs) as well as test flows (to test a representative path between PEs).

EVPN Service OAM MUST support connectivity verification on test flows and MAY support connectivity verification on user flows.

For multicast connectivity verification, EVPN Network OAM MUST support reporting on:

- the DF filtering status of specific port(s) or all the ports in a given bridge-domain.
- the Split Horizon filtering status of specific port(s) or all the ports in a given bridge-domain.

### 3.1.2.2 Fault Isolation

EVPN OAM MUST support an on-demand fault localization function. This involves the capability to narrow down the locality of a fault to a particular port, link or node. The characteristic of forward/reverse path asymmetry, in MPLS/IP, renders fault isolation into a direction-sensitive operation. That is, given two PEs A and B, localization of continuity failures between them requires running fault isolation procedures from PE A to PE B as well as from PE B to PE A.

EVPN Service OAM mechanisms only have visibility to the PEs but not the MPLS/IP P nodes. As such, they can be used to deduce whether the fault is in the customer's own network, the local CE-PE segment or remote CE-PE segment(s). EVPN Network and Transport OAM mechanisms can be used for fault isolation between the PEs and P nodes.

## 3.2 Performance Management

Performance Management functions can be performed both proactively and on-demand. Proactive management involves a recurring function, where the performance management probes are run continuously without a trigger. We cover both proactive and on-demand functions in this section.

### 3.2.1 Packet Loss

EVPN Network OAM SHOULD provide mechanisms for measuring packet loss for a given service.

Given that EVPN provides inherent support for multipoint-to-multipoint connectivity, then packet loss cannot be accurately measured by means of counting user data packets. This is because user packets can be delivered to more PEs or more ports than are necessary (e.g. due to broadcast, un-pruned multicast or unknown unicast flooding). As such, a statistical means of approximating packet loss rate is required. This can be achieved by sending "synthetic" OAM packets that are counted only by those ports (MEPs) that are required to receive them. This provides a statistical approximation of the number of data frames lost, even with multipoint-to-multipoint connectivity.

### 3.2.2 Packet Delay

EVPN Service OAM SHOULD support measurement of one-way and two-way packet delay and delay variation (jitter) across the EVPN network. Measurement of one-way delay requires clock synchronization between the probe source and target devices. Mechanisms for clock synchronization are outside the scope of this document. Note that Service OAM performance management mechanisms defined in [Y.1731] can be used.

EVPN Network OAM MAY support measurement of one-way and two-way packet delay and delay variation (jitter) across the EVPN network.

#### 4. Security Considerations

EVPN OAM must provide mechanisms for:

- Preventing denial of service attacks caused by exploitation of the OAM message channel.
- Optionally authenticate communicating endpoints (MEPs and MIPs)
- Preventing OAM packets from leaking outside of the EVPN network or outside their corresponding Maintenance Domain. This can be done by having MEPs implement a filtering function based on the Maintenance Level associated with received OAM packets.

#### 5. Acknowledgements

The authors would like to thank Gregory Mirsky for his thorough review of this work and invaluable comments.

#### 6. IANA Considerations

This document requires no IANA actions.

## Normative References

- [RFC792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, DOI 10.17487/RFC0792, September 1981, <<https://www.rfc-editor.org/info/rfc792>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI 10.17487/RFC5881, June 2010, <<https://www.rfc-editor.org/info/rfc5881>>.
- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, DOI 10.17487/RFC5883, June 2010, <<https://www.rfc-editor.org/info/rfc5883>>.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <<https://www.rfc-editor.org/info/rfc5884>>.
- [RFC6291] Andersson, L., van Helvoort, H., Bonica, R., Romascanu, D., and S. Mansfield, "Guidelines for the Use of the "OAM" Acronym in the IETF", BCP 161, RFC 6291, DOI 10.17487/RFC6291, June 2011, <<https://www.rfc-editor.org/info/rfc6291>>.
- [RFC6425] Saxena, S., Ed., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, DOI 10.17487/RFC6425, November 2011, <<https://www.rfc-editor.org/info/rfc6425>>.
- [RFC6428] Allan, D., Ed., Swallow, G., Ed., and J. Drake, Ed., "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, DOI 10.17487/RFC6428, November 2011, <<https://www.rfc-editor.org/info/rfc6428>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February



2015, <<https://www.rfc-editor.org/info/rfc7432>>.

- [RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <<https://www.rfc-editor.org/info/rfc7623>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <<https://www.rfc-editor.org/info/rfc8029>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<http://www.rfc-editor.org/info/rfc8174>>

#### Informative References

- [802.1Q] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks", 2014.
- [Y.1731] "ITU-T Recommendation Y.1731 (02/08) - OAM functions and mechanisms for Ethernet based networks", February 2008.
- [RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <<https://www.rfc-editor.org/info/rfc2544>>.
- [RFC5085] Nadeau, T., Ed., and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, DOI 10.17487/RFC5085, December 2007, <<https://www.rfc-editor.org/info/rfc5085>>.
- [RFC6136] Sajassi, A., Ed., and D. Mohan, Ed., "Layer 2 Virtual Private Network (L2VPN) Operations, Administration, and Maintenance (OAM) Requirements and Framework", RFC 6136, DOI 10.17487/RFC6136, March 2011, <<https://www.rfc-editor.org/info/rfc6136>>.

Authors' Addresses

Samer Salam  
Cisco

Email: [ssalam@cisco.com](mailto:ssalam@cisco.com)

Ali Sajassi  
Cisco  
170 West Tasman Drive  
San Jose, CA 95134, USA

Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Sam Aldrin  
Google, Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA USA

Email: [aldrin.ietf@gmail.com](mailto:aldrin.ietf@gmail.com)

John E. Drake  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089, USA

Email: [jdrake@juniper.net](mailto:jdrake@juniper.net)

Donald E. Eastlake, 3rd  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

Tel: +1-508-333-2270  
Email: [d3e3e3@gmail.com](mailto:d3e3e3@gmail.com)

