

L2VPN Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2014

Dave Allan, Jeff Tantsura
Ericsson

July 2013

mLDP extensions for integrating EVPN and multicast
draft-allan-l2vpn-mlbp-evpn-01

Abstract

This document describes how mLDP FECs can be encoded to support both service specific and shared multicast trees and describes the associated procedures for EVPN PEs. Thus, mLDP can implement multicast for EVPN.

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress".

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/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 2014.

Copyright and License Notice

Copyright (c) 2013 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. Authors.....	2
1.2. Requirements Language.....	2
2. Changes since last version.....	3
3. Conventions used in this document.....	3
3.1. Terminology.....	3
4. Solution Overview.....	4
5. Elements of Procedure.....	4
6. FEC Encoding.....	5
6.1. VLAN tagged FEC.....	5
6.2. I-SID tagged FEC.....	6
6.3. Shared FEC.....	6
7. Acknowledgements.....	7
8. Security Considerations.....	7
9. IANA Considerations.....	7
10. References.....	7
10.1. Normative References.....	7
10.2. Informative References.....	8
11. Authors' Addresses.....	8

1. Introduction

This document describes how mLDp FECs can be encoded to permit mLDp to implement multicast for EVpn. Such support can be applied to interconnecting 802.1ad, 802.1ah, 802.1aq, and 802.1Qbp based networks.

1.1. Authors

David Allan, Jeff Tantsura

1.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [1].

2. Changes since last version

- 1) Clarifications to the use of FEC encoding for RTs and VLANs added to section 6.

3. Conventions used in this document

3.1. Terminology

BCB: Backbone Core Bridge
BEB: Backbone Edge Bridge
BU: Broadcast/Unknown
B-MAC: Backbone MAC Address
B-VID: Backbone VLAN ID
CE: Customer Edge
C-MAC: Customer/Client MAC Address
DF: Designated Forwarder
ESI: Ethernet segment identifier
EVPN: Ethernet VPN
FEC: Forwarding Equivalence Class
ISIS-SPB: IS-IS as extended for SPB
I-SID: Backbone Service Instance ID
mLDP: Multicast Label Distribution Protocol
MP2MP: Multipoint to Multipoint
MVPN: Multicast VPN
NLRI: Network layer reachability information
PBBN: Provider Backbone Bridged Network
BEB-PE: Co located BEB and PE
PE: provider edge
P2MP: Point to Multipoint
P2P: Point to Point
RD: Route Distinguisher
SPB: Shortest path bridging
SPBM: Shortest path bridging MAC mode
VID: VLAN ID

VLAN: Virtual LAN

4. Solution Overview

mLDP[6] permits arbitrary FEC encodings for the naming of multicast trees to be defined. This property is leveraged to permit both service specific trees and shared trees to be utilized to augment EVPN unicast connectivity with network based multicast and avoid the inefficiencies of edge replication.

The flooding of EVPN BGP NLRI and ISIS-SPB [7] provides each PE with sufficient information to self elect as a DF, have knowledge of peer DFs, and from that construct the identifiers for the required set of multicast trees to support the current service set, which can then be encoded as mLDP FECs, and used to originate label mapping and label withdraw messages.

Both p2mp and mp2mp trees are supported with different FEC encodings for each. Service specific tree FECs encode the VID or I-SID associated with the service instance in the subtending network. Shared tree FECs encode a sorted list of the IP addresses of the leaf DFs.

5. Elements of Procedure

A PE advertises whether or not it supports shared tree (actual mechanism is TBD). Support of both shared and service specific trees is mandatory. Whether a PE supports shared trees is a network design decision.

A PE is expected to maintain a list of current multicast memberships.

A PE, upon receipt of new information from BGP or ISIS-SPB:

- 1) Evaluates it"s DF roles (as described in [5]).
- 2) On the basis of the PE"s DF role, determines the set of services it needs to support.
- 3) Determines the set of peer DFs for each service.
- 4) On the basis of requisite tree types and ESI multicast registrations (p2mp or mp2mp/service specific or shared), determines the name of the multicast tree needed for the service.

For example an ESI may only have source interest in an ISIS-SPB I-SID in which case it would:

- require a p2mp tree to the set of DFs registering receive interest in the I-SID for p2mp trees

- require an upstream label mapping to the set of DFs registering receive interest in the I_SID for mp2mp trees

5) Upon completion of evaluating the set of services, de-duplicates the required tree membership list.

6) Compares the required list with the existing list, and originates the necessary label mapping and label withdraw transactions to the network state up to date.

7) Configures the dataplane for the appropriate service to multicast tree bindings.

6. FEC Encoding

6.1. VLAN tagged FEC

VLAN tagged FEC uses the mLDP p2mp (0x06) type FEC and the mLDP mp2mp downstream (0x07) and upstream FECs (0x08)

The encoding of the opaque value is:

0																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	2	3	4	5	6	7	8	9																								
Type "x"																Length																<unused = 0>																															
																RT																																															
Ethertype																VID																= 0																															

Where:

- RT is the route target for the EVPN instance
- Ethertype identifies the tag type (C 0x8100, S or B 0x88a8)

- VID is the VLAN ID tag value. If the VID=0, then this is the default MDT for the RT and how VLAN unaware RTs are encoded, else it permits MDTs to be defined for VLAN aware services.

6.2. I-SID tagged FEC

The encoding of the opaque value is:

```

0                               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 "x+1" | Length |                                     | <unused = 0> |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     RT                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     I-SID                                     | <unused = 0> |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Where:

- RT is the route target for the EVPN instance
- I-SID corresponds to the I-SID that will use the tree

6.3. Shared FEC

The encoding of the opaque value is:

```

0                               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 "x+2" | Length |                                     | <unused = 0> |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     RT                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     ~                                     ~
|                                     <sorted list of DF ip addresses>
|                                     ~                                     ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Where:

- RT is the route target for the EVPN instance
- Sorted list of DF addresses identifies the set of leaves that have registered interest in one or more Ethernet services (either C/S or I tagged).

7. Acknowledgements

The authors would like to thank Panagiotis Saltsidis, Jakob Heitz and Janos Farkas for their detailed review of this draft.

8. Security Considerations

For a future version of this document.

9. IANA Considerations

For a future version of this document.

10. References

10.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] Fedyk et.al. "IS-IS Extensions Supporting IEEE 802.1aq Shortest Path Bridging", IETF RFC 6329, April 2012
- [3] Rosen et.al., "BGP/MPLS IP Virtual Private Networks (VPNs)", IETF RFC 4364, February 2006
- [4] Aggarwal et.al. "BGP MPLS Based Ethernet VPN", IETF work in progress, draft-ietf-l2vpn-evpn-01, July 2012
- [5] Allan et.al. "802.1aq and 802.1Qbp Support over EVPN", IETF work in progress, draft-allan-l2vpn-spb-evpn-03, February 2013
- [6] Wijnands et.al. "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths". IETF RFC 6388, November 2011

10.2. Informative References

- [7] IEEE 802.1aq "IEEE Standard for Local and Metropolitan Area Networks: Bridges and Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging", June 2012
- [8] IEEE 802.1Qbp "Draft IEEE Standard for Local and Metropolitan Area Networks---Virtual Bridged Local Area Networks - Amendment: Equal Cost Multiple Paths (ECMP), 802.1Qbp", draft 1.3, February 2013
- [9] Sajassi et.al. "PBB E-VPN", IETF work in progress, draft-ietf-l2vpn-pbb-evpn-03, June 2012
- [10] IEEE 802.1Q-2011 "IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", August 2011

11. Authors' Addresses

Dave Allan (editor)
Ericsson
300 Holger Way
San Jose, CA 95134
USA
Email: david.i.allan@ericsson.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
Email: jeff.tantsura@ericsson.com

L2VPN Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2014

Dave Allan, Jeff Tantsura
Ericsson
Don Fedyk
Alcatel-Lucent
Ali Sajassi
Cisco

July 2013

802.1aq Support over EVPN
draft-allan-l2vpn-spbm-evpn-04

Abstract

This document describes how Ethernet Shortest Path Bridging MAC mode (802.1aq) can be combined with EVPN in a way that interworks with PBB-PEs as described in the PBB-EVPN solution in a way that permits operational isolation of each Ethernet network subtending an EVPN core while supporting full interworking between the different variations of Ethernet operation.

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress".

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/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 2014.

Copyright and License Notice

Copyright (c) 2013 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.....	3
1.1. Authors.....	3
1.2. Requirements Language.....	3
2. Conventions used in this document.....	3
2.1. Terminology.....	3
3. Changes since previous version.....	4
4. Solution Overview.....	4
5. Elements of Procedure.....	5
5.1. PE Configuration.....	5
5.2. DF Election.....	6
5.3. Control plane interworking ISIS-SPB to EVPN.....	6
5.4. Control plane interworking EVPN to ISIS-SPB.....	7
5.5. Data plane Interworking 802.1aq SPBM island or PBB-PE to EVPN.....	8
5.6. Data plane Interworking EVPN to 802.1aq SPBM island.....	8
5.7. Data plane interworking EVPN to 802.1ah PBB-PE.....	8
5.8. Multicast Support.....	8
6. Other Aspects.....	8
6.1. Flow Ordering.....	8
6.2. Transit.....	9
7. Acknowledgements.....	9
8. Security Considerations.....	9
9. IANA Considerations.....	9
10. References.....	9
10.1. Normative References.....	9
10.2. Informative References.....	9
11. Authors' Addresses.....	10

1. Introduction

This document describes how Ethernet Shortest Path Bridging MAC mode (802.1aq) along with PBB-PEs and PBBNs (802.1ah) can be supported by EVPN such that each island is operationally isolated while providing full L2 connectivity between them. Each island can use its own control plane instance and multi-pathing design, be it multiple ECT sets, or multiple spanning trees.

The intention is to permit both past, current and emerging future versions of Ethernet to be seamlessly integrated to permit large scale, geographically diverse numbers of Ethernet end systems to be fully supported with EVPN as the unifying agent.

1.1. Authors

David Allan, Jeff Tantsura, Don Fedyk, Ali Sajassi

1.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [1].

2. Conventions used in this document

2.1. Terminology

BCB: Backbone Core Bridge
BEB: Backbone Edge Bridge
BU: Broadcast/Unknown
B-MAC: Backbone MAC Address
B-VID: Backbone VLAN ID
CE: Customer Edge
C-MAC: Customer/Client MAC Address
DF: Designated Forwarder
ESI: Ethernet Segment Identifier
EVPN: Ethernet VPN
ISIS-SPB: IS-IS as extended for SPB
I-SID: I-Component Service ID
MP2MP: Multipoint to Multipoint

MVPN: Multicast VPN
 NLRI: Network Layer Reachability Information
 PBBN: Provider Backbone Bridged Network
 PBB-PE: Co located BEB and PE
 PE: provider edge
 P2MP: Point to Multipoint
 P2P: Point to Point
 RD: Route Distinguisher
 SPB: Shortest path bridging
 SPBM: Shortest path bridging MAC mode

3. Changes since previous version

- 1) Removal of reference to 802.1Qbp. This will be addressed in separate document.
- 2) Determining ESI value exclusively requires configuration. This was an open item in previous drafts.

4. Solution Overview

The EVPN solution for 802.1aq SPBM incorporates control plane interworking in the PE to map ISIS-SPB [2] information elements into the EVPN NLRI information and vice versa. This requires each PE to act both as an EVPN BGP speaker and as an ISIS-SPB edge node. Associated with this are procedures for configuring the forwarding operations of the PE such that an arbitrary number of EVPN subtending SPBM islands may be interconnected without any topological or multipathing dependencies. This model also permits PBB-PEs as defined in draft-l2vpn-pbb-evpn-02[8] to seamlessly communicate with the SPB islands.

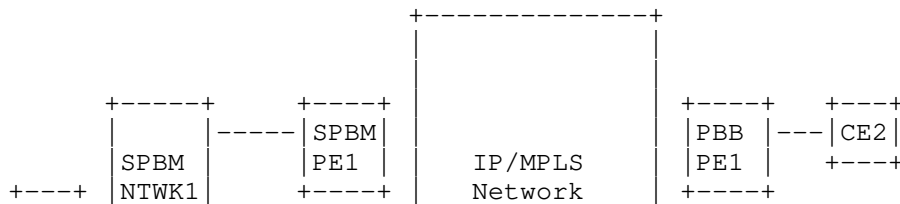




Figure 1: PBB and SPBM EVPN Network

Each EVPN is identified by a route target. The route target identifies the set of SPBM islands and BEB-PEs that are allowed to communicate. Each SPBM island is administered to have an associated Ethernet Segment ID (ESI) associated with it. This manifests itself as a set of Ethernet segments, where each Ethernet segment ID is unique within the route target.

BGP acts as a common repository of the I-SID attachment points for the set of subtending PEs/SPBM islands. This is in the form of B-MAC address/I-SID/Tx-Rx-attribute tuples. BGP leaks I-SID information into each SPBM island on the basis of locally registered interest. If an SPBM island has no BEBs registering interest in an I-SID, information about that I-SID from other SPBM islands, PBB-PEs or PBBNs will not be leaked into the local ISIS-SPB routing system.

For each B-VID in an SPBM island, a single SPBM-PE is elected the designated forwarder for the B-VID. An SPBM-PE may be a DF for more than one B-VID. This is described further in section 4.2. The SPBM-PE originates IS-IS advertisements as if it were an I-BEB or IB-BEB that proxy for the other SPBM islands and PBB PEs in the EVPN defined by the route target, but the PE typically will not actually host any I-components.

An SPBM-PE that is a DF for a B-VID strips the B-VID tag information from frames relayed towards the EVPN. The DF also inserts the appropriate B-VID tag information into frames relayed towards the SPBM island on the basis of the local I-SID/B-VID bindings advertised in ISIS-SPB.

5. Elements of Procedure

5.1. PE Configuration

At SPBM island commissioning a PE is configured with:

- 1) The route target for the service instance. Where a route target is defined as identifying the set of SPBM islands, PBBNs and PBB-PEs to be interconnected by the EVPN.
- 2) The unique ESI for the SPBM island.

And the following is configured as part of commissioning an ISIS-SPB node:

- 1) A Shortest Path Source ID (SPSourceID) used for algorithmic construction of multicast DA addresses. Note this is required for SPBM BEBs independent of the EVPN operation.
- 2) The set of VLANs (identified by B-VIDs) used in the SPBM island and multi-pathing algorithm IDs to use. The B-VID may be different in different domains and may be removed as carried over the IP/MPLS network.

A type-1 Route Distinguisher (RD) for the node can be auto-derived. This will be described in a future version of the document.

5.2. DF Election

PEs self appoint in the role of DF for a B-VID for a given SPBM island. The procedure used is as per section 9.5 of draft-ietf-l2vpn-evpn-03[4] "DF election".

5.3. Control plane interworking ISIS-SPB to EVPN

When a PE receives an SPBM service identifier and unicast address sub-TLV as part of an ISIS-SPB MT capability TLV it checks if it is the DF for the B-VID in the sub-TLV.

If it is the DF, and there is new or changed information then a MAC advertisement route NLRI is created for each new I-SID in the sub-TLV.

- the Route Distinguisher (RD) is set to that of the PE.
- the ESI is that of the SPBM island.
- the Ethernet tag ID contains the I-SID (including the Tx/Rx attributes). The encoding of I-SID information is as per figure 2.

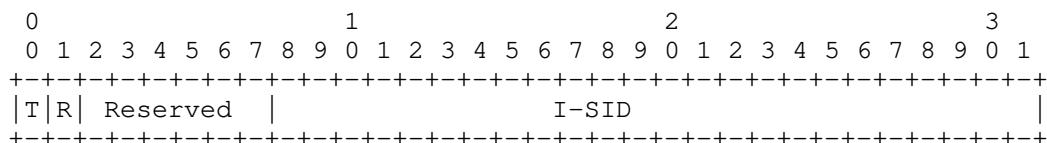


Figure 2: I-SID encoding in the Ethernet tag-ID field

- the MAC address from the sub-TLV
- an MPLS label

Similarly in the scenario where a PE became elected DF for a B-VID in an operating network, the IS-IS database would be processed in order to construct the NLRI information associated with the new role of the PE.

If the BGP database has NLRI information for the I-SID, and this is the first instance of registration of interest in the I-SID from the SPB island, the NLRI information with that tag is processed to construct an updated set of SPBM service identifier and unicast address sub-TLVs to be advertised by the PE.

The ISIS-SPB information is also used to keep a local table indexed by I-SID current to indicate the associated B-VID for processing of frames received from EVPN. When an I-SID is associated with more than one B-VID, only one entry is allowed in the table. Rules for this will be in a future version of the document.

5.4. Control plane interworking EVPN to ISIS-SPB

When a PE receives a BGP NLRI that has new information, it checks if the I-SID in the Ethernet Tag ID locally maps to the B-VID that are an elected DF. Note that if no BEBs in the SPB island have advertised any interest in the I-SID, it will not be associated with any B-VID locally, and therefore not of interest. If the I-SID is of local interest to the SPBM island and the PE is the DF for the B-VID that that I-SID is locally mapped to, a SPBM service identifier and unicast address sub-TLV is constructed/updated for advertisement into ISIS-SPB.

The NLRI information advertised into ISIS-SPB is also used to locally populate a forwarding table indexed by B-MAC+I-SID that points to the

label stack associated with the SPBM frame. The bottom label in the stack being that offered in the NLRI.

5.5. Data plane Interworking 802.1aq SPBM island or PBB-PE to EVPN

When an PE receives a frame from the SPBM island in a B-VID for which it is a DF, it looks up the B-MAC/I-SID information to determine the label stack to be added to the frame for forwarding in the EVPN. The PE strips the B-VID information from the frame, adds the label information to the frame and forwards the resulting MPLS packet.

5.6. Data plane Interworking EVPN to 802.1aq SPBM island

When a PE receives a packet from the EVPN it may infer the B-VID to overwrite in the SPBM frame from the I-SID or by other means (such as via the bottom label in the MPLS stack).

If the frame has a local multicast DA, it overwrites the SPsourceID in the frame with the local SPsourceID.

5.7. Data plane interworking EVPN to 802.1ah PBB-PE

A PBB-PE actually has no subtending PBBN nor concept of B-VID so no frame processing is required.

A PBB-PE is required to accept SPBM encoded multicast DAs as if they were 802.1ah encoded multicast DAs. The only information of interest being that it is a multicast frame, and the I-SID encoded in the lower 24 bits.

5.8. Multicast Support

Refer to "mLDP extensions for integrating EVPN and multicast"[5].

6. Other Aspects

6.1. Flow Ordering

When per I-SID multicast is implemented via PE replication, a stable network will preserve frame ordering between known unicast and BU traffic (e.g. race conditions will not exist). This cannot be guaranteed when multicast is used in the EVPN.

6.2. Transit

Any PE that does not need to participate in the tandem calculations may use the IS-IS overload bit to exclude SPBM tandem paths and behave as pure interworking platform (I-BEB).

7. Acknowledgements

The authors would like to thank Peter Ashwood-Smith, Martin Julien and Janos Farkas for their detailed review of this draft.

8. Security Considerations

For a future version of this document.

9. IANA Considerations

For a future version of this document.

10. References

10.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] Fedyk et.al. "IS-IS Extensions Supporting IEEE 802.1aq Shortest Path Bridging", IETF RFC 6329, April 2012
- [3] Rosen et.al., "BGP/MPLS IP Virtual Private Networks (VPNs)", IETF RFC 4364, February 2006
- [4] Aggarwal et.al. "BGP MPLS Based Ethernet VPN", IETF work in progress, draft-ietf-l2vpn-evpn-02, October 2012
- [5] Allan et.al. "mLDP extensions for integrating EVPN and multicast", IETF work in progress draft-allan-l2vpn-mldp-evpn-01, May 2013

10.2. Informative References

- [6] 802.1aq(2012) IEEE Standard for Local and Metropolitan Area Networks: Bridges and Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging
- [7] Sajassi et.al. "PBB E-VPN", IETF work in progress, draft-ietf-l2vpn-pbb-evpn-04, February 2013

- [8] 802.1Q (2011) IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks

11. Authors' Addresses

Dave Allan (editor)
Ericsson
300 Holger Way
San Jose, CA 95134
USA
Email: david.i.allan@ericsson.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
Email: jeff.tantsura@ericsson.com

Don Fedyk
Alcatel-Lucent
Groton, MA 01450
USA
EMail: Donald.Fedyk@alcatel-lucent.com

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.com

Network Working Group
INTERNET-DRAFT
Category: Standards Track

Sami Boutros
Ali Sajassi
Samer Salam
Dennis Cai
Samir Thoria
Cisco

John Drake
Juniper

Expires: January 16, 2014

July 16, 2013

VXLAN DCI Using EVPN
draft-boutros-l2vpn-vxlan-evpn-02.txt

Abstract

This document describes how Ethernet VPN (EVPN) technology can be used to interconnect VXLAN or NVGRE networks over an MPLS/IP network. This is to provide intra-subnet connectivity at Layer 2 and control-plane separation among the interconnected VXLAN or NVGRE networks. The scope of the learning of host MAC addresses in VXLAN or NVGRE network is limited to data plane learning in this document.

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) 2013 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	3
1.1	Terminology	3
2.	Requirements	3
2.1.	Control Plane Separation among VXLAN/NVGRE Networks	3
2.2	All-Active Multi-homing	4
2.3	Layer 2 Extension of VNIs/VSIDs over the MPLS/IP Network	4
2.4	Support for Integrated Routing and Bridging (IRB)	4
3.	Solution Overview	4
3.1.	Redundancy and All-Active Multi-homing	5
4.	EVPN Routes	6
4.1.	BGP MAC Advertisement Route	6
4.2.	Ethernet Auto-Discovery Route	7
4.3.	Per VPN Route Targets	7
4.4	Inclusive Multicast Route	7
4.5.	Unicast Forwarding	7
4.6.	Handling Multicast	8
4.6.2.	Multicast Stitching with Per-VNI Load Balancing	9
5.	NVGRE	9
6.	Acknowledgements	10
7.	Security Considerations	10
8.	IANA Considerations	10
9.	References	10
9.1	Normative References	10
9.2	Informative References	10
	Authors' Addresses	10

1 Introduction

[EVPN] introduces a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP control plane over the core MPLS/IP network. [VXLAN] defines a tunneling scheme to overlay Layer 2 networks on top of Layer 3 networks. [VXLAN] allows for optimal forwarding of Ethernet frames with support for multipathing of unicast and multicast traffic. VXLAN uses UDP/IP encapsulation for tunneling.

In this document, we discuss how Ethernet VPN (EVPN) technology can be used to interconnect VXLAN or NVGRE networks over an MPLS/IP network. This is achieved by terminating the VxLAN tunnel at the hand-off points, performing data plane MAC learning of customer traffic and providing intra-subnet connectivity for the customers at Layer 2 across the MPLS/IP core. The solution maintains control-plane separation among the interconnected VXLAN or NVGRE networks. The scope of the learning of host MAC addresses in VXLAN or NVGRE network is limited to data plane learning in this document. The distribution of MAC addresses in control plane using BGP in VXLAN or NVGRE network is outside of the scope of this document and it is covered in [EVPN-OVERLY].

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].

LDP: Label Distribution Protocol
MAC: Media Access Control
MPLS: Multi Protocol Label Switching
NVO: Network Virtualization Overlay
NVE: NVO Endpoint
OAM: Operations, Administration and Maintenance
PE: Provide Edge Node
PW: PseudoWire
TLV: Type, Length, and Value
VPLS: Virtual Private LAN Services
VXLAN: Virtual eXtensible Local Area Network
VTEP: VXLAN Tunnel End Point
VNI: VXLAN Network Identifier (or VXLAN Segment ID)
ToR: Top of Rack switch

2. Requirements

2.1. Control Plane Separation among VXLAN/NVGRE Networks

It is required to maintain control-plane separation for the underlay networks (e.g., among the various VXLAN/NVGRE networks) being interconnected over the MPLS/IP network. This ensures the following characteristics:

- scalability of the IGP control plane in large deployments and fault domain localization, where link or node failures in one site do not trigger re-convergence in remote sites.
- scalability of multicast trees as the number of interconnected networks scales.

2.2 All-Active Multi-homing

It is important to allow for all-active multi-homing of the VXLAN/NVGRE network to MPLS/IP network where traffic from a VTEP can arrive at any of the PEs and can be forwarded accordingly over the MPLS/IP network. Furthermore, traffic destined to a VTEP can be received over the MPLS/IP network at any of the PEs connected to the VXLAN/NVGRE network and be forwarded accordingly. The solution MUST support all-active multi-homing to an VXLAN/NVGRE network.

2.3 Layer 2 Extension of VNIs/VSIDs over the MPLS/IP Network

It is required to extend the VXLAN VNIs or NVGRE VSIDs over the MPLS/IP network to provide intra-subnet connectivity between the hosts (e.g. VMs) at Layer 2.

2.4 Support for Integrated Routing and Bridging (IRB)

The data center WAN edge node is required to support integrated routing and bridging in order to accommodate both inter-subnet routing and intra-subnet bridging for a given VNI/VSID. For example, inter-subnet switching is required when a remote host connected to an enterprise IP-VPN site wants to access an application resided on a VM.

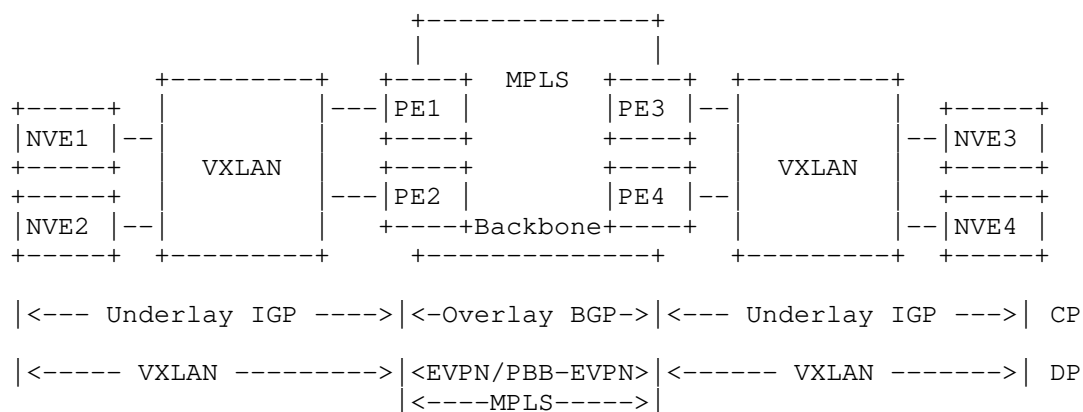
3. Solution Overview

Every VXLAN/NVGRE network, which is connected to the MPLS/IP core, runs an independent instance of the IGP control-plane. Each PE participates in the IGP control plane instance of its VXLAN/NVGRE network.

Each PE node terminates the VXLAN or NVGRE data-plane encapsulation where each VNI or VSID is mapped to a bridge-domain. The PE performs data plane MAC learning on the traffic received from the VXLAN/NVGRE network.

Each PE node implements EVPN or PBB-EVPN to distribute in BGP either the client MAC addresses learnt over the VXLAN tunnel in case of EVPN, or the PE's B-MAC addresses in case of PBB-EVPN. In the PBB-EVPN case, client MAC addresses will continue to be learnt in data plane.

Each PE node would encapsulate the Ethernet frames with MPLS when sending the packets over the MPLS core and with the VXLAN or NVGRE tunnel header when sending the packets over the VXLAN or NVGRE Network.



Legend: CP = Control Plane View

DP = Data Plane View

Figure 1: Interconnecting VXLAN Networks with VXLAN-EVPN

3.1. Redundancy and All-Active Multi-homing

When a VXLAN network is multi-homed to two or more PEs, and provided that these PEs have the same IGP distance to a given NVE, the solution MUST support load-balancing of traffic between the NVE and the MPLS network, among all the multi-homed PEs. This maximizes the use of the bisectional bandwidth of the VXLAN network. One of the main capabilities of EVPN/PBB-EVPN is the support for all-active multi-homing, where the known unicast traffic to/from a multi-homed site can be forwarded by any of the PEs attached to that site. This ensures optimal usage of multiple paths and load balancing. EVPN/PBB-EVPN, through its DF election and split-horizon filtering mechanisms, ensures that no packet duplication or forwarding loops result in such scenarios. In this solution, the VXLAN network is treated as a multi-homed site for the purpose of EVPN operation.

Since the context of this solution is VXLAN networks with data-plane

learning paradigm, it is important for the multi-homing mechanism to ensure stability of the MAC forwarding tables at the NVEs, while supporting all-active forwarding at the PEs. For example, in Figure 1 above, if each PE uses a distinct IP address for its VTEP tunnel, then for a given VNI, when an NVE learns a host's MAC address against the originating VTEP source address, its MAC forwarding table will keep flip-flopping among the VTEP addresses of the local PEs. This is because a flow associated with the same host MAC address can arrive at any of the PE devices. In order to ensure that there is no flip/flopping of MAC-to-VTEP address associations, an IP Anycast address MUST be used as the VTEP address on all PEs multi-homed to a given VXLAN network. The use of IP Anycast address has two advantages:

- a) It prevents any flip/flopping in the forwarding tables for the MAC-to-VTEP associations
- b) It enables load-balancing via ECMP for DCI traffic among the multi-homed PEs

In the baseline [EVPN] draft, the all-active multi-homing is described for a multi-homed device (MHD) using [LACP] and the single-active multi-homing is described for a multi-homed network (MHN) using [802.1Q]. In this draft, the all-active multi-homing is described for a VXLAN MHN. This implies some changes to the filtering which will be described in details in the multicast section (Section 4.6.2).

The filtering used for BUM traffic of all-active multi-homing in [EVPN] is asymmetric; where the BUM traffic from the MPLS/IP network towards the multi-homed site is filtered on non-DF PE(s) and it passes thorough the DF PE. There is no filtering of BUM traffic originating from the multi-homed site because of the use of Ethernet Link Aggregation: the MHD hashes the BUM traffic to only a single link. However, in this solution because BUM traffic can arrive at both PEs in both core-to-site and site-to-core directions, the filtering needs to be symmetric just like the filtering of BUM traffic for single-active multi-homing (on a per service instance/VLAN basis).

4. EVPN Routes

This solution leverages the same BGP Routes and Attributes defined in [EVPN], adapted as follows:

4.1. BGP MAC Advertisement Route

This route and its associated modes are used to distribute the customer MAC addresses learnt in data plane over the VXLAN tunnel in case of EVPN. Or can be used to distribute the provider Backbone MAC addresses in case of PBB-EVPN.

In case of EVPN, the Ethernet Tag ID of this route is set to zero for VNI-based mode, where there is one-to-one mapping between a VNI and an EVI. In such case, there is no need to carry the VNI in the MAC advertisement route because BD ID can be derived from the RT associated with this route. However, for VNI-aware bundle mode, where there is multiple VNIs can be mapped to the same EVI, the Ethernet Tag ID MUST be set to the VNI. At the receiving PE, the BD ID is derived from the combination of RT + VNI - e.g., the RT identifies the associated EVI on that PE and the VNI identifies the corresponding BD ID within that EVI.

4.2. Ethernet Auto-Discovery Route

When EVPN is used, the application of this route is as specified in [EVPN]. However, when PBB-EVPN is used, there is no need for this route per [PBB-EVPN].

4.3. Per VPN Route Targets

VXLAN-EVPN uses the same set of route targets defined in [EVPN].

4.4 Inclusive Multicast Route

The EVPN Inclusive Multicast route is used for auto-discovery of PE devices participating in the same tenant virtual network identified by a VNI over the MPLS network. It also enables the stitching of the IP multicast trees, which are local to each VXLAN site, with the Label Switched Multicast (LSM) trees of the MPLS network.

The Inclusive Multicast Route is encoded as follow:

- Ethernet Tag ID is set to zero for VNI-based mode and to VNI for VNI-aware bundle mode.

- Originating Router's IP Address is set to one of the PE's IP addresses.

All other fields are set as defined in [EVPN].

Please see section 4.6 "Handling Multicast"

4.5. Unicast Forwarding

Host MAC addresses will be learnt in data plane from the VXLAN network and associated with the corresponding VTEP identified by the source IP address. Host MAC addresses will be learnt in control plane if EVPN is implemented over the MPLS/IP core, or in the data-plane if PBB-EVPN is implemented over the MPLS core. When Host MAC addresses are learned in data plane over MPLS/IP core [in case of PBB-EVPN], they are associated with their corresponding BMAC addresses.

L2 Unicast traffic destined to the VXLAN network will be encapsulated with the IP/UDP header and the corresponding customer bridge VNI.

L2 Unicast traffic destined to the MPLS/IP network will be encapsulated with the MPLS label.

4.6. Handling Multicast

Each VXLAN network independently builds its P2MP or MP2MP shared multicast trees. A P2MP or MP2MP tree is built for one or more VNIs local to the VXLAN network.

In the MPLS/IP network, multiple options are available for the delivery of multicast traffic:

- Ingress replication
- LSM with Inclusive trees
- LSM with Aggregate Inclusive trees
- LSM with Selective trees
- LSM with Aggregate Selective trees

When LSM is used, the trees are P2MP.

The PE nodes are responsible for stitching the IP multicast trees, on the access side, to the ingress replication tunnels or LSM trees in the MPLS/IP core. The stitching must ensure that the following characteristics are maintained at all times:

1. Avoiding Packet Duplication: In the case where the VXLAN network is multi-homed to multiple PE nodes, if all of the PE nodes forward the same multicast frame, then packet duplication would arise. This applies to both multicast traffic from site to core as well as from core to site.

2. Avoiding Forwarding Loops: In the case of VXLAN network multi-homing, the solution must ensure that a multicast frame forwarded by a given PE to the MPLS core is not forwarded back by another PE (in the same VXLAN network) to the VXLAN network of origin. The same applies for traffic in the core to site direction.

The following approach of per-VNI load balancing can guarantee proper

stitching that meets the above requirements.

4.6.2. Multicast Stitching with Per-VNI Load Balancing

To setup multicast trees in the VXLAN network for DC applications, PIM Bidir can be of special interest because it reduces the amount of multicast state in the network significantly. Furthermore, it alleviates any special processing for RPF check since PIM Bidir doesn't require any RPF check. The RP for PIM Bidir can be any of the spine nodes. Multiple trees can be built (e.g., one tree rooted per spine node) for efficient load-balancing within the network. All PEs participating in the multi-homing of the VXLAN network join all the trees. Therefore, for a given tree, all PEs receive BUM traffic. DF election procedures of [EVPN] are used to ensure that only traffic to/from a single PE is forwarded, thus avoiding packet duplications and forwarding loops. For load-balancing of BUM traffic, when a PE or an NVE wants to send BUM traffic over the VXLAN network, it selects one of the trees based on its VNI and forwards all the traffic for that VNI on that tree. PIM SM will be described in future revision of this draft.

Multicast traffic from VXLAN/NVGRE is first subjected to filtering based on DF election procedures of [EVPN] using the VNI as the Ethernet Tag. This is similar to filtering in [EVPN] in principal; however, instead of VLAN ID, VNI is used for filtering, and instead of being 802.1Q frame, it is a VXLAN encapsulated packet. On the DF PE, where the multicast traffic is allowed to be forwarded, the VNI is used to select a bridge domain,. After the packet is de-capsulated, an L2 lookup is performed based on host MAC DA. It should be noted that the MAC learning is performed in data-plane for the traffic received from the VXLAN/NVGRE network and the host MAC SA is learnt against the source VTEP address.

The PE nodes, connected to a multi-homed VXLAN network, perform BGP DF election to decide which PE node is responsible for forwarding multicast traffic associated with a given VNI. A PE would forward multicast traffic for a given VNI only when it is the DF for this VNI. This forwarding rule applies in both the site-to-core as well as core-to-site directions.

5. NVGRE

Just like VXLAN, all the above specification would apply for NVGRE, replacing the VNI with Virtual Subnet Identifier (VSID) and the VTEP with NVGRE Endpoint.

6. Acknowledgements

TBD.

7. Security Considerations

There are no additional security aspects that need to be discussed here.

8. IANA Considerations

TBD.

9. References

9.1 Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt, work in progress, February, 2012.

[TRILL] Sajassi et al., TRILL-EVPN draft-ietf-l2vpn-trill-evpn-00, work in progress, June 2012.

[VXLAN] Mahalingam, Dutt et al., A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks draft-mahalingam-dutt-dcops-vxlan-02.txt, work in progress, August, 2012.

[NVGRE] Sridharan et al., Network Virtualization using Generic Routing Encapsulation draft-sridharan-virtualization-nvgre-01.txt, work in progress, July, 2012.

Authors' Addresses

Sami Boutros
Cisco
EMail: sboutros@cisco.com

Ali Sajassi
Cisco
EMail: sajassi@cisco.com

Samer Salam

Cisco
EMail: ssalam@cisco.com

Dennis Cai
Cisco
EMail: dcai@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

Samir Thoria
Cisco
EMail: sthoria@cisco.com

L2VPN

Internet Draft
Intended status: Standards Track
Expires: December 2013

Weiguo Hao
Yizhou Li
Pei Xu
Huawei
June 14, 2013

Multi-homed network in EVPN
draft-hao-l2vpn-evpn-mhn-00.txt

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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 December 14, 2013.

Copyright Notice

Copyright (c) 2013 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.

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.

Abstract

To enhance the reliability, bridged network is normally multi-homed to an EVPN network, there are two categories of mechanisms to avoid the layer 2 traffic loop. The first category does not require the PEs participating in the control protocol of the bridged network, while the second category requires that. [EVPN] described one of the first category mechanisms called designated forwarder (DF) election to achieve loop avoidance and vlan-based load balancing. This draft mainly focuses on the second category of mechanisms which can achieve intra-vlan MAC-based load balancing. MAC-based VLAN balancing is more applicable than DF election mechanism if all end stations in bridged network are on the same VLAN which can cause traffic congestion in DF link.

Table of Contents

1. Introduction	3
2. Conventions used in this document.....	4
3. Recap on Designated Forwarder (DF) election mechanism.....	4
4. Active/Active MAC-based load balancing mechanism	6
4.1. Emulated MSTP root bridge solution	7
4.2. Bridge control plane protocol tunneling solution.....	8
4.2.1. Scenario 1: Local bridged network is MSTP.....	10
4.2.2. Scenario 2: Local bridged network is G.8032.....	10
4.2.3. Fast convergence.....	10
5. EVPN protocol extension.....	11
6. Security Considerations.....	11
7. IANA Considerations	11
7.1. Normative References.....	12
7.2. Informative References.....	12
8. Acknowledgments	12

1. Introduction

[EVPN] introduces a solution for multipoint L2VPN services. In EVPN networks, MAC learning between PEs is not via the data plane (different from what happens in traditional bridging network) but via the control plane using multi-protocol (MP) BGP.

To enhance the reliability, the PE nodes need offer multi-homed connectivity to a CE or access Network, i.e., both multi-homed device (MHD) as well as multi-homed network (MHN) scenarios in [EVPN-REQ] should be covered by E-VPN solution. In MHN scenario, the multi-homed Ethernet network would typically run a resiliency mechanism such as Multiple Spanning Tree Protocol [802.1Q] or Ethernet Ring Protection Switching [G.8032]. For example, EVPN can be used for Data Center (DC) interconnection to provide LAN extension for each DC site and each site is an MSTP networks. Normally each site should be multi-homed to multiple EVPN PEs to ensure the reliability.

As defined in [EVPN-REQ], the following solutions should be provided for MHN scenario:

A solution MUST support multi-homed network connectivity with active/standby redundancy.

A solution MUST also support multi-homed network with active/active VLAN-based load balancing (i.e. disjoint VLAN sets active on disparate PEs).

A solution MAY support VLAN-based load balancing among PEs that are member of a redundancy group spanning multiple ASes.

A solution MAY support multi-homed network with active/active MAC-based load balancing (i.e. different MAC addresses on a VLAN are reachable via different PEs).

The former three requirements can be addressed through designated forwarder (DF) election mechanism as described in [EVPN], a brief review of DF election mechanism will be given in section 3.

This draft will mainly focus on a new mechanism to achieve active/active MAC-based load balancing to fulfil the fourth requirement. The details of the solution will be illustrated in section 4. Protocol extensions of EVPN for this mechanism will be given in section 5.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

This document uses the terminologies defined in [RFC6325] along with the following:

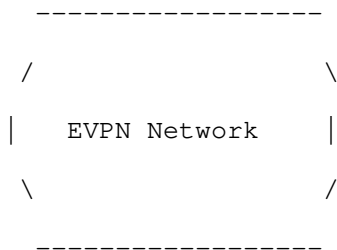
EVPN: Ethernet virtual private network.

G.8032: Ethernet ring protection switching.

NVO3: Network virtualization over layer3.

STP: Spanning Tree Protocol.

3. Recap on Designated Forwarder (DF) election mechanism



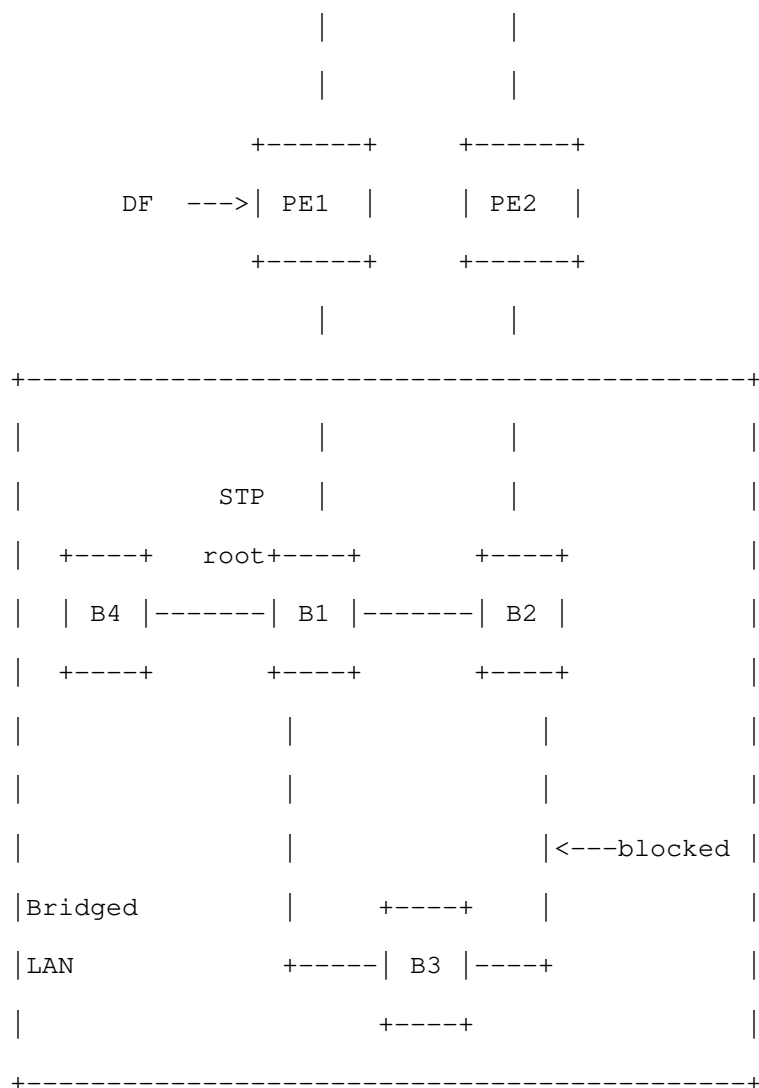


Figure 1 DF election mechanism

As described in [EVPN], designated forwarder (DF) mechanism is required for loop avoidance. Only one of the links between the switched bridged network and the PEs is active for a given Ethernet tag, as shown by Figure 1. This mechanism does not require the PEs to participate in the control protocol of the bridged network. Bridges in the local bridged network runs normal Multiple Spanning Tree Protocol [802.1Q] or Ethernet Ring Protection Switching [G.8032].

Through this method VLAN-based load balancing among PEs can be achieved. All end systems of one VLAN can access the EVPN network through only one PE.

In this case, the Ethernet A-D route per Ethernet segment MUST be advertised with the "Active-Standby" flag set to one. Only one PE is elected as DF for each EVI(E-VPN Instance). Only DF is responsible for sending multicast, broadcast and unknown unicast traffic, on a given Ethernet tag to the bridged network. In order to perform better traffic load-balancing within a given segment, multiple DFs per Ethernet segment can be elected and each PE is the DF for a disjoint set of EVIs. An EVI is an E-VPN routing and forwarding instance on a PE and consists of one or more broadcast domains which is identified by an Ethernet Tag which are assigned to the broadcast domains of a given E-VPN instance by the provider of that E-VPN. The information about an Ethernet Tag on a particular Ethernet segment is advertised using an "Ethernet Auto-Discovery route(Ethernet A-D route)". In the case of a multi-homed CE, this route MUST carry the "ESI Label Extended Community" to enable split horizon. Also, the route can be used for Designated Forwarder (DF) election and MAY be used to optimize the withdrawal of MAC addresses upon failure.

For fast convergence case, upon a failure in connectivity to the attached segment, the PE withdraws the corresponding Ethernet A-D route. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all MAC addresses associated with the Ethernet segment in question. If there is any other PE advertising an Ethernet A-D route for the same segment, the PE updates the next-hop adjacencies to point to this backup PE(s).

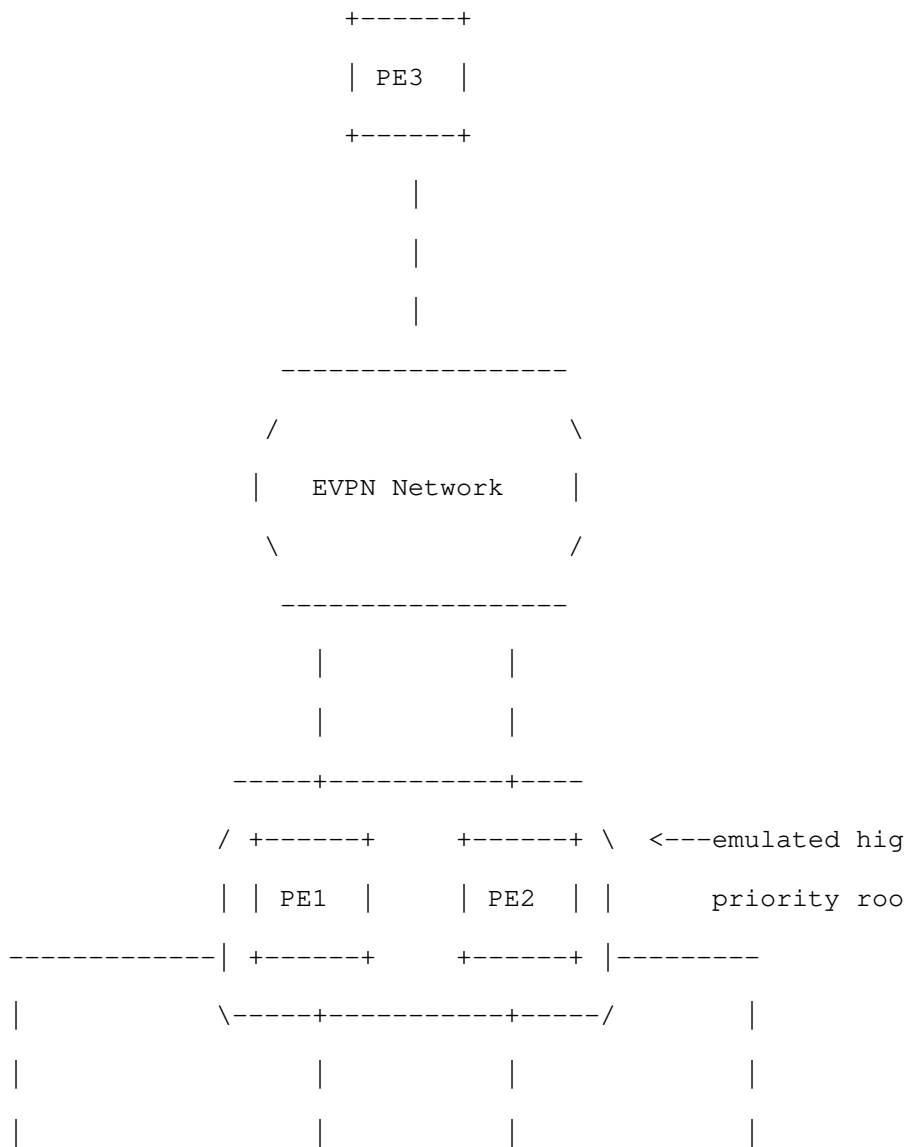
With DF mechanism, native frames enter and leave bridged network via the same designated forwarder for a given VLAN. It may cause congestion or suboptimal routing. PE and bridges should be carefully configured so that end stations on a remnant bridged LAN are separated into different VLANs that have different designated forwarders to achieve better load balancing.

4. Active/Active MAC-based load balancing mechanism

Active/Active MAC-based load balancing mechanism requires the PEs to participate in the control plane protocol of the bridged network. With this mechanism, loop avoidance and per-vlan MAC-based load balancing can be achieved. So it can achieve better load balancing than DF election, and is more applicable if all end stations in bridged network on the same VLAN may cause traffic congestion over the link to DF.

The following two solutions can be used to achieve active/active MAC-based load balancing. One is emulated MSTP root bridge solution; another is bridge control plane protocol tunneling solution. We will described them in the following subsections respectively.

4.1. Emulated MSTP root bridge solution



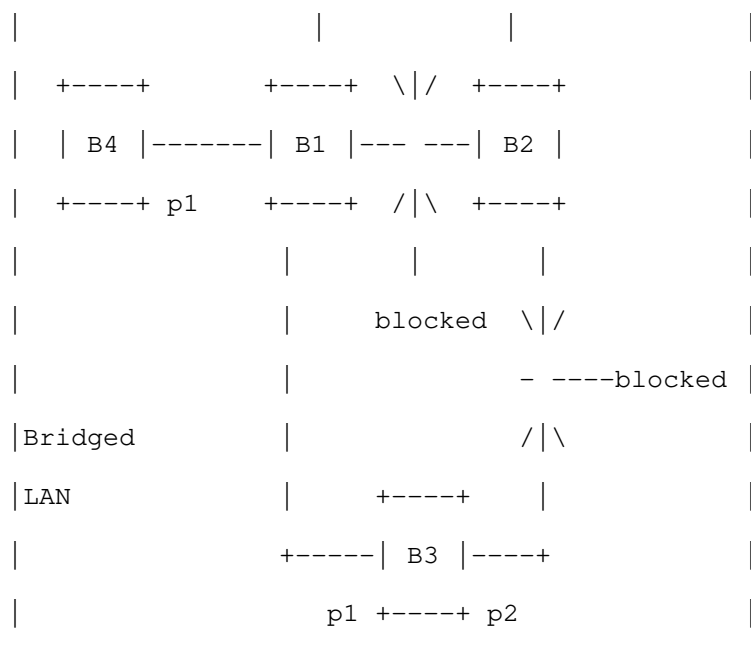


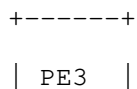
Figure 2 emulated MSTP root bridge solution

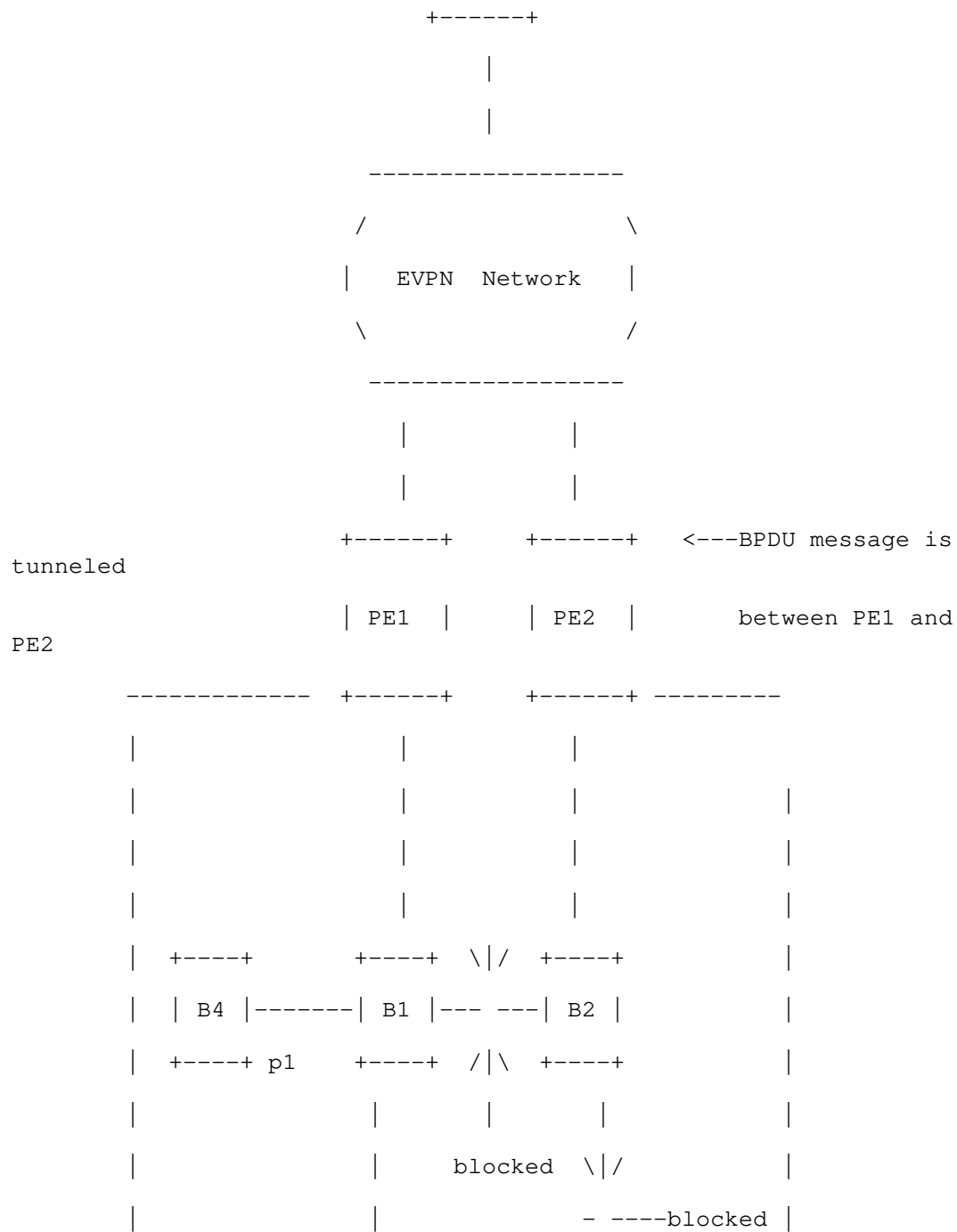
PE1 and PE2 act as an emulated MSTP root bridge. PE1 & PE2 use the same bridge ID to emit spanning tree BPDUs as the highest priority root Bx. All bridges in bridged network see PE1 and PE2 as single tree root. Therefore B1-B2 and B2-B3 links are blocked for loop avoidance by the spanning tree protocol.

When B1-B3 link fails, alternate port p2 on B3 will start to send TC BPDU and go to forwarding state. PE2 receives TC BPDU from B2 sequentially. PE2 tunnel the TC BPDU to PE1. At the same time, PE2 notifies remote PE3 to flush the MAC table through corresponding Ethernet A-D route.

With this solution, PE1 and PE2 needs to tunnel TC BPDU to each other when topology change occurs in the local bridged network.

4.2. Bridge control plane protocol tunneling solution





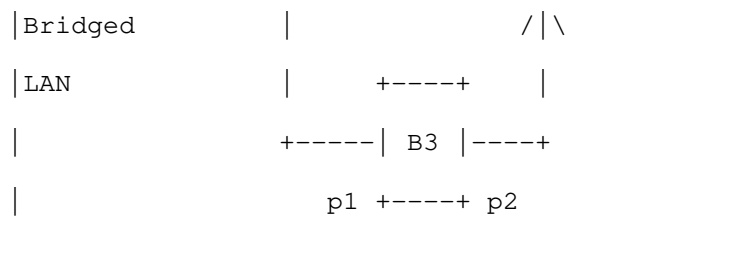


Figure 3 PE1 and PE2 act as normal MSTP bridge nodes

The solution described in the previous section is applicable for STP/MSTP domain. Now we are going to present another solution which can be used for both MSTP and G.8032 domain. The basic idea is to tunnel the control plane messages of local domain among the multi-homed PEs over EVPN network.

4.2.1. Scenario 1: Local bridged network is MSTP

PE1 and PE2 act as normal MSTP bridge nodes. MSTP root bridge can be PE or any switch in the bridged network. BPDU message can be sent through tunnel over EVPN network between PE1 and PE2. The tunnel can be MPLS P2P LSP, MPLS P2MP LSP, or NVO3 tunnel, etc. PE1 and PE2 regard the BPDU tunnel as normal physical link. To avoid BPDU tunnel blocked by MSTP, link cost of the tunnel should be set to 0 or minimum value in MSTP network. With such configuration, it is expected that the blocked port by MSTP protocol can never be the EVPN network facing port on PEs.

4.2.2. Scenario 2: Local bridged network is G.8032

Similarly, PE1 and PE2 act as normal G.8032 ring nodes. They support standard FDB MAC learning, forwarding, flush behavior and port blocking/unblocking mechanisms. G.8032 message can be sent through tunnel over EVPN network between PE1 and PE2. ring protection link(RPL) owner node can be PE or any switch in bridged network. If PE is RPL owner node, RPL can only be configured on access link and can never be configured on the EVPN network facing port on PEs.

4.2.3. Fast convergence

For fast convergence, when a PE notice a topology change event, it should flush local MAC entries and notify the remote PE of the same EVPN instance to withdraw the corresponding Ethernet A-D route. The remote PE that received the withdrawal simply invalidates the MAC entries for that segment.

5. EVPN protocol extension

ESI Label Extended Community MUST be included in EVPN Ethernet A-D route. All-Active multi-homing or active-standby multi-homing mode is decided by the "Active-Standby" bit in the flags of the ESI Label Extended Community through DF mechanism.

ESI Label Extended Community should be extended to support the mechanisms illustrated in this document. "M" bit is introduced to indicate multi-homing mode of MAC-based all active without DF Election. DF selection procedures should be skipped if "M" bit is set to be 1. When remote PE receives Ethernet A-D route withdraw message, it simply invalidates the MAC entries for the segment that corresponding to the Ethernet A-D route.

```

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=0x01 |DF|R|M|      Reserved=0          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Reserved = 0|               ESI Label                       |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

DF: As defined in [EVPN]. It should be ignored if M bit is 1.

R: The bit is already defined as the "Root-Leaf" in [EVPN].

M: The bit is defined as "MAC-based all active without DF Election" and may be set to 1. The above "DF" bit is significant only when "M" bit is set to 0. A value of 1 for M bit means that multi-homed site uses MAC-based active-active access.

6. Security Considerations

TBD

7. IANA Considerations

TBD

7.1. Normative References

- [1] [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

7.2. Informative References

- [1] [EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-01.txt.
- [2] [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt, work in progress, February, 2012.

8. Acknowledgments

The authors wish to acknowledge the important contributions of Shunwan Zhuang.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56625375
Email: liyizhou@huawei.com

Pei Xu
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623590
Email: xupeih@huawei.com

L2VPN

Internet Draft

Intended status: Standards Track

Expires: January 2014

Weiguo Hao

Yizhou Li

Huawei

July 11, 2013

Active-active access in NVO3 network
draft-hao-l2vpn-evpn-nvo3-active-active-00.txt

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BC
P 78
and BCP 79. This document may not be modified, and derivative works of it may
not
be created, and it may not be published except as an Internet-Draft.

This Internet-Draft is submitted in full conformance with the provisions of BC
P 78
and BCP 79. This document may not be modified, and derivative works of it may
not
be created, except to publish it as an RFC and to translate it into languages
other than English.

This document may contain material from IETF Documents or IETF Contributions
published or made publicly available before November 10, 2008. The person(s)
controlling the copyright in some of this material may not have granted the IE
TF

Trust the right to allow modifications of such material outside the IETF Stand
ards

Process. Without obtaining an adequate license from the person(s) controlling
the

copyright in such materials, this document may not be modified outside the IET
F

Standards Process, and derivative works of it may not be created outside the I
ETF

Standards Process, except to format it for publication as an RFC or to transla
te

it into languages other than English.

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 oth
er
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 11, 2014.

Copyright Notice

Copyright (c) 2013 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.

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.

Abstract

E-VPN can be used as a control plane protocol for NV03 network. In All-Active access scenario, loop & echo forwarding prevention among multi-homed NVEs relies on "Split horizon" filtering mechanism. ESI MPLS label is leveraged to perform split horizon filtering in MPLS based EVPN network. In VXLAN or NVGRE encapsulation based overlay network, no MPLS is used. Therefore a new method is introduced in this document to achieve split horizon filtering in non-MPLS overlay network. Data plane of the overlay network is extended with a LAGID field. The function of LAGID is similar to ESI MPLS Label in [EVPN]. It is used to identify each Ethernet segment (ES) on each NVE.

Table of Contents

1. Introduction.....	3
1.1. Terminology.....	4
2. Source IP based solution.....	4
3. LAGID extension solution.....	4
3.1. Ingress Replication.....	6
3.2. Point-to-multipoint.....	6
4. VXLAN data plane extension	6
5. NVGRE data plane extension	7
6. Security Considerations.....	8
7. IANA Considerations	8

8. References	8
8.1. Normative References.....	8
8.2. Informative References	8
9. Acknowledgments	8

1. Introduction

Network Virtualization Overlays (NVO3) is a solution to satisfy a core requirement of multi-tenancy in data center networks through an overlay-based network virtualization approach. VXLAN and NVGRE are two typical mechanisms to implement network virtualization overlays. E-VPN was originally designed for MPLS-based network. E-VPN supports the flexible multi-homing with all-active Attachment Circuits (ACs). In E-VPN, MAC learning between PEs occurs in the control plane.

In All-Active case, the following two problems for multicast packet forwarding should be solved:

1. Duplicate delivery of flooded traffic. As described in [EVPN], Designated Forwarder(DF) election mechanism can be used to prevent duplicate copies of flooded traffic from remote PE. Only one link is elected as DF per <ESI, EVI> or per ESI. DF is responsible for forwarding flooded multi-destination frames to the multi-homed Segment. If a CE is multi-homed to two or more PEs, an Ethernet segment is the set of Ethernet links and may appear to the CE as a Link Aggregation Group (LAG). EVI is E-VPN instance.
2. Loop & Echo Forwarding among multi-homed PEs. As described in [EVPN], if a CE sends a broadcast, unknown unicast or multicast (BUM) packet to one of the non-DF PEs, say PE1, PE1 will forward that packet to all or subset of the other PEs in the EVI including the DF PE for that Ethernet segment. In this case the DF PE MUST drop the packet instead of forwarding to CE. "Split horizon" filtering mechanism relying on MPLS ESI label can be used to avoid loop & echo forwarding.

In NVO3 network, NVE is equivalent to PE in [EVPN], VXLAN and NVGRE are two typical data plane encapsulations between NVEs. [NV-EVPN] analyzes the feasibility of E-VPN to be used as an control plane protocol for NVO3 network, especially the impact of various tunnel encapsulation options such as VXLAN and NVGRE on the E-VPN protocol. With some modifications on E-VPN procedures, EVPN framework can be used for NVO3 solution.

In the scenario of NVE residing in TOR switch, the servers (where VMs are residing) are normally multi-homed to ToR switches to enhance the reliability. Multi-homing may operate in All-Active redundancy mode. In All-Active access scenario, DF

election can be used to solve the duplicated delivery of flooded traffic issue
in

NVO3 network. Each EVI corresponds to single VNI or multiple VNIs. As for the
issue of loop & echo forwarding among multi-homed NVEs, as VXLAN or NVGRE
encapsulation does not include any MPLS label, method other than ESI MPLS labe
1

should be proposed. This draft introduces such mechanism to address the issue of loop & echo forwarding among multi-homed NVEs.

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].

DC: Data Center.

NVE: Network Virtualization Endpoint.

VN: Virtual Network. This is a virtual L2 or L3 domain that belongs a tenant.

TES: Tenant End System. It defines an end system of a particular tenant, which can be for instance a virtual machine (VM), a non-virtualized server, or a physical appliance.

2. Source IP based solution

There is a known solution to address split-horizon filtering problem in the NVO3-based EVPN scenario as following.

Each NVE allocates a unique IP address for each Ethernet segment which is called ESI IP address here. When an NVE receives a BUM frame from a local ESI interface, it uses the ESI IP address as source IP address for NVO3 tunnel encapsulation and sends the frame to other NVE(s).

When an egress NVE receives the multicast frame from NVO3 network, it checks the source IP address of NVO3 tunnel and filters out the frame on all local interfaces connected to Ethernet segments that are shared with the ingress NVE. Each NVE should track the IP address(es) associated with the other NVE(s) with which it has shared multi-homed Ethernet segments. The solution has IP address allocation scalability issue, as each NVE needs to allocate an IP address per Ethernet Segment.

To address the issue above, a new solution with NVO3 data plane extension is introduced in this draft. The details of the solution will be illustrated in section 4. VXLAN and NVGRE data plane extensions will be given in section 5.

3. LAGID extension solution

Link Aggregation Group Identifier(LAGID) is introduced in this solution to perform loop & echo forwarding prevention among multi-homed NVEs. LAGID is used to identify each Ethernet segment on an NVE. All NVEs operating in All-Active multi-

homing mode should announce the local assigned LAGID to other NVEs for each

Hao&Li

Expires January 11, 2014

[Page 4]

Ethernet segment, the LAGID is assigned on each NVE independently and different LAGID can be assigned on different NVE for same Ethernet segment.

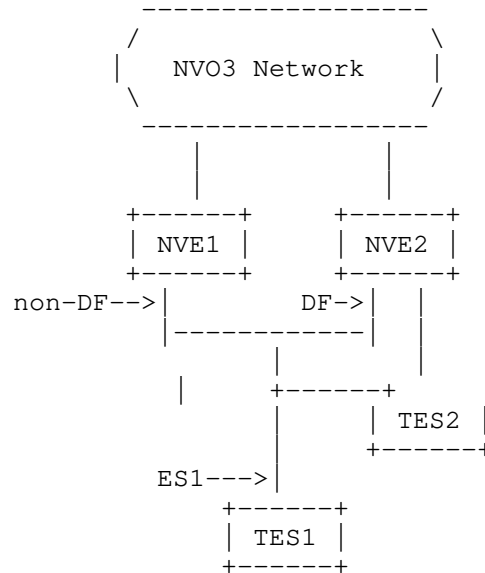


Figure 1 Active-Active access in NVO3 network

The reserved bits in VXLAN/NVGRE header can be used to carry the LAGID for each Ethernet segment, and the new header with LAGID is called LAGID extended NVO3 header. Each BUM packet originating from a non-DF NVE is encapsulated with the LAGID extended NVO3 header that identifies the Ethernet segment from which the frame entered the NVO3 network. Egress DF NVE relies on the value of the LAGID to determine whether or not a BUM frame is allowed to egress a specific Ethernet segment. If the BUM frame is originated from the DF NVE operating in All-Active multi-homing mode, then the DF NVE MAY use normal NVO3 encapsulation without the LAGID.

In NVO3 network, ingress replication or point-to-multipoint tunnels can be used to send BUM traffic destined to multiple NVEs on a per-VNI basis. LAGID extension solution can be used for both.

For ingress replication, ingress NVE sends BUM packet to each destination NVE through a unicast NVO3 tunnel, the LAGID in the extended NVO3 encapsulation is assigned by egress NVE. While for point-to-multipoint, ingress NVE sends BUM packet to all destination NVEs through a multicast NVO3 tunnel, the LAGID in the encapsulation is assigned by ingress NVE. The LAGID and ingress NVE IP address uniquely identifies the Ethernet segment sending the BUM frame in point-to-point scenario.

The following sub-sections will illustrate in more details.

LAGID information can be distributed via "Ethernet A-D route per Ethernet Segment" TLV.

As showing in figure1, TES1 is multi-homed to NVE1 and NVE2 on Ethernet segment(ES)1 and operating in All-Active multi-homing mode. TES2 is single homed to NVE2. Both TES1 and TES2 belong to VNI1. NVE1 is the non-DF for VNI1 and NVE2 is the DF for VNI1. Forwarding procedures for ingress replication and point-to-multipoint is described in the following sub-sections respectively.

3.1. Ingress Replication

1. NVE1 receives a BUM packet from TES1 on VNI1 on ES1.
2. NVE1 sends the BUM packet to egress NVE2 using unicast tunnel with LAGID extension. The LAGID is assigned by egress NVE2 in advance. The destination IP address of the unicast tunnel is NVE2 IP address.
3. Egress NVE2 receives this packet from NVO3 network. As LAGID in unicast NVO3 encapsulation is equal to the local assigned LAGID for ES1, NVE2 does not forward the packet to TES1. Because the link connects to TES2 doesn't belong to ES1, so NVE2 forwards the packet to TES2.

3.2. Point-to-multipoint

1. NVE1 receives a BUM packet from TES1 on VNI1 on ES1.
2. NVE1 sends the BUM packet to egress NVE2 using multicast with LAGID extended NVO3 encapsulation. The LAGID in the NVO3 encapsulation is assigned by ingress NVE1. the destination IP address of the multicast tunnel is the multicast IP address corresponding to VNI1.
3. Egress NVE2 receives this packet from NVO3 network. AS the source IP of the multicast NVO3 encapsulation is NVE1 and the LAGID in the encapsulation is the LAGID that announced by NVE1 for ES1, so NVE2 drops the packet to TES1. Because the link connects to TES2 doesn't belong to ES1, so PE2 forwards the packet to TES2.

4. VXLAN data plane extension

The VXLAN header can be extended to support the mechanism illustrated in this document. "'L'" flag is introduced to indicate that the LAGID field is present in the VXLAN header. "'LAGID'" is a 12 bit value to identify local Ethernet segment

nt on
ingress or egress NVE. For unicast VXLAN tunnel, the LAGID is assigned by egress
ss
NVE and is the identification of Ethernet segment on egress NVE. For multicast
VXLAN tunnel, the LAGID is assigned by ingress NVE and is the identification of
f

Ethernet segment on ingress NVE. Each BUM packet originating from a non-DF NVE to VXLAN network must carry the LAGID.

Egress DF NVE relies on the value of the LAGID to determine a BUM frame should be dropped or forwarded to a specific local Ethernet segment.

VXLAN header format is shown below:

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|L|R|R|R|I|R|R|R|      LAGID(optional)      |      Reserved      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               VXLAN Network Identifier (VNI) |      Reserved      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

L: If the bit is set to 1, then it indicates that the LAGID field is present in the VXLAN header. Otherwise, the LAGID field is not present in the VXLAN header.

LAGID: LAGID field is a 12 bit field and is used to identify an Ethernet segment on an NVE operating in All-Active multi-homing mode. LAGID is significant only when "L" bit is set to 1.

5. NVGRE data plane extension

Similar to VXLAN data plane extension, the NVGRE header can be extended to support the mechanism illustrated in this document too. "L" flag is introduced to indicate that the LAGID field is present in the GRE header. Each BUM packet originating from a non-DF NVE to NVGRE network must carry the LAGID. Egress DF NVE relies on the value of the LAGID to determine a BUM frame should be dropped or forwarded to a specific local Ethernet segment.

NVGRE header format is shown below:

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|C| |K|S|L| Reserved0      | Ver |      Protocol Type 0x6558      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Checksum (optional)      | LAGID(optional)      | Resv |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Virtual Subnet ID (VSID)      |      Reserved      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

L: If the bit is set to 1, then it indicates that the LAGID field is present in the NVGRE header. Otherwise, the LAGID field is not present in the NVGRE header.

LAGID: The LAGID field is a 12 bit field and is inserted by ingress NVE when L Bit is set. The LAGID is used to identify a ESI that is assigned on each NVE operating in an All-Active multi-homing mode. For unicast NVGRE tunnel, the LAGID is

assigned by egress NVE. For multicast NVGRE tunnel, the LAGID is assigned by ingress NVE.

6. Security Considerations

NA

7. IANA Considerations

NA

8. References

8.1. Normative References

8.2. Informative References

- [1] [NV-EVPN] Sajassi, A., Drake J, D., Bitar, N., , " A Network Virtualization Overlay Solution using E-VPN", draft-sd-l2vpn-evpn-overlay-01, February 2013.
- [2] [RFC2890] G. Dommety, "Key and Sequence Number Extensions to GRE",RFC 2890, September 2000
- [3] [NVGRE] Sridhavan, M., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-01.txt, July 8, 2012.
- [4] [VXLAN] Dutt, D., et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draftmahalingam- dutt-dcops-vxlan-02.txt, August 22, 2012.
- [5] [E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-02.txt, work in progress, February, 2012..

9. Acknowledgments

The authors wish to acknowledge the important contributions of Junlin Zhang.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56625375
Email: liyizhou@huawei.com

Network Working Group
INTERNET-DRAFT
Category: Standards Track

A. Sajassi
Cisco

N. Bitar
Verizon

R. Aggarwal
Arktan

S. Boutros
K. Patel
S. Salam
Cisco

W. Henderickx
F. Balus
Alcatel-Lucent

J. Drake
R. Shekhar
Juniper Networks

Aldrin Isaac
Bloomberg

J. Uttaro
AT&T

Expires: January 15, 2014

July 15, 2013

BGP MPLS Based Ethernet VPN
draft-ietf-l2vpn-evpn-04

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) 2013 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.

Abstract

This document describes procedures for BGP MPLS based Ethernet VPNs (EVPN).

Table of Contents

1. Specification of requirements	5
2. Terminology	5
3. Introduction	6
4. Contributors	6
5. BGP MPLS Based EVPN Overview	6
6. Ethernet Segment	7
7. Ethernet Tag	9
7.1 VLAN Based Service Interface	9
7.2 VLAN Bundle Service Interface	9
7.2.1 Port Based Service Interface	10
7.3 VLAN Aware Bundle Service Interface	10
7.3.1 Port Based VLAN Aware Service Interface	10
8. BGP EVPN NLRI	10
8.1. Ethernet Auto-Discovery Route	11
8.2. MAC Advertisement Route	12
8.3. Inclusive Multicast Ethernet Tag Route	12
8.4 Ethernet Segment Route	13
8.5 ESI Label Extended Community	13
8.6 ES-Import Route Target	14
8.7 MAC Mobility Extended Community	14
8.8 Default Gateway Extended Community	15
9. Multi-homing Functions	15
9.1 Multi-homed Ethernet Segment Auto-Discovery	15
9.1.1 Constructing the Ethernet Segment Route	15
9.2 Fast Convergence	16
9.2.1 Constructing the Ethernet A-D Route per Ethernet Segment	16
9.2.1.1. Ethernet A-D Route Targets	17
9.3 Split Horizon	17
9.3.1 ESI Label Assignment	18
9.3.1.1 Ingress Replication	18

9.3.1.2. P2MP MPLS LSPs	19
9.4 Aliasing and Backup-Path	20
9.4.1 Constructing the Ethernet A-D Route per EVI	21
9.4.1.1 Ethernet A-D Route Targets	22
9.5 Designated Forwarder Election	22
9.6. Interoperability with Single-homing PEs	24
10. Determining Reachability to Unicast MAC Addresses	25
10.1. Local Learning	25
10.2. Remote learning	26
10.2.1. Constructing the BGP EVPN MAC Address Advertisement	26
10.2.2 Route Resolution	28
11. ARP and ND	29
11.1 Default Gateway	29
12. Handling of Multi-Destination Traffic	30
12.1. Construction of the Inclusive Multicast Ethernet Tag Route	31
12.2. P-Tunnel Identification	31
13. Processing of Unknown Unicast Packets	32
13.1. Ingress Replication	33
13.2. P2MP MPLS LSPs	33
14. Forwarding Unicast Packets	34
14.1. Forwarding packets received from a CE	34
14.2. Forwarding packets received from a remote PE	35
14.2.1. Unknown Unicast Forwarding	35
14.2.2. Known Unicast Forwarding	35
15. Load Balancing of Unicast Frames	36
15.1. Load balancing of traffic from an PE to remote CEs	36
15.1.1 Single-Active Redundancy Mode	36
15.1.2 All-Active Redundancy Mode	37
15.2. Load balancing of traffic between an PE and a local CE	38
15.2.1. Data plane learning	38
15.2.2. Control plane learning	39
16. MAC Mobility	39
16.1. MAC Duplication Issue	41
16.2. Sticky MAC addresses	41
17. Multicast & Broadcast	41
17.1. Ingress Replication	41
17.2. P2MP LSPs	42
17.2.1. Inclusive Trees	42
18. Convergence	42
18.1. Transit Link and Node Failures between PEs	42
18.2. PE Failures	43
18.2. PE to CE Network Failures	43
19. Frame Ordering	43
20. Acknowledgements	44
21. Security Considerations	44
22. IANA Considerations	44
23. References	45

23.1 Normative References	45
23.2 Informative References	45
24. Author's Address	45

1. Specification of requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Terminology

Bridge Domain:

Broadcast Domain:

CE: Customer Edge device e.g., host or router or switch

EVI: An EVPN instance spanning across the PEs participating in that VPN

MAC-VRF: A Virtual Routing and Forwarding table for MAC addresses on a PE for an EVI

Ethernet Segment Identifier (ESI): If a CE is multi-homed to two or more PEs, the set of Ethernet links that attaches the CE to the PEs is an 'Ethernet segment'. Ethernet segments MUST have a unique non-zero identifier, the '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. Ethernet tag(s) are assigned to the broadcast domains of a given EVPN instance by the provider of that EVPN, and each PE in that EVPN instance performs a mapping between broadcast domain identifier(s) understood by each of its attached CEs and the corresponding Ethernet tag.

LACP: Link Aggregation Control Protocol

MP2MP: Multipoint to Multipoint

P2MP: Point to Multipoint

P2P: Point to Point

Single-Active Mode: When a device or a network is multi-homed to two or more PEs and when only a single PE in such redundancy group can forward traffic to/from the multi-homed device or network for a given VLAN, then such multi-homing or redundancy is referred to as "Single-Active".

All-Active Mode: When a device is multi-homed to two or more PEs and when all PEs in such redundancy group can forward traffic to/from the multi-homed device for a given VLAN, then such multi-homing or redundancy is referred to as "All-Active".

3. Introduction

This document describes procedures for BGP MPLS based Ethernet VPNs (EVPN). The procedures described here are intended to meet the requirements specified in [EVPN-REQ]. Please refer to [EVPN-REQ] for the detailed requirements and motivation. EVPN requires extensions to existing IP/MPLS protocols as described in this document. In addition to these extensions EVPN uses several building blocks from existing MPLS technologies.

4. Contributors

In addition to the authors listed above, the following individuals also contributed to this document:

Quaizar Vohra
Kireeti Kompella
Apurva Mehta
Nadeem Mohammad
Juniper Networks

Clarence Filsfils
Dennis Cai
Cisco

5. BGP MPLS Based EVPN Overview

This section provides an overview of EVPN. An EVPN instance comprises CEs that are connected to PEs that form the edge of the MPLS infrastructure. A CE may be a host, a router or a switch. The PEs provide virtual Layer 2 bridged connectivity between the CEs. There may be multiple EVPN instances in the provider's network.

The PEs may be connected by an MPLS LSP infrastructure which provides the benefits of MPLS technology such as fast-reroute, resiliency, etc. The PEs may also be connected by an IP infrastructure in which case IP/GRE tunneling or other IP tunneling can be used between the PEs. The detailed procedures in this version of this document are specified only for MPLS LSPs as the tunneling technology. However these procedures are designed to be extensible to IP tunneling as the

PSN tunneling technology.

In an EVPN, MAC learning between PEs occurs not in the data plane (as happens with traditional bridging) but in the control plane. Control plane learning offers greater control over the MAC learning process, such as restricting who learns what, and the ability to apply policies. Furthermore, the control plane chosen for advertising MAC reachability information is multi-protocol (MP) BGP (similar to IP VPNs (RFC 4364)). This provides greater scalability and the ability to preserve the "virtualization" or isolation of groups of interacting agents (hosts, servers, virtual machines) from each other. In EVPN, PEs advertise the MAC addresses learned from the CEs that are connected to them, along with an MPLS label, to other PEs in the control plane using MP-BGP. Control plane learning enables load balancing of traffic to and from CEs that are multi-homed to multiple PEs. This is in addition to load balancing across the MPLS core via multiple LSPs between the same pair of PEs. In other words it allows CEs to connect to multiple active points of attachment. It also improves convergence times in the event of certain network failures.

However, learning between PEs and CEs is done by the method best suited to the CE: data plane learning, IEEE 802.1x, LLDP, 802.1aq, ARP, management plane or other protocols.

It is a local decision as to whether the Layer 2 forwarding table on an PE is populated with all the MAC destination addresses known to the control plane, or whether the PE implements a cache based scheme. For instance the MAC forwarding table may be populated only with the MAC destinations of the active flows transiting a specific PE.

The policy attributes of EVPN are very similar to those of IP-VPN. A EVPN instance requires a Route-Distinguisher (RD) which is unique per PE and one or more globally unique Route-Targets (RTs). A CE attaches to a MAC-VRF on an PE, on an Ethernet interface which may be configured for one or more Ethernet Tags, e.g., VLAN IDs. Some deployment scenarios guarantee uniqueness of VLAN IDs across EVPN instances: all points of attachment for a given EVPN instance use the same VLAN ID, and no other EVPN instance uses this VLAN ID. This document refers to this case as a "Unique VLAN EVPN" and describes simplified procedures to optimize for it.

6. Ethernet Segment

If a CE is multi-homed to two or more PEs, the set of Ethernet links constitutes an "Ethernet Segment". An Ethernet segment may appear to the CE as a Link Aggregation Group (LAG). Ethernet segments have an identifier, called the "Ethernet Segment Identifier" (ESI) which is

encoded as a ten octets integer. The following two ESI values are reserved:

- ESI 0 denotes a single-homed CE.
- ESI {0xFF} (repeated 10 times) is known as MAX-ESI and is reserved.

In general, an Ethernet segment MUST have a non-reserved ESI that is unique network wide (e.g., across all EVPN instances on all the PEs). If the CE(s) constituting an Ethernet Segment is (are) managed by the network operator, then ESI uniqueness should be guaranteed; however, if the CE(s) is (are) not managed, then the operator MUST configure a network-wide unique ESI for that Ethernet Segment. This is required to enable auto-discovery of Ethernet Segments and DF election. The ESI can be assigned using various mechanisms:

1. If IEEE 802.1AX LACP is used between the PEs and CEs, then the ESI is determined from LACP by concatenating the following parameters:

- + CE LACP System Identifier comprised of two octets of System Priority and six octets of System MAC address, where the System Priority is encoded in the most significant two octets. The CE LACP identifier MUST be encoded in the high order eight octets of the ESI.
- + CE LACP two octets Port Key. The CE LACP port key MUST be encoded in the low order two octets of the ESI.

As far as the CE is concerned, it would treat the multiple PEs that it is connected to as the same switch. This allows the CE to aggregate links that are attached to different PEs in the same bundle.

This mechanism could be used only if it produces ESIs that satisfy the uniqueness requirement specified above.

2. In the case of indirectly connected hosts via a bridged LAN between the CEs and the PEs, the ESI is determined based on the Layer 2 bridge protocol as follows: If MST is used in the bridged LAN then the value of the ESI is derived by listening to BPDUs on the Ethernet segment. To achieve this the PE is not required to run MST. However the PE must learn the Root Bridge MAC address and Bridge Priority of the root of the Internal Spanning Tree (IST) by listening to the BPDUs. The ESI is constructed as follows:

{Bridge Priority (16 bits) , Root Bridge MAC Address (48 bits)}

This mechanism could be used only if it produces ESIs that satisfy the uniqueness requirement specified above.

3. The ESI may be configured.

7. Ethernet Tag

An Ethernet Tag identifies a particular broadcast domain, e.g. a VLAN, in an EVPN Instance. An EVPN Instance consists of one or more broadcast domains (one or more VLANs). VLANs are assigned to a given EVPN Instance by the provider of the EVPN service. A given VLAN can itself be represented by multiple VLAN IDs (VIDs). In such cases, the PEs participating in that VLAN for a given EVPN instance are responsible for performing VLAN ID translation to/from locally attached CE devices.

If a VLAN is represented by a single VID across all PE devices participating in that VLAN for that EVPN instance, then there is no need for VID translation at the PEs. Furthermore, some deployment scenarios guarantee uniqueness of VID across all EVPN instances; all points of attachment for a given EVPN instance use the same VID and no other EVPN instances use that VID. This allows the RT(s) for each EVPN instance to be derived automatically from the corresponding VID, as described in section 9.4.1.1.1 "Auto-Derivation from the Ethernet Tag ID".

The following subsections discuss the relationship between broadcast domains (e.g., VLANs), Ethernet Tags (e.g., VIDs), and MAC-VRFs as well as the setting of the Ethernet Tag Identifier, in the various EVPN BGP routes (defined in section 8), for the different types of service interfaces described in [EVPN-REQ].

7.1 VLAN Based Service Interface

With this service interface, an EVPN instance consists of only a single broadcast domain (e.g., a single VLAN). Therefore, there is a one to one mapping between a VID on this interface and a MAC-VRF. Since a MAC-VRF corresponds to a single VLAN, it consists of a single bridge domain corresponding to that VLAN. If the VLAN is represented by different VIDs on different PEs, then each PE needs to perform VID translation for frames destined to its attached CEs. In such scenarios, the Ethernet frames transported over MPLS/IP network SHOULD remain tagged with the originating VID and a VID translation MUST be supported in the data path and MUST be performed on the disposition PE. The Ethernet Tag Identifier in all EVPN routes MUST be set to 0.

7.2 VLAN Bundle Service Interface

With this service interface, an EVPN instance corresponds to several broadcast domains (e.g., several VLANs); however, only a single bridge domain is maintained per MAC-VRF which means multiple VLANs share the same bridge domain. This implies MAC addresses MUST be unique across different VLANs for this service to work. In other words, there is a many-to-one mapping between VLANs and a MAC-VRF, and the MAC-VRF consists of a single bridge domain. Furthermore, a single VLAN must be represented by a single VID - e.g., no VID translation is allowed for this service interface type. The MPLS encapsulated frames MUST remain tagged with the originating VID. Tag translation is NOT permitted. The Ethernet Tag Identifier in all EVPN routes MUST be set to 0.

7.2.1 Port Based Service Interface

This service interface is a special case of the VLAN Bundle service interface, where all of the VLANs on the port are part of the same service and map to the same bundle. The procedures are identical to those described in section 7.2.

7.3 VLAN Aware Bundle Service Interface

With this service interface, an EVPN instance consists of several broadcast domains (e.g., several VLANs) with each VLAN having its own bridge domain - e.g., multiple bridge domains (one per VLAN) is maintained by a single MAC-VRF corresponding to the EVPN instance. In the case where a single VLAN is represented by different VIDs on different CEs and thus tag (VID) translation is required, a normalized Ethernet Tag (VID) MUST be carried in the MPLS encapsulated frames and a tag translation function MUST be supported in the data path. This translation MUST be performed in data path on both the imposition as well as the disposition PEs (translating to normalized tag on imposition PE and translating to local tag on disposition PE). The Ethernet Tag Identifier in all EVPN routes MUST be set to the normalized Ethernet Tag assigned by the EVPN provider.

7.3.1 Port Based VLAN Aware Service Interface

This service interface is a special case of the VLAN Aware Bundle service interface, where all of the VLANs on the port are part of the same service and map to the same bundle. The procedures are identical to those described in section 7.3.

8. BGP EVPN NLRI

This document defines a new BGP NLRI, called the EVPN NLRI.

Following is the format of the EVPN NLRI:

Route Type (1 octet)
Length (1 octet)
Route Type specific (variable)

The Route Type field defines encoding of the rest of the EVPN NLRI (Route Type specific EVPN NLRI).

The Length field indicates the length in octets of the Route Type specific field of EVPN NLRI.

This document defines the following Route Types:

- + 1 - Ethernet Auto-Discovery (A-D) route
- + 2 - MAC advertisement route
- + 3 - Inclusive Multicast Route
- + 4 - Ethernet Segment Route

The detailed encoding and procedures for these route types are described in subsequent sections.

The EVPN NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an AFI of 25 (L2VPN) and a SAFI of 70 (EVPN). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the EVPN NLRI (encoded as specified above).

In order for two BGP speakers to exchange labeled EVPN NLRI, they must use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of 25 (L2VPN) and a SAFI of 70 (EVPN).

8.1. Ethernet Auto-Discovery Route

A Ethernet A-D route type specific EVPN NLRI consists of the following:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
MPLS Label (3 octets)

For procedures and usage of this route please see section 9.2 "Fast Convergence" and section 9.4 "Aliasing".

8.2. MAC Advertisement Route

A MAC advertisement route type specific EVPN NLRI consists of the following:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octet)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (4 or 16 octets)
MPLS Label (3 octets)

For the purpose of BGP route key processing, only the Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP Address Address fields are considered to be part of the prefix in the NLRI. The Ethernet Segment Identifier and MPLS Label fields are to be treated as route attributes as opposed to being part of the "route".

For procedures and usage of this route please see section 10 "Determining Reachability to Unicast MAC Addresses" and section 15 "Load Balancing of Unicast Packets".

8.3. Inclusive Multicast Ethernet Tag Route

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

RD (8 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octet)
Originating Router's IP Addr (4 or 16 octets)

For procedures and usage of this route please see section 12 "Handling of Multi-Destination Traffic", section 13 "Processing of Unknown Unicast Traffic" and section 17 "Multicast".

8.4 Ethernet Segment Route

The Ethernet Segment Route is encoded in the EVPN NLRI using the Route Type value of 4. The Route Type Specific field of the NLRI is formatted as follows:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
IP Address Length (1 octet)
Originating Router's IP Addr (4 or 16 octets)

For procedures and usage of this route please see section 9.5 "Designated Forwarder Election".

8.5 ESI Label Extended Community

This extended community is a new transitive extended community with the Type field is 0x06, and the Sub-Type of 0x01. It may be advertised along with Ethernet Auto-Discovery routes and it enables split-horizon procedures for multi-homed sites as described in section 9.3 "Split Horizon".

Each ESI Label Extended Community is encoded as a 8-octet value as follows:

```

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=0x01 | Flags (One Octet) | Reserved=0 |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Reserved = 0 |               ESI Label               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

The low order bit of the flags octet is defined as the "Active-Standby" bit and may be set to 1. A value of 0 means that the multi-homed site is operating in All-Active mode; whereas, a value of 1 means that the multi-homed site is operating in Single-Active mode.

The second low order bit of the flags octet is defined as the "Root-Leaf". A value of 0 means that this label is associated with a Root site; whereas, a value of 1 means that this label is associate with a Leaf site. The other bits must be set to 0.

8.6 ES-Import Route Target

This is a new transitive Route Target extended community carried with the Ethernet Segment route. When used, it enables all the PEs connected to the same multi-homed site to import the Ethernet Segment routes. The value is derived automatically from the ESI by encoding the 6-byte MAC address portion of the ESI in the ES-Import Route Target. The format of this extended community is as follows:

```

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=0x02 |               ES-Import               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               ES-Import Cont'd               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

This document expands the definition of the Route Target extended community to allow the value of high order octet (Type field) to be 0x06 (in addition to the values specified in rfc4360). The value of low order octet (Sub-Type field) of 0x02 indicates that this extended community is of type "Route Target". The new value for Type field of 0x06 indicates that the structure of this RT is a six bytes value (e.g., a MAC address). A BGP speaker that implements RT-Constrain (RFC4684) MUST apply the RT-Constrain procedures to the ES-import RT as-well.

For procedures and usage of this attribute, please see section 9.1 "Redundancy Group Discovery".

8.7 MAC Mobility Extended Community

This extended community is a new transitive extended community with the Type field of 0x06 and the Sub-Type of 0x00. It may be advertised along with MAC Advertisement routes. The procedures for using this Extended Community are described in section 16 "MAC Mobility".

The MAC Mobility Extended Community is encoded as a 8-octet value as follows:

```

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=0x00 | Flags(1 octet) | Reserved=0  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Sequence Number                    |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

The low order bit of the flags octet is defined as the "Sticky/static" flag and may be set to 1. A value of 1 means that the MAC address is static and cannot move.

8.8 Default Gateway Extended Community

The Default Gateway community is an Extended Community of an Opaque Type (see 3.3 of rfc4360). It is a transitive community, which means that the first octet is 0x03. The value of the second octet (Sub-Type) is 0x030d (Default Gateway) as defined by IANA. The Value field of this community is reserved (set to 0 by the senders, ignored by the receivers).

9. Multi-homing Functions

This section discusses the functions, procedures and associated BGP routes used to support multi-homing in EVPN. This covers both multi-homed device (MHD) as well as multi-homed network (MHN) scenarios.

9.1 Multi-homed Ethernet Segment Auto-Discovery

PEs connected to the same Ethernet segment can automatically discover each other with minimal to no configuration through the exchange of the Ethernet Segment route.

9.1.1 Constructing the Ethernet Segment Route

The Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value field comprises an IP address of the MES (typically, the loopback address) followed by 0's.

The Ethernet Segment Identifier MUST be set to the ten octet ESI identifier described in section 6.

The BGP advertisement that advertises the Ethernet Segment route MUST also carry an ES-Import extended community attribute, as defined in section 8.6.

The Ethernet Segment Route filtering MUST be done such that the Ethernet Segment Route is imported only by the PEs that are multi-homed to the same Ethernet Segment. To that end, each PE that is connected to a particular Ethernet segment constructs an import filtering rule to import a route that carries the ES-Import extended community, constructed from the ESI.

9.2 Fast Convergence

In EVPN, MAC address reachability is learnt via the BGP control-plane over the MPLS network. As such, in the absence of any fast protection mechanism, the network convergence time is a function of the number of MAC Advertisement routes that must be withdrawn by the PE encountering a failure. For highly scaled environments, this scheme yields slow convergence.

To alleviate this, EVPN defines a mechanism to efficiently and quickly signal, to remote PE nodes, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment. This is done by having each PE advertise an Ethernet A-D Route per Ethernet segment for each locally attached segment (refer to section 9.2.1 below for details on how this route is constructed). Upon a failure in connectivity to the attached segment, the PE withdraws the corresponding Ethernet A-D route. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all MAC addresses associated with the Ethernet segment in question. If no other PE had advertised an Ethernet A-D route for the same segment, then the PE that received the withdrawal simply invalidates the MAC entries for that segment. Otherwise, the PE updates the next-hop adjacencies to point to the backup PE(s).

9.2.1 Constructing the Ethernet A-D Route per Ethernet Segment

This section describes procedures to construct the Ethernet A-D route when a single such route is advertised by an PE for a given Ethernet Segment. This flavor of the Ethernet A-D route is used for fast convergence (as discussed above) as well as for advertising the ESI label used for split-horizon filtering (as discussed in section 9.3). Support of this route flavor is MANDATORY.

Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value

field comprises an IP address of the PE (typically, the loopback address) followed by 0.

The Ethernet Segment Identifier MUST be a ten octet entity as described in section "Ethernet Segment". This document does not specify the use of the Ethernet A-D route when the Segment Identifier is set to 0.

The Ethernet Tag ID MUST be set to 0.

The MPLS label in the NLRI MUST be set to 0.

The "ESI Label Extended Community" MUST be included in the route. If all-Active multi-homing is desired, then the "Active-Standby" bit in the flags of the ESI Label Extended Community MUST be set to 0 and the MPLS label in that extended community MUST be set to a valid MPLS label value. The MPLS label in this Extended Community is referred to as an "ESI label". This label MUST be a downstream assigned MPLS label if the advertising PE is using ingress replication for receiving multicast, broadcast or unknown unicast traffic from other PEs. If the advertising PE is using P2MP MPLS LSPs for sending multicast, broadcast or unknown unicast traffic, then this label MUST be an upstream assigned MPLS label. The usage of this label is described in section 9.3.

If the Ethernet Segment is connected to more than one PE and Single-Active multi-homing is desired, then the "Active-Standby" bit in the flags of the ESI Label Extended Community MUST be set to 1 and ESI label MUST be set to zero.

9.2.1.1. Ethernet A-D Route Targets

The Ethernet A-D route MUST carry one or more Route Target (RT) attributes. These RTs MUST be the set of RTs associated with all the EVPN instances to which the Ethernet Segment, corresponding to the Ethernet A-D route, belongs.

9.3 Split Horizon

Consider a CE that is multi-homed to two or more PEs on an Ethernet segment ES1 operating in All-Active mode. If the CE sends a broadcast, unknown unicast, or multicast (BUM) packet to one of the non-DF (Designated Forwarder) PEs, say PE1, then PE1 will forward that packet to all or subset of the other PEs in that EVPN instance including the DF PE for that Ethernet segment. In this case the DF PE that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This filtering is referred to as "split horizon" filtering in this document.

In order to achieve this split horizon function, every BUM packet originating from a non-DF PE is encapsulated with an MPLS label that identifies the Ethernet segment of origin (i.e. the segment from which the frame entered the EVPN network). This label is referred to as the ESI label, and MUST be distributed by all PEs when operating in All-Active multi-homing mode using the "Ethernet A-D route per Ethernet Segment" as per the procedures in section 9.2.1 above. This route is imported by the PEs connected to the Ethernet Segment and also by the PEs that have at least one EVPN instance in common with the Ethernet Segment in the route. As described in section 9.1.1, the route MUST carry an ESI Label Extended Community with a valid ESI label. The disposition DF PE rely on the value of the ESI label to determine whether or not a BUM frame is allowed to egress a specific Ethernet segment. It should be noted that if the BUM frame is originated from the DF PE operating in All-Active multi-homing mode, then the DF PE MAY not encapsulate the frame with the ESI label. Furthermore, if the multi-homed PEs operate in active/standby mode, then the packet MUST NOT be encapsulated with the ESI label and the label value MUST be set to zero in ESI Label Extended Community per section 9.2.1 above.

9.3.1 ESI Label Assignment

The following subsections describe the assignment procedures for the ESI label, which differ depending on the type of tunnels being used to deliver multi-destination packets in the EVPN network.

9.3.1.1 Ingress Replication

All PEs operating in an All-Active multi-homing mode that rely on ingress replication for the reception of BUM traffic, distribute to other PEs, that belong to the Ethernet segment, a downstream assigned "ESI label" in the Ethernet A-D route per ESI. This label MUST be programmed in the platform label space by the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for.

Consider PE1 and PE2 that are multi-homed to CE1 on ES1 and operating in All-Active multi-homing mode. Further consider that PE1 is using P2P or MP2P LSPs to send packets to PE2. Consider that PE1 is the non-DF for VLAN1 and PE2 is the DF for VLAN1, and PE1 receives a BUM packet from CE1 on VLAN1 on ES1. In this scenario, PE2 distributes an Inclusive Multicast Ethernet Tag route for VLAN1 corresponding to an EVPN instance. So, when PE1 sends a BUM packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that PE2 has distributed for ES1. It MUST then push on the MPLS label distributed by PE2 in the Inclusive Multicast Ethernet Tag route for

VLAN1. The resulting packet is further encapsulated in the P2P or MP2P LSP label stack required to transmit the packet to PE2. When PE2 receives this packet, it determines the set of ESIs to replicate the packet to from the top MPLS label, after any P2P or MP2P LSP labels have been removed. If the next label is the ESI label assigned by PE2 for ES1, then PE2 MUST NOT forward the packet onto ES1. If the next label is an ESI label which has not been assigned by PE2, then PE2 MUST drop the packet. It should be noted that in this scenario, if PE2 receives a BUM traffic for VLAN1 from CE1, then it doesn't need to encapsulate the packet with an ESI label when sending it to the PE1 since PE1 can use its DF logic to filter the BUM packets and thus doesn't need to use split-horizon filtering for ES1.

9.3.1.2. P2MP MPLS LSPs

The non-DF PE's operating in an All-Active multi-homing mode that is using P2MP LSPs for sending BUM traffic, distribute to other PE's, that belong to the Ethernet segment or have an EVPN instance in common with the Ethernet Segment, an upstream assigned "ESI label" in the Ethernet A-D route. This label is upstream assigned by the PE that advertises the route. This label MUST be programmed by the other PE's, that are connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for. This label MUST also be programmed by the other PE's, that import the route but are not connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must be a POP with no other associated action.

Consider PE1 and PE2 that are multi-homed to CE1 on ES1 and operating in All-Active multi-homing mode. Also consider PE3 belongs to one of the EVPN instances of ES1. Further, assume that PE1 which is the non-DF, using P2MP MPLS LSPs to send BUM packets. When PE1 sends a BUM packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI that the packet was received on. The resulting packet is further encapsulated in the P2MP MPLS label stack necessary to transmit the packet to the other PE's. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for EVPN. When PE2 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1, then PE2 MUST NOT forward the packet onto ES1. When PE3 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1 and PE3 is not

connected to ES1, then PE3 MUST pop the label and flood the packet over all local ESIs in that EVPN instance. It should be noted that when PE2 sends a BUM frame over a P2MP LSP, it does not need to encapsulate the frame with an ESI label because it is the DF for that VLAN.

9.4 Aliasing and Backup-Path

In the case where a CE is multi-homed to multiple PE nodes, using a LAG with All-Active redundancy, it is possible that only a single PE learns a set of the MAC addresses associated with traffic transmitted by the CE. This leads to a situation where remote PE nodes receive MAC advertisement routes, for these addresses, from a single PE even though multiple PEs are connected to the multi-homed segment. As a result, the remote PEs are not able to effectively load-balance traffic among the PE nodes connected to the multi-homed Ethernet segment. This could be the case, for e.g. when the PEs perform data-path learning on the access, and the load-balancing function on the CE hashes traffic from a given source MAC address to a single PE. Another scenario where this occurs is when the PEs rely on control plane learning on the access (e.g. using ARP), since ARP traffic will be hashed to a single link in the LAG.

To alleviate this issue, EVPN introduces the concept of 'Aliasing'. Aliasing refers to the ability of a PE to signal that it has reachability to a given locally attached Ethernet segment, even when it has learnt no MAC addresses from that segment. The Ethernet A-D route per EVI is used to that end. Remote PEs which receive MAC advertisement routes with non-reserved ESI SHOULD consider the advertised MAC address as reachable via all PEs which have advertised reachability to the relevant Segment using: (1) Ethernet A-D routes per EVI with the same ESI (and Ethernet Tag if applicable) AND (2) Ethernet A-D routes per ESI with the same ESI and with the Active/Standby bit set to 0 in the ESI Label Extended Community.

This flavor of Ethernet A-D route per EVI, associated with aliasing, can arrive at target PEs asynchronously relative to the flavor of Ethernet A-D route associated with split-horizon and mass-withdraw (i.e. per ESI). Therefore, if the Ethernet A-D route per EVI arrives ahead of the Ethernet A-D route per ESI, then the former must NOT be used for traffic forwarding till the latter arrives. This will take care of corner cases and race conditions where the Ethernet A-D route associated with mass-withdraw is withdrawn but a PE still receives the route associated with aliasing.

Backup-Path is a closely related function, albeit it applies to the case where the redundancy mode is Active/Standby. In this case, the

PE advertises that it has reachability to a given locally attached Ethernet Segment using the Ethernet A-D route as well. Remote PEs which receive the MAC advertisement routes, with non-reserved ESI, MUST consider the MAC address as reachable via the advertising PE. Furthermore, the remote PEs SHOULD install a Backup-Path, for said MAC, to the PE which had advertised reachability to the relevant Segment using (1) an Ethernet A-D routes per EVI with the same ESI (and Ethernet Tag if applicable) AND (2) Ethernet A-D routes per ESI with the same ESI and with the Active/Standby bit set to 1 in the ESI Label Extended Community.

9.4.1 Constructing the Ethernet A-D Route per EVI

This section describes procedures to construct the Ethernet A-D route when one or more such routes are advertised by an PE for a given EVI. This flavor of the Ethernet A-D route is used for aliasing, and support of this route flavor is OPTIONAL.

Route-Distinguisher (RD) MUST be set to the RD of the EVI that is advertising the NLRI. An RD MUST be assigned for a given EVI on an PE. This RD MUST be unique across all EVIs on an PE. It is RECOMMENDED to use the 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. This number may be generated by the PE. Or in the Unique VLAN EVPN case, the low order 12 bits may be the 12 bit VLAN ID, with the remaining high order 4 bits set to 0.

The Ethernet Segment Identifier MUST be a ten octet entity as described in section "Ethernet Segment Identifier". This document does not specify the use of the Ethernet A-D route when the Segment Identifier is set to 0.

The Ethernet Tag ID is the identifier of an Ethernet Tag on the Ethernet segment. This value may be a 12 bit VLAN ID, in which case the low order 12 bits are set to the VLAN ID and the high order 20 bits are set to 0. Or it may be another Ethernet Tag used by the EVPN. It MAY be set to the default Ethernet Tag on the Ethernet segment or to the value 0.

Note that the above allows the Ethernet A-D route to be advertised with one of the following granularities:

- + One Ethernet A-D route for a given <ESI, Ethernet Tag ID> tuple per EVI. This is applicable when the PE uses MPLS-based disposition.
- + One Ethernet A-D route per <ESI, EVI> (where the Ethernet Tag ID is set to 0). This is applicable when the PE uses

MAC-based disposition, or when the PE uses MPLS-based disposition when no VLAN translation is required.

The usage of the MPLS label is described in the section on "Load Balancing of Unicast Packets".

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

9.4.1.1 Ethernet A-D Route Targets

The Ethernet A-D route MUST carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically.

If an PE uses Route Target Constrain [RT-CONSTRAIN], the PE SHOULD advertise all such RTs using Route Target Constrains. The use of RT Constrains allows each Ethernet A-D route to reach only those PEs that are configured to import at least one RT from the set of RTs carried in the Ethernet A-D route.

9.4.1.1.1 Auto-Derivation from the Ethernet Tag ID

The following is the procedure for deriving the RT attribute automatically from the Ethernet Tag ID associated with the advertisement:

- + The Global Administrator field of the RT MUST be set to the Autonomous System (AS) number that the PE belongs to.
- + The Local Administrator field of the RT contains a 4 octets long number that encodes the Ethernet Tag-ID. If the Ethernet Tag-ID is a two octet VLAN ID then it MUST be encoded in the lower two octets of the Local Administrator field and the higher two octets MUST be set to zero.

For the "Unique VLAN EVPN" this results in auto-deriving the RT from the Ethernet Tag, e.g., VLAN ID for that EVPN.

9.5 Designated Forwarder Election

Consider a CE that is a host or a router that is multi-homed directly to more than one PE in an EVPN instance on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

- Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the CE.
- Flooding unknown unicast traffic (i.e. traffic for which an PE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the CE, if the environment requires flooding of unknown unicast traffic.

Note that this behavior, which allows selecting a DF at the granularity of <ESI, EVI> for multicast, broadcast and unknown unicast traffic, is the default behavior in this specification.

Note that a CE always sends packets belonging to a specific flow using a single link towards an PE. For instance, if the CE is a host then, as mentioned earlier, the host treats the multiple links that it uses to reach the PEs as a Link Aggregation Group (LAG). The CE employs a local hashing function to map traffic flows onto links in the LAG.

If a bridged network is multi-homed to more than one PE in an EVPN network via switches, then the support of All-Active points of attachments, as described in this specification, requires the bridge network to be connected to two or more PEs using a LAG. In this case the reasons for doing DF election are the same as those described above when a CE is a host or a router.

If a bridged network does not connect to the PEs using LAG, then only one of the links between the switched bridged network and the PEs must be the active link for a given Ethernet Tag. In this case, the Ethernet A-D route per Ethernet segment MUST be advertised with the "Active-Standby" flag set to one. Procedures for supporting All-Active points of attachments, when a bridge network connects to the PEs using LAG, are for further study.

The default procedure for DF election at the granularity of <ESI, EVI> is referred to as "service carving". With service carving, it is possible to elect multiple DFs per Ethernet Segment (one per EVI) in order to perform load-balancing of multi-destination traffic destined to a given Segment. The load-balancing procedures carve up the EVI space among the PE nodes evenly, in such a way that every PE is the DF for a disjoint set of EVIs. The procedure for service carving is as follows:

1. When a PE discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.

2. The PE then starts a timer (default value = 3 seconds) to allow the reception of Ethernet Segment routes from other PE nodes connected to the same Ethernet Segment. This timer value MUST be same across all PEs connected to the same Ethernet Segment.

3. When the timer expires, each PE builds an ordered list of the IP addresses of all the PE nodes connected to the Ethernet Segment (including itself), in increasing numeric value. Each IP address in this list is extracted from the "Originator Router's IP address" field of the advertised Ethernet Segment route. Every PE is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the numerically lowest IP address. The ordinals are used to determine which PE node will be the DF for a given EVPN instance on the Ethernet Segment using the following rule: Assuming a redundancy group of N PE nodes, the PE with ordinal i is the DF for an EVPN instance with an associated Ethernet Tag value V when $(V \bmod N) = i$. In the case where multiple Ethernet Tags are associated with a single EVPN instance, then the numerically lowest Ethernet Tag value in that EVPN instance MUST be used in the modulo function.

It should be noted that using "Originator Router's IP address" field in the Ethernet Segment route to get the PE IP address needed for the ordered list, allows for a CE to be multi-homed across different ASes if such need every arises.

4. The PE that is elected as a DF for a given EVPN instance will unblock traffic for the Ethernet Tags associated with that EVPN instance. Note that the DF PE unblocks multi-destination traffic in the egress direction towards the Segment. All non-DF PEs continue to drop multi-destination traffic (for the associated EVPN instances) in the egress direction towards the Segment.

In the case of link or port failure, the affected PE withdraws its Ethernet Segment route. This will re-trigger the service carving procedures on all the PEs in the RG. For PE node failure, or upon PE commissioning or decommissioning, the PEs re-trigger the service carving. In case of a Single-Active multi-homing, when a service moves from one PE in the RG to another PE as a result of re-carving, the PE, which ends up being the elected DF for the service, must trigger a MAC address flush notification towards the associated Ethernet Segment. This can be done, for e.g. using IEEE 802.1ak MVRP 'new' declaration.

9.6. Interoperability with Single-homing PEs

Let's refer to PEs that only support single-homed CE devices as

single-homing PEs. For single-homing PEs, all the above multi-homing procedures can be omitted; however, to allow for single-homing PEs to fully inter-operate with multi-homing PEs, some of the multi-homing procedures described above SHOULD be supported even by single-homing PEs:

- procedures related to processing Ethernet A-D route for the purpose of Fast Convergence (9.2 Fast Convergence), to let single-homing PEs benefit from fast convergence
- procedures related to processing Ethernet A-D route for the purpose of Aliasing (9.4 Aliasing and Backup-path), to let single-homing PEs benefit from load balancing
- procedures related to processing Ethernet A-D route for the purpose of Backup-path (9.4 Aliasing and Backup-path), to let single-homing PEs to benefit from the corresponding convergence improvement

10. Determining Reachability to Unicast MAC Addresses

PEs forward packets that they receive based on the destination MAC address. This implies that PEs must be able to learn how to reach a given destination unicast MAC address.

There are two components to MAC address learning, "local learning" and "remote learning":

10.1. Local Learning

A particular PE must be able to learn the MAC addresses from the CEs that are connected to it. This is referred to as local learning.

The PEs in a particular EVPN instance MUST support local data plane learning using standard IEEE Ethernet learning procedures. An PE must be capable of learning MAC addresses in the data plane when it receives packets such as the following from the CE network:

- DHCP requests
- ARP request for its own MAC.
- ARP request for a peer.

Alternatively PEs MAY learn the MAC addresses of the CEs in the control plane or via management plane integration between the PEs and the CEs.

There are applications where a MAC address that is reachable via a given PE on a locally attached Segment (e.g. with ESI X) may move such that it becomes reachable via another PE on another Segment (e.g. with ESI Y). This is referred to as a "MAC Mobility". Procedures to support this are described in section "MAC Mobility".

10.2. Remote learning

A particular PE must be able to determine how to send traffic to MAC addresses that belong to or are behind CEs connected to other PEs i.e. to remote CEs or hosts behind remote CEs. We call such MAC addresses as "remote" MAC addresses.

This document requires an PE to learn remote MAC addresses in the control plane. In order to achieve this, each PE advertises the MAC addresses it learns from its locally attached CEs in the control plane, to all the other PEs in that EVPN instance, using MP-BGP and specifically the MAC Advertisement route.

10.2.1. Constructing the BGP EVPN MAC Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC Advertisement route type in the EVPN NLRI.

The RD MUST be the RD of the EVI that is advertising the NLRI. The procedures for setting the RD for a given EVI are described in section 9.4.1.

The Ethernet Segment Identifier is set to the ten octet ESI described in section "Ethernet Segment".

The Ethernet Tag ID may be zero or may represent a valid Ethernet Tag ID. This field may be non-zero when there are multiple bridge domains in the MAC-VRF (e.g., the PE needs to perform qualified learning for the VLANs in that MAC-VRF).

When the the Ethernet Tag ID in the NLRI is set to a non-zero value, for a particular bridge domain, then this Ethernet Tag may either be the Ethernet tag value associated with the CE, e.g., VLAN ID, or it may be the Ethernet Tag Identifier, e.g., VLAN ID assigned by the EVPN provider and mapped to the CE's Ethernet tag. The latter would be the case if the CE Ethernet tags, e.g., VLAN ID, for a particular bridge domain are different on different CEs.

The MAC address length field is in bits and it is typically set to 48. However this specification enables specifying the MAC address as a prefix; in which case, the MAC address length field is set to the length of the prefix. This provides the ability to aggregate MAC

addresses if the deployment environment supports that. The encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1D-ORIG] [802.1D-REV]. If the MAC address is advertised as a prefix then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as 6 octets.

The IP Address field is optional. By default, the IP Address Length field is set to 0 and the IP address field is omitted from the route. When a valid IP address or address prefix needs to be advertised (e.g., for ARP suppression purposes or for inter-subnet switching), it is then encoded in this route.

The IP Address Length field is in bits and it is the length of the IP prefix. This provides the ability to advertise IP address prefixes when the deployment environment supports that. The encoding of an IP address MUST be either 4 octets for IPv4 or 16 octets for IPv6. When the IP address is advertised as a prefix, then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as either 4 or 16 octets. The length field of EVPN NLRI (which is in octets and is described in section 8) is sufficient to determine whether an IP address/prefix is encoded in this route and if so, whether the encoded IP address/prefix is IPV4 or IPV6.

The MPLS label field carries a single label and it is encoded as 3 octets, where the high-order 20 bits contain the label value. The MPLS label MUST be the downstream assigned that is used by the PE to forward MPLS-encapsulated Ethernet frames, where the destination MAC address in the Ethernet frame is the MAC address advertised in the above NLRI. The forwarding procedures are specified in section "Forwarding Unicast Packets" and "Load Balancing of Unicast Packets".

An PE may advertise the same single EVPN label for all MAC addresses in a given EVI. This label assignment methodology is referred to as a per EVI label assignment. Alternatively, an PE may advertise a unique EVPN label per <ESI, Ethernet Tag> combination. This label assignment methodology is referred to as a per <ESI, Ethernet Tag> label assignment. As a third option, an PE may advertise a unique EVPN label per MAC address. All of these methodologies have their tradeoffs. The choice of a particular label assignment methodology is purely local to the PE that originates the route.

Per EVI label assignment requires the least number of EVPN labels, but requires a MAC lookup in addition to an MPLS lookup on an egress PE for forwarding. On the other hand, a unique label per <ESI, Ethernet Tag> or a unique label per MAC allows an egress PE to forward a packet that it receives from another PE, to the connected CE, after looking up only the MPLS labels without having to perform a MAC lookup. This includes the capability to perform appropriate VLAN

ID translation on egress to the CE.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

The BGP advertisement for the MAC advertisement route MUST also carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically from the Ethernet Tag ID, in the Unique VLAN case, as described in section "Ethernet A-D Route per EVPN".

It is to be noted that this document does not require PEs to create forwarding state for remote MACs when they are learnt in the control plane. When this forwarding state is actually created is a local implementation matter.

10.2.2 Route Resolution

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to the reserved ESI value of 0 or MAX-ESI, then the receiving PE MUST install forwarding state for the associated MAC Address based on the MAC Advertisement route alone.

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, and the receiving PE is locally attached to the same ESI, then the PE does not alter its forwarding state based on the received route. This ensures that local routes are preferred to remote routes.

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, then the receiving PE MUST install forwarding state for a given MAC address only when both the MAC Advertisement route AND the associated Ethernet A-D route per ESI have been received.

To illustrate this with an example, consider two PEs (PE1 and PE2) connected to a multi-homed Ethernet Segment ES1. All-Active redundancy mode is assumed. A given MAC address M1 is learnt by PE1 but not PE2. On PE3, the following states may arise:

T1- When the MAC Advertisement Route from PE1 and the Ethernet A-D routes per ESI from PE1 and PE2 are received, PE3 can forward traffic destined to M1 to both PE1 and PE2.

T2- If after T1, PE1 withdraws its Ethernet A-D route per ESI, then PE3 forwards traffic destined to M1 to PE2 only.

T3- If after T1, PE2 withdraws its Ethernet A-D route per ESI, then

PE3 forwards traffic destined to M1 to PE1 only.

T4- If after T1, PE1 withdraws its MAC Advertisement route, then PE3 treats traffic to M1 as unknown unicast. Note, here, that had PE2 also advertised a MAC route for M1 before PE1 withdraws its MAC route, then PE3 would have continued forwarding traffic destined to M1 to PE2.

11. ARP and ND

The IP address field in the MAC advertisement route may optionally carry one of the IP addresses associated with the MAC address. This provides an option which can be used to minimize the flooding of ARP or Neighbor Discovery (ND) messages over the MPLS network and to remote CEs. This option also minimizes ARP (or ND) message processing on end-stations/hosts connected to the EVPN network. An PE may learn the IP address associated with a MAC address in the control or management plane between the CE and the PE. Or, it may learn this binding by snooping certain messages to or from a CE. When an PE learns the IP address associated with a MAC address, of a locally connected CE, it may advertise this address to other PEs by including it in the MAC Advertisement route. The IP Address may be an IPv4 address encoded using four octets, or an IPv6 address encoded using sixteen octets. The IP Address length field MUST be set to 32 for an IPv4 address or to 128 for an IPv6 address.

If there are multiple IP addresses associated with a MAC address, then multiple MAC advertisement routes MUST be generated, one for each IP address. For instance, this may be the case when there are both an IPv4 and an IPv6 address associated with the MAC address. When the IP address is dissociated with the MAC address, then the MAC advertisement route with that particular IP address MUST be withdrawn.

When an PE receives an ARP request for an IP address from a CE, and if the PE has the MAC address binding for that IP address, the PE SHOULD perform ARP proxy by responding to the ARP request.

11.1 Default Gateway

When a PE needs to perform inter-subnet forwarding where each subnet is represented by a different broadcast domain (e.g., different VLAN) the inter-subnet forwarding is performed at layer 3 and the PE that performs such function is called the default gateway. In this case when the PE receives an ARP Request for the IP address of the default gateway, the PE originates an ARP Reply.

Each PE that acts as a default gateway for a given EVPN instance MAY

advertise in the EVPN control plane its default gateway MAC address using the MAC advertisement route, and indicates that such route is associated with the default gateway. This is accomplished by requiring the route to carry the Default Gateway extended community defined in [Section 8.8 Default Gateway Extended Community]. The IP address field (4 octets for IPv4, 16 octets for IPv6) is set to zero when advertising the MAC route with the Default Gateway extended community. Both ESI and Ethernet Tag fields are also set to zero for this advertisement.

Unless it is known a priori (by means outside of this document) that all PEs of a given EVPN instance act as a default gateway for that EVPN instance, the MPLS label MUST be set to a valid downstream assigned label.

Furthermore, even if all PEs of a given EVPN instance do act as a default gateway for that EVPN instance, but only some, but not all, of these PEs have sufficient (routing) information to provide inter-subnet routing for all the inter-subnet traffic originated within the subnet associated with the EVPN instance, then when such PE advertises in the EVPN control plane its default gateway MAC address using the MAC advertisement route, and indicates that such route is associated with the default gateway, the route MUST carry a valid downstream assigned label.

If all PEs of a given EVPN instance act as a default gateway for that EVPN instance, and the same default gateway MAC address is used across all gateway devices, then no such advertisement is needed. However, if each default gateway uses a different MAC address, then each default gateway needs to be aware of other gateways' MAC addresses and thus the need for such advertisement. This is called MAC address aliasing since a single default GW can be represented by multiple MAC addresses.

Each PE that receives this route and imports it as per procedures specified in this document follows the procedures in this section when replying to ARP Requests that it receives if such Requests are for the IP address in the received EVPN route.

Each PE that acts as a default gateway for a given EVPN instance that receives this route and imports it as per procedures specified in this document MUST create MAC forwarding state that enables it to apply IP forwarding to the packets destined to the MAC address carried in the route.

12. Handling of Multi-Destination Traffic

Procedures are required for a given PE to send broadcast or multicast traffic, received from a CE encapsulated in a given Ethernet Tag (VLAN) in an EVPN instance, to all the other PEs that span that Ethernet Tag (VLAN) in that EVPN instance. In certain scenarios, described in section "Processing of Unknown Unicast Packets", a given PE may also need to flood unknown unicast traffic to other PEs.

The PEs in a particular EVPN instance may use ingress replication, P2MP LSPs or MP2MP LSPs to send unknown unicast, broadcast or multicast traffic to other PEs.

Each PE MUST advertise an "Inclusive Multicast Ethernet Tag Route" to enable the above. The following subsection provides the procedures to construct the Inclusive Multicast Ethernet Tag route. Subsequent subsections describe in further detail its usage.

12.1. Construction of the Inclusive Multicast Ethernet Tag Route

The RD MUST be the RD of the EVI that is advertising the NLRI. The procedures for setting the RD for a given EVPN instance on a PE are described in section 9.4.1.

The Ethernet Tag ID is the identifier of the Ethernet Tag. It MAY be set to 0 or to a valid Ethernet Tag value.

The Originating Router's IP address MUST be set to an IP address of the PE. This address SHOULD be common for all the EVIs on the PE (e.,g., this address may be PE's loopback address). The IP Address Length field is in bits.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router's IP Address field.

The BGP advertisement for the Inclusive Multicast Ethernet Tag route MUST also carry one or more Route Target (RT) attributes. The assignment of RTs described in the section on "Constructing the BGP EVPN MAC Address Advertisement" MUST be followed.

12.2. P-Tunnel Identification

In order to identify the P-Tunnel used for sending broadcast, unknown unicast or multicast traffic, the Inclusive Multicast Ethernet Tag route MUST carry a "PMSI Tunnel Attribute" as specified in [BGP MVPN].

Depending on the technology used for the P-tunnel for the EVPN instance on the PE, the PMSI Tunnel attribute of the Inclusive

Multicast Ethernet Tag route is constructed as follows.

- + If the PE that originates the advertisement uses a P-Multicast tree for the P-tunnel for EVPN, the PMSI Tunnel attribute MUST contain the identity of the tree (note that the PE could create the identity of the tree prior to the actual instantiation of the tree).
- + An PE that uses a P-Multicast tree for the P-tunnel MAY aggregate two or more Ethernet Tags in the same or different EVIs present on the PE onto the same tree. In this case, in addition to carrying the identity of the tree, the PMSI Tunnel attribute MUST carry an MPLS upstream assigned label which the PE has bound uniquely to the Ethernet Tag for the EVI associated with this update (as determined by its RTs).

If the PE has already advertised Inclusive Multicast Ethernet Tag routes for two or more Ethernet Tags that it now desires to aggregate, then the PE MUST re-advertise those routes. The re-advertised routes MUST be the same as the original ones, except for the PMSI Tunnel attribute and the label carried in that attribute.

- + If the PE that originates the advertisement uses ingress replication for the P-tunnel for EVPN, the route MUST include the PMSI Tunnel attribute with the Tunnel Type set to Ingress Replication and Tunnel Identifier set to a routable address of the PE. The PMSI Tunnel attribute MUST carry a downstream assigned MPLS label. This label is used to demultiplex the broadcast, multicast or unknown unicast EVPN traffic received over a MP2P tunnel by the PE.
- + The Leaf Information Required flag of the PMSI Tunnel attribute MUST be set to zero, and MUST be ignored on receipt.

13. Processing of Unknown Unicast Packets

The procedures in this document do not require the PEs to flood unknown unicast traffic to other PEs. If PEs learn CE MAC addresses via a control plane protocol, the PEs can then distribute MAC addresses via BGP, and all unicast MAC addresses will be learnt prior to traffic to those destinations.

However, if a destination MAC address of a received packet is not known by the PE, the PE may have to flood the packet. When flooding, one must take into account "split horizon forwarding" as follows: The principles behind the following procedures are borrowed from the split horizon forwarding rules in VPLS solutions [RFC4761] and

[RFC4762]. When an PE capable of flooding (say PEx) receives an unknown destination MAC address, it floods the frame. If the frame arrived from an attached CE, PEx must send a copy of the frame to every other attached CE participating in that EVPN instance, on a different ESI than the one it received the frame on, as long as the PE is the DF for the egress ESI. In addition, the PE must flood the frame to all other PEs participating in that EVPN instance. If, on the other hand, the frame arrived from another PE (say PEy), PEx must send a copy of the packet only to attached CEs as long as it is the DF for the egress ESI. PEx MUST NOT send the frame to other PEs, since PEy would have already done so. Split horizon forwarding rules apply to unknown MAC addresses.

Whether or not to flood packets to unknown destination MAC addresses should be an administrative choice, depending on how learning happens between CEs and PEs.

The PEs in a particular EVPN instance may use ingress replication using RSVP-TE P2P LSPs or LDP MP2P LSPs for sending unknown unicast traffic to other PEs. Or they may use RSVP-TE P2MP or LDP P2MP for sending such traffic to other PEs.

13.1. Ingress Replication

If ingress replication is in use, the P-Tunnel attribute, carried in the Inclusive Multicast Ethernet Tag routes for the EVPN instance, specifies the downstream label that the other PEs can use to send unknown unicast, multicast or broadcast traffic for that EVPN instance to this particular PE.

The PE that receives a packet with this particular MPLS label MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the PE MUST treat the packet as an unknown unicast packet.

13.2. P2MP MPLS LSPs

The procedures for using P2MP LSPs are very similar to VPLS procedures [VPLS-MCAST]. The P-Tunnel attribute used by an PE for sending unknown unicast, broadcast or multicast traffic for a particular EVPN instance is advertised in the Inclusive Ethernet Tag Multicast route as described in section "Handling of Multi-Destination Traffic".

The P-Tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree in [VPLS-MCAST]. Note that multiple Ethernet Tags, which may be in different EVPN instances, may use the same P2MP LSP, using upstream labels [VPLS-MCAST]. This is the

equivalent of an Aggregate Inclusive tree in [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

The PE that receives a packet on the P2MP LSP specified in the PMSI Tunnel Attribute MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the PE MUST treat the packet as an unknown unicast packet.

14. Forwarding Unicast Packets

This section describes procedures for forwarding unicast packets by PEs, where such packets are received from either directly connected CEs, or from some other PEs.

14.1. Forwarding packets received from a CE

When an PE receives a packet from a CE, on a given Ethernet Tag, it must first look up the source MAC address of the packet. In certain environments the source MAC address MAY be used to authenticate the CE and determine that traffic from the host can be allowed into the network. Source MAC lookup MAY also be used for local MAC address learning.

If the PE decides to forward the packet, the destination MAC address of the packet must be looked up. If the PE has received MAC address advertisements for this destination MAC address from one or more other PEs or learned it from locally connected CEs, it is considered as a known MAC address. Otherwise, the MAC address is considered as an unknown MAC address.

For known MAC addresses the PE forwards this packet to one of the remote PEs or to a locally attached CE. When forwarding to a remote PE, the packet is encapsulated in the EVPN MPLS label advertised by the remote PE, for that MAC address, and in the MPLS LSP label stack to reach the remote PE.

If the MAC address is unknown and if the administrative policy on the PE requires flooding of unknown unicast traffic then:

- The PE MUST flood the packet to other PEs. The PE MUST first encapsulate the packet in the ESI MPLS label as described in section 9.3. If ingress replication is used, the packet MUST be replicated one or more times to each remote PE with the outermost label being an MPLS label determined as follows: This is the MPLS label advertised by the remote PE in a PMSI Tunnel Attribute in the Inclusive Multicast Ethernet Tag route for an <EVPN instance, Ethernet Tag> combination. The Ethernet Tag in the route must be the same as the

Ethernet Tag associated with the interface on which the ingress PE receives the packet. If P2MP LSPs are being used the packet MUST be sent on the P2MP LSP that the PE is the root of for the Ethernet Tag in the EVPN instance. If the same P2MP LSP is used for all Ethernet Tags, then all the PEs in the EVPN instance MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP is used for a given Ethernet Tag in the EVPN instance, then only the PEs in the Ethernet Tag MUST be the leaves of the P2MP LSP. The packet MUST be encapsulated in the P2MP LSP label stack.

If the MAC address is unknown then, if the administrative policy on the PE does not allow flooding of unknown unicast traffic:

- The PE MUST drop the packet.

14.2. Forwarding packets received from a remote PE

This section describes the procedures for forwarding known and unknown unicast packets received from a remote PE.

14.2.1. Unknown Unicast Forwarding

When an PE receives an MPLS packet from a remote PE then, after processing the MPLS label stack, if the top MPLS label ends up being a P2MP LSP label associated with an EVPN instance or in case of ingress replication the downstream label advertised in the P-Tunnel attribute, and after performing the split horizon procedures described in section "Split Horizon":

- If the PE is the designated forwarder of BUM traffic on a particular set of ESIs for the Ethernet Tag, the default behavior is for the PE to flood the packet on these ESIs. In other words, the default behavior is for the PE to assume that for BUM traffic, it is not required to perform a destination MAC address lookup. As an option, the PE may perform a destination MAC lookup to flood the packet to only a subset of the CE interfaces in the Ethernet Tag. For instance the PE may decide to not flood an BUM packet on certain Ethernet segments even if it is the DF on the Ethernet segment, based on administrative policy.

- If the PE is not the designated forwarder on any of the ESIs for the Ethernet Tag, the default behavior is for it to drop the packet.

14.2.2. Known Unicast Forwarding

If the top MPLS label ends up being an EVPN label that was advertised in the unicast MAC advertisements, then the PE either forwards the packet based on CE next-hop forwarding information associated with

the label or does a destination MAC address lookup to forward the packet to a CE.

15. Load Balancing of Unicast Frames

This section specifies the load balancing procedures for sending known unicast frames to a multi-homed CE.

15.1. Load balancing of traffic from an PE to remote CEs

Whenever a remote PE imports a MAC advertisement for a given <ESI, Ethernet Tag> in an EVI, it MUST examine all imported Ethernet A-D routes for that ESI in order to determine the load-balancing characteristics of the Ethernet segment.

15.1.1 Single-Active Redundancy Mode

For a given ESI, if the remote PE has imported an Ethernet A-D route per Ethernet Segment from at least one PE, where the "Active-Standby" flag in the ESI Label Extended Community is set, then the remote PE MUST deduce that the Ethernet segment is operating in Single-Active redundancy mode. As such, the MAC address will be reachable only via the PE announcing the associated MAC Advertisement route - this is referred to as the primary PE. The set of other PE nodes advertising Ethernet A-D routes per Ethernet Segment for the same ESI serve as backup paths, in case the active PE encounters a failure. These are referred to as the backup PEs. It should be noted that the primary PE for a given <ESI, EVI> is the DF for that <ESI, EVI>.

If the primary PE encounters a failure, it MAY withdraw its Ethernet A-D route for the affected segment prior to withdrawing the entire set of MAC Advertisement routes.

In the case where only a single other backup PE in the network had advertised an Ethernet A-D route for the same ESI, the remote PE can then use the Ethernet A-D route withdrawal as a trigger to update its forwarding entries, for the associated MAC addresses, to point towards the backup PE. As the backup PE starts learning the MAC addresses over its attached Ethernet segment, it will start sending MAC Advertisement routes while the failed PE withdraws its own. This mechanism minimizes the flooding of traffic during fail-over events.

In the case where multiple other backup PE in the network had advertised an Ethernet A-D route for the same ESI, the remote PE MUST then use the Ethernet A-D route withdrawal as a trigger to start flooding traffic destined to the associated MAC addresses (as long as flooding of unknown unicast is administratively allowed). It is not possible to select a single backup path in this case.

15.1.1.2 All-Active Redundancy Mode

If for the given ESI, none of the Ethernet A-D routes per Ethernet Segment imported by the remote PE have the "Active-Standby" flag set in the ESI Label Extended Community, then the remote PE MUST treat the Ethernet segment as operating in All-Active redundancy mode. The remote PE would then treat the MAC address as reachable via all of the PE nodes from which it has received both an Ethernet A-D route per Ethernet Segment as well as an Ethernet A-D route per EVI for the ESI in question. The remote PE MUST use the MAC advertisement and eligible Ethernet A-D routes to construct the set of next-hops that it can use to send the packet to the destination MAC. Each next-hop comprises an MPLS label stack that is to be used by the egress PE to forward the packet. This label stack is determined as follows:

-If the next-hop is constructed as a result of a MAC route then this label stack MUST be used. However, if the MAC route doesn't exist, then the next-hop and MPLS label stack is constructed as a result of the Ethernet A-D routes. Note that the following description applies to determining the label stack for a particular next-hop to reach a given PE, from which the remote PE has received and imported Ethernet A-D routes that have the matching ESI and Ethernet Tag as the one present in the MAC advertisement. The Ethernet A-D routes mentioned in the following description refer to the ones imported from this given PE.

-If an Ethernet A-D route per Ethernet Segment for that ESI exists, together with an Ethernet A-D route per EVI, then the label from that latter route must be used.

The following example explains the above.

Consider a CE (CE1) that is dual-homed to two PEs (PE1 and PE2) on a LAG interface (ES1), and is sending packets with MAC address MAC1 on VLAN1. A remote PE, say PE3, is able to learn that MAC1 is reachable via PE1 and PE2. Both PE1 and PE2 may advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If this is not the case, and if MAC1 is advertised only by PE1, PE3 still considers MAC1 as reachable via both PE1 and PE2 as both PE1 and PE2 advertise a Ethernet A-D route per ESI for ES1 as well as an Ethernet A-D route per EVI for <ES1, VLAN1>.

The MPLS label stack to send the packets to PE1 is the MPLS LSP stack to get to PE1 and the EVPN label advertised by PE1 for CE1's MAC.

The MPLS label stack to send packets to PE2 is the MPLS LSP stack to get to PE2 and the MPLS label in the Ethernet A-D route advertised by PE2 for <ES1, VLAN1>, if PE2 has not advertised MAC1 in BGP.

We will refer to these label stacks as MPLS next-hops.

The remote PE (PE3) can now load balance the traffic it receives from its CEs, destined for CE1, between PE1 and PE2. PE3 may use N-Tuple flow information to hash traffic into one of the MPLS next-hops for load balancing of IP traffic. Alternatively PE3 may rely on the source MAC addresses for load balancing.

Note that once PE3 decides to send a particular packet to PE1 or PE2 it can pick one out of multiple possible paths to reach the particular remote PE using regular MPLS procedures. For instance, if the tunneling technology is based on RSVP-TE LSPs, and PE3 decides to send a particular packet to PE1, then PE3 can choose from multiple RSVP-TE LSPs that have PE1 as their destination.

When PE1 or PE2 receive the packet destined for CE1 from PE3, if the packet is a unicast MAC packet it is forwarded to CE1. If it is a multicast or broadcast MAC packet then only one of PE1 or PE2 must forward the packet to the CE. Which of PE1 or PE2 forward this packet to the CE is determined based on which of the two is the DF.

If the connectivity between the multi-homed CE and one of the PEs that it is attached to fails, the PE MUST withdraw the Ethernet Tag A-D routes, that had been previously advertised, for the Ethernet Segment to the CE. When the MAC entry on the PE ages out, the PE MUST withdraw the MAC address from BGP. Note that to aid convergence, the Ethernet Tag A-D routes MAY be withdrawn before the MAC routes. This enables the remote PEs to remove the MPLS next-hop to this particular PE from the set of MPLS next-hops that can be used to forward traffic to the CE. For further details and procedures on withdrawal of EVPN route types in the event of PE to CE failures please see section "PE to CE Network Failures".

15.2. Load balancing of traffic between an PE and a local CE

A CE may be configured with more than one interface connected to different PEs or the same PE for load balancing, using a technology such as LAG. The PE(s) and the CE can load balance traffic onto these interfaces using one of the following mechanisms.

15.2.1. Data plane learning

Consider that the PEs perform data plane learning for local MAC addresses learned from local CEs. This enables the PE(s) to learn a particular MAC address and associate it with one or more interfaces, if the technology between the PE and the CE supports multi-pathing. The PEs can now load balance traffic destined to that MAC address on the multiple interfaces.

Whether the CE can load balance traffic that it generates on the multiple interfaces is dependent on the CE implementation.

15.2.2. Control plane learning

The CE can be a host that advertises the same MAC address using a control protocol on both interfaces. This enables the PE(s) to learn the host's MAC address and associate it with one or more interfaces. The PEs can now load balance traffic destined to the host on the multiple interfaces. The host can also load balance the traffic it generates onto these interfaces and the PE that receives the traffic employs EVPN forwarding procedures to forward the traffic.

16. MAC Mobility

It is possible for a given host or end-station (as defined by its MAC address) to move from one Ethernet segment to another; this is referred to as 'MAC Mobility' or 'MAC move' and it is different from the multi-homing situation in which a given MAC address is reachable via multiple PEs for the same Ethernet segment. In a MAC move, there would be two sets of MAC Advertisement routes, one set with the new Ethernet segment and one set with the previous Ethernet segment, and the MAC address would appear to be reachable via each of these segments.

In order to allow all of the PEs in the EVPN instance to correctly determine the current location of the MAC address, all advertisements of it being reachable via the previous Ethernet segment **MUST** be withdrawn by the PEs, for the previous Ethernet segment, that had advertised it.

If local learning is performed using the data plane, these PEs will not be able to detect that the MAC address has moved to another Ethernet segment and the receipt of MAC Advertisement routes, with the MAC Mobility extended community attribute, from other PEs serves as the trigger for these PEs to withdraw their advertisements. If local learning is performed using the control or management planes, these interactions serve as the trigger for these PEs to withdraw their advertisements.

In a situation where there are multiple moves of a given MAC, possibly between the same two Ethernet segments, there may be multiple withdrawals and re-advertisements. In order to ensure that all PEs in the EVPN instance receive all of these correctly through the intervening BGP infrastructure, it is necessary to introduce a sequence number into the MAC Mobility extended community attribute.

An implementation **MUST** handle the scenarios where the sequence number

wraps around to process mobility event correctly.

Every MAC mobility event for a given MAC address will contain a sequence number that is set using the following rules:

- A PE advertising a MAC address for the first time advertises it with no MAC Mobility extended community attribute.
- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with a different Ethernet segment identifier advertises the MAC address in a MAC Advertisement route tagged with a MAC Mobility extended community attribute with a sequence number one greater than the sequence number in the MAC mobility attribute of the received MAC Advertisement route. In the case of the first mobility event for a given MAC address, where the received MAC Advertisement route does not carry a MAC Mobility attribute, the value of the sequence number in the received route is assumed to be 0 for purpose of this processing.
- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with the same non-zero Ethernet segment identifier advertises it with:
 - i. no MAC Mobility extended community attribute, if the received route did not carry said attribute.
 - ii. a MAC Mobility extended community attribute with the sequence number equal to the highest of the sequence number(s) in the received MAC Advertisement route(s), if the received route(s) is (are) tagged with a MAC Mobility extended community attribute.
- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with the same zero Ethernet segment identifier (single-homed scenarios) advertises it with MAC mobility extended community attribute with the sequence number set properly. In case of single-homed scenarios, there is no need for ESI comparison. The reason ESI comparison is done for multi-homing, is to prevent false detection of MAC move among the PEs attached to the same multi-homed site.

A PE receiving a MAC Advertisement route for a MAC address with a different Ethernet segment identifier and a higher sequence number than that which it had previously advertised, withdraws its MAC Advertisement route. If two (or more) PEs advertise the same MAC address with same sequence number but different Ethernet segment identifiers, a PE that receives these routes selects the route advertised by the PE with lowest IP address as the best route.

16.1. MAC Duplication Issue

A situation may arise where the same MAC address is learned by different PEs in the same VLAN because of two (or more hosts) being mis-configured with the same (duplicate) MAC address. In such situation, the traffic originating from these hosts would trigger continuous MAC moves among the PEs attached to these hosts. It is important to recognize such situation and avoid incrementing the sequence number (in the MAC Mobility attribute) to infinity. In order to remedy such situation, a PE that detects a MAC mobility event by way of local learning starts an M-second timer (default value of M = 5) and if it detects N MAC moves before the timer expires (default value for N = 3), it concludes that a duplicate MAC situation has occurred. The PE MUST alert the operator and stop sending and processing any BGP MAC Advertisement routes for that MAC address till a corrective action is taken by the operator. The values of M and N MUST be configurable to allow for flexibility in operator control. Note that the other PEs in the E-VPN instance will forward the traffic for the duplicate MAC address to one of the PEs advertising the duplicate MAC address.

16.2. Sticky MAC addresses

There are scenarios in which it is desired to configure some MAC addresses as static so that they are not subjected to MAC move. In such scenarios, these MAC addresses are advertised with MAC Mobility Extended Community where static flag is set to 1 and sequence number is set to zero. If a PE receives such advertisements and later learns the same MAC address(es) via local learning, then the PE MUST alert the operator.

17. Multicast & Broadcast

The PEs in a particular EVPN instance may use ingress replication or P2MP LSPs to send multicast traffic to other PEs.

17.1. Ingress Replication

The PEs may use ingress replication for flooding BUM traffic as described in section "Handling of Multi-Destination Traffic". A given broadcast packet must be sent to all the remote PEs. However a given multicast packet for a multicast flow may be sent to only a subset of the PEs. Specifically a given multicast flow may be sent to only those PEs that have receivers that are interested in the multicast flow. Determining which of the PEs have receivers for a given multicast flow is done using explicit tracking described below.

17.2. P2MP LSPs

An PE may use an "Inclusive" tree for sending an BUM packet. This terminology is borrowed from [VPLS-MCAST].

A variety of transport technologies may be used in the SP network. For inclusive P-Multicast trees, these transport technologies include point-to-multipoint LSPs created by RSVP-TE or mLDP.

17.2.1. Inclusive Trees

An Inclusive Tree allows the use of a single multicast distribution tree, referred to as an Inclusive P-Multicast tree, in the SP network to carry all the multicast traffic from a specified set of EVPN instances on a given PE. A particular P-Multicast tree can be set up to carry the traffic originated by sites belonging to a single EVPN instance, or to carry the traffic originated by sites belonging to different EVPN instances. The ability to carry the traffic of more than one EVPN instance on the same tree is termed 'Aggregation'. The tree needs to include every PE that is a member of any of the EVPN instances that are using the tree. This implies that an PE may receive multicast traffic for a multicast stream even if it doesn't have any receivers that are interested in receiving traffic for that stream.

An Inclusive P-Multicast tree as defined in this document is a P2MP tree. A P2MP tree is used to carry traffic only for EVPN CEs that are connected to the PE that is the root of the tree.

The procedures for signaling an Inclusive Tree are the same as those in [VPLS-MCAST] with the VPLS-AD route replaced with the Inclusive Multicast Ethernet Tag route. The P-Tunnel attribute [VPLS-MCAST] for an Inclusive tree is advertised in the Inclusive Multicast route as described in section "Handling of Multi-Destination Traffic". Note that an PE can "aggregate" multiple inclusive trees for different EVPN instances on the same P2MP LSP using upstream labels. The procedures for aggregation are the same as those described in [VPLS-MCAST], with VPLS A-D routes replaced by EVPN Inclusive Multicast routes.

18. Convergence

This section describes failure recovery from different types of network failures.

18.1. Transit Link and Node Failures between PEs

The use of existing MPLS Fast-Reroute mechanisms can provide failure recovery in the order of 50ms, in the event of transit link and node failures in the infrastructure that connects the PEs.

18.2. PE Failures

Consider a host host1 that is dual homed to PE1 and PE2. If PE1 fails, a remote PE, PE3, can discover this based on the failure of the BGP session. This failure detection can be in the sub-second range if BFD is used to detect BGP session failure. PE3 can update its forwarding state to start sending all traffic for host1 to only PE2. It is to be noted that this failure recovery is potentially faster than what would be possible if data plane learning were to be used. As in that case PE3 would have to rely on re-learning of MAC addresses via PE2.

18.2. PE to CE Network Failures

When an Ethernet segment connected to an PE fails or when a Ethernet Tag is decommissioned on an Ethernet segment, then the PE MUST withdraw the Ethernet A-D route(s) announced for the <ESI, Ethernet Tags> that are impacted by the failure or decommissioning. In addition, the PE MUST also withdraw the MAC advertisement routes that are impacted by the failure or decommissioning.

The Ethernet A-D routes should be used by an implementation to optimize the withdrawal of MAC advertisement routes. When an PE receives a withdrawal of a particular Ethernet A-D route from an PE it SHOULD consider all the MAC advertisement routes, that are learned from the same <ESI, Ethernet Tag> as in the Ethernet A-D route, from the advertising PE, as having been withdrawn. This optimizes the network convergence times in the event of PE to CE failures.

19. Frame Ordering

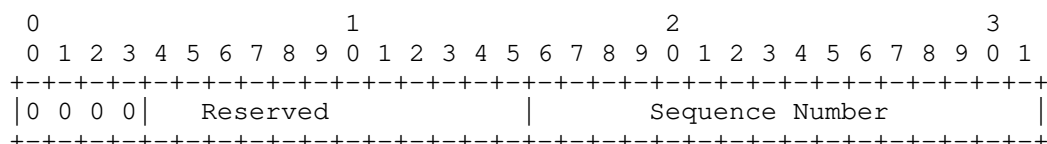
In a MAC address, bit-1 of the most significant byte is used for unicast/multicast indication and bit-2 is used for globally unique versus locally administered MAC address. If the value of the 2nd nibble (bits 4 thorough 8) of the most significant byte of the destination MAC address (which follows the last MPLS label) happens to be 0x4 or 0x6, then the Ethernet frame can be misinterpreted as an IPv4 or IPv6 packet by intermediate P nodes performing ECMP resulting in load balancing packets belonging to the same flow on different ECMP paths, thus subjecting them to different delays. Therefore, packets belonging to the same flow can arrive at the destination out of order. This out of order delivery can happen during steady state in absence of any failures resulting in significant impact to the

network operation.

In order to avoid any such mis-ordering, the usage of control word SHALL adhere to the following rules:

- A PE MUST use the control word when sending EVPN encapsulated packets over a MP2P or a P2P LSP
- A PE MUST NOT use the control word when sending EVPN encapsulated packets over a P2MP LSP

The control word is defined as follows:



In the above diagram the first 4 bits MUST be set to 0. The rest of the first 16 bits are reserved for future use. They MUST be set to 0 when transmitting, and MUST be ignored upon receipt. The next 16 bits provide a sequence number that MUST also be set to zero by default.

20. Acknowledgements

Special thanks to Yakov Rekhter for reviewing this draft several times and providing valuable comments and for his very engaging discussions on several topics of this draft that helped shape this document. We would also like to thank Pedro Marques, Kaushik Ghosh, Nischal Sheth, Robert Raszuk, Amit Shukla and Nadeem Mohammed for discussions that helped shape this document. We would also like to thank Han Nguyen for his comments and support of this work. We would also like to thank Steve Kensil and Reshad Rahman for their reviews. Last but not least, many thanks to Jakob Heitz for his help to improve several sections of this draft.

21. Security Considerations

22. IANA Considerations

23. References

23.1 Normative References

- [RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4271] Y. Rekhter et. al., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006
- [RFC4760] T. Bates et. al., "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007

23.2 Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-04.txt, July 2013.
- [VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-14.txt, July 2013.
- [RT-CONSTRAIN] P. Marques et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006

24. Author's Address

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Rahul Aggarwal
Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
e-mail: wim.henderickx@alcatel-lucent.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748
USA
Email: uttaro@att.com

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com

Ravi Shekhar
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US
Email: rshekhar@juniper.net

Florin Balus
Alcatel-Lucent
e-mail: Florin.Balus@alcatel-lucent.com

Keyur Patel
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: keyupate@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Cisco

Florin Balus
Wim Henderickx
Alcatel-Lucent

Nabil Bitar
Verizon

Clarence Filsfils
Dennis Cai
Cisco

Aldrin Isaac
Bloomberg

Lizhong Jin
ZTE

Expires: January 16, 2014

July 16, 2013

PBB-EVPN
draft-ietf-l2vpn-pbb-evpn-05

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) 2013 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.

Abstract

This document discusses how Ethernet Provider Backbone Bridging [802.1ah] can be combined with EVPN in order to reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), provide client MAC address mobility using C-MAC aggregation and B-MAC subnetting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes. The combined solution is referred to as PBB-EVPN.

Conventions

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.

Table of Contents

1. Introduction	4
2. Contributors	4
3. Terminology	4
4. Requirements	4
4.1. MAC Advertisement Route Scalability	5
4.2. C-MAC Mobility with MAC Summarization	5
4.3. C-MAC Address Learning and Confinement	5
4.4. Per Site Policy Support	6
4.5. Avoiding C-MAC Address Flushing	6
5. Solution Overview	6
6. BGP Encoding	7
6.1. BGP MAC Advertisement Route	7
6.2. Ethernet Auto-Discovery Route	7
6.3. Per VPN Route Targets	8
6.4. MAC Mobility Extended Community	8
7. Operation	8
7.1. MAC Address Distribution over Core	8
7.2. Device Multi-homing	8
7.2.1 Flow-based Load-balancing	8

7.2.1.1	PE B-MAC Address Assignment	8
7.2.1.2.	Automating B-MAC Address Assignment	10
7.2.1.3	Split Horizon and Designated Forwarder Election . .	11
7.2.2	I-SID Based Load-balancing	11
7.2.2.1	PE B-MAC Address Assignment	11
7.2.2.2	Split Horizon and Designated Forwarder Election . .	12
7.2.2.3	Handling Failure Scenarios	12
7.3.	Network Multi-homing	13
7.4.	Frame Forwarding	13
7.4.1.	Unicast	13
7.4.2.	Multicast/Broadcast	14
8.	Minimizing ARP Broadcast	14
9.	Seamless Interworking with IEEE 802.1aq/802.1Qbp	15
9.1	B-MAC Address Assignment	15
9.2	IEEE 802.1aq / 802.1Qbp B-MAC Advertisement Route	15
9.3	Operation:	16
10.	Solution Advantages	16
10.1.	MAC Advertisement Route Scalability	16
10.2.	C-MAC Mobility with MAC Sub-netting	17
10.3.	C-MAC Address Learning and Confinement	17
10.4.	Seamless Interworking with TRILL and 802.1aq Access Networks	17
10.5.	Per Site Policy Support	18
10.6.	Avoiding C-MAC Address Flushing	18
11.	Acknowledgements	19
12.	Security Considerations	19
13.	IANA Considerations	19
14.	Intellectual Property Considerations	19
15.	Normative References	19
16.	Informative References	19
17.	Authors' Addresses	19

1. Introduction

[EVPN] introduces a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the core MPLS/IP network. [802.1ah] defines an architecture for Ethernet Provider Backbone Bridging (PBB), where MAC tunneling is employed to improve service instance and MAC address scalability in Ethernet as well as VPLS networks [PBB-VPLS].

In this document, we discuss how PBB can be combined with EVPN in order to: reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), 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. The combined solution is referred to as PBB-EVPN.

2. Contributors

In addition to the authors listed above, the following individuals also contributed to this document.

Keyur Patel, Cisco
Sam Aldrin, Huawei
Himanshu Shah, Ciena

3. Terminology

BEB: Backbone Edge Bridge
B-MAC: Backbone MAC Address
CE: Customer Edge
C-MAC: Customer/Client MAC Address
DHD: Dual-homed Device
DHN: Dual-homed Network
LACP: Link Aggregation Control Protocol
LSM: Label Switched Multicast
MDT: Multicast Delivery Tree
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point
PE: Provider Edge
PoA: Point of Attachment
PW: Pseudowire
EVPN: Ethernet VPN

4. Requirements

The requirements for PBB-EVPN include all the requirements for EVPN that were described in [EVPN-REQ], in addition to the following:

4.1. MAC Advertisement Route Scalability

In typical operation, an [EVPN] PE sends a BGP MAC Advertisement Route per customer/client MAC (C-MAC) address. In certain applications, this poses scalability challenges, as is the case in virtualized data center environments where the number of virtual machines (VMs), and hence the number of C-MAC addresses, can be in the millions. In such scenarios, it is required to reduce the number of BGP MAC Advertisement routes by relying on a 'MAC summarization' scheme, as is provided by PBB. Note that the MAC summarization capability already built into EVPN is not sufficient in those environments, as will be discussed next.

4.2. C-MAC Mobility with MAC Summarization

Certain applications, such as virtual machine mobility, require support for fast C-MAC address mobility. For these applications, it is not possible to use MAC address summarization in EVPN, i.e. advertise reach-ability to a MAC address prefix. Rather, the exact virtual machine MAC address needs to be transmitted in BGP MAC Advertisement route. Otherwise, traffic would be forwarded to the wrong segment when a virtual machine moves from one Ethernet segment to another. This hinders the scalability benefits of summarization.

It is required to support C-MAC address mobility, while retaining the scalability benefits of MAC summarization. This can be achieved by leveraging PBB technology, which defines a Backbone MAC (B-MAC) address space that is independent of the C-MAC address space, and aggregate C-MAC addresses via a B-MAC address and then apply summarization to B-MAC addresses.

4.3. C-MAC Address Learning and Confinement

In EVPN, all the PE nodes participating in the same EVPN instance are exposed to all the C-MAC addresses learnt by any one of these PE nodes because a C-MAC learned by one of the PE nodes is advertised in BGP to other PE nodes in that EVPN instance. This is the case even if some of the PE nodes for that EVPN instance are not involved in forwarding traffic to, or from, these C-MAC addresses. Even if an implementation does not install hardware forwarding entries for C-MAC addresses that are not part of active traffic flows on that PE, the device memory is still consumed by keeping record of the C-MAC addresses in the routing table (RIB). In network applications with millions of C-MAC addresses, this introduces a non-trivial waste of PE resources. As such, it is required to confine the scope of

visibility of C-MAC addresses only to those PE nodes that are actively involved in forwarding traffic to, or from, these addresses.

4.4. Per Site Policy Support

In many applications, it is required to be able to enforce connectivity policy rules at the granularity of a site (or segment). This includes the ability to control which PE nodes in the network can forward traffic to, or from, a given site. PBB-EVPN is capable of providing this granularity of policy control. In the case where per C-MAC address granularity is required, the EVI can always continue to operate in EVPN mode.

4.5. Avoiding C-MAC Address Flushing

It is required to avoid C-MAC address flushing upon link, port or node failure for multi-homed devices and networks. This is in order to speed up re-convergence upon failure.

5. Solution Overview

The solution involves incorporating IEEE 802.1ah Backbone Edge Bridge (BEB) functionality on the EVPN PE nodes similar to PBB-VPLS, where BEB functionality is incorporated in the VPLS PE nodes. The PE devices would then receive 802.1Q Ethernet frames from their attachment circuits, encapsulate them in the PBB header and forward the frames over the IP/MPLS core. On the egress EVPN PE, the PBB header is removed following the MPLS disposition, and the original 802.1Q Ethernet frame is delivered to the customer equipment.

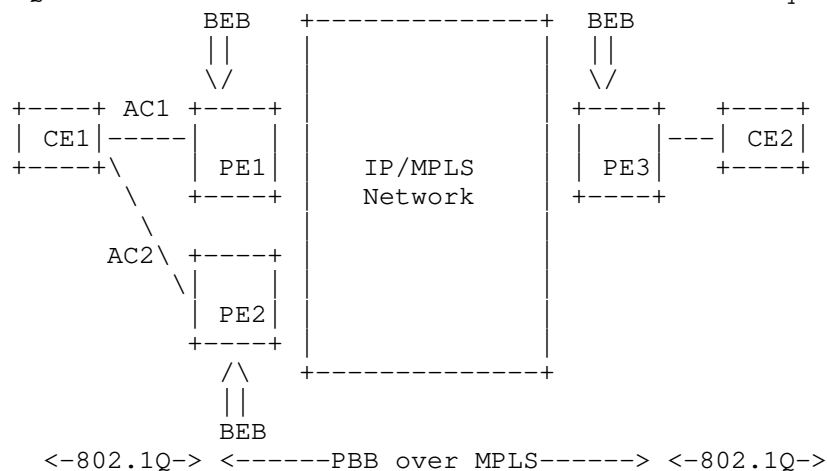


Figure 1: PBB-EVPN Network

The PE nodes perform the following functions:- Learn customer/client MAC addresses (C-MACs) over the attachment circuits in the data-plane, per normal bridge operation.

- Learn remote C-MAC to B-MAC bindings in the data-plane from traffic ingress from the core per [802.1ah] bridging operation.

- Advertise local B-MAC address reach-ability information in BGP to all other PE nodes in the same set of service instances. Note that every PE has a set of local B-MAC addresses that uniquely identify the device. More on the PE addressing in section 5.

- Build a forwarding table from remote BGP advertisements received associating remote B-MAC addresses with remote PE IP addresses and the associated MPLS label(s).

6. BGP Encoding

PBB-EVPN leverages the same BGP Routes and Attributes defined in [EVPN], adapted as follows:

6.1. BGP MAC Advertisement Route

The EVPN MAC Advertisement Route is used to distribute B-MAC addresses of the PE nodes instead of the C-MAC addresses of end-stations/hosts. This is because the C-MAC addresses are learnt in the data-plane for traffic arriving from the core. The MAC Advertisement Route is encoded as follows:

- The MAC address field contains the B-MAC address.
- The Ethernet Tag field is set to 0.
- The Ethernet Segment Identifier field must be set either to 0 (for single-homed Segments or multi-homed Segments with per-ISID load-balancing) or to MAX-ESI (for multi-homed Segments with per-flow load-balancing). All other values are not permitted.

The route is tagged with the RT corresponding to the EVI associated with the B-MAC address.

All other fields are set as defined in [EVPN].

6.2. Ethernet Auto-Discovery Route

This route and all of its associated modes are not needed in PBB-EVPN.

The receiving PE knows that it need not wait for the receipt of the Ethernet A-D route for route resolution by means of the reserved ESI

encoded in the MAC Advertisement route: the ESI values of 0 and MAX-ESI indicate that the receiving PE can resolve the path without an Ethernet A-D route.

6.3. Per VPN Route Targets

PBB-EVPN uses the same set of route targets defined in [EVPN]. The future revision of this document will describe new RT types.

6.4. MAC Mobility Extended Community

This extended community is defined in [EVPN]. When used in PBB-EVPN, it indicates that the C-MAC forwarding tables for the I-SIDs associated with the RT tagging the MAC Advertisement route must be flushed.

Note that all other BGP messages and/or attributes are used as defined in [EVPN].

7. Operation

This section discusses the operation of PBB-EVPN, specifically in areas where it differs from [EVPN].

7.1. MAC Address Distribution over Core

In PBB-EVPN, host MAC addresses (i.e. C-MAC addresses) need not be distributed in BGP. Rather, every PE independently learns the C-MAC addresses in the data-plane via normal bridging operation. Every PE has a set of one or more unicast B-MAC addresses associated with it, and those are the addresses distributed over the core in MAC Advertisement routes.

7.2. Device Multi-homing

7.2.1 Flow-based Load-balancing

This section describes the procedures for supporting device multi-homing in an all-active redundancy model with flow-based load-balancing.

7.2.1.1 PE B-MAC Address Assignment

In [802.1ah] every BEB is uniquely identified by one or more B-MAC addresses. These addresses are usually locally administered by the Service Provider. For PBB-EVPN, the choice of B-MAC address(es) for the PE nodes must be examined carefully as it has implications on the

proper operation of multi-homing. In particular, for the scenario where a CE is multi-homed to a number of PE nodes with all-active redundancy and flow-based load-balancing, a given C-MAC address would be reachable via multiple PE nodes concurrently. Given that any given remote PE will bind the C-MAC address to a single B-MAC address, then the various PE nodes connected to the same CE must share the same B-MAC address. Otherwise, the MAC address table of the remote PE nodes will keep oscillating between the B-MAC addresses of the various PE devices. For example, consider the network of Figure 1, and assume that PE1 has B-MAC BM1 and PE2 has B-MAC BM2. Also, assume that both links from CE1 to the PE nodes are part of an all-active multi-chassis Ethernet link aggregation group. If BM1 is not equal to BM2, the consequence is that the MAC address table on PE3 will keep oscillating such that the C-MAC address CM of CE1 would flip-flop between BM1 or BM2, depending on the load-balancing decision on CE1 for traffic destined to the core.

Considering that there could be multiple sites (e.g. CEs) that are multi-homed to the same set of PE nodes, then it is required for all the PE devices in a Redundancy Group to have a unique B-MAC address per site. This way, it is possible to achieve fast convergence in the case where a link or port failure impacts the attachment circuit connecting a single site to a given PE.

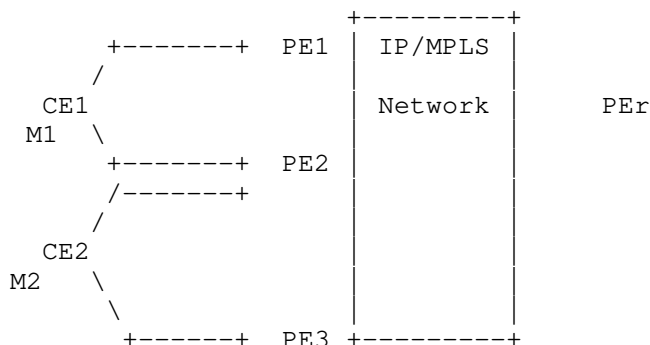


Figure 2: B-MAC Address Assignment

In the example network shown in Figure 2 above, two sites corresponding to CE1 and CE2 are dual-homed to PE1/PE2 and PE2/PE3, respectively. Assume that BM1 is the B-MAC used for the site corresponding to CE1. Similarly, BM2 is the B-MAC used for the site corresponding to CE2. On PE1, a single B-MAC address (BM1) is required for the site corresponding to CE1. On PE2, two B-MAC addresses (BM1 and BM2) are required, one per site. Whereas on PE3, a single B-MAC address (BM2) is required for the site corresponding to CE2. All three PE nodes would advertise their respective B-MAC

addresses in BGP using the MAC Advertisement routes defined in [EVPN]. The remote PE, PEr, would learn via BGP that BM1 is reachable via PE1 and PE2, whereas BM2 is reachable via both PE2 and PE3. Furthermore, PEr establishes via the normal bridge learning that C-MAC M1 is reachable via BM1, and C-MAC M2 is reachable via BM2. As a result, PEr can load-balance traffic destined to M1 between PE1 and PE2, as well as traffic destined to M2 between both PE2 and PE3. In the case of a failure that causes, for example, CE1 to be isolated from PE1, the latter can withdraw the route it has advertised for BM1. This way, PEr would update its path list for BM1, and will send all traffic destined to M1 over to PE2 only.

For single-homed sites, it is possible to assign a unique B-MAC address per site, or have all the single-homed sites connected to a given PE share a single B-MAC address. The advantage of the first model over the second model is the ability to avoid C-MAC destination address lookup on the disposition PE (even though source C-MAC learning is still required in the data-plane). Also, by assigning the B-MAC addresses from a contiguous range, it is possible to advertise a single B-MAC subnet for all single-homed sites, thereby rendering the number of MAC advertisement routes required at par with the second model.

In summary, every PE may use a unicast B-MAC address shared by all single-homed CEs or a unicast B-MAC address per single-homed CE and, in addition, a unicast B-MAC address per dual-homed CE. In the latter case, the B-MAC address MUST be the same for all PE nodes in a Redundancy Group connected to the same CE.

7.2.1.2. Automating B-MAC Address Assignment

The PE B-MAC address used for single-homed sites can be automatically derived from the hardware (using for e.g. the backplane's address). However, the B-MAC address used for multi-homed sites must be coordinated among the RG members. To automate the assignment of this latter address, the PE can derive this B-MAC address from the MAC Address portion of the CE's LACP System Identifier by flipping the 'Locally Administered' bit of the CE's address. This guarantees the uniqueness of the B-MAC address within the network, and ensures that all PE nodes connected to the same multi-homed CE use the same value for the B-MAC address.

Note that with this automatic provisioning of the B-MAC address associated with multi-homed CEs, it is not possible to support the uncommon scenario where a CE has multiple bundles towards the PE nodes, and the service involves hair-pinning traffic from one bundle to another. This is because the split-horizon filtering relies on B-MAC addresses rather than Site-ID Labels (as will be described in the

next section). The operator must explicitly configure the B-MAC address for this fairly uncommon service scenario.

Whenever a B-MAC address is provisioned on the PE, either manually or automatically (as an outcome of CE auto-discovery), the PE MUST transmit an MAC Advertisement Route for the B-MAC address with a downstream assigned MPLS label that uniquely identifies that address on the advertising PE. The route is tagged with the RTs of the associated EVIs as described above.

7.2.1.3 Split Horizon and Designated Forwarder Election

[EVPN] relies on access split horizon, where the Ethernet Segment Label is used for egress filtering on the attachment circuit in order to prevent forwarding loops. In PBB-EVPN, the B-MAC source address can be used for the same purpose, as it uniquely identifies the originating site of a given frame. As such, Segment Labels are not used in PBB-EVPN, and the egress split-horizon filtering is done based on the B-MAC source address. It is worth noting here that [802.1ah] defines this B-MAC address based filtering function as part of the I-Component options, hence no new functions are required to support split-horizon beyond what is already defined in [802.1ah]. Given that the Segment label is not used in PBB-EVPN, the PE sets the Label field in the Ethernet Segment Route to 0.

The Designated Forwarder election procedures are defined in [I-D-Segment-Route].

7.2.2 I-SID Based Load-balancing

This section describes the procedures for supporting device multi-homing in an all-active redundancy model with per-ISID load-balancing.

7.2.2.1 PE B-MAC Address Assignment

In the case where per-ISID load-balancing is desired among the PE nodes in a given redundancy group, multiple unicast B-MAC addresses are allocated per multi-homed Ethernet Segment: Each PE connected to the multi-homed segment is assigned a unique B-MAC. Every PE then advertises its B-MAC address using the BGP MAC advertisement route. In this mode of operation, two B-MAC address assignment models are possible:

- The PE may use a shared B-MAC address for multiple Ethernet Segments. This includes the single-homed segments as well as the multi-homed segments operating with per-ISID load-balancing mode.

- The PE may use a dedicated B-MAC address for each Ethernet Segment operating with per-ISID load-balancing mode.

All PE implementations MUST support the shared B-MAC address model and MAY support the dedicated B-MAC address model.

A remote PE initially floods traffic to a destination C-MAC address, located in a given multi-homed Ethernet Segment, to all the PE nodes connected to that segment. Then, when reply traffic arrives at the remote PE, it learns (in the data-path) the B-MAC address and associated next-hop PE to use for said C-MAC address.

7.2.2.2 Split Horizon and Designated Forwarder Election The procedures are similar to the flow-based load-balancing case, with the only difference being that the DF filtering must be applied to unicast as well as multicast traffic, and in both core-to-segment as well as segment-to-core directions.

7.2.2.3 Handling Failure Scenarios

When a PE connected to a multi-homed Ethernet Segment loses connectivity to the segment, due to link or port failure, it needs to notify the remote PEs to trigger C-MAC address flushing. This can be achieved in one of two ways, depending on the B-MAC assignment model:

- If the PE uses a shared B-MAC address for multiple Ethernet Segments, then the C-MAC flushing is signaled by means of having the failed PE re-advertise the MAC Advertisement route for the associated B-MAC, tagged with the MAC Mobility Extended Community attribute. The value of the Counter field in that attribute must be incremented prior to advertisement. This causes the remote PE nodes to flush all C-MAC addresses associated with the B-MAC in question. This is done across all I-SIDs that are mapped to the EVI of the withdrawn MAC route.

- If the PE uses a dedicated B-MAC address for each Ethernet Segment operating under per-ISID load-balancing mode, the the failed PE simply withdraws the B-MAC route previously advertised for that segment. This causes the remote PE nodes to flush all C-MAC addresses associated with the B-MAC in question. This is done across all I-SIDs that are mapped to the EVI of the withdrawn MAC route.

When a PE connected to a multi-homed Ethernet Segment fails (i.e. node failure) or when the PE becomes completely isolated from the EVPN network, the remote PEs will start purging the MAC Advertisement routes that were advertised by the failed PE. This is done either as an outcome of the remote PEs detecting that the BGP session to the

failed PE has gone down, or by having a Route Reflector withdrawing all the routes that were advertised by the failed PE. The remote PEs, in this case, will perform C-MAC address flushing as an outcome of the MAC Advertisement route withdrawals.

For all failure scenarios (link/port failure, node failure and PE node isolation), when the fault condition clears, the recovered PE re-advertises the associated Ethernet Segment route to other members of its Redundancy Group. This triggers the backup PE(s) in the Redundancy Group to block the I-SIDs for which the recovered PE is a DF. When a backup PE blocks the I-SIDs, it triggers a C-MAC address flush notification to the remote PEs by re-advertising the MAC Advertisement route for the associated B-MAC, with the MAC Mobility Extended Community attribute. The value of the Counter field in that attribute must be incremented prior to advertisement. This causes the remote PE nodes to flush all C-MAC addresses associated with the B-MAC in question. This is done across all I-SIDs that are mapped to the EVI of the withdrawn MAC route.

7.3. Network Multi-homing

When an Ethernet network is multi-homed to a set of PE nodes running PBB-EVPN, an all-active redundancy model can be supported with per service instance (i.e. I-SID) load-balancing. In this model, DF election is performed to ensure that a single PE node in the redundancy group is responsible for forwarding traffic associated with a given I-SID. This guarantees that no forwarding loops are created. Filtering based on DF state applies to both unicast and multicast traffic, and in both access-to-core as well as core-to-access directions (unlike the multi-homed device scenario where DF filtering is limited to multi-destination frames in the core-to-access direction). Similar to the multi-homed device scenario, with I-SID based load-balancing, a unique B-MAC address is assigned to each of the PE nodes connected to the multi-homed network (Segment).

7.4. Frame Forwarding

The frame forwarding functions are divided in between the Bridge Module, which hosts the [802.1ah] Backbone Edge Bridge (BEB) functionality, and the MPLS Forwarder which handles the MPLS imposition/disposition. The details of frame forwarding for unicast and multi-destination frames are discussed next.

7.4.1. Unicast

Known unicast traffic received from the AC will be PBB-encapsulated by the PE using the B-MAC source address corresponding to the originating site. The unicast B-MAC destination address is determined

based on a lookup of the C-MAC destination address (the binding of the two is done via transparent learning of reverse traffic). The resulting frame is then encapsulated with an LSP tunnel label and the MPLS label which uniquely identifies the B-MAC destination address on the egress PE. If per flow load-balancing over ECMPs in the MPLS core is required, then a flow label is added as the end of stack label.

For unknown unicast traffic, the PE forwards these frames over MPLS core. When these frames are to be forwarded, then the same set of options used for forwarding multicast/broadcast frames (as described in next section) are used.

7.4.2. Multicast/Broadcast

Multi-destination frames received from the AC will be PBB-encapsulated by the PE using the B-MAC source address corresponding to the originating site. The multicast B-MAC destination address is selected based on the value of the I-SID as defined in [802.1ah]. The resulting frame is then forwarded over the MPLS core using one out of the following two options:

Option 1: the MPLS Forwarder can perform ingress replication over a set of MP2P tunnel LSPs. The frame is encapsulated with a tunnel LSP label and the EVPN ingress replication label advertised in the Inclusive Multicast Route.

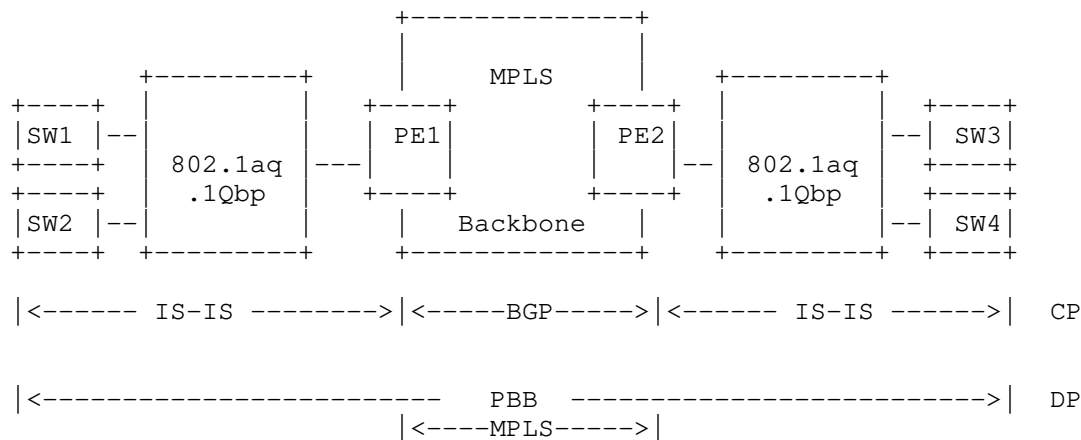
Option 2: the MPLS Forwarder can use P2MP tunnel LSP per the procedures defined in [EVPN]. This includes either the use of Inclusive or Aggregate Inclusive trees.

Note that the same procedures for advertising and handling the Inclusive Multicast Route defined in [EVPN] apply here.

8. Minimizing ARP Broadcast

The PE nodes implement an ARP-proxy function in order to minimize the volume of ARP traffic that is broadcasted over the MPLS network. This is achieved by having each PE node snoop on ARP request and response messages received over the access interfaces or the MPLS core. The PE builds a cache of IP / MAC address bindings from these snooped messages. The PE then uses this cache to respond to ARP requests ingress on access ports and targeting hosts that are in remote sites. If the PE finds a match for the IP address in its ARP cache, it responds back to the requesting host and drops the request. Otherwise, if it does not find a match, then the request is flooded over the MPLS network using either ingress replication or LSM.

9. Seamless Interworking with IEEE 802.1aq/802.1Qbp



Legend: CP = Control Plane View
DP = Data Plane View

Figure 7: Interconnecting 802.1aq/802.1Qbp Networks with PBB-EVPN

9.1 B-MAC Address Assignment

For the same reasons cited in the TRILL section, the B-MAC addresses need to be globally unique across all the IEEE 802.1aq / 802.1Qbp networks. The same hierarchical address assignment scheme depicted above is proposed for B-MAC addresses as well.

9.2 IEEE 802.1aq / 802.1Qbp B-MAC Advertisement Route

B-MAC addresses associated with 802.1aq / 802.1Qbp switches are advertised using the BGP MAC Advertisement route already defined in [EVPN].

The encapsulation for the transport of PBB frames over MPLS is similar to that of classical Ethernet, albeit with the additional PBB header, as shown in the figure below:

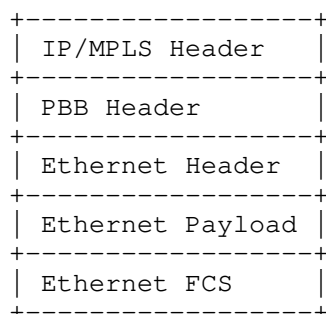


Figure 8: PBB over MPLS Encapsulation

9.3 Operation:

When a PE receives a PBB-encapsulated Ethernet frame from the access side, it performs a lookup on the B-MAC destination address to identify the next hop. If the lookup yields that the next hop is a remote PE, the local PE would then encapsulate the PBB frame in MPLS. The label stack comprises of the VPN label (advertised by the remote PE), followed by an LSP/IGP label. From that point onwards, regular MPLS forwarding is applied.

On the disposition PE, assuming penultimate-hop-popping is employed, the PE receives the MPLS-encapsulated PBB frame with a single label: the VPN label. The value of the label indicates to the disposition PE that this is a PBB frame, so the label is popped, the TTL field (in the 802.1Qbp F-Tag) is reinitialized and normal PBB processing is employed from this point onwards.

10. Solution Advantages

In this section, we discuss the advantages of the PBB-EVPN solution in the context of the requirements set forth in section 3 above.

10.1. MAC Advertisement Route Scalability

In PBB-EVPN the number of MAC Advertisement Routes is a function of the number of segments (sites), rather than the number of hosts/servers. This is because the B-MAC addresses of the PEs, rather than C-MAC addresses (of hosts/servers) are being advertised in BGP. And, as discussed above, there's a one-to-one mapping between multi-homed segments and B-MAC addresses, whereas there's a one-to-one or many-to-one mapping between single-homed segments and B-MAC addresses for a given PE. As a result, the volume of MAC Advertisement Routes in PBB-EVPN is multiple orders of magnitude less than EVPN.

10.2. C-MAC Mobility with MAC Sub-netting

In PBB-EVPN, if a PE allocates its B-MAC addresses from a contiguous range, then it can advertise a MAC prefix rather than individual 48-bit addresses. It should be noted that B-MAC addresses can easily be assigned from a contiguous range because PE nodes are within the provider administrative domain; however, CE devices and hosts are typically not within the provider administrative domain. The advantage of such MAC address sub-netting can be maintained even as C-MAC addresses move from one Ethernet segment to another. This is because the C-MAC address to B-MAC address association is learnt in the data-plane and C-MAC addresses are not advertised in BGP. To illustrate how this compares to EVPN, consider the following example:

If a PE running EVPN advertises reachability for a MAC subnet that spans N addresses via a particular segment, and then 50% of the MAC addresses in that subnet move to other segments (e.g. due to virtual machine mobility), then in the worst case, N/2 additional MAC Advertisement routes need to be sent for the MAC addresses that have moved. This defeats the purpose of the sub-netting. With PBB-EVPN, on the other hand, the sub-netting applies to the B-MAC addresses which are statically associated with PE nodes and are not subject to mobility. As C-MAC addresses move from one segment to another, the binding of C-MAC to B-MAC addresses is updated via data-plane learning.

10.3. C-MAC Address Learning and Confinement

In PBB-EVPN, C-MAC address reachability information is built via data-plane learning. As such, PE nodes not participating in active conversations involving a particular C-MAC address will purge that address from their forwarding tables. Furthermore, since C-MAC addresses are not distributed in BGP, PE nodes will not maintain any record of them in control-plane routing table.

10.4. Seamless Interworking with TRILL and 802.1aq Access Networks

Consider the scenario where two access networks, one running MPLS and the other running 802.1aq, are interconnected via an MPLS backbone network. The figure below shows such an example network.

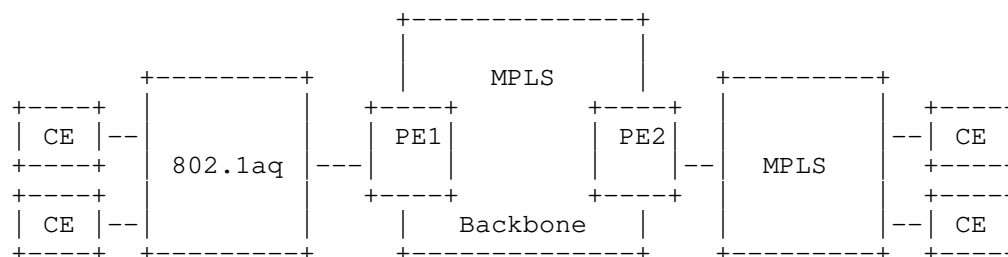


Figure 9: Interoperability with 802.1aq

If the MPLS backbone network employs EVPN, then the 802.1aq data-plane encapsulation must be terminated on PE1 or the edge device connecting to PE1. Either way, all the PE nodes that are part of the associated service instances will be exposed to all the C-MAC addresses of all hosts/servers connected to the access networks. However, if the MPLS backbone network employs PBB-EVPN, then the 802.1aq encapsulation can be extended over the MPLS backbone, thereby maintaining C-MAC address transparency on PE1. If PBB-EVPN is also extended over the MPLS access network on the right, then C-MAC addresses would be transparent to PE2 as well.

Interoperability with TRILL access network will be described in future revision of this draft.

10.5. Per Site Policy Support

In PBB-EVPN, a unique B-MAC address can be associated with every site (single-homed or multi-homed). Given that the B-MAC addresses are sent in BGP MAC Advertisement routes, it is possible to define per site (i.e. B-MAC) forwarding policies including policies for E-TREE service.

10.6. Avoiding C-MAC Address Flushing

With PBB-EVPN, it is possible to avoid C-MAC address flushing upon topology change affecting a multi-homed device. To illustrate this, consider the example network of Figure 1. Both PE1 and PE2 advertise the same B-MAC address (BM1) to PE3. PE3 then learns the C-MAC addresses of the servers/hosts behind CE1 via data-plane learning. If AC1 fails, then PE3 does not need to flush any of the C-MAC addresses learnt and associated with BM1. This is because PE1 will withdraw the MAC Advertisement routes associated with BM1, thereby leading PE3 to have a single adjacency (to PE2) for this B-MAC address. Therefore, the topology change is communicated to PE3 and no C-MAC address flushing is required.

11. Acknowledgements

TBD.

12. Security Considerations

There are no additional security aspects beyond those of VPLS/H-VPLS that need to be discussed here.

13. IANA Considerations

This document requires IANA to assign a new SAFI value for L2VPN_MAC SAFI.

14. Intellectual Property Considerations

This document is being submitted for use in IETF standards discussions.

15. Normative References

[802.1ah] "Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges", IEEE Std. 802.1ah-2008, August 2008.

16. Informative References

[PBB-VPLS] Sajassi et al., "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-pbb-vpls-interop-05.txt, work in progress, July, 2011.

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (EVPN)", draft-ietf-l2vpn-evpn-req-04.txt, work in progress, July, 2011.

[EVPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, February, 2012.

17. Authors' Addresses

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.com

Samer Salam
Cisco
595 Burrard Street, Suite # 2123
Vancouver, BC V7X 1J1, Canada
Email: ssalam@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Florin Balus
Alcatel-Lucent
701 E. Middlefield Road
Mountain View, CA, USA 94043
Email: florin.balus@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.be

Clarence Filsfils
Cisco
Email: cfilsfil@cisco.com

Dennis Cai
Cisco
Email: dcai@cisco.com

Lizhong Jin
ZTE Corporation

889, Bibo Road
Shanghai, 201203, China
Email: lizhong.jin@zte.com.cn

Network Working Group
Internet-Draft
Updates: 4761 (if approved)
Intended status: Standards Track
Expires: August 29, 2013

B. Kothari
Cohere Networks
K. Kompella
Juniper Networks
W. Henderickx
F. Balus
Alcatel-Lucent
J. Uttaro
AT&T
S. Palislaamovic
Alcatel-Lucent
W. Lin
Juniper Networks
February 25, 2013

BGP based Multi-homing in Virtual Private LAN Service
draft-ietf-l2vpn-vpls-multihoming-05.txt

Abstract

Virtual Private LAN Service (VPLS) is a Layer 2 Virtual Private Network (VPN) that gives its customers the appearance that their sites are connected via a Local Area Network (LAN). It is often required for the Service Provider (SP) to give the customer redundant connectivity to some sites, often called "multi-homing". This memo shows how BGP-based multi-homing can be offered in the context of LDP and BGP VPLS solutions.

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 August 29, 2013.

Copyright Notice

Copyright (c) 2013 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. General Terminology	4
1.2. Conventions	5
2. Background	6
2.1. Scenarios	6
2.2. VPLS Multi-homing Considerations	7
3. Multi-homing Operation	8
3.1. Multi-homing NLRI	8
3.2. Provisioning Model	9
3.3. Designated Forwarder Election	10
3.3.1. Attributes	10
3.3.2. Variables Used	11
3.3.3. Election Procedures	12
3.4. DF Election on PEs	14
4. Multi-AS VPLS	15
4.1. Route Origin Extended Community	15
4.2. VPLS Preference	15
4.3. Use of BGP-MH attributes in Inter-AS Methods	16
4.3.1. Inter-AS Method (b): EBGW Redistribution of VPLS Information between ASBRs	16
4.3.2. Inter-AS Method (c): Multi-Hop EBGW Redistribution of VPLS Information between ASes	17
5. MAC Flush Operations	19
5.1. MAC List Flush	19
5.2. Implicit MAC Flush	19
5.3. Minimizing the effects of fast link transitions	20
6. Backwards Compatibility	21
6.1. BGP based VPLS	21
6.2. LDP VPLS with BGP Auto-discovery	21
7. Security Considerations	22
8. IANA Considerations	23
9. Acknowledgments	24
10. References	25
10.1. Normative References	25
10.2. Informative References	25
Authors' Addresses	26

1. Introduction

Virtual Private LAN Service (VPLS) is a Layer 2 Virtual Private Network (VPN) that gives its customers the appearance that their sites are connected via a Local Area Network (LAN). It is often required for a Service Provider (SP) to give the customer redundant connectivity to one or more sites, often called "multi-homing". [RFC4761] explains how VPLS can be offered using BGP for auto-discovery and signaling; section 3.5 of that document describes how multi-homing can be achieved in this context. [RFC6074] explains how VPLS can be offered using BGP for auto-discovery (BGP-AD) and [RFC4762] explains how VPLS can be offered using LDP for signaling. This document provides a BGP-based multi-homing solution applicable to both BGP and LDP VPLS technologies. Note that BGP MH can be used for LDP VPLS without the use of the BGP-AD solution.

Section 2 lays out some of the scenarios for multi-homing, other ways that this can be achieved, and some of the expectations of BGP-based multi-homing. Section 3 defines the components of BGP-based multi-homing, and the procedures required to achieve this. Section 7 may someday discuss security considerations.

1.1. General Terminology

Some general terminology is defined here; most is from [RFC4761], [RFC4762] or [RFC4364]. Terminology specific to this memo is introduced as needed in later sections.

A "Customer Edge" (CE) device, typically located on customer premises, connects to a "Provider Edge" (PE) device, which is owned and operated by the SP. A "Provider" (P) device is also owned and operated by the SP, but has no direct customer connections. A "VPLS Edge" (VE) device is a PE that offers VPLS services.

A VPLS domain represents a bridging domain per customer. A Route Target community as described in [RFC4360] is typically used to identify all the PE routers participating in a particular VPLS domain. A VPLS site is a grouping of ports on a PE that belong to the same VPLS domain. A Multi-homed (MH) site is uniquely identified by a MH site ID (MH-ID). Sites are referred to as local or remote depending on whether they are configured on the PE router in context or on one of the remote PE routers (network peers). The terms "VPLS instance" and "VPLS domain" are used interchangeably in this document.

1.2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2.2. VPLS Multi-homing Considerations

The first (perhaps obvious) fact about a multi-homed VPLS CE, such as CE1 in Figure 1 is that if CE1 is an Ethernet switch or bridge, a loop has been created in the customer VPLS. This is a dangerous situation for an Ethernet network, and the loop must be broken. Even if CE1 is a router, it will get duplicates every time a packet is flooded, which is clearly undesirable.

The next is that (unlike the case of IP-based multi-homing) only one of PE1 and PE2 can be actively sending traffic, either towards CE1 or into the SP cloud. That is to say, load balancing techniques will not work. All other PEs MUST choose the same designated forwarder for a multi-homed site. Call the PE that is chosen to send traffic to/from CE1 the "designated forwarder".

In Figure 2, CE1 and CE4 must be dealt with independently, since CE1 is dual-homed, but CE4 is not.

3. Multi-homing Operation

This section describes procedures for electing a designated forwarder among the set of PEs that are multi-homed to a customer site. The procedures described in this section are applicable to BGP based VPLS, LDP based VPLS with BGP-AD or a VPLS that contains a mix of both BGP and LDP signaled PWs.

3.1. Multi-homing NLRI

Section 3.2.2 in [RFC4761] specifies a NLRI to be used for BGP based VPLS (BGP VPLS NLRI). The format of the BGP VPLS NLRI is shown below.

Length (2 octets)
Route Distinguisher (8 octets)
VE ID (2 octets)
VE Block Offset (2 octets)
VE Block Size (2 octets)
Label Base (3 octets)

BGP VPLS NLRI

For multi-homing operation, a multi-homing NLRI (MH NLRI) is proposed that uses BGP VPLS NLRI with the following fields set to zero: VE Block Offset, VE Block Size and Label Base. In addition, the VE-ID field of the NLRI is set to MH-ID. Thus, the MH NLRI contains 2 octets indicating the length, 8 octets for Route Distinguisher, 2 octets for MH-ID and 7 octets with value zero.

It is valid to have non-zero VE block offset, VE block size and label base in the VPLS NLRI for a multi-homed site. VPLS operations, including multi-homing, in such a case are outside the scope of this document. However, for interoperability with existing deployments that use non-zero VE block offset, VE block size and label base for multi-homing operation, Section 6.1 provides more detail.

3.2. Provisioning Model

It is mandatory that each instance within a VPLS domain MUST be provisioned with a unique Route Distinguisher value. Unique Route Distinguisher allows VPLS advertisements from different VPLS PEs to be distinct even if the advertisements have the same VE-ID, which can occur in case of multi-homing. This allows standard BGP path selection rules to be applied to VPLS advertisements.

Each VPLS PE must advertise a unique VE-ID with non-zero VE Block Offset, VE Block Size and Label Base values in the BGP NLRI. VE-ID is associated with the base VPLS instance and the NLRI associated with it must be used for creating PWs among VPLS PEs. Any single homed customer sites connected to the VPLS instance do not require any special addressing. Any multi-homed customer sites connected to the VPLS instance require special addressing, which is achieved by use of MH-ID. A set of customer sites are distinguished as multi-homed if they all have the same MH-ID. The following examples illustrate the use of VE-ID and MH-ID.

Figure 1 shows a customer site, CE1, multi-homed to two VPLS PEs, PE1 and PE2. In order for all VPLS PEs to set up PWs to each other, each VPLS PE must be configured with a unique VE-ID for its base VPLS instance. In addition, in order for all VPLS PEs within the same VPLS domain to elect one of the multi-homed PEs as the designated forwarder, an indicator that the PEs are multi-homed to the same customer site is required. This is achieved by assigning the same multi-homed site ID (MH-ID) on PE1 and PE2 for CE1. When remote VPLS PEs receive NLRI advertisement from PE1 and PE2 for CE1, the two NLRI advertisements for CE1 are identified as candidates for designated forwarder selection due to the same MH-ID. Thus, same MH-ID MUST be assigned on all VPLS PEs that are multi-homed to the same customer site.

Figure 2 shows two customer sites, CE1 and CE4, connected to PE1 with CE1 multi-homed to PE1 and PE2. Similar to Figure 1 provisioning model, each VPLS PE must be configured with a unique VE-ID for its base VPLS instance. CE4 does not require special addressing on PE1. However, CE1 which is multi-homed to PE1 and PE2 requires configuration of MH-ID and both PE1 and PE2 MUST be provisioned with the same MH-ID for CE1.

Note that a MH-ID=0 is invalid and a PE should discard such an advertisement.

Use of multiple VE-IDs per VPLS instance for either multi-homing operation or for any other purpose is outside the scope of this document. However, for interoperability with existing deployments

that use multiple VE-IDs, Section 6.1 provides more detail.

3.3. Designated Forwarder Election

BGP-based multi-homing for VPLS relies on standard BGP path selection and VPLS DF election. The net result of doing both BGP path selection and VPLS DF election is that of electing a single designated forwarder (DF) among the set of PEs to which a customer site is multi-homed. All the PEs that are elected as non-designated forwarders MUST keep their attachment circuit to the multi-homed CE in blocked status (no forwarding).

These election algorithms operate on VPLS advertisements, which include both the NLRI and attached BGP attributes. These election algorithms are applicable to all VPLS NLRIs, and not just to MH NLRIs. In order to simplify the explanation of these algorithms, we will use a number of variables derived from fields in the VPLS advertisement. These variables are: RD, SITE-ID, VBO, DOM, ACS, PREF and PE-ID. The notation ADV -> <RD, SITE-ID, VBO, DOM, ACS, PREF, PE-ID> means that from a received VPLS advertisement ADV, the respective variables were derived. The following sections describe two attributes needed for DF election, then describe the variables and how they are derived from fields in VPLS advertisement ADV, and finally describe how DF election is done.

3.3.1. Attributes

The procedures below refer to two attributes: the Route Origin community (see Section 4.1) and the L2-info community (see Section 4.2). These attributes are required for inter-AS operation; for generality, the procedures below show how they are to be used. The procedures also outline how to handle the case that either or both are not present.

For BGP-based Multi-homing, ADV MUST contain an L2-info extended community as specified in [RFC4761]. Within this community are various control flags. Two new control flags are proposed in this document. Figure 3 shows the position of the new 'D' and 'F' flags.

Control Flags Bit Vector

```

0 1 2 3 4 5 6 7
+---+---+---+---+
|D|Z|F|Z|Z|Z|C|S| (Z = MUST Be Zero)
+---+---+---+---+

```

Figure 3

1. 'D' (Down): Indicates connectivity status between a CE site and a VPLS PE. The bit MUST be set to one if all the attachment circuits connecting a CE site to a VPLS PE are down.
2. 'F' (Flush): Indicates when to flush MAC state. A designated forwarder must set the F bit and a non-designated forwarder must clear the F bit when sending BGP MH advertisements. A state transition from one to zero for the F bit can be used by a remote PE to flush all the MACs learned from the PE that is transitioning from designated forwarder to non-designated forwarder. Refer to Section 5.2 for more details on the use case.

3.3.2. Variables Used

3.3.2.1. RD

RD is simply set to the Route Distinguisher field in the NLRI part of ADV.

3.3.2.2. SITE-ID

SITE-ID is simply set to the VE-ID field in the NLRI part of the ADV.

Note that no distinction is made whether VE-ID is for a multi-homed site or not.

3.3.2.3. VBO

VBO is simply set to the VE Block Offset field in the NLRI part of ADV.

3.3.2.4. DOM

This variable, indicating the VPLS domain to which ADV belongs, is derived by applying BGP policy to the Route Target extended communities in ADV. The details of how this is done are outside the scope of this document.

3.3.2.5. ACS

ACS is the status of the attachment circuits for a given site of a VPLS. ACS = 1 if all attachment circuits for the site are down, and 0 otherwise.

ACS is set to the value of the 'D' bit in ADV that belongs to MH NLRI. If ADV belongs to base VPLS instance with non-zero label block values, no change must be made to ACS.

3.3.2.6. PREF

PREF is derived from the Local Preference (LP) attribute in ADV as well as the VPLS Preference field (VP) in the L2-info extended community. If the Local Preference attribute is missing, LP is set to 0; if the L2-info community is missing, VP is set to 0. The following table shows how PREF is computed from LP and VP.

VP Value	LP Value	PREF Value	Comment
0	0	0	malformed advertisement, unless ACS=1
0	1 to $(2^{16}-1)$	LP	backwards compatibility
0	2^{16} to $(2^{32}-1)$	$(2^{16}-1)$	backwards compatibility
>0	LP same as VP	VP	Implementation supports VP
>0	LP != VP	0	malformed advertisement

Table 1

3.3.2.7. PE-ID

If ADV contains a Route Origin (RO) community (see Section 4.1) with type 0x01, then PE-ID is set to the Global Administrator sub-field of the RO. Otherwise, if ADV has an ORIGINATOR_ID attribute, then PE-ID is set to the ORIGINATOR_ID. Otherwise, PE-ID is set to the BGP Identifier.

3.3.3. Election Procedures

The election procedures described in this section apply equally to BGP VPLS and LDP VPLS. A distinction MUST NOT be made on whether the NLRI is a multi-homing NLRI or not. Subset of these procedures documented in standard BGP best path selection deals with general IP Prefix BGP route selection processing as defined in [RFC4271]. A separate part of the algorithm defined under VPLS DF election is specific to designated forwarded election procedures performed on VPLS advertisements. A concept of bucketization is introduced to define route selection rules for VPLS advertisements. Note that this is a conceptual description of the process; an implementation MAY choose to realize this differently as long as the semantics are

preserved.

3.3.3.1. Bucketization for standard BGP path selection

An advertisement

ADV -> <RD, SITE-ID, VBO, ACS, PREF, PE-ID>

is put into the bucket for <RD, SITE-ID, VBO>. In other words, the information in BGP path selection consists of <RD, SITE-ID, VBO> and only advertisements with exact same <RD, SITE-ID, VBO> are candidates for BGP path selection procedure as defined in [RFC4271].

3.3.3.2. Bucketization for VPLS DF Election

An advertisement

ADV -> <RD, SITE-ID, VBO, DOM, ACS, PREF, PE-ID>

is discarded if DOM is not of interest to the VPLS PE. Otherwise, ADV is put into the bucket for <DOM, SITE-ID>. In other words, all advertisements for a particular VPLS domain that have the same SITE-ID are candidates for VPLS DF election.

3.3.3.3. Tie-breaking Rules

This section describes the tie-breaking rules for VPLS DF election. Tie-breaking rules for VPLS DF election are applied to candidate advertisements by all VPLS PEs and the actions taken by VPLS PEs based on the VPLS DF election result are described in Section 3.4.

Given two advertisements ADV1 and ADV2 from a given bucket, first compute the variables needed for DF election:

ADV1 -> <RD1, SITE-ID1, VBO1, DOM1, ACS1, PREF1, PE-ID1>
ADV2 -> <RD2, SITE-ID2, VBO2, DOM2, ACS2, PREF2, PE-ID2>

Note that SITE-ID1 = SITE-ID2 and DOM1 = DOM2, since ADV1 and ADV2 came from the same bucket. Then the following tie-breaking rules MUST be applied in the given order.

1. if (ACS1 != 1) AND (ACS2 == 1) ADV1 wins; stop
if (ACS1 == 1) AND (ACS2 != 1) ADV2 wins; stop
else continue
2. if (PREF1 > PREF2) ADV1 wins; stop;
else if (PREF1 < PREF2) ADV2 wins; stop;
else continue

3. if (PE-ID1 < PE-ID2) ADV1 wins; stop;
else if (PE-ID1 > PE-ID2) ADV2 wins; stop;
else ADV1 and ADV2 are from the same VPLS PE

If there is no winner and ADV1 and ADV2 are from the same PE, a VPLS PE MUST retain both ADV1 and ADV2.

3.4. DF Election on PEs

DF election algorithm MUST be run by all multi-homed VPLS PEs. In addition, all other PEs SHOULD also run the DF election algorithm. As a result of the DF election, multi-homed PEs that lose the DF election for a SITE-ID MUST put the ACs associated with the SITE-ID in non-forwarding state.

DF election result on the egress PEs can be used in traffic forwarding decision. Figure 2 shows two customer sites, CE1 and CE4, connected to PE1 with CE1 multi-homed to PE1 and PE2. If PE1 is the designated forwarder for CE1, based on the DF election result, PE3 can chose to not send unknown unicast and multicast traffic to PE2 as PE2 is not the designated forwarder for any customer site and it has no other single homed sites connected to it.

4. Multi-AS VPLS

This section describes multi-homing in an inter-AS context.

4.1. Route Origin Extended Community

Due to lack of information about the PEs that originate the VPLS NLRI in inter-AS operations, Route Origin Extended Community [RFC4360] is used to carry the source PE's IP address.

To use Route Origin Extended Community for carrying the originator VPLS PE's loopback address, the type field of the community MUST be set to 0x01 and the Global Administrator sub-field MUST be set to the PE's loopback IP address.

4.2. VPLS Preference

When multiple PEs are assigned the same site ID for multi-homing, it is often desired to be able to control the selection of a particular PE as the designated forwarder. Section 3.5 in [RFC4761] describes the use of BGP Local Preference in path selection to choose a particular NLRI, where Local Preference indicates the degree of preference for a particular VE. The use of Local Preference is inadequate when VPLS PEs are spread across multiple ASes as Local Preference is not carried across AS boundary. A new field, VPLS preference (VP), is introduced in this document that can be used to accomplish this. VPLS preference indicates a degree of preference for a particular customer site. VPLS preference is not mandatory for intra-AS operation; the algorithm explained in Section 3.3 will work with or without the presence of VPLS preference.

Section 3.2.4 in [RFC4761] describes the Layer2 Info Extended Community that carries control information about the pseudowires. The last two octets that were reserved now carries VPLS preference as shown in Figure 4.

Extended community type (2 octets)
Encaps Type (1 octet)
Control Flags (1 octet)
Layer-2 MTU (2 octet)
VPLS Preference (2 octets)

Figure 4: Layer2 Info Extended Community

A VPLS preference is a 2-octets unsigned integer. A value of zero indicates absence of a VP and is not a valid preference value. This interpretation is required for backwards compatibility. Implementations using Layer2 Info Extended Community as described in (Section 3.2.4) [RFC4761] MUST set the last two octets as zero since it was a reserved field.

For backwards compatibility, if VPLS preference is used, then BGP Local Preference MUST be set to the value of VPLS preference. Note that a Local Preference value of zero for a MH-ID is not valid unless 'D' bit in the control flags is set (see [I-D.kothari-l2vpn-auto-site-id]). In addition, Local Preference value greater than or equal to 2^{16} for VPLS advertisements is not valid.

4.3. Use of BGP-MH attributes in Inter-AS Methods

Section 3.4 in [RFC4761] and section 4 in [RFC6074] describe three methods (a, b and c) to connect sites in a VPLS to PEs that are across multiple AS. Since VPLS advertisements in method (a) do not cross AS boundaries, multi-homing operations for method (a) remain exactly the same as they are within an AS. However, for method (b) and (c), VPLS advertisements do cross AS boundary. This section describes the VPLS operations for method (b) and method (c). Consider Figure 5 for inter-AS VPLS with multi-homed customer sites.

4.3.1. Inter-AS Method (b): EBGp Redistribution of VPLS Information between ASBRs

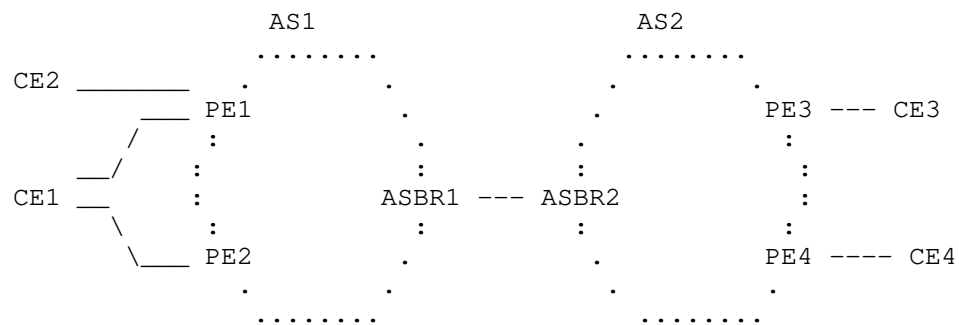


Figure 5: Inter-AS VPLS

A customer has four sites, CE1, CE2, CE3 and CE4. CE1 is multi-homed to PE1 and PE2 in AS1. CE2 is single-homed to PE1. CE3 and CE4 are also single homed to PE3 and PE4 respectively in AS2. Assume that in addition to the base LDP/BGP VPLS addressing (VSI-IDs/VE-IDs), MH ID 1 is assigned for CE1. After running DF election algorithm, all four VPLS PEs must elect the same designated forwarder for CE1 site. Since BGP Local Preference is not carried across AS boundary, VPLS preference as described in Section 4.2 MUST be used for carrying site preference in inter-AS VPLS operations.

For Inter-AS method (b) ASBR1 will send a VPLS NLRI received from PE1 to ASBR2 with itself as the BGP nexthop. ASBR2 will send the received NLRI from ASBR1 to PE3 and PE4 with itself as the BGP nexthop. Since VPLS PEs use BGP Local Preference in DF election, for backwards compatibility, ASBR2 MUST set the Local Preference value in the VPLS advertisements it sends to PE3 and PE4 to the VPLS preference value contained in the VPLS advertisement it receives from ASBR1. ASBR1 MUST do the same for the NLRIs it sends to PE1 and PE2. If ASBR1 receives a VPLS advertisement without a valid VPLS preference from a PE within its AS, then ASBR1 MUST set the VPLS preference in the advertisements to the Local Preference value before sending it to ASBR2. Similarly, ASBR2 must do the same for advertisements without VPLS Preference it receives from PEs within its AS. Thus, in method (b), ASBRs MUST update the VPLS and Local Preference based on the advertisements they receive either from an ASBR or a PE within their AS.

In Figure 5, PE1 will send the VPLS advertisements with Route Origin Extended Community containing its loopback address. PE2 will do the same. Even though PE3 receives the VPLS advertisements for VE-ID 1 and 2 from the same BGP nexthop, ASBR2, the source PE address contained in the Route Origin Extended Community is different for the CE1 and CE2 advertisements, and thus, PE3 creates two PWs, one for CE1 (for VE-ID 1) and another one for CE2 (for VE-ID 2).

4.3.2. Inter-AS Method (c): Multi-Hop EBGp Redistribution of VPLS Information between ASes

In this method, there is a multi-hop E-BGP peering between the PEs or Route Reflectors in AS1 and the PEs or Route Reflectors in AS2. There is no VPLS state in either control or data plane on the ASBRs. The multi-homing operations on the PEs in this method are exactly the same as they are in intra-AS scenario. However, since Local Preference is not carried across AS boundary, the translation of LP to VP and vice versa MUST be done by RR, if RR is used to reflect VPLS advertisements to other ASes. This is exactly the same as what

a ASBR does in case of method (b). A RR must set the VP to the LP value in an advertisement before sending it to other ASes and must set the LP to the VP value in an advertisement that it receives from other ASes before sending to the PEs within the AS.

5. MAC Flush Operations

In a service provider VPLS network, customer MAC learning is confined to PE devices and any intermediate nodes, such as a Route Reflector, do not have any state for MAC addresses.

Topology changes either in the service provider's network or in customer's network can result in the movement of MAC addresses from one PE device to another. Such events can result into traffic being dropped due to stale state of MAC addresses on the PE devices. Age out timers that clear the stale state will resume the traffic forwarding, but age out timers are typically in minutes, and convergence of the order of minutes can severely impact customer's service. To handle such events and expedite convergence of traffic, flushing of affected MAC addresses is highly desirable.

This section describes the scenarios where VPLS flush is desirable and the specific VPLS Flush TLVs that provide capability to flush the affected MAC addresses on the PE devices. All operations described in this section are in context of a particular VPLS domain and not across multiple VPLS domains. Mechanisms for MAC flush are described in [I-D.kothari-l2vpn-vpls-flush] for BGP based VPLS and in [RFC4762] for LDP based VPLS.

5.1. MAC List Flush

If multiple customer sites are connected to the same PE, PE1 as shown in Figure 2, and redundancy per site is desired when multi-homing procedures described in this document are in effect, then it is desirable to flush just the relevant MAC addresses from a particular site when the site connectivity is lost.

To flush particular set of MAC addresses, a PE SHOULD originate a flush message with MAC list that contains a list of MAC addresses that needs to be flushed. In Figure 2, if connectivity between CE1 and PE1 goes down and if PE1 was the designated forwarder for CE1, PE1 MAY send a list of MAC addresses that belong to CE1 to all its BGP peers.

It is RECOMMENDED that in case of excessive link flap of customer attachment circuit in a short duration, a PE should have a means to throttle advertisements of flush messages so that excessive flooding of such advertisements do not occur.

5.2. Implicit MAC Flush

Implicit MAC Flush refers to the use of BGP MH advertisements by the PEs to flush the MAC addresses learned from the previous designated

forwarder.

In case of a failure, when connectivity to a customer site is lost, remote PEs learn that a particular site is no longer reachable. The local PE either withdraws the VPLS NLRI that it previously advertised for the site or it sends a BGP update message for the site's VPLS NLRI with the 'D' bit set. In such cases, the remote PEs can flush all the MACs that were learned from the PE which reported the failure.

However, in cases when a designated forwarder change occurs in absence of failures, such as when an attachment circuit comes up, the BGP MH advertisement from the PE reporting the change is not sufficient for MAC flush procedures. Consider the case in Figure 2 where PE1-CE1 link is non-operational and PE2 is the designated forwarder for CE1. Also assume that Local Preference of PE1 is higher than PE2. When PE1-CE1 link becomes operational, PE1 will send a BGP MH advertisement to all its peers. If PE3 elects PE1 as the new designated forwarder for CE1 and as a result flushes all the MACs learned from PE1 before PE2 elects itself as the non-designated forwarder, there is a chance that PE3 might learn MAC addresses from PE2 and as a result may black-hole traffic until those MAC addresses are deleted due to age out timers.

A designated forwarder must set the F bit and a non-designated forwarder must clear the F bit when sending BGP MH advertisements. A state transition from one to zero for the F bit can be used by a remote PE to flush all the MACs learned from the PE that is transitioning from designated forwarder to non-designated forwarder.

5.3. Minimizing the effects of fast link transitions

Certain failure scenarios may result in fast transitions of the link towards the multi-homing CE which in turn will generate fast status transitions of one or multiple multi-homed sites reflected through multiple BGP MH advertisements and LDP MAC Flush messages.

It is recommended that a timer to damp the link flaps be used for the port towards the multi-homed CE to minimize the number of MAC Flush events in the remote PEs and the occurrences of BGP state compressions for F bit transitions. A timer value more than the time it takes BGP to converge in the network is recommended.

6. Backwards Compatibility

No forwarding loops are formed when PEs or Route Reflectors that do not support procedures defined in this section co exist in the network with PEs or Route Reflectors that do support.

6.1. BGP based VPLS

As explained in this section, multi-homed PEs to the same customer site MUST assign the same MH-ID and related NLRI SHOULD contain the block offset, block size and label base as zero. Remote PEs that lack support of multi-homing operations specified in this document will fail to create any PWs for the multi-homed MH-IDs due to the label value of zero and thus, the multi-homing NLRI should have no impact on the operation of Remote PEs that lack support of multi-homing operations specified in this document.

For compatibility with PEs that use multiple VE-IDs with non-zero label block values for multi-homing operation, it is a requirement that a PE receiving such advertisements must use the labels in the NLRIs associated with lowest VE-ID for PW creation. It is possible that maintaining PW association with lowest VE-ID can result in PW flap, and thus, traffic loss. However, it is necessary to maintain the association of PW with the lowest VE-ID as it provides deterministic DF election among all the VPLS PEs.

6.2. LDP VPLS with BGP Auto-discovery

The BGP-AD NLRI has a prefix length of 12 containing only a 8 bytes RD and a 4 bytes VSI-ID. If a LDP VPLS PEs running BGP AD lacks support of multi-homing operations specified in this document, it SHOULD ignore a MH NLRI with the length field of 17. As a result it will not ask LDP to create any PWs for the multi-homed Site-ID and thus, the multi-homing NLRI should have no impact on LDP VPLS operation. MH PEs may use existing LDP MAC Flush to flush the remote LDP VPLS PEs or may use the implicit MAC Flush procedure.

7. Security Considerations

No new security issues are introduced beyond those that are described in [RFC4761] and [RFC4762].

8. IANA Considerations

At this time, this memo includes no request to IANA.

9. Acknowledgments

The authors would like to thank Yakov Rekhter, Nischal Sheth, Mitali Singh and Ian Cowburn for their insightful comments and probing questions.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

10.2. Informative References

- [I-D.kothari-l2vpn-vpls-flush]
Kothari, B. and R. Fernando, "VPLS Flush in BGP-based Virtual Private LAN Service",
draft-kothari-l2vpn-vpls-flush-00 (work in progress),
October 2008.
- [I-D.kothari-l2vpn-auto-site-id]
Kothari, B., Kompella, K., and T. IV, "Automatic Generation of Site IDs for Virtual Private LAN Service",
draft-kothari-l2vpn-auto-site-id-01 (work in progress),
October 2008.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Authors' Addresses

Bhupesh Kothari
Cohere Networks
295 Santa Ana Court
Sunnyvale, CA 94085
US

Email: bhupesh@cohere.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti.kompella@gmail.com

Wim Henderickx
Alcatel-Lucent

Email: wim.henderickx@alcatel-lucent.be

Florin Balus
Alcatel-Lucent

Email: florin.balus@alcatel-lucent.com

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748
US

Email: uttaro@att.com

Senad Palislaamovic
Alcatel-Lucent

Email: senad.palislaamovic@alcatel-lucent.com

Wen Lin
Juniper Networks

Email: wlin@juniper.net

Layer 2 Virtual Private Networks
Internet-Draft
Intended status: Informational
Expires: January 15, 2014

O. Dornon
J. Kotalwar
Alcatel-Lucent
V. Hemige

R. Qiu
J. Zhang
Juniper Networks, Inc.
July 14, 2013

PIM Snooping over VPLS
draft-ietf-l2vpn-vpls-pim-snooping-04

Abstract

This document describes the procedures and recommendations for VPLS PEs to facilitate replication of multicast traffic to only certain ports (behind which there are interested PIM routers and/or IGMP hosts) via PIM Snooping and PIM Proxy.

With PIM Snooping, PEs passively listen to certain PIM control messages to build control and forwarding states while transparently flooding those messages. With PIM Proxy, PEs do not flood PIM Join/Prune messages but only generate their own and send out of certain ports, based on the control states built from downstream Join/Prune messages. PIM Proxy is required when PIM Join suppression is enabled on the CE devices and useful to reduce PIM control traffic in a VPLS domain.

The document also describes PIM Relay, which can be viewed as light-weight proxy, where all downstream Join/Prune messages are simply forwarded out of certain ports but not flooded to avoid triggering PIM Join suppression on CE devices.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

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 15, 2014.

Copyright Notice

Copyright (c) 2013 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	5
1.1. Multicast Snooping in VPLS	5
1.2. Assumptions	6
1.3. Definitions	7
2. PIM Snooping for VPLS	7
2.1. PIM protocol background	7
2.2. General Rules for PIM Snooping in VPLS	8
2.2.1. Preserving Assert Trigger	8
2.3. Some Considerations for PIM Snooping	9
2.3.1. Scaling	9
2.3.2. IPv6	10
2.3.3. PIM-SM (*,*,RP)	10
2.4. PIM Snooping vs PIM Proxy	10
2.4.1. Differences between PIM Snooping, Relay and Proxy	10
2.4.2. PIM Control Message Latency	11
2.4.3. When to Snoop and When to Proxy	12
2.5. Discovering PIM Routers	13
2.6. PIM-SM and PIM-SSM	14
2.6.1. Building PIM-SM Snooping States	14
2.6.2. Explanation for per (S,G,N) states	17
2.6.3. Receiving (*,G) PIM-SM Join/Prune Messages	17
2.6.4. Receiving (S,G) PIM-SM Join/Prune Messages	19
2.6.5. Receiving (S,G,rpt) Join/Prune Messages	21
2.6.6. Sending Join/Prune Messages Upstream	21
2.7. Bidirectional-PIM (PIM-BIDIR)	22
2.8. Interaction with IGMP Snooping	23
2.9. PIM-DM	23
2.9.1. Building PIM-DM Snooping States	23
2.9.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine	24
2.9.3. Triggering ASSERT election in PIM-DM	24
2.10. PIM Proxy	24
2.10.1. Upstream PIM Proxy behavior	24
2.11. Directly Connected Multicast Source	25
2.12. Data Forwarding Rules	25
2.12.1. PIM-SM Data Forwarding Rules	26
2.12.2. PIM-BIDIR Data Forwarding Rules	27
2.12.3. PIM-DM Data Forwarding Rules	28
3. IANA Considerations	29
4. Security Considerations	29
5. Contributors	29
6. Acknowledgements	30
7. References	30
7.1. Normative References	30
7.2. Informative References	30
Appendix A. PIM-BIDIR Thoughts	31
Appendix B. Example Network Scenario	31

B.1. Pim Snooping Example	32
B.2. PIM Proxy Example with (S,G) / (*,G) interaction	34
Authors' Addresses	38

1. Introduction

In Virtual Private LAN Service (VPLS), the Provider Edge (PE) devices provide a logical interconnect such that Customer Edge (CE) devices belonging to a specific VPLS instance appear to be connected by a single LAN. Forwarding Information Base for a VPLS instance is populated dynamically by source MAC address learning. Once a unicast MAC address is learned and associated with a particular Attachment Circuit (AC) or PseudoWire (PW), a frame destined to that MAC address only needs to be sent on that AC or PW.

For a frame not addressed to a known unicast MAC address, flooding has to be used. This happens with the following so called BUM traffic:

- o B: The destination MAC address is a broadcast address,
- o U: The destination MAC address is unknown (has not been learned),
- o M: The destination MAC address is a multicast address.

Multicast frames are flooded because a PE cannot know where multicast members reside. VPLS solutions (i.e., [VPLS-LDP] and [VPLS-BGP]) perform replication for multicast traffic at the ingress PE devices. As stated in the VPLS Multicast Requirements draft [VPLS-MCAST-REQ], there are two issues with VPLS Multicast today:

- o A. Multicast traffic is replicated to non-member sites.
- o B. Replication of PWs on shared physical path.

Issue A can be solved by Multicast Snooping - PEs learn sites with multicast members by snooping multicast protocol control messages and forward IP multicast traffic only to member sites. This document describes the procedures to achieve that when PIM is running between the CE devices. Issue B is outside the scope of this document and discussed in [VPLS-MCAST-TREES].

While this document is in the context of VPLS, the procedures apply to regular layer-2 switches interconnected by physical connections as well. In that case, the PW related concept/procedures are not applicable and that's all.

1.1. Multicast Snooping in VPLS

IGMP Snooping procedures described in [IGMP-SNOOP] make sure that IP multicast traffic is only sent out of the following:

- o Attachment Circuits (ACs) connecting to hosts that report related group membership
- o ACs connecting to routers
- o PseudoWires (PWs) connecting to remote PEs that have the above described ACs

Notice that traffic is always sent out of ports connecting to routers, even those on which there are no snooped group memberships, because IGMP Snooping alone can not determine if there are interested receivers beyond those routers. To further restrict traffic sent to those routers, PIM Snooping can be used, and this document describes the procedures, including the rules when both IGMP and PIM are active in a VPLS instance.

Note that for both IGMP and PIM, the term Snooping is used loosely, referring to the fact that a layer-2 device peeks into layer-3 routing protocol messages to build relevant control and forwarding states. Depending on how the control messages are handled (transparently flooded, selectively forwarded, or consumed and then regenerated), the procedure/process may be called Snooping or Proxy in different contexts.

Unless explicitly noted, the procedures in this document are used for either PIM Snooping or PIM Proxy, and we will largely refer to PIM "Snooping" in this document. The PIM Proxy specific procedures are described in Section 2.6.6. Differences that need to be observed while implementing one or the other and recommendations on which method to employ in different scenarios are noted in section Section 2.4.

This document also describes PIM Relay, which can be viewed as light-weight Proxy. Unless explicitly noted, in the rest of the document Proxy implicitly includes Relay as well.

1.2. Assumptions

The document assumes that the reader has a good understanding of the PIM protocols. The text in this draft is written in the same style as the PIM RFCs to help correlate the concepts and to make it easier to follow. In order to avoid replicating the text relating to PIM protocol handling here, this draft cross references into definitions of macros and procedures from the PIM RFCs, and assumes that the user will infer such detail from those PIM RFCs. Deviations in protocol handling specific to PIM Snooping are specified in this draft.

1.3. Definitions

There are several definitions referenced in this document that are well described in the PIM RFCs [PIM-SM], PIM-BIDIR, PIM-DM]. The following definitions and abbreviations are used throughout this document:

- o A port is defined as either an attachment circuit (AC) or a Pseudo-Wire (PW).
- o When we say a PIM message is 'received' on a port, it means that a PIM Snooping PE snooped the PIM message.

Abbreviations used in the document:

- o S: IP Address of the Multicast Source.
- o G: IP Address of the Multicast Group.
- o N: Upstream Neighbor field in a Join/Prune/Graft message.
- o Rport(N): Port on which neighbor N is learnt

Other definitions are explained in the sections where they are introduced.

2. PIM Snooping for VPLS

2.1. PIM protocol background

PIM is a multicast routing protocol running between routers, which are CE devices in a VPLS. PIM shares many of the common characteristics of a routing protocol, such as discovery messages (e.g., neighbor discovery using Hello messages), topology information (e.g., multicast tree), and error detection and notification (e.g., dead timer and designated router election). PIM does not participate in exchange of unicast routing databases, but it uses the unicast routing table to provide reverse path information for building multicast trees. There are a few variants of PIM. In [PIM-DM], multicast data is pushed towards the members similar to broadcast mechanism but routers without attached receivers will prune back towards the source. Unlike PIM-DM, other PIM flavors (PIM-SM [PIM-SM], PIM-SSM [PIM-SSM], and PIM-BIDIR [PIM-BIDIR]) employs a pull methodology via explicit joins instead of push technique.

PIM routers periodically exchange Hello messages to discover and maintain stateful sessions with neighbors. After neighbors are

discovered, PIM routers can signal their intentions to join or prune specific multicast groups. This is accomplished by having downstream routers send an explicit Join/Prune message (for the sake of generalization, consider Graft messages for PIM-DM as Join messages) to the upstream routers. The Join/Prune message can be group specific (*,G) or group and source specific (S,G).

2.2. General Rules for PIM Snooping in VPLS

The following rules for the correct operation of PIM snooping MUST be followed.

- o PIM Snooping MUST NOT affect the operation of customer layer-2 protocols (e.g., BPDUs) or layer-3 protocols.
- o PIM messages and multicast data traffic forwarded by PEs MUST follow the split-horizon rule for mesh PWs.
- o PIM snooping states in a PE MUST be per VPLS instance.
- o PIM assert triggers MUST be preserved to the extent necessary to avoid sending duplicate traffic to the same PE (see Section 2.2.1).

2.2.1. Preserving Assert Trigger

In PIM-SM/DM, there are scenarios where multiple routers could be forwarding the same multicast traffic on a LAN. When this happens, using PIM Assert Election process by sending PIM Assert Messages, routers ensure that only the Assert Winner forwards traffic on the LAN. The Assert Election is a data driven event and happens only if a router sees traffic on the interface to which it should be forwarding the traffic. In the case of VPLS with snooping, two routers may forward the same flow at the same time but each copy may reach different set of PEs, and that is acceptable from the point of view of avoiding duplicate traffic. If the two copies may reach the same PE then the sending routers must be able to see each other's traffic, in order to trigger Assert Election and stop duplicate traffic.

To achieve that, PIM-SM Snooping MUST not only forward multicast traffic for an (S,G) on the ports on which they snooped Joins(S,G)/ Joins(*,G), but also towards the upstream neighbor(s)). In other words, the ports on which the upstream neighbors are learnt must be added to the outgoing port list along with the ports on which Joins are snooped.

Similarly, PIM-DM Snooping SHOULD make sure that asserts can be

triggered (Section 2.9.3).

The above logic needs to be facilitated without breaking VPLS Split Horizon Rules. i.e. traffic should not be forwarded on the port on which it was received, and traffic arriving on a PW MUST NOT be forwarded onto other PW(s).

2.3. Some Considerations for PIM Snooping

The PIM Snooping solution described here requires a PE to examine and operate on only PIM Hello and PIM Join/Prune packets. The PE does not need to examine any other PIM packets.

Most of the procedures in PIM Snooping in the handling of PIM Hellos and PIM Join/Prune packets are very similar to that of a PIM Router.

However, the PE does not need to have any routing tables like is required in PIM Multicast Routing. It knows how to forward Join/Prunes by looking at the Upstream Neighbor field in the Join/Prune packets.

The PE does not need to know about Rendezvous Points (RP) and does not have to maintain any RP Set. All that is transparent to a PIM Snooping PE.

In the following sub-sections, we list some considerations and observations for the implementation of PIM Snooping in VPLS.

2.3.1. Scaling

Snooping needs to be employed on ACs at the downstream PEs to prevent traffic from being sent out of ACs unnecessarily. Snooping techniques can also be employed on PWs at the upstream PEs to prevent traffic from being sent to PEs unnecessarily. This may work well for small to medium scale deployments. However, if there are a large number of VPLS instances with a large number of PEs per instances, then the amount of snooping required at the upstream PEs can overwhelm the upstream PEs.

There are two methods to reduce the burden on the upstream PEs. One is to use PIM Proxy as described in Section 2.6.6, to reduce the control messages forwarded by a PE. The other is not to snoop on the PWs at all, but PEs signal the snooped states to other PEs out of band via BGP, as described in [VPLS-MCAST-TREES]. In this document, it is assumed that Snooping is performed on PWs.

2.3.2. IPv6

In VPLS, PEs forward Ethernet frames received from CEs and as such are agnostic of the layer-3 protocol used by the CEs. However, as an IGMP and PIM snooping PE, the PE would have to look deeper into the IP and IGMP/PIM packets and build snooping state based on that. The PIM Protocol specifications handle both IPv4 and IPv6. The specification for PIM Snooping in this draft can be applied to both IPv4 and IPv6 payloads.

2.3.3. PIM-SM (*,*,RP)

This draft does not address (*,*,RP) states in the VPLS network. Although [PIM-SM] specifies that routers MUST support (*,*,RP) states, there are very few implementations that actually support it in actual deployments, and it is being removed from the PIM protocol in its ongoing advancement process in IETF. Given that, this draft omits the specification relating to (*,*,RP) support.

2.4. PIM Snooping vs PIM Proxy

The document has previously alluded to PIM Snooping/Relay/Proxy. Details on the PIM Proxy/Relay solution are discussed in Section 2.6.6. In this section, a brief description and comparison are given.

2.4.1. Differences between PIM Snooping, Relay and Proxy

Differences between PIM Snooping and Proxy/Relay can be summarized as the following:

PIM Snooping	PIM Relay	PIM Proxy
Join/Prune messages snooped and flooded everywhere	Join/Prune messages snooped; forwarded as is out of certain upstream ports	Join/Prune messages consumed. Regenerated ones sent out of certain upstream ports
No PIM packets generated.	No PIM packets generated	New Join/Prune messages generated
CE Join Suppression not allowed	CE Join Suppression allowed	CE Join Suppression allowed

Note that the differences apply only to PIM Join/Prune messages. PIM

Hello messages are snooped and flooded in all cases.

Other than the above differences, most of the procedures are common to PIM Snooping and PIM Proxy/Relay, unless specifically stated otherwise.

Pure PIM Snooping PEs simply snoop on PIM packets as they are being forwarded in the VPLS. As such they truly provide transparent LAN services since no customer packets are modified or consumed or new packets introduced in the VPLS. It is also simpler to implement than PIM Proxy. However for PIM Snooping to work correctly, it is a requirement that CE routers MUST disable Join suppression in the VPLS.

Given that a large number of existing CE deployments do not support disabling of Join suppression and given the operational complexity for a provider to manage disabling of Join suppression in the VPLS, it becomes a difficult solution to deploy. Another disadvantage of PIM Snooping is that it does not scale as well as PIM Proxy. If there are a large number of CEs in a VPLS, then every CE will see every other CE's Join/Prune messages.

PIM Proxy/Relay has the advantage that it does not require Join suppression to be disabled in the VPLS. Multicast as a VPLS service can be very easily provided without requiring any changes on the CE routers. PIM Proxy/Relay helps scale VPLS Multicast since Join/Prune messages are only sent to certain upstream ports instead of flooded, and in case of full Proxy (vs. Relay) the PEs intelligently generate only one Join/Prune message for a given flow.

PIM Proxy however loses the transparency argument since Join/Prunes could get modified or even consumed at a PE. Also, new packets could get introduced in the VPLS. However, this loss of transparency is limited to PIM Join/Prune packets. It is in the interest of optimizing multicast in the VPLS and helping a VPLS network scale much better. Data traffic will still be completely transparent.

2.4.2. PIM Control Message Latency

A PIM Snooping/Proxy/Relay PE snoops on PIM Hello packets while transparently flooding them in the VPLS. As such there is no latency introduced by the VPLS in the delivery of PIM Hello packets to remote CEs in the VPLS.

A PIM Snooping PE snoops on PIM Join/Prune packets while transparently flooding them in the VPLS. There is no latency introduced by the VPLS in the delivery of PIM Join/Prune packets when PIM Snooping is employed.

A PIM Proxy/Relay PE does not simply flood PIM Join/Prune packets. This can result in additional latency for a downstream CE to receive multicast traffic after it has sent a Join. When a downstream CE prunes a multicast stream, the traffic should stop flowing to the CE with no additional latency introduced by the VPLS.

Performing only proxy of Join/Prune and not Hello messages keeps the PE behavior very similar to that of a PIM router without introducing too much additional complexity. It keeps the PIM Proxy solution fairly simple. Since Join/Prunes are forwarded by a PE along the slow-path and all other PIM packet types are forwarded along the fast-path, it is very likely that packets forwarded along the fast-path will arrive "ahead" of Join/Prune packets at a CE router (note the stress on the fact that fast-path messages will never arrive after Join/Prunes). Of particular importance are Hello packets sent along the fast-path. We can construct a variety of scenarios resulting in out of order delivery of Hellos and Join/Prune messages. However, there should be no deviation from normal expected behavior observed at the CE router receiving these messages out of order.

2.4.3. When to Snoop and When to Proxy

From the above descriptions, factors that affect the choice of Snooping/Relay/Proxy include:

- o Whether CEs do Join Suppression or not
- o Whether Join/Prune latency is critical or not
- o Whether the scale of PIM protocol message/states in a VPLS requires the scaling benefit of Proxy

Of the above factors, Join Suppression is the hard one - pure Snooping can only be used when Join Suppression is disabled on all CEs. The latency associated with Relay/Proxy is implementation dependent and may not be a concern at all with a particular implementation. The scaling benefit may not be important either, in that on a real LAN with Explicit Tracking (ET) a PIM router will need to receive and process all PIM Join/Prune messages as well.

A PIM router indicates that Join Suppression is disabled if the T-bit is set in the LAN Prune Delay option of its Hello message. If all PIM routers on a LAN set the T-bit, Explicit Tracking is possible, allowing an upstream router to track all the downstream neighbors that have Join states for any (S,G) or (*,G). That has two benefits:

- o No need for PrunePending process - the upstream router may immediately stop forwarding data when it receives a Prune from the last downstream neighbor, and immediately prune to its upstream if that's for the last downstream interface.
- o For management purpose, the upstream router knows exactly which downstream routers exist for a particular Join State.

While full Proxy can be used with or without Join Suppression on CEs and does not interfere with an upstream CE's bypass of PrunePending process, it does proxy all its downstream CEs as a single one to the upstream, removing the second benefit mentioned above.

Therefore, the general rule is that if Join Suppression is enabled on CEs then Proxy or Relay MUST be used and if Suppression is known to be disabled on all CEs then either Snooping, Relay, or Proxy MAY be used while Snooping or Relay SHOULD be used.

An implementation MAY choose dynamic determination of which mode to use, through the tracking of the above mentioned T-bit in all snooped PIM Hello messages, or MAY simply require static provisioning.

2.5. Discovering PIM Routers

A PIM Snooping PE MUST snoop on PIM Hellos received on ACs and PWs. i.e. the PE transparently floods the PIM Hello while snooping on it. PIM Hellos are used by the snooping PE to discover PIM routers and their characteristics.

For each neighbor discovered by a PE, it includes an entry in the PIM Neighbor Database with the following fields:

- o Layer 2 encapsulation for the Router sending the PIM Hello.
- o IP Address and address family of the Router sending the PIM Hello.
- o Port (AC / PW) on which the PIM Hello was received.
- o Hello TLVs

The PE should be able to interpret and act on Hello TLVs currently defined in the PIM RFCs. The TLVs of particular interest in this document are:

- o Hello-Hold-Time
- o Tracking Support

- o DR Priority

Please refer to [PIM-SM] for a list of the Hello TLVs. When a PIM Hello is received, the PE MUST reset the neighbor-expiry-timer to Hello-Hold-Time. If a PE does not receive a Hello message from a router within Hello-Hold-Time, the PE MUST remove that neighbor from its PIM Neighbor Database. If a PE receives a Hello message from a router with Hello-Hold-Time value set to zero, the PE MUST remove that router from the PIM snooping state immediately.

From the PIM Neighbor Database, a PE MUST be able to use the procedures defined in [PIM-SM] to identify the PIM Designated Router in the VPLS instance. It should also be able to determine if Tracking Support is active in the VPLS instance.

2.6. PIM-SM and PIM-SSM

The key characteristic of PIM-SM and PIM-SSM is explicit join behavior. In this model, multicast traffic is only forwarded to locations that specifically request it. The root node of a tree is the Rendezvous Point (RP) in case of a shared tree (PIM-SM only) or the first hop router that is directly connected to the multicast source in the case of a shortest path tree. All the procedures described in this section apply to both PIM-SM and PIM-SSM, except for the fact that there is no (*,G) state in PIM-SSM.

2.6.1. Building PIM-SM Snooping States

PIM-SM and PIM-SSM Snooping states are built by snooping on the PIM-SM Join/Prune messages received on AC/PWs.

The downstream state machine of a PIM-SM snooping PE very closely resembles the downstream state machine of PIM-SM routers. The downstream state consists of:

Per downstream (Port, *, G):

- o DownstreamJPState: One of { "NoInfo" (NI), "Join" (J), "Prune Pending" (PP) }

Per downstream (Port, *, G, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Per downstream (Port, S, G):

- o DownstreamJPState: One of { "NoInfo" (NI), "Join" (J), "Prune Pending" (PP) }

Per downstream (Port, S, G, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Per downstream (Port, S, G, rpt):

- o DownstreamJPRptState: One of { "NoInfo" (NI), "Pruned" (P), "Prune Pending" (PP) }

Per downstream (Port, S, G, rpt, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Where S is the address of the multicast source, G is the Group address and N is the upstream neighbor field in the Join/Prune message. Notice that unlike on PIM-SM routers where PPT and ET are per (Interface, S, G), PIM Snooping PEs have to maintain PPT and ET per (Port, S, G, N). The reasons for this are explained in Section 2.6.2

Apart from the above states, we define the following state summarization macros.

UpstreamNeighbors(*,G): If there is one or more Join(*,G) received on any port with upstream neighbor N and ET(N) is active, then N is added to UpstreamNeighbors(*,G). This set is used to determine if a Join(*,G) or a Prune(*,G) with upstream neighbor N needs to be sent upstream.

UpstreamNeighbors(S,G): If there is one or more Join(S,G) received on any port with upstream neighbor N and ET(N) is active, then N is added to UpstreamNeighbors(S,G). This set is used to determine if a Join(S,G) or a Prune(S,G) with upstream neighbor N needs to be sent upstream.

UpstreamPorts(*,G): This is the set of all Rport(N) ports where N is in the set UpstreamNeighbors(*,G). Multicast Streams forwarded using a (*,G) match MUST be forwarded to these ports in addition to downstream ports. So UpstreamPorts(*,G) MUST be added to OutgoingPortList(*,G).

UpstreamPorts(S,G): This is the set of all Rport(N) ports where N is in the set UpstreamNeighbors(S,G). UpstreamPorts(S,G) MUST be added to OutgoingPortList(S,G).

InheritedUpstreamPorts(S,G): This is the union of UpstreamPorts(S,G) and UpstreamPorts(*,G).

UpstreamPorts(S,G,rpt): If PruneDesired(S,G,rpt) becomes true, then this set is set to UpstreamPorts(*,G). Otherwise, this set is empty. UpstreamPorts(*,G) (-) UpstreamPorts(S,G,rpt) MUST be added to OutgoingPortList(S,G).

UpstreamPorts(G): This set is the union of all the UpstreamPorts(S,G) and UpstreamPorts(*,G) for a given G. Proxy (S,G) Join/Prune and (*,G) Join/Prune messages MUST be sent to a subset of UpstreamPorts(G) as specified in Section 2.6.6.1.

PWPorts: This is the set of all PWs.

OutgoingPortList(*,G): This is the set of all ports to which traffic needs to be forwarded on a (*,G) match.

OutgoingPortList(S,G): This is the set of all ports to which traffic needs to be forwarded on an (S,G) match.

See Section 2.12 on Data Forwarding Rules for the specification on how OutgoingPortList is calculated.

NumETsActive(Port,*,G): Number of (Port,*,G,N) entries that have Expiry Timer running. This macro keeps track of the number of Join(*,G)s that are received on this Port with different upstream neighbors.

NumETsActive(Port,S,G): Number of (Port,S,G,N) entries that have Expiry Timer running. This macro keeps track of the number of Join(S,G)s that are received on this Port with different upstream neighbors.

RpfVectorTlvs(*,G): RPF Vectors [RPF-VECTOR] are TLVs that may be present in received Join(*,G) messages. If present, they must be copied to RpfVectorTlvs(*,G).

RpfVectorTlvs(S,G): RPF Vectors [RPF-VECTOR] are TLVs that may be present in received Join(S,G) messages. If present, they must be copied to RpfVectorTlvs(S,G).

Since there are a few differences between the downstream state machines of PIM-SM Routers and PIM-SM snooping PEs, we specify the

details of the downstream state machine of PIM-SM snooping PEs at the risk of repeating most of the text documented in [PIM-SM].

2.6.2. Explanation for per (S,G,N) states

In PIM Routing protocols, states are built per (S,G). On a router, an (S,G) has only one RPF-Neighbor. However, a PIM Snooping PE does not have the Layer 3 routing information available to the routers in order to determine the RPF-Neighbor for a multicast flow. It merely discovers it by snooping the Join/Prune message. A PE could have snooped on two or more different Join/Prune messages for the same (S,G) that could have carried different Upstream-Neighbor fields. This could happen during transient network conditions or due to dual-homed sources. A PE cannot make assumptions on which one to pick, but instead must facilitate the CE routers decide which Upstream Neighbor gets elected the RPF-Neighbor. And for this purpose, the PE will have to track downstream and upstream Join/Prune states per (S,G,N).

2.6.3. Receiving (*,G) PIM-SM Join/Prune Messages

A Join(*,G) or Prune(*,G) is considered "received" if the following conditions are met:

- o The port on which it arrived is not Rport(N) where N is the upstream-neighbor N of the Join/Prune(*,G), or,
- o if both RPort(N) and the arrival port are PWs, then there exists at least one other (*,G,Nx) or (Sx,G,Nx) state with an AC UpstreamPort.

For simplicity, the case where both RPort(N) and the arrival port are PWs is referred to as PW-only Join/Prune in this document. The PW-only Join/Prune handling is so that the RPort(N) PW can be added to the related forwarding entries' OutgoingPortList to trigger Assert, but that is only needed for those states with AC UpstreamPort. Note that in PW-only case, it is ok for the arrival port and RPort(N) to be the same. See Appendix Appendix B for examples.

When a router receives a Join(*,G) or a Prune(*,G) with upstream neighbor N, it must process the message as defined in the state machine below. Note that the macro computations of the various macros resulting from this state machine transition is exactly as specified in the PIM-SM RFC [PIM-SM].

We define the following per-port (*,G,N) macro to help with the state machine below.

Figure 1 : Downstream per-port (*,G) state machine in tabular form

Event	Previous State		
	NoInfo (NI)	Join (J)	Prune-Pend
Receive Join(*,G)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)
Receive Prune(*,G) and NumETsActive<=1	-	-> PP state Start PPT(N)	-> PP state
Receive Prune(*,G) and NumETsActive>1	-	-> J state Start PPT(N)	-
PPT(N) expires	-	-> J state Action PPTEpiry(N)	-> NI state Action PPTEpiry(N)
ET(N) expires and NumETsActive<=1	-	-> NI state Action ETExpiry(N)	-> NI state Action ETExpiry(N)
ET(N) expires and NumETsActive>1	-	-> J state Action ETExpiry(N)	-> NI state Action ETExpiry(N)

Action RxJoin(N):

If ET(N) is not already running, then start ET(N). Otherwise restart ET(N). If N is not already in UpstreamNeighbors(*,G), then add N to UpstreamNeighbors(*,G) and trigger a Join(*,G) with upstream neighbor N to be forwarded upstream. If there are RPF Vector TLVs in the received (*,G) message and if they are different from the recorded RpfVectorTlvs(*,G), then copy them into RpfVectorTlvs(*,G).

Action PPTEpiry(N):

Same as Action ETExpiry(N) below, plus Send a Prune-Echo(*,G) with upstream-neighbor N on the downstream port.

Action ETExpiry(N):

Disable timers ET(N) and PPT(N). Delete neighbor state (Port,*,G,N). If there are no other (Port,*,G) states with NumETsActive(Port,*,G) > 0, transition DownstreamJPState to NoInfo. If there are no other (Port,*,G,N) state (different ports but for the same N), remove N from UpstreamPorts(*,G) - this also serves as a trigger for US FSM (JoinDesired(*,G,N) becomes FALSE).

2.6.4. Receiving (S,G) PIM-SM Join/Prune Messages

A Join(S,G) or Prune(S,G) is considered "received" if the following conditions are met:

- o The port on which it arrived is not Rport(N) where N is the upstream-neighbor N of the Join/Prune(S,G), or,
- o if both RPort(N) and the arrival port are PWs, then there exists at least one other (*,G,Nx) or (S,G,Nx) state with an AC UpstreamPort.

For simplicity, the case where both RPort(N) and the arrival port are PWs is referred to as PW-only Join/Prune in this document. The PW-only Join/Prune handling is so that the RPort(N) PW can be added to the related forwarding entries' OutgoingPortList to trigger Assert, but that is only needed for those states with AC UpstreamPort. See Appendix Appendix B for examples.

When a router receives a Join(S,G) or a Prune(S,G) with upstream neighbor N, it must process the message as defined in the state machine below. Note that the macro computations of the various macros resulting from this state machine transition is exactly as specified in the PIM-SM RFC [PIM-SM].

Figure 2: Downstream per-port (S,G) state machine in tabular form

Event	Previous State		
	NoInfo (NI)	Join (J)	Prune-Pend
Receive Join (S,G)	-> J state Action RxJoin (N)	-> J state Action RxJoin (N)	-> J state Action RxJoin (N)
Receive Prune (S,G) and NumETsActive<=1	-	-> PP state Start PPT(N)	-> PP state
Receive Prune (S,G) and NumETsActive>1	-	-> J state Start PPT(N)	-
PPT(N) expires	-	-> J state Action PPTExpiry (N)	-> NI state Action PPTExpiry (N)
ET(N) expires and NumETsActive<=1	-	-> NI state Action ETExpiry (N)	-> NI state Action ETExpiry (N)
ET(N) expires and NumETsActive>1	-	-> J state Action ETExpiry (N)	-> NI state Action ETExpiry (N)

Action RxJoin(N):

If ET(N) is not already running, then start ET(N). Otherwise, restart ET(N).

If N is not already in UpstreamNeighbors(S,G), then add N to UpstreamNeighbors(S,G) and trigger a Join(S,G) with upstream neighbor N to be forwarded upstream. If there are RPF Vector TLVs in the received (S,G) message and if they are different from the recorded RpfVectorTlvs(S,G), then copy them into RpfVectorTlvs(S,G).

Action PPTExpiry(N):

Same as Action ETExpiry(N) below, plus Send a Prune-Echo(S,G) with upstream-neighbor N on the downstream port.

Action ETEpiry(N):

Disable timers ET(N) and PPT(N). Delete neighbor state (Port,S,G,N). If there are no other (Port,S,G) states with NumETsActive(Port,S,G) > 0, transition DownstreamJPState to NoInfo. If there are no other (Port,S,G,N) state (different ports but for the same N), remove N from UpstreamPorts(S,G) - this also serves as a trigger for US FSM (JoinDesired(S,G,N) becomes FALSE).

2.6.5. Receiving (S,G,rpt) Join/Prune Messages

A Join(S,G,rpt) or Prune(S,G,rpt) is "received" when the port on which it was received is not also the port on which the upstream-neighbor N of the Join/Prune(S,G,rpt) was learnt.

While it is important to ensure that the (S,G) and (*,G) state machines allow for handling per (S,G,N) states, it is not as important for (S,G,rpt) states. It suffices to say that the downstream (S,G,rpt) state machine is the same as what is defined in section 4.5.4 of the PIM-SM RFC [PIM-SM].

2.6.6. Sending Join/Prune Messages Upstream

This section applies only to a PIM Proxy/Relay PE and not to a PIM Snooping PE.

A full PIM Proxy (not Relay) PE MUST implement the Upstream FSM for which the procedures are similar to what is defined in section 4.5.6 of [PIM-SM]. Similar to Downstream FSM described above, the Upstream FSM is also per Upstream Neighbor.

For the purposes of the Upstream FSM, a Join or Prune message with upstream neighbor N is "seen" on a PIM Snooping PE if the port on which the message was received is also Rport(N), and the port is an AC. The AC requirement is needed because a Join received on the Rport(N) PW must not suppress this PE's Join on that PW.

A PIM Relay PE does not implement theUpstream FSM. It simply forwards received Join/Prune messages out of the same set of upstream ports as in the PIM Proxy case.

In order to correctly facilitate assert among the CE routers, such Join/Prunes need to sent not only towards the upstream neighbor, but also on certain PWs as described below.

If RpfVectorTlvs(*,G) is not empty, then it must be encoded in a Join(*,G) message sent upstream.

If RpfVectorTlvs(S,G) is not empty, then it must be encoded in a Join(S,G) message sent upstream.

2.6.6.1. Where to send Join/Prune messages

The following rules apply, to both forwarded (in case of PIM Relay), refresh and triggered (in case of PIM Proxy) (S,G)/(*,G) Join/Prune messages.

- o The upstream neighbor field in the Join/Prune to be sent is set to the N in the corresponding Upstream FSM.
- o if Rport(N) is an AC, send the message to Rport(N).
- o Additionally, if OutgoingPortList(X,G,N) contains at least one AC, then the message MUST be sent to at least all the PWs in UpstreamPorts(G) (for (*,G)) or InheritedUpstreamPorts(S,G) (for (S,G)). Alternatively, the message MAY be sent to all PWs.

Sending to a subset of PWs as described above guarantees that if traffic (of the same flow) from two upstream routers were to reach this PE, then the two routers will receive from each other, triggering assert.

Sending to all PWs guarantees that if two upstream routers both send traffic for the same flow (even if it is to different sets of downstream PEs), then they'll receive from each other, triggering assert.

2.7. Bidirectional-PIM (PIM-BIDIR)

PIM-BIDIR is a variation of PIM-SM. The main differences between PIM-SM and Bidirectional-PIM are as follows:

- o There are no source-based trees, and source-specific multicast is not supported (i.e., no (S,G) states) in PIM-BIDIR.
- o Multicast traffic can flow up the shared tree in PIM-BIDIR.
- o To avoid forwarding loops, one router on each link is elected as the Designated Forwarder (DF) for each RP in PIM-BIDIR.

The main advantage of PIM-BIDIR is that it scales well for many-to-many applications. However, the lack of source-based trees means that multicast traffic is forced to remain on the shared tree.

As described in [PIM-BIDIR], parts of a PIM-BIDIR enabled network may forward traffic without exchanging Join/Prune messages, for instance

between DF's and the RPL.

As the described procedures for Pim snooping rely on the presence of Join/Prune messages, enabling Pim snooping on PIM-BIDIR networks could break the PIM-BIDIR functionality. Deploying Pim snooping on PIM-BIDIR enabled networks will require some further study. Some thoughts are gathered in Appendix A.

2.8. Interaction with IGMP Snooping

Whenever IGMP Snooping is enabled in conjunction with PIM Snooping in the same VPLS instance the PE SHOULD follow these rules:

- o To maintain the list of multicast routers and ports on which they are attached, the PE SHOULD NOT use the rules as described in RFC4541 [IGMP-SNOOP] but SHOULD rely on the neighbors discovered by PIM Snooping . This list SHOULD then be used to apply the forwarding rule as described in 2.1.1.(1) of RFC4541 [IGMP-SNOOP].
- o If the PE supports proxy-reporting, an IGMP membership learned only on a port to which a PIM neighbor is attached but not elsewhere SHOULD NOT be included in the summarized upstream report sent to that port.

2.9. PIM-DM

The characteristics of PIM-DM is flood and prune behavior. Shortest path trees are built as a multicast source starts transmitting.

2.9.1. Building PIM-DM Snooping States

PIM-DM Snooping states are built by snooping on the PIM-DM Join, Prune, Graft and State Refresh messages received on AC/PWs and State-Refresh Messages sent on AC/PWs. By snooping on these PIM-DM messages, a PE builds the following states per (S,G,N) where S is the address of the multicast source, G is the Group address and N is the upstream neighbor to which Prunes/Grafts are sent by downstream CEs:

Per PIM (S,G,N):

Port PIM (S,G,N) Prune State:

- * DownstreamPState(S,G,N,Port): One of {"NoInfo" (NI), "Pruned" (P), "PrunePending" (PP)}

- * Prune Pending Timer (PPT)
- * Prune Timer (PT)
- * Upstream Port (valid if the PIM(S,G,N) Prune State is "Pruned").

2.9.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine

The downstream per-port PIM(S,G,N) state machine is as defined in section 4.4.2 of [PIM-DM] with a few changes relevant to PIM Snooping. When reading section 4.4.2 of [PIM-DM] for the purposes of PIM-Snooping please be aware that the downstream states are built per (S, G, N, Downstream-Port) in PIM-Snooping and not per {Downstream-Interface, S, G} as in a PIM-DM router. As noted in the previous Section 2.9.1, the states (DownstreamPState) and timers (PPT and PT) are per (S,G,N,P).

2.9.3. Triggering ASSERT election in PIM-DM

Since PIM-DM is a flood-and-prune protocol, traffic is flooded to all routers unless explicitly pruned. Since PIM-DM routers do not prune on non-RPF interfaces, PEs should typically not receive Prunes on Rport(RPF-neighbor). So the asserting routers should typically be in pim_oiflist(S,G). In most cases, assert election should occur naturally without any special handling since data traffic will be forwarded to the asserting routers.

However, there are some scenarios where a prune might be received on a port which is also an upstream port (UP). If we prune the port from pim_oiflist(S,G), then it would not be possible for the asserting routers to determine if traffic arrived on their downstream port. This can be fixed by adding pim_iifs(S,G) to pim_oiflist(S,G) so that data traffic flows to the UP ports.

2.10. PIM Proxy

As noted earlier, PIM Snooping will work correctly only if Join Suppression is disabled in the VPLS. If Join Suppression is enabled in the VPLS, then PEs MUST do PIM Proxy/Relay for VPLS Multicast to work correctly. This section applies specifically to the full Proxy case and not Relay.

2.10.1. Upstream PIM Proxy behavior

A PIM Proxy PE consumes Join/Prune messages and regenerates PIM Join/Prune messages to be sent upstream by implementing Upstream FSM as specified in the PIM RFC, except that it is also per Upstream

Neighbor like in the Downstream FSM case. This is the only difference from PIM Relay.

The source IP address in PIM packets sent upstream SHOULD be the address of a PIM downstream neighbor in the corresponding join/prune state. The address picked MUST NOT be the upstream neighbor field to be encoded in the packet. The layer 2 encapsulation for the selected source IP address MUST be the encapsulation recorded in the PIM Neighbor database for that IP address.

2.11. Directly Connected Multicast Source

If there is a source in the CE network that connects directly into the VPLS instance, then multicast traffic from that source MUST be sent to all PIM routers on the VPLS instance apart from the igmp receivers in the VPLS. If there is already (S,G) or (*,G) snooping state that is formed on any PE, this will not happen per the current forwarding rules and guidelines. So, in order to determine if traffic needs to be flooded to all routers, a PE must be able to determine if the traffic came from a host on that LAN. There are three ways to address this problem:

- o The PE would have to do ARP snooping to determine if a source is directly connected.
- o Another option is to have configuration on all PEs to say there are CE sources that are directly connected to the VPLS instance and disallow snooping for the groups for which the source is going to send traffic. This way traffic from that source to those groups will always be flooded within the provider network.
- o A third option is to require that sources of CE multicast traffic must be behind a router.

2.12. Data Forwarding Rules

First we define the rules that are common to PIM-SM, PIM-BIDIR and PIM-DM PEs. Forwarding rules for each protocol type is specified in the sub-sections.

If there is no matching forwarding state, then the PE MAY either discard the packet or send it towards all the snooped PIM CE routers or to a configured set of ports. How this is determined is outside the scope of this document.

The following general rules MUST be followed when forwarding multicast traffic in a VPLS:

- o Traffic arriving on a port MUST NOT be forwarded back onto the same port.
- o Due to VPLS Split-Horizon rules, traffic ingressing on a PW MUST NOT be forwarded to any other PW.

2.12.1. PIM-SM Data Forwarding Rules

Per the rules in [PIM-SM] and per the additional rules specified in this document,

```
OutgoingPortList(*,G) = immediate_olist(*,G) (+)
                        UpstreamPorts(*,G) (+)
                        Rport(PimDR)
```

```
OutgoingPortList(S,G) = inherited_olist(S,G) (+)
                        UpstreamPorts(S,G) (+)
                        (UpstreamPorts(*,G) (-)
                        UpstreamPorts(S,G,rpt)) (+)
                        Rport(PimDR)
```

[PIM-SM] specifies how `immediate_olist(*,G)` and `inherited_olist(S,G)` are built. `PimDR` is the IP address of the PIM DR in the VPLS.

The PIM-SM Snooping forwarding rules are defined below in pseudocode:

```
BEGIN
  iif is the incoming port of the multicast packet.
  S is the Source IP Address of the multicast packet.
  G is the Destination IP Address of the multicast packet.

  If there is (S,G) state on the PE
  Then
    OutgoingPortList = OutgoingPortList(S,G)
  Else if there is (*,G) state on the PE
  Then
    OutgoingPortList = OutgoingPortList(*,G)
  Else
    OutgoingPortList = UserDefinedPortList
  Endif

  If iif is an AC
  Then
    OutgoingPortList = OutgoingPortList (-) iif
  Else
    ## iif is a PW
    OutgoingPortList = OutgoingPortList (-) PWPorts
  Endif

  Forward the packet to OutgoingPortList.
END
```

First if there is (S,G) state on the PE, then the set of outgoing ports is OutgoingPortList(S,G).

Otherwise if there is (*,G) state on the PE, the set of outgoing ports is OutgoingPortList(*,G).

The packet is forwarded to the selected set of outgoing ports while observing the general rules above in Section 2.12

2.12.2. PIM-BIDIR Data Forwarding Rules

The PIM-BIDIR Snooping forwarding rules are defined below in pseudocode:

```
BEGIN
    iif is the incoming port of the multicast packet.
    G is the Destination IP Address of the multicast packet.

    If there is forwarding state for G
    Then
        OutgoingPortList = olist(G)
    Else
        OutgoingPortList = UserDefinedPortList
    Endif

    If iif is an AC
    Then
        OutgoingPortList = OutgoingPortList (-) iif
    Else
        ## iif is a PW
        OutgoingPortList = OutgoingPortList (-) PWPorts
    Endif

    Forward the packet to OutgoingPortList.
END

If there is forwarding state for G, then forward the packet to
olist(G) while observing the general rules above in Section 2.12
```

[PIM-BIDIR] specifies how olist(G) is constructed.

2.12.3. PIM-DM Data Forwarding Rules

The PIM-DM Snooping data forwarding rules are defined below in pseudocode:


```
BEGIN
    iif is the incoming port of the multicast packet.
    S is the Source IP Address of the multicast packet.
    G is the Destination IP Address of the multicast packet.

    If there is (S,G) state on the PE
    Then
        OutgoingPortList = olist(S,G)
    Else
        OutgoingPortList = UserDefinedPortList
    Endif

    If iif is an AC
    Then
        OutgoingPortList = OutgoingPortList (-) iif
    Else
        ## iif is a PW
        OutgoingPortList = OutgoingPortList (-) PWPorts
    Endif

    Forward the packet to OutgoingPortList.
END
```

If there is forwarding state for (S,G), then forward the packet to olist(S,G) while observing the general rules above in section Section 2.12

[PIM-DM] specifies how olist(S,G) is constructed.

3. IANA Considerations

This document makes no request of IANA.

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

4. Security Considerations

Security considerations provided in VPLS solution documents (i.e., [VPLS-LDP] and [VPLS-BGP]) apply to this document as well.

5. Contributors

Yetik Serbest, Suresh Boddapati co-authored earlier versions.

Karl (Xiangrong) Cai and Princy Elizabeth made significant contributions to bring the specification to its current state, especially in the area of Join forwarding rules.

6. Acknowledgements

Many members of the L2VPN and PIM working groups have contributed to and provided valuable comments and feedback to this draft, including Vach Kompella, Shane Amante, Sunil Khandekar, Rob Nath, Marc Lassere, Yuji Kamite, Yiqun Cai, Ali Sajassi, Jozef Raets, Himanshu Shah (Ciena), Himanshu Shah (Alcatel-Lucent).

7. References

7.1. Normative References

- [PIM-BIDIR] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, 2007.
- [PIM-DM] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast Version 2 - Dense Mode Specification", RFC 3973, 2005.
- [PIM-SM] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast- Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, 2006.
- [PIM-SSM] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, 2006.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, 1997.
- [RPF-VECTOR] Wijnands, I., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, 2009.

7.2. Informative References

- [IGMP-SNOOP] Christensen, M., Kimball, K., and F. Solensky, "Considerations for IGMP and MLD Snooping PEs", RFC 4541, 2006.

[VPLS-BGP]

Kompella, K. and Y. Rekhter, "Virtual Private LAN Service using BGP for Auto-Discovery and Signaling", RFC 4761, 2007.

[VPLS-LDP]

Lasserre, M. and V. Kompella, "Virtual Private LAN Services using LDP Signaling", RFC 4762, 2007.

[VPLS-MCAST-REQ]

Kamite, Y., Wada, Y., Serbest, Y., Morin, T., and L. Fang, "Requirements for Multicast Support in Virtual Private LAN Services", RFC 5501, 2009.

[VPLS-MCAST-TREES]

Aggarwal, R., Kamite, Y., Fang, L., and Y. Rekhter, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-11, Work in Progress.

Appendix A. PIM-BIDIR Thoughts

This section describes some guidelines that may be used to preserve PIM-BIDIR functionality in combination with Pim Snooping.

In order to preserve PIM-BIDIR Pim snooping routers need to set up forwarding states so that :

- o on the RPL all traffic is forwarded to all Rport(N)
- o on any other interface traffic is always forwarded to the DF

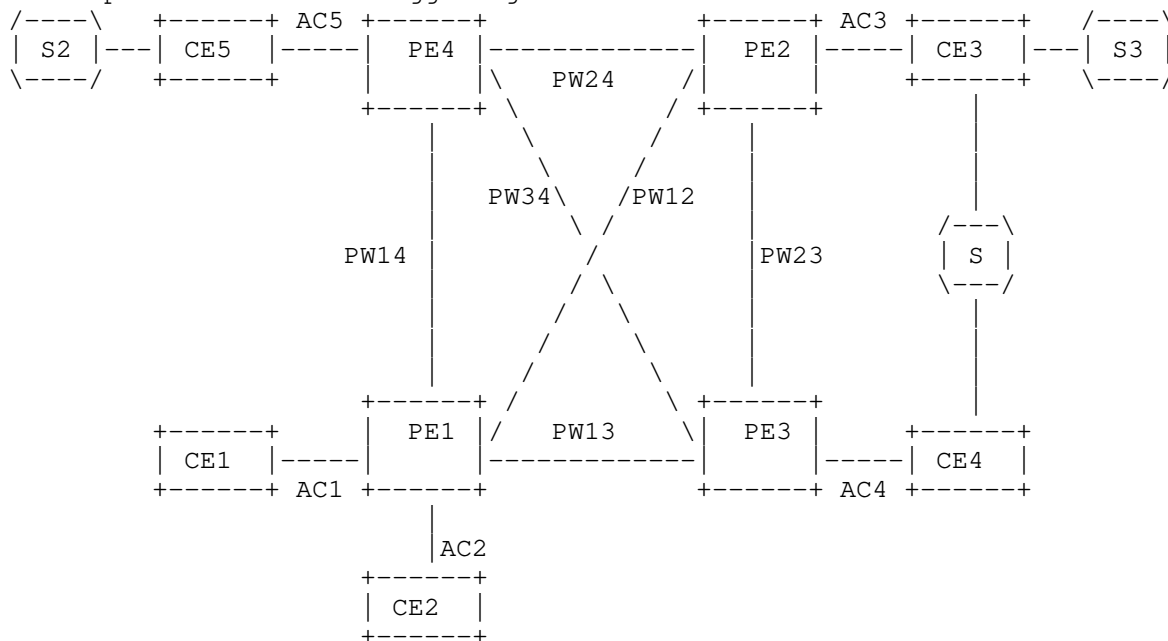
The information needed to setup these states may be obtained by :

- o determining the mapping between group(range) and RP
- o snooping and storing DF election information
- o determining where the RPL is, this could be achieved by static configuration, or by combining the information mentioned in previous bullets.

Appendix B. Example Network Scenario

Let us consider the scenario in Figure 3.

An Example Network for Triggering Assert



In the examples below, $JT(\text{Port}, S, G, N)$ is the downstream Join Expiry Timer on the specified Port for the (S, G) with upstream neighbor N .

B.1. Pim Snooping Example

In the network depicted in Figure 3, S is the source of a multicast stream (S, G) . $CE1$ and $CE2$ both have two ECMP routes to reach the source.

1. $CE1$ Sends a $Join(S, G)$ with Upstream Neighbor $(S, G) = CE3$.
2. $PE1$ snoops on the $Join(S, G)$ and builds forwarding states, since it is received on an AC and is targeting a neighbor residing across a PW it sends the join to all PW's while flooding it in the VPLS. $PE2$ and $PE3$ also snoop on the $Join(S, G)$ while flooding it in the VPLS.

The resulting states at the PEs is as follows:

At $PE1$:

```

JT(AC1, S, G, CE3)      = JP_HoldTime
UpstreamNeighbors(S, G) = { CE3 }
UpstreamPorts(S, G)     = { PW12 }
OutgoingPortList(S, G)  = { AC1, PW12 }

```

At $PE2$:

```

JT(PW12,S,G,CE3)          = JP_HoldTime
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { AC3 }
OutgoingPortList(S,G)     = { PW12, AC3 }

```

At PE3:

PE3 doesn't create a forwarding state for (S,G) because the Join(S,G) was received on a PW and the Upstream RPort is a PW too.

3. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1
4. Now CE2 sends a Join(S,G) with Upstream Neighbor(S,G) = CE4.
5. All PEs snoop on the Join(S,G).

The resulting states at the PEs:

At PE1:

```

JT(AC1,S,G,CE3)           = active
JT(AC2,S,G,CE4)           = JP_HoldTime.
UpstreamNeighbors(S,G)    = { CE3, CE4 }
UpstreamPorts(S,G)        = { PW12, PW13 }
OutgoingPortList(S,G)     = { AC1, PW12, AC2, PW13 }

```

At PE2: Note: Since PE2 already has (S,G) state, it does not ignore the Join(S,G) even though it received the Join(S,G) on a PW and the Upstream Rport is a PW.

```

JT(PW12,S,G,CE4)          = JP_HoldTime
JT(PW12,S,G,CE3)          = active
UpstreamNeighbors(S,G)    = { CE3, CE4 }
UpstreamPorts(S,G)        = { AC3, PW23 }
OutgoingPortList(S,G)     = { PW12, AC3, PW23 }

```

At PE3:

```

JT(PW13,S,G,CE4)          = JP_HoldTime
UpstreamNeighbors(S,G)    = { CE4 }
UpstreamPorts(S,G)        = { AC4 }
OutgoingPortList(S,G)     = { PW13, AC4 }

```

6. The multicast stream (S,G) flows into the VPLS from the two CEs CE3 and CE4. PE2 forwards the stream received from CE3 to PW23 and PE3 forwards the stream to AC4. This facilitates the CE routers to trigger assert election. Let us say CE3 becomes the assert winner.
7. CE3 sends an Assert message to the VPLS. The PEs flood the Assert message without examining it.
8. CE4 stops sending the multicast stream to the VPLS.
9. CE2 notices an RPF change due to Assert and sends a Prune(S,G) with Upstream Neighbor = CE4. CE2 also sends a Join(S,G) with

Upstream Neighbor = CE3.

10. All the PEs start a prune-pend timer on the ports on which they received the Prune(S,G). When the prune-pend timer expires, all PEs will remove the downstream (S,G,CE4) states.

Resulting states at the PEs:

At PE1:

JT(AC1,S,G,CE3)	= active
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { PW12 }
OutgoingPortList(S,G)	= { AC1, AC2, PW12 }

At PE2:

JT(PW12,S,G,CE3)	= active
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { AC3 }
OutgoingPortList(S,G)	= { PW12, AC3 }

At PE3: no (S,G) state.

Note that at the end of the assert election, there should be no duplicate traffic forwarded downstream and traffic should flow only on the desired path. Also note that there are no unnecessary (S,G) states on PE3 after the assert election.

B.2. PIM Proxy Example with (S,G) / (*,G) interaction

In the same network, let us assume CE4 is the Upstream Neighbor towards the RP for G.

JPST(S,G,N) is the JP sending timer for the (S,G) with upstream neighbor N.

1. CE1 Sends a Join(S,G) with Upstream Neighbor(S,G) = CE3.

PE1 consumes the Join(S,G). Since it is received on an AC and is targeting a neighbor that is residing across a PW it sends the join over all PWs.

PE2 consumes the Join(S,G). Since the join is received on a PW and targets an AC it only sends the join only over AC3.

PE3 & PE4 ignore the Join(S,G) because it is received over a PW and targets a PW, and no states exist for S,G.

The resulting states at the PEs is as follows:

PE1 states:

```

JT(AC1,S,G,CE3)           = JP_HoldTime
JPST(S,G,CE3)             = t_periodic
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { PW12 }
OutgoingPortList(S,G)     = { AC1, PW12 }

```

PE2 states:

```

JT(PW12,S,G,CE3)          = JP_HoldTime
JPST(S,G,CE3)             = t_periodic
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { AC3 }
OutgoingPortList(S,G)     = { PW12, AC3 }

```

2. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1.

3. Now let us say CE1 sends a Join(*,G) towards CE4.

PE1 snoops this Join(*,G). Since the join is received on an A and is targeting a neighbor residing on a PW it sends the join over all PWs.

PE2 consumes this Join(*,G) because it has a state for (S,G) with an AC in UpstreamPorts(S,G). Since the join is received in a PW and targets another PW it does not send the join anywhere, but adds UpstreamPorts(*,G) to OutgoingPortList(*,G) and not the downstream port PW12.

PE3 consumes the Join(*,G). Since the join is received on a PW and targets an AC it only sends the join only over AC4.

The resulting states at the PEs is as follows:

PE1 states:

```

JT(AC1,S,G,CE3)           = active
JPST(S,G,CE3)             = active
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { PW12, PW13 }
OutgoingPortList(S,G)     = { AC1, PW12, PW13 }

```

```

JT(AC1,*,G,CE4)           = JP_HoldTime
JPST(*,G,CE4)             = t_periodic
UpstreamNeighbors(*,G)    = { CE4 }
UpstreamPorts(*,G)        = { PW13 }
OutgoingPortList(*,G)     = { AC1, PW13 }

```

PE2 states:

```

JT(PW12,S,G,CE3)          = active
JPST(S,G,CE3)             = active
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { AC3, PW23 }

```

```
OutgoingPortList(S,G) = { PW12, AC3, PW23 }
```

```
JT(PW12,*,G,CE4)      = JP_HoldTime
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(G)       = { PW23 }
OutgoingPortList(*,G)  = { PW23 }
```

PE3 states:

```
JT(PW13,*,G,CE4)      = JP_HoldTime
JPST(*,G,CE4)          = t_periodic
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)     = { AC4 }
OutgoingPortList(*,G)  = { PW13, AC4 }
```

4. In the case that there is no traffic yet and PE1 sends a periodic Join(S,G) to PE2 and PE3 (step 2 is delayed after step 4).

PE1 & PE2, nothing changes except for a refresh of the timers

PE3 consumes the JOIN(S,G) because it has a (*,G) state with an AC in UpstreamPorts(*,G). Since the join is received in a PW and targets another PW it does not send the join anywhere.

PE3 States:

```
JT(PW13,*,G,CE4)      = active
JPST(S,G,CE4)          = active
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)     = { AC4 }
OutgoingPortList(*,G)  = { PW13, AC4 }

JT(PW13,S,G,CE3)       = JP_HoldTime
UpstreamNeighbors(*,G) = { CE3 }
UpstreamPorts(*,G)     = { PW23 }
OutgoingPortList(*,G)  = { PW13, AC4, PW23 }
```

5. The above state results in both (S,G) and (*,G) streams to be forwarded to AC1. The above state also results in the (S,G) stream to be forwarded from CE3 to CE4 resulting in an (S,G) assert election. Following the assert election, CE3 becomes the (S,G) assert winner. CE4 stops sending (S,G) stream down the RPT.
9. CE1 notices an RPF change due to assert. It sends a Prune(S,G,rpt) with Upstream Neighbor = CE4.
10. PE1 consumes the Prune(S,G,rpt) and since PruneDesired(S,G,Rpt,CE4) is TRUE, it needs to send the Prune(S,G,rpt) to CE4. This Prune(S,G,rpt) needs to be sent to both PW12 and PW13.

PE2 consumes the Prune(S,G,rpt), it should not send out any Prune(S,G,rpt) since this Prune(S,G,rpt) has double PW ports.

PE3 consumes the Prune(S,G,rpt) and since PruneDesired(S,G,rpt,CE4) is TRUE it sends the Prune(S,G,rpt) on AC4.

PE1 states:

```
JT(AC1,S,G,CE3)           = active
JPST(AC1,S,G,CE3)          = active
UpstreamNeighbors(S,G)    = { CE3 }

JT(AC1,S,G,CE4)           = JP_Holdtime with FLAG sgrpt prune
JPST(AC1,S,G,CE4)          = none, since JPST(AC1, *,G,CE4) is there
UpstreamPorts(S,G,rpt)    = { PW13 }
UpstreamNeighbors(S,G,rpt) = { CE4 }
UpstreamNeighbors(S,G)     = { CE3 }
UpstreamPorts(S,G)         = { PW12 }
OutgoingPortList(S,G)     = { AC1, PW12 }

JT(AC1,*,G,CE4)           = active
JPST(*,G,CE4)             = active
UpstreamNeighbors(*,G)     = { CE4 }
UpstreamPorts(*,G)         = { PW13 }
OutgoingPortList(*,G)     = { AC1, PW13 }
```

At PE2:

```
JT(PW12,S,G,CE3)          = active
JPST(PW12,S,G,CE3)         = active
UpstreamNeighbors(S,G)     = { CE3 }

JT(PW12,S,G,CE4)          = JP_Holdtime with FLAG sgrpt prune
JPST(PW12,S,G,CE4)         = none, no Prune(S,G,rpt) should be sent
UpstreamPorts(S,G,rpt)    = { PW23 }
UpstreamNeighbors(S,G,rpt) = { CE4 }

UpstreamNeighbors(S,G)     = { CE3 }
UpstreamPorts(S,G)         = { AC3 }
OutgoingPortList(*,G)     = { PW12, AC3 }

JT(PW12,*,G,CE4)          = active
UpstreamNeighbors(*,G)     = { CE4 }
UpstreamPorts(*,G)         = { PW23 }
OutgoingPortList(*,G)     = { PW23 }
```

At PE3:

```
JT(PW13,S,G,CE4)          = JP_Holdtime with S,G,rpt prune flag
JPST(PW13,S,G,CE4)         = none, no Prune(S,G,rpt) should be sent
```

```

UpstreamNeighbors(S,G,rpt) = { CE4 }
UpstreamPorts(S,G,rpt)    = { AC4 }
OutgoingPortList(S,G)     = { empty }

JT(PW13,*,G,CE4)          = active
JPST(S,G,CE4)             = active
UpstreamNeighbors(*,G)    = { CE4 }
UpstreamPorts(G)          = { AC4 }
OutgoingPortList(*,G)     = { PW13, AC4 }

```

11. If we're in case 4 for PE3

At PE3:

```

JT(PW13,S,G,CE3)          = active
JPST(PW13,S,G,CE4)        = none, this state is created by double join
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { PW23 }
OutgoingPortList(S,G)     = { PW23 }

JT(PW13,S,G,CE4)          = JP_Holdtime with S,G,rpt prune flag
JPST(PW13,S,G,CE4)        = none, no Prune(S,G,rpt) should be sent
UpstreamNeighbors(S,G,rpt) = { CE4 }
UpstreamPorts(S,G,rpt)    = { AC4 }

JT(PW13,*,G,CE4)          = active
JPST(S,G,CE4)             = active
UpstreamNeighbors(*,G)    = { CE4 }
UpstreamPorts(G)          = { AC4 }
OutgoingPortList(*,G)     = { PW13, AC4 }

```

Even in this example, at the end of the (S,G) / (*,G) assert election, there should be no duplicate traffic forwarded downstream and traffic should flow only to the desired CEs.

However, the reason we don't have duplicate traffic is because one of the CE stops sending traffic due to assert, not because we don't have any forwarding state in PE to do this forwarding. Moreover, when JP received order is different, the PE state could be different (like PE3 could have OutgoingPortList(S,G) be PW23 or empty). This is confusing, though from traffic forwarding POV it is still correct.

Other more complex scenarios exist. This draft should address in PIM-SM and the rules specified in this draft should ensure that assert is triggered among the CEs in all scenarios.

Authors' Addresses

Olivier Dornon
Alcatel-Lucent
50 Copernicuslaan
Antwerp, B2018

Email: olivier.dornon@alcatel-lucent.com

Jayant Kotalwar
Alcatel-Lucent
701 East Middlefield Rd.
Mountain View, CA 94043

Email: jayant.kotalwar@alcatel-lucent.com

Venu Hemige

Email: vhemige@gmail.com

Ray Qiu
Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089

Email: rqiujuniper.net

Jeffrey Zhang
Juniper Networks, Inc.
10 Technology Park Drive
Westford, MA 01886

Email: zzhang@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 14, 2014

Z. Li
J. Zhang
Huawei Technologies
July 13, 2013

Multicast State Advertisement in E-VPN
draft-li-l2vpn-evpn-mcast-state-ad-00

Abstract

The document defines a new extended community to advertise the active or inactive state for multicast along with the Inclusive Multicast Ethernet Tag route or Ethernet A-D route in E-VPN. The multicast state advertised can help optimization of multicast process in E-VPN to reduce unnecessary traffic replication for the broadcast, unknown unicast and multicast (BUM) traffic.

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].

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 14, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Terminology	3
3. Problem Statements	3
4. Protocol Extensions	4
5. Operations	4
5.1. Multicast State Advertisement per EVI	4
5.2. Multicast State Advertisement per <EVI, ESI>	5
6. Application	5
6.1. Ingress Replication	5
6.2. P2MP MPLS LSPs	5
7. IANA Considerations	6
8. Security Considerations	6
9. Normative References	6
Authors' Addresses	7

1. Introduction

E-VPN [I-D.ietf-l2vpn-evpn] introduces a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reachability information over the core MPLS/IP network.

In E-VPN when the PE receives the broadcast, unknown unicast, or multicast (BUM) traffic, it will forward the traffic to other PEs of the E-VPN through ingress replication or P2MP LSPs. The Inclusive Multicast Ethernet Tag routes distributed to discover the multicast membership of the E-VPN can be used to trigger setup of ingress replication tunnels or P2MP LSPs. In the actual network, the multicast service maybe use a great deal of bandwidth. It is important to save the possible bandwidth when deploy multicast service.

This document defines a new extended community to advertise the active or inactive state for multicast along with Inclusive Multicast Ethernet Tag routes or Ethernet Auto-Discovery routes in E-VPN. The multicast state advertised can help optimization of multicast process in E-VPN to reduce unnecessary traffic replication for the BUM traffic.

2. Terminology

CE: Customer Edge

E-VPN: Ethernet VPN

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

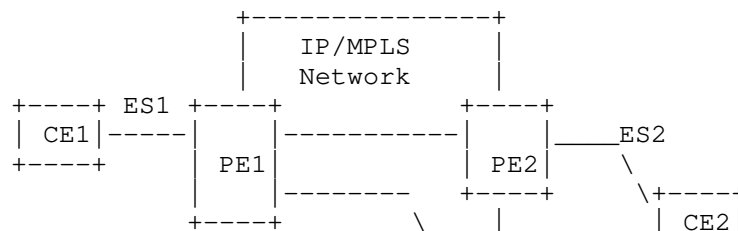
EVI: Ethernet VPN Instance

PE: Provider Edge

S-EVPN: Segment-based EVPN

3. Problem Statements

There exist multi-homing scenarios in E-VPN. As shown in the figure 1, CE2 multi-homes to two PEs (PE2 and PE3). When ingress replication is used for the BUM traffic, PE1 needs to send two copies of the same BUM traffic to both PE2 and PE3. We assume that PE2 is the Designated Forwarder (DF) for the CE2 in the E-VPN. Thus only PE2 will forward the BUM traffic to CE2 while the traffic to the PE3 will be dropped. From the example we can see that the copy of the BUM traffic sent to PE3 is not necessary in the network. If PE1 can learn that the remote PE cannot forward the BUM traffic to any CE, the bandwidth can be saved for PE1 can stop to replicate the unnecessary traffic to the remote PE. In order to achieve the object, the active or inactive state related with forwarding the BUM traffic in a EVI can be advertised by the originating PE. As to a specific EVI, the active state means that there is at least one Ethernet Segment for the EVI on the PE which needs to forward the BUM traffic to the CE and the inactive state means that none of Ethernet Segments for the EVI on the PE would forward the BUM traffic to CEs. As to a specific Ethernet Segment in a EVI, the active state means the Ethernet Segment in the EVI would forward the BUM traffic to the CE and the inactive state means that the Ethernet Segment in the EVI would not forward the BUM traffic to the CE.



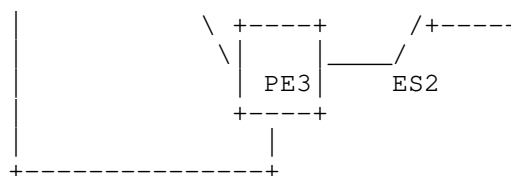


Figure 1 Multi-homing Network of E-VPN

4. Protocol Extensions

A new extended community is defined to identify the multicast state of the EVI on the leaf PE. This extended community is called as Multicast State Extended Community. It is a new transitive extended community with the Type field is 0x06, and the Sub-Type is to be defined. It may be advertised along with Inclusive Multicast Ethernet Tag routes or Ethernet Auto-Discovery routes.

Each Multicast State Extended Community is encoded as a 8-octet value as follows:

```

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(TBD) | State(One Octet) | Reserved=0 |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Reserved = 0                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The state is encoded as one octet. A value of 0 means that the multicast state is Active and a value of 1 means that the multicast state is Inactive.

5. Operations

5.1. Multicast State Advertisement per EVI

If the multicast state of a specific EVI needs to be advertised, the Multicast State Extended Community MUST be included in the Inclusive Multicast Ethernet Tag Route for the EVI. Construction of the Inclusive Multicast Ethernet Tag Route can refer to Section 12.1 in [I-D.ietf-l2vpn-evpn]. If the multicast state of the EVI is Active, the state field in the Multicast State Extended Community MUST be set to 0. If the multicast state of the EVI is Inactive, the state field in the Multicast State Extended Community MUST be set to 1. When a PE receives the Inclusive Multicast Ethernet Tag Route with the EVI State Extended Community, it can determine the multicast state of the EVI on the leaf PE originating the route is Active or Inactive.

5.2. Multicast State Advertisement per <EVI, ESI>

If the multicast state of a specific Ethernet Segment in an EVI needs to be advertised, the Multicast State Extended Community MUST be included in the Ethernet Auto-Discovery route for the Ethernet Segment in the EVI. Constructing the Ethernet A-D Route per EVI can refer to Section 9.4 of [I-D.ietf-l2vpn-evpn]. The Ethernet A-D route can be constructed for a given <ESI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> (where the Ethernet Tag ID is set to 0). If the Ethernet Segment of a specific EVI transports the BUM traffic, the state field in the Multicast State Extended Community MUST be set to 0. If the Ethernet Segment of the EVI does not transport the BUM traffic, the state field in the Multicast State Extended Community MUST be set to 1. When a PE receives Ethernet A-D routes per EVI with the EVI State Extended Community, it can determine the multicast state of the Ethernet Segment of the EVI on the leaf PE originating the route is Active or Inactive. According to the states of the Ethernet Segments of the EVI on the leaf PE, the ingress PE can determine the state of the EVI on the leaf PE. That is, if there exists one ES of the EVI which state is Active, it can determine the state of the EVI on the leaf PE is Active. If there is no ES of the EVI which state is Active, it can determine the state of the EVI is Inactive.

6. Application

6.1. Ingress Replication

When a PE determines the multicast state of the EVI on the leaf PE through the Multicast State Extended Community advertised along with the Inclusive Multicast Ethernet Tag routes or Ethernet A-D routes, it can only setup the P2P tunnels to the leaf PEs which states are Active for Ingress Replication while it will not setup the P2P tunnel to the leaf PE which EVI state is Inactive or it can stop the traffic to be replicated on the existing P2P tunnel to this leaf PE. Thus the bandwidth for the BUM traffic can be saved in the network.

6.2. P2MP MPLS LSPs

When a PE determines the multicast state of the EVI on the leaf PE through the Multicast State Extended Community advertised along with the Inclusive Multicast Ethernet Tag routes or Ethernet A-D routes, it can only setup the P2MP MPLS LSP to the leaf PEs which states are Active. Thus the bandwidth for the BUM traffic can be saved in the network.

[I-D.chen-mpls-p2mp-egress-protection] proposes a mechanism for locally protecting egress nodes of an MPLS TE P2MP LSP. In the

mechanism, the backup egress node needs to be designated for the primary egress node for a P2MP LSP. The previous hop node of the primary egress node sets up a backup Sub-LSP from itself to the backup egress node after receiving the information about the backup egress node. The multicast state advertisement proposed by the document can facilitate the provision of the local protection mechanism. The ingress PE of a specific EVI can learn the multicast state of egress PEs of the EVI through the advertised Multicast State Extended Community. Through the Ethernet A-D routes per EVI the ingress PE can also learn the information on which pair of PEs are multi-homed by one CE. Based on these information, the ingress PE can determine which egress PE can be used as the backup node to protect the primary egress node for a P2MP LSP using for the EVI. So the ingress PE can trigger to setup P2MP LSP with locally protecting egress nodes. The method saves much provision effort for this type of local protection through the auto-discovery mechanism since it need not statically designate the protection between the backup egress node and the primary egress node for a P2MP LSP.

7. IANA Considerations

This document defines a new BGP Extended Community called as Multicast State Extended Community. The sub-type value for this extended community is to be assigned by IANA.

8. Security Considerations

There are no additional security aspects beyond those of E-VPN ([I-D.ietf-l2vpn-evpn]).

9. Normative References

[I-D.chen-mpls-p2mp-egress-protection]

Chen, H., Ning, S., Liu, A., Xu, F., Toy, M., and L. Liu, "Extensions to RSVP-TE for P2MP LSP Egress Local Protection", draft-chen-mpls-p2mp-egress-protection-09 (work in progress), May 2013.

[I-D.ietf-l2vpn-evpn]

Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03 (work in progress), February 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Junlin Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jackey.zhang@huawei.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 07, 2014

Z. Li
L. Yong
J. Zhang
Huawei Technologies
July 06, 2013

Segment-Based EVPN(S-EVPN)
draft-li-l2vpn-segment-evpn-00

Abstract

This document proposes an enhanced EVPN mechanism, segment-based EVPN (S-EVPN). It satisfies the requirements of PBB-EVPN but does not require PBB implementation on PE. The solution uses a global label for each Ethernet Segment (ES) in an EVPN. It inserts the source ES label into packets at ingress PE and learns C-MAC and source ES label binding at egress PE. The solution makes the implementation easier and closer to EVPN's compared to PBB-EVPN but has the PBB-EVPN benefits.

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].

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 07, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Terminology	3
3. Challenges of PBB-EVPN Implementation	4
4. Architecture of S-EVPN	5
4.1. C-MAC Learning	6
4.2. ES Global Label Assignment	7
4.3. Ethernet A-D Route Per EVI	8
4.4. Ethernet A-D Route Per ES	9
5. Improvement on EVPN	9
5.1. Split Horizon	9
5.2. Unifying MPLS Forwarding	10
6. BGP E-VPN NLRI Extensions	10
6.1. ES Global Label Request Extended Community	11
6.2. ES Global Label Mapping Route	11
7. Operations	12
7.1. ES Global Label Request	12
7.2. ES Global Label Allocation	12
8. Solution Advantages	13
9. IANA Considerations	14
10. Security Considerations	14
11. Acknowledgements	14
12. Normative References	14
Authors' Addresses	14

1. Introduction

E-VPN [I-D.ietf-l2vpn-evpn] introduces a solution for multipoint L2VPN services. It has multi-homing capability and uses BGP for distributing customer/client MAC address reachability information over the core MPLS/IP network. PBB-EVPN [I-D.ietf-l2vpn-pbb-evpn] integrates PBB and E-VPN to achieves these:

1. reduce the number of MAC advertisement routes in BGP;
2. provide client MAC address mobility;

3. confine the scope of C-MAC learning to only active flows;
4. offer per site policies and avoid C-MAC address flushing on topology changes.

This document discusses the challenges faced by PBB-EVPN in the implementation and operation. It proposes an enhanced E-VPN mechanism, i.e. segment-based EVPN (S-EVPN), that provides the same benefits as of PBB-EVPN but does not require implementing PBB function on PE. S-EVPN mechanism allocates a global label for each Ethernet Segments in E-VPN, inserts the source ES label into the packet at ingress PE, and learns C-MAC and source ES label binding at egress PE. As a result it is not necessary to determine the source of C-MAC according to the B-MAC encapsulation which is required in PBB-EVPN. S-EVPN has simpler operation and management of EVPN and better encapsulation efficiency of packets compared to PBB-EVPN. In addition, it is easy to enhance the E-VPN to support S-EVPN and S-EVPN can unify the unicast traffic forwarding no matter C-MACs are learned by control plane or data plane.

2. Terminology

BEB: Backbone Edge Bridge

B-MAC: Backbone MAC Address

CE: Customer Edge

C-MAC: Customer/Client MAC Address

LACP: Link Aggregation Control Protocol

P2P: Point to Point

PE: Provider Edge

PBB: Provider Backbone Bridge

E-VPN: Ethernet VPN

S-EVPN: Segment-based EVPN

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

EVI: Ethernet VPN Instance

3. Challenges of PBB-EVPN Implementation

PBB-EVPN has advantages in the following aspects as [I-D.ietf-l2vpn-pbb-evpn]:

- MAC Advertisement Route Scalability
- C-MAC Mobility with MAC Sub-netting
- C-MAC Address Learning and Confinement
- Seamless Interworking with TRILL and 802.1aq Access Networks
- Per Site Policy Support
- Avoiding C-MAC Address Flushing

However, there are some challenges to implement PBB-EVPN.

1. Creation and Management B-MAC

For PBB-EVPN, the choice of B-MAC address(es) for the PE nodes must be examined carefully as it has implications on the proper operation of multi-homing. These addresses are usually locally administered by the Service Provider which involves a lot of operation and management such as design, configuration and checking. Automating B-MAC Address Assignment can be applied, but for some scenarios the method cannot work and manual provision is inevitable. A more general automated solution can be proposed to reduce manual intervention.

2. Encapsulation Efficiency of PBB-EVPN

When PBB encapsulation (shown in the figure 1) is adopted in PBB-EVPN, the B-DA, I-Tag, etc. fields in the encapsulation are useless in PBB-EVPN which reduce the effective payload.

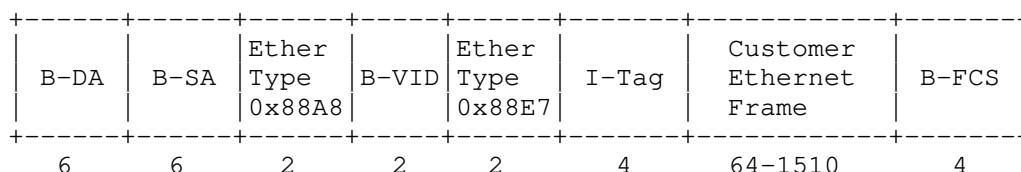


Figure 1: PBB Encapsulation

In the PBB encapsulation for PBB-EVPN, the source B-MAC is necessary since the egress PE need to learn the correspondence between C-MACs

and B-MACs. The destination B-MAC is not necessary since the destination (egress PE) is reachable through the tunnel setup in advance instead of searching routes according to the destination B-MAC.

The I-SID is also not necessary any more. PBB divides the Ethernet network into two layers: I-Component and B-Component. In the egress PE, B-Component need identify I-Component through I-SID. For PBB-VPLS, MAC learning is through the data plane which is always to use broadcast or multicast for unknown unicast traffic. In order to indentify different forwarding instance, I-SID must be adopted. For PBB-EVPN, the forwarding instance is constructed through the control plane. That is, the forwarding instance is constructed through the RT matching of EVIs and identified by the label advertised. So I-SID information in PBB encapsulation for PBB-EVPN is no use any more.

In addition B-VID in PBB encapsulation is almost never used. In a summary, in the PBB encapsulation for PBB-EVPN, only source B-MAC is indispensable. The encapsulation efficiency can be optimized.

3. Combination of PBB and E-VPN

The issues are dealt with by PBB-EVPN through the combination of two distinct technologies: PBB (layer 2 technology) and MPLS technology. In order to reduce the number of BGP MAC advertisement routes in E-VPN, PBB-EVPN can aggregate Customer/Client MAC (C- MAC) addresses via Provider Backbone MAC address (B-MAC). In fact, C-MAC addresses can be aggregated via MPLS label. Thus the issue solved by PBB-VPN can be solved in the method that is based on only MPLS technology. That is, the method is similar as E-VPN which is only based on MPLS technology. In other word, we can enhance E-VPN according to the similar way to gain PBB-EVPN benefits but not implement PBB on PE, which is a clean and simpler solution than PBB-EVPN.

4. Architecture of S-EVPN

To implement C-MAC summarization scheme, Segment-based EVPN (S-EVPN) introduces a global label for each Ethernet Segment in an EVPN regardless single homed or multi-homed CE. BGP needs to advertise the global label and Ethernet Segment binding to all PEs. In data plane, ingress PE inserts the source ES label into packets; egress PE learns the C-MAC and source ES label binding upon receiving packets. S-EVPN purely relies on BGP IP/MPLS technology.

The encapsulation of S-EVPN is shown in figure 2. The outmost label is the label for MPLS tunnel. The second label is the label which is allocated for Ethernet A-D route per EVI as E-VPN [I-D.ietf-l2vpn-evpn] and can identify a given <ESI, Ethernet Tag ID>

tuple per EVI or per <ESI, EVI> (where the Ethernet Tag ID is set to 0). The third label is a global label which identify an Ethernet Segment uniquely. The global label allocated for a specific Ethernet Segment will be described in section 4.2.

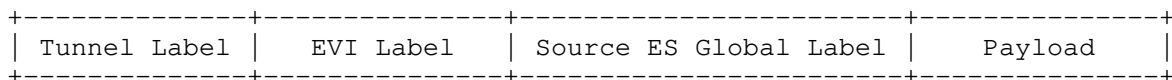


Figure 2: S-EVPN Encapsulation

4.1. C-MAC Learning

In S-EVPN, C-MACs can be learned in the data plane to determine which source Ethernet Segment they are from and which EVI they belongs to. The forwarding entry to these learned C-MACs can be installed according to the source ES and EVI information.

In S-EVPN, Ingress PE needs to send unknown traffic with source C-MACs to all remote PEs according to the encapsulation as shown in figure 2. When a specific egress PE receives the packet:

1. it can learn the C-MAC and possible VLAN Tag in the payload;
2. it can learns the EVI the C-MAC belongs to according to the EVI label which is allocated by the egress PE;
3. it can learns the Source Ethernet Segment the CMAC belongs to according to the advertised the global label and Ethernet Segment binding in BGP.

Then the egress PE needs to install the forwarding entry to the learned C-MAC. The forwarding entry to the C-MAC need two types of information: the reachability information to the ingress PE which the C-MAC belongs to; the identification for the Ethernet Segment of the EVI on the ingress PE through which the packet can send to the C-MAC.

1. Tunnel to the ingress PE: Egress PE determines PE which the Source Ethernet Segment belongs according to the advertised the global label and Ethernet Segment binding in BGP. Then egress PE can determine the tunnel to the ingress PE.
2. Label for the Ethernet Segment of the EVI on the ingress PE: The ingress PE needs to allocate label for the <ESI, EVI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> and advertise the corresponding Ethernet A-D Route per EVI to remote PEs. The egress PE can determines the Source Ethernet Segment, the EVI and the possible VLAN

which the learned the C-MAC belongs to. Then it can determine the label binded to the <ESI, EVI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> which is advertised though the Ethernet A-D Route per EVI by the ingress PE.

Besides the two types of forwarding information, when the egress PE sends a specific packet to the learned C-MAC, it needs to determine the Ethernet Segment from which the packet come and encapsulate the global label for the Ethernet Segment firstly in the packet.

According to above procedures in S-EVPN, the egress PE can learn C-MACs and install forwarding entries to these C-MACs.

4.2. ES Global Label Assignment

In S-EVPN, C-MAC summarization is done per an Ethernet Segment. The global ES label is introduced to identify the Ethernet Segment. The advantages of using global label are:

1. identify the ES globally;
2. leverage existing MPLS label stack implementation;
3. the label can be allocated dynamically to automate provision.

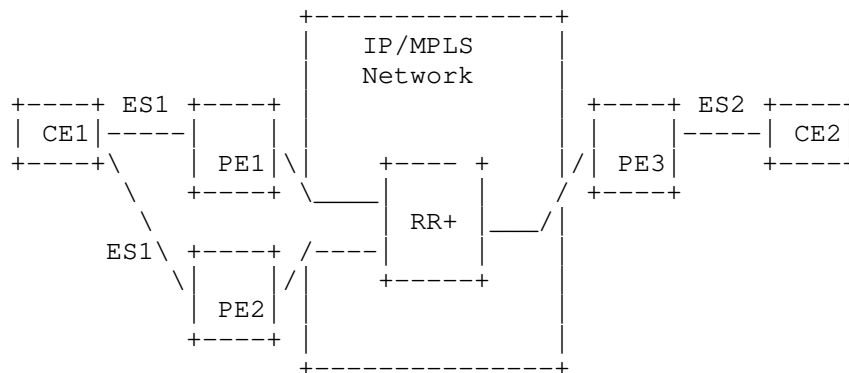


Figure 3: S-EVPN Network

In order to allocate a global label for an Ethernet Segment, there should be a centralized control point. Route Reflector (RR) of BGP may serve this role and we call this type of RR as RR+. The S-EVPN network is shown in the figure 3. All PEs of S-EVPN connects with RR+. The procedure is as follows:

1. Auto-Discovery of Ethernet Segment

RR+ can learn Ethernet Segment through the Ethernet A-D route per Ethernet Segment defined by [I-D.ietf-l2vpn-evpn]. Note that, in S-EVPN, every ES must have a unique identifier including the single-homed CEs. That is, ESI 0 cannot denote for a single-homed CE in S-EVPN. The ESI for the single-homed CE must be unique network wide and can be created automatically. The ESI is encoded as a ten octets integer. One way to generate ESI value for a single-homed CE is to use the MAC address of the Ethernet Segment with the low order 4 octets filled by value 0. The ESI value generation for multi-homed CE is specified in EVPN and can be reused in S-EVPN. Through Ethernet A-D route per Ethernet Segment, RR+ can learn all Ethernet Segments on all PEs.

2. ES Global Label Allocation

RR+ allocates global labels for the Ethernet Segments discovered and advertises <ES, label> pair to all PEs. The PEs that are members of E-VPN keep track of the global label/Ethernet Segment mappings.

The PE nodes perform the following functions:

- Learn customer/client MAC addresses (C-MACs) over the attachment circuits in the data-plane, per normal bridge operation.
- Learn remote C-MAC to B-MAC bindings in the data-plane from traffic ingress from the core per [802.1ah] bridging operation.
- Advertise local B-MAC address reachability information in BGP to all other PE nodes in the same set of service instances. Note that every PE has a set of local B-MAC addresses that uniquely identify the device. More on the PE addressing in section 5.
- Build a forwarding table from remote BGP advertisements received associating remote B-MAC addresses with remote PE IP addresses and the associated MPLS label(s).

4.3. Ethernet A-D Route Per EVI

The procedures defined for Ethernet A-D router per EVI in [I-D.ietf-l2vpn-evpn] will be reused by S-EVPN. In S-EVPN, both single home CE and multi-home CE have a unique ES identification. So for both single-homed CEs and multi-homed CEs, PEs need to allocate MPLS label for the <ESI, EVI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> and advertise corresponding Ethernet A-D routes per EVI. The MPLS label is used to identify a specific ES in an EVI.

4.4. Ethernet A-D Route Per ES

In S-EVPN, support of Ethernet A-D Route per Ethernet Segment is still MANDATORY. PEs can learn Ethernet Segments through this type of route as E-VPN. In S-EVPN, RR+ which all PEs connect to can also learn Ethernet Segments. When constructing the Ethernet A-D Route per Ethernet Segment, there are following differences from E-VPN:

-- The ESI for the single-homed CE in this route must be unique network wide instead of 0.

-- The "ESI Label Extended Community" MUST be included in the route and the "Active-Standby" bit in the flags MUST be set accordingly. But the MPLS label in the extended community can be set as 0 (Invalid MPLS label value) since ES global label is introduced in S-EVPN which can substitute ESI label.

5. Improvement on EVPN

When S-EVPN process is introduced, the E-VPN process defined by [I-D.ietf-l2vpn-evpn] can also be improved. The improvement includes split horizon, unifying unicast and multicast forwarding.

5.1. Split Horizon

ES global label is introduced to identify the Ethernet Segment globally. Thus S-EVPN can fulfill requirements proposed PBB-EVPN. Besides this, the ES global label can also be used for split horizon in EVPN. In order to achieve split horizon function, E-VPN adopts ESI label to encapsulate it in every BUM packet originating from a non-DF PE to identify the Ethernet Segment of origin. ES global label can use for the same purpose since it can identify the Ethernet Segment. Every BUM packet originating from a non-DF PE is encapsulated as the encapsulation which is shown in the figure 2. Since the original ESI label in E-VPN can be substituted by the ES global label, the ESI label in the ESI Label Extended Community can be an invalid label value. For the reason of compatibility, the ESI Label Extended Community can carry a valid ESI label. Both ESI label and ES global label should be used for split horizon no matter which label is encapsulated in the packet.

ES global label can also solve the possible issue for split horizon when MP2MP LSP is used to transport BUM traffic. When P2MP LSPs is used, the upstream label assignment mechanism is introduced for split horizon. When PE received the packet, it decapsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label allocated by the ingress PE for a specific Ethernet Segment, the

received PE will not forward the packet on the corresponding ES. In the MP2MP LSP scenarios, there are multiple roots and the upstream label allocated for Ethernet Segment maybe the same. So the received PE cannot determine a correct context label space according the top label for the MP2MP LSP. That is, the upstream label assignment mechanism for split horizon introduced in the P2MP LSP scenario can not be reused in the MP2MP LSP. But if the ES global label is used, in the MP2MP LSP scenario the received PE can also determine not to forward the packet on the specific ES which is identified by the ES global label. In one word, no matter ingress replication, P2MP LSP, or MP2MP, S-EVPN provides a unified solution based on the ES global label. It can reduce the complexity of the split horizon mechanism in E-VPN.

5.2. Unifying MPLS Forwarding

S-EVPN adopts MPLS forwarding for C-MAC learning. In the control plane, it is just to add one new route type for E-VPN. It is a smooth upgrading of E-VPN and can switch easily between C-MAC learning through control plane and C-MAC learning through data plane.

When C-MACs is learned through the control plane, the unicast forwarding uses the label for the MAC route which is shown as follows:

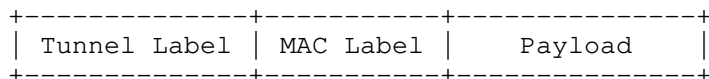


Figure 4: E-VPN Unicast Forwarding Encapsulation

When C-MACs is learned through the data plane, the unicast forwarding uses the EVI label and the Segment global label which is shown in figure 2. In fact even if the C-MAC is learned through the data plane, the data plane can also use following encapsulation. In this case, the label in MAC advertisement route should not be used. From the comparison, we can see that when E-VPN and S-EVPN are introduced, the forwarding encapsulation can be unified no matter which way C-MACs are learned by.

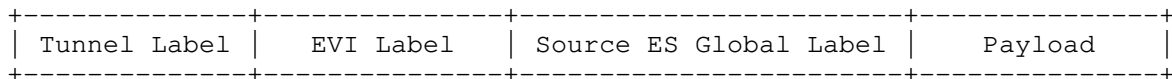


Figure 5: Unicast Forwarding Encapsulation without MAC Label

6. BGP E-VPN NLRI Extensions

6.1. ES Global Label Request Extended Community

ES Global Label Request Extended Community may be advertised along Ethernet A-D route per Ethernet Segment. ES Global Label Request Extended Community can reuse ESI Label Extended Community defined in [I-D.ietf-l2vpn-evpn] which is shown in the following figure:

```

 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=0x01 | Flags (One Octet) | Reserved=0 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Reserved = 0 | ESI Label
+-----+-----+-----+-----+-----+-----+-----+-----+

```

There defines a new bit of the flag octet as the "Global Label Request" bit.

```

+-----+
| * | * | * | * | * | 2 | 1 | 0 |
+-----+

```

Bit0: "Active-Standby" bit

Bit1: "Root-Leaf" bit

Bit2: "Global Label Request" bit

The third low order bit of the flags octet is defined as the "Global Label Request". A value of 0 means there is no global label request for the Ethernet A-D route. A value of 1 means that global label request is associated with the Ethernet A-D route.

6.2. ES Global Label Mapping Route

A new route type is defined for E-VPN NLRI to allocate global label for Ethernet Segment:

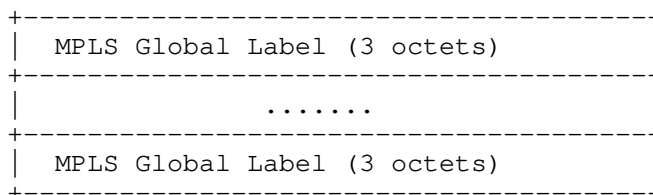
+5 - ES Global Label Mapping Route

An ES Global Label Mapping route type specific E-VPN NLRI consists of the following:

```

+-----+
| RD      (8 octets) |
+-----+
| Ethernet Segment Identifier (10 octets) |
+-----+
| Ethernet Tag ID (4 octets) |
+-----+

```



7. Operations

7.1. ES Global Label Request

Global label request is only for the Ethernet A-D route per Ethernet Segment. The Ethernet A-D route per Ethernet Segment is constructed as defined by [I-D.ietf-l2vpn-evpn]. The Ethernet Segment Identifier MUST be a unique ten octet entity. Even if the CE is single-homed, the corresponding Ethernet Segment Identifier MUST NOT be the reserved value 0.

When request a global label for a specific Ethernet Segment, ES Global Label Request Extended Community MUST be used for the Ethernet A-D route. ES Global Label Request Extended Community of S-EVPN can reuse the ESI Label Extended Community. The "Global Label Request" bit of the flag octet MUST be set as 1 for Global Label Request. According to Section 5 "Improvement on E-VPN", if ES global label is introduced, the original ESI label may not be used. The "root-leaf" bit of the flag octet and the ESI Label value in the ESI Label Extended Community can always be 0 to simplify the process.

One or more Route Target(RT) MUST be carried with the Ethernet A-D route. These RTs are the set of RTs associated with all the EVIs to which the Ethernet Segment belongs. Since the Global label is allocated per Ethernet Segment, RTs carried by the Ethernet A-D route will be ignored by the RR+ when allocate global label for the Ethernet Segment specified in the Ethernet A-D routes. The global label per Ethernet Segment is advertised to all PEs. For multi-homed Ethernet Segment, if one EVI on one PE requests label allocation for the Ethernet Segment and the ES Global Label Mapping Route has been advertised corresponding to the Ethernet Segment, other EVIs on other PEs SHOULD NOT send the global label request for the Ethernet Segment again, that is, the "Global Label Request" bit SHOULD set as 0 when advertise Ethernet A-D routes for the Ethernet Segment by these EVIs.

7.2. ES Global Label Allocation

When RR+ receives the Ethernet A-D route per Ethernet Segment and the "Global Label Request" bit of the ES Global Label Request Extended Community is set as 1, RR+ MUST allocate global label for the Ethernet Segment and advertise the ES Global Mapping route to all PEs.

The ES Global Label Mapping route is constructed as follows:

RD, Ethernet Segment Identifier and Ethernet Tag ID values can be directly derived from the corresponding Ethernet A-D route per Ethernet Segment.

The MPLS Global Label field carries one or more labels (that corresponds to the stack of labels [MPLS-ENCAPS]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [MPLS-ENCAPS]).

One or more Route Target(RT) MUST be carried with the ES Global Label Mapping route. These RTs can be directly derived from the RTs associated with the corresponding Ethernet A-D route.

For multi-homed Ethernet Segment, there maybe multiple global label request for the same Ethernet Segment advertised to RR+ by different PEs. When RR+ receives them, if RTs for these routes are same, owing to the Ethernet Segment Identifier is the same, it SHOULD advertise only one corresponding ES Global Label Mapping Route to all PEs. That is, the subsequent global label request for the same Ethernet Segment SHOULD be ignored. If RTs carried with the Ethernet A-D routes for the Ethernet Segment are different, RR+ SHOULD advertise multiple ES Global Label Mapping Routes with the same global label value and different RTs.

8. Solution Advantages

S-EVN has following advantages:

1. Remove the requirement of automating B-MAC address assignment to simplify provision of PBB-EVPN.
2. Improve the encapsulation efficiency of PBB-EVPN.
3. Seamless MPLS thoughts to solve the issue dealt with by PBB-EVPN instead of combination of two distinct technologies.
4. Be able to unify the split horizon mechanisms for ingress replication, P2MP LSP, and MP2MP LSP in E-VPN.

5. Be able to unify unicast traffic forwarding of E-VPN to implement seamless switch between C-MACs learning through control plane and C-MACs learning through data plane.

9. IANA Considerations

This document requires IANA to assign a new route type value for E-VPN NLRI.

10. Security Considerations

There are no additional security aspects beyond those of VPLS/H-VPLS that need to be discussed here.

11. Acknowledgements

TBD.

12. Normative References

[I-D.ietf-l2vpn-evpn-req]

Sajassi, A., Aggarwal, R., Bitar, N., and A. Isaac,
"Requirements for Ethernet VPN (E-VPN)", draft-ietf-l2vpn-
evpn-req-03 (work in progress), May 2013.

[I-D.ietf-l2vpn-evpn]

Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F.,
Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN",
draft-ietf-l2vpn-evpn-03 (work in progress), February
2013.

[I-D.ietf-l2vpn-pbb-evpn]

Sajassi, A., Salam, S., Boutros, S., Bitar, N., Isaac, A.,
and L. Jin, "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-04 (work
in progress), February 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Lucy Yong
Huawei Technologies
1700 Alma Dr. Suite 500
Plano, TX 75075
USA

Email: lucyyong@huawei.com

Junlin Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jackey.zhang@huawei.com

L2VPN Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan
W. Henderickx
S. Sathappan
S. Palislamovic
Alcatel-Lucent

F. Balus
Nuage Networks

Expires: January 16, 2014

July 15, 2013

Data Center Interconnect Solution for E-VPN Overlay networks
draft-rabadan-l2vpn-dci-evpn-overlay-00.txt

Abstract

This document describes how Network Virtualization Overlay networks (NVO3) can be connected to a Wide Area Network (WAN) in order to extend the layer-2 connectivity required for some tenants. The solution will analyze the interaction between NVO3 networks running E-VPN and other technologies used in the WAN, such as VPLS/PBB-VPLS or E-VPN/PBB-EVPN.

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 16, 2014.

Copyright Notice

Copyright (c) 2013 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	3
2. VPLS/PBB-VPLS based DCI for E-VPN overlay networks	3
2.1. VPLS/PBB-VPLS DCI Solution Overview	3
2.2. VPLS/PBB-VPLS DCI options	4
2.2.1. VPLS DCI with VLAN-based hand-off	4
2.2.2. VPLS DCI with Pseudowire-based hand-off	5
2.2.3. VPLS DCI with integrated Gateway and WAN Edge functions	6
2.2.4. PBB-VPLS DCI	6
2.3. Unknown MAC route on the DC GWs	7
2.4. Disabling unknown unicast flooding in a DC with VPLS DCI	8
2.5. ARP-flooding control	9
2.6. Multi-homing solution for VPLS DCI	9
2.6.1. Multi-homing solution requirements for VPLS DCI	9
2.6.2. Multi-homing solution description	10
2.6.2.1. Multi-homed Ethernet Segment Auto-Discovery	11
2.6.2.2. Designated Forwarder (DF) election and forwarding	11
2.6.2.3. Fast Convergence using the Unknown MAC Route	11
3. E-VPN DCI for E-VPN overlay networks	13
4. PBB-EVPN DCI for E-VPN overlay networks	13
5. Conventions and Terminology	14
6. Security Considerations	14
7. IANA Considerations	14
8. References	14
8.1. Normative References	14
8.2. Informative References	15
9. Acknowledgments	15
10. Authors' Addresses	15

1. Introduction

[E-VPN-Overlays] discusses the use of E-VPN as the control plane for Network Virtualization Overlay (NVO) networks, where VXLAN, NVGRE or MPLS over GRE can be used as possible data plane encapsulation options.

While this model provides a scalable and efficient multi-tenant solution within the Data Center, it might not be easily extended to the WAN in some cases due to the existing deployed technologies. For instance, a Service Provider might have an already deployed VPLS network that must be used to interconnect Data Centers.

This document describes a Data Center Interconnect (DCI) solution for E-VPN overlay Data Center networks, assuming that the L2VPN technology deployed in the WAN can be based on:

1. VPLS as defined in [RFC4761][RFC4762][RFC6074] or even PBB-VPLS, as defined in [PBB-VPLS]
2. E-VPN as defined in [E-VPN]
3. PBB-EVPN as defined in [PBB-EVPN]

Each of these DCI models is analyzed in the following sections.

2. VPLS/PBB-VPLS based DCI for E-VPN overlay networks

VPLS and PBB-VPLS are deployed in many Service Providers as the multi-point L2VPN service technology in the WAN. Those Service Providers will require integrating the new virtualized data center services with the L2VPN technology existing in the WAN, so that there is a minimum impact on the Service Provider operations.

By implementing the Data Center Gateway (DC GW) functions described in this section, a Service Provider PE will be able to connect a DC tenant segment to an existing VPLS or PBB-VPLS service, for DC-to-DC layer-2 extension and for user-to-DC layer-2 connectivity.

2.1. VPLS/PBB-VPLS DCI Solution Overview

Figure 1 depicts the reference model described in this section.

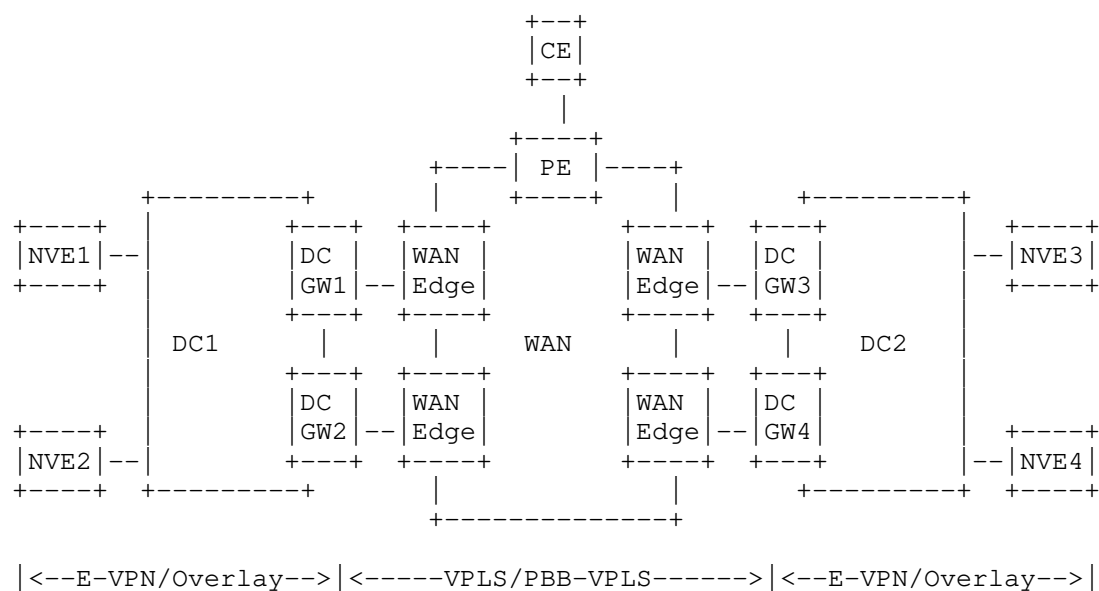


Figure 1 VPLS DCI model

In this model, the WAN Service Provider requires the use of its existing VPLS procedures to extend the layer-2 connectivity for the tenants. There are four potential options in this model:

- o VPLS DCI with VLAN-based hand-off
- o VPLS DCI with Pseudowire-based hand-off
- o VPLS DCI with integrated Gateway and WAN Edge functions
- o PBB-VPLS DCI

Section 2.2 describes each specific option.

2.2. VPLS/PBB-VPLS DCI options

2.2.1. VPLS DCI with VLAN-based hand-off

In this option, the hand-off between the DC GWs and the WAN Edge routers is based on 802.1Q VLANs. Each E-VPN Instance (EVI) in the DC GW is connected to a different VPLS Instance (VSI) in the WAN Edge router by using a different C-TAG VLAN ID or a different combination of S-TAG/C-TAG VLAN IDs that match at both sides. In this use-case, the WAN Edge router becomes a VPLS PE with regular VLAN-based Attachment Circuits.

This option is required in those cases where the WAN and DC networks are operated by different entities and a secure demarcation between both is needed (no control plane protocols are run between DC GW and WAN Edge router, and each network can apply its own security and QoS policies independently based on the incoming/outgoing VLAN ID). The disadvantages of this model are the provisioning overhead and the reduced scalability (limited to the VLAN-ID space). The provisioning in the DC GWs can be automated though by the cloud management system.

In this model, the DC GW acts as a regular Network Virtualization Edge (NVE) towards the D. Its control plane, data plane procedures and interactions are described in [E-VPN-Overlays]. From an E-VPN perspective, the connectivity to the WAN Edge routers is treated as VLAN-based service interfaces, therefore there is a 1:1 relation between DCI VLAN ID and EVI. If the data plane encapsulation in the NVO network supports VLAN tags in the encapsulated frames, a VLAN Bundle Service interface is possible in the DCI. As described in [E-VPN-Overlays] this interface type is possible if VXLAN is used and not for NVGRE. NVGRE only supports VLAN-based service interfaces.

The WAN Edge router acts as a VPLS PE. Its functions are described in [RFC4761] [RFC4762] [RFC6074].

The DC GW multi-homing functions for this model are described in section 2.6.

2.2.2. VPLS DCI with Pseudowire-based hand-off

If MPLS can be enabled between the DC GW and the WAN Edge router, a more scalable DCI solution can be deployed. In this option the hand-off between both routers is based on FEC128-based pseudowires or, alternatively, FEC129-based pseudowires for a greater level of network automation. Note that this model still provides a clear demarcation boundary between DC and WAN, and security/QoS policies may be applied on a per pseudowire basis.

In this model, besides the usual MPLS procedures between DC GW and WAN Edge router, the DC GW MUST support an interworking function in each EVI that requires extension to the WAN:

- o If a FEC128-based pseudowire is used between the EVI (DC GW) and the VSI (WAN Edge), the provisioning of the VCID for such pseudowire MUST be supported on the EVI and must match the VCID used in the peer VSI at the WAN Edge router.
- o If BGP Auto-discovery [RFC6074] and FEC129-based pseudowires are used between the DC GW EVI and the WAN Edge VSI, the provisioning of the VPLS-ID MUST be supported on the EVI and must match the

VPLS-ID used in the WAN Edge VSI. Note that the Route Distinguisher (RD) and Route Target (RT) already provisioned for its use in E-VPN, can be re-used for VPLS. The WAN Edge VSI will have to be configured with two different RT extended communities. For example, if EVI-1 in DC GW-1 (figure 1) uses RT1, the peer WAN Edge VSI will use RT1 to import/export routes from/to the DC GW and RT2 to import/export routes from/to the remote WAN Edge VSIs. The WAN Edge router will import RT1 and RT2 in two different split-horizon groups, so that traffic to/from the DC GW can be switched to/from the WAN.

The DC GW multi-homing functions for this model are described in section 2.6.

2.2.3. VPLS DCI with integrated Gateway and WAN Edge functions

When the DC and the WAN are operated by the same administrative entity, the Service Provider can decide to integrate the DC GW and WAN Edge PE functions in the same router for obvious CAPEX and OPEX saving reasons. In the example depicted in figure 1 that would mean the WAN Edge routers would be P routers and will not maintain any tenant state. Note that this model does not provide an explicit demarcation between DC and WAN anymore, and ACLs or QoS policies between both networks become a very complex task.

In this option, any EVI instance in the DC GW requiring layer-2 extension to the WAN MUST support an interworking function to VPLS. The EVI will become a VSI from the WAN perspective and will setup a full mesh of pseudowires to all the remote PEs and DC GWs (except to the DC GW of its own DC) and according to the procedures described in [RFC4761][RFC4762][RFC6074].

The DC GW multi-homing functions for this model are described in section 2.6.

2.2.4. PBB-VPLS DCI

This case is a variation of the one described in section 2.2.3. When the DC GW and WAN Edge PE functions can be integrated, PBB-VPLS can also be used as the DCI technology of choice. In this case, the DC GW EVIs will become I-components multiplexed into a B-component that will be connected to the WAN.

Since many EVIs can be multiplexed into the same B-component, this option provides significant savings in terms of pseudowires to be maintained in the WAN.

The DC GW multi-homing functions for this model are described in

section 2.6.

2.3. Unknown MAC route on the DC GWs

The use of E-VPN, as the control plane of Network Virtualization Networks in the DC, brings a significant number of benefits as described in [E-VPN-Overlays]. There are however two potential issues that SHOULD be addressed when a VPLS DCI is used for a NVO3 DC:

- o All the MAC addresses learnt from the WAN side of the VSI must be advertised by BGP E-VPN updates. Even if optimized BGP techniques like RT-constraint are used, the amount of MAC addresses to advertise or withdraw (in case of failure) from the DC GWs can be difficult to control and overwhelming for the DC network, especially when the NVEs reside in the hypervisors.
- o As described in [E-VPN-Overlays], when the NVEs reside in the hypervisors, the E-VPN BGP routes and attributes associated with multi-homing are no longer required. The simple reason is the fact that, in a hypervisor environment, there is no need for multi-homing between VMs and NVEs since both, VMs and NVEs, are part of the same hardware. This reduces the required routes to be generated and processed to only two: the MAC Advertisement Route and the Inclusive Multicast Ethernet Tag Route. While this simplification greatly helps the implementation of E-VPN in the DC, it brings back some of the issues related to Multi-Homing that were solved by the removed procedures and that are still applicable for the specific use-case of the DC, since Multi-Homing is required at the DC GWs.

The solution suggested in this document for the VPLS DCI use case is based on the use of an "Unknown MAC route" that is advertised by the two DC GWs. By using this Unknown MAC Route advertisement the user may optionally turn off the advertisement of WAN MAC addresses in the DC GW, hence reducing the control plane overhead and the size of the FDB tables in the NVEs. In addition to this, the Unknown MAC Route may provide a fast convergence solution valid for TORs and hypervisor NVEs, even if they do not support the Ethernet A-D route procedures.

If this procedure is used, when an EVI is created in the DC GWs and the Designated Forwarder (DF) is elected, the DF will send a BGP update containing an "Unknown MAC" address. The Unknown MAC address will be conveyed in an "Unknown MAC" Advertisement Route:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
VNI/VSID (4 octets)
MAC Address Length (1 octet)
Unknown MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (4 or 16 octets)
MPLS Label (3 octets)

Where the ESI identifies the WAN Ethernet Segment, the VNI/VSID is encoded in the Ethernet Tag Field as explained in [E-VPN-Overlays], the MAC address length is set to 48 and the Unknown MAC address value will be set to 00:00:00:00:00:00. The IP address length will be zero, the IP address value omitted and the MPLS label will be set to zero.

If the DC GW is DF for more than one ES within the same EVI, it will advertise an Unknown MAC route per ES, each one tagged with its corresponding ESI.

As outlined before, there are two main functions that can be carried out by using this Unknown MAC Route: fast convergence for hypervisor NVEs (described in section 2.6.2.3) and disabling unknown unicast flooding in the DC (described in section 2.4).

2.4. Disabling unknown unicast flooding in a DC with VPLS DCI

In DCs where MAC addresses are learnt through the control plane, the use of flooding for unknown destination MAC addresses can be disabled. However, when we use a VPLS DCI, the DC GW will normally learn the WAN MAC addresses in the data plane, therefore, even if the rest of the NVEs in the DC do control plane learning, disabling the unknown unicast flooding is no longer an administrative choice.

The use of the Unknown MAC route in DC GWs allows two configuration options:

- a) Disable the unknown flooding in all the NVEs in the DC (except on the DC GWs) if Data Center MACs are learnt through the

control/management plane.

- b) Disable the advertisement of the WAN MAC addresses from the DC GWs, so that the control plane overhead and the forwarding table sizes in the NVEs are both reduced.

Both options SHOULD be an administrative configuration choice supported on the DC GWs.

If option b) is enabled, the DC GW will advertise only the Unknown MAC Route for the EVIs on which it is the Designated Forwarder (DF). The NVEs will learn their local MACs through the control/management plane and advertise them in BGP. If any NVE receives a packet to an unknown destination MAC address, and option a) is enabled, the NVE will send the packet to the next-hop associated to the Unknown MAC Route (for each ESI if there is more than one), since the packet is assumed to be destined to the WAN. This assumption is valid since all the DC MACs are learnt in the control/management plane. The DC GW will receive the packet and will do an FDB lookup to find out what VPLS pseudowire or attachment circuit it has to send the packet to. If the destination MAC is unknown for the DC GW, it will flood the packet to the WAN, following standard VPLS procedures.

2.5. ARP-flooding control

The use of the Unknown MAC route may eliminate the unknown flooding within the DC and provide an extra security protection mechanism against an excessive explosion of MAC addresses in the WAN that would trigger the advertisement of a significant number of MAC addresses in the DC.

Another security mechanism, naturally provided by E-VPN in the DC GWs, is the Proxy ARP function. The DC GWs SHOULD build a Proxy ARP table with the IP and MAC address information coded in the MAC advertisement routes coming from the DC NVEs. When the active DC GW receives an ARP request coming from the WAN, the DC GW should check the Proxy ARP table for the EVI and reply to the ARP request as long as the information is available.

This mechanism is specially recommended when VPLS DCI is used on the DC GWs since it protects the DC network from external ARP-flooding.

2.6. Multi-homing solution for VPLS DCI

2.6.1. Multi-homing solution requirements for VPLS DCI

As it can be easily inferred from the scenario in figure 1, a multi-homing solution MUST be provided in the DC. The Multi-homing

requirements on the DC GWs are listed here:

- o A Multi-homing solution MUST be supported by the DC GWs independently of the capabilities of the WAN Edge routers (since they can be managed by a different Service Provider).
- o The Multi-homing solution MUST support service-based (EVI) load-balancing. No flow-based load-balancing is required when VPLS DCI is used.
- o The Multi-homing solution MUST support single-active redundancy mode per E-VPN on the DC GWs, as per [E-VPN]. All-active multi-homing is neither possible if VPLS is used in the DCI nor required since the number of EVIs on the DC GWs is supposed to be large enough so that the traffic between DC and WAN can be fairly distributed.

2.6.2. Multi-homing solution description

When the DCI model is the one described in the section 2.1, a single-active Multi-homing solution is required. Note that, since all-active Multi-homing is not required, only a subset of E-VPN routes and extended communities will be needed to be generated from the DC GWs:

- o Ethernet Segment route and ES-Import route target: required for the Ethernet Segment Auto-Discovery and Designated Forwarder (DF) election between the DC GWs. The DC GWs MUST generate an ES route per WAN network to which they are directly connected, and MUST be able to process the inbound ES routes as per [E-VPN].
- o Ethernet Auto-Discovery (A-D) route per ESI: required for fast convergence and back-up path. The DC GWs MUST generate an A-D route per ESI with an ESI Label extended community. The active-standby flag will be always set and the label field will be zero (no Split-Horizon procedures are required on the DC GWs as per [E-VPN]). The DC GWs will be able to process the received A-D routes per ESI.
- o Ethernet Auto-Discovery (A-D) route per EVI: the DC GWs will NOT generate A-D routes per EVI, since no aliasing functions are required for single-active Multi-homing. The DC GWs however MUST support processing A-D routes per EVI, since there might be some TORs in the DC supporting all-active Multi-homing.
- o MAC Advertisement route and MAC Mobility extended community: they MUST be supported at generation and reception as per [E-VPN-Overlays].
- o Inclusive Multicast Ethernet Tag route and PMSI Tunnel attribute:

they MUST be supported at generation and reception as per [E-VPN-Overlays].

The above routes and communities will be used for the following Multi-homing functions:

2.6.2.1. Multi-homed Ethernet Segment Auto-Discovery

The DC GWs will advertise an Ethernet Segment route per WAN Ethernet Segment (ES), with the corresponding ES-Import extended community. There will be a single ESI per WAN network, i.e. DC GW1 and DC GW2 will only advertise one ESI in the example of figure 1, and only the DC GWs of the DC will import the ES route for the WAN ESI, as per [E-VPN].

2.6.2.2. Designated Forwarder (DF) election and forwarding

The DF election will be carried out as described in [E-VPN]. Service carving is recommended so that there can be per EVI load-balancing to/from the WAN. Assuming DC GW1 is elected as DF for a given EVI1, DC GW1 will be the only DC GW sending/receiving traffic to/from the WAN for EVI1. DC GW2 will block the transmission and reception of any traffic (including unicast and multi-destination traffic) to/from the WAN for EVI1.

The use of OAM is recommended from the non-DF to the VPLS network, so that the VPLS PEs do not send any traffic to the non-DF DC GW for the EVI in which the DC GW is non-DF:

- o If the VPLS DCI solution is based on a VLAN hand-off, 802.1ag/Y.1731 Ethernet-CFM can be used by the non-DF DC GW so that the peer WAN Edge router do not send any traffic to the DC GW for that particular EVI.
- o If the VPLS DCI solution is based on a pseudowire hand-off, the LDP PW Status bits TLV can be used by the non-DF to signal "Standby status" to the WAN Edge router for that particular EVI.
- o If the VPLS DCI is based on an integrated DC GW and WAN Edge router solution where the EVI is part of the VPLS full mesh of pseudowires, the non-DF DC GW can also make use of the LDP PW Status bits TLV to let the remote PEs know that it is not forwarding traffic for that EVI/VSI.

2.6.2.3. Fast Convergence using the Unknown MAC Route

[E-VPN] proposes a Fast Convergence mechanism, so that when there is an ES failure on the DF router, the failover time can be uniform and

independent of the number of MACs and EVI services in the DC GWs. This is done by having the DC GWs advertising an A-D route per WAN Ethernet Segment. Upon a failure in connectivity to the WAN, the DF withdraws the Ethernet A-D route for the WAN Ethernet Segment so that the NVEs in the DC receiving the BGP withdraw can update their FDB for all the MAC addresses associated to the WAN ES.

This mechanism is valid as long as the NVEs in the DC support the Ethernet A-D route per ESI. However, as described in [E-VPN-Overlays], in the Data Center there will be a mix of NVEs supporting Ethernet A-D routes (TORs with dual-homed servers) and NVEs NOT supporting Ethernet A-D routes (hypervisors), hence a complementary fast convergence mechanism is required for those.

While the existing E-VPN Mass Withdraw procedure will be used for NVEs supporting the processing of Ethernet A-D routes, this document describes a complementary procedure for NVEs not supporting the processing of Ethernet A-D routes. The new procedure does not require the addition of any new route or extended community. It is just based on the interpretation of the Unknown MAC Route described in section 2.3 which will be sent by the DC GWs in regular MAC advertisement routes. The user MAY decide whether the Unknown MAC Route procedure is used only by the hypervisors or by the hypervisors and the TORs too.

Only one of the DC GWs will advertise the Unknown MAC Route per EVI and per WAN ESI. The DF will also advertise all the MAC addresses being learnt from the WAN Ethernet Segment (assuming option b in section 2.4 is not turned on). The hypervisor NVEs will import the Unknown MAC route as well as the rest of the WAN MAC addresses associated to the active DC GW. The Unknown MAC route is used by the active DC GW as a way of signaling that it owns the reachability to the WAN Ethernet Segment (ES) for a given EVI. The Unknown MAC address (00:...:00/48) conveyed in the Unknown MAC route will be added to the corresponding EVI forwarding table at the remote NVE.

When the WAN Ethernet Segment active path fails (due to a port or link failure), the DC GW will withdraw the Unknown MAC route on all the EVIs for which it is the DF. This triggers all the hypervisor NVEs that receive the withdraw advertisement to immediately invalidate all the MAC addresses associated to the Ethernet Segment, as opposed to having to wait for each individual MAC to be withdrawn.

This function is compatible with the E-VPN Fast Convergence procedure carried out by the use of the Ethernet A-D route. The Ethernet A-D route can still be used for TOR NVEs supporting all the E-VPN routes.

Note that while the E-VPN mass withdraw provides a fast convergence mechanism independent of the number of services and MACs, the Unknown MAC withdraw provides a fast convergence mechanism per service, independent of the number of MACs in each service, i.e. convergence is not expected to be uniform for all the services, but uniform for all the hosts within a service. The use of the Unknown MAC route can significantly speed up the convergence in hypervisor NVEs, especially in services with a fair amount of MACs.

3. E-VPN DCI for E-VPN overlay networks

Another potential DCI technology that can be used in the WAN is E-VPN. Assuming E-VPN for MPLS tunnels is used in the WAN, the use of a DC GW is required if the overlay tunneling technology deployed within the DC is not MPLS over GRE, i.e. if VXLAN or NVGRE are used.

Figure 2 illustrates this E-VPN DCI model.

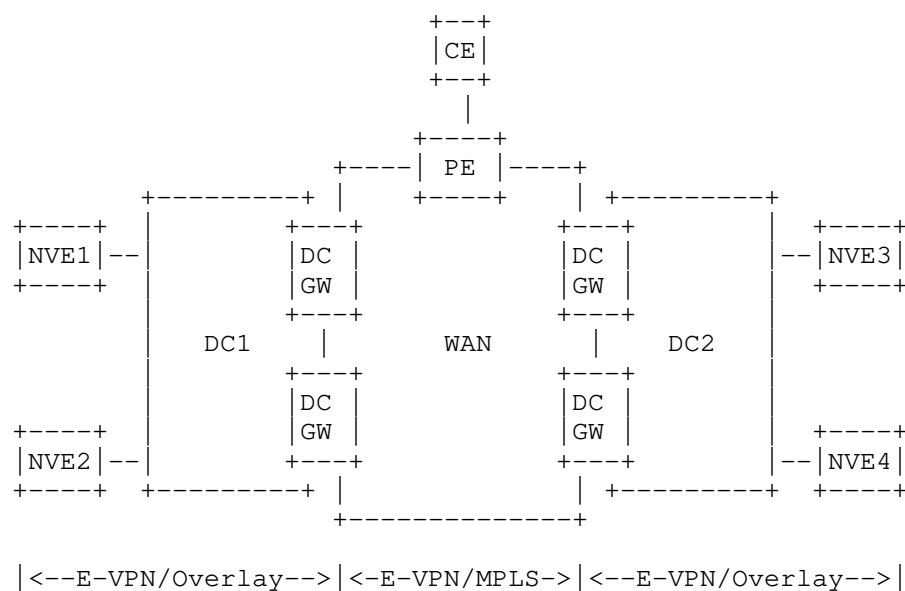


Figure 2 E-VPN DCI model

More information will be added in future versions of this document.

4. PBB-EVPN DCI for E-VPN overlay networks

[PBB-EVPN] is yet another DCI option. It requires the use of DC GWs where the multiplexing of I-components into the B-component is carried out. E-VPN will run in both components.

More information will be added in future versions of this document.

5. Conventions and 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].

DF: Designated Forwarder

DC GW: Data Center Gateway

DCI: Data Center Interconnect

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

EVI: E-VPN Instance

NVE: Network Virtualization Edge

TOR: Top-Of-Rack switch

VNI/VSID: refers to VXLAN/NVGRE virtual identifiers

6. Security Considerations

7. IANA Considerations

8. References

8.1. Normative References

[RFC4761] Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP)

Signaling", RFC 4762, January 2007.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo,
"Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual
Private Networks (L2VPNs)", RFC 6074, January 2011.

8.2. Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03.txt, work in progress, February, 2013

[E-VPN-OVERLAYS] Sajassi-Drake et al., "A Network Virtualization
Overlay Solution using E-VPN", draft-sd-l2vpn-evpn-overlay-01.txt,
work in progress, February, 2013

9. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

10. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Florin Balus
Nuage Networks
Email: florin@nuagenetworks.net

Senthil Sathappan
Alcatel-Lucent
Email: senthil.sathappan@alcatel-lucent.com

Senad Palislaamovic
Alcatel-Lucent

Email: senad.palislamovic@alcatel-lucent.com

L2VPN Workgroup
Internet Draft

Intended status: Standards Track

J. Rabadan
W. Henderickx
S. Palislaamovic
Alcatel-Lucent

F. Balus
Nuage Networks

A. Isaac
Bloomberg

Expires: January 16, 2014

July 15, 2013

IP Prefix Advertisement in E-VPN
draft-rabadan-l2vpn-evpn-prefix-advertisement-00

Abstract

E-VPN provides a flexible control plane that allows intra-subnet connectivity in an IP/MPLS and/or an NVO-based network. In Data Centers, there is also a need for a dynamic and efficient inter-subnet connectivity across Tenant Systems and End Devices that can be physical or virtual and may not support their own routing protocols. This document defines a new E-VPN route type for the advertisement of IP Prefixes and explains how E-VPN should be used to provide inter-subnet connectivity with the flexibility required by the Data Center applications.

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 16, 2014.

Copyright Notice

Copyright (c) 2013 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
1.1 Inter-subnet connectivity requirements in Data Centers	3
1.2 The requirement for advertising IP prefixes in E-VPN	5
1.3 The requirement for a new E-VPN route type	6
2. The BGP E-VPN IP Prefix route	8
2.1. IP Prefix Route encoding	9
2.2. BGP remote-next-hop attribute	9
3. Procedures associated to the advertisement of IP Prefixes . . .	10
3.1. Usage of the MAC advertisement and IP Prefix advertisement routes	10
3.2. Inter-subnet connectivity for TS	11
3.3. Inter-subnet connectivity for redundant TS (floating IP) .	13
3.4. Inter-subnet connectivity for IRB interfaces	15
3.4.1. Inter-subnet connectivity for unnumbered IRB interfaces	17
4. Conclusions	19
5. Conventions used in this document	20
6. Security Considerations	20
7. IANA Considerations	20
8. References	20
8.1. Normative References	20
8.2. Informative References	20
9. Acknowledgments	20
10. Authors' Addresses	21

1. Introduction and problem statement

Inter-subnet connectivity is required within the Data Center, therefore IP Prefixes must be advertised in the control plane. This section explains why IP-VPN [RFC4364] procedures cannot be used for such advertisements and why the existing E-VPN MAC route type does not meet the Data Center requirements for the advertisement of IP Prefixes, hence a new E-VPN route type is proposed.

Section 1.1 describes the inter-subnet connectivity requirements in Data Centers. Section 1.2 and 1.3 explain why neither IP-VPN nor the existing E-VPN route types meet the requirements for IP Prefix advertisements. Once the need for a new E-VPN route type is justified, sections 2 and 3 will describe this route type and how it is used in some specific use cases.

1.1 Inter-subnet connectivity requirements in Data Centers

[E-VPN] is used as the control plane for a Network Virtualization Overlay (NVO3) solution in Data Centers (DC), where Network Virtualization Edge (NVE) devices can be located in Hypervisors or TORs, as described in [E-VPN-OVERLAYS].

If we use the term Tenant System (TS) to designate a physical or virtual system identified by MAC and IP addresses, and connected to an E-VPN instance, the following considerations apply:

- o The Tenant Systems may be Virtual Machines (VMs) that generate traffic from their own MAC and IP.
- o The Tenant Systems may be Virtual Appliance entities (VAs) that forward traffic to/from IP addresses of different End Devices seating behind them.
 - o These VAs can be firewalls, load balancers, NAT devices, other appliances or virtual gateways with virtual routing instances.
 - o These VAs do not have their own routing protocols and hence rely on the E-VPN NVEs to advertise the routes on their behalf.
 - o In all these cases, the VA will forward traffic to the Data Center using its own source MAC but the source IP will be the one associated to the End Device seating behind or a translated IP address (part of a public NAT pool) if the VA is performing NAT.
- o Note that the same IP address could exist behind two of these

TS. One example of this would be certain appliance resiliency mechanisms, where a virtual IP or floating IP can be own by one of the two VAs running the resiliency protocol (the master VA). VRRP is one particular example of this. Another example is multi-homed subnets, i.e. the same subnet is connected to two VAs.

The following figure illustrates some of the examples described above.

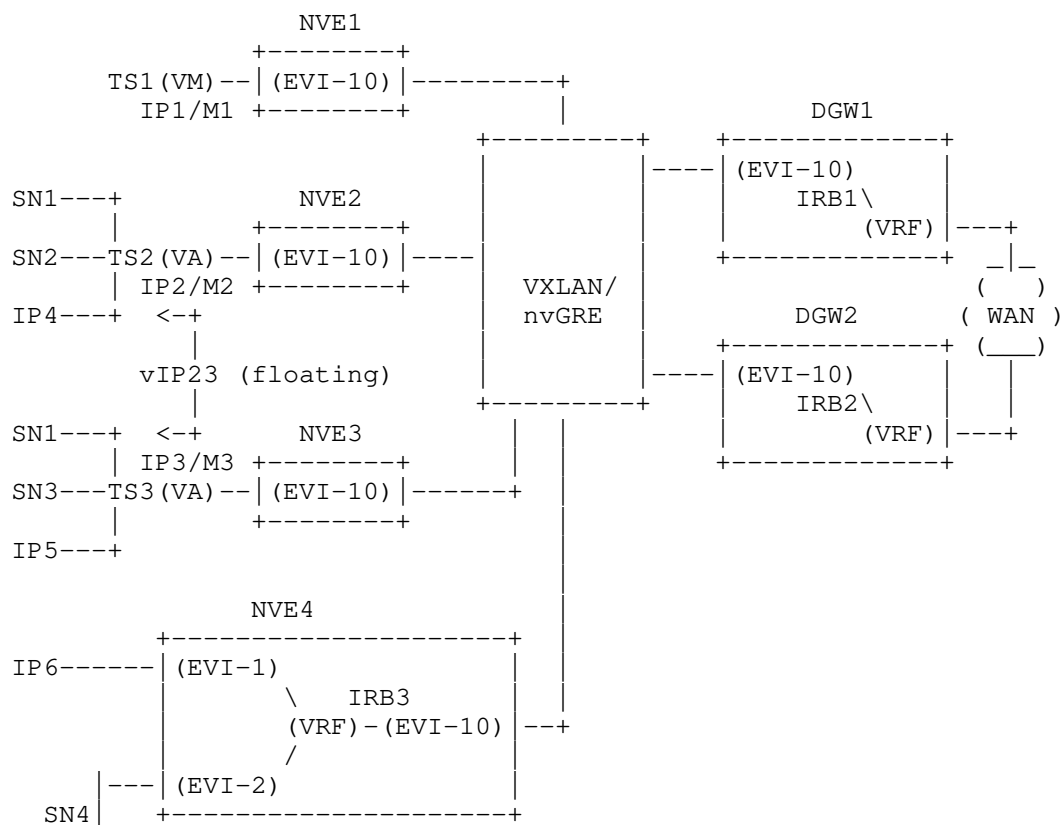


Figure 1 DC inter-subnet use-cases

Where:

NVE1, NVE2, NVE3, NVE4, DGW1 and DGW2 share the same E-VPN for a particular tenant. EVI-10 is the corresponding E-VPN instance on each element, and all the hosts connected to that instance belong to the same IP subnet. The hosts connected to E-VPN 10 are listed below:

- o TS1 is a VM that generates/receives traffic from/to IP1, where IP1 belongs to the E-VPN 10 subnet.
- o TS2 and TS3 are Virtual Appliances (VA) that generate/receive traffic from/to the subnets and hosts seating behind them (SN1, SN2, SN3, IP4 and IP5). Their IP addresses (IP2 and IP3) belong to the E-VPN subnet and they can also generate/receive traffic. When these VAs receive packets destined to their own MAC addresses (M2 and M3) they will route the packets to the proper subnet or host. These VAs do not support routing protocols to advertise the subnets connected to them and can move to a different server and NVE when the Cloud Management System decides to do so. These VAs may also support redundancy mechanisms for some subnets, similar to VRRP, where a floating IP is owned by the master VA and only the master VA forwards traffic to a given subnet. E.g.: vIP23 in figure 1 is a floating IP that can be owned by TS2 or TS3 depending on who the master is. Only the master will forward traffic to SN1.
- o Integrated Routing and Bridging interfaces IRB1, IRB2 and IRB3 have their own IP addresses that belong to the E-VPN 10 subnet too. These IRB interfaces connect the E-VPN 10 subnet to Virtual Routing and Forwarding (VRF) instances that can route the traffic to other connected subnets for the same tenant (within the DC or at the other end of the WAN). In some occasions, the IRB interfaces do not terminate IP traffic themselves and therefore they do not need any IP address configured. In such case, we will refer to these special IRB interfaces as "unnumbered" IRB interfaces.

All the above DC use cases use individual IP hosts and subnets for intra/inter connectivity. Therefore, their IP addresses MUST be advertised:

- a) From the NVEs (since VAs and VMs do not run routing protocols) and
- b) Associated to a next-hop that can be a VA IP address, a floating IP address, and IRB IP address or a MAC address.

1.2 The requirement for advertising IP prefixes in E-VPN

In all the inter-subnet connectivity cases discussed in section 1.1 there is a need to advertise IP prefixes in the control plane that cannot be satisfied by using [RFC4364] due to the following requirements, specific to NVO-based Data Centers:

- o The data plane in NVO-based Data Centers is not based on IP over a GRE or MPLS tunnel as required by [RFC4364], but Ethernet over an IP tunnel, such as VXLAN or NVGRE.

- o The IP prefixes in the DC must be advertised with a flexibility that does not exist in IP-VPNs. For instance:
 - a) The advertised next-hop for a given IP prefix can be an IRB IP address (see section 3.4), a floating IP address (see section 3.3) or even a MAC address (see section 3.4.1). In the future, the ESI could also be defined as a next-hop for the advertised prefixes.
 - b) As stated by [E-VPN-OVERLAYS], VXLAN or NVGRE virtual identifiers can have a global or a local scope. The implementation MUST support the flexibility to advertise IP Prefixes associated to a global identifier (32-bit value encoded in the E-VPN Ethernet Tag ID) or a locally significant identifier (20-bit value encoded in the MPLS label field). At the moment, [RFC4364] can only advertise Prefixes associated to a locally significant identifier (MPLS label).
- o IP prefixes must be advertised by NVE devices that have no VRF instances defined and no capability to process IP-VPN prefixes. These NVE devices just support E-VPN and advertise IP Prefixes on behalf of some connected Tenant Systems. In other words: any attempt to solve this problem by simply using [RFC4364] routes requires that any EVPN deployment must be accompanied with a concurrent IP-VPN topology, which is not possible in most of the cases.
- o Finally, Data Center providers want to use a single BGP Subsequent Address Family (AFI/SAFI) for the advertisement of addresses within the Data Center, i.e. BGP E-VPN only, as opposed to using E-VPN and IP-VPN in a concurrent topology. This minimizes the control plane overhead in TORs and Hypervisors and simplifies the operations.

E-VPN is extended - as described in this document - to advertise IP prefixes with the flexibility required by the current and future Data Center applications.

1.3 The requirement for a new E-VPN route type

[E-VPN] defines a MAC route (or route type 2) where a MAC address can be advertised together with an IP address length (IPL) and IP address (IP). While a variable IPL might be used to indicate the presence of an IP prefix in a route type 2, there are several specific use cases in which using this route type to deliver IP Prefixes is not suitable.

One example of such use cases is the "floating IP" example described in section 1.1. In this example we need to decouple the advertisement of the prefixes from the advertisement of the floating IP (vIP23 in figure 1) and MAC associated to it, otherwise the solution gets highly inefficient and does not scale.

E.g.: if we are advertising 1k prefixes from M2 (using route type 2) and the floating IP owner changes from M2 to M3, we would need to withdraw 1k routes from M2 and re-advertise 1k routes from M3. However if we use a separate route type, we can advertise the 1k routes associated to the floating IP address (vIP23) and only one route type 2 for advertising the ownership of the floating IP, i.e. vIP23 and M2 in the route type 2. When the floating IP owner changes from M2 to M3, a single route type 2 withdraw/update is required to indicate the change. The remote DGW will not change any of the 1k prefixes associated to vIP23, but will only update the ARP resolution entry for vIP23 (now pointing at M3).

Any other attempt to improve the efficiency of the solution when using non-MAC-decoupled Prefix advertisements, will derive in dependencies on the Cloud Management System (if ESIs are to be used) and changes in the current E-VPN semantics. The DC applications require mechanisms to provide IP Prefix resiliency independent of the E-VPN procedures.

Other reasons to decouple the IP Prefix advertisement from the MAC route are listed below:

- o Clean identification, operation of troubleshooting of IP Prefixes, not subject to interpretation and independent of the IPL and the IP value. E.g.: An IP address for ARP resolution must be always clearly distinguished from an /32 IP Prefix, or a default IP route 0.0.0.0/0 must always be easily and clearly distinguished from the absence of IP information.
- o MAC address information must not be compared by BGP when selecting two IP Prefix routes. If IP Prefixes are to be advertised using MAC routes, the MAC information is always present and part of the route key.
- o IP Prefix routes must not be subject to MAC route procedures such as MAC Mobility or aliasing. Prefixes advertised from two different ESIs do not mean mobility; MACs advertised from two different ESIs do mean mobility. Similarly load balancing for IP prefixes is achieved through IP mechanisms such as ECMP, and not through MAC route mechanisms such as aliasing.
- o NVEs that do not require processing IP Prefixes must have an

easy way to identify an update with an IP Prefix and ignore it, rather than processing the MAC route only to find out later that it carries a Prefix that must be ignored.

The following sections describe how E-VPN is extended with a new route type for the advertisement of prefixes and how this route is used to address the current and future inter-subnet connectivity requirements existing in the Data Center.

2. The BGP E-VPN IP Prefix route

The current BGP E-VPN NLRI as defined in [E-VPN] is shown below:

Route Type (1 octet)
Length (1 octet)
Route Type specific (variable)

Where the route type field can contain one of the following specific values:

- + 1 - Ethernet Auto-Discovery (A-D) route
- + 2 - MAC advertisement route
- + 3 - Inclusive Multicast Route
- + 4 - Ethernet Segment Route

This document defines an additional route type that will be used for the advertisement of IP Prefixes:

- + 5 - IP Prefix Route

The support for this new route type is OPTIONAL.

By using a separate route type for IP prefix advertisements, there is a clean separation of functions between route types, i.e. route type 2 or MAC Advertisement route will be used for MAC and ARP resolution advertisement, whereas route type 5 or IP Prefix route will be used for the advertisement of prefixes. Since this new route type is OPTIONAL, an implementation not supporting it will easily ignore the route, based on the route type value.

The detailed encoding of this route and associated procedures are

described in the following sections.

2.1. IP Prefix Route encoding

An IP Prefix advertisement route type specific E-VPN NLRI consists of the following fields:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octet)
IP Address (4 or 16 octets)
MPLS Label (3 octets)

Where:

- o RD, Ethernet Tag ID and MPLS Label fields will be used as defined in [E-VPN] and [E-VPN-OVERLAYS].
- o The Ethernet Segment Identifier will be zero for IP prefix advertisements in this version of the document, and be re-used in the future for other purposes.
- o The IP address length can be set to a value between 0 and 32 (bits) for ipv4 and between 0 and 128 for ipv6.
- o The IP address will be a 32 or 128-bit field (ipv4 or ipv6).
- o The total route length will indicate the type of prefix (ipv4 or ipv6).

The Eth-Tag ID, IP address length and IP address will be part of the route key used by BGP to compare routes. The rest of the fields will be out of the route key.

2.2. BGP remote-next-hop attribute

The BGP remote-next-hop attribute [BGP-REMOTE-NH] will be sent along with the IP Prefix advertisement to indicate the next-hop behind which the advertised prefix is located. The following table shows the different types of next-hops defined in this document and their

corresponding encoding in the BGP remote-next-hop attribute.

Prefix next-hop	Field in the remote-nh attribute
MAC address	sub-TLV (for VXLAN or NVGRE)
IRB IP address	tunnel address (ipv4 or ipv6)
Floating IP address	tunnel address (ipv4 or ipv6)

3. Procedures associated to the advertisement of IP Prefixes

This section describes the separate function of each E-VPN advertisement route: route type 2 for MAC/IP advertisements and route type 5 for IP Prefixes.

After defining the role of each route type and the benefits of using a separate route for IP Prefixes, the procedures associated to the advertisement of prefixes will be explained in three different use cases.

3.1. Usage of the MAC advertisement and IP Prefix advertisement routes

[E-VPN] describes the content of the BGP E-VPN route type 2 specific NLRI, i.e. MAC Advertisement Route, where the IP address length (IPL) and IP address (IP) of a specific advertised MAC are encoded. The subject of the MAC advertisement route is the MAC address (M) and MAC address length (ML) encoded in the route. The MAC mobility and other complex procedures are defined around that MAC address. The IP address information carries the host IP address required for the ARP resolution of the MAC.

The BGP E-VPN route type 5 defined in this document, i.e. IP Prefix Advertisement route, decouples the advertisement of IP prefixes from the advertisement of any MAC address related to it. This brings some major benefits to NVO-based networks where inter-subnet forwarding is required. Some of those benefits are:

- a) Upon receiving a route type 2 or type 5, an egress NVE can easily distinguish MACs and IPs for ARP resolution from IP Prefixes. E.g. an IP prefix with IPL=32 being advertised from two different ingress NVEs (as route type 5) can be identified as such and be imported in the designated routing context as two ECMP routes, as opposed to two ARP entries competing for the same IP.
- b) Similarly, upon receiving a route, an egress NVE not supporting processing IP Prefixes can easily ignore the update, based on the route type.

- c) A MAC route includes the ML, M, IPL and IP in the route key that is used by BGP to compare routes. Advertised IP Prefixes are imported into the designated routing context, where there is no MAC information associated to IP routes. In the example illustrated in figure 1, subnet SN1 should be advertised by NVE2 and NVE3 and interpreted by DGW1 as the same route coming from two different next-hops, regardless of the MAC address associated to TS2 or TS3. This is easily accomplished in the route type 5 by including only the IP information in the route key.
- d) By decoupling the MAC from the IP Prefix advertisement procedures, we can leave the IP prefix advertisements out of the MAC mobility procedures defined in [E-VPN] for MACs. In addition, this allows us to have an indirection mechanism for IP prefixes advertised from a MAC/IP that can move between hypervisors. E.g. if there are 1,000 prefixes seating behind TS2 (figure 1), NVE2 will advertise all those prefixes in type 5 routes associated to the next-hop IP2. Should TS2 move to a different NVE, a single MAC advertisement route withdraw for the M2/IP2 route from NVE2 will invalidate the 1,000 prefixes, as opposed to have to wait for each individual prefix to be withdrawn. This may be easily accomplished by using a different IP Prefix route type that is not tied to a MAC address.

3.2. Inter-subnet connectivity for TS

The following figure illustrates an example of inter-subnet forwarding for subnets seating behind Virtual Appliances (on TS2 and TS3).

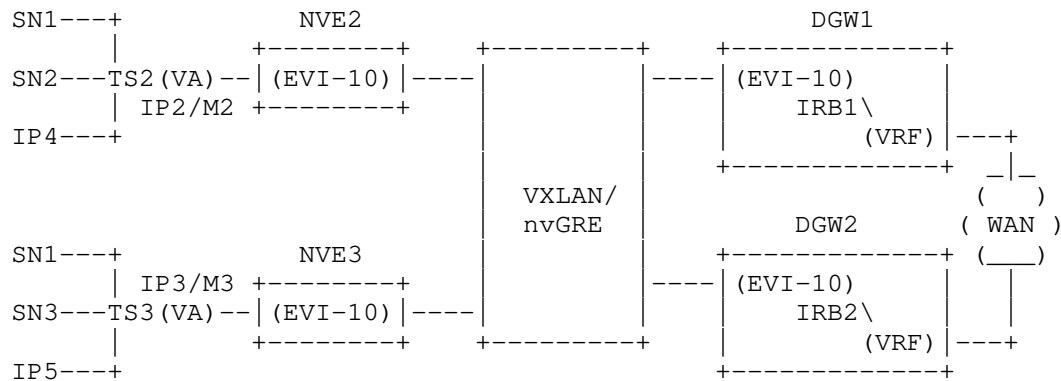


Figure 2 Inter-subnet forwarding for TS

An example of inter-subnet forwarding between subnet SN1/24 and a subnet seating in the WAN is described below. NVE2, NVE3, DGW1 and DGW2 are running BGP E-VPN. TS2 and TS3 do not support routing protocols, only a static route to forward the traffic to the WAN.

(1) NVE2 advertises the following BGP routes on behalf of TS2:

- o Route type 2 (MAC route) containing: ML=48, M=M2, IPL=32, IP=IP2
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, remote-nh tunnel address=IP2

(2) NVE3 advertises the following BGP routes on behalf of TS3:

- o Route type 2 (MAC route) containing: ML=48, M=M3, IPL=32, IP=IP3
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, remote-nh tunnel address=IP3

(3) DGW1 and DGW2 import both received routes based on the RT:

- o Based on the EVI-10 route-target in DGW1 and DGW2, the MAC route is imported and M2 is added to the EVI-10 MAC FIB along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]). IP2 - M2 is added to the ARP table.
- o Based on the EVI-10 route-target in DGW1 and DGW2, the IP Prefix route is also imported and SN1/24 is added to the designated routing context with next-hop IP2 pointing at the local EVI-10. Should ECMP be enabled in the routing context, SN1/24 would also be added to the routing table with next-hop IP3.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=IP2 is found. The tunnel information to encapsulate the packet will be derived from the route-type 2 (MAC route) received for M2/IP2.
- o IP2 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC FIB (remote VTEP and VNI for the VXLAN case).

- o The IP packet destined to IPx is encapsulated with:
 - . Source inner MAC = IRB1 MAC
 - . Destination inner MAC = M2
 - . Tunnel information provided by the MAC FIB (VNI, VTEP IPs and MACs for the VXLAN case)

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup.
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(5) Should TS2 move from NVE2 to NVE3, MAC Mobility procedures will be applied to the MAC route IP2/M2, as defined in [EVPN]. Route type 5 prefixes are not subject to MAC mobility procedures, hence no changes in the DGW VRF routing table will occur for TS2 mobility, i.e. all the prefixes will still be pointing at IP2 as next-hop. There is an indirection for e.g. SN1/24, which still points at next-hop IP2 in the routing table, but IP2 will be simply resolved to a different tunnel, based on the outcome of the MAC mobility procedures for the MAC route IP2/M2.

Note that in the opposite direction, TS2 will send traffic based on its static-route next-hop information (IRB1 and/or IRB2), and regular E-VPN procedures will be applied.

3.3. Inter-subnet connectivity for redundant TS (floating IP)

Sometimes Tenant Systems (TS) work in active/standby mode where an upstream floating IP - owned by the active TS - is used as the next-hop to get to some subnets behind. This redundancy mode, already introduced in section 1.1 and 1.3, is illustrated in Figure 3.

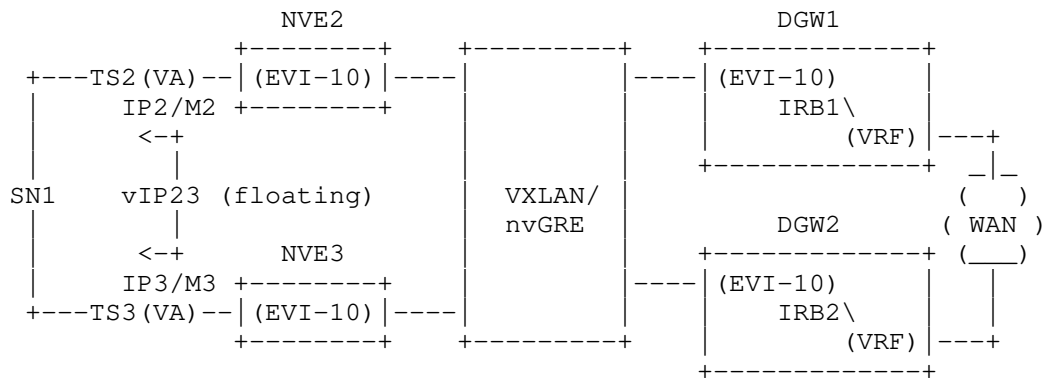


Figure 3 Inter-subnet forwarding for redundant TS

In this example, assuming TS2 is the active TS and owns IP23:

- (1) NVE2 advertises the following BGP routes for TS2:
 - o Route type 2 (MAC route) containing: ML=48, M=M2, IPL=32, IP=IP23
 - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, remote-nh tunnel address=IP23
- (2) NVE3 advertises the following BGP routes for TS3:
 - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, remote-nh tunnel address=IP23
- (3) DGW1 and DGW2 import both received routes based on the RT:
 - o M2 is added to the EVI-10 MAC FIB along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]). IP23 - M2 is added to the ARP table.
 - o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop IP23 pointing at the local EVI-10.
- (4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:
 - o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=IP23 is found. The tunnel information to encapsulate the packet will be derived from the route-type 2 (MAC route) received for M2/IP23.

- o IP23 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC FIB (remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:
 - . Source inner MAC = IRB1 MAC
 - . Destination inner MAC = M2
 - . Tunnel information provided by the MAC FIB (VNI, VTEP IPs and MACs for the VXLAN case)

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup.
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(5) When the redundancy protocol running between TS2 and TS3 appoints TS3 as the new active TS for SN1, TS3 will now own the floating IP23 and will signal this new ownership (GARP message or similar). Upon receiving the new owner's notification, NVE3 will issue a route type 2 for M3-IP23. DGW1 and DGW2 will update their ARP tables with the new MAC resolving the floating IP. No changes are carried out in the VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1 (from NVE2 and NVE3) but only the one with the same BGP next-hop as the IP23 route type 2 BGP next-hop will be valid.

3.4. Inter-subnet connectivity for IRB interfaces

In some other cases, the NVEs and DGWs will have just IRB interfaces as hosts in the E-VPN instance. Figure 4 illustrates an example.

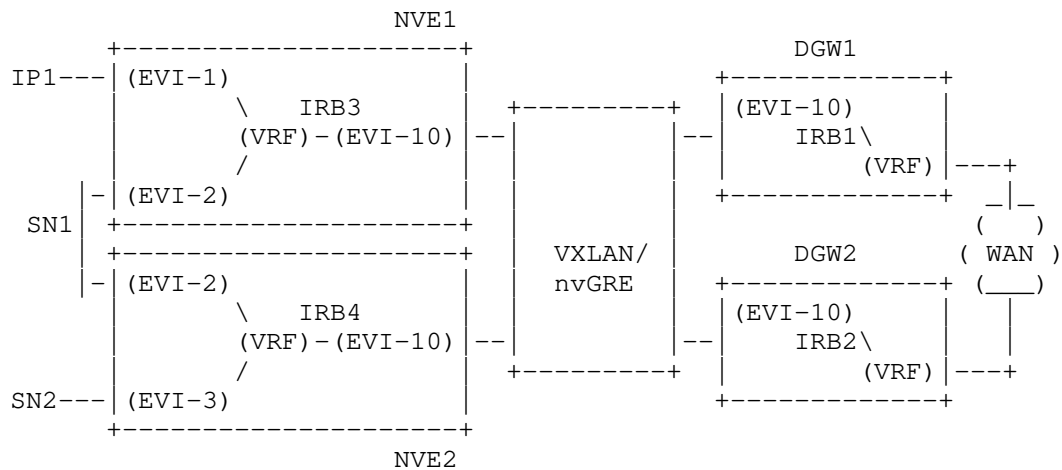


Figure 4 Inter-subnet forwarding for IRB interfaces

In this case:

- (1) NVE1 advertises the following BGP routes for SN1 resolution:
 - o Route type 2 (MAC route) containing: ML=48, M=IRB3-MAC, IPL=32, IP=IRB3-IP
 - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, remote-nh tunnel address=IRB3-IP
- (2) NVE2 advertises the following BGP routes for SN1 resolution:
 - o Route type 2 (MAC route) containing: ML=48, M=IRB4-MAC, IPL=32, IP=IRB4-IP
 - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, remote-nh tunnel address=IRB4-IP
- (3) DGW1 and DGW2 import both received routes based on the RT:
 - o IRB3-MAC and IRB4-MAC are added to the EVI-10 MAC FIB along with their corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]). IRB3-MAC - IRB3-IP and IRB4-MAC - IRB4-IP are added to the ARP table.
 - o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop IRB3-IP (and/or IRB4-IP) pointing at the

local EVI-10.

Similar forwarding procedures as the ones described in the previous use-cases are followed.

3.4.1. Inter-subnet connectivity for unnumbered IRB interfaces

In the previous example, the E-VPN instance can connect IRB interfaces and any other Tenant Systems connected to it. E-VPN provides connectivity for:

- a) Traffic destined to the IRB IP interfaces as well as
- b) Traffic destined to IP subnets seating behind the IRB interfaces, e.g. SN1 or SN2.

In order to provide connectivity for (a) we need MAC routes (route-type 2) distributing IRB MACs and IPs. Connectivity type (b) is accomplished by the exchange of IP Prefix routes (route-type 5) for IPs and subnets seating behind IRBs. As discussed in this document, prefixes are advertised along with their corresponding remote next-hop tunnel address, and those tunnel addresses are used to link prefixes to MAC/IPs advertised in MAC routes (type 2).

In some cases, connectivity type (a) (see above) is not required and the E-VPN instance is connecting only IRB interfaces, which are never the final destination of any packet. This use case is depicted in the diagram below and we refer to it as the "unnumbered IRB interface" use-case:

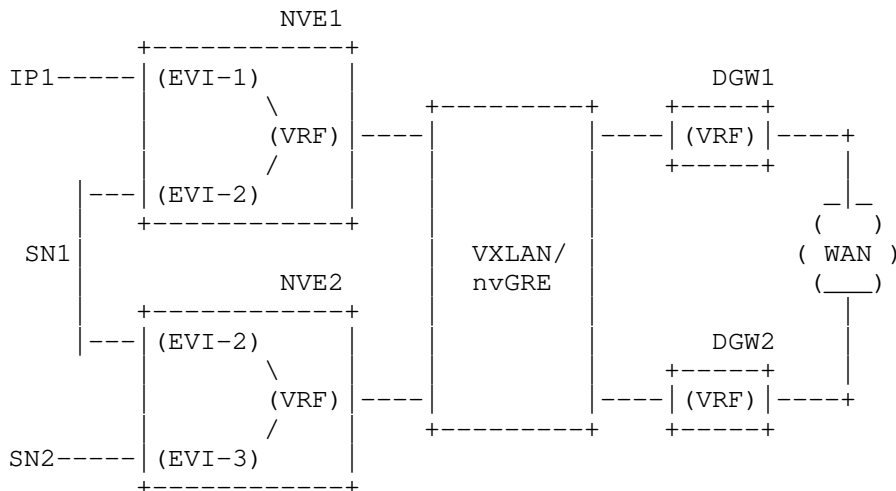


Figure 5 Inter-subnet forwarding for unnumbered IRB interfaces

In this case, we need to provide connectivity from/to IP hosts in SN1, SN2, IP1 and hosts seating at the other end of the WAN. The E-VPN in the core just connects all the IRBs in NVE1, NVE2, DGW1 and DGW2 but there will not be any IP host in this core E-VPN that is the final destination of any IP packet.

Therefore there is no need to define IRB IP addresses (IRBs are not represented in the diagram). This is the reason why we refer to this solution as "unnumbered Ethernet IRB" solution.

In this case, the proposal is to use EVPN type 5 routes and the BGP Remote-Next-Hop attribute, where the following information is carried:

- o Route type 5 Eth-Tag ID can contain the core instance VNI (if the VNI is global, otherwise, for local significant VNIs, an MPLS label field may be added with a 20-bit VNI encoded in the label space, as per [E-VPN-OVERLAYS]).
- o Route type 5 IP address length and IP address, as explained in the previous section.
- o Remote next-hop Tunnel Type is: TBD for VXLAN and TBD for NVGRE (TBD by IANA).
- o Remote next-hop Tunnel Address is populated with zeros, meaning that the prefix next-hop is an "unnumbered IRB".
- o Remote next-hop sub-TLV (for VXLAN/NVGRE) in the Tunnel Parameters field: contains the next-hop MAC address associated to the unnumbered IRB interface. This MAC address identifies the NVE/DGW and can be re-used for all the VRFs in the node.

Example of prefix advertisement for the ipv4 prefix SN1/24 advertised from NVE1:

(1) NVE1 advertises the following BGP route for SN1:

- o Route type 5 (IP Prefix route) containing: Eth-Tag=VNI=10 (assuming global VNI), IPL=24, IP=SN1. In addition to that, a Remote-NH attribute will be sent, where: Tunnel-type= VXLAN or NVGRE and a Sub-TLV will contain a MAC address= NVE1 MAC.
- o As discussed, no MAC route is advertised for this core evpn.

(2) DGW1 imports the received route from NVE1 and SN1/24 is added to the designated routing context. The next-hop for SN1/24 will be given by the route type 5 BGP next-hop (NVE1), which is resolved to a

tunnel. For instance: if the tunnel is VXLAN based, the BGP next-hop will be resolved to a VXLAN tunnel where: destination-VTEP= NVE1 IP, VNI=10, inner destination MAC = NVE1 MAC (derived from the remote-nh attribute).

(3) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop= "NVE1 IP" is found. The tunnel information to encapsulate the packet will be derived from the route-type 5 received for SN1.
- o The IP packet destined to IPx is encapsulated with: Source inner MAC = DGW1 MAC, Destination inner MAC = NVE1 MAC, Source outer IP (source VTEP) = DGW1 IP, Destination outer IP (destination VTEP) = NVE1 IP

(4) When the packet arrives at NVE1:

- o Based on the tunnel information (VNI for the VXLAN case), the routing context is identified for an IP lookup.
- o An IP lookup is performed in the routing context, where SN1 turns out to be a local subnet associated to EVI-2. A subsequent lookup in the ARP table and the EVI-2 MAC FIB will return the forwarding information for the packet in EVI-2.

4. Conclusions

A new E-VPN route type 5 for the advertisement of IP Prefixes is proposed in this document. This new route type will have a differentiated role from the route type 2, i.e. MAC advertisement route, and will address all the inter-subnet connectivity scenarios which are required in the Data Center. As discussed throughout the document, IP-VPN cannot be used in an NVO-based DC to advertise IP Prefixes and the existing E-VPN route type 2 does not meet the requirements for all the DC use cases, therefore a new E-VPN route type is required.

This new E-VPN route type 5 decouples the IP Prefix advertisements from the MAC route advertisements in E-VPN, hence:

- a) Allows the clean and clear announcements of ipv4 or ipv6 prefixes in an NLRI with no MAC addresses in the route key, so that only IP information is used in BGP route comparisons.
- b) Since the route type is different from the MAC advertisement

route, the advertisement of prefixes will be excluded from all the procedures defined for the advertisement of VM MACs, e.g. MAC Mobility or aliasing. As a result of that, the current E-VPN procedures do not need to be modified.

c) Allows a flexible implementation where the prefix can be linked to different types of next-hops: MAC address, IP address, IRB IP address, ESI, etc. and these MAC or IP addresses do not need to reside in the advertising NVE.

d) An E-VPN implementation not requiring IP Prefixes can simply discard them by looking at the route type value.

5. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

6. Security Considerations

7. IANA Considerations

8. References

8.1. Normative References

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

8.2. Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03.txt, work in progress, February, 2013

[E-VPN-OVERLAYS] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using E-VPN", draft-sd-l2vpn-evpn-overlay-01.txt, work in progress, February, 2013

[BGP-REMOTE-NH] Van de Velde et al., "BGP Remote-Next-Hop", draft-vandeveld-idr-remote-next-hop-03.txt, work in progress, October, 2012

9. Acknowledgments

The authors would like to thank Mukul Katiyar and Senthil Sathappan for their valuable feedback and contributions.

10. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Florin Balus
Nuage Networks
Email: florin@nuagenetworks.net

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Senad Palislamovic
Alcatel-Lucent
Email: senad.palislamovic@alcatel-lucent.com

L2VPN Workgroup
Internet Draft

Intended status: Standards Track

J. Uttaro
AT&T

A. Isaac
T. Boyes
Bloomberg

J. Rabadan
S. Palislamovic
W. Henderickx
F. Balus
Alcatel-Lucent

K. Patel
A. Sajassi
Cisco

Expires: December 2013

June 26, 2013

Usage and applicability of BGP MPLS based Ethernet VPN
draft-rp-l2vpn-evpn-usage-00.txt

Abstract

This document discusses the usage and applicability of BGP MPLS based Ethernet VPN (E-VPN) in a simple and fairly common deployment scenario. The different E-VPN procedures will be explained on the example scenario, analyzing the benefits and trade-offs of each option. Along with [E-VPN], this document is intended to provide a simplified guide for the deployment of E-VPN in Service Provider 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), 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 December 28, 2013.

Copyright Notice

Copyright (c) 2013 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. Use-case scenario description	4
3. Provisioning Model	6
3.1. Common provisioning tasks	6
3.1.1. Non-service specific parameters	7
3.1.2. Service specific parameters	7
3.2. Service interface dependent provisioning tasks	8
3.2.1. VLAN-based service interface EVI	8
3.2.2. VLAN-bundle service interface EVI	9
3.2.3. VLAN-aware bundling service interface EVI	9
4. BGP E-VPN NLRI usage	9
5. MAC-based forwarding model use-case	10
5.1. E-VPN Network Startup procedures	10
5.2. VLAN-based service procedures	11
5.2.1. Service startup procedures	11
5.2.2. Packet walkthrough	12
5.3. VLAN-bundle service procedures	15
5.3.1. Service startup procedures	15
5.3.2. Packet Walkthrough	16
5.4. VLAN-aware bundling service procedures	19
5.4.1. Service startup procedures	19
5.4.2. Packet Walkthrough	20
6. MPLS-based forwarding model use-case	24

6.1. Impact of MPLS-based forwarding on the E-VPN network startup	24
6.2. Impact of MPLS-based forwarding on the VLAN-based service procedures	24
6.3. Impact of MPLS-based forwarding on the VLAN-bundle service procedures	25
6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures	26
7. Comparison between MAC-based and MPLS-based forwarding models	27
8. Traffic flow optimization	28
8.1. Control Plane Procedures	28
8.1.1. MAC learning options	28
8.1.2. Proxy ARP	29
8.1.3. Unknown Unicast flooding suppression	29
8.1.4. Optimization of Inter-subnet forwarding	29
8.2. Packet Walkthrough Examples	30
8.2.1. Proxy-ARP example for CE2 to CE3 traffic	30
8.2.2. Flood suppression example for CE1 to CE3 traffic	31
8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic	32
9. Conventions used in this document	33
10. Security Considerations	33
11. IANA Considerations	33
12. References	34
12.1. Normative References	34
12.2. Informative References	34
13. Acknowledgments	34
14. Authors' Addresses	34

1. Introduction

This document complements [E-VPN] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in section 2.

After describing the topology of the use-case scenario and the characteristics of the service to be deployed, the following section will describe the provisioning model, comparing the E-VPN procedures with the provisioning tasks required for other VPN technologies, such as VPLS or IP-VPN.

Once the provisioning model is analyzed, the following sections will describe the control plane and data plane procedures for the traffic in the example scenario, for the two potential disposition/forwarding models: MAC-based and MPLS-based models. While both models can interoperate in the same network, each one has different trade-offs that are analyzed in this document.

Finally, E-VPN provides some potential traffic flow optimization tools that are also described in the context of the example scenario.

2. Use-case scenario description

The following figure depicts the scenario that will be referenced throughout the rest of the document.

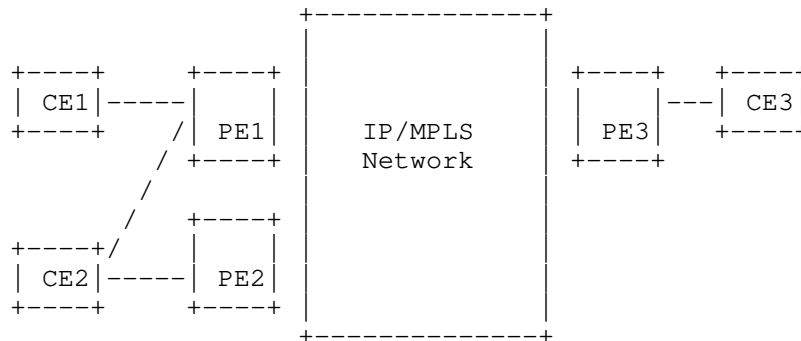


Figure 1 E-VPN use-case scenario

There are three PEs and three CEs considered in this example: PE1, PE2, PE3, as well as CE1, CE2 and CE3. Layer-2 traffic must be extended among the three CEs. The following service requirements are assumed in this scenario:

- o Redundancy requirements: CE1 and CE3 are single-homed to PE1 and

PE3 respectively. CE2 requires multi-homing connectivity to PE1 and PE2, not only for redundancy purposes, but also for adding more upstream/downstream connectivity bandwidth to/from the network. If CE2 has a single CE-VID (or a few CE-VIDs) the current VPLS multi-homing solutions (based on load-balancing per CE-VID or service) do not provide the optimized link utilization required in this example. Another redundancy requirement that must be met is fast convergence. E.g.: if the link between CE2 and PE1 goes down, a fast convergence mechanism must be supported so that PE3 can immediately send the traffic to PE2, irrespectively of the number of affected services and MAC addresses. E-VPN provides the flow-based load-balancing multi-homing solution required in this scenario to optimize the upstream/downstream link utilization between CE2 and PE1-PE2. E-VPN also provides a fast convergence solution so that PE3 can immediately send the traffic to PE2 upon failure on the link between CE2 and PE1.

- o Service interface requirements: service definition must be flexible in terms of CE-VID-to-broadcast-domain assignment and service contexts in the core. The following three services are required in this example:

EVI100 - It will use VLAN-based service interfaces in the three CEs with a 1:1 mapping (VLAN-to-EVI). The CE-VIDs at the three CEs can be the same, e.g.: VID 100, or different at each CE, e.g.: VID 101 in CE1, VID 102 in CE2 and VID 103 in CE3. A single broadcast domain needs to be created for EVI100 in any case; therefore CE-VIDs will require translation at the egress PEs if they are not consistent across the three CEs. The case when the same CE-VID is used across the three CEs for EVI100 is referred in [E-VPN] as the "Unique VLAN" E-VPN case. This term will be used throughout this document too.

EVI200 - It will use VLAN-bundle service interfaces in CE1, CE2 and CE3, based on an N:1 VLAN-to-EVI mapping. In this case, the service provider just needs to assign a pre-configured number of CE-VIDs on the ingress PE to EVI200, and send the customer frames with the original CE-VIDs. The Service Provider will build a single broadcast domain for the customer. The customer will be responsible for the CE-VID handling.

EVI300 - It will use VLAN-aware bundling service interfaces in CE1, CE2 and CE3. At the ingress PE, an N:1 VLAN-to-EVI mapping will be done, however and as opposed to EVI200, a separate core broadcast domain is required per CE-VID. In addition to that, the CE-VIDs can be different (hence CE-VID translation is required). Note that, while the requirements stated for EVI100 and EVI200 might be met with the current VPLS solutions, the VLAN-aware

bundling service interfaces required by EVI300 are not supported by the current VPLS tools.

- o BUM (Broadcast, Unknown unicast, Multicast) optimization requirements: The solution must be able to support ingress replication, P2MP MPLS LSPs and MP2MP MPLS LSPs and the user must be able to decide what kind of provider tree will be used by each EVI service. For example, if we assume that EVI100 and EVI200 will not carry much BUM traffic, we can use ingress replication for those service instances. The benefit is that the core will not need to maintain any states for the multicast trees associated to EVI100 and EVI200. On the contrary, if EVI300 is presumably carrying a significant amount of multicast traffic, P2MP MPLS LSPs or MP2MP LSPs can be used for this service. Note that ingress replication and P2MP LSPs are supported by VPLS solutions (see [VPLS-MCAST]), however VPLS solutions do not support MP2MP LSPs, since the source of the tree must be identified for the data plane MAC learning, and that identification is challenging when using MP2MP LSPs. Since E-VPN uses the control plane for MAC learning, any type of provider multicast tree is supported in the core.

As already outlined above, the current VPLS solutions, based on [RFC4761][RFC4762][RFC6074], cannot meet all the above set of requirements and therefore a new solution is needed. The following sections will describe how E-VPN can be used to meet those service requirements and even optimize the network further by:

- o Providing the user with an option to reduce (and even suppress) the ARP-flooding.
- o Supporting ARP termination for inter-subnet forwarding

3. Provisioning Model

One of the requirements stated in [E-VPN-REQ] is the ease of provisioning. BGP parameters and service context parameters should be auto-provisioned so that the addition of a new EVI to the E-VPN requires a minimum number of single-sided provisioning touches. However this is only possible in a limited number of cases. This section describes the provisioning tasks required for the services described in section 2, i.e. EVI100 (VLAN-based service interfaces), EVI200 (VLAN-bundle service interfaces) and EVI300 (VLAN-aware bundling service interfaces).

3.1. Common provisioning tasks

Regardless of the service interface type (VLAN-based, VLAN-bundle or VLAN-aware), the following sub-sections describe the parameters to be

provisioned in the three PEs.

3.1.1. Non-service specific parameters

The multi-homing function in E-VPN requires the provisioning of certain parameters which are not service-specific and that are shared by all the EVIs using the multi-homing capabilities. In our use-case, these parameters are only provisioned in PE1 and PE2, and are listed below:

- o Ethernet Segment Identifier (ESI): only the ESI associated to CE2 needs to be considered in our example. Single-homed CEs such as CE1 and CE3 do not require the provisioning of an ESI (the ESI will be coded as zero in the BGP NLRI). In our example, a LAG is used between CE2 and PE1-PE2 (since all-active multi-homing is a requirement) therefore the ESI can be auto-derived from the LACP information as described in [E-VPN]. Note that the ESI MUST be unique across all the PEs in the network, therefore the auto-provisioning of the ESI is only recommended in case the CEs are managed by the Service Provider. Otherwise the ESI should be manually provisioned in order to avoid potential conflicts.
- o ES-Import Route Target (ES-Import RT): this is the RT that will be sent by PE1 and PE2, along with the ES route. Regardless of how the ESI is provisioned in PE1 and PE2, the ES-Import RT must always be auto-derived from the 6-byte MAC address portion of the ESI value.
- o Ethernet Segment Route Distinguisher (ES RD): this is the RD to be encoded in the ES route and Ethernet Auto-Discovery (A-D) route to be sent by PE1 and PE2 for the CE2 ESI. This RD should always be auto-derived from the PE IP address, as described in [E-VPN].
- o Multi-homing type: the user must be able to provision the multi-homing type to be used in the network. In our use-case, the multi-homing type will be set to all-active for the CE2 ESI. This piece of information is encoded in the ESI Label extended community flags and sent by PE1 and PE2 along with the Ethernet A-D route for the CE2 ESI.

In our use-case, besides the above parameters, all the corresponding LAG and LACP parameters will be configured in PE1 and PE2, so that CE2 can send different flows to PE1 and PE2 for the same CE-VID as though they were forming a single system from the CE2 perspective.

3.1.2. Service specific parameters

The following parameters must be provisioned in PE1, PE2 and PE3 per

EVI service:

- o EVI identifier: global identifier per EVI that is shared by all the PEs part of the EVI, i.e. PE1, PE2 and PE3 will be provisioned with EVI100, 200 and 300. The EVI identifier can be associated to (or be the same value as) the EVI default Ethernet Tag (4-byte default broadcast domain identifier for the EVI). The Ethernet Tag is different from zero in the E-VPN BGP routes only if the service interface type (of the source PE) is VLAN-aware.
- o EVI Route Distinguisher (EVI RD): This RD is a unique value across all the EVIs in a PE. Auto-derivation of this RD might be possible depending on the service interface type being used in the EVI. Next section discusses the specifics of each service interface type.
- o EVI Route Target(s) (EVI RT): one or more RTs can be provisioned per EVI. The RT(s) imported and exported can be equal or different, just as the RT(s) in IP-VPNs. Auto-derivation of this RT(s) might be possible depending on the service interface type being used in the EVI. Next section discusses the specifics of each service interface type.
- o CE-VID and port/LAG binding to EVI identifier or Ethernet Tag: see the next section.

3.2. Service interface dependent provisioning tasks

Depending on the service interface type being used in the EVI, a specific CE-VID binding provisioning must be specified.

3.2.1. VLAN-based service interface EVI

In our use-case, EVI100 is a VLAN-based service interface EVI.

EVI100 can be a "unique-VLAN" E-VPN if the CE-VID being used for this service in CE1, CE2 and CE3 is equal, e.g. VID 100. In that case, the VID 100 binding must be provisioned in PE1, PE2 and PE3 for EVI100 and the associated port or LAG. The EVI RD and EVI RT can be auto-derived from the CE-VID:

- o The auto-derived EVI RD will be a Type 1 RD, as recommended in [E-VPN], and it will be comprised of [PE-IP]:[zero-padded-VID]; where PE-IP is the IP address of the PE (normally a loopback address) and [zero-padded-VID] is a 2-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 4 bits are zero.

- o The auto-derived EVI RT will be composed of [AS]:[zero-padded-VID]; where AS is the Autonomous System that the PE belongs to and [zero-padded-VID] is a 4-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 20 bits are zero. Note that auto-deriving the EVI RT implies supporting a basic any-to-any topology in the E-VPN and using the same import and export RT in the EVI.

If EVI100 is not a "unique-VLAN" E-VPN, each individual CE-VID must be configured in each PE, and EVI RDs and EVI RTs cannot be auto-derived, hence they must be provisioned by the user.

3.2.2. VLAN-bundle service interface EVI

Assuming EVI200 is a VLAN-bundle service interface EVI, and VIDs 200-250 are assigned to EVI200, the CE-VID bundle 200-250 must be provisioned on PE1, PE2 and PE3. Note that this model does not allow CE-VID translation and the CEs must use the same CE-VIDs for EVI200. No auto-derived EVI RDs or EVI RTs are possible.

3.2.3. VLAN-aware bundling service interface EVI

If EVI300 is a VLAN-aware bundling service interface EVI, CE-VID binding to EVI300 does not have to match on the three PEs (only on PE1 and PE2, since they are part of the same ES). E.g.: PE1 and PE2 CE-VID binding to EVI300 can be set to the range 300-310 and PE3 to 321-330. Note that each individual CE-VID will be assigned to a core broadcast domain, i.e. Ethernet Tag, which will be encoded in the BGP E-VPN routes.

Therefore, besides the CE-VID bundle range bound to EVI300 in each PE, associations between each individual CE-VID and the E-VPN Ethernet Tag must be provisioned by the user. No auto-derived EVI RDs/RTs are possible.

4. BGP E-VPN NLRI usage

[E-VPN] defines four different types of routes and four different extended communities advertised along with the different routes. However not all the PEs in a network must generate and process all the different routes and extended communities. The following table shows the routes that must be exported and imported in the use-case described in this document. "Export", in this context, means that the PE must be capable of generating and exporting a given route, assuming there are no BGP policies to prevent it. In the same way, "Import" means the PE must be capable of importing and processing a given route, assuming the right RTs and policies. "N/A" means neither import nor export actions are required.

BGP E-VPN routes	PE1-PE2	PE3
ES	Export/import	N/A
A-D per ESI	Export/import	Import
A-D per EVI	Export/import	Import
MAC	Export/import	Export/import
Inclusive mcast	Export/import	Export/import

PE3 is only required to export MAC and Inclusive multicast routes and be able to import and process A-D routes, as well as MAC and Inclusive multicast routes. If PE3 did not support importing and processing A-D routes per ESI and per EVI, fast convergence and aliasing functions (respectively) would not be possible in this use-case.

5. MAC-based forwarding model use-case

This section describes how the BGP E-VPN routes are exported and imported by the PEs in our use-case, as well as how traffic is forwarded assuming that PE1, PE2 and PE3 support a MAC-based forwarding model. In order to compare the control and data plane impact in the two forwarding models (MAC-based and MPLS-based) and different service types, we will assume that CE1, CE2 and CE3 need to exchange traffic for up to 4k CE-VIDs.

5.1. E-VPN Network Startup procedures

Before any EVI is provisioned in the network, the following procedures are required:

- o Infrastructure setup: the proper MPLS infrastructure must be setup among PE1, PE2 and PE3 so that the E-VPN services can make use of P2P, P2MP and/or MP2MP LSPs. In addition to the MPLS transport, PE1 and PE2 must be properly configured to create a multi-chassis LAG to CE2. Details are provided in [E-VPN]. Once the LAG is properly setup, as discussed in section 3.1, the ESI for the CE2 Ethernet Segment, e.g. ESI12, can be auto-generated by PE1 and PE2 from the LACP information exchanged with CE2. Alternatively, the ESI can also be manually provisioned on PE1 and PE2. PE1 and PE2 will auto-configure a BGP policy that will import any ES route matching the auto-derived ES-import RT for ESI12.
- o Ethernet Segment route exchange and DF election: PE1 and PE2 will advertise a BGP Ethernet Segment route for ESI12, where the ESI RD and ES-Import RT will be auto-generated as discussed in section 3.1.1. PE1 and PE2 will import the ES routes of each other and

will run the DF election algorithm for any existing EVI (if any, at this point). PE3 will simply discard the route. Note that the DF election algorithm can support service carving, so that the downstream BUM traffic from the network to CE2 can be load-balanced across PE1 and PE2 on a per-service basis.

At the end of this process, the network infrastructure is ready to start deploying E-VPN services. PE1 and PE2 are aware of the existence of a shared Ethernet Segment, i.e. ESI12.

5.2. VLAN-based service procedures

Assuming that the E-VPN network must carry traffic among CE1, CE2 and CE3 for up to 4k CE-VIDs, the Service Provider can decide to implement VLAN-based service interface EVIs to accomplish it. In this case, each CE-VID will be individually mapped to a different EVI. While this means a total number of 4k EVIs is required per PE, the advantages of this approach are the auto-provisioning of most of the service parameters if no VLAN translation is needed (see section 3.2.1) and great control over each individual customer broadcast domain. We assume in this section that the range of EVIs from 1 to 4k is provisioned in the network.

5.2.1. Service startup procedures

As soon as the EVIs are created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (up to 4k routes per PE) so that the flooding tree per EVI can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the EVIs have the same core topology, PMSI aggregation makes sense in order to save some multicast forwarding states in the core.
- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a list of 4k RTs (one per EVI) and an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). These routes will be imported by the three PEs, since the RTs match the EVI RTs locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as discussed in [E-VPN].
- o Ethernet A-D routes per EVI (4k routes): An A-D route per EVI will

be sent by PE1 and PE2 for ESI12. Each individual route includes the corresponding EVI RT and an MPLS label to be used by PE3 for the aliasing function. These routes will be imported by the three PEs.

5.2.2. Packet walkthrough

Once the services are setup, the traffic can start flowing. Assuming there are no MAC addresses learnt yet and that MAC learning at the access is performed in the data plane in our use-case, this is the process followed upon receiving packets from each CE (example for EVI1).

(1) BUM packet example from CE1:

- a) An ARP-request with CE-VID=1 is issued from source MAC CE1-MAC (MAC address coming from CE1 or from a device connected to CE1) to find the MAC address of CE3-IP.
- b) Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI1 context and if CE1-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port and CE-VID, (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI1 RD and RT, ESI=0, Ethernet-Tag=0 and CE1-MAC/CE1-IP along with an MPLS label assigned to EVI1 from the PE1 label space. Since we assume a MAC forwarding model, a label per EVI is normally allocated and signaled by the three PEs for MAC advertisement routes. Based on the RT, the route is imported by PE2 and PE3 and the forwarding state plus ARP entry are added to their EVI1 context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding for CE1-IP is not needed in the core.
- c) Since the ARP packet is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI1 (CE-VID=1 is kept if translation is required). Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication and no aggregation, the packet is replicated to PE2 and PE3 with the downstream allocated labels and the P2P LSP transport labels. No other labels are added to the stack.
- d) Assuming PE1 is the DF for EVI1 on ESI12, the packet is locally replicated to CE2.

- e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI1 on ESI12, and there is no other CE connected to PE2, the packet is discarded. At PE3, the packet is de-encapsulated, CE-VID translated if needed and replicated to CE3.

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

- a) An ARP-request with CE-VID=1 is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.
- b) CE2 will hash the packet and will forward it to e.g. PE2. Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI1 context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI and CE-VID, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI1 RD and RT, ESI=12, Ethernet-Tag=0 and CE2-MAC/CE2-IP along with an MPLS label assigned from the PE2 label space (one label per EVI). Note that, since PE3 is not part of ESI12, it will install a forwarding state for CE2-MAC as long as the A-D route per ESI for ESI12 is also active on PE3. On the contrary, PE1 is part of ESI12, therefore PE1 will not modify the forwarding state for CE2-MAC if it has previously learnt CE2-MAC locally attached to ESI12. Otherwise it will add forwarding state for CE2-MAC.
- c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI1 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI1, adding the ESI label for ESI12 at the bottom of the stack. The ESI label has been previously allocated and signaled by the A-D routes for ESI12. Note that if the result of the CE2 hashing had been different and the packet sent to PE1, PE1 would not have added the ESI label to the label stack (PE1 is the DF for EVI1 on ESI12).
- d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI

identified by the ESI label. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3. If a P2MP LSP is used as Inclusive multicast tree for EVI1, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will simply be ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

- a) A unicast packet with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE1-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE1-IP, and has added CE3-MAC/CE3-IP to its ARP proxy table).
- b) Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the EVI1 context and this time, since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and the label stack associated to the MAC CE1-MAC is found (including the label associated to EVI1 in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1.
- c) At PE1, the packet is identified to be part of EVI1 (based on the bottom of the stack label) and a destination MAC lookup is performed in the EVI1 context. The labels are popped and the packet forwarded to CE1 with CE-VID=1. Unicast packets from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

(4) Unicast packet example from CE3 to CE2:

- a) A unicast packet with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE2-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).
- b) Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the EVI1 context and since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and PE3 finds CE2-MAC associated to PE2 on ESI12, an Ethernet Segment for which PE3 has two active A-D routes per ESI (from PE1 and PE2) and two active A-D routes for EVI1 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI1). Either way, the packet is encapsulated and sent to the remote PE.

- c) At PE2 (or PE1), the packet is identified to be part of EVI1 based on the bottom label, and a destination MAC lookup is performed. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per EVI, hence a MAC lookup for the EVI1 context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12, hence the packet is forwarded to ESI12.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

5.3. VLAN-bundle service procedures

Instead of using VLAN-based interfaces, the Service Provider can choose to implement VLAN-bundle interfaces to carry the traffic for the 4k CE-VIDs among CE1, CE2 and CE3. If that is the case, the 4k CE-VIDs can be mapped to the same EVI, e.g. EVI200, at each PE. The main advantage of this approach is the low control plane overhead (reduced number of routes and labels) and easiness of provisioning, at the expense of no control over the customer broadcast domains, i.e. a single inclusive multicast tree for all the CE-VIDs and no CE-VID translation in the Provider network.

5.3.1. Service startup procedures

As soon as the EVI200 is created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI (one route): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (hence only one route per PE) so that the flooding tree per EVI can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the CE-VIDs are part of the same EVI, a single tree is created for all of them.
- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI200), an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the EVI200 RT locally configured. The A-D routes per ESI will be used for fast

convergence and split-horizon functions, as described in [E-VPN].

- o Ethernet A-D routes per EVI (one route): An A-D route (EVI200) will be sent by PE1 and PE2 for ESI12. This route includes the EVI200 RT and an MPLS label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

5.3.2. Packet Walkthrough

The packet walkthrough for the VLAN-bundle case is similar to the one described for EVI1 in the VLAN-based case except for some differences. The main difference is the fact that no VLAN translation is allowed and the CE-VIDs are kept untouched from CE to CE.

(1) BUM packet example from CE1:

- a) An ARP-request tagged with any CE-VID is issued from source MAC CE1-MAC to find the MAC address of CE3-IP.
- b) The packet is identified to be forwarded in the EVI200 context as long as its CE-VID belongs to the VLAN-bundle defined in the PE1 port to CE1. This case is a special VLAN-bundle case, since the entire CE-VID range is defined in the ports, therefore any CE-VID would be part of EVI200. A source MAC lookup is done next, in the MAC FIB and ARP proxy table within the EVI200 context and if CE1-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port (CE-VID is not taken into account), (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI200 RD and RT, ESI=0, Ethernet-Tag=0 and CE1-MAC/CE1-IP along with an MPLS label assigned from the PE1 label space. Since we assume a MAC forwarding model, a label per EVI is normally allocated and signaled by the three PEs for MAC advertisement routes. Based on the RT, the route is imported by PE2 and PE3 and the forwarding state plus ARP entry are added to their EVI200 context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding for CE1-IP is not needed in the core.
- c) Since the ARP is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI200. Note that the ingress CE-VID MUST be kept at the imposition PE and the disposition PE. Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication, the packet is replicated to PE2 and PE3 with the downstream allocated labels (by PE2 and PE3 respectively) and the P2P LSP transport labels. No other labels

are added to the stack.

- d) Assuming PE1 is the DF for EVI200 on ESI12, the packet is locally replicated to CE2.
- e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI200 on ESI12 and there is no other CE connected, the packet is discarded. At PE3, the packet is de-encapsulated and replicated to CE3. The CE-VID remains untouched throughout the whole process.

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

- a) An ARP-request, tagged with any CE-VID, is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.
- b) CE2 will hash the packet and will forward it to e.g. PE2. The packet CE-VID is identified to be forwarded in the EVI200 context, since the CE-VID belongs to the defined VLAN-bundle on the port. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI200 context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI200 RD and RT, ESI=12, Ethernet-Tag=0 and CE2-MAC/CE2-IP along with an MPLS label assigned from the PE2 label space (one label per EVI). Note that since PE3 is not part of ESI12, it will install a forwarding state for CE2-MAC as long as the A-D route per ESI for ESI12 is also active on PE3. On the contrary, PE1 is part of ESI12, therefore PE1 will not modify the forwarding state for CE2-MAC if it has previously learnt CE2-MAC locally attached to ESI12. Otherwise it will add a forwarding state for CE2-MAC.
- c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI200 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI200, adding the ESI label for ESI12 at the bottom of the stack. The ESI label has been previously allocated and signaled by the A-D routes for ESI12. Note that if the result of the CE2 hashing had been different and the packet sent to PE1, PE1 would not have added the ESI label to the label stack (PE1 is the DF for EVI200 on ESI12).

- d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI identified by the ESI label. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3. If a P2MP LSP is used as Inclusive multicast tree for EVI200, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will simply be ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

- a) A unicast packet, tagged with any CE-VID is issued from source MAC CE3-MAC and destination MAC CE1-MAC (PE3 has previously resolved an ARP request from CE3 to find the MAC of CE1-IP, and has added CE3-MAC/CE3-IP to its ARP proxy table).
- b) The packet is identified to be forwarded in the EVI200 context, since the CE-VID belongs to the defined VLAN-bundle on the port. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI200 context and, this time, since we assume CE3-MAC and CE3-IP are known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and the label stack associated to the MAC CE1-MAC is found (this includes the label associated to EVI200 in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1. The CE-VID is kept.
- c) At PE1, the packet is identified to be part of EVI200 (based on the bottom label) and a destination MAC lookup is performed in the EVI200 context. The labels are popped and the packet forwarded to CE1. The CE-VID remains untouched throughout the whole process.

Unicast packets from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

(4) Unicast packet example from CE3 to CE2:

- a) A unicast packet, tagged with any CE-VID, is issued from source MAC CE3-MAC and destination MAC CE2-MAC (PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).
- b) The packet is identified to be forwarded in the EVI200 context, since the CE-VID belongs to the defined VLAN-bundle on the ingress port. A source MAC lookup is done in the MAC FIB within the EVI200 context and since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination

MAC lookup is performed next and PE3 finds CE2-MAC associated to PE2 on ESI12, an Ethernet Segment for which PE3 has two active A-D routes per ESI (from PE1 and PE2) and two active A-D routes for EVI200 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI200). Either way, the packet is encapsulated and sent to the remote PE.

- c) At PE2 (or PE1), the packet is identified to be part of EVI200 based on the bottom label, and a destination MAC lookup is performed at the MAC FIB. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per EVI, hence a MAC lookup for the EVI200 context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12, hence the packet is forwarded to ESI12.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

5.4. VLAN-aware bundling service procedures

The last potential service type analyzed in this document is VLAN-aware bundling. When these types of service interfaces are used to carry the 4k CE-VIDs among CE1, CE2 and CE3, all the CE-VIDs will be mapped to the same EVI, e.g. EVI300. The difference, compared to the VLAN-bundle service type in the previous section, is that each incoming CE-VID will also be mapped to a different "normalized" Ethernet-Tag in addition to EVI300. If no translation is required, the Ethernet-tag will match the CE-VID. Otherwise a translation between CE-VID and Ethernet-tag will be needed at the imposition PE and at the disposition PE. The main advantage of this approach is the ability to control customer broadcast domains while providing a single EVI to the customer.

5.4.1. Service startup procedures

As soon as the EVI300 is created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI per Ethernet-Tag (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI and per Ethernet-Tag (hence 4k routes per PE) so that the flooding tree per customer broadcast domain can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created.

In the described use-case, since all the CE-VIDs and Ethernet-Tags are defined on the three PEs, multicast tree aggregation might make sense in order to save forwarding states.

- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI300), an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the EVI300 RT locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [E-VPN].
- o Ethernet A-D routes per EVI (one route): An A-D route (EVI300) will be sent by PE1 and PE2 for ESI12. This route includes the EVI300 RT and an MPLS label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

5.4.2. Packet Walkthrough

The packet walkthrough for the VLAN-aware case is similar to the ones described before. Compared to the other two cases, VLAN-aware services allow for CE-VID translation and for an N:1 CE-VID to EVI mapping. Note that this model requires qualified learning on the MAC FIBs.

(1) BUM packet example from CE1:

- a) An ARP-request tagged with CE-VID=x is issued from source MAC CE1-MAC to find the MAC address of CE3-IP.
- b) The packet is identified to be forwarded in the EVI300 context as long as its CE-VID belongs to the range defined in the PE1 port to CE1. In addition to it, CE-VID=x is mapped to Ethernet-Tag=y at the EVI300 (where x and y might be equal if no translation is needed). A source MAC lookup is done next, in the MAC FIB and ARP proxy table within the EVI300/Ethernet-Tag=y context and if CE1-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port and Ethernet-Tag, (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI300 RD and RT, ESI=0, Ethernet-Tag=y and CE1-MAC/CE1-IP along with an MPLS label assigned from the PE1 label space. Since we assume a MAC forwarding model, a label per EVI is normally allocated and signaled by the three PEs for MAC advertisement routes. Based on

the RT, the route is imported by PE2 and PE3 and the forwarding state plus ARP entry are added to their EVI300/Ethernet-Tag=y context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding is not needed in the core.

- c) Since the ARP is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI300/Ethernet-Tag=y. Note that the ingress CE-VID=x MUST be translated to the Ethernet-Tag=y at the imposition PE, assuming x and y are not equal. Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication, the packet is replicated to PE2 and PE3 with the downstream allocated labels (by PE2 and PE3 respectively) and the P2P LSP transport labels. No other labels are added to the stack.
- d) Assuming PE1 is the DF for EVI300 on ESI12, the packet is locally replicated to CE2. Note that the Ethernet-Tag MUST be translated to the egress CE-VID (if they are different).
- e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI300 on ESI12 and there are no other CEs connected, the packet is discarded. At PE3, the packet is de-encapsulated and replicated to CE3. The Ethernet-Tag in the packet is translated to the egress CE-VID (if different).

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

- a) An ARP-request, tagged with CE-VID=x, is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.
- b) CE2 will hash the packet and will forward the packet to e.g. PE2. The packet CE-VID=x is identified to be forwarded in the EVI300/Ethernet-Tag=y context, since the CE-VID belongs to the defined range on the port/Ethernet-Tag. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI300/Ethernet-Tag=y context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI and Ethernet-Tag, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI300 RD and RT, ESI=12, Ethernet-Tag=y and CE2-MAC/CE2-IP along with an MPLS label assigned from the PE2 label space (one label per EVI). Note that since PE3 is not part

of ESI12, it will install a forwarding state for CE2-MAC in the EVI300/Ethernet-Tag=y context as long as the A-D route per ESI for ESI12 is also active on PE3. On the contrary, PE1 is part of ESI12, therefore PE1 will not modify the forwarding state for CE2-MAC if it has previously learnt CE2-MAC locally attached to ESI12. Otherwise it will add a forwarding state for CE2-MAC.

- c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI300 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI300/Ethernet-Tag=y, adding the ESI label for ESI12 at the bottom of the stack. The ESI label has been previously allocated and signaled by the A-D routes for ESI12. Note that if the result of the CE2 hashing had been different and the packet sent to PE1, PE1 would not have added the ESI label to the label stack (PE1 is the DF for EVI300 on ESI12).
- d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI identified by the ESI label. The Ethernet-Tag will be translated, if needed, to the egress CE-VID. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3 after translating the Ethernet-Tag to the egress CE-VID. If a P2MP LSP is used as Inclusive multicast tree for EVI300/Ethernet-Tag=y, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will be simply ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

- a) A unicast packet, tagged with CE-VID=x is issued from source MAC CE3-MAC and destination MAC CE1-MAC (PE3 has previously resolved an ARP request from CE3 to find the MAC of CE1-IP, and has added CE3-MAC/CE3-IP to its ARP proxy table).
- b) The packet is identified to be forwarded in the EVI300/Ethernet-Tag=y context, since the CE-VID belongs to the defined range on the port/Ethernet-Tag. A source MAC lookup is done in the MAC FIB within the EVI300/Ethernet-Tag=y context and, this time, since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and the label stack associated to the MAC CE1-MAC is found (this includes the label associated to EVI300/Ethernet-Tag=y in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1. The CE-VID=x is translated

to the Ethernet-Tag=y value.

- c) At PE1, the packet is identified to be part of EVI300 (based on the bottom of the stack label) and a destination MAC lookup is performed in the EVI300/Ethernet-Tag=y context. The labels are popped and the packet forwarded to CE1 after translating the Ethernet-Tag value to the egress CE-VID.

Unicast packets from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

(4) Unicast packet example from CE3 to CE2:

- a) A unicast packet, tagged with CE-VID=x, is issued from source MAC CE3-MAC and destination MAC CE2-MAC (PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).
- b) The packet is identified to be forwarded in the EVI300/Ethernet-Tag=y context, since the CE-VID belongs to the defined range on the ingress port/Ethernet-Tag. A source MAC lookup is done in the MAC FIB table within the EVI300/Ethernet-Tag=y context and since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and PE3 finds CE2-MAC associated to PE2 on ESI12/Ethernet-Tag=y, an Ethernet Segment for which PE3 has two active A-D routes per ESI (from PE1 and PE2) and two active A-D routes for EVI300 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI300). Either way, the packet is encapsulated, CE-VID translated to Ethernet-Tag and sent to the remote PE.
- c) At PE2 (or PE1), the packet is identified to be part of EVI300/Ethernet-Tag=y based on the bottom label and the packet Ethernet-Tag, and a destination MAC lookup is performed at the MAC FIB. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per EVI and the Ethernet-Tag=y, hence a MAC lookup for the EVI300/Ethernet-Tag=y context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12 and the packet Ethernet-Tag the pointer at the egress CE-VID, hence the packet is forwarded to ESI12, with a translated tag from the Ethernet-Tag=y to the egress CE-VID=x.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

6. MPLS-based forwarding model use-case

E-VPN supports an alternative forwarding model, usually referred to as MPLS-based forwarding or disposition model as opposed to the MAC-based forwarding or disposition model described in section 5. Using MPLS-based forwarding model instead of the MAC-based one might have an impact on:

- o The number of forwarding states required
- o The FIB where the forwarding states are handled: MAC FIB or MPLS LFIB.

The MPLS-based forwarding model avoids the destination MAC lookup at the egress PE MAC FIB, at the expense of increasing the number of next-hop forwarding states at the egress MPLS LFIB. This also has an impact on the control plane and the label allocation model, since an MPLS-based disposition PE MUST send as many routes and labels as required next-hops in the egress EVI. This concept is equivalent to the forwarding models supported in IP-VPNs at the egress PE, where an IP lookup in the IP-VPN FIB might be necessary or not depending on the available next-hop forwarding states in the LFIB.

The following sub-sections highlight the impact on the control and data plane procedures described in section 5 when and MPLS-based forwarding model is used.

Note that both forwarding models are compatible and interoperable in the same network. The implementation of either model in each PE is a decision local to the PE node.

6.1. Impact of MPLS-based forwarding on the E-VPN network startup

The MPLS-based forwarding model has no impact on the procedures explained in section 5.1.

6.2. Impact of MPLS-based forwarding on the VLAN-based service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes, when all the service interfaces are VLAN-based. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact

compared to the MAC-based model.

- o Ethernet A-D routes per EVI (4k routes per PE/ESI): no impact compared to the MAC-based model.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same EVI, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS label=x can be re-used for both MAC advertisements since they both share the same source ESI12. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment.
- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

6.3. Impact of MPLS-based forwarding on the VLAN-bundle service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes when all the service interfaces are VLAN-bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (one route): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (one route per PE/ESI): no impact compared to the MAC-based model since no VLAN translation is required.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same EVI, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be

advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS label=x can be re-used for both MAC advertisements since they both share the same source ESI12. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment.

- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has definitively an impact in terms of number of A-D routes when all the service interfaces are VLAN-aware bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (4k routes per PE/ESI): PE1 and PE2 will send 4k routes for EVI300, one per <ESI, Ethernet-Tag ID> tuple. This will allow the egress PE to find out all the forwarding information in the MPLS LFIB and even support Ethernet-Tag to CE-VID translation at the egress. The MAC-based forwarding model would allow the PEs to send a single route per PE/ESI for EVI300, since the packet with the embedded Ethernet-Tag would be used to perform a MAC lookup and find out the egress CE-VID.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same EVI, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS

label=x can be re-used for both MAC advertisements since they both share the same source ESI12. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment. Note that, in this model, the Ethernet-Tag will be set to a non-zero value for the MAC-advertisement routes. The same MAC address can be announced with different Ethernet-Tag value. This will make the advertising PE install two different forwarding states in the MPLS LFIB.

- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

7. Comparison between MAC-based and MPLS-based forwarding models

Both forwarding models are possible in a network deployment and each one has its own trade-offs.

The MAC-based forwarding model can save A-D routes per EVI when VLAN-aware bundling services are deployed and therefore reduce the control plane overhead. A MAC FIB lookup at the egress PE is required in order to do so.

The MPLS-based forwarding model can save forwarding states at the egress PEs if labels per next hop CE (as opposed to per MAC) are implemented. No egress MAC lookup is required. An A-D route per <EVI, Ethernet-Tag> is required for VLAN-aware services, as opposed to an A-D route per EVI.

The following table summarizes the implementation details of both models for the VLAN-aware bundling service type.

4k CE-VID VLANs	MAC-based Model	MPLS-based Model
A-D routes/EVI	1 per ESI/EVI	4k per ESI/EVI
Egress PE Forwarding states	1 per MAC	1 per next-hop
Egress PE Lookups	2 (MPLS+MAC)	1 (MPLS)

The egress forwarding model is an implementation local to the egress PE and is independent of the model supported on the rest of the PEs, i.e. in our use-case, PE1, PE2 and PE3 could have either egress

forwarding model without any dependencies.

8. Traffic flow optimization

In addition to the procedures described across sections 1 through 7, E-VPN [E-VPN] procedures allow for optimized traffic handling in order to minimize unnecessary flooding across the entire infrastructure. Optimization is provided through specific ARP termination and the ability to block unknown unicast flooding. Additionally, E-VPN procedures allow for intelligent, closest to the source, inter-subnet forwarding and solves the commonly known sub-optimal routing problem. Besides the traffic efficiency, ingress based inter-subnet forwarding also optimizes packet forwarding rules and implementation at the egress nodes as well. Details of these procedures are outlined in the following sections.

8.1. Control Plane Procedures

8.1.1. MAC learning options

The fundamental premise of [E-VPN] is the notion of a different approach to MAC address learning compared to traditional IEEE 802.1 bridge learning methods; specifically E-VPN differentiates between data and control plane driven learning mechanisms.

Data driven learning implies that there is no separate communication channel used to advertise and propagate MAC addresses. Rather, MAC addresses are learned through IEEE defined bridge-learning procedures as well as by snooping on DHCP and ARP requests. As different MAC addresses show up on different ports, the L2 FIB is populated with the appropriate MAC addresses.

Control plane driven learning implies that there is a communication channel could be either a control-plane protocol or a management-plane mechanism. In the context of E-VPN, two different learning procedures are defined, i.e. local and remote procedures:

- o Local learning defines the procedures used for learning the MAC addresses of network elements locally connected to EVI. Local learning could be implemented through all three learning procedures: control plane, management plane as well as data plane. However, the expectation is that for most of the use cases, local learning through data plane should be sufficient.
- o Remote learning defines the procedures used for learning MAC addresses of network elements remotely connected to EVI, i.e. far-end PEs. Remote learning procedures defined in [E-VPN] advocate using only control plane learning; specifically BGP. Through the

use of BGP E-VPN NLRIs, the remote PE has the capability of advertising all the MAC addresses present in its local FIB.

8.1.2. Proxy ARP

In E-VPN, MAC addresses are advertised via the MAC Advertisement Route, as discussed in [E-VPN]. Optionally an IP address can be advertised along with the MAC address announcement. However, there are certain rules put in place in terms of IP address usage: if the MAC Advertisement Route contains an IP address, and the IP Address Length is 32 bits (or 128 in the IPv6 case), this particular IP address correlates directly with the advertised MAC address. Such advertisement allows us to build a Proxy ARP table populated with the IP<>MAC bindings received from all the remote nodes.

Furthermore, based on these bindings, a local EVI can now provide Proxy-ARP functionality for all ARP requests directed to the IP address pool learned through BGP. Therefore, the amount of unnecessary L2 flooding, ARP requests in this case, can be further reduced by the introduction of Proxy-ARP functionality across all E-VPN EVIs.

8.1.3. Unknown Unicast flooding suppression

Given that all locally learned MAC addresses are advertised through BGP to all remote PEs, suppressing flooding of any Unknown Unicast traffic towards the remote PEs is a feasible network optimization.

The assumption in the use case is made that any network device that appears on the remote EVI network will somehow signal its presence to the network. This signaling can be either done through gratuitous events. Once the remote PE acknowledges the presence of the node in the EVI, it will do two things: install its MAC address in its local FIB and advertise this MAC address to all other BGP speakers via E-VPN NLRI. Therefore, we can assume that any active MAC address is propagated and learnt through the entire E-VPN domain. Given that MAC addresses become pre-populated - once nodes are alive on the network - there is no need to flood any unknown unicast towards the remote PEs. If the owner of a given destination MAC is active, the BGP route will be present in the local RIB and FIB, assuming that the BGP import policies are successfully applied; otherwise, the owner of such destination MAC is not present on the network.

It is worth noting that unless control or management plane learning is used in all the PEs for a given EVI, unknown unicast flooding MUST be enabled.

8.1.4. Optimization of Inter-subnet forwarding

In a scenario in which both L2 and L3 services are needed over the same physical topology, some interaction between E-VPN and IP-VPN is required. A common way of stitching the two service planes is through the use of an IRB interface, which allows for traffic to be either routed or bridged depending on its destination MAC address. If the destination MAC address is the one of the IRB interface, traffic needs to be passed through a routing module and potentially be either routed to a remote PE or forwarded to a local subnet. If the destination MAC address is not the one of the IRB, the EVI follows standard bridging procedures.

A typical example of E-VPN inter-subnet forwarding would be a scenario in which multiple IP subnets are part of a single or multiple EVIs, and they all belong to a single IP-VPN. In such topologies, it is desired that inter-subnet traffic can be efficiently routed without any tromboning effects in the network. Due to the overlapping physical and service topology in such scenarios, all inter-subnet connectivity will be locally routed through the IRB interface.

In addition to optimizing the traffic patterns in the network, local inter-subnet forwarding also optimizes greatly the amount of processing needed to cross the subnets: standard VPLS to IP-VPN stitching through IRB interfaces forces the traffic to pass through IRB interfaces twice, once locally, as the traffic gets into the routing domain for a given IP VPN, and once remotely as the traffic exits the routing domain and enters the remote VPLS instance at the egress PE.

Through E-VPN MAC advertisements, the local PE learns the real destination MAC address associated with the remote IP address and the inter-subnet forwarding can happen locally. When the packet is received at the egress PE, it is directly mapped to an egress EVI, bypassing any egress IP-VPN processing.

8.2. Packet Walkthrough Examples

Assuming that the services are setup according to figure 1 in section 2, the following flow optimization processes will take place in terms of creating, receiving and forwarding packets across the network.

8.2.1. Proxy-ARP example for CE2 to CE3 traffic

Using figure 1 in section 2, consider EVI 400 residing on PE1, PE2 and PE3 connecting CE2 and CE3 networks. Also, consider that PE1 and PE2 are part of the all-active multi-homing ES for CE2, and that PE2 is elected designated-forwarder for EVI400. We assume that all the PEs implement the Proxy-ARP functionality in the EVI 400 context.

In this scenario, PE3 will not only advertise the MAC addresses through the E-VPN MAC Advertisement Route but also IP addresses of individual hosts, i.e. /32 prefixes, behind CE3. Upon receiving the E-VPN routes, PE1 and PE2 will install the MAC addresses in the EVI 400 FIB and based on the associated received IP addresses, PE1 and PE2 can now build a Proxy-ARP table within the context of EVI 400.

From the forwarding perspective, when a node behind CE2 sends a packet destined to a node behind CE3, it will first send an ARP request to e.g. PE2 (based on the result of the CE2 hashing). Assuming that PE2 has populated its Proxy-ARP table for all active nodes behind the CE3, and that the IP address in the ARP message matches the entry in the table, PE2 will respond to the ARP request with the actual MAC address on behalf of the node behind CE3.

Once the nodes behind CE2 learn the actual MAC address of the nodes behind CE3, all the MAC-to-MAC communications between the two networks will be unicast.

8.2.2. Flood suppression example for CE1 to CE3 traffic

Using figure 1 in section 2, consider EVI 500 residing on PE1 and PE3 connecting CE1 and CE3 networks. Consider that both PE1 and PE3 have disabled unknown unicast flooding for this specific EVI context. Once the network devices behind CE3 come online they will learn their MAC addresses and create local FIB entries for these devices. Note that local FIB entries could also be created through either a control or management plane between PE and CE as well. Consequently, PE3 will automatically create E-VPN Type 2 MAC Advertisement Routes and advertise all locally learned MAC addresses. The routes will also include the MPLS label associated with the corresponding egress EVI or egress next-hop, depending on the forwarding model scheme being used by PE3.

Given that PE1 automatically learns and installs all MAC addresses behind CE3, its EVI FIB will already be pre-populated with the respective next-hops and label assignments associated with the MAC addresses behind CE3. As such, as soon as the traffic sent by CE1 to nodes behind CE3 is received into the context of EVI 500, PE1 will push the MPLS Label(s) onto the original Ethernet frame and send the packet to the MPLS network. As usual, once PE3 receives this packet, and depending on the forwarding model, PE3 will either do a next-hop lookup in the EVI 500 context, or will just forward the traffic directly to the CE3. In the case that PE1 EVI 500 does not have a MAC entry for a specific destination that CE1 is trying to reach, PE1 will drop the packet since unknown unicast flooding is disabled.

Based on the assumption that all the MAC entries behind the CEs are

pre-populated through gratuitous-ARP and/or DHCP requests, if one specific MAC entry is not present in the EVI 500 FIB on PE1, the owner of that MAC is not alive on the network behind the CE3, hence the traffic can be dropped at PE1 instead of be flooded and consume network bandwidth.

8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic

Using figure 1 in section 2 consider that there is an IP-VPN 666 context residing on PE1, PE2 and PE3 which connects CE1, CE2 and CE3 into a single IP-VPN domain. Also consider that there are two EVIs present on the PEs, EVI 600 and EVI 60. Each IP subnet is associated to a different E-VPN context. Thus there is a single subnet, subnet 600, between CE1 and CE3 that is established through EVI 600. Similarly, there is another subnet, subnet 60, between CE2 and CE3 that is established through EVI 60. Since both subnets are part of the same IP VPN, there is a mapping of each EVI (or individual subnet) to a local IRB interface on the three PEs.

If a node behind CE2 wants to communicate with a node on the same subnet seating behind CE3, the communication flow will follow the standard E-VPN procedures, i.e. FIB lookup within the PE1 (or PE2) after adding the corresponding E-VPN label to the MPLS label stack (downstream label allocation from PE3 for EVI 60).

When it comes to crossing the subnet boundaries, the ingress PE implements local inter-subnet forwarding. For example, when a node behind CE2 (EVI 60) sends a packet to a node behind CE1 (EVI 600) the destination IP address will be in the subnet 600, but the destination MAC address will be the address of source node's default gateway, which in this case will be an IRB interface on PE1 (connecting EVI 60 to IP-VPN 666). Once PE1 sees the traffic destined to its own MAC address, it will route the packet to EVI 600, i.e. it will change the source MAC address to the one of the IRB interface in EVI 600 and change the destination MAC address to the address belonging to the node behind CE1, which is already populated in the EVI 600 FIB, either through data or control plane learning.

An important optimization to be noted is the local inter-subnet forwarding in lieu of IP VPN routing. If the node from subnet 60 (behind CE2) is sending a packet to the remote end node on subnet 600 (behind CE3), the mechanism in place still honors the local inter-subnet (inter-EVI) forwarding. In a typical IP-VPN-to-VPLS scenario, once the packet leaves the L2 domain on PE1, it would be routed through the IP-VPN procedures and consequently, through a remote PE3 IRB interface, routed back into the remote VPLS domain for further processing. However, in the E-VPN case, traffic locally routed and

forwarded to the egress PE within the E-VPN EVI context.

In our use-case, therefore, when node from subnet 60 behind CE2 sends traffic to the node on subnet 600 behind CE3, the destination MAC address is the PE1 EVI 60 IRB MAC address. However, once the traffic locally crosses EVIs, to EVI 600, via the IRB interface on PE1, the source MAC address is changed to that of the IRB interface and the destination MAC address is changed to the one advertised by PE3 via E-VPN and already installed in EVI 600. The rest of the forwarding through PE1 is using the EVI 600 forwarding context and label space.

Another very relevant optimization is due to the fact that traffic between PEs is forwarded through E-VPN, rather than through IP-VPN. In the example described above for traffic from EVI 60 on CE2 to EVI 600 on CE3, there is no need for IP-VPN processing on the egress PE3. Traffic is forwarded either to the EVI 600 context in PE3 for further MAC lookup and next-hop processing, or directly to the node behind CE3, depending on the egress forwarding model being used.

9. Conventions used in this document

In the examples, the following conventions are used:

- o CE-VIDs refer to the VLAN tag identifiers being used at CE1, CE2 and CE3 to tag customer traffic sent to the Service Provider E-VPN network
- o CE1-MAC, CE2-MAC and CE3-MAC refer to source MAC addresses "behind" each CE respectively. Those MAC addresses can belong to the CEs themselves or to devices connected to the CEs.
- o CE1-IP, CE2-IP and CE3-IP refer to IP addresses associated to the above MAC addresses.

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

10. Security Considerations

11. IANA Considerations

12. References

12.1. Normative References

[RFC4761]Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762]Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6074]Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

[RFC4364]Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

12.2. Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03.txt, work in progress, February, 2013

[EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-02.txt

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-13.txt

13. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

14. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Senad Palislamovic
Alcatel-Lucent
Email: senad.palislamovic@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.be

Florin Balus
Alcatel-Lucent
Email: Florin.Balus@alcatel-lucent.com

Keyur Patel
Cisco
Email: keyupate@cisco.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

James Uttaro
AT&T
Email: uttaro@att.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Truman Boyes
Bloomberg
Email: tboyes@bloomberg.net

L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Cisco

Wim Henderickx
Alcatel-Lucent

Yakov Rekhter
John Drake
Juniper

Florin Balus
Nuage Networks

Lucy Yong
Linda Dunbar
Huawei

Expires: January 15, 2014

July 15, 2013

IP Inter-Subnet Forwarding in EVPN
draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-02

Abstract

EVPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among hosts/VMs over an MPLS/IP network. However, there are scenarios in which inter-subnet forwarding among hosts/VMs across different IP subnets is required, while maintaining the multi-homing capabilities of EVPN. This document describes an IRB solution based on EVPN to address such requirements.

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) 2013 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	Traditional Inter-Subnet Forwarding	4
1.2.	Scenarios of EVPN NVEs as L3GW	4
2	Inter-Subnet Forwarding Scenarios	5
2.1	Switching among EVIs within a DC	6
2.2	Switching among EVIs in different DCs without route aggregation	7
2.3	Switching among EVIs in different DCs with route aggregation	7
2.4	Switching among IP-VPN sites and EVIs with route aggregation	7
3	Default L3 Gateway Addressing	8
3.1	Homogeneous Environment	8
3.1	Heterogeneous Environment	9
4	Operational Models for Inter-Subnet Forwarding	9
4.1	Among EVPN NVEs within a DC	9
4.2	Among EVPN NVEs in Different DCs Without Route Aggregation	10
4.3	Among EVPN NVEs in Different DCs with Route Aggregation	12
4.4	Among IP-VPN Sites and EVPN NVEs with Route Aggregation	13
4.5	Use of Centralized Gateway	14
5	VM Mobility	14
5.1	VM Mobility & Optimum Forwarding for VM's Outbound Traffic	14
5.2	VM Mobility & Optimum Forwarding for VM's Inbound Traffic	15
5.2.1	Mobility without Route Aggregation	15
5.2.2	Mobility with Route Aggregation	15

6	Acknowledgements	15
7	Security Considerations	15
8	IANA Considerations	15
9	References	16
9.1	Normative References	16
9.2	Informative References	16
	Authors' Addresses	16

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].

IRB: Integrated Routing and Bridging

IRB Interface: A virtual interface that connects the bridging module and the routing module on an NVE.

NVE: Network Virtualization Endpoint

1 Introduction

EVPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among hosts/VMs over an MPLS/IP network. However, there are scenarios where, in addition to intra-subnet forwarding, inter-subnet forwarding is required among hosts/VMs across different IP subnets at the EVPN PE nodes, also known as EVPN NVE nodes throughout this document, while maintaining the multi-homing capabilities of EVPN. This document describes an IRB solution based on EVPN to address such requirements.

1.1 Traditional Inter-Subnet Forwarding

The inter-subnet communication is traditionally achieved at the L3 Gateway nodes where all the inter-subnet communication policies are enforced. Even for different subnets belonging to one IP-VPN or tenant, traffic may need to go through FW or IPS between the trusted and un-trusted zones.

Some operators may prefer centralized approach, i.e. only have a set of default L3 gateways (whose redundancy is typically achieved by VRRP) for all inter-subnet traffic to go through. Usually there are FW, IPS, or other network appliances directly attached to the centralized L3 Gateway nodes. The centralized approach makes it easier for maintaining consistent policies and less prone to configuration errors. However, such centralized approach suffers from a major drawback of requiring all traffic to be hair-pinned to the L3GW nodes.

Some operators may prefer fully distributed L3 gateway design, e.g. allowing all NVEs to have the policies to route traffic across subnets. Under this design, all traffic between hosts attached to one NVE can be routed locally, thus avoiding traffic hair-pinning issue at the centralized L3GW. The perceived drawback of this fully distributed approach may be the extra effort required in maintaining policy consistence across all the NVEs.

Some operators may prefer somewhere in the middle, i.e. allowing NVEs to route traffic across only selected subnets. For example, allow NVEs to route traffic among subnets belonging to one tenant or one security zone.

1.2. Scenarios of EVPN NVEs as L3GW

When an EVPN NVE node is not the L3GW for the subnets attached, the EVPN NVE performs only L2 switching function for the traffic initiated from or destined to the hosts attached to the NVE.

Some EVPN NVEs can be the default L3GWs for some subnets. In this situation, the EVPN NVEs can route traffic across the subnets for which they are default L3GWs.

When there are multiple subnets attached to an EVPN NVE, some of the subnets could have the EVPN NVE as their L3GW, some other subnets don't have the NVE as their L3GW. For example: "Subnet-X" can communicate with "Subnet-Y" via NVE "A", but "Subnet-X" can't communicate with "Subnet-Z" via NVE "A". So when the "Subnet-X" needs to communicate with "Subnet-Z", the traffic might need to be routed through another device (e.g. FW, IPS, or another L3GW node).

1. When the EVPN NVE is the L3GW for "Subnet -X", hosts within "Subnet-X" will have the NVE's IRB MAC address as their default GW MAC address when they send data frames towards targets in different subnets.
2. When the EVPN NVE is not the L3GW for "Subnet-Y", hosts within "Subnet-Y", (even though still attached to the NVE), will use their own designated L3GW MAC address (that is different from the NVE's IRB address) in data frames destined towards targets in different subnets.

2 Inter-Subnet Forwarding Scenarios

The inter-subnet forwarding scenarios performed by an EVPN NVE can be divided into the following five categories. The last scenario, along with their corresponding solutions, are described in [EVPN-IPVPN-INTEROP]. The solutions for the first four scenarios are the focus of this document.

1. Switching among EVPN instances within a DC
2. Switching among EVPN instances in different DCs without route aggregation
3. Switching among EVPN instances in different DCs with route aggregation
4. Switching among IP-VPN sites and EVPN instances with route aggregation
5. Switching among IP-VPN sites and EVPN instances without route aggregation

In the above scenario, the term "route aggregation" refers to the

case where for a given EVI/VRF a node situated at the WAN edge of the data center network behaves as a default gateway for all the destinations that are outside the data center. The absence of route aggregation refers to the scenario where a given EVI/VRF within a data center has (host) routes to individual VMs that are outside of the data center.

In the case (4) the WAN edge node also performs route aggregation for all the destinations within its own data center, and acts as an interworking unit between EVPN and IP VPN (it implements both EVPN and IP VPN functionality).

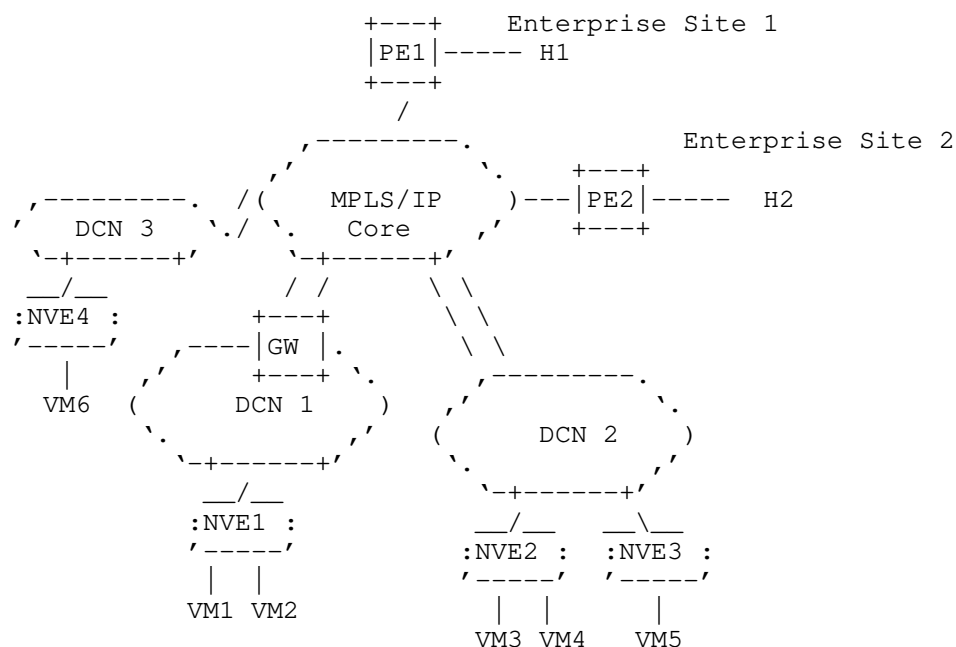


Figure 2: Interoperability Use-Cases

In what follows, we will describe scenarios 3 through 6 in more detail.

2.1 Switching among EVIs within a DC

In this scenario, connectivity is required between hosts (e.g. VMs) in the same data center, where those hosts belong to different IP subnets. All these subnets are part of the same IP VPN. Each subnet is associated with a single EVPN, where each such EVPN is realized by a collection of EVIs residing on appropriate NVEs.

As an example, consider VM3 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the IP3 subnet whereas VM5 belongs to the IP5 subnet. Both IP3 and IP5 subnets are part of the same IP VPN. NVE2 has an EVI3 associated with IP3 subnet and NVE3 has an EVI5 associated with the IP5 subnet.

2.2 Switching among EVIs in different DCs without route aggregation

This case is similar to that of section 2.1 above albeit for the fact that the hosts belong to different data centers that are interconnected over a WAN (e.g. MPLS/IP PSN). The data centers in question here are seamlessly interconnected to the WAN, i.e., the WAN edge does not maintain any host/VM-specific addresses in the forwarding path.

As an example, consider VM3 and VM6 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the IP3 subnet whereas VM6 belongs to the IP6 subnet. NVE2 has an EVI3 associated with IP3 subnet and NVE4 has an EVI6 associated with the IP6 subnet. Both IP3 and IP6 subnets are part of the same IP VPN and both EVI3 and EVI6 are associated with their VRFs for that IP VPN.

2.3 Switching among EVIs in different DCs with route aggregation

In this scenario, connectivity is required between hosts (e.g. VMs) in different data centers, and those hosts belong to different IP subnets. What makes this case different from that of Section 2.2 is that (in the context of a given EVI/VRF) at least one of the data centers in question has a gateway as the WAN edge switch. Because of that, the EVIs/VRFs within each data center need not maintain (host) routes to individual VMs outside of the data center.

As an example, consider VM1 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM1 belongs to the IP1 subnet whereas VM5 belongs to the IP5 subnet thus IP1 and IP5 subnets belong to the same IP VPN. NVE3 has an EVI5 associated with the IP5 subnet and NVE1 has an EVI1 associated with the IP1 subnet. Both EVI1 and EVI5 have associated with their VRFs that belong to the IP VPN that includes IP1 and IP5 subnets. Due to the gateway at the edge of DCN 1, NVE1 does not have the address of VM5 in its VRF table but instead it has a default route in its VRF with the next-hop being the GW.

2.4 Switching among IP-VPN sites and EVIs with route aggregation

In this scenario (within a context of a particular EVPN instance), connectivity is required between hosts (e.g. VMs) in a data center and hosts in an enterprise site that belongs to a given IP-VPN. The NVE within the data center is an EVPN NVE, whereas the enterprise site has an IP-VPN PE. Furthermore, the data center in question has a gateway as the WAN edge switch. Because of that, the NVE in the data center does not need to maintain individual IP prefixes advertised by enterprise sites (by IP-VPN PEs).

As an example, consider end-station H1 and VM2 of Figure 2. Assume that connectivity is required between the end-station and the VM, where VM2 belongs to the IP2 subnet that is realized using EVPN, whereas H1 belongs to an IP VPN site connected to PE1 (PE1 maintains an IP VPN VRF associated with that IP VPN). NVE1 has an EVI2 associated with the IP2 subnet. Moreover, NVE1 maintains a VRF associated with EVI2. PE1 originates a VPN-IP route that covers H1. The gateway at the edge of DCN1 performs interworking function between IP-VPN and EVPN. As a result of this, a default route in the VRF associated with EVI2, pointing to the gateway as the next hop, and a route to the VM2 (or maybe IP2 subnet) on the H1's VRF on PE1 are sufficient for the connectivity between H1 and VM2.

3 Default L3 Gateway Addressing

3.1 Homogeneous Environment

This is an environment where all NVEs to which an EVPN instance could potentially be attached (or moved), perform inter-subnet switching. Therefore, inter-subnet traffic can be locally switched by the EVPN NVE connecting the VMs belonging to different subnets.

To support such inter-subnet forwarding, the NVE behaves as an IP Default Gateway from the perspective of the attached end-stations (e.g. VMs). Two models are possible, as discussed in [DC-MOBILITY]:

1. All the EVIs of a given EVPN instance use the same anycast default gateway IP address and the same anycast default gateway MAC address. On each NVE, this default gateway IP/MAC address correspond to the IRB interface of the EVI associated with that EVPN instance.
2. Each EVI of a given EVPN instance uses its own default gateway IP and MAC addresses, and these addresses are aliased to the same conceptual gateway through the use of the Default Gateway extended community as specified in [EVPN], which is carried in the EVPN MAC Advertisement routes. On each NVE, this default gateway IP/MAC address correspond to the IRB interface of the EVI associated with that EVPN instance.

Both of these models enable a packet forwarding paradigm where inter-subnet traffic can bypass the VRF processing on the egress (i.e. disposition) NVE. The egress NVE merely needs to perform a lookup in the associated EVI and forward the Ethernet frames unmodified, i.e. without rewriting the source MAC address. This is different from traditional IRB forwarding where a packet is forwarded through the bridge module followed by the routing module on the ingress NVE, and then forwarded through the routing module followed by the bridging module on the egress NVE. For inter-subnet forwarding using EVPN, the routing module on the egress NVE can be completely bypassed.

It is worth noting that if the applications that are running on the hosts (e.g. VMs) are employing or relying on any form of MAC security, then the first model (i.e. using anycast addresses) would be required to ensure that the applications receive traffic from the same source MAC address that they are sending to.

3.1 Heterogeneous Environment

For large data centers with thousands of servers and ToR (or Access) switches, some of them may not have the capability of maintaining or enforcing policies for inter-subnet switching. Even though policies among multiple subnets belonging to same tenant can be simpler, hosts belonging to one tenant can also send traffic to peers belonging to different tenants or security zones. A L3GW not only needs to enforce policies for communication among subnets belonging to a single tenant, but also it needs to know how to handle traffic destined towards peers in different tenants. Therefore, there can be a mixed environment where an NVE performs inter-subnet switching for some EVPN instances but not others.

4 Operational Models for Inter-Subnet Forwarding

4.1 Among EVPN NVEs within a DC

When an EVPN MAC advertisement route is received by the NVE, the IP address associated with the route is used to populate the VRF, whereas the MAC address associated with the route is used to populate both the bridge-domain MAC table, as well as the adjacency associated with the IP route in the VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup identifies both the next-hop (i.e. egress) NVE to which the

packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop NVE. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) NVE after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the EVI label that was advertised by the egress NVE. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVI label to identify the bridge-domain table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 2 below depicts the packet flow, where NVE1 and NVE2 are the ingress and egress NVEs, respectively.

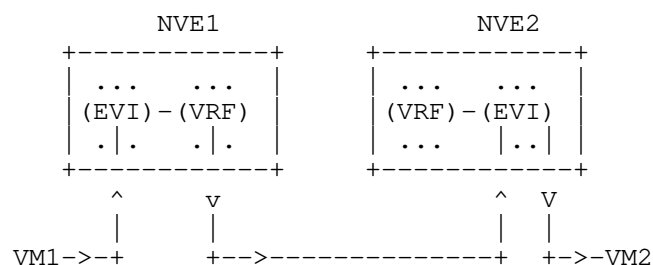


Figure 2: Inter-Subnet Forwarding Among EVPN NVEs within a DC

Note that the forwarding behavior on the egress NVE is similar to EVPN intra-subnet forwarding. In other words, all the packet processing associated with the inter-subnet forwarding semantics is confined to the ingress NVE.

It should also be noted that [EVPN] provides different level of granularity for the EVI label. Besides identifying bridge domain table, it can be used to identify the egress interface or a destination MAC address on that interface. If EVI label is used for egress interface or destination MAC address identification, then no MAC lookup is needed in the egress EVI and the packet can be directly forwarded to the egress interface just based on EVI label lookup.

4.2 Among EVPN NVEs in Different DCs Without Route Aggregation

When an EVPN MAC advertisement route is received by the NVE, the IP address associated with the route is used to populate the VRF,

whereas the MAC address associated with the route is used to populate both the bridge-domain MAC table, as well as the adjacency associated with the IP route in the VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup identifies both the next-hop (i.e. egress) Gateway to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop Gateway. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) Gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as an EVI label. The EVI label could be either advertised by the ingress Gateway, if inter-AS option B is used, or advertised by the egress NVE, if inter-AS option C is used. When the MPLS encapsulated packet is received by the ingress Gateway, the processing again differs depending on whether inter-AS option B or option C is employed: in the former case, the ingress Gateway swaps the EVI label in the packets with the EVI label value received from the egress Gateway. In the latter case, the ingress Gateway does not modify the EVI label and performs normal label switching on the LSP label. Similarly on the egress Gateway, for option B, the egress Gateway swaps the EVI label with the value advertised by the egress NVE. Whereas, for option C, the egress Gateway does not modify the EVI label, and performs normal label switching on the LSP label. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVI label to identify the bridge-domain table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 3 below depicts the packet flow.

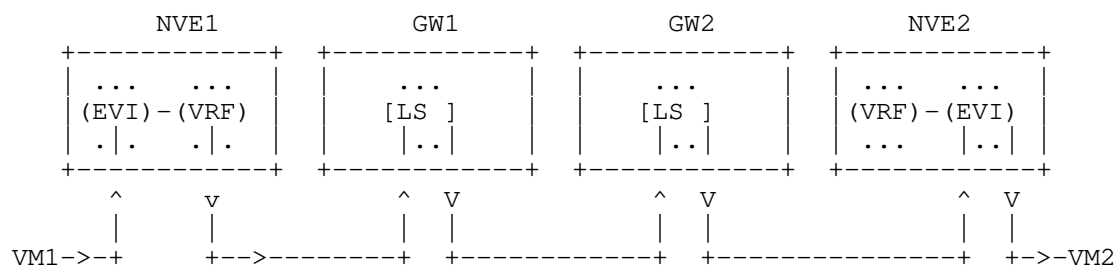


Figure 3: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs without Route Aggregation

4.3 Among EVPN NVEs in Different DCs with Route Aggregation

In this scenario, the NVEs within a given data center do not have entries for the MAC/IP addresses of hosts in remote data centers. Rather, the NVEs have a default IP route pointing to the WAN gateway for each VRF. This is accomplished by the WAN gateway advertising for a given EVPN that spans multiple DC a default VPN-IP route that is imported by the NVEs of that EVPN that are in the gateway's own DC.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup, in this case, matches the default route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the IRB Interface MAC address of the local WAN gateway. It also rewrites the source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to the WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the IP-VPN label that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the IP-VPN label to identify the VRF table. It then performs an IP lookup in that table. The lookup identifies both the remote WAN gateway (of the remote data center) to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the ultimate destination host (as populated by the EVPN MAC route). The local WAN gateway then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The local WAN gateway, then, forwards the frame to the remote WAN gateway after encapsulating it with the MPLS label stack. Note that

this label stack includes the LSP label as well as a VPN label that was advertised by the remote WAN gateway. When the MPLS encapsulated packet is received by the remote WAN gateway, it simply swaps the VPN label with the EVI label advertised by the egress NVE. This implies that the remote WAN gateway must allocate the VPN label at least at the granularity of a (VRF, egress NVE) tuple. The remote WAN gateway then forward the packet to the egress NVE. The egress NVE then performs a MAC lookup in the EVI (identified by the received EVI label) to determine the outbound port to send the traffic on.

Figure 4 below depicts the forwarding model.

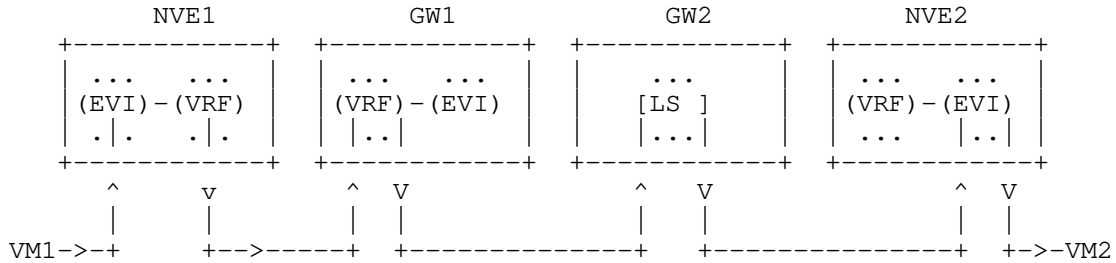


Figure 4: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs with Route Aggregation

4.4 Among IP-VPN Sites and EVPN NVEs with Route Aggregation

In this scenario, the NVEs within a given data center do not have entries for the IP addresses of hosts in remote enterprise sites. Rather, the NVEs have a default IP route pointing to the WAN gateway for each VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup, in this case, matches the default route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the IRB Interface MAC address of the local WAN gateway. It also rewrites the source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to the WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the IP-VPN label that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the

IP-VPN label to identify the VRF table. It then performs an IP lookup in that table. The lookup identifies the next hop ASBR to which the packet must be forwarded. The local gateway in this case strips the Ethernet encapsulation and forwards the IP packet to the ASBR using a label stack comprising of an LSP label and a VPN label that was advertised by the ASBR. When the MPLS encapsulated packet is received by the ASBR, it simply swaps the VPN label with the IP-VPN label advertised by the egress PE. This implies that the remote WAN gateway must allocate the VPN label at least at the granularity of a (VRF, egress PE) tuple. The ASBR then forwards the packet to the egress PE. The egress PE then performs an IP lookup in the VRF (identified by the received IP-VPN label) to determine where to forward the traffic.

Figure 5 below depicts the forwarding model.

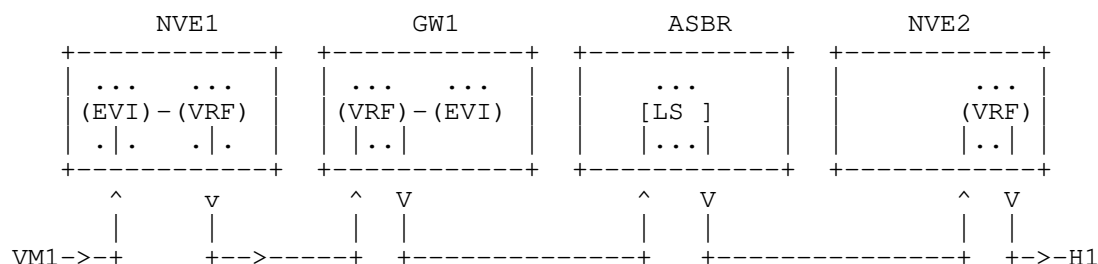


Figure 5: Inter-Subnet Forwarding Among IP-VPN Sites and EVPN NVEs with Route Aggregation

4.5 Use of Centralized Gateway

In this scenario, the NVEs within a given data center need to forward traffic in L2 to a centralized L3GW for a number of reasons: a) they don't have IRB capabilities or b) they don't have required policy for switching traffic between different tenants or security zones. The centralized L3GW performs both the IRB function for switching traffic among different EVPN instances as well as it performs interworking function when the traffic needs to be switched between IP-VPN sites and EVPN instances.

5 VM Mobility

5.1 VM Mobility & Optimum Forwarding for VM's Outbound Traffic

Optimum forwarding for the VM's outbound traffic, upon VM mobility, can be achieved using either the anycast default Gateway MAC and IP

addresses, or using the address aliasing as discussed in [DC-MOBILITY].

5.2 VM Mobility & Optimum Forwarding for VM's Inbound Traffic

For optimum forwarding of the VM's inbound traffic, upon VM mobility, all the NVEs and/or IP-VPN PEs need to know the up to date location of the VM. Two scenarios must be considered, as discussed next.

In what follows, we use the following terminology:

- source NVE refers to the NVE behind which the VM used to reside prior to the VM mobility event.
- target NVE refers to the new NVE behind which the VM has moved after the mobility event.

5.2.1 Mobility without Route Aggregation

In this scenario, when a target NVE detects that a MAC mobility event has occurred, it initiates the MAC mobility handshake in BGP as specified in [EVPN]. The WAN Gateways, acting as ASBRs in this case, re-advertise the MAC route of the target NVE with the MAC Mobility extended community attribute unmodified. Because the WAN Gateway for a given data center re-advertises BGP routes received from the WAN into the data center, the source NVE will receive the MAC Advertisement route of the target NVE (with the next hop attribute adjusted depending on which inter-AS option is employed). The source NVE will then withdraw its original MAC Advertisement route as a result of evaluating the Sequence Number field of the MAC Mobility extended community in the received MAC Advertisement route. This is per the procedures already defined in [EVPN].

5.2.2 Mobility with Route Aggregation

This section will be completed in the next revision.

6 Acknowledgements

The authors would like to thank Sami Boutros for his valuable comments.

7 Security Considerations

8 IANA Considerations

9 References

9.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, July, 2014.

[EVPN-IPVPN-INTEROP] Sajassi et al., "EVPN Seamless Interoperability with IP-VPN", draft-sajassi-l2vpn-evpn-ipvpn-interop-01, work in progress, October, 2012.

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-05.txt, work in progress, June, 2013.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

Yakov Rekhter
Juniper Networks
Email: yakov@juniper.net

John E. Drake
Juniper Networks
Email: jdrake@juniper.net

Lucy Yong
Huawei Technologies
Email: lucy.yong@huawei.com

Linda Dunbar
Huawei Technologies
Email: linda.dunbar@huawei.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Florin Balus
Alcatel-Lucent
Email: Florin.Balus@alcatel-lucent.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 14, 2014

L. Zheng
Z. Li
Huawei Technologies
July 13, 2013

A Framework for E-VPN Performance Monitoring
draft-zheng-l2vpn-evpn-pm-framework-00

Abstract

The capability of Ethernet VPN performance monitoring (PM) is important to meet the Service Level Agreement (SLA) for the service beared. Since multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, flow identifying is a big challenge for E-VPN PM. This document specifies the framework and mechanisms for the application of E-VPN PM.

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].

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 14, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Terminology	3
3. Overview and Concepts	4
3.1. EVI-to-EVI Tunnel	4
4. Control Plane	4
4.1. E-VPN Membership Auto-Discovery	4
4.2. EVI-to-EVI Tunnel Label Allocation	4
5. Data Plane	5
5.1. Additional Label for Ingress EVI Identification	5
5.2. Replace MAC Label with ET Label	6
6. E-VPN Performance Monitoring	6
7. IANA Considerations	7
8. Security Considerations	7
9. Acknowledgements	7
10. References	7
10.1. Normative References	7
10.2. Informative References	7
Authors' Addresses	8

1. Introduction

Virtual Private LAN Service (VPLS) is a proven and widely deployed Ethernet L2VPN solution. However, it has a number of limitations when it comes to redundancy, multicast optimization and provisioning simplicity. Also, new applications are driving several new requirements for other L2VPN services such as E-TREE, VPWS, and VPMS. Furthermore, data center interconnect applications are driving the need for new service interface types, the "VLAN-aware Bundling" service interfaces. Then the Ethernet VPN (E-VPN) solution (defined in [I-D.ietf-l2vpn-evpn]) has been proposed to meet these requirements which is documented in [I-D.ietf-l2vpn-evpn-req].

An E-VPN comprises PEs that form the edge of the MPLS infrastructure and CEs that are connected to PEs. The PEs provide virtual Layer 2 bridged connectivity between the CEs. In E-VPN, MAC learning between PEs occurs not in the data plane but in the control plane. PEs advertise the MAC addresses learned from the CEs, along with the associated MPLS label, to other PEs in the control plane by using MP-BGP.

The capability of E-VPN to measure and monitor performance metrics for packet loss, delay, as well as related metrics is essential for meeting the Service Level Agreement (SLA). This measurement capability also provides operators with greater visibility into the performance characteristics of the services in their networks, and provides diagnostic information in case of performance degradation or failure and helps for fault localization. To perform the measurement of packet loss, delay and other metrics on a particular E-VPN flow, the egress PE needs to determine which specific ingress EVI packets belongs to. There exists complete and mature performance monitoring mechanism for the traditional L2VPN based on the point-to-point PW. But in the case of E-VPN, multipoint-to-point (MP2P) or multipoint-to-multipoint (MP2MP) network model applies, it makes the flow identifying a big challenge for packets loss and delay measurement. This MP2P or MP2MP model also apply to L3VPN, please refer to [I-D.zheng-l3vpn-pm-analysis] for detailed description of the challenge for performance monitoring of such network model.

This document defines the framework for performance monitoring of E-VPN. The point-to-point connection named as EVI-to-EVI tunnel is introduced in E-VPN. And the corresponding process of control plane and data plane is defined.

2. Terminology

E-VPN: Ethernet VPN

EVI: Ethernet VPN Instance

ET: EVI-to-EVI Tunnel

MP2P: Multi-Point to Point

MP2MP: Multi-Point to Multi-Point

P2P: Point to Point

PM: Performance Monitoring

3. Overview and Concepts

Based on the mechanisms in [I-D.ietf-l2vpn-evpn], for a particular MAC address route, the directly connected PE allocates the same MPLS label to all the remote PEs which maintain the MAC routing and forwarding instance (EVI) of that E-VPN. Thus for the egress PE, it is unable to identify the source EVI of the received E-VPN packets.

To perform the packet loss or delay measurement on a specific E-VPN flow, it is critical to establish the Point-to-Point connection between the two EVIs. Once the Point-to-Point connection is built up, current measurement mechanisms for MPLS networks may be applied to E-VPN. A new concept "EVI-to-EVI Tunnel" is introduced in the following section to establish such Point-to-Point connection in E-VPN.

3.1. EVI-to-EVI Tunnel

In order to perform performance monitoring in E-VPN, a point-to-point connection between any two EVIs of a particular E-VPN needs to be established. This point-to-point connection enables the egress PE identifying the ingress EVI of the received E-VPN packet, thus enables the measurement of the packet loss and delay between the ingress and egress EVIs. Such point-to-point connection between an ingress EVI and an egress EVI is called "EVI-to-EVI Tunnel (ET)".

4. Control Plane

This section describes the control plane mechanisms for E-VPN performance monitoring.

4.1. E-VPN Membership Auto-Discovery

Before the Point-to-Point connections between EVIs could be established, each PE attaching a given E-VPN needs to learn all the remote PEs that attach to the same E-VPN. This could be achieved by the Ethernet A-D route per EVI defined in [I-D.ietf-l2vpn-evpn]. Please refer to section 9.4.1 [I-D.ietf-l2vpn-evpn] for details.

4.2. EVI-to-EVI Tunnel Label Allocation

After obtaining the E-VPN membership information, each PE needs to allocate MPLS labels to identify the EVI-to-EVI tunnel from the remote EVI to the local EVI. We call such labels as ET labels in this document. For each local EVI, the egress PE SHOULD allocate different ET labels for each remote EVI in PEs belonging to the same E-VPN. As such, the egress PE could identify the E-VPN flow received from different ingress EVIs, and the packet loss and delay

measurement could be performed between each ingress EVI and the local EVI.

5. Data Plane

This section introduces two new MPLS label stack encapsulations when ET label applies.

5.1. Additional Label for Ingress EVI Identification

When a E-VPN data packet is to be sent on the ingress PE, firstly the label advertised by the MP-BGP for the Mac address route is pushed onto the label stack. The ET label allocated by the egress EVI for the ingress EVI should then be pushed onto the label stack to identify the Point-to-Point connection between the sending and receiving EVI. Finally the MPLS tunnel label is pushed onto the label stack. The process of TTL and COS fields between the E-VPN label encapsulation and the tunnel label encapsulation is done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. The value of the TTL and COS field in the MAC label's encapsulation SHOULD be copied to the corresponding fields of the ET label's encapsulation. As such, one extra label is carried in the label stack compared with E-VPN data plane defined in [I-D.ietf-l2vpn-evpn].

When the E-VPN data packet received by the egress PE, the outermost tunnel label is popped, then the egress PE could use the ET label to identify the ingress EVI of the packet. The process of TTL and COS fields at the egress node should be done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. Since the value of the TTL and COS fields of the MAC label encapsulation and the ET label encapsulation are the same, the TTL and COS fields of the ET label encapsulation could be ignored during the course of the TTL and COS process at the egress node.

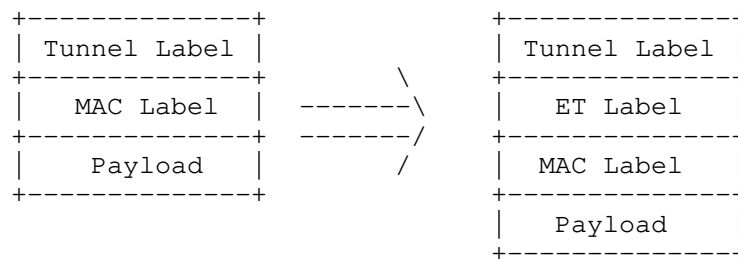


Fig.1 Additional Label for Ingress EVI Identification

5.2. Replace MAC Label with ET Label

Since the ET label identifies the connection between the ingress EVI and egress EVI, it could also be used to identify the egress EVI forwarding table in which the MAC prefix lookup should be performed. Thus when encapsulating the E-VPN data packets, the ingress PE could simply replace the MAC label with the ET label, then push the tunnel label. The process of TTL and COS fields between the MAC label encapsulation and the tunnel label encapsulation is done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. The TTL and COS value of the MAC label entry should be copied to the TTL and COS field of the ET label entry respectively. In this way the depth of the MPLS label stack is unchanged.

The encapsulation method would require the egress PE to perform MAC prefix lookup in the egress EVI forwarding table before the packet can be forwarded to a specific CE. The similar procedure is also required when per-instance EVI label allocation mechanism is used. The process of TTL and COS fields at the egress node should be done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. Since the MAC label encapsulation is replaced with the ET label encapsulation, the TTL and COS fields of the VT label encapsulation should be used as those of the MAC label encapsulation during the course of the TTL and COS process at the egress node.

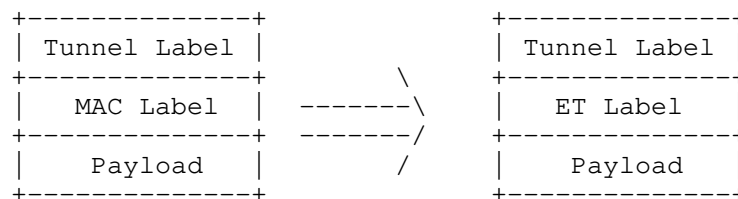


Fig.2 Replace the MAC Label with ET Label

6. E-VPN Performance Monitoring

[RFC6374] defines procedure and protocol mechanisms to enable the efficient and accurate measurement of packet loss, delay, as well as related metrics in MPLS networks. It provides either point-to-point or point-to-multipoint measurement capabilities. Once the point-to-point connection EVI-to-EVI Tunnel is established between the ingress and egress EVIs, the procedures for the packet loss and delay measurement as defined in [RFC6374] can be utilized for E-VPN performance monitoring. The main difference between performance monitoring of E-VPN and MPLS is the format of identifiers in the Loss Measurement (LM) and Delay Measurement (DM) messages. Specifically,

for E-VPN, the source and destination addresses of the LM and DM messages should be set to the concatenation of the Route Distinguisher (RD) of the particular EVI and the IP address of the ingress and egress PE respectively.

7. IANA Considerations

This document makes no request of IANA.

8. Security Considerations

TBD

9. Acknowledgements

TBD

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

10.2. Informative References

[I-D.ietf-l2vpn-evpn-req]
Sajassi, A., Aggarwal, R., Bitar, N., and A. Isaac,
"Requirements for Ethernet VPN (E-VPN)", draft-ietf-l2vpn-evpn-req-03 (work in progress), May 2013.

[I-D.ietf-l2vpn-evpn]
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F.,
Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN",
draft-ietf-l2vpn-evpn-03 (work in progress), February
2013.

[I-D.zheng-l3vpn-pm-analysis]
Zheng, L., Li, Z., and B. Parise, "Performance Monitoring
Analysis for L3VPN", draft-zheng-l3vpn-pm-analysis-01
(work in progress), April 2013.

[RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen,
P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-
Protocol Label Switching (MPLS) Support of Differentiated
Services", RFC 3270, May 2002.

- [RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, January 2003.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

Authors' Addresses

Lianshu Zheng
Huawei Technologies
Huawei Campus, No.156 Beiqing Rd.
Beijing 100095
China

Email: vero.zheng@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Campus, No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

L3 VPN Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 11, 2013

J. Zhang
Juniper Networks, Inc.
January 07, 2013

L2L3 VPN Multicast MIB
draft-zzhang-l2l3-vpn-mcast-mib-00

Abstract

This memo defines an experimental portion of the Management Information Base for use with network management protocols in the Internet community.

In particular, it describes managed objects common to both VPLS and VPN Multicast.

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 July 11, 2013.

Copyright Notice

Copyright (c) 2013 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	3
2. The Internet-Standard Management Framework	3
3. Conventions	3
4. Summary of MIB Module	3
5. Definitions	3
6. Security Considerations	8
7. IANA Considerations	8
8. References	8
8.1. Normative References	8
8.2. Informative References	9

1. Introduction

Multicast in VPLS and VPN can be achieved by using provider tunnels to deliver to all or a subset of PEs. The signaling of provider tunnel choice is very similar for both VPLS and VPN multicast (aka MVPN), and this memo describes managed objects common to both VPLS Multicast [I-D.ietf-l2vpn-vpls-mcast] and MVPN [RFC 6513/6514].

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC 3410 [RFC3410].

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIV2, which is described in STD 58, RFC 2578 [RFC2578], STD 58, RFC 2579 [RFC2579] and STD 58, RFC 2580 [RFC2580].

3. Conventions

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].

4. Summary of MIB Module

L2L3-VPN-MCAST-MIB contains a Textual Convention, L2L3VpnMcastProviderTunnelType, and a L2L3VpnMcastPmsiTunnelAttributeTable. Other MIB objects ([I-D.ietf-l2vpn-vpls-mcast], [I-D.ietf-l3vpn-mvpn-mib]) may point to entries in the L2L3VpnMcastPmsiTunnelAttributeTable.

5. Definitions

```
L2L3-VPN-MCAST-MIB DEFINITIONS ::= BEGIN
```

```
IMPORTS
```

```
    MODULE-IDENTITY, OBJECT-TYPE, NOTIFICATION-TYPE,  
    experimental, Unsigned32  
    FROM SNMPv2-SMI
```

```
    MODULE-COMPLIANCE, OBJECT-GROUP, NOTIFICATION-GROUP
```

```
FROM SNMPv2-CONF

TruthValue, RowPointer, RowStatus, TimeStamp, TimeInterval
FROM SNMPv2-TC

SnmAdminString
FROM SNMP-FRAMEWORK-MIB

InetAddress, InetAddressType
FROM INET-ADDRESS-MIB

MplsLabel
FROM MPLS-TC-STD-MIB

l2L3VpnMcastMIB MODULE-IDENTITY
    LAST-UPDATED "201301071200Z" -- 07 January 2013 12:00:00 GMT
    ORGANIZATION "IETF Layer-3 Virtual Private
        Networks Working Group."
    CONTACT-INFO

        "
        Comments and discussion to l3vpn@ietf.org
        Jeffrey (Zhaohui) Zhang
        Juniper Networks, Inc.
        10 Technology Park Drive
        Westford, MA 01886
        USA
        Email: zzhang@juniper.net
        "

    DESCRIPTION
        "This MIB contains common managed object definitions for
        multicast in Layer 2 and Layer 3 VPNs, defined by
        [I-D.ietf-l2vpn-vpls-mcast] and RFC 6513/6514.
        Copyright (C) The Internet Society (2013)."
```

-- Revision history.

```
REVISION "201301071200Z" -- 07 January 2013 12:00:00 GMT
DESCRIPTION
    "Initial version of the draft."
 ::= { experimental 99 } -- number to be assigned

-- Textual convention

l2L3VpnMcastProviderTunnelType ::= TEXTUAL-CONVENTION
    SYNTAX          INTEGER { unconfigured (0),
                             pim-asm (1),
```

```

        pim-ssm (2),
        pim-bidir (3),
        rsvp-p2mp (4),
        ldp-p2mp (5),
        ingress-replication (6),
        ldp-mp2mp (7)
    }
    STATUS          current
    DESCRIPTION
        "Types of provider tunnels used for multicast in a l2/l3vpn."
    REFERENCE
        "[RFC6514]"

-- Top level components of this MIB.
-- tables, scalars

l2L3VpnMcastObjects OBJECT IDENTIFIER ::= { l2L3VpnMcastMIB 1 }
l2L3VpnMcastStates  OBJECT IDENTIFIER ::= { l2L3VpnMcastObjects 1 }

-- Table of PMSI attributes

l2L3VpnMcastPmsiTunnelAttributeTable OBJECT-TYPE
    SYNTAX          SEQUENCE OF L2L3VpnMcastPmsiTunnelAttributeEntry
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "This table is for advertised/received PMSI attributes,
         to be referred to by I-PMSI or S-PMSI table entries"
        ::= {l2L3VpnMcastStates 1 }

l2L3VpnMcastPmsiTunnelAttributeEntry OBJECT-TYPE
    SYNTAX          L2L3VpnMcastPmsiTunnelAttributeEntry
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "An entry in this table corresponds to an PMSI attribute
         that is advertised/received on this router.
         For BGP-based signaling (for I-PMSI via auto-discovery
         procedure, or for S-PMSI via S-PMSI A-D routes),
         they are just as signaled by BGP (RFC 6514 section 5,
         'PMSI Tunnel attribute').
         For UDP-based S-PMSI signaling for PIM-MVPN,
         they're derived from S-PMSI Join Message
         (RFC 6513 section 7.4.2, 'UDP-based Protocol')..

         Note that BGP-based signaling may be used for
         PIM-MVPN as well."
    INDEX {

```



```

        12L3VpnMcastPmsiTunnelAttributeFlags,
        12L3VpnMcastPmsiTunnelAttributeType,
        12L3VpnMcastPmsiTunnelAttributeLabel,
        12L3VpnMcastPmsiTunnelAttributeId
    }
 ::= { 12L3VpnMcastPmsiTunnelAttributeTable 1 }

L2L3VpnMcastPmsiTunnelAttributeEntry ::= SEQUENCE {
    12L3VpnMcastPmsiTunnelAttributeFlags    OCTET STRING,
    12L3VpnMcastPmsiTunnelAttributeType      Unsigned32,
    12L3VpnMcastPmsiTunnelAttributeLabel     MplsLabel,
    12L3VpnMcastPmsiTunnelAttributeId        OCTET STRING,
    12L3VpnMcastPmsiTunnelPointer            RowPointer,
    12L3VpnMcastPmsiTunnelIf                 RowPointer
}

12L3VpnMcastPmsiTunnelAttributeFlags OBJECT-TYPE
    SYNTAX      OCTET STRING (SIZE (1))
    MAX-ACCESS   not-accessible
    STATUS       current
    DESCRIPTION
        "For UDP-based S-PMSI signaling for PIM-MVPN, this is 0.
        For BGP-based I/S-PMSI signaling,
        per RFC 6514 section 5, 'PMSI Tunnel Attribute':

```

The Flags field has the following format:

```

    0 1 2 3 4 5 6 7
    +--+--+--+--+--+--+
    | reserved |L|
    +--+--+--+--+--+--+

```

This document defines the following flags:

```

    + Leaf Information Required (L)"
 ::= { 12L3VpnMcastPmsiTunnelAttributeEntry 1 }

12L3VpnMcastPmsiTunnelAttributeType OBJECT-TYPE
    SYNTAX      L2L3VpnMcastProviderTunnelType
    MAX-ACCESS   not-accessible
    STATUS       current
    DESCRIPTION
        "For BGP-based I/S-PMSI signaling for either PIM or BGP-MVPN,
        per RFC 6514 section 5, 'PMSI Tunnel Attribute':

```

The Tunnel Type identifies the type of the tunneling technology used to establish the PMSI tunnel. The type determines the syntax and semantics of the Tunnel Identifier field. This document defines the

following Tunnel Types:

- 0 - No tunnel information present
- 1 - RSVP-TE P2MP LSP
- 2 - mLDP P2MP LSP
- 3 - PIM-SSM Tree
- 4 - PIM-SM Tree
- 5 - PIM-Bidir Tree
- 6 - Ingress Replication
- 7 - mLDP MP2MP LSP

For UDP-based S-PMSI signaling for PIM-MVPN, RFC 6513 does not specify if a PIM provider tunnel is SSM, SM or Bidir, and an agent can use either type 3, 4, or 5 based on its best knowledge."

```
::= { l2L3VpnMcastPmsiTunnelAttributeEntry 2 }
```

l2L3VpnMcastPmsiTunnelAttributeLabel OBJECT-TYPE

```
SYNTAX          MplsLabel
MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
```

```
"For BGP-based I/S-PMSI signaling,
per RFC 6514 section 5, 'PMSI Tunnel Attribute':
```

If the MPLS Label field is non-zero, then it contains an MPLS label encoded as 3 octets, where the high-order 20 bits contain the label value. Absence of MPLS Label is indicated by setting the MPLS Label field to zero.

For UDP-based S-PMSI signaling for PIM-MVPN, this is not applicable for now, as RFC 6513 does not specify mpls encapsulation and tunnel aggregation with UDP-based signaling."

```
::= { l2L3VpnMcastPmsiTunnelAttributeEntry 3 }
```

l2L3VpnMcastPmsiTunnelAttributeId OBJECT-TYPE

```
SYNTAX          OCTET STRING ( SIZE (4|8|12) )
MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
```

```
"For BGP-based signaling, as defined in RFC 6514 section 5,
'PMSI Tunnel Attribute'.
```

For UDP-based S-PMSI signaling for PIM-MVPN, RFC 6513 only specifies the 'P-Group' address, and that is filled into the first four octets of this field."

```
::= { l2L3VpnMcastPmsiTunnelAttributeEntry 4 }
```

l2L3VpnMcastPmsiTunnelPointer OBJECT-TYPE

SYNTAX RowPointer

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"If the tunnel exists in some MIB table, this is the row pointer to it."

::= { l2L3VpnMcastPmsiTunnelAttributeEntry 5 }

l2L3VpnMcastPmsiTunnelIf OBJECT-TYPE

SYNTAX RowPointer

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"If the tunnel has a corresponding interface, this is the row pointer to the ifName table."

::= { l2L3VpnMcastPmsiTunnelAttributeEntry 6 }

END

6. Security Considerations

N/A

7. IANA Considerations

IANA is requested to root MIB objects in the MIB module contained in this document under the transmission subtree.

.

8. References

8.1. Normative References

- [RFC3418] Presuhn, R., "Management Information Base (MIB) for the Simple Network Management Protocol (SNMP)", STD 62, RFC 3418, December 2002.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2578] McCloghrie, K., Ed., Perkins, D., Ed.,

- and J. Schoenwaelder, Ed., "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.
- [RFC2579] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Textual Conventions for SMIv2", STD 58, RFC 2579, April 1999.
- [RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Conformance Statements for SMIv2", STD 58, RFC 2580, April 1999.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [I-D.ietf-l2vpn-vpls-mcast] Aggarwal, R., Rekhter, Y., Kamite, Y., and L. Fang, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-11 (work in progress), July 2012.
- [I-D.ietf-l2vpn-vpls-mib] Nadeau, T., Koushik, K., and R. Mediratta, "Virtual Private Lan Services (VPLS) Management Information Base", draft-ietf-l2vpn-vpls-mib-07 (work in progress), September 2012.

8.2. Informative References

- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.

Author's Address

Zhaohui Zhang
Juniper Networks, Inc.
10 Technology Park Drive
Westford, MA 01886
USA

EMail: zzhang@juniper.net