

L2VPN Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2013

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

July 2012

802.1aq and 802.1Qbp Support over EVPN
draft-allan-l2vpn-spbm-evpn-01

Abstract

This document describes how Ethernet Shortest Path Bridging - MAC mode (802.1aq) and (802.1Qbp) can be combined with EVPN in a way that interworks with PBB-MESs 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 3 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 2013.

Copyright and License Notice

Copyright (c) 2012 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.....	3
1.2. Requirements Language.....	3
2. Conventions used in this document.....	3
2.1. Terminology.....	3
3. Solution Overview.....	4
4. Elements of Procedure.....	5
4.1. MES Configuration.....	5
4.2. DF Election.....	6
4.3. Control plane interworking ISIS-SPB to EVPN.....	6
4.4. Control plane interworking EVPN to ISIS-SPB.....	7
4.5. Data plane Interworking 802.1aq SPBM island or PBB-MES to EVPN.....	7
4.6. Data plane Interworking EVPN to 802.1aq SPBM island.....	8
4.7. Data plane interworking EVPN to 802.1ah PBB-MES.....	8
4.8. Dataplane interworking between 802.1Qbp islands and EVPN.....	8
4.9. Multicast Stitching.....	8
5. Other Aspects.....	8
5.1. Flow Ordering.....	8
5.2. Loop Avoidance and Black Holing...Error! Bookmark not defined.	
5.3. Transit.....	8
6. Acknowledgements.....	8
7. Security Considerations.....	9
8. IANA Considerations.....	9
8.1. Normative References.....	9
8.2. Informative References.....	9
9. Authors' Addresses.....	10

1. Introduction

This document describes how Ethernet Shortest Path Bridging - MAC mode (802.1aq) and (802.1Qbp) along with PBB-MESs 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, multiple spanning trees, or ECMP.

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
MES: MPLS Edge Switch
MP2MP: Multipoint to Multipoint
MVPN: Multicast VPN
NLRI: Network layer reachability information

PBBN: Provider Backbone Bridged Network
 PBB-MES: Co located BEB and MES
 P2MP: Point to Multipoint
 P2P: Point to Point
 RD: Route Distinguisher
 SPB: Shortest path bridging
 SPBM: Shortest path bridging MAC mode

3. Solution Overview

The EVPN solution for 802.1aq SPBM incorporates control plane interworking in the MES to map ISIS-SPB [2] information elements into the EVPN NLRI information and vice versa. This requires each MES 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 MES such that an arbitrary number of EVPN subtending SPB islands may be interconnected without any topological or multipathing dependencies. This requires each MES connected to an SPBM island to act both as an EVPN BGP speaker and as an ISIS-SPB edge node. This model also permits PBB-MESs as defined in draft-l2vpn-pbb-evpn-02[6] to be seamlessly communicate with the SPB islands. The next version of this document will add support for 802.1Qbp permitting seamless interworking between 802.1ah, 802.1aq and 802.1Qbp as well as supporting subtending 802.1ad based PBNS.

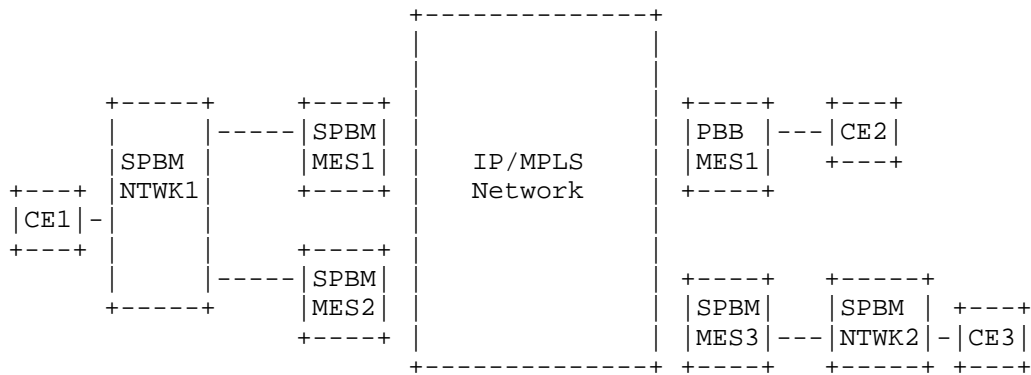


Figure 1: PBB and SPBM EVPN Network

Each EVPN is identified by a route target. The route target identifies the set of SPB islands and BEB-MESs that are allowed to communicate. 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 MESs/SPBM islands. This is in the form of B-MAC address/I-SID/Tx-Rx-attribute tuples. BGP filters leaking 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 island, PBB-MESs or PBBNs will not be leaked into the local ISIS-SPB routing system. Each SPBM island is administered to have an associated Ethernet Segment ID (ESI) associated with it.

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

An SPBM-MES 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.

4. Elements of Procedure

4.1. MES Configuration

At SPBM island commissioning a MES is configured with:

- 1) The route target for the service instance. Where a service instance is defined as the set of SPBM islands, PBBNs and PBB-MESs to be interconnected by the EVPN.
- 2) The unique ESI for the SPBM island. Mechanisms for deriving a common ESI for the SPBM island are for a future version of the document.

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
- 2) The set of VLANs (identified by B-VIDs Ethernet frames) used in the SPBM island and multipathing 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 RD for the node can be auto-derived. This will be described in a future version of the document.

4.2. DF Election

MESs self appoint in the role of DF for a B-VID for a given SPBM island. The procedure used is as per section 9.5.2 of draft-ietf-l2vpn-evpn-01[4] "DF election with service carving".

4.3. Control plane interworking ISIS-SPB to EVPN

When a MES 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 MES
- 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.

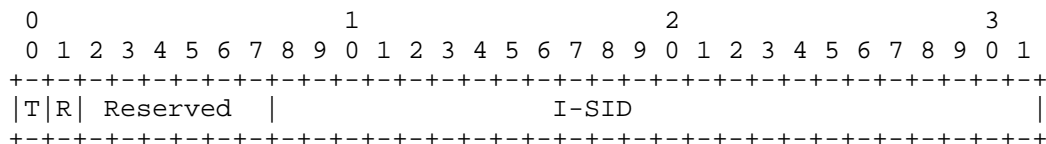


Figure 2: I-SID encoding in the Ethernet tag-ID field

- the MAC address from the sub-TLV

- an MPLS label

Similarly in the scenario where a MES 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 MES.

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 MES.

The ISIS-SPB information is also used to keep current a local table indexed by I-SID 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.

4.4. Control plane interworking EVPN to ISIS-SPB

When a MES receives a BGP NLRI that is new information, it checks if the I-SID in the Ethernet Tag ID locally maps to the B-VID it is an elected DF for. 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 MES 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 IS-IS.

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 to impose on the SPBM frame. The bottom label being that offered in the NLRI.

4.5. Data plane Interworking 802.1aq SPBM island or PBB-MES to EVPN

When an MES 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 MES strips the B-VID information from the frame, adds the label information to the frame and forwards the resulting MPLS packet.

4.6. Data plane Interworking EVPN to 802.1aq SPBM island

When a MES 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.

4.7. Data plane interworking EVPN to 802.1ah PBB-MES

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

A PBB-MES 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.

4.8. Dataplane interworking between 802.1Qbp islands and EVPN

For a future version of the document

4.9. Multicast Stitching

For a future version of the document

5. Other Aspects

5.1. Flow Ordering

When per I-SID multicast is implemented via MES 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.

5.2. Transit

Any MES 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.

6. Acknowledgements

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

7. Security Considerations

For a future version of this document.

8. IANA Considerations

For a future version of this document.

8.1. 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

8.2. Informative References

- [5] IEEE Standard for Local and Metropolitan Area Networks: Bridges and Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging
- [6] Draft IEEE Standard for Local and Metropolitan Area Networks---Virtual Bridged Local Area Networks - Amendment: Equal Cost Multiple Paths (ECMP), 802.1Qbp draft 1.0
- [7] Sajassi et.al. "PBB E-VPN", IETF work in progress, draft-ietf-l2vpn-pbb-evpn-03, June 2012
- [8] 802.1Q (2011) IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks

9. 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

INTERNET-DRAFT
Intended Status: Informational
Expires: January 1, 2013

Sami Boutros
Ali Sajassi
Samer Salam
June 30, 2012

VPWS support in E-VPN
draft-boutros-l2vpn-evpn-vpws-00.txt

Abstract

This document describes how E-VPN can be used to support virtual private wire service (VPWS) in MPLS/IP networks. E-VPN enables the following characteristics for VPWS: 1) active/standby redundancy, 2) active/active multi-homing with flow-based load-balancing, 3) eliminates the need for single-segment and multi-segment PW signaling, and 4) provides faster convergence using data-plane prefix independent convergence upon node or link failure in comparison to control-plane convergence with PW redundancy.

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) 2012 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.	BGP Extensions	4
3	Operation	4
4	E-VPN Comparison to PW Signaling	5
5	VPWS with multiple sites	5
6	Security Considerations	6
7	IANA Considerations	6
8	References	6
8.1	Normative References	6
8.2	Informative References	6
	Authors' Addresses	6

1 Introduction

This document describes how E-VPN can be used to support virtual private wire service (VPWS) in MPLS/IP networks. The use of E-VPN mechanisms for VPWS introduces all the benefits of E-VPN to p2p services. These benefits include active/standby AC redundancy, active/active multi-homing with flow-based load-balancing. Furthermore, the use of E-VPN for VPWS eliminates the need for signaling single-segment and multi-segment PWs for p2p Ethernet services.

[E-VPN] has the ability to forward customer traffic to/from a given customer Attachment Circuit (aka Ethernet AD route) without any MAC lookup. This capability is ideal in providing P2P services (aka VPWS services). [MEF] defines EVPL service as P2P service between a pair of ACs (designated by VLANs). EVPL can be considered as a VPWS with only two ACs. In delivering an EVPL service, traffic forwarding capability of E-VPN between a pair of Ethernet AD routes is used; whereas, for more general VPWS, traffic forwarding capability of E-VPN among a group of Ethernet AD routes (one Ether AD route per AC/site) is used. Since in VPWS services, the traffic from an originating Ether AD route can go only to a single destination Ether AD route, no MAC lookup is needed and MPLS label associated with the destination Ether AD route can be used in forwarding user traffic to the destination AC.

In current PW redundancy mechanisms, convergence time is a function of control plane convergence characteristics. However, with E-VPN it is possible to attain faster convergence through the use of data-plane prefix independent convergence upon node or link failure.

This document proposes the use of the Ethernet AD route to signal labels for P2P Ethernet services. As with E-VPN, the Ethernet Segment route can be used to synchronize LACP and other state between the PEs attached to the same multi-homed device.

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].

MAC: Media Access Control

MPLS: Multi Protocol Label Switching.

OAM: Operations, Administration and Maintenance.

PE: Provide Edge Node.

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

EVI: E-VPN Instance.

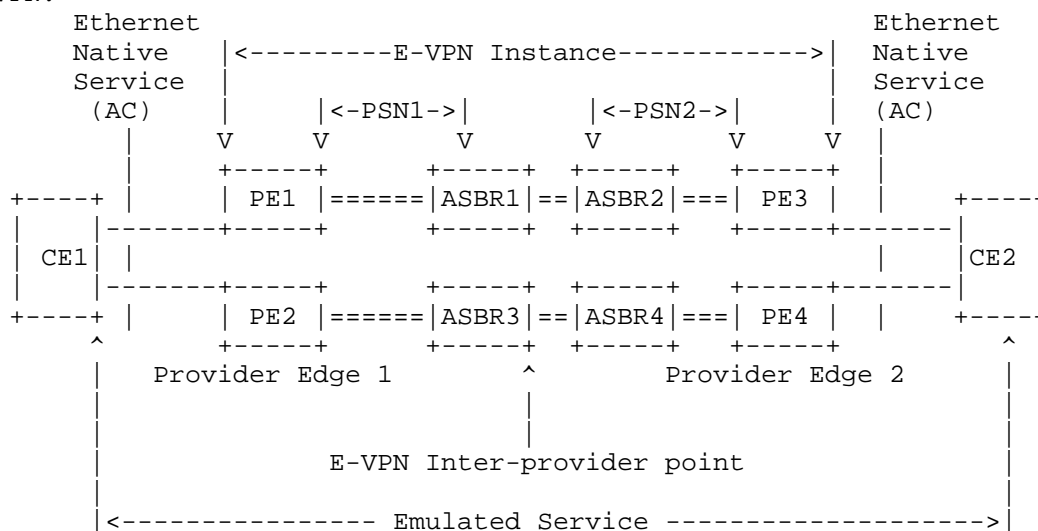
2. BGP Extensions

[E-VPN] defines a new BGP NLRI for advertising different route types for E-VPN operation. This document does not define any new BGP messages, but rather repurposes one of the routes as described next.

This document proposes the use of the Ethernet AD route to signal P2P services. The Ethernet Segment Identifier field is set to the ESI of the attachment circuit of the VPWS service instance. The Ethernet Tag field is set to 0 in the case of an Ethernet Private Wire service, and to the VLAN identifier associated with the service for Ethernet Virtual Private Wire service. The route is associated with a Route-Target (RT) extended community attribute that identifies the service instance (together with the Ethernet Tag field when non-zero

3 Operation

The following figure shows an example of a P2P service deployed with E-VPN.



iBGP sessions will be established between PE1, PE2, ASBR1 and ASBR3, possibly via a BGP route-reflector. Similarly, iBGP sessions will be established between PE3, PE4, ASBR2 and ASBR4. eBGP sessions will be established among ASBR1, ASBR2, ASBR3, and ASBR4.

All PEs and ASBRs are enabled for the E-VPN SAFI, and exchange E-VPN Ethernet A-D routes - one route per AC. The ASBRs re-advertise the Ethernet A-D routes with Next Hop attribute set to their IP addresses. The link between the CE and the PE is an C-TAG or S-TAG interface as described in [802.1Q] that can carry a single vlan tag or two vlan tags nested in each other. This interface is setup as a trunk with multiple VLANs.

A VPWS with multiple sites or multiple EVPL services on the same CE port can be included in one EVI between 2 or more PEs. An Ethernet Tag corresponding to each P2P connection and known to both PEs is used to identify the services multiplexed in the same EVI. For CE multi-homing, the Ethernet AD Route encodes the ESI associated with the CE. This allows flow-based load-balancing of traffic between PEs connected to the same multi-homed CE. The VPN ID MUST be the same on both PEs attached to the site. The Ethernet Segment route may be used too, for discovery of multi-homed CEs. In all cases traffic follows the transport paths, which may be asymmetric.

4 E-VPN Comparison to PW Signaling

In E-VPN, service endpoint discovery and label signaling are done concurrently using BGP. Whereas, with VPWS based on [RFC4448], label signaling is done via LDP and service endpoint discovery is either through manual provisioning or through BGP. In VPWS, redundancy is limited to Active/Standby mode, while with E-VPN both Active/Active and Active/Standby redundancy modes can be supported. In VPWS, backup PWs are not used to carry traffic, while E-VPN traffic can be load-balanced among primary and secondary PEs. On link or node failure, E-VPN can trigger failover with the withdrawal of a single BGP route per service, whereas with VPWS PW redundancy, the failover sequence requires exchange of two control plane messages: one message to deactivate the group of primary PWs and a second message to activate the group of backup PWs associated with the access link. Finally, E-VPN may employ data plane local repair mechanisms not available in VPWS.

5 VPWS with multiple sites

The future revision of this draft will describe how a VPWS among multiple sites (full mesh of P2P connections - one per pair of sites) can be setup automatically without any explicit provisioning of P2P connections among the sites.

6 Security Considerations

This document does not introduce any additional security constraints.

7 IANA Considerations

TBD

8 References

8.1 Normative References

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

8.2 Informative References

[EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-00.txt.

[EVPN] A. Sajassi, R. Aggarwal et. al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt.

Authors' Addresses

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

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

INTERNET-DRAFT
Intended Status: Informational Track

Dennis Cai
Sami Boutros
Samer Salam
Reshad Rahman
June 28, 2012

Expires: December 30, 2012

VLAN Aware EVPN services
draft-cai-l2vpn-evpn-vlan-aware-bundling-00.txt

Abstract

This document specifies E-VPN extensions to support the new VLAN aware bundling service interface type defined in [EVPN-REQ]. The new service interface type provides advantages in reducing provisioning overhead as well as E-VPN instances scale in environments where a large number of VLANs need to be extended over an MPLS/IP network, while maintaining traffic segregation among those VLANs. The VLAN aware bundling service interface can handle the high scale requirements of today's Data Centers by bundling different VLANs over a single WAN E-VPN instance used to interconnect those Data Center sites.

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) 2012 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.	VLAN-aware-bundling E-VPN	3
3.	Operation	4
3.1	Packet forwarding, MAC learning, aging and flushing	5
3.2	Multicast Pruning	5
3.3	OAM	5
3.4	VLAN translation	5
4	Security Considerations	6
5	References	6
5.1	Normative References	6
5.2	Informative References	6
6	Appendix Vlan Aware VPLS	6
6.1	VLAN-aware-bundling PW	7
6.2	PW VLAN Vector TLV	7
6.3	LDP Capability Negotiation	8
6.4	Multicast Pruning	9
	Authors' Addresses	9

1 Introduction

The high scale requirements of Layer 2 data center interconnect services mandate the signaling of a large number of WAN E-VPN instances. As such, network operators are looking for solutions whereby they can extend multiple Ethernet VLANs over a WAN using a single E-VPN instance, while maintaining traffic segregation among these VLANs in the data-plane. This gives rise to a requirement for a new service interface types: the VLAN aware bundling service interfaces.

These new VLAN aware bundling service interfaces MUST: - Provide the ability to bundle multiple customer VLANs - Guarantee customer VLAN transparency end-to-end.- Maintain data-plane separation between the customer VLANs by creating a dedicated bridge-domain per VLAN.- Support customer VLAN translation to handle the scenario where different VLAN Identifiers (VIDs) are used on different sites to designate the same customer VLAN.

As discussed in [EVPN-REQ], two new service interface types are defined for VLAN aware bundling: with and without translation. The new service interfaces maintain data-plane separation, per VLAN, while sharing one L2VPN E-VPN instance. This document describes the use of different E-VPN routes as defined in [E-VPN] for implementing the VLAN-aware bundling service.

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].

MAC: Media Access Control

MPLS: Multi Protocol Label Switching.

OAM: Operations, Administration and Maintenance.

PE: Provide Edge Node.

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

EVI: E-VPN Instance.

2. VLAN-aware-bundling E-VPN

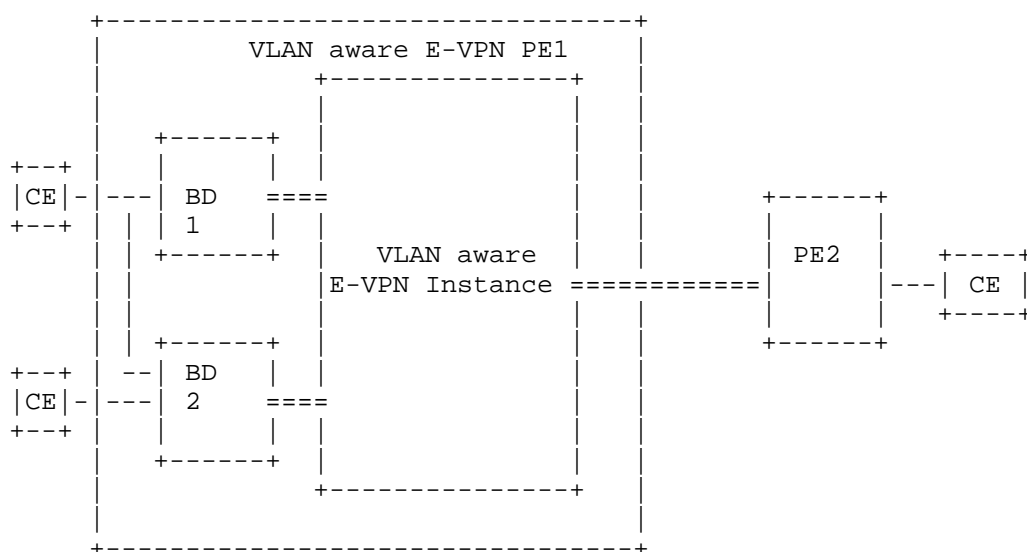
[E-VPN] uses a new BGP NLRI for advertising different route types for

E-VPN operation.

This document discusses how the Ethernet Tag field in the Ethernet Auto-Discovery Route, Mac Advertisement route, and inclusive multicast Ethernet tag route can be used to multiplex several VLANs over the same EVI.

3. Operation

The following figure shows an example of how a VLAN aware Bundling service type over E-VPN could be deployed.



One E-VPN instance has been set up between two sites to extend multiple customer VLANs. On each site, multiple CE devices could be connected to the PE. The link between the CE and the PE could be C-tag or S-tag interface per [802.1Q], carrying several VLANs. Only a single E-VPN instance has been set up to carry customer VLANs between the two sites. The use of two sites in the above figure is for illustration; however, this could be extended to many sites. In order to quantify the benefit of the approach, let's assume N data center sites, with M customer VLANs. With the new VLAN aware service interface type, the solution would require one E-VPN instance, instead of M E-VPN instances. To maintain data-plane separation among the customer VLANs, each PE will create a bridge-domain per customer VLAN. As well, a customer VLAN on each CE port will represent a unique bridge port in the customer bridge-domain. Only one E-VPN instance would be signaled in the core and will be used to carry multiple customer bridge-domains (or customer VLANs) as long as those

customer VLANs need to be extended to the same set of sites. On the egress PE, the E-VPN label + the VLAN-tag would identify the customer-bridge domain.

3.1 Packet forwarding, MAC learning, aging and flushing

Given the data-plane separation, packet forwarding in the scope of one bridge-domain will remain unchanged. When sending traffic over the E-VPN instance, a qualifying VLAN tag MUST be present on the packet. This VLAN tag has global significance across all sites connected to the E-VPN instance and is used to identify the customer bridge domain in all sites. MAC learning, aging and flushing per bridge-domain will remain un-changed. A mass withdraw for MAC routes learned over the EVPN instance can be done by withdrawing the Ethernet AD route with the tag ID corresponding to the bridge domain.

3.2 Multicast Pruning

Efficient multicast replication in the core can be achieved via the use of the Inclusive Multicast Ethernet Tag Route, to prune the flooding on a per VLAN basis. It is possible to only replicate traffic to PEs that have advertised the Inclusive Multicast Ethernet Tag Route with the Tag value set to the VLAN value. If VLAN value is set to zero, then a single multicast LSM is setup to be used for all VLAN traffic for that E-VPN instance. Multicast snooping protocols such as IGMP and PIM MAY be used to further prune the replication scope for a given multicast group in one customer bridge-domain.

3.3 OAM

Customer Ethernet OAM frames (e.g. CFM [802.1ag]) will be carried transparently over the shared E-VPN instance by the customer's bridge-domains. Current MPLS OAM mechanisms need to be extended to verify connectivity in the E-VPN instance shared by the customer bridge-domains, service level OAM monitoring should be performed according to [RFC-6136], MPLS OAM extensions is out of scope of the document.

3.4 VLAN translation

As mentioned above, the VLAN tag carried across the E-VPN instance for the new VLAN aware bundling E-VPN instance MUST have network wide significance within the scope of the E-VPN instance. As such, VLAN translation may be performed at each PE attached to the E-VPN instance to translate between the global VLAN tag identifying the customer bridge-domain and the local VLAN tag used by the customer

bridge-domain on this PE.

4 Security Considerations

This document does not introduce any additional security constraints.

4 IANA Considerations

TBD

5 References

5.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC1776] Crocker, S., "The Address is the Message", RFC 1776, April 1 1995.
- [TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, April 1 1996.

5.2 Informative References

- [EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-00.txt.
- [EVPN] A. Sajassi, R. Aggarwal et. al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt.
- [RFC-6136] Layer 2 Virtual Private Network (L2VPN) Operations, Administration, and Maintenance (OAM) Requirements and Framework.

6 Appendix Vlan Aware VPLS

It is possible to extend VPLS to support VLAN aware bundling type service, a new PW VLAN Vector TLV to be included the LDP PW FEC label mapping messages for the VPLS service, using the mechanisms specified in RFC 4762, as well as a new LDP capability by which a PE can specify its ability to support this new VLAN aware bundling service interface type. The new PW VLAN Vector TLV would allow multiple VLANs

to share a single VPLS instance, while maintaining data plane segregation among these VLANs. This document defines extension to the PWE3 control protocol [RFC4447] to set up the new VLAN aware bundling type service in MPLS networks. An extension to the MAC Withdrawal mechanisms would allow per VLAN service MAC flushing for this new VLAN aware bundling service.

6.1 VLAN-aware-bundling PW

[RFC4447] uses LDP Label Mapping message [RFC5036] for advertising the FEC-to-PW Label binding. Two types of PW FEC, FEC-128 and FEC-129, can be used for this purpose. Both types of PW FEC contain a PW type Field.

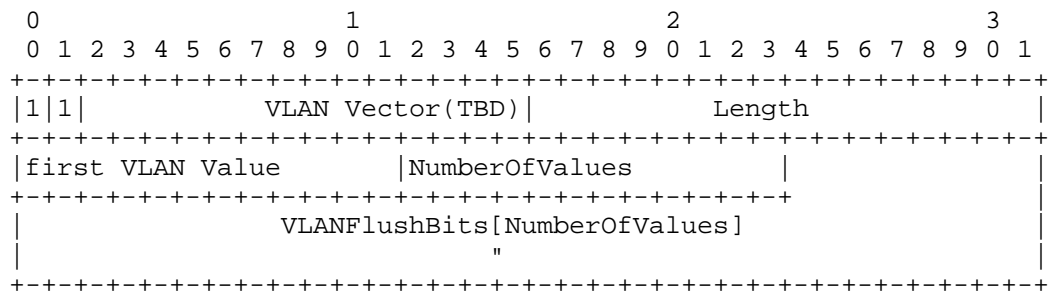
PW type port or raw mode will be used for the VLAN aware bundling interface type service.

Use of control word is optional and frame encapsulation follows the same rules as in [RFC4448].

A new PW VLAN vector TLV is defined, the new PW VLAN Vector TLV will be included in LDP PW label mapping messages, as well it can be included in the MAC flush message.

6.2 PW VLAN Vector TLV

The PW VLAN Vector TLV is described as below:



The U and F bits are set to forward if unknown so that potential intermediate VPLS PES unaware of the new TLV can just propagate it transparently.

The MAC Flush VLAN Vector TLV type is to be assigned by IANA from the LDP standard [RFC5036] "TLV type name space", as described in section 7.

The TLV value field is of variable length. The first 12 bits encode

the starting VLAN value. The second 12 bits contain the number of values. The VLANFlushBits is an array of bits of length = NumberOfValues, each bit in the array represents a VLAN flush state starting from the 1st VLAN value. A bit value of 1 means flush and a bit value of 0 means don't flush

A Starting VLAN value of 0, SHOULD mean include all VLANs, in this case the NumberOfValues SHOULD be 0.

The PW VLAN Vector TLV SHOULD be placed after the PW FEC TLV in the label mapping message as specified in [RFC4447], and SHOULD be placed after the existing TLVs in MAC Flush message as specified in [RFC4762].

6.3 LDP Capability Negotiation

The capability of supporting VLAN Aware Bundling interface type Service MUST be advertised to all LDP peers. This is achieved by using the methods in [RFC5561] and advertising the LDP "VLAN aware Bundling Capability" TLV. If an LDP peer supports the dynamic capability advertisement, it can send a new Capability message with the S bit set for the VLAN Aware Bundling capability TLV. If the peer does not supports dynamic capability advertisement, then the VLAN aware Bundling Capability TLV MUST be included in the LDP Initialization message during the session establishment. An LSR having VLAN Aware Bundling capability MUST recognize the new PW VLAN Vector TLV in LDP label messages.

In line with requirements listed in [RFC5561], the following TLV is defined to indicate the VLAN Aware Bundling capability:

```

      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
      +-----+-----+-----+-----+-----+-----+-----+-----+
      |U|F| VLAN Aware Capability TBD |                               |
      +-----+-----+-----+-----+-----+-----+-----+-----+
      |S| Reserved      |      Reserved      |                               |
      +-----+-----+-----+-----+-----+-----+-----+-----+

```

Note: TLV number pending IANA allocation.

* U-bit: SHOULD be 1 (ignore if not understood).

* F-bit: SHOULD be 0 (don't forward if not understood).

* VLAN Aware Bundling Capability TLV Code Point:

The TLV type, which identifies a specific capability. The VLAN Aware capability code point is requested in the IANA allocation section below.

* S-bit:

The State Bit indicates whether the sender is advertising or withdrawing the VLAN Aware capability. The State bit is used as follows:

1 - The TLV is advertising the capability specified by the TLV Code Point.

0 - The TLV is withdrawing the capability specified by the TLV Code Point.

* Length: MUST be set to 2 (octet).

6.4 Multicast Pruning

Efficient multicast replication in the core can be achieved via the use of the new VLAN vector TLV, to prune the flooding on a per VLAN basis. It is possible to only replicate traffic to PEs that have advertised a given VLAN in their Vector TLV. Multicast snooping protocols such as IGMP and PIM MAY be used to further prune the replication scope for a given multicast group in one customer bridge-domain.

Authors' Addresses

Dennis Cai
Cisco Systems

EMail: dcai@cisco.com

Sami Boutros
Cisco Systems

EMail: sboutros@cisco.com

Samer Salam
Cisco Systems

EMail: ssalam@cisco.com

Reshad Rahman
Cisco Systems

EMail: rrahman@cisco.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

Aldrin Isaac
Bloomberg

J. Drake
R. Shekhar
Juniper Networks

J. Uttaro
AT&T

Expires: January 14, 2012

July 14, 2012

BGP MPLS Based Ethernet VPN
draft-ietf-l2vpn-evpn-01

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) 2012 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 (E-VPN).

Table of Contents

1. Specification of requirements	5
2. Contributors	5
3. Introduction	5
4. Terminology	5
5. BGP MPLS Based E-VPN Overview	6
6. Ethernet Segment	7
7. Ethernet Tag	8
7.1 VLAN Based Service Interface	9
7.2 VLAN Bundle Service Interface	9
7.2.1 Port Based Service Interface	9
7.3 VLAN Aware Bundle Service Interface	9
8. BGP E-VPN NLRI	10
8.1. Ethernet Auto-Discovery Route	10
8.2. MAC Advertisement Route	11
8.3. Inclusive Multicast Ethernet Tag Route	11
8.4 Ethernet Segment Route	12
8.5 ESI MPLS Label Extended Community	12
8.6 ES-Import Extended Community	13
8.7 MAC Mobility Extended Community	13
9. Multi-homing Functions	13
9.1 Multi-homed Ethernet Segment Auto-Discovery	13
9.1.1 Constructing the Ethernet Segment Route	14
9.2 Fast Convergence	14
9.2.1 Constructing the Ethernet A-D Route per Ethernet Segment	15
9.2.1.1. Ethernet A-D Route Targets	15
9.3 Split Horizon	16
9.3.1 ESI MPLS Label Assignment	16
9.3.1.1 Ingress Replication	16
9.3.1.2. P2MP MPLS LSPs	17
9.3.1.3. MP2MP LSPs	18

9.4 Aliasing	18
9.4.1 Constructing the Ethernet A-D Route per EVI	18
9.4.1.1 Ethernet A-D Route Targets	19
9.5 Designated Forwarder Election	20
9.5.1 Default DF Election Procedure	21
9.5.2 DF Election with Service Carving	21
10. Determining Reachability to Unicast MAC Addresses	22
10.1. Local Learning	23
10.2. Remote learning	23
10.2.1. Constructing the BGP E-VPN MAC Address Advertisement	23
11. ARP and ND	25
12. Handling of Multi-Destination Traffic	26
12.1. Construction of the Inclusive Multicast Ethernet Tag Route	26
12.2. P-Tunnel Identification	27
13. Processing of Unknown Unicast Packets	28
13.1. Ingress Replication	28
13.2. P2MP MPLS LSPs	29
14. Forwarding Unicast Packets	29
14.1. Forwarding packets received from a CE	29
14.2. Forwarding packets received from a remote PE	30
14.2.1. Unknown Unicast Forwarding	30
14.2.2. Known Unicast Forwarding	31
15. Load Balancing of Unicast Frames	31
15.1. Load balancing of traffic from an PE to remote CEs	31
15.1.1 Active-Standby Redundancy Mode	31
15.1.2 All-Active Redundancy Mode	32
15.2. Load balancing of traffic between an PE and a local CE	33
15.2.1. Data plane learning	34
15.2.2. Control plane learning	34
16. MAC Mobility	34
17. Multicast	36
17.1. Ingress Replication	36
17.2. P2MP LSPs	36
17.3. MP2MP LSPs	36
17.3.1. Inclusive Trees	36
17.3.2. Selective Trees	37
17.4. Explicit Tracking	38
18. Convergence	38
18.1. Transit Link and Node Failures between PEs	38
18.2. PE Failures	38
18.2.1. Local Repair	38
18.3. PE to CE Network Failures	39
19. LACP State Synchronization	39
20. Acknowledgements	40
21. References	40
21. Author's Address	41

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. 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

3. Introduction

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

4. Terminology

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

E-VPN Instance (EVI): An E-VPN routing and forwarding instance on a PE.

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 E-VPN instance consists of one or more broadcast domains. Ethernet tag(s) are assigned to the broadcast domains of a given E-VPN instance by the provider of that E-VPN, and each PE in that E-VPN instance performs a mapping between broadcast

domain identifier(s) understood by each of its attached CEs and the corresponding Ethernet tag.

Link Aggregation Control Protocol (LACP):

Multipoint to Multipoint (MP2MP):

Point to Multipoint (P2MP):

Point to Point (P2P):

5. BGP MPLS Based E-VPN Overview

This section provides an overview of E-VPN.

An E-VPN 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 E-VPNs 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 E-VPN, 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 E-VPN, 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 E-VPN are very similar to those of IP-VPN. An EVI requires a Route-Distinguisher (RD) and one or more Route-Targets (RTs). A CE attaches to an E-VPN instance (EVI) on an PE, on an Ethernet interface which may be configured for one or more Ethernet Tags, e.g., VLANs. Some deployment scenarios guarantee uniqueness of VLANs across E-VPNs: all points of attachment of a given EVI use the same VLAN, and no other EVI uses this VLAN. This document refers to this case as a "Unique VLAN E-VPN" 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. A single-homed CE is considered to be attached to an Ethernet segment with ESI 0. Otherwise, an Ethernet segment MUST have a unique non-zero ESI. The ESI can be assigned using various mechanisms:

1. The ESI may be configured. For instance when E-VPNs are used to provide a VPLS service the ESI is fairly analogous to the Multi-homing site ID in [BGP-VPLS-MH].

2. 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.

3. If LLDP is used between the PEs and CEs that are hosts, then the ESI is determined by LLDP. The ESI will be specified in a following version.

4. 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)}

7. Ethernet Tag

An Ethernet Tag identifies a particular broadcast domain, e.g. a VLAN, in an EVI. An EVI consists of one or more broadcast domains. Ethernet Tags are assigned to the broadcast domains of a given EVI by the provider of the E-VPN service. Each PE, in a given EVI, performs a mapping between the Ethernet Tag and the corresponding broadcast domain identifier(s) understood by each of its attached CEs (e.g. CE VLAN Identifiers or CE-VIDs).

If the broadcast domain identifier(s) are understood consistently by all of the CEs in an EVI, the broadcast domain identifier(s) MAY be used as the corresponding Ethernet Tag(s). In other words, the Ethernet Tag ID assigned by the provider is numerically equal to the broadcast domain identifier (e.g., CE-VID = Ethernet Tag).

Further, some deployment scenarios guarantee uniqueness of broadcast domain identifiers across all EVIs; all points of attachment of a given EVI use the same broadcast domain identifier(s) and no other EVI uses these broadcast domain identifier(s). This allows the RT(s) for each EVI to be derived automatically, as described in section 9.4.1.1.1 "Auto-Derivation from the Ethernet Tag ID".

The following subsections discuss the relationship between Ethernet Tags, EVIs and broadcast domain identifiers as well as the setting of the Ethernet Tag Identifier, in the various E-VPN 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, there is a one-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there is a single bridge domain per PE for the EVI. Different CEs connected to different PE ports MAY use different broadcast domain identifiers (e.g. CE-VIDs) for the same EVI. If said identifiers are different, the frames SHOULD remain tagged with the originating CE's broadcast domain identifier (e.g. CE-VID). When the CE broadcast domain identifiers are not consistent, a tag translation function MUST be supported in the data path and MUST be performed on the disposition PE. The Ethernet Tag Identifier in all E-VPN routes MUST be set to 0.

7.2 VLAN Bundle Service Interface

With this service interface, there is a many-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there is a single bridge domain per PE for the EVI. Different CEs connected to different PE ports MUST use the same broadcast domain identifiers (e.g. CE-VIDs) for the same EVI. The MPLS encapsulated frames MUST remain tagged with the originating CE's broadcast domain identifier (e.g. CE-VID). Tag translation is NOT permitted. The Ethernet Tag Identifier in all E-VPN 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, there is a many-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there are multiple bridge domains per PE for the EVI: one broadcast domain per CE broadcast domain identifier. In the case where the CE broadcast domain identifiers are not consistent for different CEs, a normalized Ethernet Tag 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 on both the imposition as well as the disposition PEs. The Ethernet Tag Identifier in all E-VPN routes MUST be set to the normalized Ethernet Tag assigned by the E-VPN provider.

8. BGP E-VPN NLRI

This document defines a new BGP NLRI, called the E-VPN NLRI.

Following is the format of the E-VPN NLRI:

```
+-----+
|   Route Type (1 octet)   |
+-----+
|   Length (1 octet)      |
+-----+
| Route Type specific (variable) |
+-----+
```

The Route Type field defines encoding of the rest of the E-VPN NLRI (Route Type specific E-VPN NLRI).

The Length field indicates the length in octets of the Route Type specific field of E-VPN 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 E-VPN NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an AFI of TBD and an SAFI of E-VPN (To be assigned by IANA). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the E-VPN NLRI (encoded as specified above).

In order for two BGP speakers to exchange labeled E-VPN 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 TBD and an SAFI of E-VPN.

8.1. Ethernet Auto-Discovery Route

A Ethernet A-D route type specific E-VPN 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 E-VPN 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 ($n * 3$ octets)

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 E-VPN 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 E-VPN 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)

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

8.5 ESI MPLS Label Extended Community

This extended community is a new transitive extended community. 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 MPLS 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
0x44								Sub-Type								Flags (One Octet)								Reserved=0							
Reserved = 0								ESI MPLS label																							

The low order bit of the flags octet is defined as the "Active-

Standby" bit and may be set to 1. The other bits must be set to 0.

8.6 ES-Import Extended Community

This is a new transitive 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 Extended Community. 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
+-----+-----+-----+-----+-----+-----+-----+-----+
| 0x44          | Sub-Type          | ES-Import          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     ES-Import Cont'd         |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

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. 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
+-----+-----+-----+-----+-----+-----+-----+-----+
| 0x44          | Sub-Type          | Reserved=0          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Sequence Number          |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

9. Multi-homing Functions

This section discusses the functions, procedures and associated BGP routes used to support multi-homing in E-VPN. 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.

Note that the new ES-Import extended community is not the same as the Route Target Extended Community. The Ethernet Segment route carries this new ES-Import extended community. The PEs apply filtering on this new extended community. As a result the Ethernet Segment route is imported only by the PEs that are connected to the same Ethernet segment.

9.2 Fast Convergence

In E-VPN, 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, E-VPN 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 MPLS label used for split-horizon filtering (as discussed in section 9.2). 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 reason for such encoding is that the RD cannot be that of a given EVI since the ESI can span across one or more EVIs.

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 MPLS 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 MPLS 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 9.2.

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

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 EVIs 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. If the CE sends a multicast, broadcast or unknown unicast packet to a particular PE, say PE1, then PE1 will forward that packet to all or subset of the other PEs in the EVI. In this case the PEs, other than PE1, that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This is referred to as "split horizon" filtering in this document.

In order to achieve this split horizon function, every multicast, broadcast or unknown unicast packet is encapsulated with an MPLS label that identifies the Ethernet segment of origin (i.e. the segment from which the frame entered the E-VPN network). This label is referred to as the ESI MPLS label, and is distributed using the "Ethernet A-D route per Ethernet Segment" as per the procedures in section 9.1.1 above. This route is imported by the PEs connected to the Ethernet Segment and also by the PEs that have at least one EVI in common with the Ethernet Segment in the route. As described in section 9.1.1, the route MUST carry an ESI MPLS Label Extended Community with a valid ESI MPLS label. The disposition PEs rely on the value of the ESI MPLS label to determine whether or not a flooded frame is allowed to egress a specific Ethernet segment.

9.3.1 ESI MPLS Label Assignment

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

9.3.1.1 Ingress Replication

An PE that is using ingress replication for sending broadcast, multicast or unknown unicast traffic, distributes to other PEs, that belong to the Ethernet segment, a downstream assigned "ESI MPLS label" in the Ethernet A-D route. 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. Further consider that PE1 is using P2P or MP2P LSPs to send packets to PE2.

Consider that PE1 receives a multicast, broadcast or unknown unicast packet from CE1 on VLAN1 on ESI1. In this scenario, PE2 distributes an Inclusive Multicast Ethernet Tag route for VLAN1 in the associated EVI. So, when PE1 sends a multicast, broadcast or unknown unicast packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that PE2 has distributed for ESI1. 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 ESI1, then PE2 MUST NOT forward the packet onto ESI1.

9.3.1.2. P2MP MPLS LSPs

An PE that is using P2MP LSPs for sending broadcast, multicast or unknown unicast traffic, distributes to other PEs, that belong to the Ethernet segment or have an E-VPN in common with the Ethernet Segment, an upstream assigned "ESI MPLS 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 PEs, 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 PEs, 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 ESI. Also consider PE3 that is in the same EVI as one of the EVIs to which ESI belongs. Further, assume that PE1 is using P2MP MPLS LSPs to send broadcast, multicast or unknown unicast packets. When PE1 sends a multicast, broadcast or unknown unicast 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 PEs. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for E-VPN. 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 ESI1, then PE2 MUST NOT forward the packet onto ESI1. 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 ESI1 and PE3 is not connected to ESI1, then PE3 MUST pop the label and flood the packet over all local ESIs in the EVI.

9.3.1.3. MP2MP LSPs

The procedures for ESI MPLS Label assignment and usage for MP2MP LSPs will be described in a future version.

9.4 Aliasing

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, E-VPN introduces the concept of 'Aliasing'. Aliasing refers to the ability of an 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-zero ESI SHOULD consider the advertised MAC address as reachable via all PEs which have advertised reachability to the relevant Segment using Ethernet A-D routes with the same ESI (and Ethernet Tag if applicable).

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

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 E-VPN 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. Optional mechanisms, which will be specified in the future, will allow selecting a DF at the granularity of <ESI, EVI, S, G>.

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 E-VPN 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 granularity of the DF election MUST be at least the Ethernet segment via which the CE is multi-homed to the PEs. If the DF election is done at the Ethernet segment granularity then a single PE MUST be elected as the DF on the Ethernet segment.

If there are one or more EVIs enabled on the Ethernet segment, then the granularity of the DF election SHOULD be the combination of the Ethernet segment and EVI on that Ethernet segment. In this case a single PE MUST be elected as the DF for a particular EVI on that Ethernet segment.

The detailed procedures for DF election are described next.

9.5.1 Default DF Election Procedure

As a PE discovers the other PEs that are connected to the same Ethernet Segment, using the Ethernet Segment routes, it starts building an ordered list based on the originating PE IP addresses. This list is used to select a DF and a backup DF (BDF) on a per Ethernet Segment basis. By default, the PE with the numerically highest IP address is considered the DF for that Ethernet Segment and the next PE in the list is considered the BDF.

If the Ethernet Segment is a multi-homed device, then the elected DF is the only PE that must forward flooded multi-destination packets towards the segment. All other PE nodes must not permit multi-destination packets to egress to the segment. In the case where the DF fails, the BDF takes over its functionality.

This procedure enables the election of a single DF per Ethernet Segment, for all EVIs enabled on the segment. It is possible to achieve more granular load-balancing of traffic among the PE nodes by employing Service Carving, as discussed in the next section.

9.5.2 DF Election with 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 to allow the reception of Ethernet Segment routes from other PE nodes 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. 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 EVI 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 EVI V when $(V \bmod N) = i$.

The above procedure results in the entire EVI range being divided up among the PEs in the RG, regardless of whether a given EVI is configured/enabled on the associated Ethernet Segment or not.

4. The PE that is elected as a DF for a given EVI will unblock traffic for that EVI only if the EVI is configured/enabled on the Segment. 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 EVIs) 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. 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.

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 E-VPN 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 the same PE or 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 the EVI, using MP-BGP and specifically the MAC Advertisement route.

10.2.1. Constructing the BGP E-VPN MAC Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC Advertisement route type in the E-VPN 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 EVI (e.g., the PE needs to perform qualified learning for the VLANs in that EVI).

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 E-VPN 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 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 Length field value is set to the number of octets in the IP Address field.

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 is included, it is encoded as specified in section 12.

The MPLS 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]). The MPLS label stack MUST be the downstream assigned E-VPN MPLS label stack that is used by the PE to forward MPLS-encapsulated Ethernet frames received from remote PEs, 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 E-VPN 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 E-VPN 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 E-VPN label per MAC address. All of these methodologies have their tradeoffs.

Per EVI label assignment requires the least number of E-VPN 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 E-VPN".

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.

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 E-VPN 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 and respond to the ARP request.

Further detailed procedures will be specified in a later version.

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 in an EVI, to all the other PEs that span that Ethernet Tag in the EVI. 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 E-VPN 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 E-VPN 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 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 E-VPN 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 E-VPN 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 E-VPN, 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 E-VPN, 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 E-VPN 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. Flooding 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 [RFC 4761, RFC 4762]. When an PE capable of flooding (say PEx) receives a broadcast or multicast Ethernet frame, or one with an unknown destination MAC address, it must flood 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 the EVI, 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 the EVI. 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 broadcast and multicast packets, as well as packets to an unknown MAC address.

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 E-VPN may use ingress replication using RSVP-TE P2P LSPs or LDP MP2P LSPs for sending broadcast, multicast and unknown unicast traffic to other PEs. Or they may use RSVP-TE P2MP or LDP P2MP or LDP MP2MP LSPs 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 EVI, specifies

the downstream label that the other PEs can use to send unknown unicast, multicast or broadcast traffic for the EVI 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 EVI 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 EVIs, 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

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 E-VPN 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 <EVI, 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 EVI. If the same P2MP LSP is used for all Ethernet Tags, then all the PEs in the EVI MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP is used for a given Ethernet Tag in the EVI, 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

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 EVI or 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 unknown unicast, broadcast or multicast 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 the destination MAC address is unknown unicast, broadcast or multicast and 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 unknown unicast 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 E-VPN 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 Active-Standby 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 MPLS Label Extended Community is set, then the remote PE MUST deduce that the Ethernet segment is operating in Active-Standby 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.

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.

15.1.2 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 MPLS 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

<ESI1, VLAN1>.

The MPLS label stack to send the packets to PE1 is the MPLS LSP stack to get to PE1 and the E-VPN 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 <ESI1, 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 E-VPN 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 E-VPN 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 E-VPN 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 E-VPN 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.

Since the sequence number is an unsigned 32 bit integer, all sequence number comparisons must be performed modulo 2^{32} . This unsigned arithmetic preserves the relationship of sequence numbers as they cycle from $2^{32} - 1$ to 0.

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 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 sequence number in the received MAC Advertisement route, if the received route is tagged with a MAC Mobility extended community attribute.

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.

17. Multicast

The PEs in a particular E-VPN 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 unknown unicast, multicast or broadcast traffic as described in section "Handling of Multi-Destination Traffic". A given unknown unicast or 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 unknown unicast, broadcast or multicast packet or a "Selective" tree. 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. For selective P-Multicast trees, only unicast PE-PE tunnels (using MPLS or IP/GRE encapsulation) and P2MP LSPs are supported, and the supported P2MP LSP signaling protocols are RSVP-TE, and mLDP.

17.3. MP2MP LSPs

The root of the MP2MP LDP LSP advertises the Inclusive Multicast Tag route with the PMSI Tunnel attribute set to the MP2MP Tunnel identifier. This advertisement is then sent to all PEs in the E-VPN.

Upon receiving the Inclusive Multicast Tag routes with a PMSI Tunnel attribute that contains the MP2MP Tunnel identifier, the receiving PEs initiate the setup of the MP2MP tunnel towards the root using the procedures in [MLDP].

17.3.1. 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 EVIs on a given PE. A particular P-Multicast tree can be set up to carry the traffic originated by sites belonging to a single E-VPN, or to carry the traffic originated by sites belonging to different E-VPNs. The

ability to carry the traffic of more than one E-VPN on the same tree is termed 'Aggregation'. The tree needs to include every PE that is a member of any of the E-VPNs 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 E-VPN 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 EVIs 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 E-VPN Inclusive Multicast routes.

17.3.2. Selective Trees

A Selective P-Multicast tree is used by an PE to send IP multicast traffic for one or more specific IP multicast streams, originated by CEs connected to the PE, that belong to the same or different E-VPNs, to a subset of the PEs that belong to those E-VPNs. Each of the PEs in the subset should be on the path to a receiver of one or more multicast streams that are mapped onto the tree. The ability to use the same tree for multicast streams that belong to different E-VPNs is termed an PE the ability to create separate SP multicast trees for specific multicast streams, e.g. high bandwidth multicast streams. This allows traffic for these multicast streams to reach only those PE routers that have receivers in these streams. This avoids flooding other PE routers in the E-VPN.

An SP can use both Inclusive P-Multicast trees and Selective P-Multicast trees or either of them for a given E-VPN on an PE, based on local configuration.

The granularity of a selective tree is <RD, PE, S, G> where S is an IP multicast source address and G is an IP multicast group address or G is a multicast MAC address. Wildcard sources and wildcard groups are supported. Selective trees require explicit tracking as described below.

A E-VPN PE advertises a selective tree using a E-VPN selective A-D route. The procedures are the same as those in [VPLS-MCAST] with S-

PMSI A-D routes in [VPLS-MCAST] replaced by E-VPN Selective A-D routes. The information elements of the E-VPN selective A-D route are similar to those of the VPLS S-PMSI A-D route with the following differences. A E-VPN Selective A-D route includes an optional Ethernet Tag field. Also an E-VPN selective A-D route may encode a MAC address in the Group field. The encoding details of the E-VPN selective A-D route will be described in the next revision.

Selective trees can also be aggregated on the same P2MP LSP using aggregation as described in [VPLS-MCAST].

17.4. Explicit Tracking

[VPLS-MCAST] describes procedures for explicit tracking that rely on Leaf A-D routes. The same procedures are used for explicit tracking in this specification with VPLS Leaf A-D routes replaced with E-VPN Leaf A-D routes. These procedures allow a root PE to request multicast membership information for a given (S, G), from leaf PEs. Leaf PEs rely on IGMP snooping or PIM snooping between the PE and the CE to determine the multicast membership information. Note that the procedures in [VPLS-MCAST] do not describe how explicit tracking is performed if the CEs are enabled with join suppression. The procedures for this case will be described in a future version.

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.

18.2.1. Local Repair

It is possible to perform local repair in the case of PE failures. Details will be specified in the future.

18.3. 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. LACP State Synchronization

This section requires review and discussion amongst the authors and will be revised in the next version.

To support CE multi-homing with multi-chassis Ethernet bundles, the PEs connected to a given CE should synchronize [802.1AX] LACP state amongst each other. This ensures that the PEs can present a single LACP bundle to the CE. This is required for initial system bring-up and upon any configuration change.

This includes at least the following LACP specific configuration parameters:

- System Identifier (MAC Address): uniquely identifies a LACP speaker.
- System Priority: determines which LACP speaker's port priorities are used in the Selection logic.
- Aggregator Identifier: uniquely identifies a bundle within a LACP speaker.
- Aggregator MAC Address: identifies the MAC address of the bundle.
- Aggregator Key: used to determine which ports can join an Aggregator.
- Port Number: uniquely identifies an interface within a LACP speaker.
- Port Key: determines the set of ports that can be bundled.
- Port Priority: determines a port's precedence level to join

a bundle in case the number of eligible ports exceeds the maximum number of links allowed in a bundle.

Furthermore, the PEs should also synchronize operational (run-time) data, in order for the LACP Selection logic state-machines to execute. This operational data includes the following LACP operational parameters, on a per port basis:

- Partner System Identifier: this is the CE System MAC address.
- Partner System Priority: the CE LACP System Priority
- Partner Port Number: CE's AC port number.
- Partner Port Priority: CE's AC Port Priority.
- Partner Key: CE's key for this AC.
- Partner State: CE's LACP State for the AC.
- Actor State: PE's LACP State for the AC.
- Port State: PE's AC port status.

The above state needs to be communicated between PEs forming a multi-chassis bundle during LACP initial bringup, upon any configuration change and upon the occurrence of a failure.

It should be noted that the above configuration and operational state is localized in scope and is only relevant to PEs which connect to the same multi-homed CE over a given Ethernet bundle.

Furthermore, the communication of state changes, upon failures, must occur with minimal latency, in order to minimize the switchover time and consequent service disruption. The protocol details for synchronizing the LACP state will be described in the following version.

20. Acknowledgements

We would like to thank Yakov Rekhter, 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 for his review.

21. References

- [E-VPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt
- [RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006
- [VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-

l2vpn-vpls-mcast-04.txt

[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.

[VPLS-MULTIHOMING] "BGP based Multi-homing in Virtual Private LAN Service", K. Kompella et. al., draft-ietf-l2vpn-vpls-multihoming-00.txt

[PIM-SNOOPING] "PIM Snooping over VPLS", V. Hemige et. al., draft-ietf-l2vpn-vpls-pim-snooping-01

[IGMP-SNOOPING] "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", M. Christensen et. al., RFC4541,

[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

[EVPN-SEGMENT-ROUTE] A. Sajassi et. al., "E-VPN Ethernet Segment Route", draft-sajassi-l2vpn-evpn-segment-route-00.txt, work in progress.

21. Author's Address

Rahul Aggarwal
Email: raggarwa_1@yahoo.com

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.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
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1, Canada

Email: ssalam@cisco.com

Layer 2 Virtual Private Networks
Internet-Draft
Intended status: Informational
Expires: January 17, 2013

O. Dornon
J. Kotalwar
Alcatel-Lucent
J. Zhang
Juniper Networks, Inc.
V. Hemige
Alcatel-Lucent
July 16, 2012

PIM Snooping over VPLS
draft-ietf-l2vpn-vpls-pim-snooping-02

Abstract

In Virtual Private LAN Service (VPLS), as also in IEEE Bridged Networks, the switches simply flood multicast traffic on all ports in the LAN by default. IGMP Snooping is commonly deployed to ensure multicast traffic is not forwarded on ports without IGMP receivers. The procedures and recommendations for IGMP Snooping are defined in [IGMP-SNOOP]. But when any protocol other than IGMP is used, the common practice is to simply flood multicast traffic to all ports. PIM-SM, PIM-SSM, PIM-BIDIR are widely deployed routing protocols. PIM Snooping procedures are important to restrict multicast traffic to only the routers interested in receiving such traffic.

While most of the PIM Snooping procedures defined here also apply to IEEE Bridged Networks, VPLS demands certain special procedures due to the split-horizon rules that require the Provider Edge (PE) devices to co-operate. This document describes the procedures and recommendations for PIM-Snooping in VPLS to facilitate replication to only those ports behind which there are interested PIM routers and/or IGMP hosts. This document also describes procedures for PIM Proxy. PIM Proxy is required on PEs for VPLS Multicast to work correctly when Join suppression is enabled in the VPLS. PIM Proxy also helps scale VPLS Multicast much better than just PIM Snooping.

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 17, 2013.

Copyright Notice

Copyright (c) 2012 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. Assumptions	6
1.2. PIM Snooping and PIM Proxy Complexity	6
1.3. Definitions	6
2. Multicast Traffic over VPLS	7
2.1. Constraining of IP Multicast in a VPLS	8
2.2. IPv6 Considerations	9
2.3. PIM-SM (*,*,RP) Considerations	9
2.4. PIM Packet Types to Snoop	9
2.5. PIM Snooping vs PIM Proxy	9
2.5.1. Differences between PIM Snooping and PIM Proxy	10
2.5.2. PIM Control Message Latency	11
2.5.3. When to Snoop and When to Proxy	11
3. PIM Snooping for VPLS	12
3.1. General Rules for PIM Snooping in VPLS	13
3.1.1. Snooping PIM Packets	13
3.1.2. Preserving Assert Trigger	13
3.2. Discovering PIM Routers	14
3.3. PIM-SM and PIM-SSM	15
3.3.1. Building PIM-SM Snooping States	15
3.3.2. Explanation for per (S,G,N) states	17
3.3.3. Receiving (*,G) PIM-SM Join/Prune Messages	18
3.3.4. Receiving (S,G) PIM-SM Join/Prune Messages	20
3.3.5. Receiving (S,G,rpt) Join/Prune Messages	22
3.3.6. Sending Join/Prune Messages Upstream	22
3.4. Bidirectional-PIM (PIM-BIDIR)	23
3.5. Interaction with IGMP Snooping	24
3.6. PIM-DM	24
3.6.1. Building PIM-DM Snooping States	24
3.6.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine	25
3.6.3. Triggering ASSERT election in PIM-DM	25
3.7. PIM Proxy	25
3.7.1. Downstream PIM Proxy behavior	26
3.7.2. Upstream PIM Proxy behavior	26
3.7.3. Source IP Address in Proxy PIM Join/Prune Packets	26
3.8. Directly Connected Multicast Source	27
3.9. Data Forwarding Rules	27
3.9.1. PIM-SM Data Forwarding Rules	28
3.9.2. PIM-BIDIR Data Forwarding Rules	29
3.9.3. PIM-DM Data Forwarding Rules	30
4. IANA Considerations	31
5. Security Considerations	31
6. Contributors	31
7. Acknowledgements	32
8. References	32
8.1. Normative References	32

8.2. Informative References	32
Appendix A. PIM-BIDIR Thoughts	33
Appendix B. Example Network Scenario	33
B.1. Pim Snooping Example	34
B.2. PIM Proxy Example with (S,G) / (*,G) interaction	36
Authors' Addresses	39

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 particular VPLS instance is populated dynamically by source MAC address learning. This is a straightforward solution to support unicast traffic, with reasonable flooding for unicast unknown traffic. Since a VPLS provides LAN emulation for IEEE bridges as well as for routers, the unicast and multicast traffic need to follow the same path for layer-2 protocols to work properly. As such, multicast traffic is treated as broadcast traffic and is flooded to every site in the VPLS instance. 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: A. Multicast traffic is replicated to non-member sites. B. Replication of PWS on shared physical path.

This document solves Issue A of [VPLS-MCAST-REQ] by ensuring that IP multicast traffic is not forwarded to non-member sites. Issue B is outside the scope of this document. The different mechanisms to tunnel IP multicast traffic in a VPLS from the ingress PE to the egress PEs are discussed in [VPLS-MCAST-TREES]. The solution in this document when combined with the solutions proposed in the working group to solve Issue B will provide a complete VPLS Multicast solution set.

Using IGMP/PIM Snooping in VPLS has the following advantages:

- o It improves IP Multicast bandwidth usage in the VPLS core by ensuring traffic is replicated only to PEs with member sites. Note that this is not necessarily optimum, as there can still be bandwidth waste if traffic from a PE to other PE(s) is not forwarded along a minimum cost spanning tree.
- o It prevents sending multicast traffic to non-member sites.

Procedures for IGMP Snooping are specified in [IGMP-SNOOP]. This document describes the procedures for Protocol Independent Multicast (PIM) snooping over VPLS for efficient distribution of IP multicast traffic. It also describes the rules when both IGMP and PIM are active in a VPLS instance.

This document also describes procedures for PIM Proxy. PIM Proxy is required on PEs for VPLS Multicast to work correctly when Join suppression is enabled in the VPLS. PIM Proxy also helps scale VPLS Multicast much better than just PIM Snooping.

1.1. Assumptions

Since this draft describes the procedures for PIM Snooping and PIM Proxy, the draft 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 assumes that the user will infer such detail from the PIM RFC referenced in this document. Deviations in protocol handling specific to PIM Snooping and PIM Proxy are specified in this draft. There could be cross references into definitions of macros and procedures from the PIM RFCs.

1.2. PIM Snooping and PIM Proxy Complexity

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

The switch 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 switch 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 switch.

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

The solutions described here provide complete separation of control and data planes.

A PIM Proxy solution minimizes the control plane messages received at CE routers by proxying one message upstream on behalf of a large number of downstream CEs. As such control plane messaging is very similar to that of a PIM Router.

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 switch 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. Multicast Traffic over VPLS

In VPLS, if a PE receives a frame from an Attachment Circuit (AC) with no matching entry in the forwarding information base for that particular VPLS instance, it floods the frame to all other PEs (which are part of this VPLS instance) and to directly connected ACs (other than the one that the frame is received from). The flooding of a frame occurs when:

- o The destination MAC address has not been learned,
- o The destination MAC address is a broadcast address,
- o The destination MAC address is a multicast address.

Malicious attacks (e.g., receiving unknown frames constantly) aside, the first situation is handled by VPLS solutions as long as destination MAC address can be learned. After that point on, the frames will not be flooded. A PE is REQUIRED to have safeguards, such as unknown unicast limiting and MAC table limiting, against malicious unknown unicast attacks.

There is no way around flooding broadcast frames. To prevent runaway broadcast traffic from adversely affecting the VPLS service and the SP network, a PE is REQUIRED to have tools to rate limit the broadcast traffic as well.

Similar to broadcast frames, multicast frames are flooded as well, as

a PE cannot know where multicast members reside. Rate limiting multicast traffic, while possible, should be done carefully since several network control protocols relies on multicast. For one thing, layer-2 and layer-3 protocols utilize multicast for their operation. For instance, Bridge Protocol Data Units (BPDUs) use an IEEE assigned all bridges multicast MAC address, and OSPF is multicast to all OSPF routers multicast MAC address. If the rate-limiting of multicast traffic is not done properly, the customer network will experience instability and poor performance. For the other, it is not straightforward to determine the right rate limiting parameters for multicast.

A VPLS solution MUST NOT affect the operation of customer layer-2 protocols (e.g., BPDUs). Additionally, a VPLS solution MUST NOT affect the operation of layer-3 protocols.

In the following section, we describe procedures to constrain the flooding of IP multicast traffic in a VPLS.

2.1. Constraining of IP Multicast in a VPLS

For a PE in a VPLS (a layer-2 device) to constrain IP multicast traffic, it needs to be able to learn which CEs are interested in receiving multicast traffic for what flows.

The most obvious solution is to snoop IP multicast control traffic at the PEs. Snooping as a solution to constrain multicast traffic makes sense under the following circumstances:

- o The CE-CE protocol the PEs snoop is a popular and widely deployed protocol.
- o It does not require any changes on the CEs and it should be completely transparent to the CEs.

IGMP/MLD and PIM are the popular IP Multicast Routing protocols today. Other routing protocols such as DVMRP or MOSPF are outside the scope of this document.

This document describes the guidelines for PIM Snooping and PIM Proxy in VPLS. The specifications in this document could be used for either PIM Snooping or PIM Proxy. The PIM Proxy solution is described in section Section 3.7. 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 2.5. We will largely refer to PIM "Snooping" in this document. Unless specifically specified, the same procedures should apply to a Proxy solution as well.

In the following sub-sections, we provide some guidelines for the implementation of PIM snooping in VPLS. Snooping techniques need to be employed on ACs at the downstream PEs. Snooping techniques can also be employed on PWs at the upstream PEs. 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.

2.2. IPv6 Considerations

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 switch, 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. PIM-SM (*,*,RP) Considerations

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. Given the complexity of supporting (*,*,RP) states and knowing that there is little to no use to supporting it, this draft omits the specification relating to (*,*,RP) support.

2.4. PIM Packet Types to Snoop

A PIM Snooping switch need only snoop on PIM Hellos and PIM Join/Prune packets. All other PIM packets can be transparently flooded unexamined.

2.5. PIM Snooping vs PIM Proxy

PIM Snooping switches simply snoop on PIM packets as they are being forwarded in the VPLS. As such it truly provides transparent LAN services since no customer packets are modified or consumed or new packets introduced in the VPLS. It is also slightly 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 as a solution 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 on the PEs 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 be provided without requiring any changes on the CE routers. It also helps scale VPLS Multicast very well since the PEs intelligently forward only one Join/Prune message for a given flow and only to the upstream CE.

PIM Proxy as a solution 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.

2.5.1. Differences between PIM Snooping and PIM Proxy

For PIM-SM and PIM-BIDIR, a PIM Snooping/Proxy Switch only needs to examine PIM Hello and Join/Prune messages. PIM Proxy for PIM-DM is for future study and is not currently specified in this draft.

A proxy switch performs proxy only for the Join/Prune messages. PIM Hello messages are snooped by both PIM Snooping and PIM Proxy switches.

Details on the PIM Proxy solution are discussed in section Section 3.7. This section is presented here to say that most of the procedures to follow (unless explicitly specified) are common to both PIM Snooping and PIM Proxy. Differences between a PIM Snooping switch and a PIM Proxy switch can be summarized as the following:

PIM Snooping	PIM Proxy
1. PIM Snooping switches snoop Hello and Join/Prune messages while they are transparently flooded in the VPLS.	1. PIM Proxy switches also snoop PIM Hello messages while they are transparently flooded in the VPLS. But they consume PIM Join/Prune messages and do not flood them as is in the VPLS.
2. PIM Snooping switches do not originate any PIM packets.	2. PIM Proxy switches may originate new or modified Join/Prune packets.

Other than the above simple differences, most of the procedures are common to PIM Snooping and PIM Proxy. In the text to follow, we describe the procedures for PIM "Snooping". Unless specifically stated otherwise, such procedures apply to PIM Proxy as well.

2.5.2. PIM Control Message Latency

A PIM Snooping or PIM Proxy switch snoops on PIM Hello packets while transparently flooding it 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 Proxy switch consumes PIM Join/Prune packets and generates proxy Join/Prune packets to be sent upstream. 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.

A PIM Snooping switch 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.

2.5.3. When to Snoop and When to Proxy

Explicit Tracking (ET) is enabled in a VPLS when all PIM CE Routers in the VPLS advertise Tracking Support in their PIM Hello messages. If even one does not advertise Tracking Support, then all PIM CE routers disable ET in the VPLS. When ET is enabled, it implies that Join Suppression is disabled and vice versa.

PIM Snooping PEs can determine if ET is enabled or disabled in a VPLS by examining PIM Hellos. If ET is disabled, PIM Proxy MUST be used. If ET is enabled, either PIM Snooping or PIM Proxy can be used, unless the PIM control message latency due to proxy is a concern, in which case PIM Snooping SHOULD be used.

3. PIM Snooping for VPLS

IGMP snooping procedures described in [IGMP-SNOOP] provide efficient delivery of IP multicast traffic in a given VPLS service when end stations are connected to the VPLS. However, when VPLS is offered as a WAN service it is likely that the CE devices are routers and would run PIM between them. To provide efficient IP multicasting in such cases, it is necessary that the PE routers offering the VPLS service do PIM snooping.

PIM is a multicast routing protocol, which runs exclusively between routers. 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). On the other hand, PIM does not participate in any kind of exchange of databases, as 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. PIM-DM constructs a separate delivery tree for each multicast group. As opposed to PIM-DM, other PIM flavors (PIM-SM [PIM-SM], PIM-SSM [PIM-SSM], and PIM-BIDIR [PIM-BIDIR]) invoke a pull methodology 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).

In PIM snooping, a PE snoops on the PIM message exchanged between routers, and builds its multicast states.

Based on the multicast states, it forwards IP multicast traffic accordingly to avoid unnecessary flooding.

In the following sub-sections, snooping mechanisms for each variety

of PIM are specified.

3.1. General Rules for PIM Snooping in VPLS

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

- 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 3.1.2).

3.1.1. Snooping PIM Packets

PIM-SM snooping PEs need to snoop on just the PIM Hello and PIM Join/Prune messages to build its multicast states.

- o PIM-DM snooping PEs have to also snoop on PIM Graft and PIM State Refresh messages.

3.1.2. 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 3.6.3).

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).

3.2. 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 switch 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.

3.3. 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.

3.3.1. 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 switch 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 switches have to maintain PPT and ET per (Port, S, G, N). The reasons for this are explained in section Section 3.3.2

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 Section 3.3.6.1.

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 Section 3.9 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(*,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 switches, we specify the details of the downstream state machine of PIM-SM snooping switches at the risk of repeating most of the text documented in [PIM-SM].

3.3.2. 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 switch 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).

3.3.3. 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 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(*,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 PPTExpiry(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).

3.3.4. 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 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(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 PPTEpiry(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).

3.3.5. 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].

3.3.6. Sending Join/Prune Messages Upstream

This section applies only to a PIM Proxy Switch and not to a PIM Snooping Switch.

A PIM Proxy 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 switch 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.

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.

3.3.6.1. Where to send Join/Prune messages

The following rules apply, to both refresh and triggered (S,G)/(*,G) Join/Prune messages.

- o The upstream neighbor field N 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 set of downstream PEs), then they'll receive from each other, triggering assert.

3.4. 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.

3.5. Interaction with IGMP Snooping

Whenever IGMP Snooping is enabled in conjunction with PIM Snooping in the same VPLS instance the switch SHOULD follow these rules:

- o To maintain the list of multicast routers and ports on which they are attached, the switch SHOULD NOT use the rules as described in section 2.1.1.(1) of 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 switch supports proxy-reporting, as described in section 2.1.1.(2) of RFC4541 [IGMP-SNOOP], all IGMP membership information learned on a port to which a PIM neighbor is attached SHOULD NOT be included in the summarized upstream report.

3.6. PIM-DM

The characteristics of PIM-DM is flood and prune behavior. Shortest path trees are built as a multicast source starts transmitting.

3.6.1. 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").

3.6.2. 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 Section 3.6.1, the states (DownstreamPState) and timers (PPT and PT) are per (S,G,N,P).

3.6.3. 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.

3.7. 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 for VPLS Multicast to work correctly.

A PIM Proxy switch behaves like a PIM Router by doing most of the functionality of a PIM Router. The complexity however is much lesser on a switch since many of the issues that a PIM Router has to deal with are not relevant on a switch. A PIM Router needs to be able to build and maintain RP-Sets. They also have to deal with the Register and Assert State Machines. There are other complexities for a PIM Router resulting from inter-domain multicast. A PIM Snooping or PIM Proxy switch can be agnostic of all of this. All that a PIM Proxy switch cares about is building multicast states using PIM Hellos and PIM Join/Prune message. As such it's complexity is greatly reduced.

Other than the procedures defined here, the rest of the procedures that apply to PIM Snooping apply to PIM Proxy as well.

3.7.1. Downstream PIM Proxy behavior

Only PIM Join/Prune messages are proxied. Hellos MUST be snooped while being flooded in the VPLS. i.e. PIM Hellos MUST NOT be consumed at a PE and regenerated.

All other PIM packet types are flooded in the VPLS without any processing.

Performing only proxy of Join/Prune messages keeps the switch 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.

The other option for a PIM Proxy solution is to proxy both Hello and Join/Prune messages that a PE is interested in building states for. If Hellos are being proxied, then it becomes necessary that the PE proxy all other PIM packet types also. Because if Hellos are received after other packet types are received at a CE router, then bad things will happen. That means every PIM packet has to be sent along the slow-path. This greatly increases the complexity on the CE router, it is very compute intensive and does not scale well. Also, proxying Hellos will result in added latency to delivery of Hello messages to a CE and that affects multicast convergence in the VPLS.

3.7.2. Upstream PIM Proxy behavior

Since a PIM Proxy switch consumes Join/Prune messages, it must also originate PIM Join/Prune messages to be sent upstream. On ACs, both triggered and refresh Join/Prunes are forwarded as PIM packets.

3.7.3. Source IP Address in Proxy PIM Join/Prune Packets

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.

If Explicit Tracking (ET) is disabled in the VPLS, then it does not matter what Source IP Address is picked in the packets sent upstream as long as we adhere to the rule in the previous paragraph.

If ET is enabled, it means that a CE router is interested in tracking every CE that wishes to join a stream. If a PE determines that ET is enabled, then it SHOULD use PIM Snooping procedures instead of PIM Proxy.

3.8. 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 routers must appear behind a router.

3.9. 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.

3.9.1. 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 Section 3.9

3.9.2. 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 Section 3.9

[PIM-BIDIR] specifies how olist(G) is constructed.

3.9.3. 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 3.9

[PIM-DM] specifies how olist(S,G) is constructed.

4. IANA Considerations

This document makes no request of IANA.

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

5. Security Considerations

Security considerations provided in VPLS solution documents (i.e., [VPLS-LDP] and [VPLS-BGP]) apply to this document as well.

6. Contributors

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.

Yetik Serbest, Ray Qiu, Suresh Boddapati co-authored earlier versions.

7. 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).

8. References

8.1. 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.

8.2. Informative References

- [IGMP-SNOOP] Christensen, M., Kimball, K., and F. Solensky, "Considerations for IGMP and MLD Snooping Switches", 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-10, 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

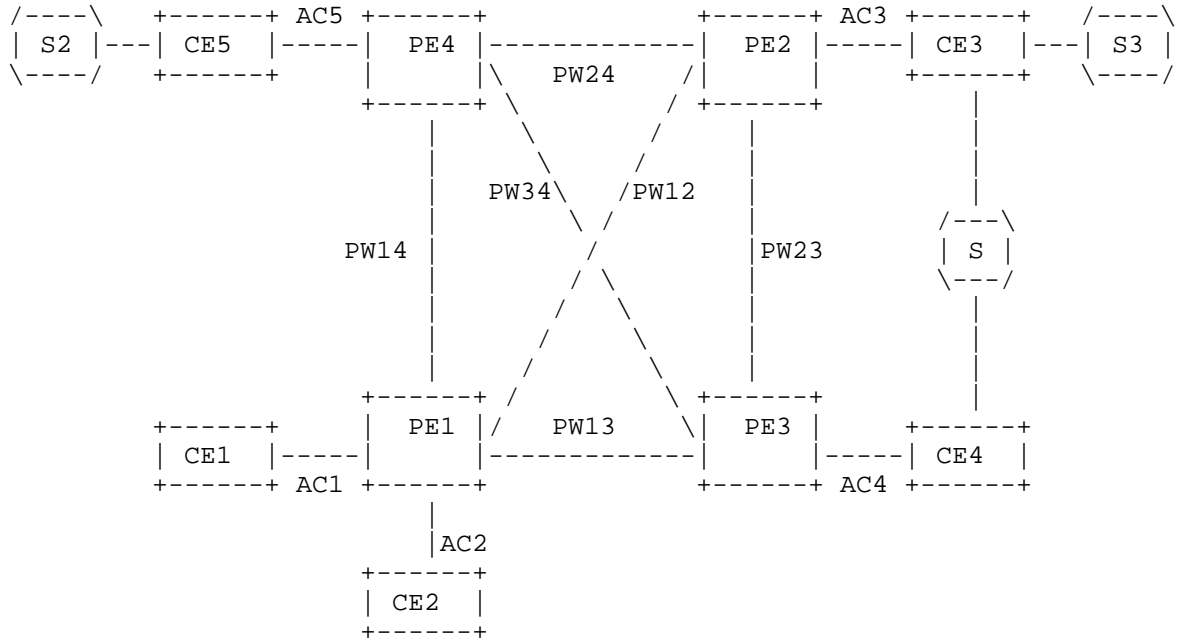
The information needed to setup these states may be obtained by :

- 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 $UpstreamNeighbor(S, G) = CE3$.
2. $PE1$ snoops on the $Join(S, G)$ 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.

1. CE1 Sends a Join(S,G) with Upstream Neighbor(S,G) = CE3.
2. PE1 consumes the Join(S,G). PE1 looks up the neighbor database and determines CE3 was learnt on PW12. PE1 sends a Proxy Join(S,G) to the resulting UpstreamPorts(G). i.e. it sends the proxy Join(S,G) on PW12.
3. Likewise, PE2 consumes the Join(S,G) and sends a proxy Join(S,G) on AC3 with Upstream Neighbor = CE3.

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 did not receive any PIM Join(S,G). So it has no (S,G) state.

4. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1.

5. Now let us say CE1 sends a Join(*,G) towards CE4.

6. PE1 consumes the Join(*,G). PE1 sends a Proxy Join(*,G) to the resulting UpstreamPorts(G). Since UpstreamPorts(G) now has both PW12 and PW13, the Join(*,G) gets sent on both PW12 and PW13.

Note that the UpstreamPorts(S,G) and OutgoingPortList(S,G) inherit the corresponding (*,G) sets, but not vice versa.

remove "but not vice versa"

COMMENT : > Original "but not vice versa" applies to OutgoingPortList(S,G) only, I assume, because of the earlier definition:

UpstreamPorts(G): This set is the union of all the UpstreamPorts(S,G) and UpstreamPorts(*,G) for a given G

7. PE2 and PE3 perform a similar function. PE2 received the Join(*,G) on a PW and the Upstream Neighbor is also on a PW. Hence PE2 only adds UpstreamPorts(*,G) to OutgoingPortList(*,G) and not the downstream port PW12.

At PE1:

```
JT(AC1,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.
UpstreamNeighbors(*,G)   = { CE4 }
UpstreamPorts(*,G)       = { PW13 }
OutgoingPortList(*,G)    = { AC1, PW13 }
```

```
UpstreamPorts(G)         = { PW12, PW13 }
```

At PE2:

```
JT(PW12,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 }
```

At PE3:

```
JT(PW13,*,G,CE4) = JP_HoldTime
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)    = { AC4 }
OutgoingPortList(*,G) = { PW13, AC4 }
```

8. 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 forwards the Prune(S,G,rpt) to both PW12 and PW13. PE2 consumes the Prune(S,G,rpt) and updates its states. PE3 updates its states and forwards the Prune(S,G,rpt) on AC4.

At PE1:

```
JT(AC1,S,G,CE3)      = active
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { PW12 }
OutgoingPortList(S,G) = { AC1, PW12 }

JT(AC1,*,G,CE4)      = active.
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)    = { PW13 }
OutgoingPortList(*,G) = { AC1, PW13 }
```

At PE2:

```
JT(PW12,S,G,CE3)      = active
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(*,G)    = { AC3 }
OutgoingPortList(S,G) = { PW12, AC3 }

JT(PW12,*,G,CE4)      = JP_HoldTime
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)    = { PW23 }
OutgoingPortList(*,G) = { PW23 }
```

At PE3:

```
JT(PW13,*,G,CE4)      = JP_HoldTime
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.

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

Jeffrey Zhang
Juniper Networks, Inc.
10 Technology Park Drive
Westford, MA 01886

Email: zzhang@juniper.net

Venu Hemige
Alcatel-Lucent
701 East Middlefield Rd.
Mountain View, CA 94043

Email: Venu.hemige@alcatel-lucent.com

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: January 30, 2013

Parag Jain, Ed.
Sami Boutros
Samer Salam
Cisco Systems

July 9, 2012

LSP-Ping Mechanisms for E-VPN and PBB-EVPN
draft-jain-l2vpn-evpn-lsp-ping-00.txt

Abstract

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

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), 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>

Copyright Notice

Copyright (c) 2012 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 described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Terminology	3
4. Proposed Target FEC Stack Sub-TLVs	4
4.1. E-VPN MAC Sub-TLV	4
4.2. E-VPN Inclusive Multicast Sub-TLV	5
4.3. E-VPN Auto-Discovery Sub-TLV	6
5. Operations	6
5.1. Unicast Data-plane connectivity checks	6
5.2. Inclusive Multicast Data-plane Connectivity Checks	8
5.2.1. Ingress Replication	8
5.2.2. Using P2MP P-tree	9
5.2.3. Controlling Echo Responses when using P2MP P-tree	10
5.3. E-VPN Aliasing Data-plane connectivity check	10
6. Security Considerations	10
7. IANA Considerations	10
8. References	11
8.1. Normative References	11
8.2. Informative References	11
9. Acknowledgments	12

1. Introduction

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

[PBBEVPN] describes the use of Provider Backbone Bridging [802.1ah] with E-VPN. PBB-EVPN maintains the C-MAC learning in data plane and only advertises Provider Backbone MAC (B-MAC) addresses in control plane using BGP.

Procedures for simple and efficient mechanisms to detect data-plane failures using LSP Ping in MPLS network are well defined in [RFC4379][RFC6425]. This document defines procedures to detect data-plane failures using LSP Ping in MPLS networks deploying E-VPN and PBB-EVPN. This draft defines 3 new Sub-TLVs for Target FEC Stack TLV with the purpose of identifying the FEC on the Peer PE.

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].

The term FEC-Type is used to refer to a tuple consisting of <FEC Element Type, Address Family>.

3. Terminology

B-MAC: Backbone MAC Address

CE: Customer Edge Device

C-MAC: Customer MAC Address

DF: Designated Forwarder

ESI: Ethernet Segment Identifier

EVI: E-VPN Instance

E-VPN: Ethernet Virtual Private Network

MPLS-OAM: MPLS Operations, Administration and Maintenance

P2MP: Point-to-Multipoint

PBB: Provider Backbone Bridge

PE: Provider Edge Device

4. Proposed Target FEC Stack Sub-TLVs

This document introduces three new Target FEC Stack sub-TLVs that are included in the LSP-Ping Echo Request packet sent for detecting faults in data-plane connectivity in E-VPN and PBB-EVPN networks. These Target FEC Stack sub-TLVs are described next.

4.1. E-VPN MAC Sub-TLV

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

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

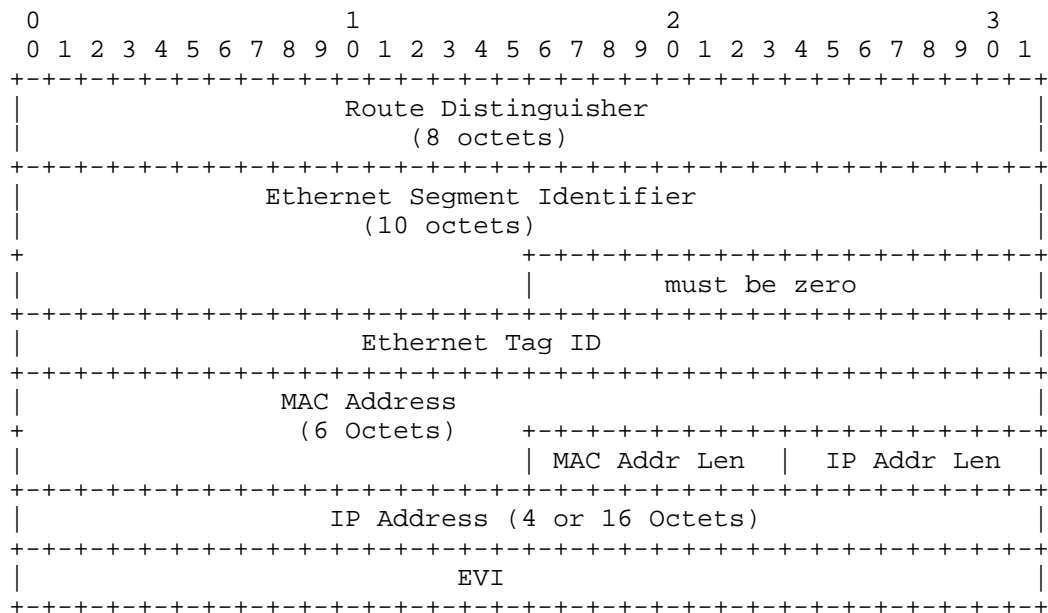


Figure 1: E-VPN MAC sub-TLV format

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

4.2. E-VPN Inclusive Multicast Sub-TLV

The E-VPN Inclusive Multicast sub-TLV fields are based on the E-VPN Inclusive Multicast route defined in [EVPN].

The E-VPN Inclusive Multicast sub-TLV has the format as shown in Figure 2. This TLV is included in the echo request sent to the E-VPN peer PE by the originator of request to verify the multicast connectivity state on the peer PE(s) in E-VPN and PBB-EVPN.

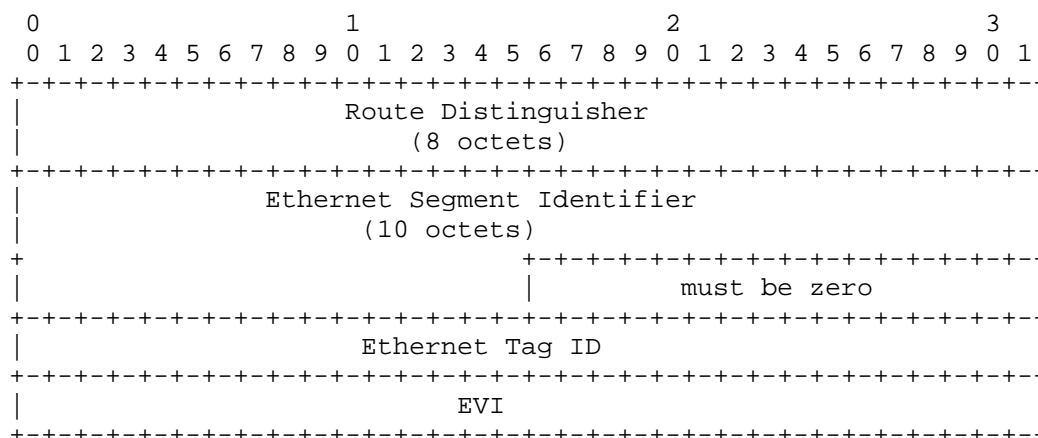


Figure 2: E-VPN Inclusive Multicast sub-TLV format

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

When using P2MP P-tree in E-VPN or PBB-EVPN, the Echo Request is sent using P2MP P-tree transport label for inclusive P-tree arrangement or using a label stack of <P2MP P-tree transport label,

upstream assigned EVPN Inclusive Multicast label> for aggregate inclusive P2MP P-tree arrangement as described in Section 5.

In case of E-VPN, an additional, E-VPN Auto-Discovery sub-TLV and ESI MPLS label as the bottom label, may also be included in the Echo Request as is described in Section 5.

4.3. E-VPN Auto-Discovery Sub-TLV

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

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

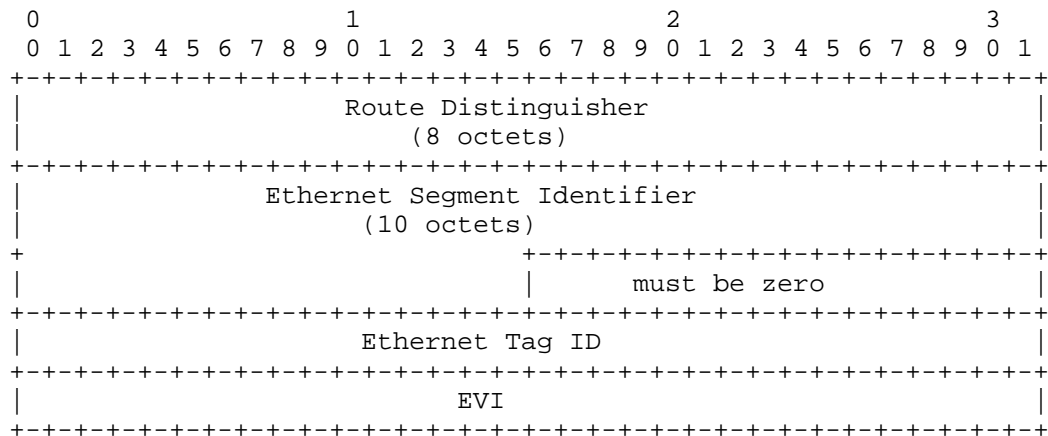


Figure 3: E-VPN Auto-Discovery sub-TLV format

5. Operations

5.1. Unicast Data-plane connectivity checks

Figure 4 is an example of a PBB-EVPN network. CE1 is dual-homed to PE1 and PE2. Assume, PE1 announced a MAC route with RD 1.1.1.1:00 and B-MAC 00aa.00bb.00cc and with MPLS label 16001 for EVI 10. Similarly PE2 announced a MAC route with RD 2.2.2.2:00 and B-MAC 00aa.00bb.00cc and with MPLS label 16002.

On PE3, when a operator performs a connectivity check for the B-MAC address 00aa.00bb.00cc on PE1, the operator initiates an LSP Ping

request with the target FEC stack TLV containing E-VPN MAC sub-TLV in the Echo Request packet. The Echo Request packet is sent with the {Transport Label(s) to reach PE1 + E-VPN Label = 16001} MPLS label stack. Once the echo request packet reaches PE1, it will process the packet and perform checks for the E-VPN MAC sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC4379] and respond according to [RFC4379] processing rules.

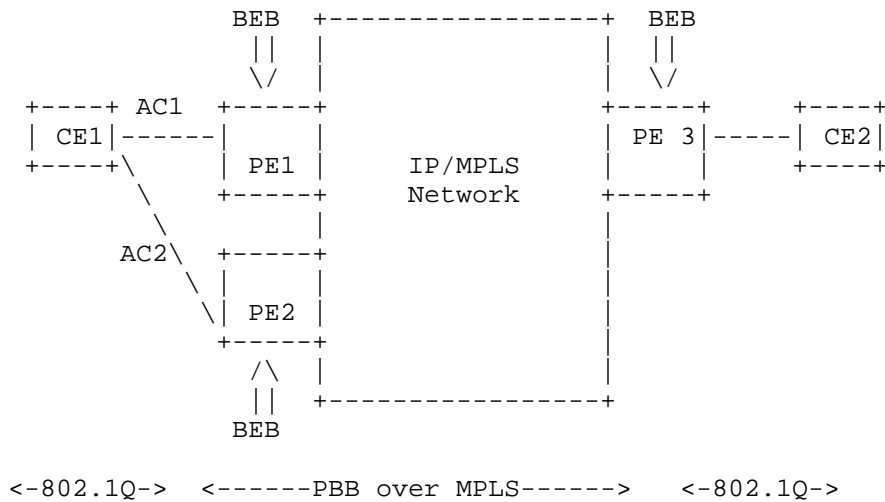


Figure 4: PBB EVPN network

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

LSP Ping operation for unicast data-plane connectivity checks in E-VPN, are similar to as described above for PBB-EVPN except that the checks are for C-MACs and not for B-MACs.

5.2. Inclusive Multicast Data-plane Connectivity Checks

5.2.1. Ingress Replication

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

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

When an operator at PE3 initiates a connectivity check for the inclusive multicast on PE1, the operator initiates an LSP Ping request with the target FEC stack TLV containing E-VPN Inclusive Multicast sub-TLV in the Echo Request packet. The Echo Request packet is sent with the {Transport Label(s) to reach PE1 + E-VPN Incl. Multicast Label = 17001} MPLS label stack. Once the packet reaches PE1, the packet will have E-VPN Inclusive multicast label. PE1 will process the packet and perform checks for the E-VPN Inclusive Multicast sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC4379] and respond according to [RFC4379] processing rules.

Operator at PE3, may similarly also initiate an LSP Ping to PE2 with the target FEC stack TLV containing E-VPN Inclusive Multicast sub-TLV in the echo request packet. The echo request packet is sent with the {transport Label(s) to reach PE2 + E-VPN Incl. Multicast Label = 17002} MPLS label stack. Since PE2 is not the DF for ISID 10 for the port corresponding to the ESI value in the Inclusive Multicast sub-TLV in the Echo Request, PE2 will reply with special code indicating that FEC exists on the router and the behavior is to drop the packet because of not DF as described in Section 7.

In case of E-VPN, in the Echo Request packet, an Ethernet AD sub-TLV and the associated MPLS Split Horizon Label at the bottom of the MPLS label stack, may be added to emulate traffic coming from a MH site, this label is used by leaf PE(s) attached to the same MH site not to forward packets back to the MH site. If the behavior on a leaf PE is to drop the packet because of Split Horizon filtering, the PE2 will reply with special code indicating that FEC exists on the router and the behavior is to drop the packet because of Split Horizon Filtering as described in Section 7.

5.2.2. Using P2MP P-tree

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

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

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

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

For an Aggregate Inclusive P-tree arrangement, when an operator performs a connectivity check for the multicast L2 service, the operator initiates an LSP Ping request with the target FEC stack TLV containing E-VPN Inclusive Multicast sub-TLV in the echo request packet. The echo request packet is sent with the {P2MP P-tree label + E-VPN Upstream assigned Multicast Label} MPLS label stack.

The Leaf PE(s) of the p2mp tree will process the packet and perform checks for the E-VPN Inclusive Multicast sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC4379] and respond according to [RFC4379] processing rules. A PE that is not the DF for the EVI on the ESI in the Inclusive Multicast sub-TLV, will reply with a special code indicating that FEC exists on the router and the behavior is to drop the packet because of not DF as described in Section 7.

In case of E-VPN, in the Echo Request packet, an Ethernet AD sub-TLV and the associated MPLS Split Horizon Label at the bottom of the MPLS label stack, may be added to emulate traffic coming from a MH site, this label is used by leaf PE(s) attached to the same MH site not to forward packets back to the MH site. If the behavior on a leaf PE is to drop the packet because of Split Horizon filtering, the PE2 will reply with special code indicating that FEC exists on the router and the behavior is to drop the packet because of Split Horizon Filtering as described in Section 7.

5.2.3. Controlling Echo Responses when using P2MP P-tree

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

5.3. E-VPN Aliasing Data-plane connectivity check

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

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

When PE1 receives the packet it will process the packet and perform checks for the E-VPN Ethernet AD sub-TLV present in the Target FEC Stack TLV as described in Section 4.4 in [RFC4379] and respond according to [RFC4379] processing rules.

6. Security Considerations

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

7. IANA Considerations

This document defines 3 new sub-TLV type to be included in Target FEC Stack TLV (TLV Type 1) [RFC4379] in LSP Ping.

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

- o E-VPN MAC route sub-TLV.
- o E-VPN Inclusive Multicast route sub-TLV
- o E-VPN Auto-Discovery Route sub-TLV

Proposed new Return Codes

[RFC4379] defines values for the Return Code field of Echo Reply. This document proposes two new Return Codes, which SHOULD be included in the Echo Reply message by a PE in response to LSP Ping Echo Request message:

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

8. References

8.1. Normative References

- [EVPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt, work in progress, February 2012.
- [PBBEVPN] Sajassi et al., "PBB E-VPN", draft-ietf-l2vpn-pbb-evpn-03.txt, work in progress, March 2012.
- [RFC4379] K. Kompella, G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC6425] Saxena, S et al, Detecting Data Plane Failures in Point-to-Multipoint Multiprotocol Label Switching (MPLS) - Extensions to LSP. RFC 6425, November 2011.

8.2. Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.
- [RFC5085] T. Nadeau, et. al, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires ", RFC 5085, December 2007.

- [RFC6388] Minei, I., Kompella, K., Wijnands, I., and Thomas, B.,
"LDP Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths, RFC 6388, November 2011.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and Yasukawa, S.,
"Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

9. Acknowledgments

The authors would like to thank Patrice Brissette for his valuable input and comments.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Parag Jain
Cisco Systems, Inc.,
2000 Innovation Drive,
Kanata, ON K2K3E8, Canada.
E-mail: paragj@cisco.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way,
San Jose, CA 95134, USA.
E-mail: sboutros@cisco.com

Samer Salam
Cisco Systems, Inc.
595 Burrard Street, Suite 2123,
Vancouver, BC V7X 1J1, Canada.
E-mail: ssalam@cisco.com

Internet Working Group

Y. Jiang

Internet Draft

L. Yong
Huawei

Intended status: Standards Track

M. Paul
Deutsche Telekom

F. Jounay
Orange CH

F. Balus
W. Henderickx
Alcatel-Lucent

A. Sajassi
Cisco

Expires: December 2012

June 14, 2012

VPLS PE Model for E-Tree Support
draft-jiang-l2vpn-vpls-pe-etree-06.txt

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 December 14, 2012.

Copyright Notice

Copyright (c) 2012 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.

Abstract

A generic VPLS solution for E-Tree services is proposed which uses VLANs to indicate root/leaf traffic. A VPLS Provider Edge (PE) model is illustrated as an example for the solution. In the solution, E-Tree VPLS PEs are interconnected by PWs which carry the VLAN indicating the E-Tree attribute, the MAC address based Ethernet forwarding engine and the PW work in the same way as before. A signaling mechanism for E-Tree capability and VLAN mapping negotiation is further described.

Table of Contents

1.	Introduction	3
2.	Conventions used in this document	4
3.	Terminology	4
4.	PE Model with E-Tree Support	5
4.1.	Existing PE Models	5
4.2.	A New PE Model with E-Tree Support	8
5.	PW for E-Tree Support	9
5.1.	PW Encapsulation	9
5.2.	VLAN Mapping	9
5.3.	PW Processing	11
5.3.1.	PW Processing in the VLAN Mapping Mode	11
5.3.2.	PW Processing in the Compatible Mode	12
5.3.3.	PW Processing in the Optimized Mode	13
6.	LDP Extensions for E-Tree Support	14
7.	BGP Extensions for E-Tree Support	16
8.	OAM Considerations	17
9.	Applicability	18
10.	Security Considerations	18
11.	IANA Considerations	18
12.	References	19
12.1.	Normative References	19
12.2.	Informative References	19

13. Acknowledgments	20
Appendix A. Other PE Models for E-Tree	21
A.1. A PE Model With a VSI and No bridge	21
A.2. A PE Model With external E-Tree interface	22

1. Introduction

The E-Tree service is defined in Metro Ethernet Forum (MEF) as a Rooted-Multipoint EVC service. It is a multipoint Ethernet service with special restrictions: the frames from a root may be received by any other root or leaf, and the frames from a leaf may be received by any root, but MUST not be received by a leaf. Further, an E-Tree service may include multiple roots and multiple leaves. Although VPMS or P2MP multicast is a somewhat simplified version of this service, in fact, there is no exact corresponding terminology in IETF.

[Etree-req] gives the requirements for providing E-Tree solutions in the VPLS and the need to filter leaf-to-leaf traffic.

[Vpls-etree] describes a PW control word based E-Tree solution, where a bit in the PW control word is used to indicate the root/leaf attribute for a packet. The Ethernet forwarder in the VPLS is also extended to filter the leaf-to-leaf traffic based on the <ingress port, egress port, CW L-bit> tuple.

[Etree-2PW] proposes another E-Tree solution where root and leaf traffic are classified and forwarded in the same VSI but with two separate PWs.

Both solutions are only applicable to "VPLS only" networks.

In fact, VPLS PE usually consists of a bridge module itself (see [RFC4664] and [RFC6246]); moreover, E-Tree services may cross both Ethernet and VPLS domains. Therefore, it is necessary to develop an E-Tree solution both for "VPLS only" scenarios and for interworking between Ethernet and VPLS.

IEEE 802.1 has incorporated the generic E-Tree solution in the latest version of 802.1Q [802.1aq], which is just an improvement on the traditional asymmetric VLAN mechanism the use of different VLANs to indicate E-Tree root/leaf attributes and prohibiting leaf-to-leaf traffic with the help of VLANs was first standardized in IEEE 802.1Q-2003 . In the solution, VLANs are used to indicate root/leaf attribute of a packet: one VLAN ID is used to indicate the frames originated from the roots and another VLAN ID is used to indicate the

frames originated from the leaves. At a leaf port, the bridge can then filter out all the frames from other leaf ports based on the VLAN ID. It is better to reuse the same mechanism in VPLS than to develop a new mechanism. The latter will introduce more complexity to interwork with IEEE 802.1Q solution.

This document introduces how the Ethernet VLAN solution can be used to support generic E-Tree services in the VPLS. The solution proposed here is fully compatible with the IEEE bridge architecture and the IETF PWE3 technology, thus it will not change the FIB (such as installing E-Tree attributes in the FIB), or need any specially tailored implementation. Furthermore, VPLS scalability and simplicity is also well kept. With this mechanism, it is also convenient to deploy a converged E-Tree service across both Ethernet and MPLS networks.

Firstly, a typical VPLS PE model is introduced as an example; the model is then extended in which a Tree VSI is connected to a VLAN bridge with a dual-VLAN interface.

This document then discusses the PW encapsulation and PW processing such as VLAN mapping options for transporting E-Tree services in a VPLS.

Finally, it describes the signaling extensions for E-Tree support and PE processing procedures.

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 [RFC2119].

3. Terminology

E-Tree: a Rooted-Multipoint EVC service as defined in MEF 6.1

EVC: Ethernet Virtual Connection, as defined in MEF 4.0

FIB: Forwarding Information Base, or forwarding table

T-VSI: Tree VSI, a VSI with E-Tree support

Root AC, an AC attached with a root

Leaf AC, an AC attached with a leaf

C-VLAN, Customer VLAN

S-VLAN, Service VLAN

B-VLAN, Backbone VLAN

Root VLAN, a VLAN ID used to indicate all the frames that are originated at a root AC

Leaf VLAN, a VLAN ID used to indicate all the frames that are originated at a leaf AC

I-SID, Backbone Service Instance Identifier, as defined in IEEE 802.1ah

4. PE Model with E-Tree Support

"VPLS only" PE architecture as shown in Fig. 1 of [Etree-req] is a simplification of the VPLS and PWE3 architecture, several common VPLS PE architectures are discussed in more details in [RFC4664] and [RFC6246].

Therefore, VLAN based E-Tree solution are demonstrated with the help of a typical VPLS PE model. It can also be used by other PE models which are discussed in Appendix A.

4.1. Existing PE Models

According to [RFC4664], there are at least three models possible for a VPLS PE, including:

- o A single bridge module, a single VSI;
- o A single bridge module, multiple VSIs;
- o Multiple bridge modules, each attaches to a VSI.

The second PE model is commonly used. A typical example is further depicted in Fig. 1 and Fig. 2 [RFC6246], where an S-VLAN bridge module is connected to multiple VSIs each with a single VLAN virtual interface.

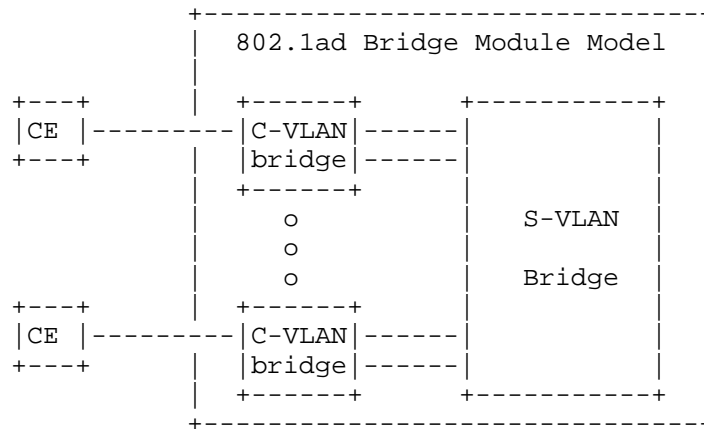


Figure 1 A model of 802.1ad Bridge Module

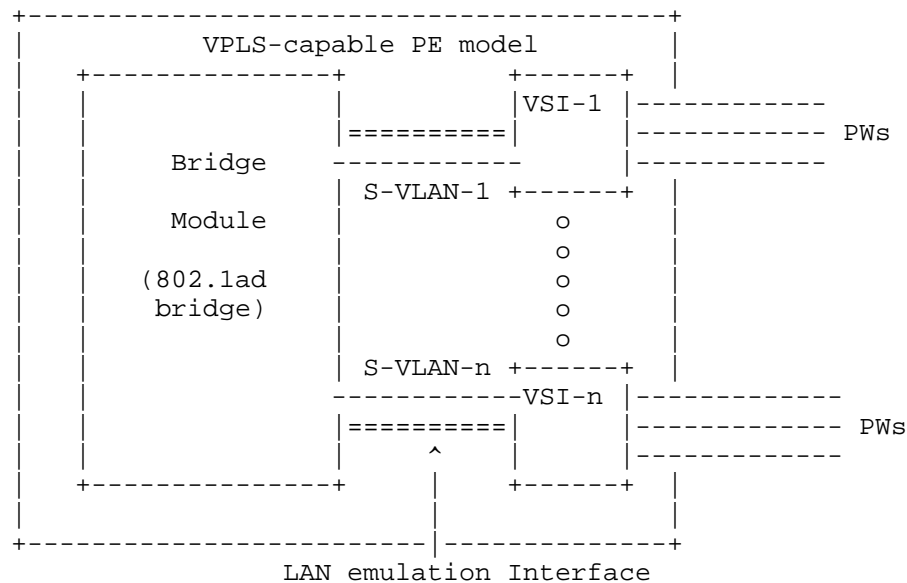


Figure 2 A VPLS-capable PE Model

In this PE model, Ethernet frames from Customer Edges (CEs) will cross multiple stages of bridge modules (i.e., C-VLAN and S-VLAN bridge) and a VSI in a PE before being sent on the PW to a remote PE. Therefore, the association between an AC port and a PW on a VSI as

required in [Vpls-etree] or [Etree-2PW] is difficult, sometimes even impossible.

This model could be further enhanced: When Ethernet frames arrive at a PE, a root VLAN or a leaf VLAN tag is added. Then the frames with the root VLAN tag are transmitted both to the roots and the leaves, while the frames with the leaf VLAN tag are transmitted to the roots but dropped for the leaves (these VLAN tags are removed before the frames are transmitted over the wire). It was demonstrated in [802.1aq] that the E-Tree service in Ethernet networks can be well supported with this mechanism.

Assuming this mechanism is implemented in the bridge module, it is quite straightforward to infer a VPLS PE model with two VSIs to support the E-Tree (as shown in Fig. 3). But this model will require two VSIs per PE and two sets of PWs per E-Tree service, which is poorly scalable in a large MPLS/VPLS network; in addition, both these VSIs have to share their learned MAC addresses.

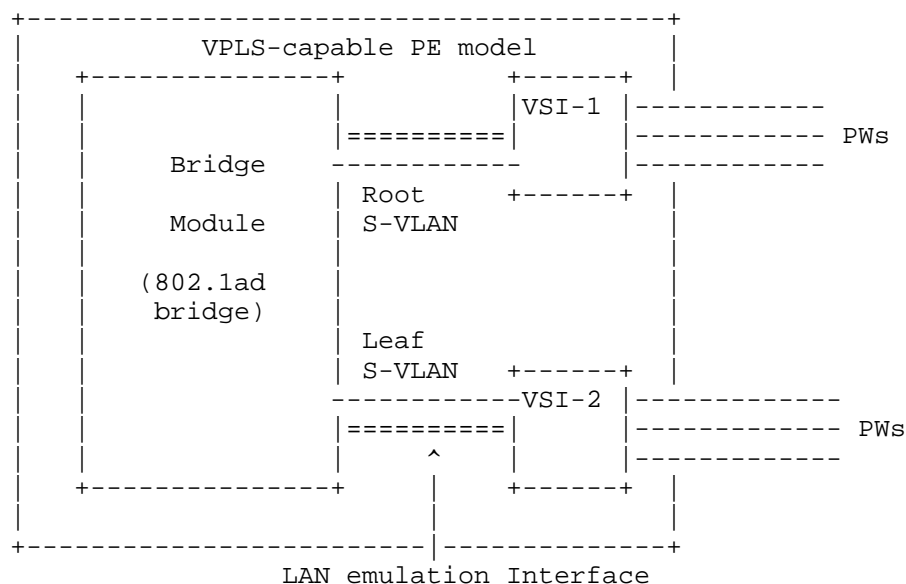


Figure 3 A VPLS PE Model for E-Tree with 2 VSIs

4.2. A New PE Model with E-Tree Support

In order to support the E-Tree in a more scalable way, a new VPLS PE model with a single Tree VSI (T-VSI, a VSI with E-Tree support) is proposed. As depicted in Fig. 4, the bridge module is connected to the T-VSI with a dual-VLAN virtual interface, i.e., both the root VLAN and the leaf VLAN are connected to the same T-VSI, and they share the same FIB and work in shared VLAN learning. In this way, only one VPLS instance and one set of PWs is needed per E-Tree service, and the scalability of VPLS is improved.

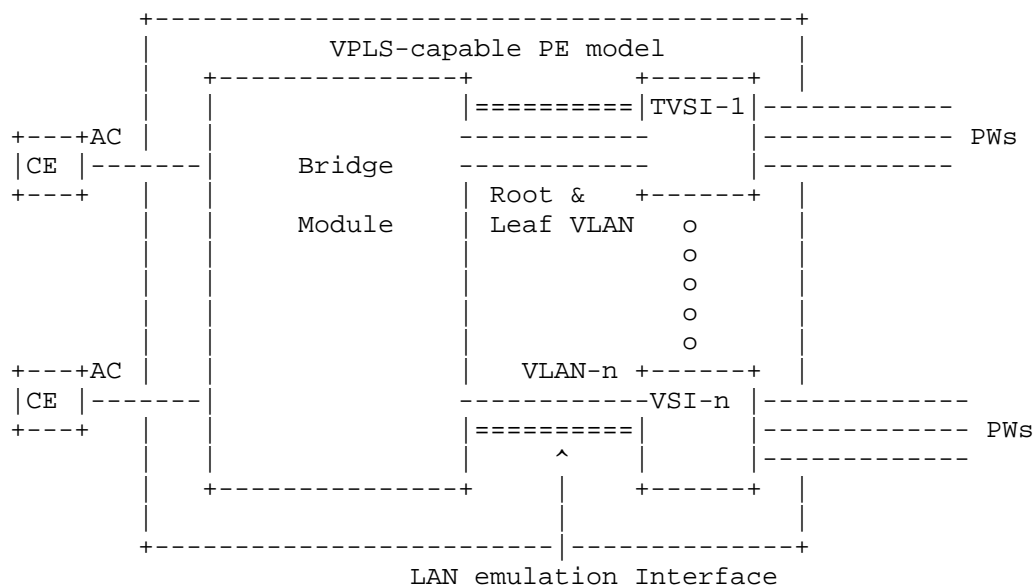


Figure 4 A VPLS PE Model for E-Tree with a Single T-VSI

For an untagged port (customer sites attached to the PEs with untagged ports), the Ethernet frames received from the root ACs can be tagged with a root C-VLAN, and optionally be added with another root S-VLAN. Alternatively, the frames from the root ACs can be tagged with the root S-VLAN tag directly in the VPLS network domain.

For a C-VLAN tagged port, the Ethernet frames received from the root ACs can be added with a root S-VLAN. Alternatively, the C-VLAN can be translated to the root S-VLAN in the VPLS network domain.

For an S-VLAN tagged port, the S-VLAN tag in the Ethernet frames received from the root ACs can be translated to the root S-VLAN in the VPLS network domain. Alternatively, the PBB VPLS PE model (where

an IEEE 802.1ah bridge module is embedded in the PE) as described in [PBB-VPLS] can be used, and a root B-VLAN or leaf B-VLAN can be added in this case (the E-Tree attribute may also be indicated with two I-SID tags in the bridge module, and the frames are further encapsulated and transported transparently over a single B-VLAN, thus the PBB VPLS works just in the same way as described in [PBB-VPLS] and will be discussed no more in this document). When many S-VLANs are multiplexed in a single AC, the 2nd option has an advantage of both VLAN scalability and MAC address scalability.

In a similar way, the traffic from the leaf ACs is tagged and transported on the leaf C-VLAN, S-VLAN or B-VLAN.

In all cases, the outermost VLAN in the resulted Ethernet header is used to indicate the E-Tree attribute of an Ethernet frame; this document will use VLAN to refer to this outermost VLAN for simplicity in the latter sections.

5. PW for E-Tree Support

5.1. PW Encapsulation

To support an E-Tree service, T-VSIs in a VPLS must be interconnected with a bidirectional Ethernet PW. The Ethernet PW may work in the tagged mode (PW type 0x0004) as described in [RFC4448], and a VLAN tag must be carried in each frame in the PW to indicate the frame originated from either root or leaf (the VLAN tag indicating the frame originated from either root or leaf can be translated by a bridge module in the PE or added by an outside Ethernet edge device, even by a customer device). In the tagged PW mode, two service delimiting VLANs must be allocated in the VPLS domain for an E-Tree. PW processing for the tagged PW will be described in Section 5.3 of this document.

Raw PW (PW type 0x0005 in [RFC4448]) may be used to carry E-Tree service for a PW in Compatible mode as shown in Section 5.3.2.

5.2. VLAN Mapping

There are two ways of manipulating VLANs for an E-Tree in VPLS:

- o Global VLAN based, that is, provisioning two global VLANs (Root VLAN, Leaf VLAN) across the VPLS network, thus no VLAN mapping is needed at all, or the VLAN mapping is done completely in the Ethernet domains.

- o Local VLAN based, that is, provisioning two local VLANs for each PE (which participates in the E-Tree) in the VPLS network independently.

The first method requires no VLAN mapping in the PW, but two unique service delimiting VLANs must be allocated across the VPLS domain.

The second method is more scalable in the use of VLANs, but needs a VLAN mapping mechanism in the PW similar to what is already described in Section 4.3 of [RFC4448].

Global or local VLANs can be manually configured or provisioned by an OSS system. Alternatively, some automatic VLAN allocation algorithm may be provided in the management plane, but it is out scope of this document.

For both methods, VLAN mapping parameters from a remote PE can be provisioned or determined by a signaling protocol as described in Section 6 when a PW is being established.

5.3. PW Processing

5.3.1. PW Processing in the VLAN Mapping Mode

In the VLAN Mapping mode, two VPLS PE with E-Tree capability are inter-connected with a PW (For example, the scenario of Fig. 5 depicts the interconnection of two PEs miscellaneously attached with both root and leaf nodes).

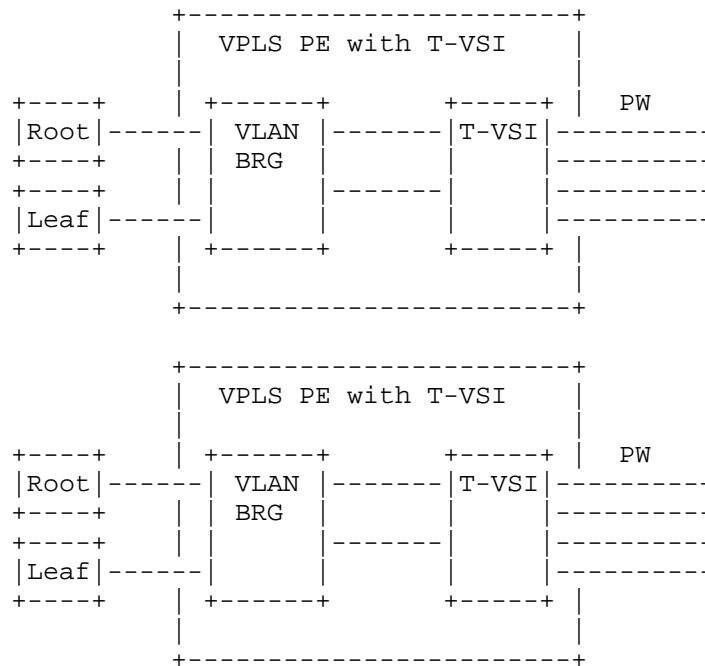


Figure 5 T-VSI Interconnected in the Normal Mode

If a PE is in the VLAN mapping mode for a PW, then in the data plane the PE MUST map the VLAN in each frame as follows:

- o Upon transmitting frames on the PW, map from local VLAN to remote VLAN (i.e., the local leaf VLAN in a frame is translated to the remote leaf VLAN; the local root VLAN in a frame is translated to the remote root VLAN).
- o Upon receiving frames on the PW, map from remote VLAN to local VLAN, and the frames are further forwarded or dropped in the egress bridge module using the filtering mechanism as described in [802.1aq].

5.3.2. PW Processing in the Compatible Mode

The new VPLS PE model can work in a traditional VPLS network seamlessly in the compatibility mode. As shown in Fig. 6, the VPLS PE with T-VSI can be attached with root and/or leaf nodes, while the VPLS PE with a traditional VSI can only be attached with root nodes. Raw PW should be used to connect with a traditional PE.

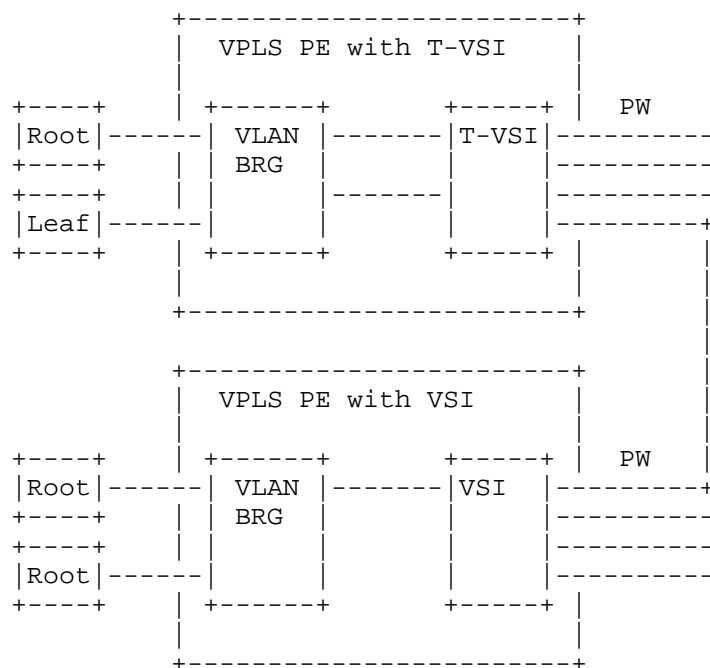


Figure 6 T-VSI interconnected with Traditional VSI

If a PE is in the Compatible mode for a PW, then in the data plane the PE MUST process the frame as follows:

- o Upon transmitting frames on the PW, remove the root or leaf VLAN in the frames.
- o Upon receiving frames on the PW, add a VLAN tag with a value of the local root VLAN to the frames.

5.3.3.PW Processing in the Optimized Mode

When two PEs are connected with their T-VSIs and one PE (e.g., PE2) is attached with only leaf nodes, as shown in the scenario of Fig. 6, the peer PE (e.g., PE1) should then work in the optimization mode. In this case, PE1 should not send the frames originated from the local leaf VLAN to PE2, i.e., these frames are dropped rather than transported over the PW. The bandwidth efficiency of the VPLS can thus be improved. The signaling for the PE attached with only leaf nodes is specified in Section 6.

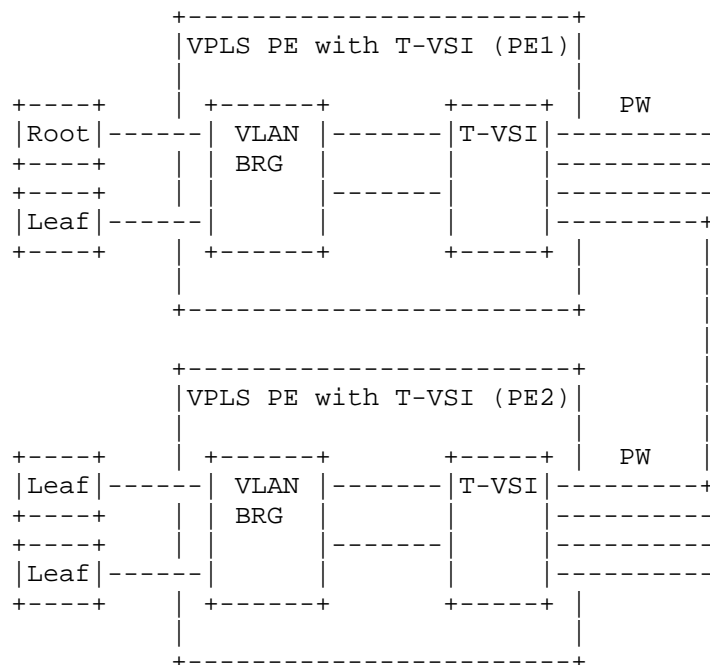


Figure 7 T-VSI interconnected with PE attached with only leaf nodes

If a PE is in the Optimized Mode for a PW, upon transmit, the PE SHOULD first operate as follows:

- o Drop a frame if its VLAN ID matches the local leaf VLAN ID.

6. LDP Extensions for E-Tree Support

In addition to the signaling procedures as specified in [RFC4447], this document proposes a new interface parameter sub-TLV to provision an E-Tree service and negotiate the VLAN mapping function, as follows:

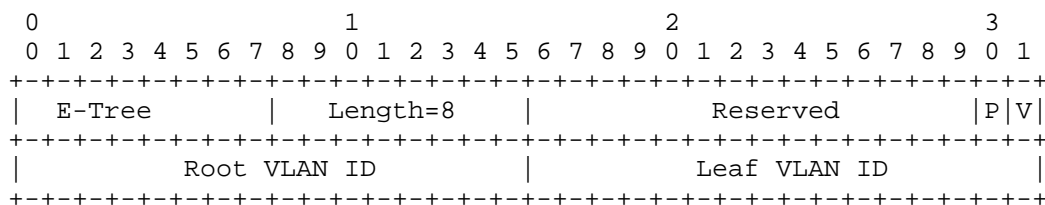


Figure 8 E-Tree Sub-TLV

Where:

- o E-Tree is the sub-TLV identifier to be assigned by IANA.
- o Length is the length of the sub TLV in octets.
- o Reserved bits MUST be set to zero on transmit and be ignored on receive.
- o P is a Leaf-only bit, it is set to 1 to indicate that the PE is attached with only leaf nodes, and set to 0 otherwise.
- o V is a bit indicating the sender's VLAN mapping capability. A PE capable of VLAN mapping MUST set this bit, and clear it otherwise.
- o Root VLAN ID is the value of the local root VLAN.
- o Leaf VLAN ID is the value of the local leaf VLAN.

When setting up a PW for the E-Tree based VPLS, two PEs negotiate the E-Tree support using the above E-Tree sub-TLV. Note PW type of 0x0004 should be used during the PW negotiation.

A PE that wishes to support E-Tree service MUST include an E-Tree Sub-TLV in its PW label mapping message and include its local root VLAN ID and leaf VLAN ID in the TLV. A PE that has the VLAN mapping capability MUST set the V bit to 1, and a PE is attached with only leaf nodes SHOULD set the P bit to 1.

In default, for each PW, VLAN-Mapping-Mode, Compatible-Mode, and Optimized-Mode are all set to FALSE.

A PE that receives a PW label mapping message with an E-Tree Sub-TLV from its peer PE must process it as follows:

- 1) if the root and leaf VLAN ID in the message match the local root and leaf VLAN ID, then continue to 3);
 - 2) else {
 - if the bit V is cleared, then {
 - if the PE is capable of VLAN mapping, then it MUST set VLAN-Mapping-Mode to TRUE;
 - else {

A label release message with the error code "E-Tree VLAN mapping not supported" is sent to the peer PE and exit the process;
 - }
 - if the bit V is set, and the PE is capable of VLAN mapping, then the PE with the minimum IP address MUST set VLAN-Mapping-Mode to TRUE;
- 3) If the P bit is set, then:
 - {
 - If the PE is a leaf-only node itself, then a label release message with a status code "Leaf to Leaf PW released" is sent to the peer PE and exit the process;
 - Else the PE SHOULD set the Optimized-Mode to TRUE.
 - }

If a PE has sent an E-Tree Sub-TLV but does not receive any E-Tree Sub-TLV in its peer's PW label mapping message, The PE SHOULD then

establish a raw PW with this peer as in traditional VPLS and set Compatible-Mode to TRUE for this PW.

Data plane processing for this PW is as following:

If Optimized-Mode is TRUE, then data plane processing as described in Section 5.3.3 applies.

If VLAN-Mapping-Mode is TRUE, then data plane processing as described in Section 5.3.1 applies.

If Compatible-Mode is TRUE, then data plane processing is as described in Section 5.3.2.

PW processing as described in [RFC4448] proceeds as usual for all cases.

7. BGP Extensions for E-Tree Support

A new E-Tree extended community is proposed for E-Tree signaling in BGP VPLS:

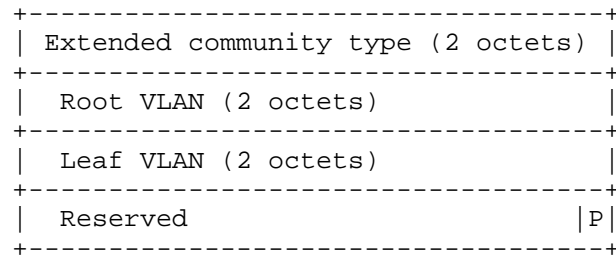


Figure 9 E-Tree Extended Community

Where:

- o Root VLAN ID is the value of the local root VLAN.
- o Leaf VLAN ID is the value of the local leaf VLAN.
- o Reserved, 15 bits MUST be set to zero on transmit and be ignored on receive.
- o P is a Leaf-only bit, it is set to 1 to indicate that the PE is attached with only leaf nodes, and set to 0 otherwise.

The PEs attached with both leaf and root nodes must support BGP E-Tree signaling as described in this document, and must support VLAN mapping in their data planes. The traditional PE attached with only root nodes may also participate in an E-Tree service.

In BGP VPLS signaling, besides attaching a Layer2 Info Extended Community as detailed in [RFC4761], an E-Tree Extended Community MUST be further attached if a PE wishes to participate in an E-Tree service. The PE MUST include its local root VLAN ID and leaf VLAN ID in the E-Tree Extended Community. A PE attached with only leaf nodes of an E-Tree SHOULD set the P bit in the E-Tree Extended Community to 1.

A PE that receives a BGP UPDATE message with an E-Tree Extended Community from its peer PE must process it as follows (after processing procedures as specified in Section 3.2 of [RFC4761]):

- 1) if the root and leaf VLAN ID in the E-Tree Extended Community match the local root and leaf VLAN ID, then continue to 3);
- 2) else {

the PE with the minimum IP address MUST set VLAN-Mapping-Mode to TRUE;

}
- 3) If the P bit is set, then the PE SHOULD set the Optimized-Mode to TRUE.

A PE which does not recognize this attribute shall ignore it silently. If a PE has sent an E-Tree Extended Community but does not receive any E-Tree Extended Community from its peer, the PE SHOULD then establish a raw PW with this peer as in traditional VPLS, and set Compatible-Mode to TRUE for this PW.

Data plane in the VPLS is the same as described in Section 4.2 of [RFC4761], and data plane processing for a PW is the same as described at the end of Section 6.

8. OAM Considerations

VPLS OAM requirements and framework as specified in [RFC6136] are applicable to E-Tree, as both Ethernet OAM frames and data traffic are transported over the same PW.

Ethernet OAM for E-Tree including both service OAM and segment OAM frames shall undergo the same VLAN mapping as the data traffic; and root VLAN SHOULD be applied to segment OAM frames so that they are not filtered.

9. Applicability

The solution is applicable to both LDP VPLS [RFC4762] and BGP VPLS [RFC4761].

The solution is applicable to both "VPLS Only" networks and VPLS with Ethernet aggregation networks.

The solution is also applicable to PBB VPLS networks.

10. Security Considerations

Besides security considerations as described in [RFC4448], [RFC4761] and [RFC4762], this solution prevents leaf to leaf communication in the data plane of VPLS when its PEs are interconnected with PWs. In this regard, security can be enhanced for customers with this solution.

11. IANA Considerations

IANA is requested to allocate a value for E-Tree in the registry of Pseudowire Interface Parameters Sub-TLV type.

Parameter ID	Length	Description
=====		
TBD	8	E-Tree

IANA is requested to allocate two new LDP status codes from the registry of name "STATUS CODE NAME SPACE". The following values are suggested:

Range/Value	E	Description

TBD	1	E-Tree VLAN mapping not supported
TBD	0	Leaf to Leaf PW released

IANA is requested to allocate a value for E-Tree in the registry of BGP Extended Community.

Type	Value	Name
=====		
TBD		E-Tree Info

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4447] Martini, L., and et al, "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4448] Martini, L., and et al, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [RFC4761] Kompella, K. and Rekhter, Y., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007
- [RFC4762] Lasserre, M. and Kompella, V., "Virtual Private LAN Services using LDP", RFC 4762, January 2007.
- [RFC6136] Sajassi, A. and Mohan, D., "L2VPN OAM Requirements and Framework", RFC 6136, March 2011

12.2. Informative References

- [RFC3985] Bryant, S., and Pate, P., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4664] Andersson, L., and Rosen, E., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.
- [RFC6246] Sajassi, A., and et al, "Virtual Private LAN Service (VPLS) Interoperability with Customer Edge (CE) Bridges", RFC 6246, June 2011
- [ETree-req] Key, R., et al, "Requirements for MEF E-Tree Support in VPLS", draft-ietf-l2vpn-etree-reqt-01, Work in Progress
- [Vpls-etree] Key, R., and et al, "Extension to VPLS for E-Tree", draft-key-l2vpn-vpls-etree-06, October 2011

[802.1aq] IEEE 802.1aq D4.3, Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging, September 2011

[Etree-2PW] Ram, R., and et al., Extension to LDP-VPLS for E-Tree Using Two PW, draft-ram-l2vpn-ldp-vpls-etree-2pw-02, May 2011

[PBB-VPLS] Balus, F., and et al., Extensions to VPLS PE model for Provider Backbone Bridging, draft-ietf-l2vpn-pbb-vpls-pe-model-04, October 2011

13. Acknowledgments

The authors would like to thank Adrian Farrel, Susan Hares and Shane Amante for their valuable advices, thank Ben Mack-crane, Edwin Mallette, Donald Fedyk, Dave Allan, Giles Heron, Raymond Key, Josh Rogers, Sam Cao and Daniel Cohn for their valuable comments and discussions.

Appendix A. Other PE Models for E-Tree

A.1. A PE Model With a VSI and No bridge

If there is no bridge module in a PE, the PE may consist of Native Service Processors (NSPs) as shown in Figure A.1 (adapted from Fig. 5 of [RFC3985]) where any transformation operation for VLANs (e.g., VLAN insertion/removal or VLAN mapping) may be applied. Thus a root VLAN or leaf VLAN can be added by the NSP depending on the UNI type (root/leaf) associated with the AC over which the packet arrives.

Further, when a packet with a leaf VLAN exits a forwarder and arrives at the NSP, the NSP must drop the packet if the egress AC is associated with a leaf UNI.

Tagged PW and VLAN mapping work in the same way as in the typical PE model.

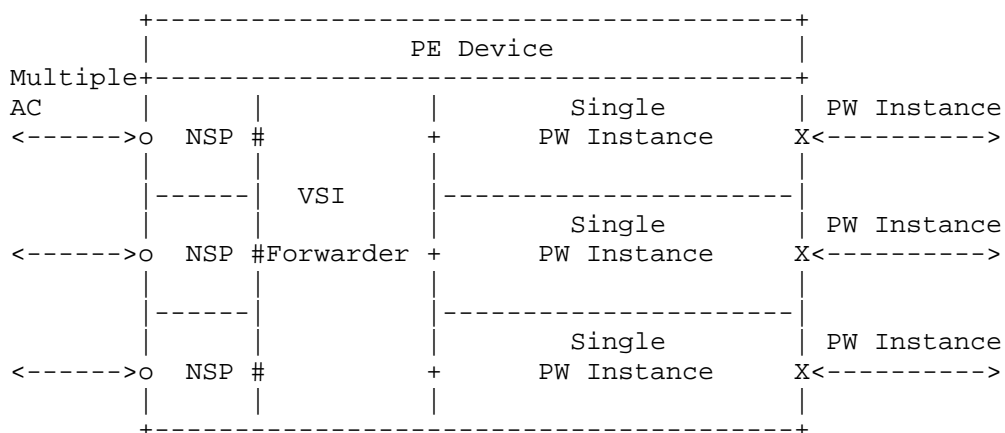


Figure A.1 A PE model with a VSI and no bridge module

This PE model may be used by an MTU-s in an H-VPLS network, or an N-PE in an H-VPLS network with non-bridging edge devices, wherein a spoke PW can be treated as an AC in this model.

A.2. A PE Model With external E-Tree interface

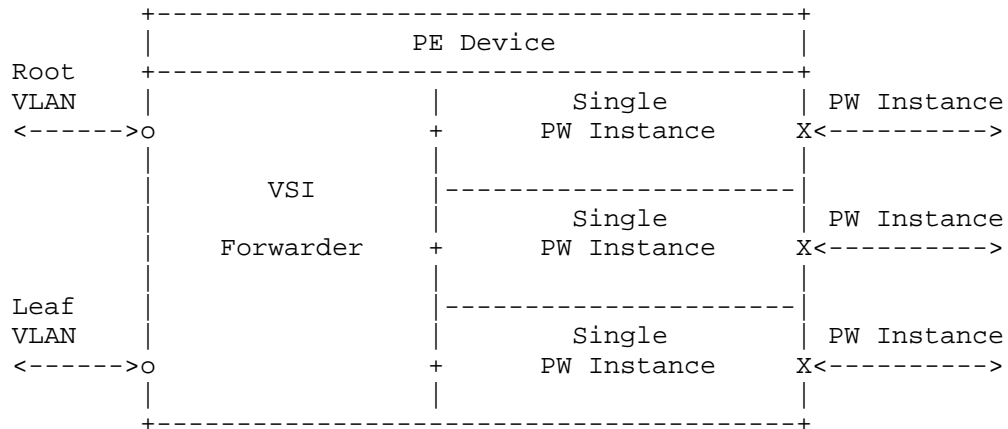


Figure A.2 A PE model with external E-Tree interface

A more simplified PE model is depicted in A.2, where Root/Leaf VLANs are directly or indirectly over a single PW connected to a same VSI forwarder in a PE, any transformation of E-Tree VLANs, e.g., VLAN insertion/removal or VLAN mapping, can be performed by some outer equipments, and the PE may further translate these VLANs into its own local VLANs. This PE model may be used by an N-PE in an H-VPLS network with bridging-capable devices, or scenarios such as providing E-Tree Network-to-Network (NNI) interfaces.

Authors' Addresses

Yuanlong Jiang
Huawei Technologies Co., Ltd.
Bantian, Longgang district
Shenzhen 518129, China
Email: jiangyuanlong@huawei.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075, USA
Email: lucyyong@huawei.com

Manuel Paul
Deutsche Telekom
Winterfeldtstr. 21
10781 Berlin, Germany
Email: manuel.paul@telekom.de

Frederic Jounay
Orange CH
4 rue caudray 1020 Renens, Switzerland
Email: frederic.jounay@orange.ch

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

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp, Belgium
Email: wim.henderickx@alcatel-lucent.com

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

L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Cisco

Expires: December 29, 2012

June 29, 2012

E-TREE Support in E-VPN
draft-sajassi-l2vpn-evpn-etree-00

Abstract

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). [ETREE-FRAMEWORK] proposes a solution framework for supporting this service in MPLS networks. This document discusses how those functional requirements can be easily met with E-VPN.

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) 2012 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	E-Tree Scenarios and E-VPN Support	3
2.1	Scenario 1: Leaf OR Root site(s) per PE	3
2.2	Scenario 2: Leaf AND Root site(s) per PE	5
2.3	Scenario 3: Leaf AND Root site(s) per Ethernet Segment	6
3	Operation	7
3.1	E-Tree with MAC Learning	7
3.2	E-Tree without MAC Learning	8
4	Acknowledgement	8
5	Security Considerations	8
6	IANA Considerations	8
7	References	8
7.1	Normative References	8
7.2	Informative References	8
	Authors' Addresses	9

1 Introduction

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). In an E-Tree service, endpoints are labeled as either Root or Leaf sites. Root sites can communicate with all other sites. Leaf sites can communicate with Root sites but not with other Leaf sites.

[ETREE-FRAMEWORK] proposes the solution framework for supporting E-Tree service in MPLS networks. The document identifies the functional components of the overall solution to emulate E-Tree services in addition to Ethernet LAN (E-LAN) services on an existing MPLS network.

[EVPN] is a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the MPLS/IP network.

This document discusses how the functional requirements for E-Tree service can be easily met with E-VPN.

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].

2 E-Tree Scenarios and E-VPN Support

In this section, we will categorize support for E-Tree into three different scenarios, depending on the nature of the site association (Root/Leaf) per PE or per Ethernet Segment:

- Leaf OR Root site(s) per PE
- Leaf AND Root site(s) per PE
- Leaf AND Root site(s) per Ethernet Segment

For each scenario, we will describe the E-VPN mechanism for supporting the E-Tree service.

2.1 Scenario 1: Leaf OR Root site(s) per PE

In this scenario, a PE may have Root sites OR Leaf sites for a given VPN instance, but not both concurrently. The PE may have both Root and Leaf sites albeit for different VPNs. Every Ethernet Segment

connected to the PE is uniquely identified as either a Root or a Leaf site.

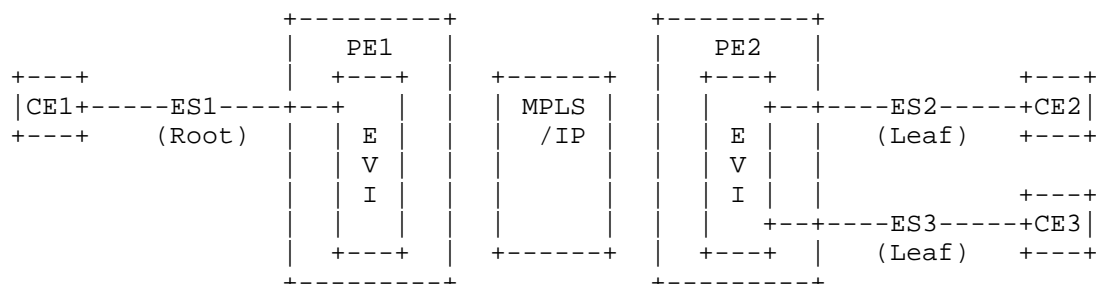
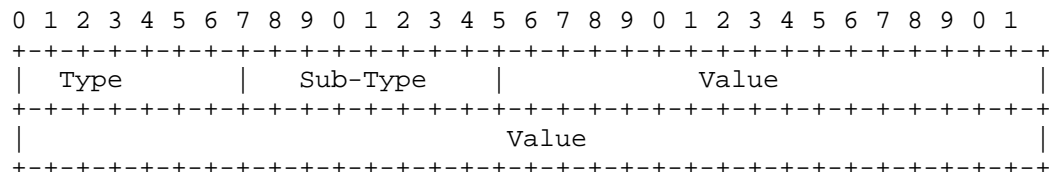


Figure 1: Scenario 1

One approach for addressing this scenario involves associating two BGP Route-Targets (RTs) with every E-VPN Instance (EVI): one RT is associated with the Root sites and the other is associated with the Leaf sites. On a per EVI basis, every PE exports the single RT associated with its type of site(s). Furthermore, a PE with Root site(s) imports both Root and Leaf RTs, whereas a PE with Leaf site(s) only imports the Root RT. This approach suffers from two shortcomings:

- Additional configuration overhead, as it requires the network operator to configure two RTs per EVI.
- Introduces a scalability limitation where only 32K E-Tree EVIs can be supported (due to 2 bytes RT value, and the fact that two RTs are required per EVI).

To alleviate both of these issues, we propose a new BGP Extended Community attribute encoded as follows:



Where,

Type = To be assigned by IANA

Sub-Type = 1 byte, value TBA1 denotes Root, value TBA2 denotes Leaf

Value = 6 bytes uniquely identifying an EVI.

This extended community is a new transitive extended community, and will be referred to as the EVI-Import Extended Community. This extended community is used in lieu of the RT on the following E-VPN routes:

- MAC Advertisement Routes
- Ethernet A-D Routes
- Inclusive Multicast Routes

On a per EVI basis, every PE exports routes with the single EVI-Import extended community associated with its type of site(s). Furthermore, a PE with Root site(s) imports routes with both Root and Leaf EVI-Import extended community. Whereas, a PE with Leaf site(s) only imports the Root EVI-Import extended community.

2.2 Scenario 2: Leaf AND Root site(s) per PE

In this scenario, a PE may have a set of one or more Root sites AND a set of one or more Leaf sites for a given VPN instance. Every Ethernet Segment connected to the PE is uniquely identified as either a Root or a Leaf site.

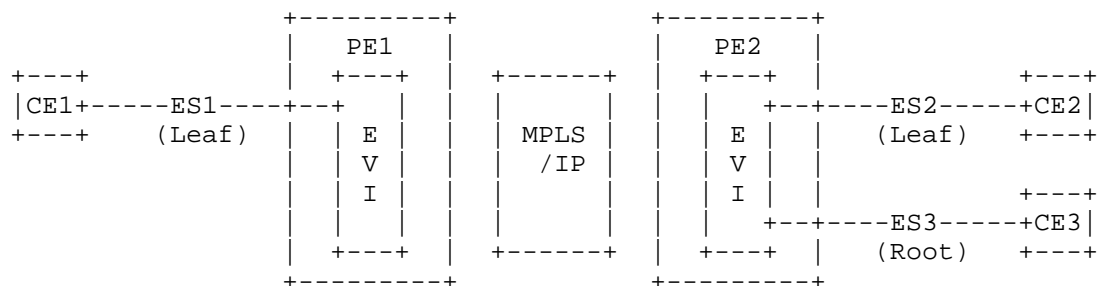


Figure 2: Scenario 2

This scenario requires that the MPLS-encapsulated frames be tagged with an indication of whether they originated from a Root or a Leaf Ethernet Segment, so that the proper connectivity constraints can be enforced. This can be achieved in E-VPN through the use of the ESI MPLS label, since this label identifies the Ethernet Segment of origin of a given frame. For E-Tree service, the ESI MPLS label must be used to encapsulate not only multi-destination frames (i.e. broadcast, multicast & unknown unicast), but also known unicast frames. The egress PE determines whether or not to forward a particular frame to an Ethernet Segment depending on a combination of the split-horizon rule defined in [EVPN] and on the E-Tree

connectivity constraints:

- If the ESI Label indicates that the source Ethernet Segment is a Root, then the frame can be forwarded on a segment granted that it passes the split-horizon check.
- If the ESI Label indicates that the source Ethernet Segment is a Leaf, then the frame can be forwarded only on a Root segment, granted that it passes the split-horizon check.

When advertising the ESI MPLS label for a given Ethernet Segment, a PE must indicate whether the corresponding ESI is a Root or a Leaf site. This can be done by re-purposing one of the Reserved bits in the Flags field of the ESI MPLS label Extended Community attribute ([EVPN] Section 8) to indicate Root/Leaf status.

2.3 Scenario 3: Leaf AND Root site(s) per Ethernet Segment

In this scenario, a PE may have a set of one or more Root sites AND a set of one or more Leaf sites for a given VPN instance. An Ethernet Segment connected to the PE may be identified as both a Root and a Leaf site concurrently.

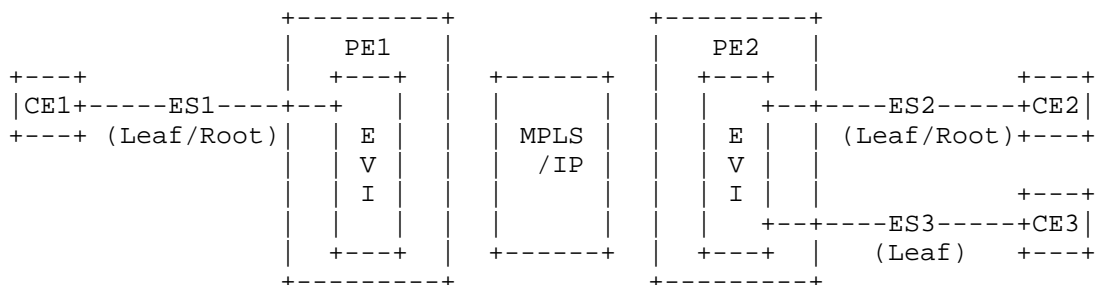


Figure 3: Scenario 3

This scenario can be addressed by extending the use of the ESI MPLS label, as described in the previous section, so that for an Ethernet Segment that has both Root and Leaf sites attached, two ESI MPLS labels are allocated and advertised: one ESI MPLS label denotes Root and the other denotes Leaf. The ingress PE imposes the right ESI MPLS label depending on whether the Ethernet frame originated from the Root or Leaf site on that Ethernet Segment. The mechanism by which the PE identifies whether a given frame originated from a Root or Leaf site on the segment is outside the scope of this document.

In addition to advertising two ESI MPLS labels per Ethernet Segment (for segments that have both Root and Leaf attached), a PE advertises

two special ESI MPLS labels: one for Root and another for Leaf. These are used by remote PEs for traffic originating from single-homed segments and for multi-homed segments that are not connected to the advertising PE.

3 Operation

Per [ETREE-FRAMEWORK], a generic E-Tree service supports all of the following traffic flows:

- Ethernet Unicast from Root to Leaf
- Ethernet Unicast from Leaf to Root
- Ethernet Unicast from Root to Root
- Ethernet Broadcast/Multicast from Root to Roots & Leafs
- Ethernet Broadcast/Multicast from Leaf to Roots

A particular E-Tree service may need to support all of the above types of flows or only a select subset, depending on the target application. In the case where unicast flows need not be supported, the L2VPN PEs can avoid performing any MAC learning function.

In the subsections that follow, we will describe the operation of E-VPN to support E-Tree service with and without MAC learning.

3.1 E-Tree with MAC Learning

The PEs implementing an E-Tree service must perform MAC learning when unicast traffic flows must be supported from Root to Leaf or from Leaf to Root sites. In this case, the PE with Root sites performs MAC learning in the data-path over the Ethernet Segments, and advertises reachability in E-VPN MAC Advertisement routes. These routes will be imported by PEs that have Leaf sites as well as by PEs that have Root sites, in a given EVI. Similarly, the PEs with Leaf sites perform MAC learning in the data-path over their Ethernet Segments, and advertise reachability in E-VPN MAC Advertisement routes which are imported only by PEs with at least one Root site in the EVI. A PE with only Leaf sites will not import these routes. PEs with Root and/or Leaf sites may use the Ethernet A-D routes for aliasing (in the case of multi-homed segments) and for mass MAC withdrawal.

To support multicast/broadcast from Root to Leaf sites, either a P2MP tree rooted at the PE(s) with the Root site(s) or ingress replication can be used. The multicast tunnels are set up through the exchange of the E-VPN Inclusive Multicast route, as defined in [E-VPN].

To support multicast/broadcast from Leaf to Root sites, ingress replication should be sufficient for most scenarios where there is a single Root or few Roots. If the number of Roots is large, a P2MP

tree rooted at the PEs with Leaf sites may be used.

3.2 E-Tree without MAC Learning

The PEs implementing an E-Tree service need not perform MAC learning when the traffic flows between Root and Leaf sites are multicast or broadcast. In this case, the PEs do not exchange E-VPN MAC Advertisement routes. Instead, the Ethernet A-D routes are used to exchange the E-VPN labels.

The fields of the Ethernet A-D route are populated per the procedures defined in [E-VPN], and the route import rules are as described in previous sections.

4 Acknowledgement

We would like to thank Sami Boutros for his comments.

5 Security Considerations

Same security considerations as [E-VPN].

6 IANA Considerations

Allocation of Extended Community Type and Sub-Type for E-VPN.

7 References

7.1 Normative References

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

[ETREE-FRAMEWORK] Key et al., "A Framework for E-Tree Service over MPLS Network", draft-ietf-l2vpn-etree-frwk-00, work in progress, January 2012.

7.2 Informative References

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

[ETREE-REQ] Key et al., "Requirements for MEF E-Tree Support in VPLS", draft-ietf-l2vpn-etree-req-01.txt, work in progress, April, 2012.

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