

INTERNET-DRAFT
Intended Status: Standard Track

Sami Boutros
Ali Sajassi
Samer Salam
Cisco Systems
John Drake
Juniper Networks
Jeff Tantsura
Ericsson
February 24, 2013

Expires: August 28, 2013

VPWS support in E-VPN
draft-boutros-l2vpn-evpn-vpws-01.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: active/standby as well as active/active multi-homing with flow-based load-balancing, eliminates the need for single-segment and multi-segment PW signaling, and provides fast protection using data-plane prefix independent convergence upon node or link failure.

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	BGP Extensions	4
3	Operation	4
4	E-VPN Comparison to PW Signaling	5
5	ESI Bandwidth Attribute	5
6	VPWS with multiple sites	6
7	Security Considerations	6
8	IANA Considerations	6
9	References	6
9.1	Normative References	6
9.2	Informative References	7
	Authors' Addresses	7

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 applies the benefits of E-VPN to p2p services. These benefits include active/standby AC redundancy as well as 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 (AC), aka Ethernet Segment in E-VPN terminology, 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, the traffic forwarding capability of E-VPN using only a pair of Ethernet AD routes is used; whereas, for more general VPWS, traffic forwarding capability of E-VPN using a group of Ethernet AD routes (one Ethernet AD route per AC/segment) is used. Since in VPWS services, the traffic from an originating Ethernet Segment can go only to a single destination Ethernet Segment, no MAC lookup is needed and the MPLS label associated with the per-EVI Ethernet 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 state between the PEs attached to the same multi-homed Segment.

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.

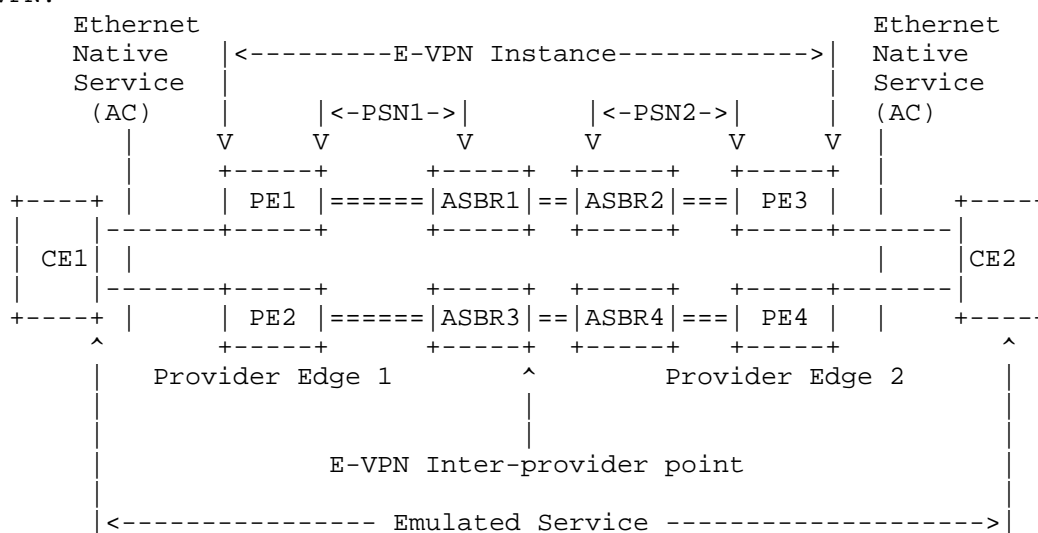
2. BGP Extensions

[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 re-purposes one of the routes as described next.

This document proposes the use of the per EVI 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 either a C-tagged or S-tagged interface, as described in [802.1Q], that can carry a single VLAN tag or two nested VLAN tags. This interface is set up 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 ESI Bandwidth Attribute

The ESI Bandwidth Attribute is a new optional BGP attribute that will be associated with the Ethernet AD route used to realize the EVPL services.

Type	(2 octets)
Length	(2 octets)
Flags	(1 Octet)
Reserved=0	(1 Octet)
Reverse SENDER_TSPEC	

The content of the SENDER_TSPEC are as defined in [RFC 2210] section 3.1.

When a PE receives this attribute for a given EVPL it MUST request the appropriate resources described in the SENDER_TSPEC from the PSN towards the other EVPL service destination PE originating the message. When resources are allocated from the PSN for a given EVPL service, then the PSN SHOULD account for the Bandwidth requested by this EVPL service.

In the case where PSN resources are not available, the PE receiving this attribute MUST re-send its local Ethernet AD routes for this EVPL service with the ESI Bandwidth attribute and with the Flags set to 1 "PSN Resources Unavailable".

6 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.

7 Security Considerations

This document does not introduce any additional security constraints.

8 IANA Considerations

TBD

9 References

9.1 Normative References

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

[RFC 2210] Wroclawski, J. "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997

9.2 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

John Drake
Juniper Networks
Email: jdrake@juniper.net

Jeff Tantsura
Ericsson
Email: jeff.tantsura@ericsson.com

INTERNET-DRAFT
Intended Status: Informational

Sami Boutros
Ali Sajassi
Samer Salam
Dennis Cai
February 24, 2013

Expires: August 28, 2013

VXLAN DCI Using EVPN
draft-boutros-l2vpn-vxlan-evpn-01.txt

Abstract

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

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

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

1 Introduction

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

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

1.1 Terminology

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

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

2. Requirements

2.1. Control Plane Separation among VXLAN/NVGRE Networks

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

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

- scalability of multicast trees as the number of interconnected networks scales.

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

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

2.3 Support for Integrated Routing and Bridging (IRB)

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

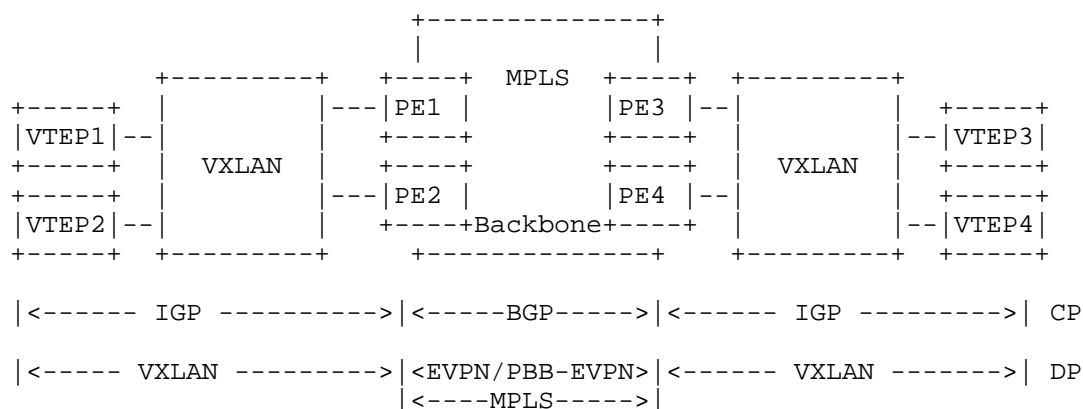
3. Solution Overview

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

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

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

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



Legend: CP = Control Plane View

DP = Data Plane View

Figure 1: Interconnecting VXLAN Networks with VXLAN-EVPN

4. E-VPN Routes This solution leverages the same BGP Routes and Attributes defined in [E-VPN], adapted as follows:

4.1. BGP MAC Advertisement Route

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

4.2. Ethernet Auto-Discovery Route

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

4.3. Per VPN Route Targets

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

4.4 Inclusive Multicast Route

The E-VPN Inclusive Multicast route is used to distribute the VNI information over the MPLS network. This is required to perform the discovery of the PEs participating in a given VNI. It also enables the stitching of the IP multicast trees, which are local to each VXLAN site, with the Label Switched Multicast (LSM) trees of the MPLS network.

The Inclusive Multicast Route is encoded as follow:

- Ethernet Tag ID is set to VXLAN Network Identifier (VNI).
- Originating Router's IP Address is set to one of the PE's IP addresses.

All other fields are set as defined in [E-VPN].

Please see section 4.6 "Handling Multicast"

4.5. Unicast Forwarding

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

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

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

4.6. Handling Multicast

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

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

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

When LSM is used, the trees are P2MP.

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

1. Avoiding Packet Duplication: In the case where the VXLAN network

is multi-homed to multiple PE nodes, if all of the PE nodes forward the same multicast frame, then packet duplication would arise. This applies to both multicast traffic from site to core as well as from core to site.

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

The following approach of per-VNI load balancing can guarantee proper stitching that meets the above requirements.

4.6.2. Multicast Stitching with Per-VNI Load Balancing

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

5. NVGRE

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

6. Acknowledgements

TBD.

7. Security Considerations

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

8. IANA Considerations

TBD.

9. References

9.1 Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2 Informative References

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

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

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

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

Authors' Addresses

Sami Boutros
Cisco Systems

E-Mail: sboutros@cisco.com

Ali Sajassi
Cisco Systems

E-Mail: sajassi@cisco.com

Samer Salam
Cisco Systems

E-Mail: ssalam@cisco.com

Dennis Cai
Cisco Systems

E-Mail: dcai@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: August 25, 2013

February 25, 2013

BGP MPLS Based Ethernet VPN
draft-ietf-l2vpn-evpn-03

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

9.3.1.2. P2MP MPLS LSPs	19
9.3.1.3. MP2MP LSPs	20
9.4 Aliasing and Backup-Path	20
9.4.1 Constructing the Ethernet A-D Route per EVI	21
9.4.1.1 Ethernet A-D Route Targets	22
9.5 Designated Forwarder Election	22
10. Determining Reachability to Unicast MAC Addresses	24
10.1. Local Learning	25
10.2. Remote learning	25
10.2.1. Constructing the BGP E-VPN MAC Address Advertisement	25
10.2.2 Route Resolution	27
11. ARP and ND	28
11.1 Default Gateway	29
12. Handling of Multi-Destination Traffic	29
12.1. Construction of the Inclusive Multicast Ethernet Tag Route	30
12.2. P-Tunnel Identification	30
13. Processing of Unknown Unicast Packets	31
13.1. Ingress Replication	32
13.2. P2MP MPLS LSPs	32
14. Forwarding Unicast Packets	32
14.1. Forwarding packets received from a CE	32
14.2. Forwarding packets received from a remote PE	34
14.2.1. Unknown Unicast Forwarding	34
14.2.2. Known Unicast Forwarding	34
15. Load Balancing of Unicast Frames	34
15.1. Load balancing of traffic from an PE to remote CEs	34
15.1.1 Single-Active Redundancy Mode	34
15.1.2 All-Active Redundancy Mode	35
15.2. Load balancing of traffic between an PE and a local CE	37
15.2.1. Data plane learning	37
15.2.2. Control plane learning	37
16. MAC Mobility	37
16.1. MAC Duplication Issue	39
17. Multicast	39
17.1. Ingress Replication	40
17.2. P2MP LSPs	40
17.3. MP2MP LSPs	40
17.3.1. Inclusive Trees	40
17.3.2. Selective Trees	41
17.4. Explicit Tracking	42
18. Convergence	42
18.1. Transit Link and Node Failures between PEs	42
18.2. PE Failures	42
18.2.1. Local Repair	42
18.3. PE to CE Network Failures	42
19. LACP State Synchronization	43
20. Acknowledgements	44

21. Security Considerations	44
22. IANA Considerations	44
23. References	44
23.1 Normative References	44
23.2 Informative References	45
24. Author's Address	45

1. Specification of requirements

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

2. 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 [EVPN-REQ]. Please refer to [EVPN-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

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

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):

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

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. The following two ESI values are reserved:

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

In general, an Ethernet segment MUST have a non-reserved ESI that is unique network wide (e.g., across all EVPNs on all the PEs). If the

CE(s) constituting an Ethernet Segment is (are) managed by the network operator, then ESI uniqueness should be guaranteed; however, if the CE(s) is (are) not managed, then the operator MUST configure a network-wide unique ESI for that Ethernet Segment. This is required to enable auto-discovery of Ethernet Segments and DF election. The ESI can be assigned using various mechanisms:

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

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

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

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

2. 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.

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

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

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

This mechanism could be used only if it produces ESIs that satisfy

the uniqueness requirement specified above.

4. The ESI may be configured.

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.

7.3.1 Port Based VLAN Aware Service Interface

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

8. BGP 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 25 (L2VPN) and a SAFI of 70 (E-VPN). 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 25 (L2VPN) and a SAFI of 70 (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 (3 octets)	
+-----+		

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

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

8.3. Inclusive Multicast Ethernet Tag Route

An Inclusive Multicast Ethernet Tag route type specific 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 Label Extended Community

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

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

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type=0x06   | Sub-Type=0x01 | Flags (One Octet) | Reserved=0 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Reserved = 0 |               ESI Label               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

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

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

8.6 ES-Import Route Target

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

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type=0x06   | Sub-Type=0x02 |               ES-Import               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               ES-Import Cont'd               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

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

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

8.7 MAC Mobility Extended Community

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

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

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type=0x06      | Sub-Type=0x00 | Reserved=0      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Sequence Number         |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

8.8 Default Gateway Extended Community

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

9. Multi-homing Functions

This section discusses the functions, procedures and associated BGP routes used to support multi-homing in 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 label used for split-horizon filtering (as discussed in section 9.3). Support of this route flavor is MANDATORY.

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

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

The Ethernet Tag ID MUST be set to 0.

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

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

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

9.2.1.1. Ethernet A-D Route Targets

The Ethernet A-D route MUST carry one or more Route Target (RT) attributes. These RTs MUST be the set of RTs associated with all the 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 operating in All-Active mode. If the CE sends a broadcast, unknown unicast, or multicast (BUM) packet to one of the non-DF (Designated Forwarder) PEs, say PE1, then PE1 will forward that packet to all or subset of the other PEs in the EVI including the DF PE for that Ethernet segment. In this case the DF PE that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This filtering is referred to as "split horizon" filtering in

this document.

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

9.3.1 ESI Label Assignment

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

9.3.1.1 Ingress Replication

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

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

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

9.3.1.2. P2MP MPLS LSPs

The non-DF PEs operating in an All-Active multi-homing mode that is using P2MP LSPs for sending BUM traffic, distribute to other PEs, that belong to the Ethernet segment or have an E-VPN in common with the Ethernet Segment, an upstream assigned "ESI label" in the Ethernet A-D route. This label is upstream assigned by the PE that advertises the route. This label MUST be programmed by the other 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 ES1 and operating in All-Active multi-homing mode. Also consider PE3 that is in the same EVI as one of the EVIs to which ES1 belongs. Further, assume that PE1 which is the non-DF, using P2MP MPLS LSPs to send BUM packets. When PE1 sends a BUM packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI that the packet was received on. The resulting packet is further encapsulated in the P2MP MPLS label stack necessary to transmit the packet to the other 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-encapsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1, then PE2 MUST NOT forward the packet onto ES1. When PE3 receives this packet, it de-encapsulates the

top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1 and PE3 is not connected to ES1, then PE3 MUST pop the label and flood the packet over all local ESIs in the EVI. It should be noted that when PE2 sends a BUM frame over a P2MP LSP, it does not need to encapsulate the frame with an ESI label because it is the DF for that VLAN.

9.3.1.3. MP2MP LSPs

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

9.4 Aliasing and Backup-Path

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

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

This flavor of Ethernet A-D route per EVI, associated with aliasing, can arrive at target PEs asynchronously relative to the flavor of Ethernet A-D route associated with split-horizon and mass-withdraw (i.e. per ESI). Therefore, if the Ethernet A-D route per EVI arrives ahead of the Ethernet A-D route per ESI, then the former must NOT be used for traffic forwarding till the latter arrives. This will take

care of corner cases and race conditions where the Ethernet A-D route associated with mass-withdraw is withdrawn but a PE still receives the route associated with aliasing.

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

9.4.1 Constructing the Ethernet A-D Route per EVI

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

Route-Distinguisher (RD) MUST be set to the RD of the EVI that is advertising the NLRI. An RD MUST be assigned for a given EVI on an PE. This RD MUST be unique across all EVIs on an PE. It is RECOMMENDED to use the Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE. This number may be generated by the PE. Or in the Unique VLAN 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 default procedure for DF election at the granularity of <ESI, EVI> is referred to as "service carving". With service carving, it is possible to elect multiple DFs per Ethernet Segment (one per EVI) in order to perform load-balancing of multi-destination traffic destined to a given Segment. The load-balancing procedures carve up the EVI space among the PE nodes evenly, in such a way that every PE is the

DF for a disjoint set of EVIs. The procedure for service carving is as follows:

1. When a PE discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.
2. The PE then starts a timer (default value = 3 seconds) to allow the reception of Ethernet Segment routes from other PE nodes connected to the same Ethernet Segment. This timer value MUST be same across all PEs connected to the same Ethernet Segment.
3. When the timer expires, each PE builds an ordered list of the IP addresses of all the PE nodes connected to the Ethernet Segment (including itself), in increasing numeric value. 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 an EVI with an associated Ethernet Tag value V when $(V \bmod N) = i$. In the case where multiple Ethernet Tags are associated with a single EVI, then the numerically lowest Ethernet Tag value in that EVI MUST be used in the modulo function.
4. The PE that is elected as a DF for a given EVI will unblock traffic for the Ethernet Tags associated with that EVI. 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. In case of a Single-Active multi-homing, when a service moves from one PE in the RG to another PE as a result of re-carving, the PE, which ends up being the elected DF for the service, must trigger a MAC address flush notification towards the associated Ethernet Segment. This can be done, for e.g. using IEEE 802.1ak MVRP 'new' declaration.

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 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 needs to be advertised (e.g., for ARP suppression purposes or for inter-subnet switching), it is then encoded in this route.

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.

10.2.2 Route Resolution

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

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

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, then the receiving PE MUST install forwarding state for a given MAC address only when

both the MAC Advertisement route AND the associated Ethernet A-D route per ESI have been received.

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

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

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

T3- If after T1, PE2 withdraws its Ethernet A-D route per ESI, then PE3 forwards traffic destined to M1 to PE1 only.

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

11. ARP and ND

The IP address field in the MAC advertisement route may optionally carry one of the IP addresses associated with the MAC address. This provides an option which can be used to minimize the flooding of ARP or Neighbor Discovery (ND) messages over the MPLS network and to remote CEs. This option also minimizes ARP (or ND) message processing on end-stations/hosts connected to the 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.

11.1 Default Gateway

A PE MAY choose to terminate ARP messages instead of performing ARP proxy for them. Such scenarios arises when the PE needs to perform inter-subnet forwarding where each subnet is represented by a different bridge domain/EVI. In such scenarios the inter-subnet forwarding is performed at layer 3 and the PE that performs such function is called the default gateway.

Each PE that acts as a default gateway for a given E-VPN advertises in the E-VPN control plane its default gateway IP and MAC address using the MAC advertisement route, and indicates that such route is associated with the default gateway. This is accomplished by requiring the route to carry the Default Gateway extended community defined in [Section 8.8 Default Gateway Extended Community].

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

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

12. Handling of Multi-Destination Traffic

Procedures are required for a given PE to send broadcast or multicast traffic, received from a CE encapsulated in a given Ethernet Tag 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 [RFC4761] and [RFC4762]. 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 Single-Active Redundancy Mode

For a given ESI, if the remote PE has imported an Ethernet A-D route per Ethernet Segment from at least one PE, where the "Active-Standby"

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

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

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

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

15.1.2 All-Active Redundancy Mode

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

-If the next-hop is constructed as a result of a MAC route then this label stack MUST be used. However, if the MAC route doesn't exist, then the next-hop and MPLS label stack is constructed as a result of

the Ethernet A-D routes. Note that the following description applies to determining the label stack for a particular next-hop to reach a given PE, from which the remote PE has received and imported Ethernet A-D routes that have the matching ESI and Ethernet Tag as the one present in the MAC advertisement. The Ethernet A-D routes mentioned in the following description refer to the ones imported from this given PE.

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

The following example explains the above.

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

The MPLS label stack to send the packets to PE1 is the MPLS LSP stack to get to PE1 and the 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 <ES1, VLAN1>, if PE2 has not advertised MAC1 in BGP.

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

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

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

When PE1 or PE2 receive the packet destined for CE1 from PE3, if the packet is a unicast MAC packet it is forwarded to CE1. If it is a

multicast or broadcast MAC packet then only one of PE1 or PE2 must forward the packet to the CE. Which of PE1 or PE2 forward this packet to the CE is determined based on which of the two is the DF.

If the connectivity between the multi-homed CE and one of the PEs that it is attached to fails, the PE MUST withdraw the Ethernet Tag A-D routes, that had been previously advertised, for the Ethernet Segment to the CE. When the MAC entry on the PE ages out, the PE MUST withdraw the MAC address from BGP. Note that to aid convergence, the Ethernet Tag A-D routes MAY be withdrawn before the MAC routes. This enables the remote PEs to remove the MPLS next-hop to this particular PE from the set of MPLS next-hops that can be used to forward traffic to the CE. For further details and procedures on withdrawal of 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 highest of the sequence number(s) in the received MAC Advertisement route(s), if the received route(s) is (are) tagged with a MAC Mobility extended community attribute.

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

16.1. MAC Duplication Issue

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

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 and Reshad Rahman for their reviews. Last but not least, many thanks to Jakob Heitz for his help to improve several sections of this draft.

21. Security Considerations

22. IANA Considerations

23. References

23.1 Normative References

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

23.2 Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-01.txt
- [VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-11.txt
- [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
- [BGP-VPLS-MH] "BGP based Multi-homing in Virtual Private LAN Service", K. Kompella et. al., draft-ietf-l2vpn-vpls-multihoming-04.txt

24. Author's Address

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Rahul Aggarwal
Email: raggarwa_1@yahoo.com

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

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

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

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

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

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

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

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

Samer Salam
Cisco
Email: ssalam@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Cisco

Florin Balus
Wim Henderickx
Alcatel-Lucent

Nabil Bitar
Verizon

Clarence Filsfils
Dennis Cai
Cisco

Aldrin Isaac
Bloomberg

Lizhong Jin
ZTE

Expires: August 25, 2013

February 25, 2013

PBB-EVPN
draft-ietf-l2vpn-pbb-evpn-04

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

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

Conventions

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

Table of Contents

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

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

1. Introduction

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

In this document, we discuss how PBB can be combined with E-VPN in order to: reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes. The combined solution is referred to as PBB-EVPN.

2. Contributors

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

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

3. Terminology

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

4. Requirements

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

4.1. MAC Advertisement Route Scalability

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

4.2. C-MAC Mobility with MAC Summarization

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

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

4.3. C-MAC Address Learning and Confinement

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

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

4.4. Per Site Policy Support

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

4.5. Avoiding C-MAC Address Flushing

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

5. Solution Overview

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

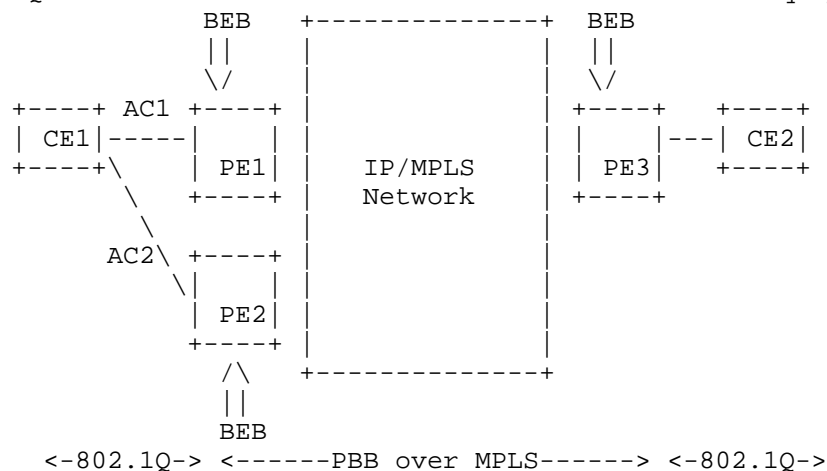


Figure 1: PBB-EVPN Network

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

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

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

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

6. BGP Encoding

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

6.1. BGP MAC Advertisement Route

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

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

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

All other fields are set as defined in [E-VPN].

6.2. Ethernet Auto-Discovery Route

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

The receiving PE knows that it need not wait for the receipt of the Ethernet A-D route for route resolution by means of the reserved ESI encoded in the MAC Advertisement route: the ESI values of 0 and MAX-

ESI indicate that the receiving PE can resolve the path without an Ethernet A-D route.

6.3. Per VPN Route Targets

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

6.4. MAC Mobility Extended Community

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

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

7. Operation

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

7.1. MAC Address Distribution over Core

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

7.2. Device Multi-homing

7.2.1 Flow-based Load-balancing

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

7.2.1.1 PE B-MAC Address Assignment

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

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

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

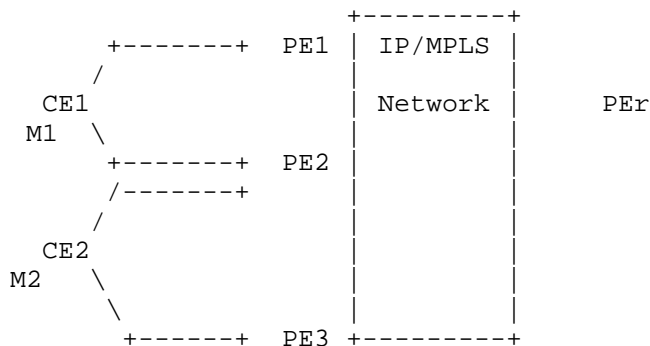


Figure 2: B-MAC Address Assignment

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

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

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

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

7.2.1.2. Automating B-MAC Address Assignment

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

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

address for this fairly uncommon service scenario.

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

7.2.1.3 Split Horizon and Designated Forwarder Election

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

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

7.2.2 I-SID Based Load-balancing

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

7.2.2.1 PE B-MAC Address Assignment

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

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

previously advertised for that segment. This causes the remote PE nodes to flush all C-MAC addresses associated with the B-MAC in question. This is done across all I-SIDs that are mapped to the EVI of the withdrawn MAC route.

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

7.3. Network Multi-homing

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

7.4. Frame Forwarding

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

7.4.1. Unicast

Known unicast traffic received from the AC will be PBB-encapsulated by the PE using the B-MAC source address corresponding to the originating site. The unicast B-MAC destination address is determined based on a lookup of the C-MAC destination address (the binding of the two is done via transparent learning of reverse traffic). The resulting frame is then encapsulated with an LSP tunnel label and the MPLS label which uniquely identifies the B-MAC destination address on the egress PE. If per flow load-balancing over ECMPs in the MPLS core is required, then a flow label is added as the end of stack label.

For unknown unicast traffic, the PE forwards these frames over MPLS

core. When these frames are to be forwarded, then the same set of options used for forwarding multicast/broadcast frames (as described in next section) are used.

7.4.2. Multicast/Broadcast

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

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

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

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

8. Minimizing ARP Broadcast

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

9. Seamless Interworking with IEEE 802.1aq/802.1Qbp

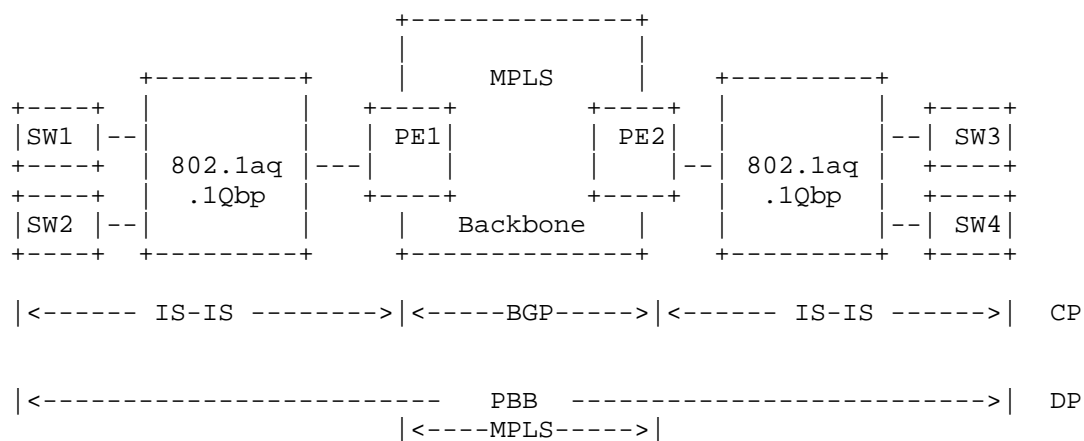


Figure 7: Interconnecting 802.1aq/802.1Qbp Networks with PBB-EVPN

9.1 B-MAC Address Assignment

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

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

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

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

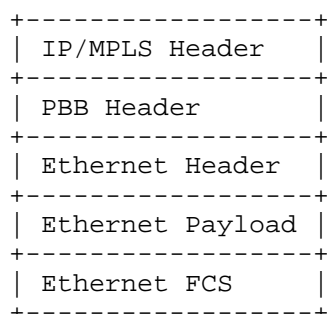


Figure 8: PBB over MPLS Encapsulation

9.3 Operation:

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

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

10. Solution Advantages

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

10.1. MAC Advertisement Route Scalability

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

10.2. C-MAC Mobility with MAC Sub-netting

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

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

10.3. C-MAC Address Learning and Confinement

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

10.4. Seamless Interworking with TRILL and 802.1aq Access Networks

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

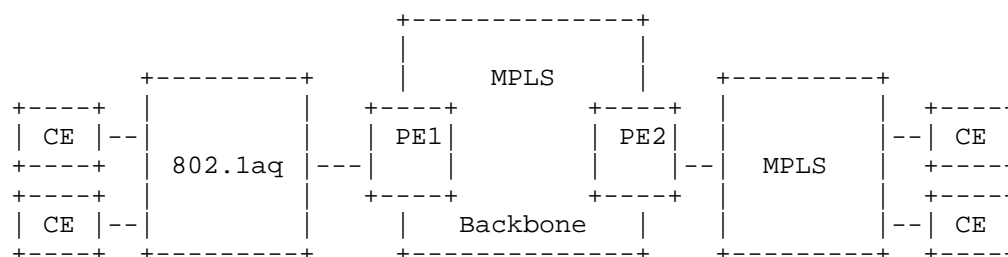


Figure 9: Interoperability with 802.1aq

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

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

10.5. Per Site Policy Support

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

10.6. Avoiding C-MAC Address Flushing

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

11. Acknowledgements

TBD.

12. Security Considerations

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

13. IANA Considerations

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

14. Intellectual Property Considerations

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

15. Normative References

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

16. Informative References

[PBB-VPLS] Sajassi et al., "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-vpls-pbb-interop-02.txt, work in progress, July, 2011.

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (E-VPN)", draft-sajassi-raggarwa-l2vpn-evpn-req-01.txt, work in progress, July, 2011.

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

17. Authors' Addresses

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

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

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

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

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

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

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

Clarence Filsfils
Cisco
Email: cfilsfil@cisco.com

Dennis Cai
Cisco
Email: dcai@cisco.com

Lizhong Jin
ZTE Corporation

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

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

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

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

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

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

1. Introduction

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

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

1.1. General Terminology

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

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

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

1.2. Conventions

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

2. Background

This section describes various scenarios where multi-homing may be required, and the implications thereof. It also describes some of the singular properties of VPLS multi-homing, and what that means from both an operational point of view and an implementation point of view. There are other approaches for providing multi-homing such as Spanning Tree Protocol, and this document specifies use of BGP for multi-homing. Comprehensive comparison among the approaches is outside the scope of this document.

2.1. Scenarios

CE1 is a VPLS CE that is dual-homed to both PE1 and PE2 for redundant connectivity.

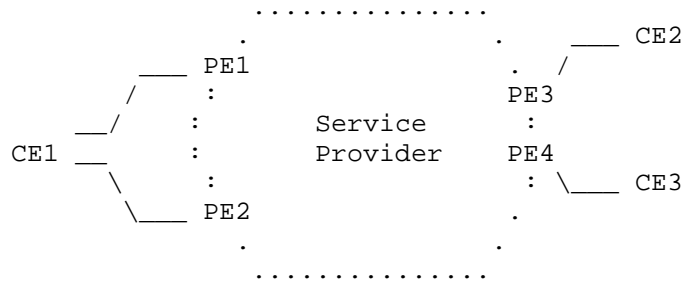


Figure 1: Scenario 1

CE1 is a VPLS CE that is dual-homed to both PE1 and PE2 for redundant connectivity. However, CE4, which is also in the same VPLS domain, is single-homed to just PE1.

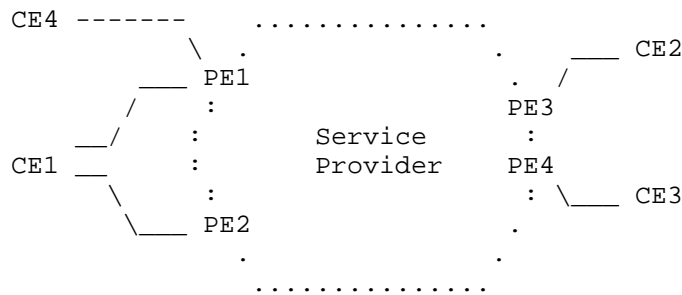


Figure 2: Scenario 2

2.2. VPLS Multi-homing Considerations

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

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

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

3. Multi-homing Operation

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

3.1. Multi-homing NLRI

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

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

BGP VPLS NLRI

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

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

3.2. Provisioning Model

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

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

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

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

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

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

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

3.3. Designated Forwarder Election

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

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

3.3.1. Attributes

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

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

Control Flags Bit Vector

```

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

```

Figure 3

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

3.3.2. Variables Used

3.3.2.1. RD

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

3.3.2.2. SITE-ID

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

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

3.3.2.3. VBO

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

3.3.2.4. DOM

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

3.3.2.5. ACS

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

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

3.3.2.6. PREF

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

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

Table 1

3.3.2.7. PE-ID

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

3.3.3. Election Procedures

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

preserved.

3.3.3.1. Bucketization for standard BGP path selection

An advertisement

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

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

3.3.3.2. Bucketization for VPLS DF Election

An advertisement

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

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

3.3.3.3. Tie-breaking Rules

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

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

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

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

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

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

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

3.4. DF Election on PEs

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

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

4. Multi-AS VPLS

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

4.1. Route Origin Extended Community

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

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

4.2. VPLS Preference

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

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

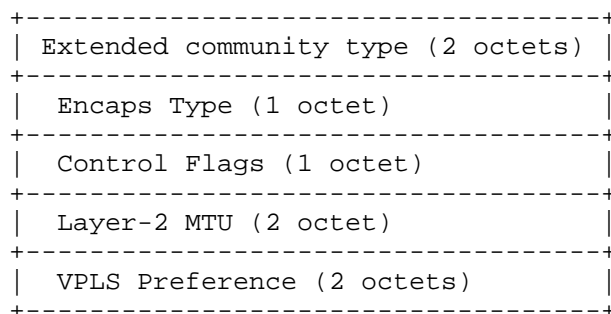


Figure 4: Layer2 Info Extended Community

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

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

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

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

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

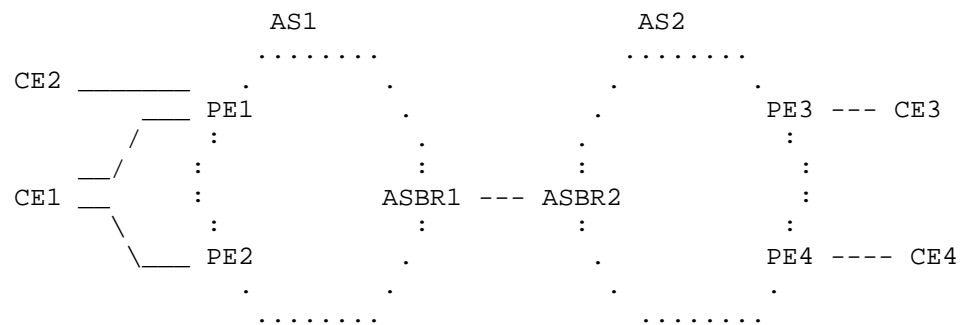


Figure 5: Inter-AS VPLS

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

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

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

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

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

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

5. MAC Flush Operations

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

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

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

5.1. MAC List Flush

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

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

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

5.2. Implicit MAC Flush

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

forwarder.

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

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

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

5.3. Minimizing the effects of fast link transitions

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

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

6. Backwards Compatibility

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

6.1. BGP based VPLS

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

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

6.2. LDP VPLS with BGP Auto-discovery

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

7. Security Considerations

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

8. IANA Considerations

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

9. Acknowledgments

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

10. References

10.1. Normative References

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

10.2. Informative References

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

Authors' Addresses

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

Email: bhupesh@cohere.net

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

Email: kireeti.kompella@gmail.com

Wim Henderickx
Alcatel-Lucent

Email: wim.henderickx@alcatel-lucent.be

Florin Balus
Alcatel-Lucent

Email: florin.balus@alcatel-lucent.com

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

Email: uttaro@att.com

Senad Palislaamovic
Alcatel-Lucent

Email: senad.palislaamovic@alcatel-lucent.com

Wen Lin
Juniper Networks

Email: wlin@juniper.net

Internet Working Group

Internet Draft

Intended status: Standards Track

Y. Jiang

L. Yong

Huawei

M. Paul
Deutsche Telekom

F. Jounay

Orange CH

F. Balus
W. Henderickx
Alcatel-Lucent

A. Sajassi
Cisco

Expires: August 2013

February 5, 2013

VPLS PE Model for E-Tree Support
draft-ietf-l2vpn-vpls-pe-etree-01.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 August 5, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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.	Signaling for E-Tree Support	14
6.1.	LDP Extensions for E-Tree Support	14
6.2.	BGP Extensions for E-Tree Support	16
7.	OAM Considerations	18
8.	Applicability	18
9.	Security Considerations	18

10. IANA Considerations	18
11. References	19
11.1. Normative References	19
11.2. Informative References	19
12. 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.

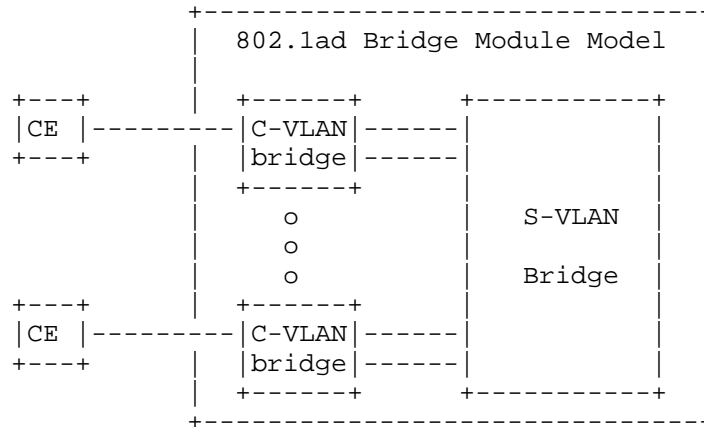


Figure 1 A model of 802.1ad Bridge Module

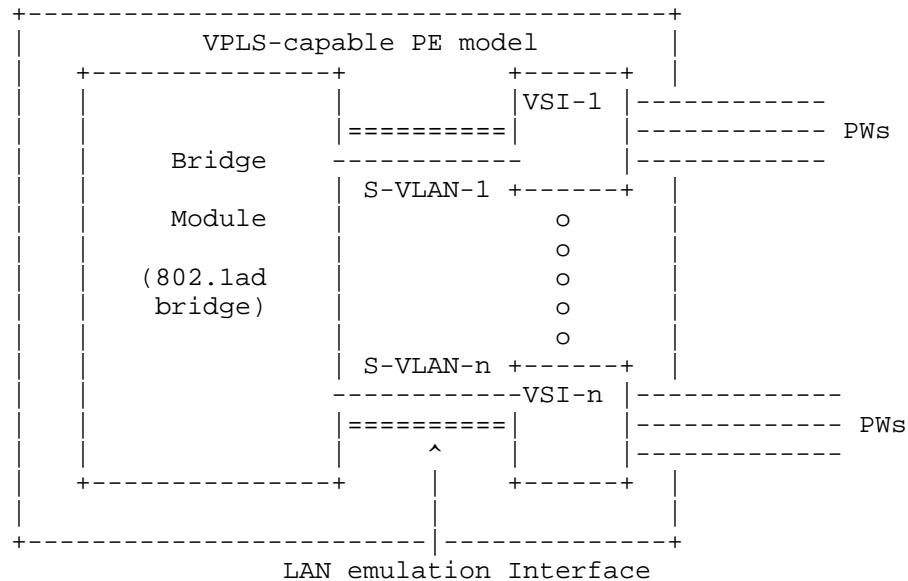


Figure 2 A VPLS-capable PE Model

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.

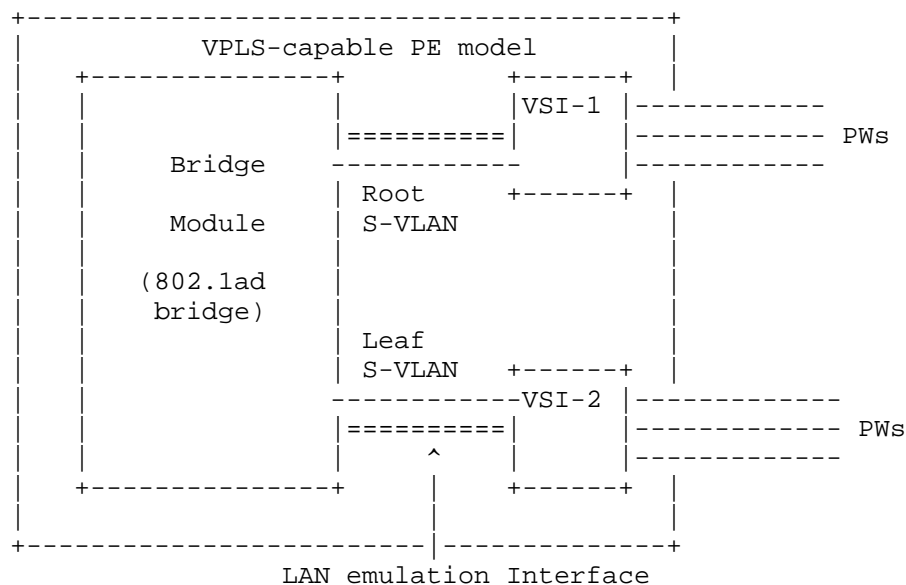


Figure 3 A VPLS PE Model for E-Tree with 2 VSIs

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.

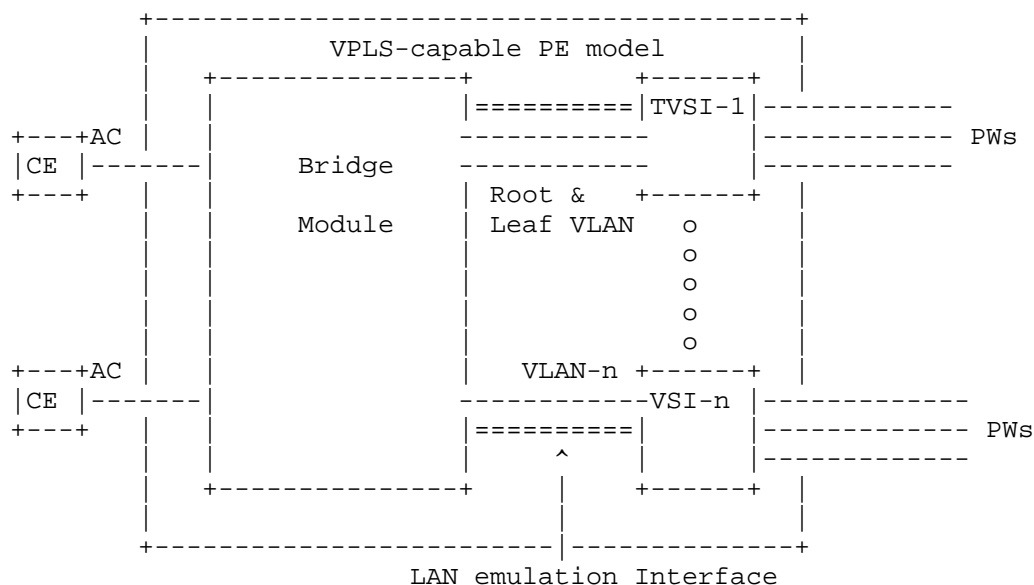


Figure 4 A VPLS PE Model for E-Tree with a Single T-VSI

For an untagged port (frames over this port are untagged) or VLAN-unaware port (VLAN tags in the frames are ignored), the Ethernet frames received from the root ACs SHOULD be tagged with a root C-VLAN, and optionally MAY be added with another root S-VLAN.

For a C-VLAN tagged port, the Ethernet frames received from the root ACs SHOULD be added with a root S-VLAN.

For an S-VLAN tagged port, the S-VLAN tag in the Ethernet frames received from the root ACs SHOULD 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] MAY be used, and a root B-VLAN or leaf B-VLAN MAY 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).

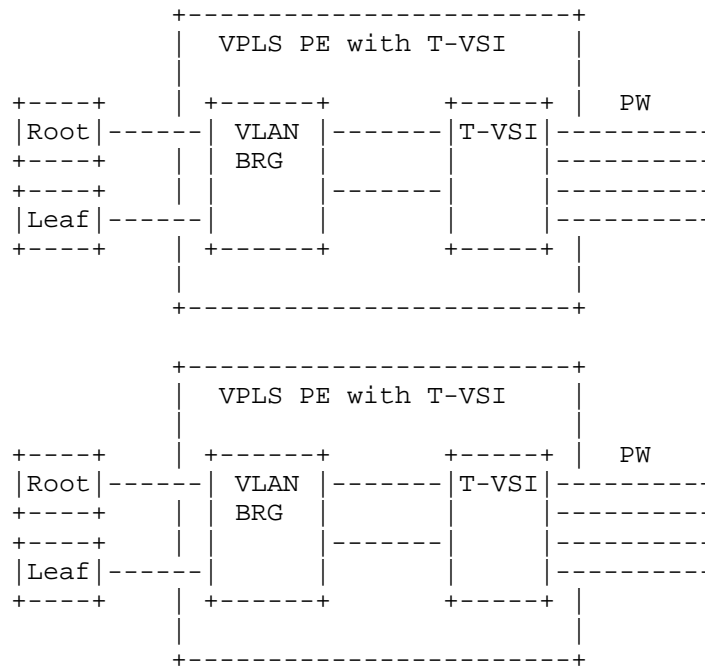


Figure 5 T-VSI Interconnected in the Normal Mode

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].

The signaling for VLANs is specified in Section 6.

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. A raw PW should be used to connect them.

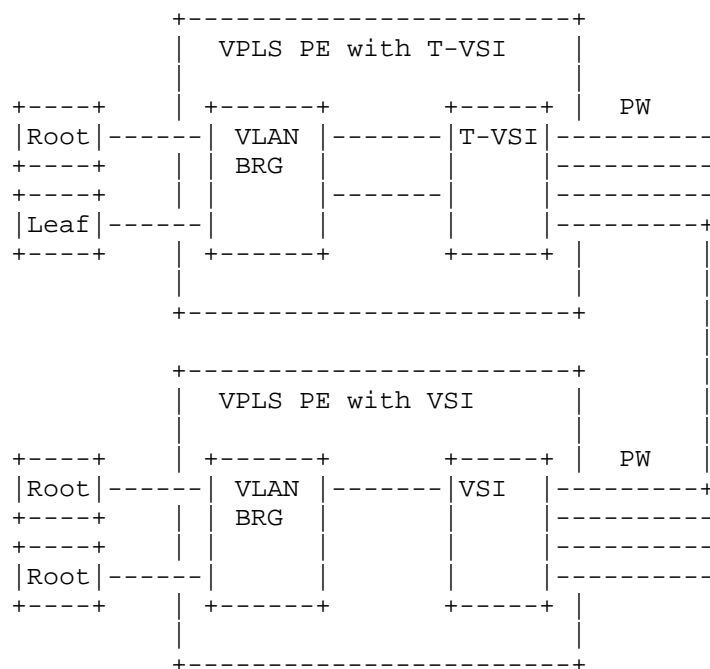


Figure 6 T-VSI interconnected with Traditional VSI

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 (both have E-Tree capability) are inter-connected and one of them (e.g., PE2) is attached with only leaf nodes, as shown in the scenario of Fig. 7, its peer PE (e.g., PE1) should then work in the optimized 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.

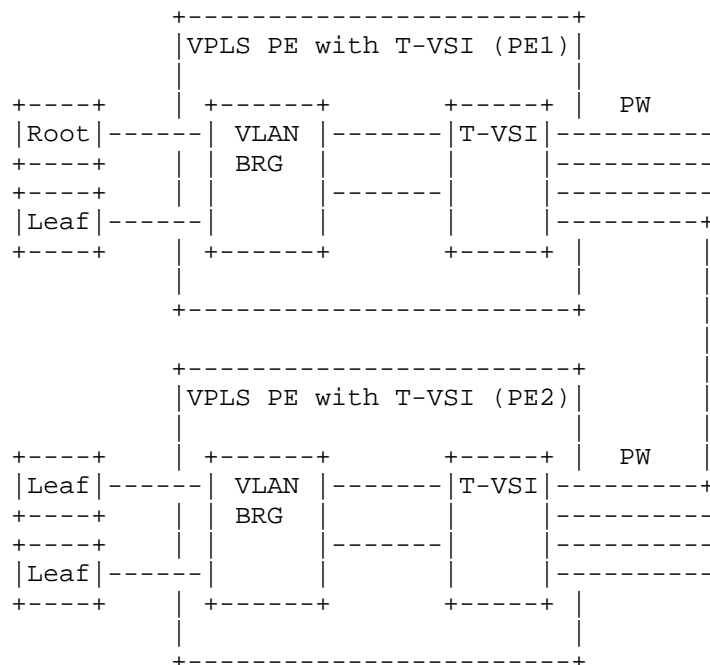


Figure 7 T-VSI interconnected with PE attached with only leaf nodes

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. Signaling for E-Tree Support

6.1. 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:

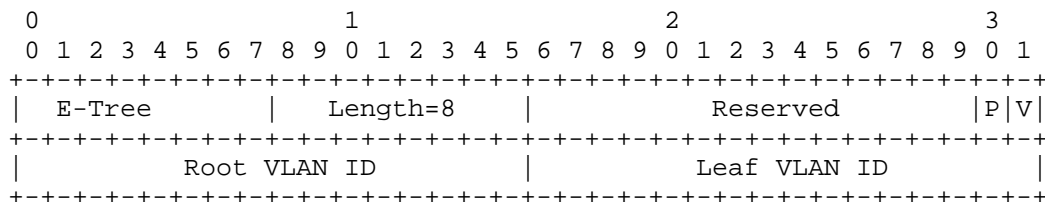


Figure 8 E-Tree Sub-TLV

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, after saving the VLAN information for the PW, 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.

6.2. BGP Extensions for E-Tree Support

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

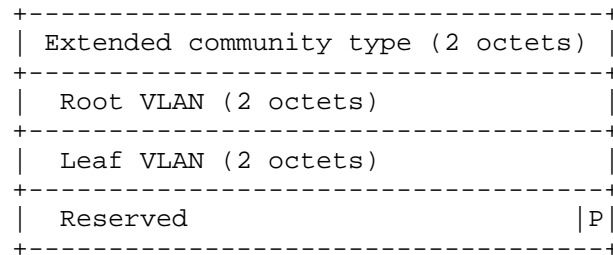


Figure 9 E-Tree Extended Community

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, after saving the VLAN information for the PW, 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 {

If the PE is a leaf-only PE itself, then forbids any traffic on the PW;

Else 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 6.1.

7. 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.

8. 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.

9. 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.

10. 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

11. References

11.1. 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

11.2. 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

12. 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.

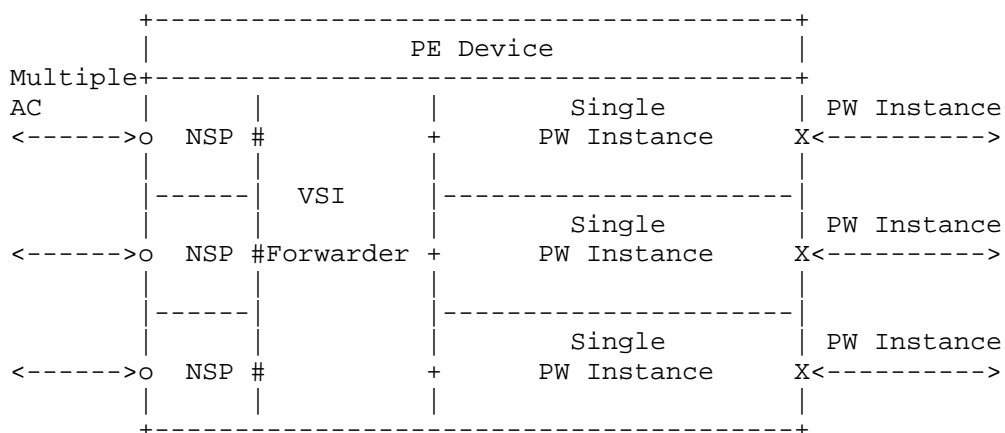


Figure A.1 A PE model with a VSI and no bridge module

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

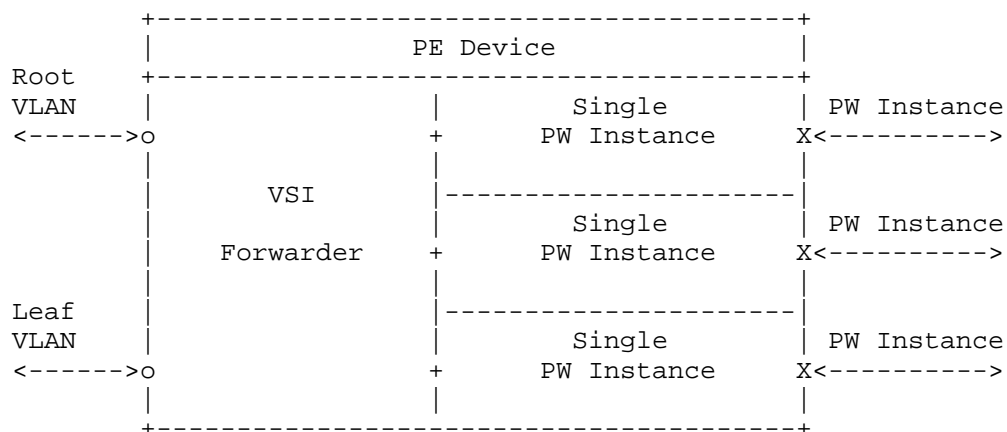


Figure A.2 A PE model with external E-Tree interface

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

Wim Henderickx
Alcatel-Lucent

Jim Uttaro
AT&T

Expires: April 22, 2012

October 22, 2012

E-TREE Support in E-VPN
draft-sajassi-l2vpn-evpn-etree-01

Abstract

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). [ETREE-FMWK] 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	4
2.3	Scenario 3: Leaf AND Root site(s) per Ethernet Segment	4
3	Operation	5
3.1	E-Tree with MAC Learning	7
3.2	E-Tree without MAC Learning	7
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	8

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-FMWK] 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.

[E-VPN] 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 [KEYWORDS].

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

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.

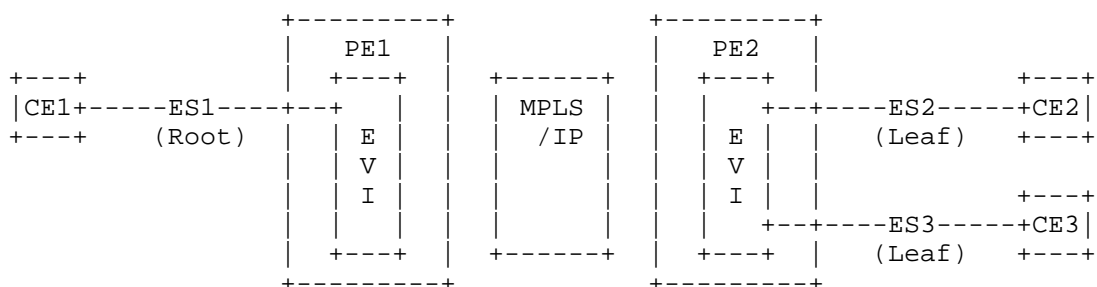


Figure 1: Scenario 1

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.

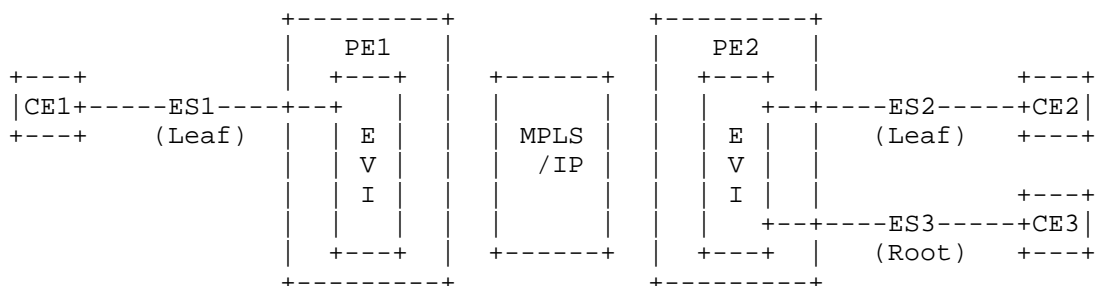


Figure 2: Scenario 2

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.

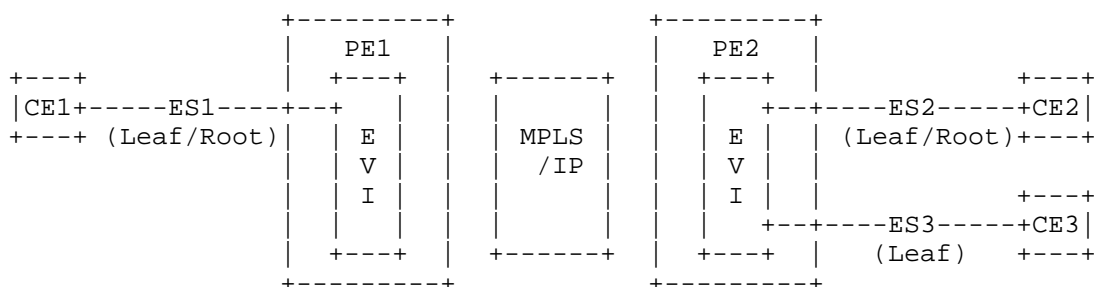


Figure 3: Scenario 3

3 Operation

[E-VPN] defines the notion of an Ethernet Segment which can be readily used to identify a Root and/or Leaf site in E-TREE services. In other words, [E-VPN] has inherent capability to support E-TREE services without defining any new BGP routes and/or attributes. It only requires a minor modification to the existing procedures as shown in this section.

The following procedure is used consistently for all the scenarios highlighted in the previous section. In order to apply the proper egress filtering, which varies based on whether a packet is sent from a Root or a Leaf, the MPLS-encapsulated frames MUST be tagged with an indication of whether they originated from a Root or a Leaf Ethernet Segment. 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 the split-horizon rule defined in [E-VPN]:

- 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 encoding the Root or Leaf indication in the

Flags field of the ESI MPLS label Extended Community attribute ([E-VPN] Section 8) to indicate Root/Leaf status.

In the case where a multi-homed Ethernet Segment 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 the case where a multi-homed Ethernet Segment has either Root or Leaf sites attached, then a single ESI MPL label is allocated and advertised.

Furthermore, 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. Note that these special labels are advertised on a per PE basis (i.e. each PE advertises only two such special labels).

In addition to egress filtering (which is a MUST requirement), an E-VPN PE implementation MAY provide topology constraint among the PEs belonging to the same EVI associated with an E-TREE service. The purpose of this topology constraint is to avoid having PEs with only host Leaf sites importing and processing BGP MAC routes from each other, thereby unnecessarily exhausting their RIB tables. However, as soon as a Root site is added to a Leaf PE, then that PE needs to process MAC routes from all other Leaf PEs and add them to its forwarding table. To support such topology constrain in E-VPN, two BGP Route-Targets (RTs) are used for 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. If for a given EVI, the PEs can eventually have both Leaf and Root sites attached, even though they may start as Root-only or Leaf-only PEs, then it is recommended to use a single RT per EVI and avoid additional configuration and operational overhead. If the number of EVIs is very large (e.g., more than 32K or 64K), then RT type 0 as defined in [RFC4360] SHOULD be used; otherwise, RT type 2 is sufficient.

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

- Ethernet Unicast from Root to Roots & Leaf

- Ethernet Unicast from Leaf 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 and Dennis Cai for their 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.

[RFC4360] S. Sangli et al, "'BGP Extended Communities Attribute", February, 2006.

7.2 Informative References

[ETREE-FMWK] Key et al., "A Framework for E-Tree Service over MPLS Network", draft-ietf-l2vpn-etree-frwk-01, work in progress, January 2012.

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-01.txt, work in progress, February, 2012.

[ETREE-REQ] Key et al., "Requirements for MEF E-Tree Support in L2VPN", draft-ietf-l2vpn-etree-req-03, work in progress, October 2012.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

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

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

L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Cisco

Yakov Rekhter
John Drake
Juniper

Expires: August 25, 2013

February 25, 2013

IP Inter-Subnet Forwarding in E-VPN
draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-01

Abstract

E-VPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among hosts/VMs over an MPLS/IP network. However, there are scenarios in which inter-subnet forwarding among hosts/VMs across different IP subnets is required, while maintaining the multi-homing capabilities of E-VPN. This document describes an IRB solution based on E-VPN to address such requirements.

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
2	Inter-Subnet Forwarding Scenarios	4
2.1	Connecting E-VPN NVEs within a DC	5
2.2	Connecting E-VPN NVEs in different DCs without route aggregation	5
2.3	Connecting E-VPN NVEs in different DCs with route aggregation	6
2.4	Connecting IP-VPN sites and E-VPN NVEs with route aggregation	6
3	Default Gateway Addressing	7
4	Operational Models for Inter-Subnet Forwarding	7
4.1	Among E-VPN NVEs within a DC	7
4.2	Among E-VPN NVEs in Different DCs Without Route Aggregation	9
4.3	Among E-VPN NVEs in Different DCs with Route Aggregation	10
4.4	Among IP-VPN Sites and E-VPN NVEs with Route Aggregation	11
5	VM Mobility	12
5.1	VM Mobility & Optimum Forwarding for VM's Outbound Traffic	12
5.2	VM Mobility & Optimum Forwarding for VM's Inbound Traffic	12
5.2.1	Mobility without Route Aggregation	13
5.2.2	Mobility with Route Aggregation	13
6	Acknowledgements	13
7	Security Considerations	13
8	IANA Considerations	13
9	References	13
9.1	Normative References	13
9.2	Informative References	14
	Authors' Addresses	14

Terminology

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

IRB: Integrated Routing and Bridging

IRB Interface: A virtual interface that connects the bridging module and the routing module on an NVE.

NVE: Network Virtualization Endpoint

1 Introduction

E-VPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among hosts/VMs over an MPLS/IP network. However, there are scenarios where, in addition to intra-subnet forwarding, inter-subnet forwarding is required among hosts/VMs across different IP subnets, while maintaining the multi-homing capabilities of E-VPN. This document describes an IRB solution based on E-VPN to address such requirements.

2 Inter-Subnet Forwarding Scenarios

The inter-subnet forwarding scenarios for E-VPN can be divided into the following five categories. The last scenario, along with their corresponding solutions, are described in [EVPN-IPVPN-INTEROP]. The solutions for the first four scenarios are the focus of this document. For the following inter-subnet forwarding scenarios, the E-VPN sites are assumed to belong to different E-VPN instances.

1. Connecting E-VPN sites within a DC
2. Connecting E-VPN sites in different DCs without route aggregation
3. Connecting E-VPN sites in different DCs with route aggregation
4. Connecting IP-VPN sites and E-VPN sites with route aggregation
5. Connecting IP-VPN sites and E-VPN sites without route aggregation

In the above scenario, the term "route aggregation" refers to the case where for a given EVI/VRF a node situated at the WAN edge of the data center network behaves as a default gateway for all the destinations that are outside the data center. The absence of route aggregation refers to the scenario where a given EVI/VRF within a data center has (host) routes to individual VMs that are outside of the data center.

In the case (4) the WAN edge node also performs route aggregation for all the destinations within its own data center, and acts as an interworking unit between E-VPN and IP VPN (it implements both E-VPN and IP VPN functionality).

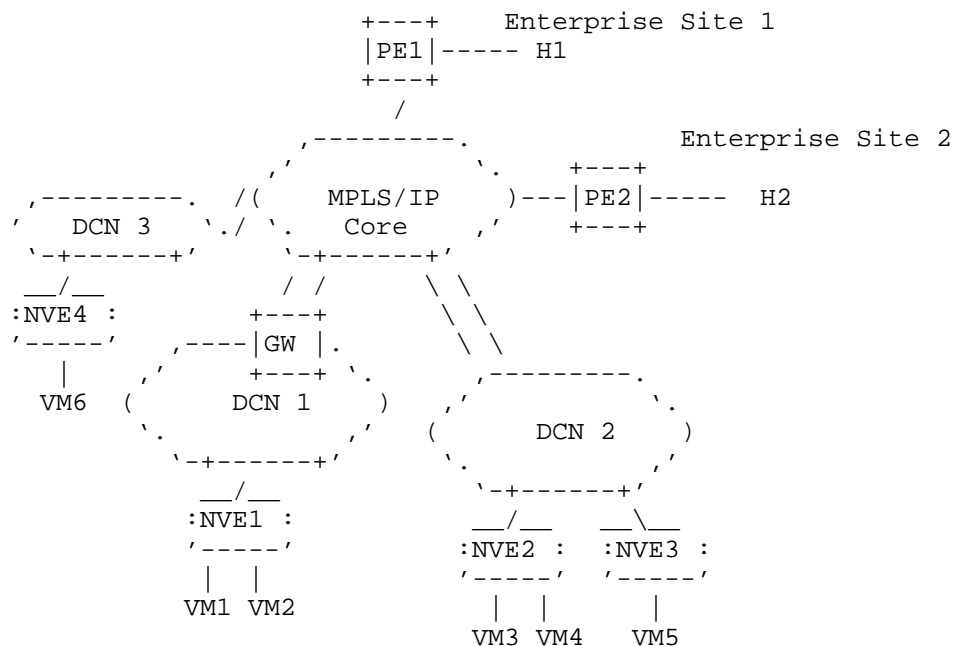


Figure 2: Interoperability Use-Cases

In what follows, we will describe scenarios 3 through 6 in more detail.

2.1 Connecting E-VPN NVEs within a DC

In this scenario, connectivity is required between hosts (e.g. VMs) in the same data center, where those hosts belong to different IP subnets. All these subnets are part of the same IP VPN. Each subnet is associated with a single EVPN, where each such EVPN is realized by a collection of EVIs residing on appropriate NVEs.

As an example, consider VM3 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the IP3 subnet whereas VM5 belongs to the IP5 subnet. Both IP3 and IP5 subnets are part of the same IP VPN. NVE2 has an EVI3 associated with IP3 subnet and NVE3 has an EVI5 associated with the IP5 subnet.

2.2 Connecting E-VPN NVEs in different DCs without route aggregation

This case is similar to that of section 2.1 above albeit for the fact that the hosts belong to different data centers that are interconnected over a WAN (e.g. MPLS/IP PSN). The data centers in

question here are seamlessly interconnected to the WAN, i.e., the WAN edge does not maintain any host/VM-specific addresses in the forwarding path.

As an example, consider VM3 and VM6 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the IP3 subnet whereas VM6 belongs to the IP6 subnet. NVE2 has an EVI3 associated with IP3 subnet and NVE4 has an EVI6 associated with the IP6 subnet. Both IP3 and IP6 subnets are part of the same IP VPN. Both EVI3 and EVI6 have VRFs associated with that IP VPN.

2.3 Connecting E-VPN NVEs in different DCs with route aggregation

In this scenario, connectivity is required between hosts (e.g. VMs) in different data centers, and those hosts belong to different IP subnets. What makes this case different from that of Section 2.2 is that (in the context of a given EVI/VRF) at least one of the data centers in question has a gateway as the WAN edge switch. Because of that, the EVIs/VRFs within each data center need not maintain (host) routes to individual VMs outside of the data center.

As an example, consider VM1 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM1 belongs to the IP1 subnet whereas VM5 belongs to the IP5 subnet thus IP1 and IP5 subnets belong to the same IP VPN. NVE3 has an EVI5 associated with the IP5 subnet and NVE1 has an EVI1 associated with the IP1 subnet. Both EVI1 and EVI5 have associated with their VRFs that belong to the IP VPN that includes IP1 and IP5 subnets. Due to the gateway at the edge of DCN 1, NVE1 does not have the address of VM5 in its VRF table.

2.4 Connecting IP-VPN sites and E-VPN NVEs with route aggregation

In this scenario (within a context of a particular E-VPN instance), connectivity is required between hosts (e.g. VMs) in a data center and hosts in an enterprise site that belongs to a given IP-VPN. The NVE within the data center is an E-VPN NVE, whereas the enterprise site has an IP-VPN PE. Furthermore, the data center in question has a gateway as the WAN edge switch. Because of that, the NVE in the data center does not need to maintain individual IP prefixes advertised by enterprise sites (by IP-VPN PEs).

As an example, consider end-station H1 and VM2 of Figure 2. Assume that connectivity is required between the end-station and the VM, where VM2 belongs to the IP2 subnet that is realized using EVPN, whereas H1 belongs to an IP VPN site connected to PE1 (PE1 maintains an IP VPN VRF associated with that IP VPN). NVE1 has an EVI2

associated with the IP2 subnet. Moreover, NVE1 maintains a VRF associated with EVI2. PE1 originates a VPN-IP route that covers H1. The gateway at the edge of DCN1 performs interworking function between IP-VPN and E-VPN. As a result of this, a default route in the VRF associated with EVI2, pointing to the gateway as the next hop, and a route to the VM2 (or maybe IP2 subnet) on the H1's VRF on PE1 are sufficient for the connectivity between H1 and VM2.

3 Default Gateway Addressing

To support inter-subnet forwarding, the NVE behaves as an IP Default Gateway from the perspective of the attached end-stations (e.g. VMs). Two models are possible, as discussed in [DC-MOBILITY]:

1. All the EVIs of a given E-VPN instance use the same anycast default gateway IP address and the same anycast default gateway MAC address. On each NVE, this default gateway IP/MAC address correspond to the IRB interface of the EVI associated with that E-VPN instance.
2. Each EVI of a given E-VPN instance uses its own default gateway IP and MAC addresses, and these addresses are aliased to the same conceptual gateway through the use of the Default Gateway extended community as specified in [E-VPN], which is carried in the E-VPN MAC Advertisement routes. On each NVE, this default gateway IP/MAC address correspond to the IRB interface of the EVI associated with that E-VPN instance.

Both of these models enable a packet forwarding paradigm where inter-subnet traffic can bypass the VRF processing on the egress (i.e. disposition) NVE. The egress NVE merely needs to perform a lookup in the associated EVI and forward the Ethernet frames unmodified, i.e. without rewriting the source MAC address. This is different from traditional IRB forwarding where a packet is forwarded through the bridge module followed by the routing module on the ingress NVE, and then forwarded through the routing module followed by the bridging module on the egress NVE. For inter-subnet forwarding using E-VPN, the routing module on the egress NVE can be completely bypassed.

It is worth noting that if the applications that are running on the hosts (e.g. VMs) are employing or relying on any form of MAC security, then the first model (i.e. using anycast addresses) would be required to ensure that the applications receive traffic from the same source MAC address that they are sending to.

4 Operational Models for Inter-Subnet Forwarding

4.1 Among E-VPN NVEs within a DC

When an E-VPN MAC advertisement route is received by the NVE, the IP address associated with the route is used to populate the VRF, whereas the MAC address associated with the route is used to populate both the bridge-domain MAC table, as well as the adjacency associated with the IP route in the VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet must be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup identifies both the next-hop (i.e. egress) NVE to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the E-VPN MAC route), instead of the MAC address of the next-hop NVE. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) NVE after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the EVI label that was advertised by the egress NVE. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVI label to identify the bridge-domain table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 2 below depicts the packet flow, where NVE1 and NVE2 are the ingress and egress NVEs, respectively.

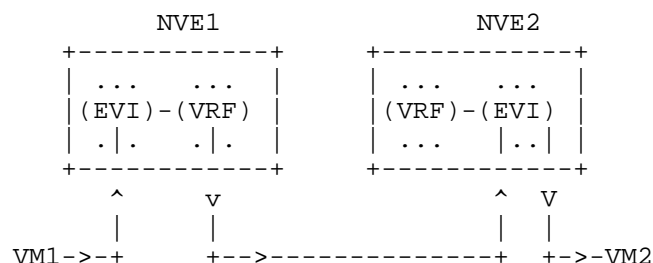


Figure 2: Inter-Subnet Forwarding Among E-VPN NVEs within a DC

Note that the forwarding behavior on the egress NVE is similar to E-VPN intra-subnet forwarding. In other words, all the packet processing associated with the inter-subnet forwarding semantics is confined to the ingress NVE.

It should also be noted that [E-VPN] provides different level of granularity for the EVI label. Besides identifying bridge domain table, it can be used to identify the egress interface or a destination MAC address on that interface. If EVI label is used for egress interface or destination MAC address identification, then no MAC lookup is needed in the egress EVI and the packet can be directly forwarded to the egress interface just based on EVI label lookup.

4.2 Among E-VPN NVEs in Different DCs Without Route Aggregation

When an E-VPN MAC advertisement route is received by the NVE, the IP address associated with the route is used to populate the VRF, whereas the MAC address associated with the route is used to populate both the bridge-domain MAC table, as well as the adjacency associated with the IP route in the VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet must be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup identifies both the next-hop (i.e. egress) Gateway to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the E-VPN MAC route), instead of the MAC address of the next-hop Gateway. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) Gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as an EVI label. The EVI label could be either advertised by the ingress Gateway, if inter-AS option B is used, or advertised by the egress NVE, if inter-AS option C is used. When the MPLS encapsulated packet is received by the ingress Gateway, the processing again differs depending on whether inter-AS option B or option C is employed: in the former case, the ingress Gateway swaps the EVI label in the packets with the EVI label value received from the egress Gateway. In the latter case, the ingress Gateway does not modify the EVI label and performs normal label switching on the LSP label. Similarly on the egress Gateway, for option B, the egress Gateway swaps the EVI label with the value advertised by the egress NVE. Whereas, for option C, the egress Gateway does not modify the EVI label, and performs normal label switching on the LSP label. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVI label to identify the bridge-domain table. It then performs a MAC lookup in that table, which yields the outbound interface to which

the Ethernet frame must be forwarded. Figure 3 below depicts the packet flow.

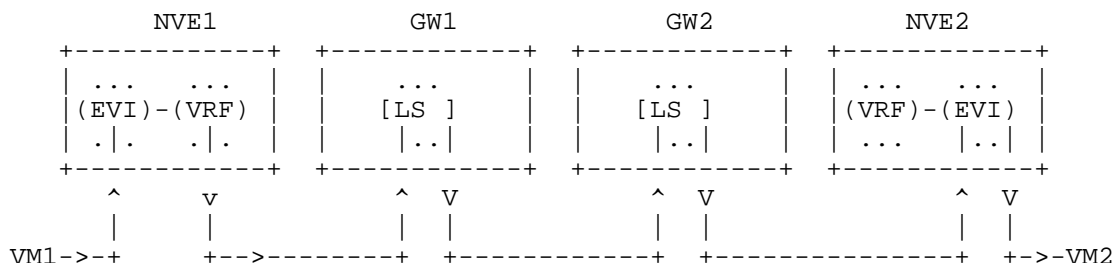


Figure 3: Inter-Subnet Forwarding Among E-VPN NVEs in Different DCs without Route Aggregation

4.3 Among E-VPN NVEs in Different DCs with Route Aggregation

In this scenario, the NVEs within a given data center do not have entries for the MAC/IP addresses of hosts in remote data centers. Rather, the NVEs have a default IP route pointing to the WAN gateway for each VRF. This is accomplished by the WAN gateway advertising for a given E-VPN that spans multiple DC a default VPN-IP route that is imported by the NVEs of that E-VPN that are in the gateway's own DC.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet must be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup, in this case, matches the default route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the IRB Interface MAC address of the local WAN gateway. It also rewrites the source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to the WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the IP-VPN label that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the IP-VPN label to identify the VRF table. It then performs an IP lookup in that table. The lookup identifies both the remote WAN gateway (of the remote data center) to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the ultimate destination host (as populated by the E-VPN MAC route). The local WAN gateway then rewrites the destination MAC address in

the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The local WAN gateway, then, forwards the frame to the remote WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as a VPN label that was advertised by the remote WAN gateway. When the MPLS encapsulated packet is received by the remote WAN gateway, it simply swaps the VPN label with the EVI label advertised by the egress NVE. This implies that the remote WAN gateway must allocate the VPN label at least at the granularity of a (VRF, egress NVE) tuple. The remote WAN gateway then forward the packet to the egress NVE. The egress NVE then performs a MAC lookup in the EVI (identified by the received EVI label) to determine the outbound port to send the traffic on.

Figure 4 below depicts the forwarding model.

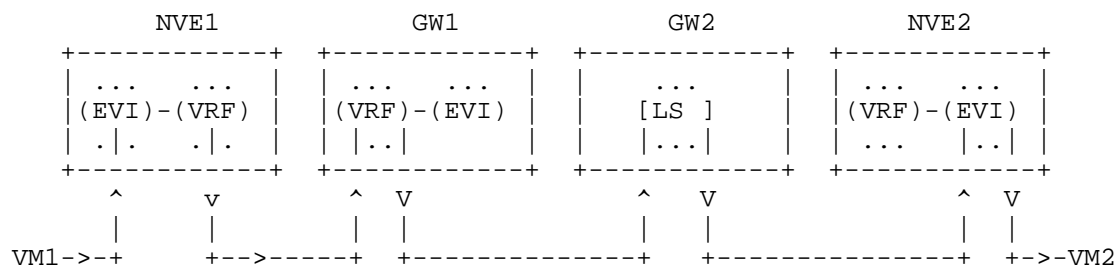


Figure 4: Inter-Subnet Forwarding Among E-VPN NVEs in Different DCs with Route Aggregation

4.4 Among IP-VPN Sites and E-VPN NVEs with Route Aggregation

In this scenario, the NVEs within a given data center do not have entries for the IP addresses of hosts in remote enterprise sites. Rather, the NVEs have a default IP route pointing to the WAN gateway for each VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet must be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated VRF table. The lookup, in this case, matches the default route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the IRB Interface MAC address of the local WAN gateway. It also rewrites the source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to

the WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the IP-VPN label that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the IP-VPN label to identify the VRF table. It then performs an IP lookup in that table. The lookup identifies the next hop ASBR to which the packet must be forwarded. The local gateway in this case strips the Ethernet encapsulation and forwards the IP packet to the ASBR using a label stack comprising of an LSP label and a VPN label that was advertised by the ASBR. When the MPLS encapsulated packet is received by the ASBR, it simply swaps the VPN label with the IP-VPN label advertised by the egress PE. This implies that the remote WAN gateway must allocate the VPN label at least at the granularity of a (VRF, egress PE) tuple. The ASBR then forwards the packet to the egress PE. The egress PE then performs an IP lookup in the VRF (identified by the received IP-VPN label) to determine where to forward the traffic.

Figure 5 below depicts the forwarding model.

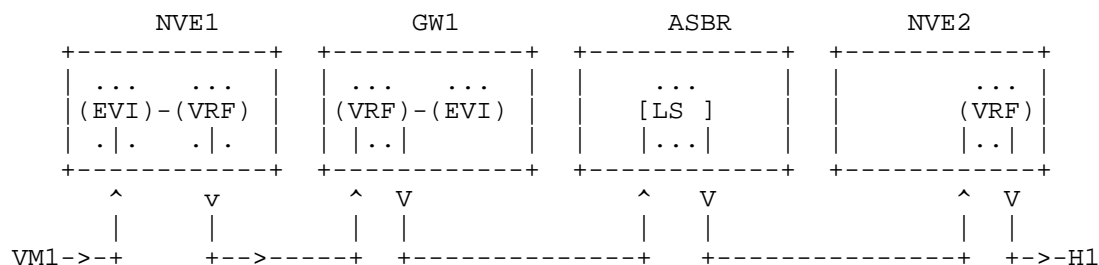


Figure 5: Inter-Subnet Forwarding Among IP-VPN Sites and E-VPN NVEs with Route Aggregation

5 VM Mobility

5.1 VM Mobility & Optimum Forwarding for VM's Outbound Traffic

Optimum forwarding for the VM's outbound traffic, upon VM mobility, can be achieved using either the anycast default Gateway MAC and IP addresses, or using the address aliasing as discussed in [DC-MOBILITY].

5.2 VM Mobility & Optimum Forwarding for VM's Inbound Traffic

For optimum forwarding of the VM's inbound traffic, upon VM mobility, all the NVEs and/or IP-VPN PE's need to know the up to date location of the VM. Two scenarios must be considered, as discussed next.

In what follows, we use the following terminology:

- source NVE refers to the NVE behind which the VM used to reside prior to the VM mobility event.
- target NVE refers to the new NVE behind which the VM has moved after the mobility event.

5.2.1 Mobility without Route Aggregation

In this scenario, when a target NVE detects that a MAC mobility event has occurred, it initiates the MAC mobility handshake in BGP as specified in [E-VPN]. The WAN Gateways, acting as ASBRs in this case, re-advertise the MAC route of the target NVE with the MAC Mobility extended community attribute unmodified. Because the WAN Gateway for a given data center re-advertises BGP routes received from the WAN into the data center, the source NVE will receive the MAC Advertisement route of the target NVE (with the next hop attribute adjusted depending on which inter-AS option is employed). The source NVE will then withdraw its original MAC Advertisement route as a result of evaluating the Sequence Number field of the MAC Mobility extended community in the received MAC Advertisement route. This is per the procedures already defined in [E-VPN].

5.2.2 Mobility with Route Aggregation

This section will be completed in the next revision.

6 Acknowledgements

The authors would like to thank Sami Boutros for his valuable comments.

7 Security Considerations

8 IANA Considerations

9 References

9.1 Normative References

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

9.2 Informative References

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

[EVPN-IPVPN-INTEROP] Sajassi et al., "E-VPN Seamless Interoperability with IP-VPN", draft-sajassi-l2vpn-evpn-ipvpn-interop-01, work in progress, October, 2012.

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-04.txt, work in progress, December, 2012.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

Yakov Rekhter
Juniper Networks
Email: yakov@juniper.net

John E. Drake
Juniper Networks
Email: jdrake@juniper.net

L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi (Editor)
Cisco

J. Drake (Editor)
Juniper

Y. Rekhter
R. Shekhar
B. Schliesser
Juniper

Nabil Bitar
Verizon

S. Salam
K. Patel
D. Rao
Cisco

Aldrin Isaac
Bloomberg

James Uttaro
AT&T

L. Yong
Huawei

W. Henderickx
Alcatel-Lucent

Expires: August 25, 2013

February 25, 2013

A Network Virtualization Overlay Solution using E-VPN
draft-sd-l2vpn-evpn-overlay-01

Abstract

This document describes how E-VPN can be used as an NVO solution and explores the various tunnel encapsulation options over IP and their impact on the E-VPN control-plane and procedures. In particular, the following encapsulation options are analyzed: MPLS over GRE, VXLAN, and NVGRE.

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	4
1.1	Terminology	5
2	E-VPN Features	5
3	Encapsulation Options for E-VPN Overlays	6
3.1	VXLAN/NVGRE Encapsulation	6
3.1.1	Virtual Identifiers Scope	7
3.1.1.1	Data Center Interconnect with Gateway	7
3.1.1.2	Data Center Interconnect without Gateway	8
3.1.2	Virtual Identifiers to EVI Mapping	8
3.1.3	Constructing E-VPN BGP Routes	9
3.2	MPLS over GRE	10
4	E-VPN with Multiple Data Plane Encapsulations	10
5	NVE Residing in Hypervisor	11
5.1	Impact on E-VPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation	11
5.2	Impact on E-VPN Procedures for VXLAN/NVGRE Encapsulation	12
6	NVE Residing in ToR Switch	12
6.1	E-VPN Multi-Homing Features	13
6.1.1	Multi-homed Ethernet Segment Auto-Discovery	13
6.1.2	Fast Convergence and Mass Withdraw	13
6.1.3	Split-Horizon	13
6.1.4	Aliasing and Backup-Path	13

6.1.5 DF Election	14
6.2 Impact on E-VPN BGP Routes & Attributes	15
6.3 Impact on E-VPN Procedures	15
6.3.1 Split Horizon	16
6.3.2 Aliasing and Backup-Path	16
7 Support for Multicast	17
8 Inter-AS	17
10 Acknowledgement	18
11 Security Considerations	18
12 IANA Considerations	19
13 References	19
11.1 Normative References	19
11.2 Informative References	20
Authors' Addresses	20

1 Introduction

In the context of this document, a Network Virtualization Overlay (NVO) is a solution to address the requirements of a multi-tenant data center, especially one with virtualized hosts, e.g., Virtual Machines (VMs). The key requirements of such a solution, as described in [Problem-Statement], are:

- Isolation of network traffic per tenant
- Support for a large number of tenants (tens or hundreds of thousands)
- Extending L2 connectivity among different VMs belonging to a given tenant segment (subnet) across different PODs within a data center or between different data centers
- Allowing a given VM to move between different physical points of attachment within a given L2 segment

The underlay network for NVO solutions is assumed to provide IP connectivity between NVO endpoints (NVEs).

This document describes how E-VPN can be used as an NVO solution and explores applicability of E-VPN functions and procedures. In particular, it describes the various tunnel encapsulation options for E-VPN over IP, and their impact on the E-VPN control-plane and procedures for two main scenarios:

- a) when the NVE resides in the hypervisor, and
- b) when the NVE resides in a ToR device

Note that the use of E-VPN as an NVO solution does not necessarily mandate that the BGP control-plane be running on the NVE. For such scenarios, it is still possible to leverage the E-VPN solution by using XMPP, or alternative mechanisms, to extend the control-plane to the NVE as discussed in [L3VPN-ENDSYSTEMS].

The possible encapsulation options for E-VPN overlays that are analyzed in this document are:

- VXLAN and NVGRE
- MPLS over GRE

Before getting into the description of the different encapsulation options for E-VPN over IP, it is important to highlight the E-VPN solution's main features, how those features are currently supported,

and any impact that the encapsulation has on those features.

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 [KEYWORDS].

NVE: Network Virtualization Endpoint

Virtual Identifier: refers to a VXLAN VNI or NVGRE VSID

2 E-VPN Features

E-VPN was originally designed to support the requirements detailed in [EVPN-REQ] and therefore has the following attributes which directly address control plane scaling and ease of deployment issues.

- 1) Control plane traffic is distributed with BGP and Broadcast and Multicast traffic is sent using a shared multicast tree or with ingress replication.
- 2) Control plane learning is used for MAC (and IP) addresses instead of data plane learning. The latter requires the flooding of unknown unicast and ARP frames; whereas, the former does not require any flooding.
- 3) Route Reflector is used to reduce a full mesh of BGP sessions among PE devices to a single BGP session between a PE and the RR. Furthermore, RR hierarchy can be leveraged to scale the number BGP routes on the RR.
- 4) Auto-discovery via BGP is used to discover PE devices participating in a given VPN, PE devices participating in a given redundancy group, tunnel encapsulation types, multicast tunnel type, multicast members, etc.
- 5) Active-active multi-homing is used. This allows a given customer device (CE) to have multiple links to multiple PEs, and traffic to/from that CE fully utilizes all of these links. This set of links is termed an Ethernet Segment (ES).
- 6) Mass withdraw is used. When a link between a CE and a PE fails, the PEs in all E-VPNs configured on that failed link are notified via the withdrawal of a single E-VPN route regardless of how many MAC addresses are located at the CE.
- 7) Route filtering and constrained route distribution are used to

ensure that the control plane traffic for a given E-VPN is only distributed to the PEs in that E-VPN.

8) The internal identifier of a broadcast domain, the Ethernet Tag, is a 32 bit number, which is mapped into whatever broadcast domain identifier, e.g., VLAN ID, is understood by the attaching CE device. This means that when 802.1q interfaces are used, there are up to 4096 distinct VLAN IDs for each attaching CE device in a given E-VPN.

9) VM Mobility mechanisms ensure that all PEs in a given E-VPN know the ES with which a given VM, as identified by its MAC and IP addresses, is currently associated.

10) Route Targets are used to allow the operator (or customer) to define a spectrum of logical network topologies including mesh, hub & spoke, and extranets (e.g., a VPN whose sites are owned by different enterprises), without the need for proprietary software or the aid of other virtual or physical devices.

11) Because the design goal for NVO is millions of instances per common physical infrastructure, the scaling properties of the control plane for NVO are extremely important. E-VPN and the extensions described herein, are designed with this level of scalability in mind.

3 Encapsulation Options for E-VPN Overlays

3.1 VXLAN/NVGRE Encapsulation

Both VXLAN and NVGRE are examples of technologies that provide a data plane encapsulation which is used to transport a packet over the common physical infrastructure between NVEs, VXLAN Tunnel End Point (VTEPs) in VXLAN and Network Virtualization Endpoint (NVEs) in NVGRE. Both of these technologies include the identifier of the specific NVO instance, Virtual Network Identifier (VNI) in VXLAN and Virtual Subnet Identifier (VSID), NVGRE, in each packet.

Note that a Provider Edge (PE) is equivalent to a VTEP/NVE.

[VXLAN] encapsulation is based on UDP, with an 8-byte header following the UDP header. VXLAN provides a 24-bit VNI, which typically provides a one-to-one mapping to the tenant VLAN ID, as described in [VXLAN]. In this scenario, the VTEP does not include an inner VLAN tag on frame encapsulation, and discards decapsulated frames with an inner VLAN tag. This mode of operation in [VXLAN] maps to VLAN Based Service in [E-VPN], where a tenant VLAN ID gets mapped to an Ethernet VPN instance (EVI).

[VXLAN] also provides an option of including an inner VLAN tag in the encapsulated frame, if explicitly configured at the VTEP. This mode of operation maps to VLAN Bundle Service in [E-VPN], where the VLANs of a given tenant get mapped to an EVI.

[NVGRE] encapsulation is based on [GRE] and it mandates the inclusion of the optional GRE Key field which carries the VSID. There is a one-to-one mapping between the VSID and the tenant VLAN ID, as described in [NVGRE] and the inclusion of an inner VLAN tag is prohibited. This mode of operation in [NVGRE] maps to VLAN Based Service in [E-VPN]. In other words, [NVGRE] prohibits the application of VLAN Bundle Service in [E-VPN] and it only requires VLAN Based Service in [E-VPN].

As described in the next section there is no change to the encoding of E-VPN routes to support VXLAN or NVGRE encapsulation except for the use of BGP Encapsulation extended community. However, there is potential impact to the E-VPN procedures depending on where the NVE is located (i.e., in hypervisor or TOR) and whether multi-homing capabilities are required.

3.1.1 Virtual Identifiers Scope

Although VNI or VSID are defined as 24-bit globally unique values, there are scenarios in which it is desirable to use a locally significant value for VNI or VSID, especially in the context of data center interconnect:

3.1.1.1 Data Center Interconnect with Gateway

In the case where NVEs in different data centers need to be interconnected, and a Gateway is employed at the edge of the data center network, the NVEs may treat the VNI or VSID as a globally unique identifier. This is because the Gateway will provide the functionality of translating the VNI or VSID when crossing network boundaries, which may align with operator span of control boundaries. As an example, consider the network of Figure 1 below. Assume there are three network operators: one for each of the DC1, DC2 and WAN networks. The Gateways at the edge of the data centers are responsible for translating the VNIs / VSIDs between the values used in each of the data center networks and the values used in the WAN.

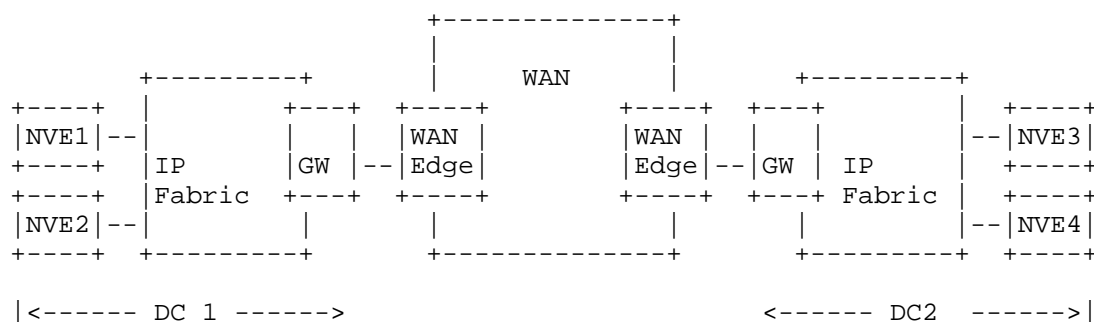


Figure 1: Data Center Interconnect with Gateway

3.1.1.2 Data Center Interconnect without Gateway

In the case where NVEs in different data centers need to be interconnected, and Gateways are not employed at the edge of the data center network, it is useful to treat the VNIs or VSIDs as locally significant identifiers (e.g., as an MPLS label). More specifically, the VNI or VSID value that is used by the transmitting NVE is allocated by the NVE that is receiving the traffic (in other words, this is a "downstream assigned" model). This allows the VNI or VSID space to be decoupled between different data center networks without the need for a dedicated Gateway at the edge of the data centers.

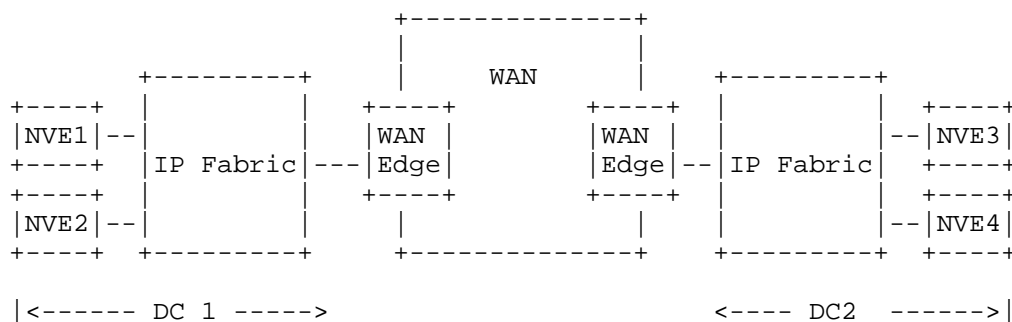


Figure 2: Data Center Interconnect without Gateway

3.1.2 Virtual Identifiers to EVI Mapping

When the E-VPN control plane is used in conjunction with VXLAN or NVGRE, two options for mapping the VXLAN VNI or NVGRE VSID to an E-VPN Instance (EVI) are possible:

1. Option 1: Single Virtual Identifier per EVI

In this option, every VNI or VSID is mapped to a unique EVI. As such, a BGP RD and RT needs to be configured per VNI / VSID on every VTEP. The advantage of this model is that it allows the BGP RT constraint mechanisms to be used in order to limit the propagation and import of routes to only the VTEPs that are interested in a given VNI or VSID. The disadvantage of this model is the provisioning overhead.

2. Option 2: Multiple Virtual Identifiers per EVI

In this option, multiple VNIs or VSIDs are mapped to a unique EVI. For example, if a tenant has multiple segments/subnets each represented by a VNI or VSID, then all the VNIs (or VSIDs) for that tenant are mapped to a single EVI. In this latter case, an EVI is equivalent to an NVO instance. The advantage of this model is that it doesn't require the provisioning of RD/RT per VNI or VSID. The disadvantage of this model is that routes would be advertised and imported by VTEPs that are not interested in a given VNI or VSID.

3.1.3 Constructing E-VPN BGP Routes

In E-VPN an MPLS label distributed by the egress PE via the E-VPN control plane and placed in the MPLS header of a given packet by the ingress PE is used upon receipt of that packet by the egress PE to disposition that packet. This is very similar to the use of the VNI or VSID by the egress VTEP or NVE, respectively, with the difference being that an MPLS label has local significance and is distributed by the E-VPN control plane, while a VNI or VSID typically has global significance.

As discussed in Section 3.1.1 above, there are scenarios in which it is desirable to use a locally significant value for VNI or VSID and in such such scenarios, MPLS label is advertised in E-VPN BGP routes and it is used in VXLAN or NVGRE encapsulation as a 20-bit value for VNI or VSID.

This memo specifies that when E-VPN is to be used with a VXLAN or NVGRE data plane and when a globally significant VNI or VSID is desirable, then Ethernet Tag field of E-VPN BGP routes (which is a 4-octet field) MUST be used and MPLS label field MUST be set to zero; however, when a locally significant VNI or VSID is desirable, then MPLS field of E-VPN BGP routes (which is a 3-octet field) MUST be used and Ethernet Tag field MUST be set to zero.

In order to indicate that a VXLAN or NVGRE data plane encapsulation rather than MPLS label stack encapsulation is to be used, the BGP Encapsulation extended community defined in [RFC5512] is included

with E-VPN MAC Advertisement or Per EVI Ethernet AD routes advertised by an egress PE. Two new values, one for VXLAN and one for NVGRE, will be defined.

3.2 MPLS over GRE

The E-VPN data-plane is modeled as an E-VPN MPLS client layer sitting over an MPLS PSN tunnel. Some of the E-VPN functions (split-horizon, aliasing and repair-path) are tied to the MPLS client layer. If MPLS over GRE encapsulation is used, then the E-VPN MPLS client layer can be carried over an IP PSN tunnel transparently. Therefore, there is no impact to the E-VPN procedures and associated data-plane operation.

The existing standards for MPLS over GRE encapsulation as defined by [RFC4023] can be used for this purpose; however, when it is used in conjunction with E-VPN the key field MUST be present, and SHOULD be used to provide a 32-bit entropy field. The Checksum and Sequence Number fields are not needed and their corresponding C and S bits MUST be set to zero.

4 E-VPN with Multiple Data Plane Encapsulations

The use of the BGP Encapsulation extended community allows each PE in a given E-VPN to know whether the other PEs in that E-VPN support MPLS label stack, VXLAN, and/or NVGRE data plane encapsulations. I.e., PEs in a given E-VPN may support multiple data plane encapsulations.

If BGP Encapsulation Extended community is not present, then the default MPLS encapsulation (or statically configured encapsulation) is used. However, if this attribute is present, then an ingress PE can send a frame to an egress PE only if the set of encapsulations advertised by the egress PE in the subject MAC Advertisement or Per EVI Ethernet AD route, forms a non-empty intersection with the set of encapsulations supported by the ingress PE, and it is at the discretion of the ingress PE which encapsulation to choose from this intersection.

An ingress node that uses shared multicast trees for sending broadcast or multicast frames MUST maintain distinct trees for each different encapsulation type.

It is the responsibility of the operator of a given E-VPN to ensure that all of the PEs in that E-VPN support at least one common encapsulation. If this condition is violated, it could result in service disruption or failure. The use of the BGP Encapsulation

extended community provides a method to detect when this condition is violated but the actions to be taken are at the discretion of the operator and are outside the scope of this document.

5 NVE Residing in Hypervisor

When a PE and its CEs are co-located in the same physical device, e.g., when the PE resides in a server and the CEs are its VMs, the links between them are virtual and they typically share fate; i.e., the subject CEs are typically not multi-homed or if they are multi-homed, the multi-homing is a purely local matter to the server hosting the VM, and need not be "visible" to any other PEs, and thus does not require any specific protocol mechanisms. The most common case of this is when the NVE resides in the hypervisor.

In the sub-sections that follow, we will discuss the impact on E-VPN procedures for the case when the NVE resides on the hypervisor and the VXLAN or NVGRE encapsulation is used.

5.1 Impact on E-VPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation

As discussed above, both [NVGRE] and [VXLAN] do not require the tenant VLAN tag to be sent in BGP routes. Therefore, the 4-octet Ethernet Tag field in the E-VPN BGP routes can be used to represent the globally significant value for VXLAN VNI or NVGRE VSID and MPLS field can be used to represent the locally significant value for VNI or VSID.

When the VXLAN VNI or NVGRE VSID is assumed to be a global value, one might question the need for the Route Distinguisher (RD) in the E-VPN routes. In the scenario where all data centers are under a single administrative domain, and there is a single global VNI/VSID space, the RD MAY be set to zero in the E-VPN routes. However, in the scenario where different groups of data centers are under different administrative domains, and these data centers are connected via one or more backbone core providers as described in [NOV3-Framework], the RD must be a unique value per EVI or per NVE as described in [E-VPN]. In other words, whenever there is more than one administrative domain for global VNI or VSID, then a non-zero RD MUST be used, or whenever the VNI or VSID value have local significance, then a non-zero RD MUST be used. It is recommend to use a non-zero RD at all time.

When the NVEs reside on the hypervisor, the E-VPN BGP routes and attributes associated with multi-homing are no longer required. This reduces the required routes and attributes to the following subset of five out of the set of eight :

- MAC Advertisement Route
- Inclusive Multicast Ethernet Tag Route
- MAC Mobility Extended Community
- Default Gateway Extended Community

As mentioned in section 3.1.1, BGP Encapsulation Extended Community attribute as defined in [RFC5512] SHOULD be used along with MAC Advertisement Route or Ethernet AD Route to indicate the supported encapsulations.

5.2 Impact on E-VPN Procedures for VXLAN/NVGRE Encapsulation

When the NVEs reside on the hypervisors, the E-VPN procedures associated with multi-homing are no longer required. This limits the procedures on the NVE to the following subset of the E-VPN procedures:

1. Local learning of MAC addresses received from the VMs per section 10.1 of [E-VPN].
2. Advertising locally learned MAC addresses in BGP using the MAC Advertisement routes.
3. Performing remote learning using BGP per Section 10.2 of [E-VPN].
4. Discovering other NVEs and constructing the multicast tunnels using the Inclusive Multicast Ethernet Tag routes.
5. Handling MAC address mobility events per the procedures of Section 16 in [E-VPN].

6 NVE Residing in ToR Switch

In this section, we discuss the scenario where the NVEs reside in the Top of Rack (ToR) switches AND the servers (where VMs are residing) are multi-homed to these ToR switches. The multi-homing may operate in All-Active or Active/Standby redundancy mode. If the servers are single-homed to the ToR switches, then the scenario becomes similar to that where the NVE resides in the hypervisor, as discussed in Section 5, as far as the required E-VPN functionality.

[E-VPN] defines a set of BGP routes, attributes and procedures to support multi-homing. We first describe these functions and procedures, then discuss which of these are impacted by the encapsulation (such as VXLAN or NVGRE) and what modifications are required.

6.1 E-VPN Multi-Homing Features

In this section, we will recap the multi-homing features of E-VPN to highlight the encapsulation dependencies. The section only describes the features and functions at a high-level. For more details, the reader is to refer to [E-VPN].

6.1.1 Multi-homed Ethernet Segment Auto-Discovery

E-VPN NVEs (or PEs) connected to the same Ethernet Segment (e.g. the same server via LAG) can automatically discover each other with minimal to no configuration through the exchange of BGP routes.

6.1.2 Fast Convergence and Mass Withdraw

E-VPN defines a mechanism to efficiently and quickly signal, to remote NVEs, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment (e.g., a link or a port failure). This is done by having each NVE advertise an Ethernet A-D Route per Ethernet segment for each locally attached segment. Upon a failure in connectivity to the attached segment, the NVE withdraws the corresponding Ethernet A-D route. This triggers all NVEs that receive the withdrawal to update their next-hop adjacencies for all MAC addresses associated with the Ethernet segment in question. If no other NVE had advertised an Ethernet A-D route for the same segment, then the NVE that received the withdrawal simply invalidates the MAC entries for that segment. Otherwise, the NVE updates the next-hop adjacencies to point to the backup NVE(s).

6.1.3 Split-Horizon

Consider a station that is multi-homed to two or more NVEs on an Ethernet segment ES1, with all-active redundancy. If the station sends a multicast, broadcast or unknown unicast packet to a particular NVE, say NE1, then NE1 will forward that packet to all or subset of the other NVEs in the E-VPN instance. In this case the NVEs, other than NE1, that the station is multi-homed to MUST drop the packet and not forward back to the station. This is referred to as "split horizon" filtering.

6.1.4 Aliasing and Backup-Path

In the case where a station is multi-homed to multiple NVEs, it is possible that only a single NVE learns a set of the MAC addresses associated with traffic transmitted by the station. This leads to a situation where remote NVEs receive MAC advertisement routes, for these addresses, from a single NVE even though multiple NVEs are connected to the multi-homed station. As a result, the remote NVEs

are not able to effectively load-balance traffic among the NVEs connected to the multi-homed Ethernet segment. This could be the case, for e.g. when the NVEs perform data-path learning on the access, and the load-balancing function on the station hashes traffic from a given source MAC address to a single NVE. Another scenario where this occurs is when the NVEs 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. This refers to the ability of an NVE 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 NVEs which receive MAC advertisement routes with non-zero ESI SHOULD consider the MAC address as reachable via all NVEs that advertise reachability to the relevant Segment using Ethernet A-D routes with the same ESI and with the Active-Standby flag reset.

Backup-Path is a closely related function, albeit it applies to the case where the redundancy mode is Active/Standby. In this case, the NVE signals that it has reachability to a given locally attached Ethernet Segment using the Ethernet A-D route as well. Remote NVEs which receive the MAC advertisement routes, with non-zero ESI, SHOULD consider the MAC address as reachable via the advertising NVE. Furthermore, the remote NVEs SHOULD install a Backup-Path, for said MAC, to the NVE which had advertised reachability to the relevant Segment using an Ethernet A-D route with the same ESI and with the Active-Standby flag set.

6.1.5 DF Election

Consider a station that is a host or a VM that is multi-homed directly to more than one NVE 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 station.
- Flooding unknown unicast traffic (i.e. traffic for which an NVE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the station, if the environment requires flooding of unknown unicast traffic.

This is required in order to prevent duplicate delivery of multi-

destination frames to a multi-homed host or VM, in case of all-active redundancy.

6.2 Impact on E-VPN BGP Routes & Attributes

Since multi-homing is supported in this scenario, then the entire set of BGP routes and attributes defined in [E-VPN] are used. As discussed in Section 3.1, the VSID or VNI is encoded in the Ethernet Tag field of the routes if globally significant or in the MPLS label field if locally significant.

As mentioned in section 3.1.1, BGP Encapsulation Extended Community attribute as defined in [RFC5512] SHOULD be used along with MAC Advertisement Route or Ethernet AD Route to indicate the supported encapsulations.

6.3 Impact on E-VPN Procedures

Two cases need to be examined here, depending on whether the NVEs are operating in Active/Standby or in All-Active redundancy.

First, let's consider the case of Active/Standby redundancy, where the hosts are multi-homed to a set of NVEs, however, only a single NVE is active at a given point of time for a given VNI or VSID. In this case, the Split-Horizon and Aliasing functions are not required but other functions such as multi-homed Ethernet segment auto-discovery, fast convergence and mass withdraw, repair path, and DF election are required. In this case, the impact of the use of the VXLAN/NVGRE encapsulation on the E-VPN procedures is when the Backup-Path function is supported, as discussed next:

In E-VPN, the NVEs connected to a multi-homed site using Active/Standby redundancy optionally advertise a VPN label, in the Ethernet A-D Route per EVI, used to send traffic to the backup NVE in the case where the primary NVE fails. In the case where VXLAN or NVGRE encapsulation is used, some alternative means that does not rely on MPLS labels is required to support Backup-Path. This is discussed in Section 4.3.2 below. If the Backup-Path function is not used, then the VXLAN/NVGRE encapsulation would have no impact on the E-VPN procedures.

Second, let's consider the case of All-Active redundancy. In this case, out of the E-VPN multi-homing features listed in section 4.1, the use of the VXLAN or NVGRE encapsulation impacts the Split-Horizon and Aliasing features, since those two rely on the MPLS client layer. Given that this MPLS client layer is absent with these types of encapsulations, alternative procedures and mechanisms are needed to

provide the required functions. Those are discussed in detail next.

6.3.1 Split Horizon

In E-VPN, an MPLS label is used for split-horizon filtering to support active/active multi-homing where an ingress ToR switch adds a label corresponding to the site of origin (aka ESI MPLS Label) when encapsulating the packet. The egress ToR switch checks the ESI MPLS label when attempting to forward a multi-destination frame out an interface, and if the label corresponds to the same site identifier (ESI) associated with that interface, the packet gets dropped. This prevents the occurrence of forwarding loops.

Since the VXLAN or NVGRE encapsulation does not include this ESI MPLS label, other means of performing the split-horizon filtering function MUST be devised. The following approach is recommended for split-horizon filtering when VXLAN or NVGRE encapsulation is used.

Every NVE track the IP address(es) associated with the other NVE(s) with which it has shared multi-homed Ethernet Segments. When the NVE receives a multi-destination frame from the overlay network, it examines the source IP address in the tunnel header (which corresponds to the ingress NVE) and filters out the frame on all local interfaces connected to Ethernet Segments that are shared with the ingress NVE. With this approach, it is required that the ingress NVE performs replication locally to all directly attached Ethernet Segments (regardless of the DF Election state) for all flooded traffic ingress from the access interfaces (i.e. from the hosts). This approach is referred to as "Local Bias", and has the advantage that only a single IP address needs to be used per NVE for split-horizon filtering, as opposed to requiring an IP address per Ethernet Segment per NVE.

In order to prevent unhealthy interactions between the split horizon procedures defined in [E-VPN] and the local bias procedures described in this memo, a mix of MPLS over GRE encapsulations on the one hand and VXLAN/NVGRE encapsulations on the other on a given Ethernet Segment is prohibited. The use of the BGP Encapsulation extended community provides a method to detect when this condition is violated but the actions to be taken are at the discretion of the operator and are outside the scope of this document.

6.3.2 Aliasing and Backup-Path

The Aliasing and the Backup-Path procedures for VXLAN/NVGRE encapsulation is very similar to the ones for MPLS. In case of MPLS, two different Ethernet AD routes are used for this purpose. The one used for Aliasing has a VPN scope and carries a VPN label but the one

used for Backup-Path has Ethernet segment scope and doesn't carry any VPN specific info (e.g., Ethernet Tag and MPLS label are set to zero). The same two routes are used when VXLAN or NVGRE encapsulation is used with the difference that when Ethernet AD route is used for Aliasing with VPN scope, the Ethernet Tag field is set to VNI or VSID to indicate VPN scope (and MPLS field may be set to a VPN label if needed).

7 Support for Multicast

The E-VPN Inclusive Multicast BGP route is used to discover the multicast tunnels among the endpoints associated with a given VXLAN VNI or NVGRE VSID. The Ethernet Tag field of this route is used to encode the VNI or VSID. This route is tagged with the PMSI Tunnel attribute, which is used to encode the type of multicast tunnel to be used as well as the multicast tunnel identifier. The following tunnel types can be used for VXLAN/NVGRE:

- PIM-SSM Tree
- PIM-SM Tree
- BIDIR-PIM Tree
- Ingress Replication

Except for Ingress Replication, this multicast tunnel is used by the PE originating the route for sending multicast traffic to other PEs, and is used by PEs that receive this route for receiving the traffic originated by CEs connected to the PE that originated the route.

In the scenario where the multicast tunnel is a tree, both the Inclusive as well as the Aggregate Inclusive variants may be used. In the former case, a multicast tree is dedicated to a VNI or VSID. Whereas, in the latter, a multicast tree is shared among multiple VNIs or VSIDs. This is done by having the NVEs advertise multiple Inclusive Multicast routes with different VNI or VSID encoded in the Ethernet Tag field, but with the same tunnel identifier encoded in the PMSI Tunnel attribute.

8 Inter-AS

For inter-AS operation, two scenarios must be considered:

- Scenario 1: The tunnel endpoint IP addresses are public
- Scenario 2: The tunnel endpoint IP addresses are private

In the first scenario, inter-AS operation is straight-forward and follows existing BGP inter-AS procedures. However, in the first scenario where the tunnel endpoint IP addresses are public, there may

be security concern regarding the distribution of these addresses among different ASes. This security concern is one of the main reasons for having the so called inter-AS "option-B" in MPLS VPN solutions such as E-VPN.

The second scenario is more challenging, because the absence of the MPLS client layer from the VXLAN encapsulation creates a situation where the ASBR has no fully qualified indication within the tunnel header as to where the tunnel endpoint resides. To elaborate on this, recall that with MPLS, the client layer labels (i.e. the VPN labels) are downstream assigned. As such, this label implicitly has a connotation of the tunnel endpoint, and it is sufficient for the ASBR to look up the client layer label in order to identify the label translation required as well as the tunnel endpoint to which a given packet is being destined. With the VXLAN encapsulation, the VNI is globally assigned and hence is shared among all endpoints. The destination IP address is the only field which identifies the tunnel endpoint in the tunnel header, and this address is privately managed by every data center network. Since the tunnel address is allocated out of a private address pool, then we either need to do a lookup based on VTEP IP address in context of a VRF (e.g., use IP-VPN) or terminate the VXLAN tunnel and do a lookup based on the tenant's MAC address to identify the egress tunnel on the ASBR. This effectively mandates that the ASBR to either run another overlay solution such as IP-VPN over MPLS/IP core network or to be aware of the MAC addresses of all VMs in its local AS, at the very least.

If VNIs/VSIDs have local significance, then the inter-AS operation can be simplified to that of MPLS and thus MPLS inter-AS option B and C can be leveraged in here. That's why the use of local significance VNIs/VSIDs (e.g., MPLS labels) are recommended for inter-AS operation of DC networks without gateways.

10 Acknowledgement

The authors would like to thank David Smith, John Mullooly, Thomas Nadeau for their valuable comments and feedback.

11 Security Considerations

This document uses IP-based tunnel technologies to support data plane transport. Consequently, the security considerations of those tunnel technologies apply. This document defines support for [VXLAN] and [NVGRE]. The security considerations from those documents as well

as [RFC4301] apply to the data plane aspects of this document.

As with [RFC5512], any modification of the information that is used to form encapsulation headers, to choose a tunnel type, or to choose a particular tunnel for a particular payload type may lead to user data packets getting misrouted, misdelivered, and/or dropped.

More broadly, the security considerations for the transport of IP reachability information using BGP are discussed in [RFC4271] and [RFC4272], and are equally applicable for the extensions described in this document.

If the integrity of the BGP session is not itself protected, then an imposter could mount a denial-of-service attack by establishing numerous BGP sessions and forcing an IPsec SA to be created for each one. However, as such an imposter could wreak havoc on the entire routing system, this particular sort of attack is probably not of any special importance.

It should be noted that a BGP session may itself be transported over an IPsec tunnel. Such IPsec tunnels can provide additional security to a BGP session. The management of such IPsec tunnels is outside the scope of this document.

12 IANA Considerations

13 References

13.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Y. Rekhter, Ed., T. Li, Ed., S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", January 2006.
- [RFC4272] S. Murphy, "BGP Security Vulnerabilities Analysis.", January 2006.
- [RFC4301] S. Kent, K. Seo., "Security Architecture for the Internet Protocol.", December 2005.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

11.2 Informative References

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (E-VPN)", draft-ietf-l2vpn-evpn-req-01.txt, work in progress, October 21, 2012.

[NVGRE] Sridhavan, M., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-01.txt, July 8, 2012.

[VXLAN] Dutt, D., et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-02.txt, August 22, 2012.

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-02.txt, work in progress, February, 2012.

[Problem-Statement] Narten et al., "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-01, September 2012.

[L3VPN-ENDSYSTEMS] Marques et al., "BGP-signaled End-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress, October 2012.

[NOV3-FRWK] Lasserre et al., "Framework for DC Network Virtualization", draft-ietf-nvo3-framework-01.txt, work in progress, October 2012.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

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

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

James Uttaro
AT&T
Email: uttaro@att.com

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

Ravi Shekhar
Juniper Networks
Email: rshekhar@juniper.net

Samer Salam
Cisco
Email: ssalam@cisco.com

Keyur Patel
Cisco
Email: Keyupate@cisco.com

Dhananjaya Rao
Cisco
Email: dhrao@cisco.com